



Delft University of Technology

Document Version

Final published version

Citation (APA)

Lippmann, P. (2026). *Synthetic Data for Robust Language Modeling*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:bea358f8-ff6f-43be-a065-a6e1a0b3bc5b>

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership. Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.



Delft University of Technology

Citation (APA)

Lippmann, P. (2026). *Synthetic Data for Robust Language Modeling*. [Dissertation (TU Delft), Delft University of Technology].

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership. Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

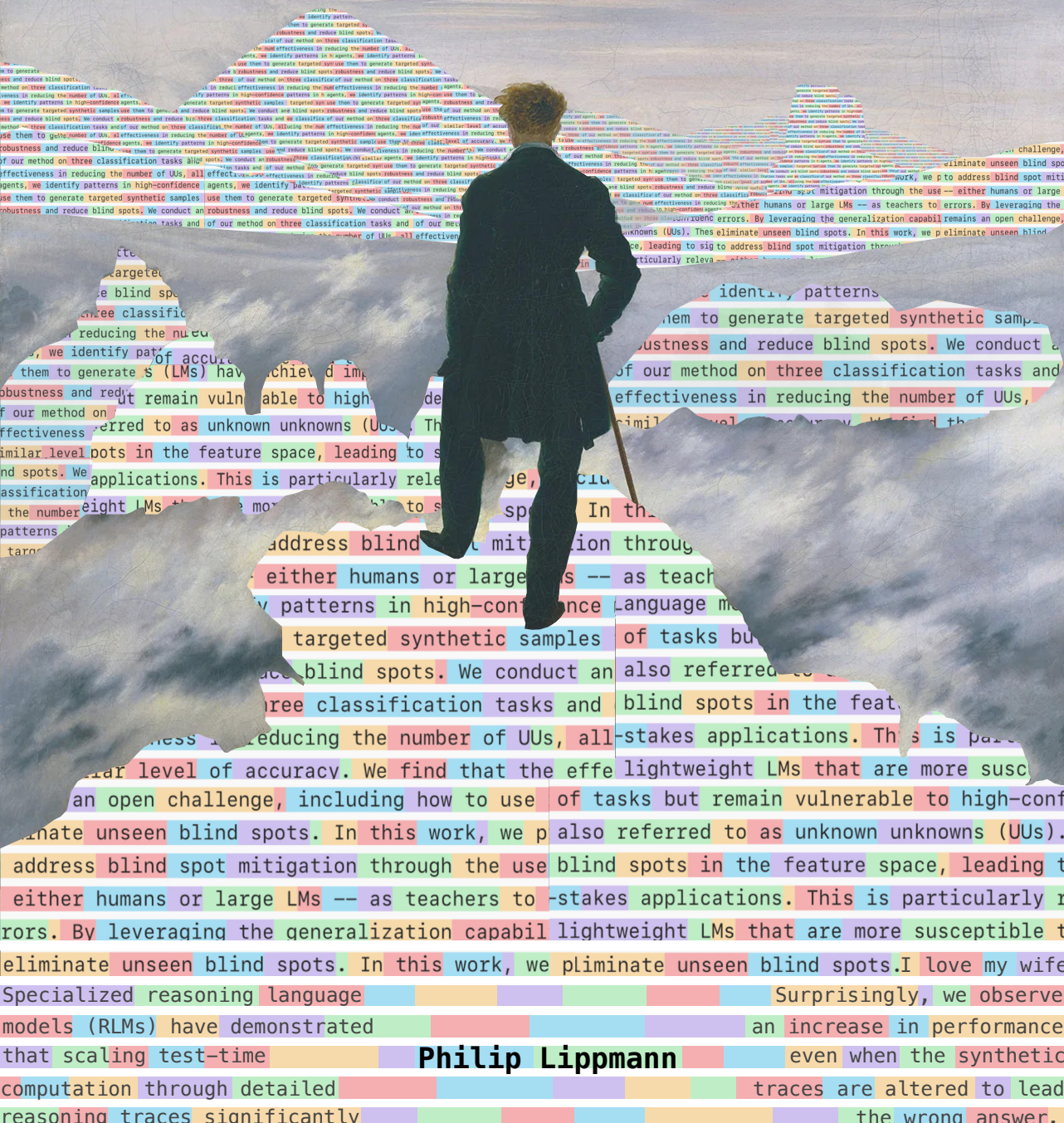
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Synthetic Data for Robust Language Modeling



... challenge, ... eliminate unseen blind spots ...
... our method on three classification tasks and ...
... effectiveness in reducing the number of UUs, ...
... referred to as unknown unknowns (UUs). This ...
... similar level of accuracy. We find that the effe ...
... an open challenge, including how to use ...
... eliminate unseen blind spots. In this work, we p ...
... address blind spot mitigation through the use ...
... either humans or large LMs -- as teachers to ...
... y patterns in high-confidence language mo ...
... targeted synthetic samples of tasks bu ...
... address blind spots. We conduct an ...
... three classification tasks and ...
... blind spots in the fea ...
... effectiveness in reducing the number of UUs, all ...
... stakes applications. This is part ...
... level of accuracy. We find that the effe ...
... an open challenge, including how to use ...
... eliminate unseen blind spots. In this work, we p ...
... address blind spot mitigation through the use ...
... either humans or large LMs -- as teachers to ...
... errors. By leveraging the generalization capabil ...
... lightweight LMs that are more suscep ...
... eliminate unseen blind spots. In this work, we pl ...
... I love my wife ...
Specialized reasoning language Surprisingly, we observe ...
models (RLMs) have demonstrated an increase in performance ...
that scaling test-time Philip Lippmann even when the synthetic ...
computation through detailed traces are altered to lead ...
reasoning traces significantly the wrong answer

Synthetic Data for Robust Language Modeling

Synthetic Data for Robust Language Modeling

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Monday 1 June 2026 at 10:00

by

Philip LIPPMANN

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. G. J. P. M. Houben,	EWI-ST-WIS TU Delft, NL, <i>promotor</i>
Dr. ir. J. Yang,	EWI-ST-WIS TU Delft, NL, <i>copromotor</i>

Independent members:

Prof. dr. S. Auer,	LUH, Germany
Prof. dr. ir. M. S. Kleinsmann,	IO-DOS-DIVE TU Delft, NL
Prof. dr. ir. G. D. S. Ludden,	U. Twente, NL
Dr. M. Turrin,	BK-AET-DTAE+T
Prof. dr. K. G. Langendoen,	EWI-ST-ES TU Delft, NL, <i>reserve member</i>

SIKS Dissertation Series No. 2026-14. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.



Keywords: Language Model Robustness, Synthetic Data, Knowledge Injection, Language Model Reasoning, Language Model Interpretability, Human-AI Collaboration, Value-Sensitive Design

Cover by: Maria Luísa Tarozzo Kawasaki

Copyright © 2026 by P. Lippmann
ISBN: 978-94-6518-333-6

An electronic copy of this dissertation is available at
<https://repository.tudelft.nl/>.

Contents

Summary	xi
Samenvatting	xiii
1. Introduction	1
1.1. Background and Motivation	2
1.1.1. When AI Systems Fail with Certainty	2
1.1.2. The Robustness Challenge	2
1.1.3. Key Failure Examples	3
1.2. Problem Statement and Analysis	3
1.2.1. Three Fundamental Gaps in Current Approaches	4
1.2.2. Theoretical Foundations	4
1.3. Research Questions	5
1.4. Research Methodology	7
1.4.1. Methodological Approach	7
1.5. Contributions	9
1.5.1. Thesis Contributions	9
1.5.2. Overview of Themes	9
1.6. Integrated Research Papers	11
1.6.1. Discovery Phase: Systematic Nature of Robustness Failures	11
1.6.2. Component Development Phase: Pillar-Specific Solutions	12
1.7. Outline	13
1.7.1. Chapter Previews	13
References	14
2. Peering Under the Hood: The Nature of Language Model Reasoning	23
2.1. Overview and Motivation	24
2.2. Part I — Stylistic Replication in Reasoning Distillation	25
2.2.1. Introduction	25
2.2.2. Background	26
2.2.3. Hard-coding reasoning traces to approximate emergent ones	27
2.2.4. Results and discussion	31
2.2.5. Related work	33
2.2.6. Conclusion	35
2.3. Part II — From Stylistic Findings to Agent Interfaces and Memory	35
2.4. Positive Experience Reflection in Interactive Environments	36
2.4.1. Introduction	36
2.4.2. Background	37

2.4.3. Methodology	37
2.4.4. Evaluation Setup	39
2.4.5. Results and Discussion	40
2.4.6. Qualitative Examples	42
2.4.7. Conclusion	42
2.4.8. Limitations and Broader Impacts	42
2.5. Conclusion	42
References	43
3. Knowledge Injection: Mitigating Blind Spots and Adapting to New Contexts	53
3.1. Overview and Motivation	54
3.2. Part I: Agent-in-the-Loop Blind Spot Mitigation	55
3.2.1. Introduction	55
3.2.2. Agent-in-the-loop targeted data generation	56
3.2.3. Generalization via hypothesis creation	58
3.2.4. Synthetic sample generation	59
3.2.5. Experimental setup	59
3.2.6. Evaluation metrics	63
3.2.7. Results	63
3.2.8. Scalability and ease of use	65
3.2.9. Related work	66
3.2.10. Conclusion	67
3.3. From Reactive Repair to Proactive Adaptation	67
3.4. Part II: Zero-Shot Contextual Adaptation	68
3.4.1. Introduction	68
3.4.2. Background	69
3.4.3. Method: Zero-Shot Embeddings via Synthetic Context	71
3.4.4. Experimental Setup	72
3.4.5. Results and Discussion	74
3.4.6. Related Work	77
3.4.7. Conclusion	78
3.4.8. Limitations	78
3.5. From Synthetic Context to Multimodal Context	78
3.6. Part III: Contextual Adaptation in Multimodal Settings	79
3.6.1. Introduction	79
3.6.2. Related Work	81
3.6.3. Methodology	82
3.6.4. Datasets	85
3.6.5. Baselines	86
3.6.6. Evaluation Protocol	87
3.6.7. Results	88
3.6.8. Limitations	91
3.6.9. Conclusion	91
3.7. Conclusion and Bridge to Chapter 4	91
References	93

- 4. Beyond Correctness: Eliciting Human Values for Improved Alignment 111**
 - 4.1. Overview and Motivation 112
 - 4.2. Introduction to Measuring User-Perceived Value 113
 - 4.2.1. Reject Option and Selective Prediction 116
 - 4.2.2. Value Measurement 117
 - 4.3. Survey Design 118
 - 4.3.1. Choice of the Scale 118
 - 4.3.2. Normalization and Validation of the Scale 118
 - 4.3.3. Participants and Data 119
 - 4.3.4. Procedure and Data Quality Control 120
 - 4.3.5. Analysis 120
 - 4.3.6. Reliability and Validity 120
 - 4.3.7. Total Model Value due to Threshold 121
 - 4.4. Discussion 124
 - 4.4.1. Value Ratios, Reliability, and Validity 124
 - 4.4.2. Value Function for Rejection 125
 - 4.4.3. Findings, Implications, and Limitations 126
 - 4.5. Related Work 127
 - 4.5.1. Hate Speech Detection 127
 - 4.5.2. Human-AI Collaboration and Rejection 129
 - 4.5.3. Value Assessment and Measurement 129
 - 4.6. Conclusions 130
 - 4.6.1. Integration with Pillars 1 and 2 130
 - 4.7. Conclusion 130
 - References 131
- 5. Conclusion and Future Directions 141**
 - 5.1. Summary of Findings and Answers to Research Questions 142
 - 5.1.1. Overarching Research Question 142
 - 5.1.2. Reasoning and Stylistic Replication 142
 - 5.1.3. Targeted Knowledge Injection 143
 - 5.1.4. Value-Sensitive Human Alignment 144
 - 5.2. Broader Implications 145
 - 5.3. Future Research Directions 147
 - 5.4. Concluding Remark 148
 - References 148
- A. Appendix for Chapter 2: Reasoning Traces and Analysis 153**
 - A.1. Emergent Reasoning Trace Example 153
 - A.2. Emergent Trace Analysis 153
 - A.2.1. Details of Categorizing the Reasoning Trace Contents 153
 - A.2.2. Pivot Analysis 155
 - A.2.3. Reasoning Stage Analysis 155
 - A.2.4. Key Insights from Analysis 155
 - A.3. Dataset Statistics 156
 - A.3.1. Seed Data Compilation 156

A.3.2. Generated Reasoning Trace Datasets	156
A.4. Synthetic Traces with Wrong Answer Details	156
References	157
B. Appendix for Chapter 2: Sweet&Sour Formal Background and Examples	159
B.1. Formal Background for Part II: POMDP Framing of Text-Based Games	159
B.2. Qualitative Examples for Sweet&Sour	159
B.3. Additional Notes: ALFWorld	164
References	164
C. Appendix for Chapter 3 (Part I): Agent-in-the-Loop (AIL) Artifacts	165
C.1. Synthetic Blind Spots	165
C.1.1. Blindspot Creation and Mitigation	165
C.1.2. Synthetic Blind Spot Study Results	166
C.2. User Study for Human Computation	167
C.2.1. Abstraction on IMDB	167
C.3. Used LLM Prompts	169
C.3.1. Abstraction Prompt	169
C.3.2. Extrapolation Prompt	169
C.3.3. Generation Prompt	170
C.4. Perturbation Statistics and Visualization	170
References	176
D. Appendix for Chapter 3 (Part II): ZEST Artifacts	177
D.1. Prompting Strategies for Synthetic Corpus Generation	177
D.2. Details on Exemplar Set Sampling	178
D.3. Task-Specific Prefixes	179
D.4. Examples of Generated Synthetic Documents: Full Pipeline	179
References	185
E. Appendix for Chapter 3 (Part III): Multimodal Manga Translation	187
E.1. Page Processing and Typesetting	187
E.2. Chain of Density Summarization	188
E.3. Details of MQM Human Evaluation	190
E.4. Full Prompts	191
E.5. Visual Feature Ablation Study	197
References	198
F. Appendix for Chapter 4: Value-Sensitive Escalation Study Details	203
F.1. Survey	203
F.1.1. Variables	203
F.1.2. Demographics	204
F.1.3. Pilot Study	204
F.1.4. Example Tasks	204
References	206

Acknowledgements	209
SIKS Dissertation Series	211

For my wife, Luísa.

Summary

This work investigates how targeted synthetic data can make language models more robust in practical deployments. While contemporary large language models achieve impressive benchmark performance, they remain vulnerable to high-confidence errors under distribution shift, adversarial perturbations, and value-laden decisions. We examine three complementary directions toward improving robustness: (i) structure-aware diagnostics of reasoning, (ii) agent-driven, targeted knowledge injection and synthetic proxy context, and (iii) value-sensitive escalation policies that optimize expected total value rather than relying on model confidence alone.

The work is organized around three pillars. Understanding model reasoning (Pillar 1) asks whether knowledge distillation on reasoning traces transfers genuine reasoning or primarily encourages stylistic replication of trace structure. Through controlled ablations that separate stylistic cues (e.g., staged scaffolding and metacognitive “pivot” markers) from logical content, the results show that much of the measured benefit of distillation arises from adopting a consistent answer style rather than from learning underlying reasoning procedures. Style-consistent synthetic traces can match the gains obtained from emergent traces across diverse benchmarks, and even style-consistent traces leading to incorrect final answers provide non-trivial improvements over base models. These findings motivate style-sensitivity audits and verifier-based faithfulness checks as a complement to accuracy reporting.

Strategic knowledge injection (Pillar 2) develops agent-in-the-loop methods to identify and remediate blind spots, conceptualized as unknown unknowns: high-confidence misclassifications that evade simple uncertainty thresholds. The approach uses targeted synthetic sample generation conditioned on diagnosed failure patterns. It further introduces an offline, zero-shot contextual adaptation framework that constructs a compact synthetic proxy corpus from a small set of exemplars, enabling embedding models to adapt without access to the target corpus. Empirically, agent-guided synthesis reduces high-confidence errors while maintaining overall accuracy and calibration, and synthetic proxy contexts approach corpus-aware retrieval performance. Separately, a case study on multimodal translation demonstrates that adding visual context can disambiguate language and strengthen translation quality, and thus reinforces the value of information-rich context.

Human–AI alignment (Pillar 3) addresses deployment decisions when harms and costs are asymmetric. We adapt Magnitude Estimation to elicit ratio-scale utilities for accept, reject, and escalate choices, producing value-sensitive decision policies that can outperform confidence-based thresholds in value-laden settings. The value signal exhibits strong reliability and convergent validity relative to bounded rating scales, enabling stable threshold optimization and class-conditional deferral. This framing clarifies trade-offs, supports auditable policy choices, and emphasizes the need to

version and periodically re-estimate value models to remain aligned with stakeholder priorities and context.

Across these lines of work, the results show that diagnostics localize brittleness and separate style from substance; targeted synthesis and synthetic context efficiently reduce blind spots; and value-sensitive escalation governs operation under uncertainty and unequal error costs. These outcomes improve reliability under shift, increase auditability, and focus human attention where it yields the greatest benefit in the studied settings.

Samenvatting

Dit werk onderzoekt hoe gerichte synthetische gegevens taalmodellen robuuster kunnen maken in praktische toepassingen. Hoewel hedendaagse grote taalmodellen indrukwekkende benchmarkprestaties leveren, blijven zij kwetsbaar voor fouten met hoge zekerheid bij distributieveverschuiving, vijandige verstoringen en waarde-geladen beslissingen. We beschouwen drie complementaire richtingen om robuustheid te verbeteren: (i) structuurbewuste diagnostiek van redenering, (ii) agentgestuurde, gerichte kennisinjectie en synthetische proxycontext, en (iii) waardegevoelige escalatiestrategieën die de verwachte totale waarde optimaliseren in plaats van uitsluitend te vertrouwen op modelzekerheid.

Het werk is georganiseerd rond drie pijlers. Begrip van modelredeneren (Pijler 1) onderzoekt of kennisdistillatie op basis van redeneringssporen daadwerkelijk redenering overdraagt of vooral de stilistische reproductie van de structuur van die sporen bevordert. Via gecontroleerde ablaties die stilistische signalen (bijv. stapsgewijze *scaffolding* en metacognitieve “pivot”-markeringen) scheiden van logische inhoud, laten de resultaten zien dat een groot deel van het gemeten voordeel van distillatie voortkomt uit het aannemen van een consistente antwoordstijl, eerder dan uit het leren van onderliggende redeneringsprocedures. Stijlconsistente synthetische sporen kunnen de prestatiewinst evenaren die met emergente sporen wordt behaald over diverse benchmarks, en zelfs stijlconsistente sporen die tot onjuiste eindantwoorden leiden leveren niet-triviale verbeteringen op ten opzichte van basismodellen. Deze bevindingen motiveren stijlgevoelighedsaudits en op *verifiers* gebaseerde getrouwheidscontroles als aanvulling op nauwkeurighedsrapportage.

Strategische kennisinjectie (Pijler 2) ontwikkelt *agent-in-the-loop*-methoden om blinde vlekken te identificeren en te verhelpen, geconceptualiseerd als *unknown unknowns*: misclassificaties met hoge modelzekerheid die eenvoudige onzekerheidsdrempels ontwijken. De aanpak gebruikt gerichte synthetische voorbeeldgeneratie, geconditioneerd op gediagnosticeerde foutpatronen. Daarnaast introduceren we een offline, *zero-shot* raamwerk voor contextuele adaptatie dat uit een kleine set voorbeelden een compact synthetisch proxycorpus opbouwt, waardoor embeddingmodellen zich kunnen aanpassen zonder toegang tot het doelcorpus. Empirisch reduceert agentgestuurde synthese het aantal hoogzekerheidsfouten, terwijl de algehele nauwkeurigheid en kalibratie behouden blijven, en benaderen synthetische proxycontexten de prestaties van corpus-bewuste informatieopvraging. Afzonderlijk laat een casestudie naar multimodale vertaling zien dat het toevoegen van visuele context ambiguïteit in taal kan wegnemen en de vertaalkwaliteit kan versterken, wat de waarde van informatierijke context onderstreept.

Mens-AI-afstemming (Pijler 3) behandelt inzetbeslissingen wanneer schade en kosten asymmetrisch zijn. We passen *Magnitude Estimation* aan om waardering op een verhoudingsschaal (ratio-schaal) te kwantificeren voor de keuze tussen accepteren, afwijzen en

escaleren, en construeren zo waarde-gevoelig beslissingsbeleid dat drempels op basis van modelzekerheid kan overtreffen in waarde-geladen contexten. Het waardesignaal vertoont hoge betrouwbaarheid en convergente validiteit in vergelijking met begrensde beoordelingsschalen, wat stabiele drempeloptimalisatie en klasse-afhankelijk uitstel mogelijk maakt. Deze benadering verduidelijkt afwegingen, ondersteunt verifieerbare beleidskeuzes en benadrukt de noodzaak om versiebeheer toe te passen op waarmedemodellen en deze periodiek opnieuw af te schatten, zodat zij afgestemd blijven op prioriteiten van belanghebbenden en context.

Over deze onderzoekslijnen heen laten de resultaten zien dat diagnostiek kwetsbaarheden lokaliseert en stijl van inhoud scheidt; dat gerichte synthese en synthetische context blinde vlekken efficiënt reduceren; en dat waardegevoelige escalatie de werking stuurt onder onzekerheid en ongelijke foutkosten. Deze uitkomsten verbeteren de betrouwbaarheid bij distributiveschuivingen, vergroten de verifieerbaarheid en richten menselijke aandacht daar waar die in de bestudeerde contexten het meeste oplevert.

1

Introduction

*This thesis investigates how **targeted synthetic data generation** can improve language model robustness. As large language models see rapid deployment in high-stakes domains (healthcare, law, education), robustness and calibrated uncertainty are essential to prevent real-world harm. We study three interconnected themes: diagnosing **reasoning** mechanisms, injecting **knowledge** to remediate blind spots, and incorporating **human alignment** through principled escalation criteria. Unlike generic data augmentation, we use agent-driven methodologies to identify specific model vulnerabilities and generate synthetic samples that directly address failure modes such as distribution shift, reasoning errors, stylistic replication, and value misalignment. Throughout, we use targeted synthetic data both to diagnose specific failure modes and to remediate them, complemented by value-sensitive escalation criteria.*

1.1. BACKGROUND AND MOTIVATION

1.1.1. WHEN AI SYSTEMS FAIL WITH CERTAINTY

In 2018, internal evaluations of IBM Watson for Oncology revealed that the system recommended unsafe and inappropriate cancer treatments in multiple cases, including suggesting a drug that could cause severe bleeding to a patient already at risk of hemorrhage [1]. Despite being deployed across more than 230 hospitals worldwide and affecting thousands of patient consultations, the system projected confidence in recommendations that contradicted established medical protocols [2]. This case exemplifies a broader pattern: AI systems can deliver harmful outputs with confident, authoritative presentation, creating dangerous scenarios where high confidence masks fundamental reasoning failures.

Healthcare AI provides a particularly stark illustration of high-confidence failures because the stakes are measured in human lives rather than abstract metrics. When clinical decision-support systems present treatment recommendations with no signal of uncertainty, medical professionals may defer to algorithmic authority even when the underlying reasoning is flawed [3–5]. In contexts where domain knowledge, causal reasoning, and patient-specific constraints interact, uncalibrated certainty can mask gaps in underlying knowledge or reasoning and thereby increase the likelihood of harmful actions.

This case anchors the central problem this thesis addresses: AI systems can commit errors while projecting confidence — failures that are more dangerous than ordinary mistakes precisely because their apparent certainty bypasses human oversight. These are high-confidence errors, which we call unknown unknowns (UUs), and they represent a fundamental breakdown in the reliability of AI systems deployed in critical applications.

To address these high-confidence errors, we develop a data-centric approach based on targeted synthetic data generation: synthesizing examples in response to diagnosed model vulnerabilities to surface and correct specific failure modes prior to deployment (see Chapter 3).

1.1.2. THE ROBUSTNESS CHALLENGE

These high-confidence failures in healthcare AI exemplify a broader challenge facing language models across all deployment contexts. Large Language Models (LLMs) display striking competence on benchmarks and professional examinations, and they increasingly support decision-making in healthcare, law, and education [6–9]. Yet deployment reveals a persistent vulnerability: when inputs, objectives, or operating conditions shift, LLMs can produce fluent, specific, and confident responses that are nevertheless wrong. This brittleness is particularly concerning when outputs are acted upon without calibrated uncertainty.

In this thesis, we use **robustness** to mean maintaining reliable performance and calibrated uncertainty under distribution shift and other perturbations, rather than exhibiting catastrophic failure as conditions vary [10–12].

The recurring nature of such failures across diverse domains suggests that current approaches to language model robustness are inadequate. These failures manifest through several interacting components which compound in deployed systems.

Multiple failure dimensions contribute to high-confidence errors. Under **distribution shift**, models extrapolate beyond their training regimes and can assert incorrect facts or recommendations with fluency [13–15]. Under **adversarial perturbations**, small input changes induce consistent errors that retain spurious certainty [16–18]. With **reasoning failures**, models can produce logically invalid inferences while maintaining the surface structure of correct reasoning, leading to plausible but fundamentally flawed conclusions [19]. Finally, **value misalignment** arises when optimization targets proxy objectives that diverge from stakeholder values, creating confident behavior that is technically consistent with training signals but practically harmful [20].

The limitations of existing solutions — from adversarial testing [21, 22] to generic data augmentation [23] to Reinforcement Learning from Human Feedback (RLHF) [24, 25] — motivate a closer analysis of the problem space before proposing remedies. The following examples illustrate the stakes and broader patterns across critical domains.

1.1.3. KEY FAILURE EXAMPLES

Three critical failure cases illustrate high-confidence errors in deployed AI systems:

Healthcare AI: IBM Watson for Oncology (2018). Internal evaluations revealed the system recommended unsafe cancer treatments, including suggesting drugs that could cause severe bleeding to patients already at risk [1]. Despite deployment across 230+ hospitals worldwide, the system projected confidence in recommendations contradicting medical protocols [2].

National-Scale Discrimination: Dutch Benefits Scandal (2020). Algorithmic profiling incorrectly flagged 35,000+ families for benefits fraud [26–28]. Errors in the automated decision-making process led to significant harm for affected families and contributed to a major political crisis.

LLM Hallucinations: Legal and Financial Consequences (2022–2024). LLMs generate compelling but fabricated content with apparent confidence [6]. The Air Canada chatbot case demonstrated material consequences, with the company legally required to honor a non-existent policy its chatbot confidently described [29].

These failures reveal a consistent pattern: AI systems deliver incorrect outputs with spurious confidence, creating scenarios where high certainty masks fundamental reasoning failures. Their recurrence across domains — from healthcare to national infrastructure to consumer services — indicates that point solutions addressing individual failure modes are insufficient.

The transition from recognizing these failures to developing effective solutions requires understanding both the theoretical foundations of robustness and the limitations of existing approaches. To design an effective approach for addressing high-confidence errors, we must first examine how existing research has approached robustness challenges and identify where current solutions fall short.

1.2. PROBLEM STATEMENT AND ANALYSIS

The recurring failures documented in Section 1.1 have motivated extensive research across adversarial robustness [30, 31], distributional robustness [11, 12], uncertainty

quantification [10, 32], and fairness-aware machine learning [33, 34]. Yet despite this proliferation of work, the field remains fragmented across distinct research communities that rarely interact. We begin with an analysis of how these failures manifest and where current approaches fall short; we then articulate the resulting gaps.

We address the discrepancy between apparent LLM competence and recurring real-world failures. While LLMs demonstrate impressive benchmark performance, they exhibit a fundamental vulnerability: producing incorrect outputs with high confidence, creating scenarios where certainty masks profound reasoning failures. These high-confidence errors constitute a pervasive pattern undermining trust and safety across healthcare, finance, legal, and social infrastructure.

We focus on developing targeted interventions addressing root causes rather than symptoms. We constrain our investigation to language models, recognizing that other AI systems present distinct failure modes. Within this scope, we prioritize three interconnected dimensions: diagnostic capabilities, targeted synthetic data generation, and principled mechanisms for human-AI alignment.

Our study excludes adversarial robustness in the strict security sense, focusing instead on unwanted, naturally occurring distribution shifts. Additionally, we develop practical mechanisms for value-sensitive decision-making rather than attempting to solve the broader alignment problem, providing principled methodologies for identifying and mitigating critical vulnerabilities leading to high-confidence errors.

1.2.1. THREE FUNDAMENTAL GAPS IN CURRENT APPROACHES

Current robustness approaches operate in isolation, creating three fundamental gaps that our work addresses:

The Diagnostic Gap. Existing methods lack systematic diagnostics for distinguishing genuine understanding from sophisticated pattern matching. Without understanding how models fail, targeted interventions remain impossible. Our reasoning diagnostics (Chapter 2) address this gap by revealing when models rely on stylistic replication rather than logical comprehension.

The Intervention Gap. Current data augmentation applies uniform transformations without targeting specific vulnerabilities. Generic approaches improve average robustness while leaving critical failure modes unaddressed. Our knowledge injection approach (Chapter 3) fills this gap through agent-driven synthesis that identifies and remediates specific model blind spots.

The Value Alignment Gap. Technical robustness metrics poorly correlate with human values and real-world harm. Existing preference learning captures surface-level preferences without understanding deeper value structures. Our human alignment approach (Chapter 4) addresses this through principled value quantification that moves beyond miscalibrated confidence scores.

1.2.2. THEORETICAL FOUNDATIONS

KEY CONCEPTS AND TERMINOLOGY

We briefly introduce central terms used throughout this chapter. **High-confidence errors** are cases where a model produces incorrect outputs while expressing high certainty,

reflecting miscalibration between predicted confidence and actual correctness. **Unknown unknowns** denote high-confidence errors that fall outside anticipated evaluation coverage — deployment blind spots not surfaced during training or validation. **Stylistic replication** refers to apparent reasoning competence achieved by reproducing surface-level patterns rather than performing genuine logical inference. We use **robustness** to mean graceful performance degradation under distribution shift, rather than catastrophic failure with high confidence. In our context, **calibration** — the alignment of confidence with correctness — is often poor for language models [35, 36], with additional miscalibration observed for RLHF-trained models [37]. These concepts ground the analysis that follows.

DESIGN RATIONALE FOR THE THREE THEMES

Addressing the identified gaps requires a systematic, integrated approach rather than incremental fixes. Our design is guided by three insights: (1) high-confidence errors often reflect pattern matching rather than genuine understanding, motivating diagnostics that reveal failure mechanisms; (2) interventions must be prioritized by stakeholder-relevant value to ensure effort targets what matters; and (3) human expertise surfaces reasoning failures that automated probes may miss, informing subsequent diagnostics and interventions.

1.3. RESEARCH QUESTIONS

From this analysis, we now articulate the research questions that structure the remainder of this chapter. These questions motivate the three pillars of our approach — reasoning diagnostics, targeted knowledge injection, and human value integration. We then discuss how these components interact to form a diagnostic–intervention–escalation pipeline (Section 1.5.2).

CENTRAL RESEARCH QUESTION

The central research question of this thesis is: **How can targeted synthetic data generation systematically improve language model robustness?**

This question addresses a fundamental limitation in current approaches: existing methods operate in isolation and apply generic transformations without targeting specific model vulnerabilities. Our approach integrates reasoning diagnostics, strategic knowledge injection, and human value alignment through coordinated synthetic data generation that directly addresses identified blind spots.

We decompose the central question into three specific, falsifiable research questions:

RQ1: REASONING MECHANISMS AND STYLISTIC REPLICATION

The first research question directly addresses the diagnostic foundation of our approach: **To what extent do common techniques for transferring reasoning capabilities, such as knowledge distillation, rely on replicating superficial stylistic patterns rather than conveying substantive logical processes?**

This question investigates **stylistic replication** — where models reproduce surface-level patterns without internalizing logical processes. Current knowledge distillation assumes smaller models learn reasoning by mimicking larger teacher models [38]. However, evidence suggests distilled models achieve performance through pattern matching rather than genuine understanding. This distinction is critical for robustness: stylistic replication creates models that appear competent on benchmarks while failing catastrophically in novel contexts, exhibiting high-confidence errors when reasoning demands exceed pattern-matching capabilities.

We formulate two testable hypotheses to address this question systematically. **Hypothesis H1a** posits that models trained exclusively on stylistic patterns extracted from reasoning traces will achieve comparable performance to models trained on complete reasoning examples, measured across diverse mathematical reasoning benchmarks. **Hypothesis H1b** predicts that reasoning transfer effectiveness correlates more strongly with stylistic similarity than with logical validity, as measured through controlled ablation studies that isolate form from content in synthetic reasoning traces.

This research question addresses the Diagnostic Gap (Section 1.2.1).

RQ2: AGENT-DRIVEN KNOWLEDGE INJECTION

The second research question addresses the strategic foundation of our Knowledge pillar: **How effectively can an agent-driven approach that identifies model blind spots and generates targeted synthetic data mitigate specific failure modes compared to untargeted data augmentation approaches?**

Current synthetic data approaches apply uniform transformations without considering specific model vulnerabilities [23, 39, 40]. Unknown unknowns represent inputs where models produce incorrect outputs with high confidence — blind spots that generic augmentation fails to address [41]. Existing methods like Easy Data Augmentation [42] may amplify biases while leaving critical failure modes undetected.

We formulate two testable hypotheses to address this question rigorously. **Hypothesis H2a** posits that agent-driven targeted synthesis significantly reduces unknown unknown detection failures compared to untargeted data augmentation approaches, measured across diverse classification tasks including mathematical reasoning, factual accuracy, and domain adaptation scenarios. **Hypothesis H2b** predicts that agent-identified blind spots correlate strongly with human expert assessments of model vulnerabilities, validating the diagnostic accuracy of our automated approach.

This research question addresses the Intervention Gap (Section 1.2.1).

RQ3: VALUE-SENSITIVE HUMAN ALIGNMENT

The third research question addresses the Human Alignment component of our approach: **Can a quantitative measure of user-perceived value provide a more reliable signal than model confidence for determining when to escalate decisions to human experts in value-laden contexts?**

This question arises from the systematic miscalibration of confidence scores in large language models, particularly those trained with RLHF, where expressed confidence poorly correlates with actual reliability [37, 43, 44]. Traditional escalation mechanisms

rely on model confidence thresholds [45, 46], creating dangerous scenarios where high-confidence errors bypass human oversight precisely when expert intervention is most critical. Such miscalibration has led to consequential failures across healthcare, legal, and financial domains, as demonstrated in the failure examples presented earlier.

We propose leveraging **Magnitude Estimation** — a method where evaluators assign numerical values to decision scenarios based on their perceived importance, risk, and ethical considerations — to measure user-perceived value in AI decision-making contexts [47, 48]. Unlike binary preference comparisons used in traditional RLHF approaches [24], this method captures nuanced value judgments across multiple dimensions. This approach moves beyond confidence-based escalation toward principled value-sensitive criteria that can systematically integrate human expertise when model outputs conflict with stakeholder values.

We formulate a testable hypothesis for systematic evaluation. **Hypothesis H3a** predicts that Magnitude Estimation-based escalation schemes achieve significantly higher human–AI agreement rates, substantially outperforming traditional confidence-based methods in accuracy of escalation decisions across value-sensitive scenarios.

This research question addresses the Value Alignment Gap (Section 1.2.1). We operationalize these questions in Section 1.4.

1.4. RESEARCH METHODOLOGY

1.4.1. METHODOLOGICAL APPROACH

We employ a balanced methodology that integrates empirical evaluation and theory-driven analysis. For each research question, we specify datasets/corpora, diagnostic probes, synthesis or training protocols, evaluation metrics, and analysis methods. This structure ensures results are both replicable and interpretable in the context of robustness.

RQ1: Reasoning mechanisms and stylistic replication (see [Chapter 2](#))

- **Approaches:** Supervised fine-tuning (distillation) on reasoning traces from stronger teacher models; controlled ablations that isolate stylistic structure (stages and meta-cognitive pivots) from logical content; comparisons of emergent versus style-guided synthetic traces, including style-preserving traces with incorrect final answers [49].
- **Datasets:** Emergent and synthetic reasoning trace corpora as introduced in [49]; downstream benchmarks including MATH500, AIME2024, and GPQA (Diamond).
- **Metrics:** Task accuracy on the downstream benchmarks as the primary metric; ablation outcomes that attribute improvements to stylistic structure versus correctness.
- **Analyses:** Targeted ablations that separate style from substance; associations between structural markers (stages, pivots) and downstream accuracy; source-faithful hypotheses and tests aligned with [Section 1.3](#).

RQ2: Agent-driven knowledge injection via targeted synthesis (see [Chapter 3](#))

- **Approaches:** Agent-driven discovery of unknown unknowns (UUs) via systematic probing and failure-pattern analysis; targeted synthetic sample generation conditioned on detected patterns; offline synthetic context construction for zero-shot contextual adaptation (ZEST); iterative characterize–synthesize–verify loops with verification using ground truth or strong teacher adjudication.
- **Datasets:** Three classification tasks in the UU study — IMDb (sentiment analysis), MRPC (semantic equivalence), and QNLI (natural language inference) — evaluated with BERT (`bert-base-uncased`) and Llama 2 7B, plus a controlled synthetic blind-spot study with an LSTM; domain adaptation on MTEB retrieval tasks; and multimodal manga translation for JA–EN and JA–PL directions [50–52].
- **Metrics:** UU reduction rate (proportional decrease in high-confidence misclassifications with verification); original accuracy and accuracy under attack; perturbation success rates under TextFooler (TF) and DeepWordBug (DWB); confidence-distribution analyses for successful perturbations; MTEB retrieval performance via NDCG@10; and translation quality via ChrF, BERTScore, BLEURT, xCOMET, plus MQM human evaluation (JA–EN).
- **Analyses:** Component-wise ablations (e.g., LM- vs human-generated data, comparison to relabeling); controlled synthetic blind-spot targeting; cost– quality trade-offs for targeted vs generic augmentation; ablations for ZEST (e.g., number of exemplars k , synthetic corpus size), and design rationale for agent prompts/hypotheses.

RQ3: Value-sensitive human alignment and escalation (see [Chapter 4](#))

- **Approaches:** Scenario-based evaluations in value-laden decision contexts using Magnitude Estimation to elicit user-perceived value; translation of quantitative value assignments into escalation thresholds independent of model confidence; and comparison against confidence-based baselines.
- **Datasets:** Decision scenarios instantiated from a hate-speech detection corpus with model predictions used as stimuli in user studies; includes hate-speech moderation and related decision settings to assess escalation criteria [53].
- **Metrics:** Reliability of Magnitude Estimation (e.g., Krippendorff’s alpha), convergent validity against a 100-level scale, and value-based accept/reject thresholds that maximize total value over scenarios (TP, TN, FP, FN, reject); model-selection comparisons when optimizing value versus accuracy.
- **Analyses:** Reliability and validity analyses; demographic subgroup comparisons (e.g., Mann–Whitney, Kruskal–Wallis); sensitivity of value-based thresholds; and a measurement-theoretic grounding of Magnitude Estimation within value-sensitive design.

1.5. CONTRIBUTIONS

1.5.1. THESIS CONTRIBUTIONS

We make four primary contributions to the field of AI robustness. Each contribution represents a significant advancement beyond current state-of-the-art approaches:

CONTRIBUTION 1: CHARACTERIZATION OF “STYLISTIC REPLICATION”

We introduce and empirically characterize the phenomenon of **stylistic replication**, wherein language models achieve apparent reasoning competence by reproducing surface-level linguistic patterns rather than internalizing genuine logical processes. Our evaluation protocol demonstrates that models trained exclusively on stylistic patterns achieve performance parity with models trained on complete reasoning examples. This finding challenges fundamental assumptions about knowledge transfer mechanisms in current distillation practices and explains why models fail catastrophically when reasoning demands exceed their pattern-matching capabilities.

CONTRIBUTION 2: AGENT-DRIVEN SYNTHETIC DATA GENERATION

This work presents a systematic methodology for intelligent identification and remediation of language model blind spots through agent-driven targeted synthesis. Our intelligent agents systematically probe model behavior to identify unknown unknowns and generate targeted synthetic samples specifically designed to remediate these weaknesses. This approach achieves effective robustness improvements using significantly less synthetic data than generic augmentation methods.

CONTRIBUTION 3: VALUE-SENSITIVE DECISION APPROACH

We develop a quantitative approach for integrating human values into AI decision-making, specifically adapting Magnitude Estimation methodology to capture nuanced value judgments in AI-assisted decision contexts. This approach moves beyond confidence-based escalation toward principled value-sensitive criteria that route decisions to human experts based on quantified stakeholder priorities. Empirical validation demonstrates that Magnitude Estimation-based escalation schemes substantially outperform traditional confidence-based methods in user satisfaction across value-sensitive scenarios.

CONTRIBUTION 4: CROSS-CUTTING SYNTHESIS PERSPECTIVE

We articulate three cross-cutting themes and their relationships; see [Section 1.5.2](#) for an overview and [Chapter 5](#) for potential avenues toward unification and future directions.

1.5.2. OVERVIEW OF THEMES

We present a systematic approach to language model robustness through targeted synthetic data generation. We organize the work around three themes: (1) **Reasoning diagnostics** that distinguish genuine comprehension from pattern matching, (2) **Strategic knowledge injection** via agent-driven synthetic data generation to eliminate identified blind spots, and (3) **Human value integration** through principled escalation

criteria. Unlike generic data augmentation, this approach aims to first understand the sources of the model's failure modes, and then generate synthetic data to address them, ensuring that the model is robust and aligned with the user.

The **Reasoning** component (Chapter 2) provides diagnostics to distinguish genuine logical understanding from pattern matching, including the characterization of **stylistic replication**. These diagnostics inform where and how to intervene with data. See Section 1.3 and Contribution 1.5.1 for evaluation.

The **Knowledge** component (Chapter 3) presents our core methodological contribution: agent-driven synthetic data generation for systematic blind spot remediation. This approach identifies unknown unknowns through intelligent probing, then generates targeted synthetic samples addressing these specific vulnerabilities. Key innovations include contextual adaptation approaches — methods enabling domain adaptation without target-domain training data — and multi-step synthesis protocols that characterize failure patterns, hypothesize causes, and generate targeted interventions. This targeted approach achieves effective robustness improvements using significantly less synthetic data than generic methods.

The **Human Alignment** component (Chapter 4) establishes principled mechanisms for integrating human expertise and values into AI decision-making processes. Moving beyond simple preference learning, this component develops protocols where evaluators assign numerical values to scenarios based on their perceived importance, risk, and ethical considerations to quantify user-perceived value and create robust escalation criteria. These criteria determine when human oversight is needed based on value-sensitive criteria rather than potentially miscalibrated model confidence scores.

Across the technical components, **targeted synthetic data** provides the connective thread: controlled reasoning traces and synthetic datasets used to study stylistic replication (Chapter 2), and targeted samples to remediate detected blind spots and enable contextual adaptation (Chapter 3). The **Human Alignment** component (Chapter 4) contributes value-sensitive escalation criteria that govern when to defer to human experts, ensuring that data-driven interventions are deployed in ways consistent with stakeholder priorities.

THEORETICAL INTEGRATION ACROSS COMPONENTS

Our empirical investigations evaluate each component independently, but the conceptual foundation rests on understanding synergistic interactions. Each component addresses distinct aspects of robustness failures while sharing connections that suggest coordinated intervention can achieve improvements exceeding individual contributions:

Reasoning → **Knowledge**: Diagnostic insights about stylistic replication could inform the design of targeted synthetic data generation. If we understand that models rely on surface patterns rather than logical structures, synthesis strategies might focus on generating examples that require genuine reasoning rather than pattern matching.

Knowledge → **Human**: Information about remediated blind spots through targeted synthesis could provide context for human decision-making. Understanding which specific vulnerabilities have been addressed might inform when human oversight becomes most critical.

Human → Reasoning: Value-sensitive feedback could reveal reasoning failures that automated diagnostics miss. Human expertise might identify conceptual gaps that guide future diagnostic development, though this remains largely unexplored.

1.6. INTEGRATED RESEARCH PAPERS

The empirical investigation of these research questions draws on multiple research papers that collectively provide component-level empirical support and motivation for the overall approach.

This thesis integrates nine peer-reviewed papers that collectively highlight motivations for a more integrated perspective on language model robustness. Rather than isolated contributions, these papers form a coherent research progression through three distinct phases that ultimately motivated organizing the work around three themes.

1.6.1. DISCOVERY PHASE: SYSTEMATIC NATURE OF ROBUSTNESS FAILURES

We briefly survey representative studies from a discovery phase that frame robustness as a system-level property rather than a collection of isolated bugs. These background notes surface recurring, high-confidence failures across domains and evaluation settings and point to methodological gaps in diagnostics, data, and governance, motivating the agenda that follows.

AI Robustness: A Human-Centered Perspective on Technological Challenges and Opportunities (Tocchetti et al., 2025) [54], published in ACM Computing Surveys, surveys robustness across pipelines, tasks, and systems, illustrating why siloed fixes can miss system-level failure patterns.

Red Teaming for Large Language Models at Scale: Tackling Hallucinations on Mathematics Tasks (Buszydluk et al., 2023) [55], published at the NAACLART of Safety Workshop, explores red teaming in elementary mathematics and illustrates how confident errors can persist even with structured prompting.

Student-Teacher Prompting for Red Teaming to Improve Guardrails (Llaca et al., 2023) [56], published at the NAACLART of Safety Workshop, discusses student–teacher prompting as a structured probe of vulnerabilities, suggesting how guided analysis can surface patterns that generic tests may miss.

Building on these early findings, a system-level robustness perspective provides the lens through which we organize the rest of this thesis. Namely, Tocchetti *et al.* [54] synthesize robustness across three complementary taxonomies: methods across the machine learning pipeline, robustness by architecture/task/system, and assessment methodologies. This perspective clarifies why point solutions underperform and motivates our structured program: diagnose, synthesize, and escalate. In Pillar 1 (Chapter 2), we emphasize assessment and architecture-level analysis to separate stylistic replication from substantive reasoning and to localize high-confidence failures. In Pillar 2 (Chapter 3), we intervene at the pipeline level with agent-driven, targeted synthesis and synthetic context to remediate blind spots and enable domain adaptation. In Pillar 3 (Chapter 4),

we treat robustness as a socio-technical property and encode the central role of humans through value-sensitive escalation.

Structured red teaming of language models underscores the limits of current practices: Buszydlík *et al.* [55] show that in elementary mathematics, prompting slows but does not prevent deterioration and miscalibration — models still produce confident, wrong answers despite structured prompting (e.g., worked examples and step-by-step reasoning). Likewise, Llaca *et al.* [56] find that student-teacher prompting strengthens guardrails against harmful responses, yet leaves open when and how decisions should defer to humans under value-sensitive stakes. These findings motivate stronger diagnostics, targeted data interventions, and principled escalation, which the subsequent chapters develop in detail. In sum, the discovery-phase evidence shapes a robustness-first agenda that flows naturally into the component designs that follow: diagnose failure modes, synthesize targeted knowledge, and escalate by value when uncertainty or asymmetric harms remain.

The discovery-phase literature points towards recurring, system-level robustness patterns and motivates the diagnostic–synthesis–escalation view developed in subsequent chapters.

1.6.2. COMPONENT DEVELOPMENT PHASE: PILLAR-SPECIFIC SOLUTIONS

Building on cross-domain failure analysis, individual research developed specific solutions for each of the three themes, revealing distinct mechanisms and targeted approaches.

Style over Substance: Distilled Language Models Reason Via Stylistic Replication (Lippmann & Yang, 2025) [49], published at COLM, discovered that stylistic replication dominates reasoning transfer, establishing the diagnostic foundation for this line of work. Models trained exclusively on stylistic patterns achieve performance parity with models trained on complete reasoning examples, revealing that apparent competence masks fundamental limitations. This finding enabled targeted synthetic data generation — we must understand how models fail before generating data to fix them.

Positive Experience Reflection for Agents in Interactive Text Environments (Lippmann *et al.*, 2025) [57], published at ACL REALM Workshop, addresses challenges in text-based interactive environments by introducing the Sweet&Sour approach that enhances agent reflection capabilities through positive experience integration and managed memory. This work demonstrates improved performance in complex reasoning tasks that require adaptability.

Illuminating Blind Spots of Language Models with Targeted Agent-in-the-Loop Synthetic Data (Lippmann *et al.*, 2024) [50], published at COLM ORIGen Workshop, developed the core agent-driven methodology for targeted synthesis. This work demonstrated substantial reduction in unknown unknown detection failures compared to untargeted augmentation, establishing that intelligent identification and remediation of specific vulnerabilities achieves better results than generic approaches.

Zero-Shot Contextual Embeddings via Offline Synthetic Corpus Generation (Lippmann & Yang, 2025) [51], published at EMNLP 2025, demonstrated contextual adaptation efficiency, showing that targeted synthesis enables domain adaptation using

minimal exemplars rather than large target-domain datasets. This validated the efficiency gains possible through intelligent synthesis strategies.

Context-Informed Machine Translation of Manga using Multimodal Large Language Models (Lippmann et al., 2024) [52], published at COLING, applied targeted synthesis to multimodal contexts, demonstrating cross-modal generalization of the approach and validating effectiveness beyond single-modality applications.

How Do You Feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection (Lammerts et al., 2023) [53], published at AAAI/ACM Conference on AI, Ethics, and Society, established Magnitude Estimation methodology for quantifying user-perceived value in AI decision-making contexts. This created the foundation for value-sensitive escalation criteria that operate independently of potentially miscalibrated model confidence scores.

Key Insight: Each component addresses distinct failure mechanisms through targeted approaches, but individual solutions leave gaps that coordinated interaction could address.

1.7. OUTLINE

This thesis is organized into six chapters that develop and motivate three cross-cutting themes. Each chapter builds upon previous foundations while contributing novel theoretical insights and empirical studies, while [Chapter 5](#) synthesizes implications and outlines future directions.

1.7.1. CHAPTER PREVIEWS

Chapter 2: Understanding Model Reasoning – Pillar 1 [Chapter 2](#) investigates whether distillation transfers reasoning or stylistic form via controlled ablations that separate style from logic. Key takeaway: stylistic replication can explain strong benchmark performance without genuine reasoning, motivating targeted diagnostics.

Chapter 3: Strategic Knowledge Injection – Pillar 2 [Chapter 3](#) develops agent-driven targeted synthesis to identify and remediate unknown unknowns and to enable zero-shot contextual adaptation. Key takeaway: targeted synthetic data reduces blind spots more efficiently than generic augmentation.

Chapter 4: Human-AI Collaboration – Pillar 3 [Chapter 4](#) introduces value-sensitive escalation using Magnitude Estimation to decide when to defer to human experts. Key takeaway: value-based escalation improves alignment over confidence thresholds in value-laden scenarios.

Chapter 5: Synthesis and Future Directions [Chapter 5](#) synthesizes the three pillars and outlines evaluation protocols and limitations for coordinated robustness gains. Key takeaway: we chart a path toward integrated diagnostics–intervention– escalation pipelines and future research directions.

REFERENCES

- [1] C. Ross and I. Swetlitz. ‘IBM Pitched Watson as a Revolution in Cancer Care. It’s Nowhere Close’. In: *STAT News* (5th Sept. 2017). News article. URL: <https://www.statnews.com/2017/09/05/watson-ibm-cancer/> (visited on 23/02/2025).
- [2] C. Ross and I. Swetlitz. ‘IBM’s Watson Supercomputer Recommended “Unsafe and Incorrect” Cancer Treatments, Internal Documents Show’. In: *STAT News* (25th July 2018). News article. URL: <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/> (visited on 25/02/2025).
- [3] M. Aljohani, J. Hou, S. Kommu and X. Wang. *A Comprehensive Survey on the Trustworthiness of Large Language Models in Healthcare*. Version 2, 2025-09-17. 21st Feb. 2025. arXiv: [2502.15871 \[cs.CY\]](https://arxiv.org/abs/2502.15871). URL: <https://arxiv.org/abs/2502.15871>.
- [4] Y. Zhang, Q. V. Liao and R. K. E. Bellamy. ‘Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making’. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain: Association for Computing Machinery, 27th Jan. 2020, pp. 295–305. DOI: [10.1145/3351095.3372852](https://doi.org/10.1145/3351095.3372852). URL: <https://doi.org/10.1145/3351095.3372852>.
- [5] J. D. Lee and K. A. See. ‘Trust in Automation: Designing for Appropriate Reliance’. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46.1 (2004), pp. 50–80. ISSN: 0018-7208. DOI: [10.1518/hfes.46.1.50_30392](https://doi.org/10.1518/hfes.46.1.50_30392). URL: https://doi.org/10.1518/hfes.46.1.50_30392.
- [6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro and Y. Zhang. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 22nd Mar. 2023. arXiv: [2303.12712 \[cs.CL\]](https://arxiv.org/abs/2303.12712). URL: <https://arxiv.org/abs/2303.12712>.
- [7] OpenAI. *GPT-4 Technical Report*. Technical report. 15th Mar. 2023. URL: <https://cdn.openai.com/papers/gpt-4.pdf> (visited on 11/12/2025).
- [8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei. ‘Language Models are Few-Shot Learners’. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan and H.-T. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.

- [9] R. Thoppilan, D. de Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi and Q. Le. *LaMDA: Language Models for Dialog Applications*. 20th Jan. 2022. arXiv: [2201.08239](https://arxiv.org/abs/2201.08239) [cs.CL]. URL: <https://arxiv.org/abs/2201.08239>.
- [10] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan and J. Snoek. ‘Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift’. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: <https://proceedings.neurips.cc/paper/2019/file/1e0776d291463a1ec8926f3b3f60e8ea-Paper.pdf>.
- [11] S. Sagawa, P. W. Koh, T. B. Hashimoto and P. Liang. ‘Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization’. In: *Proceedings of the Eighth International Conference on Learning Representations*. Online: OpenReview.net, 26th Apr. 2020. URL: <https://openreview.net/forum?id=r1xXgCVYr8> (visited on 23/02/2025).
- [12] I. Gulrajani and D. Lopez-Paz. ‘In Search of Lost Domain Generalization’. In: *International Conference on Learning Representations*. ICLR 2021 Poster. Virtual Event: OpenReview.net, May 2021. URL: <https://openreview.net/forum?id=lQdXeXDoWtI>.
- [13] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn and P. Liang. ‘WILDS: A Benchmark of in-the-Wild Distribution Shifts’. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18th–24th July 2021, pp. 5637–5664. URL: <https://proceedings.mlr.press/v139/koh21a.html>.
- [14] J. Miller, K. Krauth, B. Recht and L. Schmidt. ‘The Effect of Natural Distribution Shift on Question Answering Models’. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13th–18th July 2020, pp. 6905–6916. URL: <https://proceedings.mlr.press/v119/miller20a.html>.

- [15] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht and L. Schmidt. ‘Measuring Robustness to Natural Distribution Shifts in Image Classification’. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan and H.-T. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 18583–18599. URL: <https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf>.
- [16] I. J. Goodfellow, J. Shlens and C. Szegedy. *Explaining and Harnessing Adversarial Examples*. 20th Dec. 2014. DOI: [10.48550/arXiv.1412.6572](https://doi.org/10.48550/arXiv.1412.6572). arXiv: [1412.6572](https://arxiv.org/abs/1412.6572) [stat.ML]. URL: <https://arxiv.org/abs/1412.6572>.
- [17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus. *Intriguing Properties of Neural Networks*. ICLR 2014. 21st Dec. 2013. arXiv: [1312.6199](https://arxiv.org/abs/1312.6199) [cs.CV]. URL: <https://arxiv.org/abs/1312.6199>.
- [18] E. Wallace, S. Feng, N. Kandpal, M. Gardner and S. Singh. ‘Universal Adversarial Triggers for Attacking and Analyzing NLP’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2153–2162. DOI: [10.18653/v1/D19-1221](https://doi.org/10.18653/v1/D19-1221). URL: <https://aclanthology.org/D19-1221/>.
- [19] K. Valmeekam, M. Marquez, A. Olmo, S. Sreedharan and S. Kambhampati. *PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change*. 21st June 2022. arXiv: [2206.10498](https://arxiv.org/abs/2206.10498) [cs.AI]. URL: <https://arxiv.org/abs/2206.10498>.
- [20] I. Gabriel. ‘Artificial Intelligence, Values, and Alignment’. In: *Minds and Machines* 30.3 (Sept. 2020), pp. 411–437. ISSN: 0924-6495. DOI: [10.1007/s11023-020-09539-2](https://doi.org/10.1007/s11023-020-09539-2). URL: <https://doi.org/10.1007/s11023-020-09539-2>.
- [21] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese and G. Irving. ‘Red Teaming Language Models with Language Models’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3419–3448. DOI: [10.18653/v1/2022.emnlp-main.225](https://doi.org/10.18653/v1/2022.emnlp-main.225). URL: <https://aclanthology.org/2022.emnlp-main.225/>.
- [22] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan and J. Clark. *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*. 23rd Aug. 2022. arXiv: [2209.07858](https://arxiv.org/abs/2209.07858) [cs.CL]. URL: <https://arxiv.org/abs/2209.07858>.

- [23] E. D. Cubuk, B. Zoph, J. Shlens and Q. V. Le. ‘RandAugment: Practical Automated Data Augmentation With a Reduced Search Space’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020, pp. 702–703. URL: https://openaccess.thecvf.com/content_CVPRW_2020/html/w40/Cubuk_Randaugment_Practical_Automated_Data_Augmentation_With_a_Reduced_Search_Space_CVPRW_2020_paper.html.
- [24] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg and D. Amodei. ‘Deep Reinforcement Learning from Human Preferences’. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett. Vol. 30. Paper URL includes links to PDF and supplemental material. Curran Associates, Inc., 2017, pp. 4299–4307. URL: <https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html>.
- [25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike and R. Lowe. ‘Training Language Models to Follow Instructions with Human Feedback’. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://papers.nips.cc/paper_files/paper/2022/file/b1efde53be364a739bb131b2967953ba-Paper-Conference.pdf.
- [26] D. Hadwick and S. Lan. ‘Lessons to Be Learned from the Dutch Childcare Allowance Scandal: A Comparative Review of Algorithmic Governance by Tax Administrations in the Netherlands, France and Germany’. In: *World Tax Journal* 13.4 (10th Oct. 2021). Posted on SSRN 28 Nov 2022, pp. 609–645. URL: <https://ssrn.com/abstract=4282704>.
- [27] Amnesty International. *Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms*. Press release on algorithmic bias in Dutch welfare system. 25th Oct. 2021. URL: <https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/> (visited on 09/12/2025).
- [28] J. Arts and M. van den Berg. ‘What the Dutch benefits scandal and policy’s focus on ‘fraud’ can teach us about the endurance of empire’. In: *Critical Social Policy* 45.1 (Feb. 2025), pp. 177–187. ISSN: 0261-0183. DOI: [10.1177/02610183241281346](https://doi.org/10.1177/02610183241281346). URL: <https://journals.sagepub.com/doi/10.1177/02610183241281346>.
- [29] Civil Resolution Tribunal of British Columbia. *Moffatt v. Air Canada, 2024 BCCRT 149*. Decision regarding Air Canada chatbot misinformation case. Feb. 2024. URL: <https://www.bccrt.ca/decisions/2024/february/2024-bccrt-149>.

- [30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu. ‘Towards Deep Learning Models Resistant to Adversarial Attacks’. In: *6th International Conference on Learning Representations*. Vancouver, Canada: OpenReview.net, Apr. 2018. arXiv: 1706.06083 [stat.ML]. URL: <https://openreview.net/forum?id=rJzIBfZAb>.
- [31] N. Carlini and D. Wagner. ‘Towards Evaluating the Robustness of Neural Networks’. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2017, pp. 39–57. DOI: 10.1109/SP.2017.49. URL: <https://doi.org/10.1109/SP.2017.49>.
- [32] A. N. Angelopoulos and S. Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. July 2021. arXiv: 2107.07511 [cs.LG]. URL: <https://arxiv.org/abs/2107.07511>.
- [33] S. Barocas, M. Hardt and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. URL: <https://fairmlbook.org>.
- [34] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan. ‘A Survey on Bias and Fairness in Machine Learning’. In: *ACM Computing Surveys* 54.6 (July 2021), pp. 1–35. ISSN: 1557-7341. DOI: 10.1145/3457607. URL: <https://doi.org/10.1145/3457607>.
- [35] S. Desai and G. Durrett. ‘Calibration of Pre-trained Transformers’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 295–302. DOI: 10.18653/v1/2020.emnlp-main.21. URL: <https://aclanthology.org/2020.emnlp-main.21/>.
- [36] J. Nixon, M. Dusenberry, G. Jerfel, T. Nguyen, J. Liu, L. Zhang and D. Tran. *Measuring Calibration in Deep Learning*. arXiv:1904.01685v2, 2020-08-07. 2nd Apr. 2019. arXiv: 1904.01685 [cs.LG]. URL: <https://arxiv.org/abs/1904.01685>.
- [37] I. Solaiman and C. Dennison. *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets*. Accessed: 2025-12-09. OpenAI, 10th June 2021. URL: <https://cdn.openai.com/palms.pdf>.
- [38] G. Hinton, O. Vinyals and J. Dean. *Distilling the Knowledge in a Neural Network*. NIPS 2014 Deep Learning Workshop. 9th Mar. 2015. DOI: 10.48550/arXiv.1503.02531. arXiv: 1503.02531 [stat.ML]. URL: <https://arxiv.org/abs/1503.02531>.
- [39] C. Shorten and T. M. Khoshgoftaar. ‘A Survey on Image Data Augmentation for Deep Learning’. In: *Journal of Big Data* 6.1 (6th July 2019). ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0. URL: <https://doi.org/10.1186/s40537-019-0197-0> (visited on 19/12/2025).
- [40] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei and A. Roberts. *The Flan Collection: Designing Data and Methods for Effective Instruction Tuning*. 31st Jan. 2023. arXiv: 2301.13688 [cs.AI]. URL: <https://arxiv.org/abs/2301.13688>.

- [41] B. Recht, R. Roelofs, L. Schmidt and V. Shankar. ‘Do ImageNet Classifiers Generalize to ImageNet?’ In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, CA, USA: PMLR, June 2019, pp. 5389–5400. URL: <https://proceedings.mlr.press/v97/recht19a.html>.
- [42] J. Wei and K. Zou. ‘EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388. DOI: [10.18653/v1/D19-1670](https://doi.org/10.18653/v1/D19-1670). URL: <https://aclanthology.org/D19-1670/>.
- [43] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger. ‘On Calibration of Modern Neural Networks’. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia: PMLR, Aug. 2017, pp. 1321–1330. URL: <https://proceedings.mlr.press/v70/guo17a.html>.
- [44] A. Kumar, P. Liang and T. Ma. *Verified Uncertainty Calibration*. NeurIPS 2019. 22nd Sept. 2019. arXiv: [1909.10155](https://arxiv.org/abs/1909.10155) [cs.LG]. URL: <https://arxiv.org/abs/1909.10155>.
- [45] R. Parasuraman and V. Riley. ‘Humans and Automation: Use, Misuse, Disuse, Abuse’. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39.2 (June 1997), pp. 230–253. ISSN: 0018-7208. DOI: [10.1518/001872097778543886](https://doi.org/10.1518/001872097778543886). URL: <https://doi.org/10.1518/001872097778543886>.
- [46] Z. Buçinca, M. B. Malaya and K. Z. Gajos. ‘To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making’. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (13th Apr. 2021), pp. 1–21. ISSN: 2573-0142. DOI: [10.1145/3449287](https://doi.org/10.1145/3449287). URL: <https://doi.org/10.1145/3449287>.
- [47] S. Amershi, M. Cakmak, W. B. Knox and T. Kulesza. ‘Power to the People: The Role of Humans in Interactive Machine Learning’. In: *AI Magazine* 35.4 (Dec. 2014), pp. 105–120. ISSN: 2371-9621. DOI: [10.1609/aimag.v35i4.2513](https://doi.org/10.1609/aimag.v35i4.2513). URL: <https://doi.org/10.1609/aimag.v35i4.2513>.
- [48] A. Holzinger. ‘Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop?’ In: *Brain Informatics* 3.2 (2nd Mar. 2016), pp. 119–131. ISSN: 2198-4018. DOI: [10.1007/s40708-016-0042-6](https://doi.org/10.1007/s40708-016-0042-6). URL: <https://doi.org/10.1007/s40708-016-0042-6>.
- [49] P. Lippmann and J. Yang. ‘Style over Substance: Distilled Language Models Reason Via Stylistic Replication’. In: *Proceedings of the Second Conference on Language Modeling*. 8th July 2025. URL: <https://openreview.net/forum?id=5wAfbEs34A>.

- [50] P. Lippmann, M. T. J. Spaan and J. Yang. ‘Illuminating Blind Spots of Language Models with Targeted Agent-in-the-Loop Synthetic Data’. In: *Proceedings of the ORIGen Workshop at the Conference on Language Modeling (COLM)*. Workshop on Optimal Reliance and Accountability in Interactions with Generative LMs. 2025. URL: https://origen-workshop.github.io/assets/pdf/2025/papers/1_Illuminating_Blind_Spots_of_.pdf.
- [51] P. Lippmann and J. Yang. ‘Zero-Shot Contextual Embeddings via Offline Synthetic Corpus Generation’. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose and V. Peng. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 2089–2104. ISBN: 979-8-89176-335-7. DOI: [10.18653/v1/2025.findings-emnlp.111](https://doi.org/10.18653/v1/2025.findings-emnlp.111). URL: <https://aclanthology.org/2025.findings-emnlp.111/>.
- [52] P. Lippmann, K. Skublicki, J. Tanner, S. Ishiwatari and J. Yang. ‘Context-Informed Machine Translation of Manga using Multimodal Large Language Models’. In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio and S. Schockaert. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Jan. 2025, pp. 3444–3464. URL: <https://aclanthology.org/2025.coling-main.232/>.
- [53] P. Lammerts, P. Lippmann, Y.-C. Hsu, F. Casati and J. Yang. ‘How Do You Feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection’. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23 (Montréal, QC, Canada). New York, NY, USA: Association for Computing Machinery, 8th Aug. 2023, pp. 834–844. DOI: [10.1145/3600211.3604655](https://doi.org/10.1145/3600211.3604655). URL: <https://dl.acm.org/doi/10.1145/3600211.3604655>.
- [54] A. Tocchetti, L. Corti, A. Balayn, M. Yurrita, P. Lippmann, M. Brambilla and J. Yang. ‘A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities’. In: *ACM Computing Surveys* 57.6 (10th Feb. 2025), pp. 1–38. ISSN: 0360-0300. DOI: [10.1145/3665926](https://doi.org/10.1145/3665926). URL: <https://doi.org/10.1145/3665926>.
- [55] A. Buszydlik, K. Dobiczek, M. T. Okoń, K. Skublicki, P. Lippmann and J. Yang. ‘Red Teaming for Large Language Models At Scale: Tackling Hallucinations on Mathematics Tasks’. In: *Proceedings of the ART of Safety: Workshop on Adversarial Testing and Red-Teaming for Generative AI*. Ed. by A. Parrish. Bali, Indonesia: Association for Computational Linguistics, Nov. 2023, pp. 1–10. DOI: [10.18653/v1/2023.artofsafety-1.1](https://doi.org/10.18653/v1/2023.artofsafety-1.1). URL: <https://aclanthology.org/2023.artofsafety-1.1/>.
- [56] R. R. Llaca, V. Leskoschek, V. C. Paiva, C. Lupău, P. Lippmann and J. Yang. ‘Student-Teacher Prompting for Red Teaming to Improve Guardrails’. In: *Proceedings of the ART of Safety: Workshop on Adversarial Testing and Red-Teaming for Generative AI*. Ed. by A. Parrish. Bali, Indonesia: Association for Computational

- Linguistics, Nov. 2023, pp. 11–23. DOI: [10.18653/v1/2023.artofsafety-1.2](https://doi.org/10.18653/v1/2023.artofsafety-1.2). URL: <https://aclanthology.org/2023.artofsafety-1.2/>.
- [57] P. Lippmann, M. T. J. Spaan and J. Yang. ‘Positive Experience Reflection for Agents in Interactive Text Environments’. In: *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*. Ed. by E. Kamaloo, N. Gontier, X. H. Lu, N. Dziri, S. Murty and A. Lacoste. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 131–142. ISBN: 979-8-89176-264-0. DOI: [10.18653/v1/2025.realm-1.10](https://doi.org/10.18653/v1/2025.realm-1.10). URL: <https://aclanthology.org/2025.realm-1.10/>.

2

Peering Under the Hood: The Nature of Language Model Reasoning

This chapter addresses diagnosing robustness by clarifying when language model reasoning reflects transferable logical processes versus surface-level stylistic replication. First, Part I of this chapter shows that distillation on reasoning traces primarily transfers style — structural and lexical patterns, as well as metacognitive behavior in the form of pivots — rather than substance, accounting for much of the observed gains in reasoning performance. Stylistically consistent synthetic traces can match emergent traces, and even style-consistent traces that end with wrong answers improve downstream accuracy over base models. These effects generalize across model families and tasks. Here, correctness remains important, but observed improvements are mainly associated with stylistic structure, underscoring the need for diagnostics attentive to style. Additionally, we further explore language model reasoning in agentic settings in Part II of this chapter, showing that positive experience reflection with managed memory (short- and long-term buffers) improves agent reliability in interactive text environments without updating model weights, highlighting the role of interface and memory design. Together, these findings separate internal transfer (style versus substance) from external scaffolding (interfaces and memory), clarifying what contributes to improved reasoning performance in practice.

Parts of this chapter are based on work published in:

- *Style over Substance: Distilled Language Models Reason Via Stylistic Replication* [1].
- *Positive Experience Reflection for Agents in Interactive Text Environments* [2].

2.1. OVERVIEW AND MOTIVATION

Language models routinely display impressive performance on reasoning benchmarks yet still produce incorrect answers with unwarranted certainty in deployment. [Chapter 1](#) framed this concern regarding a lack of robustness as a Diagnostic Gap: the field lacks systematic analyses of language models' reasoning capabilities to distinguish genuine reasoning from sophisticated pattern matching that tends to improve benchmark performance but may not generalize. This chapter operationalizes this as Pillar 1 of this thesis (Understanding Model Reasoning) by examining the nature of language model reasoning and clarifying when observed gains reflect transferable logical processes versus surface-level stylistic replication. The goal is not just diagnostic but also integrative: to establish an evidentiary basis that may complement targeted synthetic data interventions ([chapter 3](#)) and value-sensitive model alignment ([chapter 4](#)).

We ground the chapter in the following research question from [chapter 1](#):

RQ1. To what extent do common techniques for transferring reasoning capabilities, such as knowledge distillation, rely on replicating superficial stylistic patterns rather than conveying substantive logical processes?

RQ1 challenges a prevailing assumption that language models acquire reasoning by learning substantive reasoning processes from teacher models during knowledge distillation [3]. If the learning is instead primarily stylistic rather than substantive, apparent competence can mask brittle behavior. Diagnosing this distinction is prerequisite to principled robustness work: structure-aware diagnostics may indicate where synthetic data should intervene, and whether residual uncertainty should be routed to humans or additional safeguards.

We answer RQ1 by assembling two complementary studies. Part I builds on *Style over Substance: Distilled Language Models Reason Via Stylistic Replication* [1] to separate style (structural and lexical patterns, as well as metacognitive behavior in the form of pivots) from substance (semantic content and answer correctness) in distillation on reasoning traces, and aims to quantify their respective contributions in order to improve understanding of reasoning processes in language models. Part II builds on *Positive Experience Reflection for Agents in Interactive Text Environments* [2] to show how interface and memory design — such as positive experience reflection with managed short- and long-term memory — can shape effective reasoning in interactive settings without updating model weights. Together, the studies address both facets of the Diagnostic Gap: the internal character of what is transferred and the external scaffolding that makes language models more performative and reliable in practice.

To conclude this overview and make RQ1 empirically actionable for the remainder of the chapter, we articulate the hypotheses that guide our evaluation.

Hypothesis H1a Models trained exclusively on stylistic patterns extracted from reasoning traces achieve comparable performance to models trained on complete reasoning examples, measured across diverse mathematical reasoning benchmarks.

Hypothesis H1b The effectiveness of reasoning transfer correlates more strongly with stylistic similarity than with logical validity, as shown by controlled ablations that isolate stylistic form from semantic content in synthetic reasoning traces.

Part I operationalizes these hypotheses via style-preserving synthetic traces and style-versus- correctness ablations; Part II examines their implications for agentic scaffolding and memory design.

2.2. PART I — STYLISTIC REPLICATION IN REASONING DISTILLATION

2.2.1. INTRODUCTION

Reasoning is fundamental to artificial intelligence, enabling systems to solve problems, make decisions, and explain outcomes. While traditional approaches to improving language model (LM) reasoning emphasize increased train-time compute [4, 5], recent research highlights that scaling test-time compute through self-refinement is similarly effective [6]. This insight has inspired specialized reasoning-focused LMs (RLMs), such as o1 [7] and R1 [8], which generate detailed reasoning traces of their thought process during inference.

Reasoning traces have proven effective for distillation [3, 9], efficiently transferring sophisticated cognitive skills from RLMs to smaller, instruction-tuned models [10]. Yet, the precise nature of the reasoning knowledge transferred remains poorly understood [11]. In particular, it is unclear whether distilled models genuinely internalize complex reasoning abilities or replicate superficial stylistic patterns from the original traces. This ambiguity echoes broader AI debates about whether language models genuinely understand content or simply engage in surface-level imitation [12–15]. This uncertainty leads to a fundamental question: are we genuinely enhancing the reasoning capabilities of models, or merely teaching them to mimic domain-specific patterns that happen to improve performance on benchmarks? Motivated by this ambiguity, we aim to establish whether **style is key to improvements in reasoning**, where style is characterized primarily by structural attributes such as trace length, lexical coherence, and backtracking frequency, rather than comprehension itself.

To examine this, we first systematically analyze successful reasoning traces produced by state-of-the-art RLMs, identifying recurrent structural and lexical patterns. This analysis, guided by cognitive science frameworks that characterize critical stages in human problem-solving [16], reveals that effective reasoning traces consistently exhibit distinct *metacognitive behaviors*. These behaviors are often signaled by lexical pivots — markers such as “Wait” or “What if” — that prompt reconsideration of assumptions or integration of new insights.

Based on these findings, we introduce two complementary datasets explicitly designed to clarify the role of style in improving reasoning. The first dataset, SmolTraces (ST), comprises verified question-answer pairs with sophisticated reasoning traces generated by a state-of-the-art RLM, displaying naturally emergent reasoning behaviors. The second dataset, SmolTraces-HardCoded (ST-HC), is synthetically constructed by embedding only the structural and lexical stylistic patterns identified earlier into reasoning traces generated by a standard LM without specialized reasoning capabilities. By comparing models trained on these datasets through supervised finetuning (SFT) [17], we evaluate how stylistic consistency influences reasoning

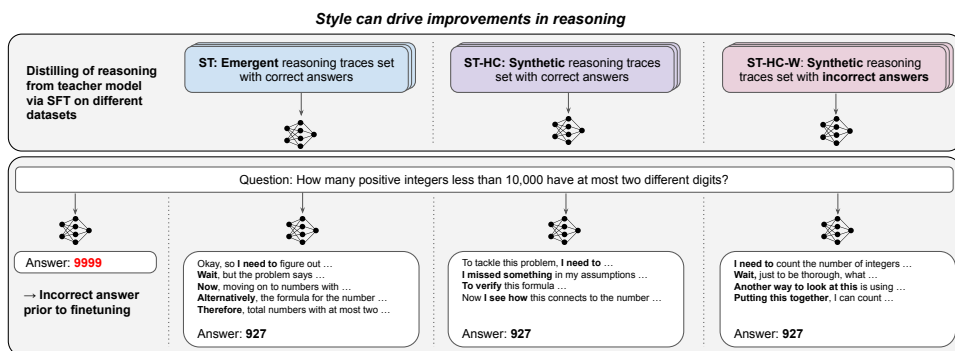


Figure 2.1.: Reasoning trace style has noticeable influence on distilled model performance. We show how different types of reasoning traces — emergent traces with correct answers, synthetic traces with correct answers, and synthetic traces with incorrect answers — affect model performance after finetuning. All three approaches give the right answer due to improved reasoning capabilities compared to base models. Pivots highlighted in bold.

performance. Our experiments demonstrate that even stylistically consistent synthetic traces from a weaker model achieve comparable downstream reasoning performance, underscoring the importance of style as a critical factor in model training.

Further experiments contextualize these findings through targeted ablation studies, which examine the relative importance of trace correctness versus style. Remarkably, we find that stylistically consistent reasoning traces — even those explicitly designed to lead to incorrect conclusions — still substantially enhance downstream reasoning performance over the base model, as shown in [figure 2.1](#). These results underscore that stylistic consistency significantly influences LM reasoning capabilities, providing an explanation for the effectiveness of reasoning distillation from RLMs to regular LMs.

In summary, our work makes two major contributions: (1) we demonstrate that distilled reasoning improvements rely heavily on stylistic patterns present in reasoning traces; and (2) we identify specific structural and lexical features indicative of successful reasoning. Collectively, these contributions deepen our understanding of how language models perform reasoning tasks, and the datasets we release can serve as valuable resources for future research into synthetic data generation and finetuning methodologies that explicitly target the relationship between reasoning trace style and substance.

2.2.2. BACKGROUND

Chain-of-Thought Early approaches to elicit reasoning from LMs, such as Chain-of-Thought (CoT) [18], demonstrate that intermediate reasoning steps are key for LMs to improve their problem-solving abilities [19]. While CoT improves performance on reasoning tasks, it primarily focuses on generating a linear sequence of steps towards a solution. On the other hand, RLM reasoning traces (sometimes called “long CoT”) —

which are the focus of this chapter – differ from CoT as they do not just try to build toward the solution linearly, but instead actively backtrack, verify, and explore different lines of thinking.

Reasoning traces Reasoning traces are semi-structured textual representations that capture a model’s thought process while working toward the solution of a problem during inference. Here, additional tokens are generated before the final answer is given to reason about the problem. These traces typically include explicit markers of metacognition, such as planning statements, hypothesis testing, and self-correction. For an example of a full reasoning trace, see [section A.1](#). A key characteristic of effective reasoning traces is the presence of *pivots*, points where the model explicitly moves between different categories of metacognition. We delve into the specific types of pivots in [section 2.2.3](#).

Finetuning on reasoning traces Recent research has demonstrated that finetuning language models on reasoning traces significantly enhances their reasoning capabilities [8, 10, 20]. In this approach, detailed reasoning traces generated by RLMs are used as training data to transfer structured reasoning behaviors into smaller, instruction-tuned models via SFT [21]. Models finetuned on these traces consistently outperform those trained only on final answers or simpler step-by-step solutions, suggesting that the explicit structure and content of reasoning traces play a critical role in improving model performance [22, 23]. Despite these successes, the precise mechanisms underlying the effectiveness of reasoning trace distillation remain unclear, particularly regarding the balance between structural stylistic cues and the cognitive complexity of the reasoning itself.

2.2.3. HARD-CODING REASONING TRACES TO APPROXIMATE EMERGENT ONES

In examining whether distilled models internalize genuine reasoning capabilities or primarily benefit from structural and lexical patterns, we investigate the relationship between *style* and *substance* in reasoning traces. Here, style encompasses the structural and lexical features while substance refers to factual correctness and semantic content. To examine stylistic influence, we develop a methodology that replicates the structural patterns of successful reasoning while varying content, allowing us to assess style’s contribution to reasoning distillation. We hypothesize that emergent reasoning behaviors can be effectively approximated by encoding the metacognitive pivots characteristic of RLM traces into synthetic reasoning traces. We define *emergent traces* as those naturally produced by RLMs after training via reinforcement learning, while *synthetic traces* refer to our approximations that incorporate the stylistic elements we identify as key, but originate from standard LMs without specialized reasoning capabilities. This approach reduces reliance on costly RLM inference¹ while enabling us to control stylistic components that potentially drive performance improvements in distilled models.

¹At time of writing, the difference in API costs per token between flagship LMs and RLMs approach an order of magnitude from the same provider: input/output pricing of \$2.50/\$10.00 for GPT-4o compared to \$15.00/\$60.00 for o1 <https://openai.com/api/pricing/> [Accessed: 2025-02-21]

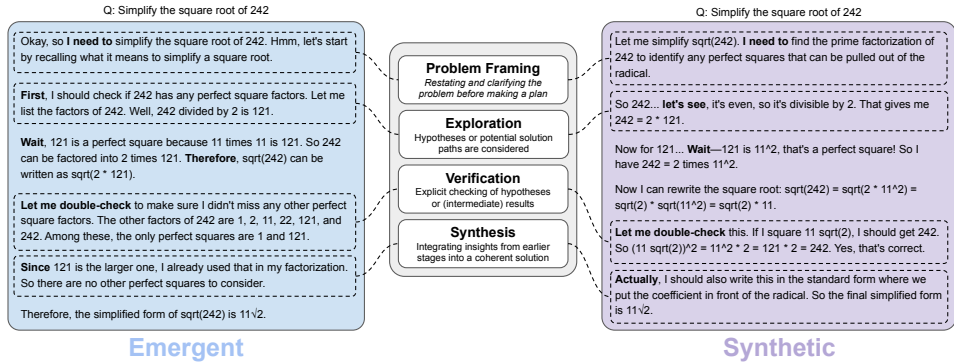


Figure 2.2.: Comparison of emergent and synthetic reasoning traces for solving the same problem. The left side shows a reasoning trace generated by an RLM, while the right side displays a synthetic trace created using our hard-coded template with predefined pivots. Both approaches follow similar cognitive stages (center): problem framing, exploration, verification, and synthesis. The dashed boxes highlight each stage, showing that synthetic traces can effectively replicate the style of emergent reasoning.

REASONING TRACE STRUCTURE AND PIVOT TYPES

Trace analysis Effective reasoning traces exhibit systematic structural patterns reflective of human-like problem-solving strategies. Cognitive science literature characterizes human reasoning as a structured process comprising distinct stages: *problem framing*, *hypothesis exploration*, *verification*, and *synthesis* [16]. Guided by this framework, we systematically analyze 17K successful reasoning traces produced by a state-of-the-art RLM (see section A.2 for details) and find that effective RLM-generated reasoning traces consistently align with these cognitive stages. Specifically, reasoning traces begin with explicit *problem framing*, restating and clarifying key aspects of the problem leading to a plan, followed by an *exploration* stage in which hypotheses or potential solution paths are considered. This is complemented by a *verification* stage, characterized by explicit checking of hypotheses or intermediate results, culminating in a final *synthesis*, integrating insights from earlier stages into a coherent solution.

Pivot types Crucially, effective reasoning traces are rarely linear; instead, they frequently revisit previous stages to correct errors, validate assumptions, or explore alternative strategies. This non-linear metacognitive behavior is operationalized through *pivots* – lexical markers signaling explicit shifts between reasoning stages. Our analysis identifies four primary pivot categories, each aligning closely with a corresponding reasoning stage: (1) *Realization pivots*, such as “Wait” or “Oh,” signal recognition of errors or oversights during the exploration stage; (2) *Verification pivots*, initiated by phrases such as “Let me check,” explicitly validate intermediate hypotheses; (3) *Exploration pivots*, introduced with phrases such as “What if” or “Another approach,” prompt the consideration of alternative solution paths; and (4) *Integration pivots*, typically signaled by expressions such as “Now I see how,” synthesize previously explored ideas

into a coherent final solution. We visualize what these stages look like in practice for emergent and synthetic traces in [figure 2.2](#)

Our analysis reveals that successful reasoning traces commonly employ multiple pivot types (96.1% contain at least three pivot categories, additional information per type given in [section A.2](#)), whereas unsuccessful traces frequently lack such pivots or exhibit limited diversity.

This underscores the critical importance of structured, metacognitive transitions in effective reasoning. Motivated by these insights, we explicitly encode these structural and lexical patterns into a reasoning template to facilitate the generation of synthetic reasoning traces that capture RLM-like reasoning behaviors. We show the associated prompt structure in [figure 2.3](#), which we subsequently use to guide reasoning trace generation ([section 2.2.3](#)), enabling us to systematically control and isolate stylistic reasoning elements in order to examine their impact on reasoning performance. First, the pivot categories are explicitly defined, followed by the general stages of problem solving that they correspond to. To generate the synthetic data, in addition to this prompt, the teacher LM is given the question and instructions on how to format its answer.

REASONING TRACE DATA GENERATION

Collect seed data Initially we curate seed data consisting of questions and their correct answers, ensuring that the accuracy of the eventual synthetic data can be verified. While previous works on RLM distillation often focus exclusively on math [20], we aim to cover a wider range of additional domains that benefit from reasoning, such as coding, science, and logic. Specifically, for questions from several scientific domains, we select OlympicArena [24]. For logic and coding, we select AGIEval [25] and LiveCodeBench v4 [26], respectively. We select NuminaMATH [27], where we randomly select a subset of 20,000 samples, and OmniMath [28] as sources of quantitative reasoning problems for our seed data, resulting a total of 31,586 question-answer pairs.

Generate synthetic traces We use the seed data to generate high-quality synthetic reasoning traces via state-of-the-art RLMs and LMs. For ST, we choose R1 as, at the time of writing, it is the best performing RLM that provides its full reasoning traces as part of its response. For ST-HC, we choose GPT-4o as our teacher model – using the prompt structure specified in [figure 2.3](#). We perform up to five rollouts per seed sample, discarding incorrect responses, stopping if the model provides the correct answer. The rollouts are done in a zero-shot manner, i.e. we provide only the question to the model and do not keep previous attempts as context.

Filtering synthetic samples First, we filter out samples that are of short length (less than 50 tokens for the entire trace) to prioritize sample quality, as shorter samples typically contain few pivots and are therefore not as impactful in training. In an effort to provide a fairer comparison, we align the datasets to the same number of samples by downsizing the larger dataset. As the final ST-HC dataset contained fewer samples ($N = 18,242$), we downsample the larger ST dataset to match this size, randomly removing questions not present in ST-HC and ensuring both datasets used for finetuning contain an equal number of samples. This results in a final 18K samples for both ST and ST-HC, each in the form of a triple (question, reasoning trace, answer). We provide a more detailed overview of the resulting datasets in [section A.3](#).

Prompt for Structured Reasoning with Explicit Pivots

This task requires solving problems using structured, real-time reasoning, including explicit self-monitoring and self-correction. Mimic the thought process of an agent that regularly pauses to reconsider assumptions, verify intermediate results, explore alternatives, and integrate findings into coherent solutions. Use explicit lexical pivots to signal shifts in thinking or corrections to your reasoning.

When solving the problem, follow a structured reasoning trace that clearly moves through the following stages:

- 1. Problem Framing:** Restate the problem and identify key elements clearly.
- 2. Exploration:** Consider one or more potential solution paths, openly weighing alternatives.
- 3. Verification:** Explicitly test intermediate results or assumptions; if inconsistencies arise, pivot explicitly to clarify or correct.
- 4. Synthesis:** Clearly integrate findings into a coherent solution, explicitly connecting back to the original problem.

Revisit stages as needed, backtracking in your thinking when appropriate.

When moving from one stage to the next, begin with a lexical pivot to signal the shift:

Realization pivots (recognizing errors or oversights): “Wait—”, “Oh—”, “Actually—”, “I missed something—”.

Verification pivots (testing assumptions or results): “Let me double-check—”, “To verify—”, “Checking again—”.

Exploration pivots (considering alternatives): “What if—”, “Another way to look at this—”, “Alternatively—”.

Integration pivots (synthesizing ideas): “Now I see how—”, “This connects back to—”, “Putting this together—”.

Use concise language. Be explicit about confusion or uncertainty when it arises. The goal is to clearly capture structured, non-linear reasoning and self-correction while making the logic easy to follow.

Figure 2.3.: Structured prompt used to elicit staged reasoning with explicit pivot markers.

EXPERIMENTAL DETAILS

Model finetuning We finetune a range of already instruction-tuned base LMs using our contributed reasoning datasets. For this, we select recent models of different families and sizes, namely: Llama 3.2 3B [29], Ministral 8B [30], and Qwen2.5 32B [31]. We choose these models as they are all high performing for their parameter count, come with open-source weights, and have permissive licenses. The learning rate used during SFT varies for each LM in line with the model’s parameter count. For 3B models we use a peak learning rate of 6×10^{-5} , for 8B models we use 4×10^{-5} , and for 32B models we

use 1×10^{-5} . For all models we use an effective batch size of 16. All models are trained for five epochs using a linear warmup for the first 10% of steps followed by cosine annealing. We use the AdamW optimizer [32] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$ and a weight decay of 1×10^{-4} . Training is performed on a system comprising 8 Nvidia H100 GPUs using `bfloat16` precision.

Baselines To contextualize our findings, we evaluate (1) the base instruction-tuned model, (2) the model after SFT on ST, and (4) the model after SFT on ST-HC. Additionally, to gauge the impact of the particular style replication we propose versus distilling from regular CoT, we add another baseline, where we instruct the generating LLM to think step-by-step (SBS), following Kojima *et al.* [33]. We then use the resulting CoT for distillation – similarly to how we use the emergent (ST) and synthetic (ST-HC) reasoning traces. This comprehensive evaluation allows us to isolate the impact of our hard-coded reasoning approach across different model sizes and compare it against the reasoning capabilities of the model itself. For completeness, we evaluate the teacher models used to generate our datasets as well.

Evaluation To evaluate the reasoning capabilities of all models, we select challenging, widely used benchmarks that test reasoning capabilities. These include MATH500 [34], AIME2024, and GPQA [35]; covering math and a range of scientific domains. Specifically, MATH500 and AIME2024 feature 500 and 30 competition math problems, respectively. GPQA consists of 198 questions from a range of scientific fields such as Biology and Chemistry. We focus only on the hardest (“Diamond”) subset of this particular dataset.

2.2.4. RESULTS AND DISCUSSION

HARD-CODED REASONING RESULTS

The performance across all combinations of models and datasets evaluated in our study is presented in [table 2.1](#). Our results clearly indicate that models finetuned with structured reasoning traces, whether emergent (ST) or synthetic (ST-HC), consistently and significantly outperform their baseline instruction-tuned counterparts across all benchmarks. Notably, even the smaller 3B and 8B parameter models exhibit substantial performance improvements. For instance, the Llama 3.2 3B model gains over 31 absolute percentage points on MATH500 when finetuned on ST, and the Ministral 8B model sees its AIME2024 score more than triple with either ST or ST-HC finetuning. These findings demonstrate that smaller models can markedly benefit from structured reasoning finetuning, challenging previous suggestions of minimal improvements for models of this scale [10]. Comparing the two finetuning approaches, we observe that models trained on synthetic traces achieve performance that approaches, and in some cases matches those trained on emergent RLM traces. This demonstrates that replicating the style of reasoning, even using a less capable teacher model guided by our prompt, is highly effective for distilling reasoning capabilities. Finally, the results show that while generating synthetic data with a SBS prompt is beneficial compared to the base model, there is a significant performance gap between SBS and our ST and ST-HC methods. This demonstrates that the performance improvements are indeed substantially driven by the specific stylistic patterns we identified, rather than solely by distilling the generating LLM’s general reasoning abilities.

Model	Variant	Params	MATH500	AIME2024	GPQA (D)
Llama 3.2	Base	3B	36.4	6.7	26.3
	SBS	3B	45.8	10.0	28.3
	ST	3B	68.4	23.3	31.3
	ST-HC	3B	64.2	16.7	29.3
Ministral	Base	8B	52.8	10.0	28.8
	SBS	8B	60.6	16.7	31.3
	ST	8B	78.2	33.3	38.9
	ST-HC	8B	77.0	33.3	34.8
Qwen2.5	Base	32B	76.8	16.7	49.0
	SBS	32B	78.2	20.0	49.5
	ST	32B	89.0	53.3	56.1
	ST-HC	32B	83.4	46.7	53.0
Teacher Models	R1	671B	96.8	76.7	71.7
	GPT-4o	—	75.4	13.3	53.0
	GPT-4o HC	—	81.2	16.7	55.1

Table 2.1.: Performance comparison of language models finetuned on reasoning traces. We compare base models against versions finetuned on SBS (step-by-step thinking with a structured prompt), ST (emergent traces from R1), and ST-HC (synthetic traces from GPT-4o using the structured prompt in Figure 2.3). This evaluates the impact of reasoning trace style on downstream performance across model scales and benchmarks. GPT-4o HC refers to the base LM prompted with our structured style. All results are pass@1 accuracy as a percentage.

When analyzing the thinking process during evaluation, we observe a clear correlation between successful reasoning and longer reasoning traces, as illustrated in figure 2.4. Models finetuned on either emergent or synthetic reasoning traces consistently produce substantially longer outputs compared to their respective base models across all evaluation benchmarks. This suggests that adopting and replicating a structured, elaborate reasoning style – whether learned from emergent traces or via synthetic ones – is a key mechanism driving the enhanced downstream reasoning capabilities we observed. Notably, figure 2.4 shows that finetuning on ST yields slightly longer reasoning traces than ST-HC, which correlates with their relative performance to one another – corroborating that a larger number of tokens spent thinking typically correlates with improved reasoning [36].

ABLATIVE STUDY

To further isolate the impact of reasoning style versus factual correctness, we conduct an ablation study using two modified datasets. First, we create ST-HC-W by adapting the synthetic ST-HC traces to retain their stylistic structure while leading to incorrect final answers generated by GPT-4o-mini (details given in section A.4). Second, we

Model	Variant	Params	MATH500	AIME2024	GPQA (D)
Llama 3.2	ST-HC-W	3B	48.2	10.0	28.4
	ST-NT	3B	40.6	6.7	26.9
Ministral	ST-HC-W	8B	62.8	20.0	29.9
	ST-NT	8B	56.2	13.3	30.0
Qwen2.5	ST-HC-W	32B	80.2	26.7	51.3
	ST-NT	32B	78.8	20.0	49.8

Table 2.2.: Ablation study evaluating the distinct contributions of reasoning trace style and answer correctness. We compare model performance after finetuning on: (1) ST-HC-W, featuring stylistically consistent synthetic traces from ST-HC but deliberately leading to *incorrect* answers, and (2) ST-NT, containing only the question-answer pairs from ST *without* reasoning traces. All results are pass@1 accuracy as a percentage.

create ST-NT, which contains only the question-answer pairs from the original ST data, removing the reasoning traces entirely. The results, presented in [table 2.2](#), reveal several key insights. Models finetuned on ST-HC-W consistently outperform the base instruction-tuned models across all evaluation benchmarks. This demonstrates that learning the stylistic patterns of reasoning enhances problem-solving capabilities even when the training data’s final conclusion is incorrect. However, ST-HC-W models perform noticeably worse than those trained on the stylistically similar but factually correct ST-HC dataset, confirming the value of accurate data. Unsurprisingly, models trained on ST-NT show only modest gains over the base models, as they do not learn to generate additional thinking tokens in the style of an RLM during inference.

These ablations confirm that while factual correctness is important for optimal performance, the stylistic patterns inherent in reasoning traces play a critical role in enhancing the reasoning abilities of distilled models. Crucially, we do not claim that style alone improves reasoning as the body of the traces still contains correct reasoning up until the answer, but rather emphasize that style is essential for improving reasoning capabilities. Thus, unlike traditional distillation methods relying predominantly on sample correctness, our results suggest that explicitly transferring a specific output structure is important.

2.2.5. RELATED WORK

Language model reasoning Language model reasoning has received increasing interest in recent years [37]. Initially, LMs were conditioned on reasoning examples during pretraining, post-training or in-context to improve their reasoning capabilities [18, 19, 38, 39], but test-time scaling [6, 36] has introduced a new paradigm for improving LM reasoning. Further, DeepSeek-AI *et al.* [8] explore training RLMs with reinforcement learning, finding that they mimic human reasoning processes like self-reflection and verification. The resulting RLMs provide reasoning traces as training data to enhance LM reasoning ability. Our work demonstrates that the structural and lexical patterns in

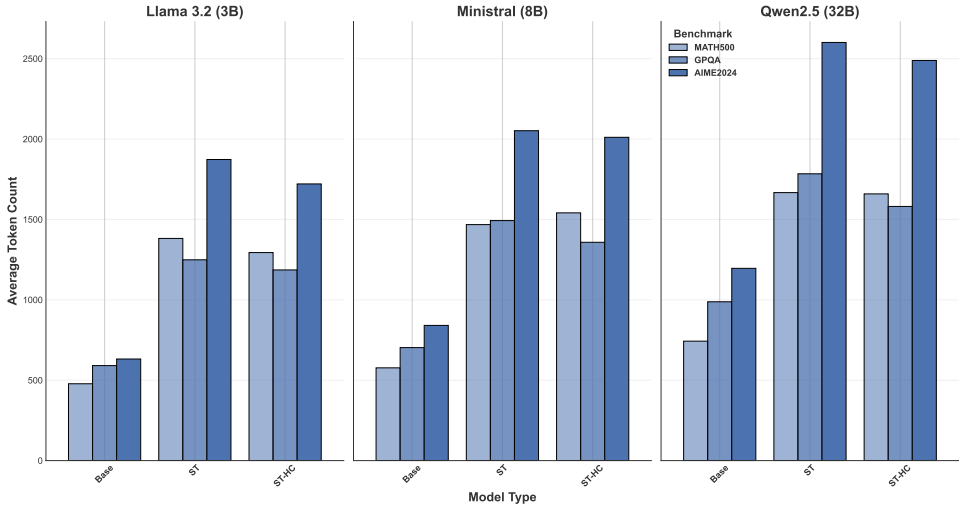


Figure 2.4.: Average token count by model family and training dataset across evaluations.

these traces play a significant role in improving reasoning performance, offering insights into what is actually transferred during reasoning distillation.

Generalization in language models Generalization in language models implies the ability to tackle unseen problems rather than simply reciting training data [40]. LMs struggle with problems that differ from their training distribution, often closely following observed patterns down to individual terms [41]. Recent research shows LMs rely heavily on memorized patterns rather than developing generalizable reasoning capabilities [42]. This pattern-matching behavior is especially evident in mathematical reasoning, where LMs show variance across different instantiations of the same question and declining performance when only numerical values change [15]. Data contamination has also been identified as a source of apparent but false generalization [43]. Our work extends this discussion by investigating how stylistic patterns in reasoning traces influence model performance, revealing that structural elements of reasoning may be as important as factual content for enhancing problem-solving capabilities.

Reasoning distillation Distillation has long been used to improve various aspects of machine learning models [3, 9, 44]. For reasoning tasks with verifiable solutions, researchers have implemented rejection sampling methodologies that extract and validate advanced models’ reasoning processes [45]. More recently, significant performance improvements have been achieved through SFT on synthetic datasets generated by superior LMs [46]. With the emergence of RLMs, reasoning performance of instruction-tuned LMs can be substantially enhanced by finetuning on high-quality reasoning traces [8]. Multiple efforts have demonstrated success in distilling RLMs via SFT on reasoning traces that contain step-by-step thinking [10, 20–23]. Our work attempts to decompose what is actually being transferred during this distillation process, revealing that stylistic elements of reasoning traces contribute significantly to

performance gains independent of their factual correctness.

2.2.6. CONCLUSION

Our work challenges existing assumptions by demonstrating that improvements in reasoning capabilities in distilled language models can be driven by stylistic replication. By analyzing reasoning traces through the lens of established principles from cognitive science on how humans solve problems, we identified structural patterns and key lexical markers (*pivots*) characteristic of effective LM reasoning. We validated these findings by synthetically replicating the stylistic patterns found in emergent reasoning traces using a weaker model, achieving comparable downstream reasoning performance after distillation. Ablation studies confirmed that even reasoning traces with incorrect solutions enhance downstream performance, highlighting the effectiveness of stylistic mimicry in distilled models. Collectively, our results contribute to a deeper understanding of reasoning in language models and offer practical insights into efficient finetuning strategies and data generation practices.

2.3. PART II — FROM STYLISTIC FINDINGS TO AGENT INTERFACES AND MEMORY

Part I (Section 2.2) established a central diagnostic: a significant share of reasoning transfer in distillation can be accounted for by effective *stylistic replication*, irrespective of answer correctness [1]. To further pursue the Diagnostic Gap, better understand model reasoning, and make robustness actionable we must separate the *internal* signal that models learn from the *external* scaffolding that determines what context is available to the model at decision time. This is addressed in Part II (Section 2.4).

If the presence and arrangement of structured intermediate thinking tokens are key correlates of success, then agent interfaces and tools that shape *when* and *how* such structure is expressed should matter, even without updating model weights. Concretely, prompts that cue thinking processes, reflections that elicit particular thinking processes, and memory can surface and leverage the very regularities that proved beneficial in Part I. Conversely, cluttered, unstructured, or failure-only contexts can suppress those regularities, leading to the same brittle behavior that Chapter 1 flagged as part of the Diagnostic Gap.

We therefore read the Part I results as generating testable design hypotheses for agents in interactive environments: (i) interfaces that encourage staged reasoning and pivot diversity should stabilize decisions across steps; (ii) success-aware context should help retain useful structure after early wins, mitigating tilt; and (iii) retrieval should be budgeted and de-duplicated to preserve the signal rather than drown it. This line of thinking directly motivates the two elements studied in Part II: *experience reflection* and *managed memory*.

Experience reflection augments a language model's context by collecting succinct takeaways when subgoals are reached in an interactive environment, thereby reinforcing the patterns (verification, synthesis, reframing) that Part I associated with stronger performance. Managed memory organizes these takeaways and makes them available

to the language model when needed. Together, these choices aim not to manufacture reasoning but to surface the structured behavior models already express when given the right context using external scaffolding.

This perspective also clarifies scope and limits. Part II assesses reliability at the *behavioral* level without updating model parameters. The result is a conceptual through-line for Pillar 1: Part I explains what internally tends to transfer through distillation (style versus substance), while Part II examines how external scaffolding in the form of interface and memory design governs whether that structure leads to improved reasoning and robustness. In doing so, Part II operationalizes the agenda from [Chapter 1](#) by turning diagnosis into concrete, testable design decisions for language model-based agents.

2.4. POSITIVE EXPERIENCE REFLECTION IN INTERACTIVE ENVIRONMENTS

2.4.1. INTRODUCTION

Intelligent agents, designed to interact with and make decisions in dynamic environments, have become a central focus in AI research, with text-based games (TBGs) emerging as a particularly challenging domain for evaluating these agents' reasoning, adaptability, and learning abilities [47, 48]. Originally popular in the 1970s as text adventure games,² TBGs present players with textual descriptions of environments, requiring them to input natural language commands to achieve objectives [49]. For instance, determining if a metal fork is conductive involves locating the fork, assembling a circuit, and analyzing the result. Navigating TBGs demands that agents exhibit a combination of abilities, including planning, memory retention, spatial reasoning, and common sense knowledge [50, 51].

Previously, deep reinforcement learning and behavior cloning were the primary approaches to develop agents to play TBGs [51, 52]. However, recent research shows that agents based on pretrained large language models (LLMs) are more effective at navigating TBGs [53]. A key factor in their success is the integration of internal *reflection* to improve planning [54–56].

Self-reflection, closely related to self-refinement, is a form of reasoning that occurs after receiving binary or scalar feedback from the environment [57]. In this process, the LLM reviews its actions and their outcomes, considering what went wrong and potential ways to improve [58]. By iteratively adjusting its strategy based on verbal reinforcement, conveyed through textual feedback, the agent refines its planning for subsequent attempts [59]. However, reflection also has several limitations, including (1) underwhelming performance when agents are correct initially [39], (2) significantly worse efficacy when using smaller LLMs [53], and (3) dependence on external feedback [60].

In this work we conduct a comprehensive analysis of LLM-based agents employing reflection approaches in TBGs and evaluate their performance across various LLMs. To address the limitations of reflection when agents are initially successful and the

²Try it yourself: <https://www.microsoft.com/en-us/research/project/textworld/try-it/>

diminished efficacy of smaller LLMs, we propose *Sweet&Sour* (S&S) to leverage positive experiences to create a richer context for self-reflection. We supplement this by proposing a managed memory approach to build context across multiple rollouts. Our findings demonstrate that our method improves the performance of agents using reflection, particularly in scenarios where they previously struggled, enabling more robust and generalizable learning.

2.4.2. BACKGROUND

Text-based games (TBGs) provide a controlled setting for studying decision-making under partial observability. They can be modeled as a partially observable Markov decision process (POMDP) [61] with tuple $\langle S, T, A, \Omega, R, \gamma \rangle$, where S is the set of latent states, A the (textual) action space, Ω the observation function, R the reward function, and γ a discount factor. At time t , the agent receives observation o_t sampled from $\Omega(o_t | s_t, a_{t-1})$ and chooses a text action a_t ; both observations and actions are sequences of tokens. Admissible actions are those that change the environment state; non-admissible actions leave the state unchanged and yield no progress.

Prior approaches organize language-model agents for these environments along a continuum of reasoning and memory. CALM leverages imitation learning with a DRRN reranker to improve action selection [52]. ReAct interleaves reasoning and acting to compose useful context at each step, but does not persist auxiliary structure across attempts [62]. Reflexion augments ReAct with failure-based self-reflection, writing takeaways to a long-term memory (LTM) that is retrieved in subsequent attempts [59]. More broadly, recent surveys highlight the role of reflection and planning for LLM agents [54–56].

Within this landscape, we consider two complementary design elements that will be developed in Section 2.4.3: (i) *positive experience reflection*, which captures succinct insights on subgoal success rather than only failures, and (ii) *managed memory*, which uses a dual-buffer design with short-term memory (STM) available within the current episode and LTM persisted across episodes. These choices target regimes where failure-only reflection underperforms (e.g., when initial attempts succeed or models are smaller), while preserving the simplicity and generality of text-based interfaces.

2.4.3. METHODOLOGY

We summarize the action–observation–reflection loop, define the Sweet&Sour reflection mechanism, and describe managed memory.

LLMs PLAYING TEXT-BASED GAMES

Assuming an LLM behaving as an actor model as part of our agent—i.e., generating actions based on the current state and policy, analogous to traditional policy-based RL setups—we sample an action a_t from the current policy π_θ at time t and receive an observation from the environment o_t . Each task consists of a number of sub-tasks (such as finding a key object), the completion of which grants the agent a sparse reward, which adds to its current reward r_t . The game continues until the agent has achieved the goal

d and receives the full reward as final score, or the maximum number of steps—which we set to 150—is reached, in which case r_t becomes the final score. A detailed problem formulation is given in [Section 2.4.2](#); formal POMDP details appear in [Appendix B.1](#).

SELF-REFLECTION

Reflection occurs in addition to the acting LLM. Here, the agent reviews a_t and o_t associated with previous unsuccessful attempts to verbalize the reason for failure. This process typically involves maintaining a persistent history of insights gained across attempts, which the LLM uses as additional context for its reflections to improve future decision making for the next attempt [59]. However, since other self-reflection methods focus on learning from failures [63–67], they overlook the importance of reinforcing successful behaviors in a similar way.

SWEET&SOUR

Positive experience reflection. To address the limitations of existing self-reflection methods, we introduce a structured approach to leverage reflections from both positive (*sweet*) and negative (*sour*) outcomes. Unlike Reflexion, which passively accumulates failure-based insights, S&S actively queries the agent for generalizable insights both from failure and success cases, promoting a more balanced context-building mechanism. Here, we draw inspiration from RL, where rewards steer the agent towards reinforcing advantageous behaviors and preventing over-reliance on error correction. When the current policy is achieving rewards, we query the agent to extrapolate, encouraging the agent to articulate what made its current policy successful and what can be generalized from this, reinforcing strategies that lead to positive outcomes while still learning from failures. This is visualized in [figure 2.5](#). The full algorithm is shown in [Algorithm 1](#); qualitative examples are provided in [Appendix B.2](#). Our method is broadly applicable to agents in environments with feedback using selfreflection, including those that build additional complexity on top of the core reflection loop, such as grounding [53] or gradient learning [67].

MANAGED MEMORY

Previous works store reflections gained from unsuccessful attempts in something akin to a long-term memory and make them available as additional context across attempts [59, 66]. This implies that the agent only has access to additional context upon failing the task—essentially brute-forcing the problem across rounds. We instead adopt a dual-buffer design with *short-term memory* (STM) within an episode and *long-term memory* (LTM) across episodes, with controlled migration. Positive reflections on subgoal completion are immediately written to STM so they are available in the next decision window; on episode end, STM is migrated to LTM if the episode was successful. If the episode fails, a concise negative reflection is written to LTM and STM is reset. At each step, the agent retrieves a small, budgeted subset from (STM, LTM) to avoid prompt bloat and keep context relevant.

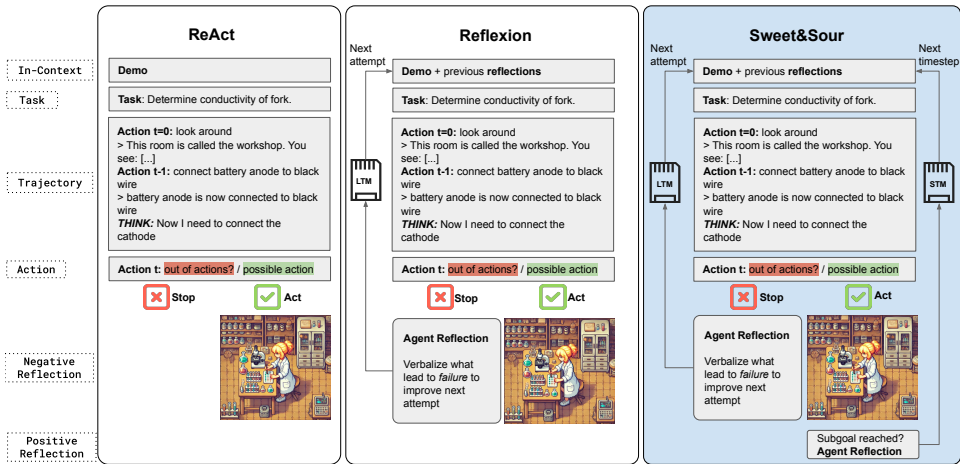


Figure 2.5.: Comparison of methods used to play ScienceWorld. ReAct introduces a THINK action to explicitly reason about the next step, which does not persist across attempts. Reflexion leverages self-reflection across attempts to learn from unsuccessful tries and stores these in long-term memory (LTM). Sweet&Sour not only performs self-reflection after failures but also after each completed subgoal, making its reflection instantly available at the next time step as part of short-term memory (STM), thus incorporating positive experiences.

2.4.4. EVALUATION SETUP

We evaluate in ScienceWorld, compare against standard agent baselines, and run across multiple LLMs to assess size and model-family effects.

DATA AND ENVIRONMENT (SCIENCEWORLD)

We use the ScienceWorld benchmark [48], which provides a versatile setting for evaluating agents in science experiment tasks across interconnected locations (e.g., greenhouse, workshop) with hundreds of objects and a diverse set of action templates, creating a large and dynamic search space. We evaluate on the test set, which offers up to 10 variations of each of 30 distinct tasks with an average optimal decision depth of about 50 steps. We measure performance using the success score, where completing all sub-tasks earns a score of 100. We elect ScienceWorld over earlier interactive text benchmarks due to their relative simplicity for current LLM-based agents. We provide additional results on the simpler environment ALFWorld [68] in Appendix B.3.

BASELINES

CALM [52] integrates a deep reinforced relevance network (DRRN) [69] with a causal language model trained via imitation learning on oracle trajectories; DRRN reranks LM action candidates. ReAct [62] is our LLM-based agent baseline, interleaving explicit

Algorithm 1 Sweet&Sour Reflection Process with Short- And Long-Term Memory

Require: Initial observation o_0 containing goal description d

Require: Maximum steps $T = 150$

```

1: Initialize short-term memory  $M_{ST} \leftarrow \emptyset$ 
2: Initialize long-term memory  $M_{LT} \leftarrow \emptyset$ 
3: Initialize current reward  $r \leftarrow 0$ 
4: for each episode do
5:    $t \leftarrow 0$ 
6:   while  $t < T$  and goal not achieved do
7:     Retrieve reflections from  $M_{ST}$  and  $M_{LT}$ 
8:     Sample action  $a_t \sim \pi_\theta(a_t \mid o_t, M_{ST}, M_{LT})$ 
9:     Execute  $a_t$ , observe  $o_{t+1}$ , potentially receive reward  $r_t$ 
10:     $r \leftarrow r + r_t$ 
11:    if subgoal achieved then
12:      Generate positive reflection  $R_t^+$ 
13:       $M_{ST} \leftarrow M_{ST} \cup \{(R_t^+, o_t, a_t, r_t)\}$ 
14:    end if
15:     $t \leftarrow t + 1$ 
16:  end while
17:  if goal achieved then
18:     $M_{LT} \leftarrow M_{LT} \cup M_{ST}$  Transfer successful memories
19:  else
20:    Generate negative reflection  $R_t^-$ 
21:     $M_{LT} \leftarrow M_{LT} \cup \{(R_t^-, o_t, a_t, r_t)\}$ 
22:  end if
23:   $M_{ST} \leftarrow \emptyset$  Reset short-term memory
24: end for

```

reasoning with acting at each step and carrying useful information forward as in-context examples. *Reflexion* [59] augments ReAct with self-reflection across attempts based on environment feedback, accumulating a long-term memory over up to four rounds before termination.

MODELS AND PROTOCOL

We assess each method with multiple LLMs to probe scaling and family effects. In descending parameter order, we use GPT-4o, Mistral Large 2, and Llama 3.1 8B. Across methods we standardize decoding and stopping criteria per benchmark guidance, and we report average scores across task variations for each task, then average across tasks.

2.4.5. RESULTS AND DISCUSSION

The main ScienceWorld results are summarized in [table 2.3](#).

Task	CALM	ReAct			Reflexion			Sweet&Sour (ours)		
	CALM	L8B	ML2	GPT	L8B	ML2	GPT	L8B	ML2	GPT
1-1 (Boil)	0.0	0.0	0.0	3.8	0.0	0.0	5.1	0.0	7.2	9.6
1-2 (Melt)	0.0	8.4	10.3	11.8	0.0	0.0	10.0	11.4	12.1	12.8
1-3 (Freeze)	0.0	1.5	0.0	8.1	0.0	2.3	8.3	2.4	3.1	8.9
1-4 (Change state)	0.0	1.0	4.7	10.0	0.0	0.0	4.2	1.7	2.9	9.2
2-1 (Thermometer)	1.0	5.1	7.8	7.7	3.4	4.2	7.6	7.8	9.7	10.9
2-2 (Melting)	1.0	6.7	6.3	5.9	3.3	3.3	26.2	7.9	36.8	46.0
2-3 (Melting)	5.0	9.1	11.8	23.4	13.2	14.7	22.6	15.2	29.0	38.3
3-1 (Power 1)	7.0	18.8	24.6	57.2	21.2	51.5	78.4	28.6	75.4	81.1
3-2 (Power 2)	2.0	10.2	24.7	55.6	9.5	11.9	24.7	23.3	44.5	58.0
3-3 (Conductivity 1)	2.0	52.4	51.7	73.0	9.2	25.8	72.1	59.1	69.2	75.7
3-4 (Conductivity 2)	10.0	54.2	64.9	89.7	35.4	41.6	75.1	62.7	60.3	67.3
4-1 (Find 1)	54.0	17.3	18.7	27.5	44.6	48.1	62.3	41.7	71.7	74.2
4-2 (Find 2)	10.0	69.1	71.6	80.3	68.4	75.7	87.3	76.8	100.0	100.0
4-3 (Find 3)	8.0	21.3	42.8	47.7	18.4	16.5	17.3	20.9	21.5	34.3
4-4 (Find 4)	2.0	15.7	15.2	19.3	39.6	46.6	100.0	55.1	87.8	100.0
5-1 (Grow plant)	4.0	10.8	10.8	10.0	7.2	7.2	7.9	14.2	14.6	17.4
5-2 (Grow fruit)	3.0	18.1	18.5	19.2	30.8	51.4	34.6	51.5	55.6	60.2
6-1 (Chemistry 1)	6.0	37.8	42.9	58.6	27.1	29.7	70.2	37.9	61.1	70.2
6-2 (Chemistry 2)	3.0	25.0	27.1	50.6	14.4	28.0	69.8	27.2	51.9	83.1
6-3 (Chemistry 3)	6.0	14.4	17.5	39.7	38.9	31.1	16.7	45.3	53.7	61.5
7-1 (Lifespan 1)	10.0	37.0	41.7	60.0	75.0	75.0	100.0	75.0	88.2	100.0
7-2 (Lifespan 2)	4.0	50.5	50.7	67.5	60.0	71.9	81.4	70.5	77.0	80.0
7-3 (Lifespan 3)	4.0	33.7	38.2	50.0	29.5	33.7	75.0	51.1	54.2	84.6
8-1 (Identify life 1)	0.0	5.1	18.9	25.3	1.7	1.7	3.4	11.1	10.3	14.2
8-2 (Identify life 2)	0.0	6.4	7.4	8.0	7.4	8.0	8.0	5.0	7.4	7.4
9-1 (Measure angle)	0.0	28.5	33.0	42.5	56.9	55.1	57.1	68.4	70.3	75.0
9-2 (Friction 1)	3.0	14.5	22.6	43.1	23.4	29.3	100.0	33.3	36.7	62.0
9-3 (Friction 2)	2.0	2.9	14.5	42.8	1.3	33.6	59.6	7.2	51.9	63.1
10-1 (Genetics 1)	2.0	25.7	27.3	26.4	5.6	9.8	50.4	38.9	48.6	78.8
10-2 (Genetics 2)	2.0	13.2	19.1	17.2	6.2	21.5	22.7	23.6	24.0	54.8
Average	5.1	20.5	24.8	36.0	21.7	27.6	45.3	32.5	44.6	54.6

Table 2.3.: Results on the ScienceWorld benchmark. For each method, we use Llama 8B (L8B), Mistral Large 2 (ML2), and GPT-4o (GPT). Each value is an average of across all task variations. Best average values per model are underlined.

Sweet&Sour outperforms baselines. S&S outperforms other methods across all LLMs, achieving the highest average score of 54.6 with GPT-4o. The gap widens for smaller models: for example, S&S reaches 44.6 versus Reflexion’s 27.6 on Mistral Large 2, and 32.5 versus 21.7 on Llama 8B, indicating greater suitability when compute is limited.

Ablations demonstrate complementary contributions. We ablate positive and negative reflections to assess their separate contributions. With failure-only reflection, performance drops to Reflexion-like levels (averages 24.6, 31.1, 44.9 on Llama 8B, Mistral Large 2, and GPT-4o, respectively). Using positive-only reflections remains above ReAct (25.8, 32.4, 42.3) but is less effective than negative-only. Removing managed memory in favor of a simple LTM decreases scores to 28.2, 38.5, and 48.7% still exceeding Reflexion but averaging 12.4% below S&S. Together, these results indicate that negative reflections, positive reflections, and managed memory each help, and their combination is strongest.

Robustness to tilt. On very hard tasks (e.g., 1-1, 8-2), all methods struggle, and on easy tasks, most methods succeed given model capacity. Medium-difficulty tasks (e.g.,

6-3, 10-2) reveal a gap: Reflexion may start strong but get stuck after the first error. We hypothesize that failure-only methods lack context from early successes, making them prone to “tilt”. By contrast, S&S’s balanced reflection injects success context, sustaining performance in these regimes.

2.4.6. QUALITATIVE EXAMPLES

Qualitative transcripts illustrating the reflection loop and managed memory effects are shown in Appendix B.2.

2.4.7. CONCLUSION

We introduced Sweet&Sour, a reflection approach that leverages both successes and failures and organizes context through managed memory (STM/LTM). In ScienceWorld (with ALFWorld corroboration), S&S outperforms standard baselines across model sizes, with especially strong relative gains for smaller LLMs. These findings highlight the value of positive experience reflection and memory design in developing more robust agents for interactive text environments.

2.4.8. LIMITATIONS AND BROADER IMPACTS

Limitations. While results are promising, our evaluation focuses on ScienceWorld with additional ALFWorld corroboration; broader interactive settings may present different challenges. LLMs provide no formal guarantees regarding reasoning quality, and reliance on textual reflection may introduce biases or inconsistencies in behavior.

Broader impacts. LLM agents carry risks of misuse and unintended negative consequences [70]. Improving agent performance can be dual-use. Our study mitigates risks by relying on simulated benchmarks and by aiming to reduce error rates; nonetheless, responsible deployment requires careful oversight and domain-appropriate safeguards.

2.5. CONCLUSION

This chapter advanced Pillar 1 by clarifying what is learned and what is expressed when language models appear to reason, and by showing how interface and memory design govern the reliability of that behavior in interactive settings. Part I (section 2.2) disentangled style and substance during distillation on reasoning traces. We found that structural regularities found in the reasoning traces account for a substantial share of the measured improvement, with synthetic traces that preserve these regularities approaching the gains of emergent traces. While correctness still matters, style alone explained large differences in performance across model families and tasks. Further, we found that longer, more structured outputs correlated with higher accuracy. These results sharpen the Diagnostic Gap from Chapter 1: without distinguishing stylistic replication from transferable logical processes, we risk overestimating language model robustness.

Part II (section 2.4) translated those diagnostics into agent design in interactive text environments. Positive experience reflection (Sweet&Sour) and managed memory

improved reliability across model sizes, particularly for smaller models and in regimes prone to tilt. Crucially, these gains were achieved without updating model parameters: the interface supported structural patterns that are conducive to reasoning.

Taken together, the evidence yields three implications. First, supervision and evaluation should be style-aware: encourage staged thinking processes and track structure (e.g., trace length, pivot diversity) alongside accuracy, while maintaining strict quality control to ensure correctness. Second, agent interfaces are part of the robustness story: positive experience reflections and managed memory can improve reasoning behavior when initial attempts succeed and reduce susceptibility to brittle behavior after encountering errors. Third, style is a means, not an end: it is a diagnostic and a design signal for data and interfaces, not a substitute for factual or domain adequacy.

Limitations remain. Our analysis focuses on math and scientific reasoning benchmarks and on two interactive environments. We do not make claims about internal representations beyond what is evidenced by behavioral correlates, and we caution that style-driven gains can obscure deficits in semantic competence if not paired with verification. These constraints motivate care in generalization claims and underscore the need for effective robustness interventions.

This chapter therefore provides the conceptual through-line for the remainder of the thesis. [Chapter 3](#) leverages the stylistic diagnostics to guide *targeted synthetic data* that addresses discovered weaknesses while preserving in-domain performance. [Chapter 4](#) turns to governance in value-laden settings, using elicited user-perceived value to determine when to trust or escalate, rather than relying on confidence alone.

REFERENCES

- [1] P. Lippmann and J. Yang. ‘Style over Substance: Distilled Language Models Reason Via Stylistic Replication’. In: *Proceedings of the Second Conference on Language Modeling*. 8th July 2025. url: <https://openreview.net/forum?id=5wAfbEs34A>.
- [2] P. Lippmann, M. T. J. Spaan and J. Yang. ‘Positive Experience Reflection for Agents in Interactive Text Environments’. In: *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*. Ed. by E. Kamaloo, N. Gontier, X. H. Lu, N. Dziri, S. Murty and A. Lacoste. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 131–142. isbn: 979-8-89176-264-0. doi: [10.18653/v1/2025.realm-1.10](https://doi.org/10.18653/v1/2025.realm-1.10). url: <https://aclanthology.org/2025.realm-1.10/>.
- [3] G. Hinton, O. Vinyals and J. Dean. *Distilling the Knowledge in a Neural Network*. NIPS 2014 Deep Learning Workshop. 9th Mar. 2015. doi: [10.48550/arXiv.1503.02531](https://doi.org/10.48550/arXiv.1503.02531). arXiv: [1503.02531](https://arxiv.org/abs/1503.02531) [stat.ML]. url: <https://arxiv.org/abs/1503.02531>.

- [4] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu and D. Amodei. *Scaling Laws for Neural Language Models*. 23rd Jan. 2020. doi: [10.48550/arXiv.2001.08361](https://doi.org/10.48550/arXiv.2001.08361). arXiv: [2001.08361](https://arxiv.org/abs/2001.08361) [cs.LG]. url: <https://arxiv.org/abs/2001.08361>.
- [5] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals and L. Sifre. *Training Compute-Optimal Large Language Models*. 29th Mar. 2022. doi: [10.48550/arXiv.2203.15556](https://doi.org/10.48550/arXiv.2203.15556). arXiv: [2203.15556](https://arxiv.org/abs/2203.15556) [cs.CL]. url: <https://arxiv.org/abs/2203.15556>.
- [6] C. Snell, J. Lee, K. Xu and A. Kumar. *Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Model Parameters*. 6th Aug. 2024. arXiv: [2408.03314](https://arxiv.org/abs/2408.03314) [cs.LG]. url: <https://arxiv.org/abs/2408.03314> (visited on 19/12/2025).
- [7] OpenAI. *Learning to Reason with LLMs*. OpenAI. 2024. url: <https://openai.com/index/learning-to-reason-with-llms/> (visited on 29/12/2025).
- [8] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 22nd Jan. 2025. arXiv: [2501.12948](https://arxiv.org/abs/2501.12948) [cs.CL]. url: <https://arxiv.org/abs/2501.12948>.
- [9] J. Schmidhuber. ‘Learning Complex, Extended Sequences Using the Principle of History Compression’. In: *Neural Computation* 4.2 (Mar. 1992), pp. 234–242. doi: [10.1162/neco.1992.4.2.234](https://doi.org/10.1162/neco.1992.4.2.234). url: <https://doi.org/10.1162/neco.1992.4.2.234>.
- [10] N. Team. *Sky-T1: Train your own O1 preview model within \$450*. <https://novasky-ai.github.io/posts/sky-t1>. Accessed: 2025-01-09. 2025.
- [11] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer and H. Sun. *Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters*. arXiv:2212.10001v2. 20th Dec. 2022. arXiv: [2212.10001](https://arxiv.org/abs/2212.10001) [cs.CL]. url: <https://arxiv.org/abs/2212.10001>.
- [12] E. M. Bender and A. Koller. ‘Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schlueter and J. Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 5185–5198. doi: [10.18653/v1/2020.acl-main.463](https://doi.org/10.18653/v1/2020.acl-main.463). url: <https://aclanthology.org/2020.acl-main.463/>.
- [13] E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell. ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?’ In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, Mar. 2021, pp. 610–623. doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). url: <https://doi.org/10.1145/3442188.3445922>.

- [14] M. Shanahan. ‘Talking About Large Language Models’. In: *Communications of the ACM* 67.2 (Feb. 2024), pp. 68–79. issn: 0001-0782. doi: [10.1145/3624724](https://doi.org/10.1145/3624724). url: <https://doi.org/10.1145/3624724> (visited on 19/12/2025).
- [15] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio and M. Farajtabar. *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. arXiv:2410.05229v2. 27th Aug. 2025. arXiv: [2410.05229](https://arxiv.org/abs/2410.05229) [cs.LG]. url: <https://arxiv.org/abs/2410.05229> (visited on 16/12/2025).
- [16] A. Newell and H. A. Simon. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall, 1972. 920 pp. isbn: 0134454030.
- [17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei. ‘Language Models are Few-Shot Learners’. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan and H.-T. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. url: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfbcb4967418bfb8ac142f64a-Paper.pdf>.
- [18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le and D. Zhou. ‘Chain-of-Thought Prompting Elicits Reasoning in Large Language Models’. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh. Vol. 35. Curran Associates, Inc., 6th Dec. 2022, pp. 24824–24837. url: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf (visited on 25/02/2025).
- [19] Z. Zhang, A. Zhang, M. Li and A. Smola. *Automatic Chain of Thought Prompting in Large Language Models*. Version 1, 2022-10-07. 7th Oct. 2022. arXiv: [2210.03493](https://arxiv.org/abs/2210.03493) [cs.CL]. url: <https://arxiv.org/abs/2210.03493> (visited on 22/12/2025).
- [20] Y. Qin, X. Li, H. Zou, Y. Liu, S. Xia, Z. Huang, Y. Ye, W. Yuan, H. Liu, Y. Li and P. Liu. *O1 Replication Journey: A Strategic Progress Report – Part 1*. Version 1. 8th Oct. 2024. arXiv: [2410.18982](https://arxiv.org/abs/2410.18982) [cs.AI]. url: <https://arxiv.org/abs/2410.18982>.
- [21] Y. Min, Z. Chen, J. Jiang, J. Chen, J. Deng, Y. Hu, Y. Tang, J. Wang, X. Cheng, H. Song, W. X. Zhao, Z. Liu, Z. Wang and J.-R. Wen. *Imitate, Explore, and Self-Improve: A Reproduction Report on Slow-thinking Reasoning Systems*. arXiv:2412.09413v2. 22nd Dec. 2024. arXiv: [2412.09413](https://arxiv.org/abs/2412.09413) [cs.AI]. url: <https://arxiv.org/abs/2412.09413> (visited on 16/12/2025).

- [22] H. Xu, X. Wu, W. Wang, Z. Li, D. Zheng, B. Chen, Y. Hu, S. Kang, J. Ji, Y. Zhang, Z. Guo, Y. Yang, M. Zhang and D. Zhang. *RedStar: Does Scaling Long-CoT Data Unlock Better Slow-Reasoning Systems?* 20th Jan. 2025. arXiv: [2501.11284](https://arxiv.org/abs/2501.11284) [cs.LG]. url: <https://arxiv.org/abs/2501.11284> (visited on 25/02/2025).
- [23] Bespoke Labs. *Bespoke-Stratos: The unreasonable effectiveness of reasoning distillation*. Bespoke Labs. 6th Dec. 2025. url: <https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation> (visited on 11/12/2025).
- [24] Z. Huang, Z. Wang, S. Xia, X. Li, H. Zou, R. Xu, R.-Z. Fan, L. Ye, E. Chern, Y. Ye, Y. Zhang, Y. Yang, T. Wu, B. Wang, S. Sun, Y. Xiao, Y. Li, F. Zhou, S. Chern, Y. Qin, Y. Ma, J. Su, Y. Liu, Y. Zheng, S. Zhang, D. Lin, Y. Qiao and P. Liu. *OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI*. 18th June 2024. arXiv: [2406.12753](https://arxiv.org/abs/2406.12753) [cs.CL]. url: <https://arxiv.org/abs/2406.12753>.
- [25] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen and N. Duan. *AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models*. arXiv:2304.06364v2. 13th Apr. 2023. arXiv: [2304.06364](https://arxiv.org/abs/2304.06364) [cs.CL]. url: <https://arxiv.org/abs/2304.06364>.
- [26] N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen and I. Stoica. *LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code*. 6th June 2024. arXiv: [2403.07974](https://arxiv.org/abs/2403.07974) [cs.SE]. url: <https://arxiv.org/abs/2403.07974>.
- [27] J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. Huang, K. Rasul, L. Yu, A. Q. Jiang, Z. Shen, Z. Qin, B. Dong, L. Zhou, Y. Fleureau, G. Lample and S. Polu. *NuminaMath: The largest public dataset in AI4Maths with 860k pairs of competition math problems and solutions*. Technical Report. Numina, 22nd July 2024. url: https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf (visited on 29/12/2025).
- [28] B. Gao, F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, Z. Tang, B. Wang, D. Zan, S. Quan, G. Zhang, L. Sha, Y. Zhang, X. Ren, T. Liu and B. Chang. *Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models*. 10th Oct. 2024. arXiv: [2410.07985](https://arxiv.org/abs/2410.07985) [cs.CL]. url: <https://arxiv.org/abs/2410.07985>.
- [29] Meta AI. *Llama 3.2 Lightweight Models (1B/3B)*. Model card describing the architecture, training recipe, and evaluation metrics for the Llama 3.2 1B and 3B models. 2024. url: https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/ (visited on 16/12/2025).
- [30] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix and W. E. Sayed.

- Mistral 7B*. 10th Oct. 2023. doi: [10.48550/arXiv.2310.06825](https://doi.org/10.48550/arXiv.2310.06825). arXiv: [2310.06825 \[cs.CL\]](https://arxiv.org/abs/2310.06825). url: <https://arxiv.org/abs/2310.06825>.
- [31] Qwen Team. *Qwen2.5 Technical Report*. 3rd Jan. 2025. arXiv: [2412.15115 \[cs.CL\]](https://arxiv.org/abs/2412.15115). url: <https://arxiv.org/abs/2412.15115>.
- [32] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. Published as a conference paper at ICLR 2019. 14th Nov. 2017. arXiv: [1711.05101 \[cs.LG\]](https://arxiv.org/abs/1711.05101). url: <https://arxiv.org/abs/1711.05101>.
- [33] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa. *Large Language Models are Zero-Shot Reasoners*. 24th May 2022. arXiv: [2205.11916 \[cs.CL\]](https://arxiv.org/abs/2205.11916). url: <https://arxiv.org/abs/2205.11916>.
- [34] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever and K. Cobbe. *Let's Verify Step by Step*. Introduces the MATH500 verification benchmark. 31st May 2023. doi: [10.48550/arXiv.2305.20050](https://doi.org/10.48550/arXiv.2305.20050). arXiv: [2305.20050 \[cs.LG\]](https://arxiv.org/abs/2305.20050). url: <https://arxiv.org/abs/2305.20050>.
- [35] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael and S. R. Bowman. *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*. 20th Nov. 2023. arXiv: [2311.12022 \[cs.AI\]](https://arxiv.org/abs/2311.12022). url: <https://arxiv.org/abs/2311.12022> (visited on 23/02/2025).
- [36] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candés and T. Hashimoto. *s1: Simple test-time scaling*. arXiv:2501.19393v3. 31st Jan. 2025. arXiv: [2501.19393 \[cs.CL\]](https://arxiv.org/abs/2501.19393). url: <https://arxiv.org/abs/2501.19393>.
- [37] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse and J. Schulman. *Training Verifiers to Solve Math Word Problems*. 27th Oct. 2021. doi: [10.48550/arXiv.2110.14168](https://doi.org/10.48550/arXiv.2110.14168). arXiv: [2110.14168 \[cs.LG\]](https://arxiv.org/abs/2110.14168). url: <https://arxiv.org/abs/2110.14168>.
- [38] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery and D. Zhou. 'Self-Consistency Improves Chain of Thought Reasoning in Language Models'. In: *The Eleventh International Conference on Learning Representations*. 2023. url: <https://openreview.net/forum?id=1PL1NIMMrw>.
- [39] Y. Li, C. Yang and A. Ettinger. *When Hindsight is Not 20/20: Testing Limits on Reflective Thinking in Large Language Models*. NAACL 2024 Findings (camera-ready). 14th Apr. 2024. arXiv: [2404.09129 \[cs.CL\]](https://arxiv.org/abs/2404.09129). url: <https://arxiv.org/abs/2404.09129>.
- [40] K. Kang, A. Setlur, D. Ghosh, J. Steinhardt, C. Tomlin, S. Levine and A. Kumar. *What Do Learning Dynamics Reveal About Generalization in LLM Reasoning?* 12th Nov. 2024. doi: [10.48550/arXiv.2411.07681](https://doi.org/10.48550/arXiv.2411.07681). arXiv: [2411.07681 \[cs.LG\]](https://arxiv.org/abs/2411.07681). url: <https://arxiv.org/abs/2411.07681>.

- [41] Y. Razeghi, R. L. Logan IV, M. Gardner and S. Singh. ‘Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning’. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Y. Goldberg, Z. Kozareva and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 840–854. doi: [10.18653/v1/2022.findings-emnlp.59](https://doi.org/10.18653/v1/2022.findings-emnlp.59). url: <https://aclanthology.org/2022.findings-emnlp.59/>.
- [42] A. Schwarzschild, Z. Feng, P. Maini, Z. C. Lipton and J. Z. Kolter. ‘Rethinking LLM Memorization through the Lens of Adversarial Compression’. In: *Advances in Neural Information Processing Systems 37*. NeurIPS 2024 (10th–15th Dec. 2024). Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Dec. 2024, pp. 56244–56267. doi: [10.52202/079017-1790](https://doi.org/10.52202/079017-1790). url: <https://doi.org/10.52202/079017-1790> (visited on 19/12/2025).
- [43] M. Jiang, K. Z. Liu, M. Zhong, R. Schaeffer, S. Ouyang, J. Han and S. Koyejo. *Investigating Data Contamination for Pre-training Language Models*. 11th Jan. 2024. doi: [10.48550/arXiv.2401.06059](https://doi.org/10.48550/arXiv.2401.06059). arXiv: [2401.06059](https://arxiv.org/abs/2401.06059) [cs.CL]. url: <https://arxiv.org/abs/2401.06059>.
- [44] V. Sanh, L. Debut, J. Chaumond and T. Wolf. *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*. 2nd Oct. 2019. arXiv: [1910.01108](https://arxiv.org/abs/1910.01108) [cs.CL]. url: <https://arxiv.org/abs/1910.01108> (visited on 23/02/2025).
- [45] E. Zelikman, Y. Wu, J. Mu and N. D. Goodman. *STaR: Bootstrapping Reasoning With Reasoning*. Version 2, 2022-05-20. 28th Mar. 2022. arXiv: [2203.14465](https://arxiv.org/abs/2203.14465) [cs.LG]. url: <https://arxiv.org/abs/2203.14465> (visited on 22/12/2025).
- [46] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee and Y. Li. *Textbooks Are All You Need*. 20th June 2023. doi: [10.48550/arXiv.2306.11644](https://doi.org/10.48550/arXiv.2306.11644). arXiv: [2306.11644](https://arxiv.org/abs/2306.11644) [cs.CL]. url: <https://arxiv.org/abs/2306.11644>.
- [47] M.-A. Côté, Á. Kádár, X. Yuan, B. Kybartas, T. Barnes, E. Fine, J. Moore, R. Y. Tao, M. Hausknecht, L. E. Asri, M. Adada, W. Tay and A. Trischler. ‘TextWorld: A Learning Environment for Text-based Games’. In: *CoRR* abs/1806.11532 (2018).
- [48] R. Wang, P. Jansen, M.-A. Côté and P. Ammanabrolu. ‘ScienceWorld: Is your Agent Smarter than a 5th Grader?’ In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11279–11298. doi: [10.18653/v1/2022.emnlp-main.775](https://doi.org/10.18653/v1/2022.emnlp-main.775). url: <https://aclanthology.org/2022.emnlp-main.775/>.

- [49] M. Hausknecht, P. Ammanabrolu, M.-A. Côté, X. Yuan and K. Narasimhan. 'Interactive Fiction Games: A Colossal Adventure'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (3rd Apr. 2020), pp. 7903–7910. issn: 2374-3468. doi: [10.1609/aaai.v34i05.6297](https://doi.org/10.1609/aaai.v34i05.6297). url: <https://ojs.aaai.org/index.php/AAAI/article/view/6297>.
- [50] K. Narasimhan, T. Kulkarni and R. Barzilay. 'Language Understanding for Text-based Games using Deep Reinforcement Learning'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015.
- [51] P. Ammanabrolu and M. Riedl. 'Playing Text-Adventure Games with Graph-Based Deep Reinforcement Learning'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3557–3565. doi: [10.18653/v1/n19-1358](https://doi.org/10.18653/v1/n19-1358). url: <https://aclanthology.org/N19-1358>.
- [52] S. Yao, R. Rao, M. Hausknecht and K. Narasimhan. 'Keep CALM and Explore: Language Models for Action Generation in Text-based Games'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 8736–8754. doi: [10.18653/v1/2020.emnlp-main.704](https://doi.org/10.18653/v1/2020.emnlp-main.704). url: <https://aclanthology.org/2020.emnlp-main.704/>.
- [53] B. Y. Lin, Y. Fu, K. Yang, F. Brahman, S. Huang, C. Bhagavatula, P. Ammanabrolu, Y. Choi and X. Ren. *SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks*. arXiv preprint (NeurIPS 2023 spotlight version). 27th May 2023. doi: [10.48550/arXiv.2305.17390](https://doi.org/10.48550/arXiv.2305.17390). arXiv: [2305.17390](https://arxiv.org/abs/2305.17390) [cs.CL]. url: <https://arxiv.org/abs/2305.17390>.
- [54] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang and T. Gui. *The Rise and Potential of Large Language Model Based Agents: A Survey*. 14th Sept. 2023. arXiv: [2309.07864](https://arxiv.org/abs/2309.07864) [cs.AI]. url: <https://arxiv.org/abs/2309.07864> (visited on 25/02/2025).
- [55] X. Huang, W. Liu, X. Chen, X. Wang, H. Wang, D. Lian, Y. Wang, R. Tang and E. Chen. *Understanding the Planning of LLM Agents: A Survey*. 5th Feb. 2024. doi: [10.48550/arXiv.2402.02716](https://doi.org/10.48550/arXiv.2402.02716). arXiv: [2402.02716](https://arxiv.org/abs/2402.02716) [cs.AI]. url: <https://arxiv.org/abs/2402.02716>.
- [56] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei and J. Wen. 'A Survey on Large Language Model Based Autonomous Agents'. In: *Frontiers of Computer Science* 18.6, 186345 (22nd Mar. 2024). issn: 2095-2228. doi: [10.1007/s11704-024-40231-1](https://doi.org/10.1007/s11704-024-40231-1). url: <https://link.springer.com/article/10.1007/s11704-024-40231-1>.

- [57] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhume, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh and P. Clark. *Self-Refine: Iterative Refinement with Self-Feedback*. Code and data: <https://selfrefine.info/>. 30th Mar. 2023. arXiv: [2303.17651](https://arxiv.org/abs/2303.17651) [cs.CL]. url: <https://arxiv.org/abs/2303.17651>.
- [58] Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma and Y. Liang. *Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents*. 3rd Feb. 2023. arXiv: [2302.01560](https://arxiv.org/abs/2302.01560) [cs.AI]. url: <https://arxiv.org/abs/2302.01560> (visited on 25/02/2025).
- [59] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan and S. Yao. *Reflexion: Language Agents with Verbal Reinforcement Learning*. 20th Mar. 2023. arXiv: [2303.11366](https://arxiv.org/abs/2303.11366) [cs.AI]. url: <https://arxiv.org/abs/2303.11366> (visited on 23/02/2025).
- [60] W. Zhang, Y. Shen, L. Wu, Q. Peng, J. Wang, Y. Zhuang and W. Lu. 'Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives'. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by L.-W. Ku, A. Martins and V. Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3602–3622. doi: [10.18653/v1/2024.acl-long.197](https://doi.org/10.18653/v1/2024.acl-long.197). url: <https://aclanthology.org/2024.acl-long.197/>.
- [61] M. T. J. Spaan. 'Partially Observable Markov Decision Processes'. In: *Reinforcement Learning*. Ed. by M. Wiering and M. van Otterlo. Adaptation, Learning, and Optimization. Springer Berlin Heidelberg, 2012, pp. 387–414. isbn: 978-3-642-27645-3. doi: [10.1007/978-3-642-27645-3_12](https://doi.org/10.1007/978-3-642-27645-3_12). url: https://doi.org/10.1007/978-3-642-27645-3_12.
- [62] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. R. Narasimhan and Y. Cao. 'ReAct: Synergizing Reasoning and Acting in Language Models'. In: *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview.net, 1st May 2023. url: https://openreview.net/forum?id=WE_vluYUL-X (visited on 22/12/2025).
- [63] R. Li, Iañes Machado, Z. Lin, H. Jin, K. Kahatapitiya, A. Shih and A. Talwalkar. *The Effects of Self-Reflection of LLM Agents*. 2024.
- [64] Z. Zhang et al. *PROAgent: Building Proactive Cooperative Agents with Multi-level Planning*. 2024.
- [65] Z. Zhang et al. *Agent-Pro: Learning to Evolve Policy-Level Agent Skills*. 2024.
- [66] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian and X. Xie. *Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations*. Version 3, 2024-01-30. 31st Aug. 2023. arXiv: [2308.16505](https://arxiv.org/abs/2308.16505) [cs.IR]. url: <https://arxiv.org/abs/2308.16505>.

- [67] W. Yao, S. Heinecke, J. C. Niebles, Z. Liu, Y. Feng, L. Xue, R. N. Rithesh, Z. Chen, J. Zhang, D. Arpit, R. Xu, P. L. Mui, H. Wang, C. Xiong and S. Savarese. ‘Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization’. In: *The Twelfth International Conference on Learning Representations*. Vienna, Austria: OpenReview.net, 7th May 2024. url: <https://openreview.net/forum?id=K0Zu91CzbK> (visited on 22/12/2025).
- [68] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler and M. Hausknecht. *ALFWorld: Aligning Text and Embodied Environments for Interactive Learning*. ICLR 2021. 8th Oct. 2020. arXiv: [2010.03768](https://arxiv.org/abs/2010.03768) [cs.CL]. url: <https://arxiv.org/abs/2010.03768> (visited on 19/12/2025).
- [69] J. He, J. Chen, X. He, J. Gao, L. Li, L. Deng and M. Ostendorf. ‘Deep Reinforcement Learning with a Natural Language Action Space’. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by K. Erk and N. A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1621–1630. doi: [10.18653/v1/P16-1153](https://doi.org/10.18653/v1/P16-1153). url: <https://aclanthology.org/P16-1153/>.
- [70] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving and I. Gabriel. *Ethical and Social Risks of Harm from Language Models*. 8th Dec. 2021. arXiv: [2112.04359](https://arxiv.org/abs/2112.04359) [cs.CL]. url: <https://arxiv.org/abs/2112.04359> (visited on 25/02/2025).

3

Knowledge Injection: Mitigating Blind Spots and Adapting to New Contexts

This chapter develops targeted knowledge injection as a mechanism to improve language model robustness. First, we propose an agent-in-the-loop framework that diagnoses blind spots — clusters of unknown unknowns (UUs) where models err with high confidence — and synthesizes targeted training data to mitigate them and thus improve model robustness. Across classification tasks, we show that this approach reduces UUs substantially while preserving overall accuracy. Second, we propose zero-shot contextual adaptation for embedding models. Here, we construct a compact synthetic proxy corpus from a handful of exemplar documents to emulate domain statistics and provide contextual knowledge at inference time. Without accessing the target corpus, this approach attains near-parity with corpus-aware methods. We further motivate the role of contextual knowledge in multimodal settings, particularly in domains where visual cues disambiguate language. Together, these reactive and proactive mechanisms show how targeted synthetic knowledge can measurably alleviate failure modes and enable adaptation under constrained settings.

Parts of this chapter are based on work published in:

- *Illuminating Blind Spots of Language Models with Targeted Agent-in-the-Loop Synthetic Data* [1].
- *Zero-Shot Contextual Embeddings via Offline Synthetic Corpus Generation* [2].
- *Context-Informed Machine Translation of Manga using Multimodal Large Language Models* [3].

3.1. OVERVIEW AND MOTIVATION

Robust language model performance depends not only on increasing the amount of training data, but also on incorporating the right knowledge or context at the right time. Diagnostics from Pillar 1 ([Chapter 2](#)) show that models can succeed by replicating stylistic patterns rather than by explicitly learning to reason from first principles. Models that lack robustness are more likely to exhibit *blind spots*, where they fail confidently on inputs that differ from learned content [4–7]. This risk aligns with long-standing observations on out-of-distribution data and adversarial fragility in language models [8–10]. This chapter presents Pillar 2, operationalizing *knowledge injection* as the targeted introduction of critical information — via synthesized data at training time or (synthetic) context at inference time — to remediate such gaps and improve model performance and robustness without indiscriminately changing model behavior.

We ground the chapter in the following research question from [Chapter 1](#):

RQ2. How effectively can an agent-driven approach that identifies model blind spots and generates targeted synthetic data mitigate specific failure modes compared to untargeted data augmentation approaches?

Unlike generic data augmentation, targeted knowledge injection is *conditioned* on observed failure modes and is designed to either adjust the model’s parameters or context to alleviate any such failures, while preserving or improving overall accuracy.

When focusing on factual knowledge, knowledge injection corrects or enriches specific concepts by adding minimally sufficient data that disambiguates entities, relations, or facts; for example, inserting a counterexample that clarifies an acronym’s expansion in a domain where multiple possible expansions exist. When focusing on contextual knowledge, knowledge injection supplies distributional and situational cues that condition representations at inference time to resolve ambiguity. Structurally, knowledge injection shapes thinking processes so that the model attends to task-relevant information in the best possible way, such as template tags or schema fields that make roles explicit. Across all of these, knowledge injection addresses a distinct absence of crucial knowledge — missing facts, missing context, or missing structure.

Two complementary mechanisms instantiate this approach. First, we propose an *agent-in-the-loop* pipeline that discovers, characterizes, and mitigates blind spots by synthesizing minimal, targeted examples that improve language model robustness when added to the training data. Second, *zero-shot contextual adaptation* is shown to be effective in replacing direct access to a target corpus with a compact, synthetic proxy corpus generated offline from a handful of exemplar documents, enabling context-aware embeddings and zero-shot adaptation without accessing potentially private or unavailable data. Finally, a *multimodal* case study in manga translation studies how and why visual cues function as contextual knowledge that resolves linguistic ambiguity and supports high performance in low-resource settings, and proposes a multimodal approach to best inject visual context for robust translation performance.

In this chapter, [Section 3.2](#) introduces a knowledge injection pipeline for blind spot discovery and targeted sample synthesis. Next, [Section 3.4](#) introduces zero-shot contextual adaptation and its offline synthetic proxy-corpus generation. Finally,

Section 3.6 motivates contextual knowledge through a multimodal translation case and studies its effectiveness. Section 3.7 synthesizes findings and bridges to Pillar 3, value-sensitive AI alignment.

To make RQ2 empirically actionable for the remainder of the chapter, we articulate the hypotheses that guide our evaluation.

Hypothesis H2a An agent-driven targeted synthesis approach significantly reduces unknown-unknown (UU) failures compared to untargeted data augmentation, while preserving overall accuracy, as measured across diverse tasks including classification, factuality, and domain adaptation scenarios.

Hypothesis H2b Agent-identified blind spots correlate strongly with human expert assessments of model vulnerabilities, indicating that the diagnostic patterns surfaced by the agent faithfully capture the underlying failure modes.

Part I (Section 3.2) tests H2a directly by comparing UU rates and accuracy against augmentation baselines, and probes H2b by aligning agent-discovered patterns with human assessments. Part II (Section 3.4) evaluates H2a in the context of zero-shot adaptation, asking whether compact synthetic proxy corpora close the performance gap to corpus-aware methods without degrading global accuracy. The multimodal case study of Part III (Section 3.6) provides complementary evidence on injecting contextual knowledge to remediate ambiguity.

3.2. PART I: AGENT-IN-THE-LOOP BLIND SPOT MITIGATION

3.2.1. INTRODUCTION

Language models (LMs) have achieved remarkable accuracy across a wide range of predictive tasks, but remain vulnerable to out-of-distribution data [8–10]. Small, lightweight LMs – while easier to train and run on limited hardware, and therefore favored in domain-specific applications – are especially prone to UUs due to their reduced robustness [11, 12]. Larger LMs, although generally more robust, require significant computational resources for both training and inference, limiting their usability [13]. This vulnerability often leads to prediction errors, including in high-stakes applications such as suicide prevention [14] and criminal justice sentencing [15], where reliable and unbiased predictions are critical. A particularly challenging class of errors, referred to as *unknown unknowns* (UUs), occurs when the model confidently misclassifies an input as the incorrect label [4]. These UUs tend to *cluster* into *blind spots* in the feature space, areas where the model consistently produces high-confidence misclassifications due to biases in the training data [7, 16]. On the left side of Figure 3.1 we show an example of a mispredicted label at a high confidence, resulting in a UU, that forms part of a blind spot.

The identification of UUs and blind spots has been extensively studied [4–7], including approaches involving human oversight to aid in detection [17, 18]. *Mitigating* blind spots – especially how to move from identified blind spots to unseen ones – remains an unresolved challenge. Simple approaches to tackling only *already discovered* blind spots, such as relabeling previously identified UUs and using them for additional

training [18], do not scale and fall short of ensuring a holistic reduction in blind spots. Thus the only blind spots of the model that can be illuminated using such reactive approaches are those that correspond to seen data, with those that correspond to unseen data remaining out of reach.

In this chapter, we introduce an agent-in-the-loop workflow that proactively mitigates blind spots of LMs by employing intelligent agents – either humans or large LMs – to *characterize* blind spots and subsequently generate targeted synthetic data. We pose that the key to mitigating these blind spots lies in the generalization abilities of the agent, allowing them to hypothesize patterns of discovered UUs and similarities between seen and unseen UUs using prior knowledge [19–21]. To this end, we guide agents to formulate these hypotheses in natural language, either describing the found blind spot consisting of discovered UUs (abstraction) or reasoning about undiscovered blind spots (extrapolation), as is shown in Figure 3.1. Using these hypotheses, we guide agents toward the generation of synthetic samples targeted at blind spots, improving the robustness of LMs through subsequent retraining by reducing the number of high-confidence misclassifications without sacrificing overall predictive accuracy. Our workflow is designed to flexibly integrate intelligence from both humans and LMs, with specific mechanisms to incorporate human computation or LMs. Additionally, the workflow can incorporate existing adversarial attack methods to proactively illuminate blind spots, further enhancing its adaptability and effectiveness.

Our workflow proves to be a viable means of distilling knowledge from intelligent agents to small LMs, making them more robust while maintaining their lightweight advantages. Through our comprehensive experiments, we find that our method is capable of substantially reducing the number of high-confidence misclassifications without decreasing accuracy. On average, we are able to reduce the number of UUs by 19.08%. Further, we show that for our method LMs are more effective overall than human agents, achieving a 22.37% reduction in UUs compared to a 15.78% reduction when using human-generated data. Additionally, LM-generated data are far more economical, making them a more scalable solution for improving the robustness of small models. Finally, we observe that humans surpass LMs in certain tasks, particularly those that align more closely with human intuition due to participants’ greater familiarity with them.

3.2.2. AGENT-IN-THE-LOOP TARGETED DATA GENERATION

Our proposed approach to blind spot mitigation involves engaging a human or LM in three tasks: *hypothesis generation via abstraction*, *hypothesis generation via extrapolation*, and *synthetic sample generation*. These tasks are designed to characterize and mitigate blind spots, ultimately reducing high-confidence misclassification. The workflow is schematically illustrated in Figure 3.1. The human computation component of our study is implemented through a survey study, the details of which are provided in Section C.2, while the equivalent LM prompts are given in Section C.3.

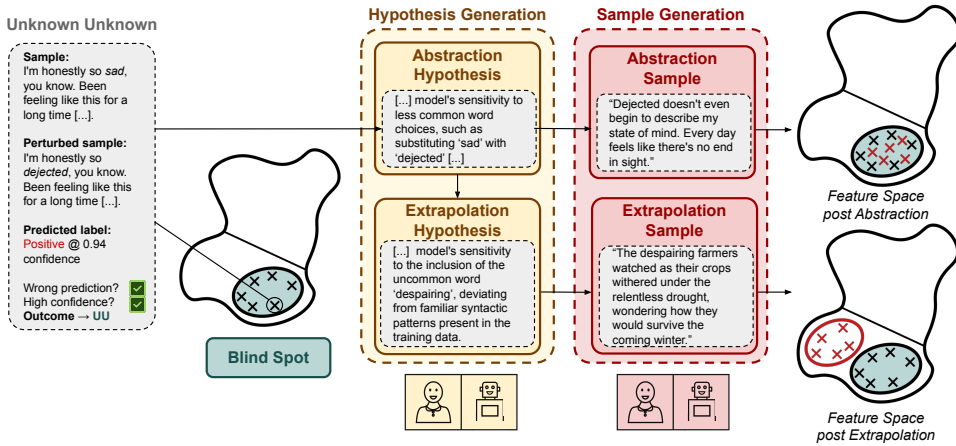


Figure 3.1.: In a sentiment classification task, we begin with a UU resulting from a perturbation — denoted by a cross in the feature space. This UU is then used to generate an initial hypothesis via abstraction through human computation or an LM. This abstraction hypothesis can then either be used to generate a synthetic sample that target the existing blind spot or to generate a new hypothesis via extrapolation, which in turn is then used to generate synthetic samples targeting an unseen blind spot.

PROBLEM FORMULATION

For UU discovery, let the dataset be $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where x is the original text sample and y the original ground truth label. Without having access to y , a predictive model θ is tasked with generating a label prediction $y_p = \theta(x)$ at a confidence $c \in [0, 1]$. Formally, a UU occurs when (1) θ predicts the wrong label $y_p \neq y$ and (2) the prediction is made with high confidence $c \geq \tau$.

In this work, in addition to dealing with the blind spots that naturally occur in models as a result of training, we make use of adversarial UU discovery, where we increase the number of misclassifications by introducing perturbations. For this, a black-box adversarial perturbation model G generates perturbed samples $\bar{x} = G(x)$, where $\bar{x} \neq x$. The model θ is then used to predict new labels $y'_i = \theta(\bar{x}_i)$ at a confidence c . The resulting perturbed dataset, denoted \mathcal{P} , consists of the new samples and predicted labels (\bar{x}, y') . If a perturbation occurs, there is an additional requirement for a misclassification to be considered a UU: (3) \bar{x} , regardless of its label indicated by θ , maintains the same underlying true label y as x post perturbation.

Given a predictive model θ trained on a dataset \mathcal{D} , our objective is to mitigate UUs produced by θ . To systematically reduce high-confidence misclassification, we seek to identify patterns in discovered UUs and generate targeted synthetic data $\{x^s, y^s\}$ for a set of UUs, where x^s is the synthetic sample and y^s represents the corresponding ground truth label for the synthetic sample. This data is then used to further train θ and thus reduce the blind spots present.

3.2.3. GENERALIZATION VIA HYPOTHESIS CREATION

For UU mitigation, we employ intelligent agents (humans or large LMs) to generalize from identified UUs to create hypotheses in natural language regarding the underlying causes of these UUs. As we use perturbations, such hypotheses are based on pairs of original and perturbed samples, $(x_i, y_i) \sim \mathcal{D}$ and $(\tilde{x}_i, y'_i) \sim \mathcal{P}$. Humans are adept at using sparse data to generalize [22], and this task exploits that capability by focusing on subsets of UUs. Each hypothesis describes the shared characteristics that explain why certain UUs occur and how these characteristics might generalize to other, unseen UUs. The goal is not merely to explain individual failure cases but to construct hypotheses that address multiple UUs clustering together into a blind spot. In doing so, we can illuminate patterns within the feature space that the model is consistently misclassifying. To this end, we pursue two distinct but complementary strategies: abstraction and extrapolation.

Abstraction Abstraction involves generating a hypothesis on why a specific UU occurred that generalizes across a set of closely related UUs, revealing underlying patterns within a blind spot. In this step, the intelligent agent is provided with an original sample (x_i, y_i) and, if adversarial perturbations are used, its perturbed counterpart (\tilde{x}_i, y'_i) . Then the agent is tasked with reasoning abstractly about the factors leading to this UU. Specifically, we instruct them to consider whether these factors involve semantics, syntax, specific words, or something else in the samples that could be the cause of the high-confidence misclassification. This is to guide the agent to identify what most likely contributes to the UU without prescribing rigid criteria, leaving room for creative thinking and allowing the agent to explore unforeseen or nuanced factors. The hypothesis is in natural language and should generalize across other UUs that share these characteristics, expanding our understanding of the particular blind spot the UU corresponds to. Compared to a mitigation approach that only makes use of a simple reactive relabeling of found UUs, our method comes with the additional advantage that it builds up a corpus of human-interpretable error reports on seen errors of the classification model.

Extrapolation Extrapolation extends the process of hypothesis creation beyond trying to describe discovered blind spots, encouraging the agent to use existing hypotheses and sample pairs (used during abstraction) to uncover new blind spots. This task emphasizes extrapolation, asking the agent to hypothesize new failure modes – also in natural language – that differ from those previously identified. Extrapolative thinking has previously been shown to be a human strong suit [23]. By ensuring that the new hypotheses are dissimilar from those used for abstraction, we aim to discover new regions in the feature space where the model may be prone to high-confidence misclassification. To avoid the agent overextrapolating, we specifically instruct them to focus on the same topic but to reason about whether a different factor from semantics, syntax, specific words might be responsible that was not mentioned in the abstraction hypothesis. In this step, we present only human-generated hypotheses to human participants and vice versa. An example of hypothesis generation via abstraction and extrapolation is shown in [Figure 3.3](#).

3.2.4. SYNTHETIC SAMPLE GENERATION

Once hypotheses have been generated via abstraction or extrapolation, the agent is tasked with generating synthetic samples. These synthetic samples must align with the structure and context of the original dataset while reflecting the characteristics of the generated hypotheses. For instance, if the dataset consists of movie reviews, the synthetic samples should maintain the form and tone of movie review-related text. The goal of this step is to create new data points that correspond to the blind spots identified during hypothesis generation. These synthetic samples are added to the training dataset, resulting in a dataset that is extended for each synthetic sample and its corresponding label $\mathcal{E} = \mathcal{D} \cup \{x_i^s, y_i^s\}$, where the label is provided by the agent. By incorporating these new samples into training, we aim to enhance the robustness of the predictive model θ by reducing its susceptibility to high-confidence misclassifications. The sample generation process is uniform, regardless of whether the hypothesis was obtained through abstraction or extrapolation. Humans generate samples based on human-created hypotheses, and LMs do the same for LM-generated hypotheses. An example of this type of sample generation from human and LM agents for abstraction and extrapolation is shown in [Figure 3.3](#).

3.2.5. EXPERIMENTAL SETUP

In this section, we present an overview of our experimental design. A schematic illustration of the workflow can be found in [Figure 3.2](#). First, we obtain our initial set of UUs of the finetuned classification model from the validation set. Following this, we characterize the blind spots corresponding to these UUs by making the intelligent agent perform generalization as described in [Section 3.2.2](#), culminating in new synthetic data that we use to retrain the model. Finally, we evaluate this retrained model with respect to accuracy and UU count. As a preliminary study, to verify that our method does indeed address blind spots, we successfully demonstrate that it is possible to artificially create blind spots by hand (*i.e.*, ground truth blind spots) in a model and then illuminate these using our approach in [Section C.1](#). In our main study, our experiments instead address mitigating both natural blind spots that occur during normal model training and those created by adversarial attacks. For this, we do not have access to the ground truth blind spots and as such just have indirect evidence that some blind spots are illuminated as the number of occurring UUs is decreased.

DATASETS, MODELS, AND PERTURBATIONS

To evaluate the generality and effectiveness of our approach, we select a diverse set of classification tasks, each representing varying levels of task complexity. Specifically, we focus on sentiment analysis (SA) using the IMDB dataset [24], semantic equivalence (SE) using the MRPC dataset [25], and natural language inference (NLI) using the QNLI dataset [26]. The statistics of the dataset for each task are shown in [Table 3.1](#). For blind spot mitigation, we use the validation set to obtain our UUs that are then used to perform the hypotheses generalization. These hypotheses are then used in turn to generate synthetic samples and extend the training set, as shown in [Figure 3.2](#). We limit the number of hypotheses derived from each of abstraction and extrapolation

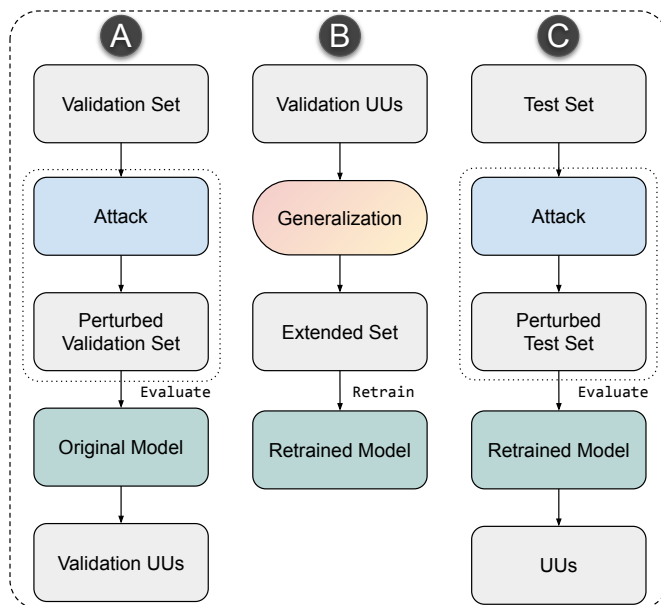


Figure 3.2.: Workflow for targeted blind spot mitigation with synthetic data: (A) Obtain UUs from the validation set on the original finetuned model; (B) use UUs to extend the training data via generalization (Figure 3.1) and thus obtain a more robust model; (C) evaluate this retrained model. The adversarial perturbations in the dotted box are optional.

to 1% of the training set size, leading to an additional 73, 500, and 2095 training samples after applying our method for MRPC, IMDB, and QNLI, respectively. These values are treated as hyperparameters and are chosen to balance computational efficiency and effectiveness. We leave further optimization of this split between abstraction- and extrapolation-derived hypotheses to future work. We employ two classification models in our experiments, finetuned for each classification task: BERT (`bert-base-uncased`) [27] and Llama 2 7B [13], selected for their contrasting architecture and size. We choose BERT for its known performance on sentence-level classification tasks and its low number of parameters, while Llama 7B was chosen for its larger (but still manageable) scale and capability in handling more complex language understanding tasks. GPT-3.5 Turbo is incorporated as the teacher model to perform hypothesis and sample generation, as it is superior to both classification models that we use.

In a black-box setting, where we assume no access to the model's internal parameters, we employ adversarial perturbation techniques to yield more UUs for our method to use. Note that while perturbations aid proactive discovery of blind spots, they are not strictly necessary to our overall approach. Perturbations are generated using TextAttack [28], specifically with TextFooler (TF) [29] for word-level perturbations and DeepWordBug (DWB) [30] for character-level perturbations. Using these two methods, we cover a wide

Example of full workflow for a single UU sample using human computation and LMs on the IMDB Dataset

Original Sample:

Footage of busy highways, planes flying overhead, poor lightning effects, soldiers doing their thing, they're all used in a [[haphazard]] manner, sometimes repeatedly, and they often only have a loose connection to the story Wood is trying to tell. [...]

Perturbed Sample:

Footage of busy highways, planes flying overhead, poor lightning effects, soldiers doing their thing, they're all used in a [[thoughtless]] manner, sometimes repeatedly, and they often only have a loose connection to the story Wood is trying to tell. [...]

Human Abstraction Hypothesis:

"Haphazard can be intentional, vs thoughtless is unplanned. It feels more of an unwillingness rather than a lack of ability."

LM Abstraction Hypothesis:

"The replacement of specific adjectives in the perturbed sample contributes to misclassification. The adjective "thoughtless" in the perturbed sample, may change the perception of the director's skills and use of stock footage. This alteration in sentiment could lead to a different classification outcome."

Human Extrapolation Hypothesis:

"The hypothesis could be that the words were potentially swapped or replaced, such as "bad" being replaced by "evil" which could lead to an incorrect prediction as it is similar but could have a slightly different meaning in some contexts."

LM Extrapolation Hypothesis:

"The use of the word "hypocrite" may have caused the misclassification. The word "hypocrite" implies that the person is saying one thing but doing another, which may be considered a negative trait. However, some people may not interpret the person in the sample as a hypocrite, leading to a difference in sentiment analysis. This difference in interpretation may have caused the misclassification by the algorithm."

Human Abstraction Sample:

"The scenes felt like they were slapped together with no care at all. It's not that the director didn't have the skill they just didn't bother to try harder. Haphazard is at least creative but thoughtless just shows they didn't wanna put the effort in."

LM Abstraction Sample:

"Skyward Dreams had potential, but the director's use of stock footage feels thoughtless. Scenes of bustling cities and planes flying overhead are inserted without any real purpose, disrupting the flow of the narrative. [...]"

Human Extrapolation Sample:

"The CGI in was straight up evil. The way the effects looked completely ruined the immersion for me, and it felt like the creators didn't even care about quality. I get that sometimes budget is an issue, but this was just on another level. [...]"

LM Extrapolation Sample:

"The protagonist of The Final Betrayal is a true hypocrite. Throughout the film, he preaches loyalty and honesty to his friends, yet secretly manipulates and betrays them behind their backs. This hypocrisy is central to the film's conflict, as the character's outward morality sharply contrasts with his deceitful actions. Despite this glaring flaw, some viewers may interpret his behavior as a survival tactic in a harsh world, rather than outright hypocrisy. [...]"

Figure 3.3.: Example of hypothesis generalization using *abstraction* for the IMDB dataset. The abstraction is performed by a human or LM based on original and perturbed samples.

spectrum of adversarial attack types, revealing additional blind spots. We focus on perturbations that maintain semantic integrity, ensuring that the true underlying label remains consistent after perturbation. Manual inspection of 100 random perturbed samples revealed that none had a different underlying true label, affirming that our perturbations are faithful.

BASELINE

As a baseline, we use a reactive relabeling approach based on the previous work by Han, Dong and Demartini [18], where identified UUs are given a ground truth label, before being reintroduced to the classification model for additional training. This method directly targets blind spots by adding these correctly labeled samples to the extended set. While Han, Dong and Demartini [18] perform this reintroduction in smaller, iterative batches to identify more UUs, we pool all relabeled UUs in a single batch, as we only concern ourselves with the mitigation of UUs and assume that we have knowledge of whether a sample is a UU or not post classification. This is similar to how we perform the retraining for our method. For a fair comparison, we apply this baseline approach with the same budgetary constraints as our proposed method, with new samples making up 2% of the initial training set size. We pose that our method, which uses hypotheses to synthesize new data, will outperform this method by uncovering additional failure modes not captured by relabeling alone.

IMPLEMENTATION

Following Lakkaraju *et al.* [16], we set the confidence threshold for determining high-confidence misclassifications to $\tau = 0.65$. All BERT models were trained for 10 epochs, using a learning rate of 2×10^{-4} , and a batch size of 64. We fine-tune all Llama 7B models using the Low-Rank Adaptation (LoRA) [31] method with the following configuration: a LoRA scaling factor of 16, dropout of 0.1, and rank $r = 64$. The target modules are all linear layers in the model, and no bias adjustment is applied. The training is performed over 3 epochs, with a batch size of 8, and gradient accumulation set to 8 steps. We employ AdamW [32] as our optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to 2×10^{-4} with a warmup ratio of 0.1, followed by a cosine decay. We apply a maximum gradient norm of 0.3 to ensure stability during training and use a weight decay of 0.001 to prevent overfitting.

The human computation component of our study is implemented through a survey study, the details of which are provided in Section C.2. A key procedural difference between human and LM-based experiments is the number of examples provided. The human participants receive two examples, while no examples are given to LMs (*i.e.*, zero-shot). This design choice aims to minimize guidance for the LM since few-shot prompting tends to result in overly homogeneous samples, even when using higher temperature settings. The LM prompts for the teacher model are given in Section C.3. When prompting the teacher model, we always ask it to explicitly give its reasoning, which we find not only increases performance but also improves interpretability. To ensure the quality of human-generated hypotheses and synthetic samples, we include attention checks [33] in each survey to eliminate inattentive or low-effort responses. For

Dataset	Task	#Classes	#Train	#Validation	#Test
MRPC	SE	2	3,668	408	1,725
IMDB	SA	2	25,000	12,500	12,500
QNLI	NLI	2	104,743	5,463	5,463

Table 3.1.: Datasets used, including the task type, number of classes, and number of samples in each of the test, validation, and training sets. Note the split of the original IMDB test set into new validation and test sets.

both human- and LM-generated hypotheses and samples, we implement automated quality checks for this purpose. We do not focus on selecting the high-quality responses, but filter out bad-faith ones such as repeated or nonsensical submissions. To be included, all text entries are required to meet a minimum character threshold ($\text{char}_{\min} = 40$) to ensure sufficient content. Additionally, we employed BERTScore [34] to automatically evaluate the similarity of new samples against a reference set in the form of samples from the training set. If the similarity score falls below a threshold of $S_{\min} = 0.5$, the sample is discarded.

3.2.6. EVALUATION METRICS

We use two key metrics to assess the effectiveness of our approach and the comparative approach. These include the accuracy of the model on the test set and the number of UUs observed during evaluation. Accuracy provides a basic measure of model performance, while the UU count reflects the model’s robustness and allows us to reason about the prevalence of blind spots. Note that the accuracy we report is the accuracy of the model before any perturbations are applied, while the number of UUs is post perturbation. Ideally, our goal is to maximize accuracy while minimizing the number of UUs. Our evaluation compares the performance of the original finetuned model with that of the models retrained on their respective extended dataset \mathcal{E} . This allows us to quantify the impact of our approach on mitigating blind spots and improving model robustness.

3.2.7. RESULTS

In this section, we report the experimental results on the effectiveness of our proposed method in reducing blind spots across the classification tasks. The results of our methods configured with human- and LM-generated data as well as those of the baselines are shown in Table 3.2. Additionally, we compare human-generated samples to those produced by LMs in terms of effectiveness, scalability, and ease of use.

IMPACT OF SYNTHETIC SAMPLES

Observation 1: Our approach leads to a significant and consistent UU reduction across tasks. As part of our evaluation, we find that our method successfully reduces UUs, with a maximum reduction of 56.09% when using human computation on the

BERT model with TF for the MRPC task. LMs generally offer more consistent UU reductions, though performance varies by task. On average, across perturbation methods and classification models, our method with LM-based data generation reduced UUs by 23.43%, while human-based data generation led to a reduction of 21.68%. Similarly, regardless of what type of agent generates the data, our method achieves an average reduction in UUs of 35.71%, 21.27%, and 10.70% for MRPC, IMDB, and QNLI, respectively. The only configuration where our method does not reduce UUs is the BERT model on the QNLI dataset, where human-based retraining with TF actually increases UUs by 5.46%. We elaborate on this in observation 3.

Observation 2: Relabeling of UU samples is effective but not as impactful. Simply relabeling UU samples from the validation set and reintroducing them as the extended set leads to a decrease in the number of UUs, albeit a more modest one compared to our method. Relabeling achieves a decrease in UUs of 8.86% and 7.06% on average for BERT and Llama 7B, respectively. This compares to the average decreases achieved by our method: 21.68% when using human-generated data and 23.43% when using LM-generated data. This confirms that only reactive illumination of blind spots using seen data is less effective than our method, regardless of agent type. While the average decrease is lower, the relabeling method is very consistent across tasks, as it is not dependent on an agent grasping the task and delivering high quality data. Additionally, it is very cost effective as no human computation or LM querying is necessary. The obvious limitation of this approach is that it only scales to blind spots that have been discovered and therefore has very little transfer learning potential, as it is unlikely that the found UUs will generalize to unseen UUs.

Observation 3: Human performance is very task dependent. We find that human-generated samples may outperform LMs in tasks that align with human intuition. For tasks such as SE and SA – which are more intuitive to humans compared to NLI, as they more closely resemble everyday tasks – human performance tends to be better, yielding more significant reductions in UUs. In particular, on the MRPC dataset we see a greater reduction in UUs using human-generated data, 35.38% and 52.19% on BERT and Llama 7B, respectively, when compared to when using LM-generated hypotheses and samples 8.21% and 47.31%. In less intuitive tasks such as NLI, humans can generate data of poor quality, leading to a reduction in model robustness, which may even result in an increase in UUs. When analyzing participants' responses for QNLI, we find that several participants did not fully grasp the natural language inference task, which was not the case for SE and SA. Note that these are not purposefully low-effort responses and are therefore not filtered out as described in [Section 3.2.5](#). This shows that irrespective of classification model, there is a task-specific advantage of human computation compared to LM teacher models when there exists a higher degree of familiarity with the task and vice versa. Although LMs provide samples of acceptable quality consistently, rare but high-quality human responses, such as a crowdworker correctly identifying that changing the date “June 15” to “John 15” referenced a Bible verse – an insight that the LM missed – can significantly reduce UUs and thus be more impactful. This suggests that while human-generated responses can have a higher ceiling in certain contexts, LMs deliver more consistent results overall as incorrect responses from just a few human participants can reduce the effectiveness of our

		BERT				Llama 7B			
		TF		DWB		TF		DWB	
		Acc. (%) ↑	UUs (#) ↓	Acc. (%) ↑	UUs (#) ↓	Acc. (%) ↑	UUs (#) ↓	Acc. (%) ↑	UUs (#) ↓
MRPC	Original Model	82.38	952	82.38	936	90.84	301	90.66	293
	Relabeling Baseline	82.49	911	82.55	898	90.61	277	90.73	268
	Hypothesis (LM)	81.57	851	82.23	882	89.86	149	89.73	164
	Hypothesis (Human)	81.58	418	82.10	802	90.20	144	89.91	140
IMDB	Original Model	94.84	1882	95.40	1682	95.20	892	95.33	810
	Relabeling Baseline	93.94	1732	94.26	1621	94.86	781	95.10	742
	Hypothesis (LM)	95.40	1241	94.41	1448	94.96	604	95.13	689
	Hypothesis (Human)	94.43	1518	95.74	1412	94.67	658	94.90	702
QNLI	Original Model	89.88	1923	89.88	2597	90.08	879	90.72	952
	Relabeling Baseline	88.24	1796	88.98	1907	89.90	856	90.60	929
	Hypothesis (LM)	89.31	1536	89.21	1746	89.58	741	90.10	890
	Hypothesis (Human)	89.42	2028	89.38	2325	89.16	857	89.73	924

Table 3.2.: Results for ALL blindspot mitigation across datasets and models. Accuracy is unchanged within $\pm 1\%$ while UUs drop substantially after targeted knowledge injection via synthetic data. TF: TextFooler; DWB: DeepWordBug.

method.

Observation 4: Accuracy does not decrease despite improved robustness. In terms of accuracy, extending the training set with human- or LM-generated data did not have a significant effect. Across tasks, accuracy fluctuations of the models with extended training sets remain within $\pm 1\%$ compared to the original models. This contrasts with previous findings that improvements in robustness often come at the expense of accuracy [35]. Detailed perturbation statistics are made available in Section C.4, as well as a visualization of prediction confidences for misclassified samples after perturbation. We observe a reduction in high-confidence misclassifications, particularly at the highest prediction confidences. Additionally, there is a clear reduction across the entire confidence range towards lowering the classifier’s confidence in its misclassifications. This, in combination with our overall results, indicates that we improve the calibration of the classification models.

3.2.8. SCALABILITY AND EASE OF USE

Observation 5: Our method scales well per sample and by parameter count. Despite only adding a small amount (2% for each task) of synthetic data relative to the total training set size, we achieve significant results in the reduction of UUs. This indicates that our method can scale to large datasets, as only a small number of synthetic samples relative to the total dataset size are required to have a significant impact in terms of improving robustness. We study classification models that use a different architecture and have an order of magnitude difference in size (110M parameters for BERT and 7B for Llama). Here, we find that models with a lower number of parameters achieve a performance similar to that of large generative LMs, with comparable accuracy on the IMDB and QNLI tasks, indicating that smaller models may be more suitable for text classification tasks when considering their other advantages, which corroborates previous findings [36]. This is especially encouraging for use cases where computational resources are limited or speed and transparency are critical.

Observation 6: Obtaining samples via LM is easier and more cost effective. When considering the practical aspects of our study, significant insights emerge regarding the costs and time involved in conducting human- and LM-based generalization experiments. The human study, which included 168 participants, resulted in a total cost of \$1072, with an hourly compensation rate of \$12 per participant. In contrast, the LM experiment incurred a much lower cost of \$46 for generating an equivalent number of generalizations and samples. Although it is challenging to provide precise estimates, the data collection process via human surveys also took substantially longer than the LM-based approach. This highlights the fact that when using LMs, our method is far more cost-effective and generates data almost instantaneously, in stark contrast to the considerable delays associated with human-based study design and data collection. Thus, from a scalability perspective, the LM-based procedure offers clear advantages, being both faster and less expensive. However, in certain high-stakes or specialized applications such as suicide prevention and criminal justice sentencing, human involvement, including via a hybrid approach where human intuition supplements the efficiency of LM-generated data, may be more advantageous. This is especially true when considering that LM outputs come with no guarantees and may be biased.

3.2.9. RELATED WORK

Unknown unknowns Attenberg, Ipeiritis and Provost [4] introduce the concept of querying humans to find UUs in a game-like setting and show that there were patterns to the found UUs. Vandenhof [6] proposes an approach to identify UUs where human-interpretable decision rules are learned to approximate how a model makes high-confidence predictions. Crowdworkers then contradict these rules by finding an instance that would classify as a UU. Cabrera *et al.* [17] explore the use of crowdworkers to generate failure reports for computer vision models to describe how or why the model failed. Han, Dong and Demartini [18] propose an approach where crowdworkers continuously extend a dataset with relabeled UUs, on which the chosen model is iteratively trained. Instead, we go beyond simple relabeling and characterize found blind spots and explore new, previously unseen blind spots. There are also algorithmic approaches to finding UUs, such as Lakkaraju *et al.* [16], who propose utilizing an explore-exploit approach to find groups of UUs. Bansal and Weld [5] extend this by proposing a utility model that rewards the degree to which the found UUs cover a sample distribution, thus encouraging the discovery of new blind spots. Instead, we do not find the UUs algorithmically, but instead use an LM or crowdworkers to find existing UUs, extrapolate from these to unseen UUs, and generate synthetic data targeting both of these.

Model calibration and robust training The concept of UUs and blind spots is connected to model calibration [37–39]. A model that is well-calibrated will have its prediction confidence aligned with the likelihood of the correctness of the prediction and, as such, a model with blind spots is a poorly calibrated model. In the case where the UUs are specifically generated through adversarial attacks, illumination of model blind spots is also related to robust training. UUs that populate these blind spots, when created by such attacks, may be identified as adversarial examples [40–42]. This underscores the relationship between our proposed method and robust training

practices with the aim of improving the robustness of the model [43, 44]. Our method focuses not on general robustness but rather on high-confidence misclassifications and is not limited to just adversarial samples, as we consider UUs that occur naturally without perturbation as well. Several approaches have been proposed to utilize synthetic data to expand training sets [45, 46]. He *et al.* [47] explore few-shot prompting LMs to generate task specific synthetic training data. Unlike prior work, we propose a method to generate targeted synthetic data with the purpose of eliminating blind spots that lead to high confidence misclassifications.

3.2.10. CONCLUSION

We propose a method to identify and mitigate blind spots in classification models by leveraging human- and LM-generated generalizations, followed by synthetic sample generation to target UUs and enhance model robustness. Our evaluation demonstrates that our method is effective at addressing model blind spots and achieves a significant reduction in UUs across datasets, while not altering the general performance of the model and therefore maintaining accuracy. Our study sheds light on the notable task dependency of the human ability to characterize blind spots and generate new data and how this ability compares to that of an LM. Future work will focus on optimizing the balance between accuracy and robustness to further enhance model performance.

3.3. FROM REACTIVE REPAIR TO PROACTIVE ADAPTATION

Part I (Section 3.2) operationalized knowledge injection as a *reactive* repair mechanism, where we diagnose high-confidence errors that cluster in the feature space to form blind spots, before hypothesizing what the underlying error in model reasoning is, and injecting targeted synthetic samples to locally adjust the decision boundary and thus illuminate the blind spot, all while keeping global behavior stable. This approach is well suited when we can surface UUs and are able to update parameters. It primarily addresses *systematic* and *structural* shortcomings of language models, where precise, hypothesis-guided data interventions nudge the model away from confident mistakes without broad distributional changes.

In many deployments, however, robustness hinges on *contextual* gaps: the model lacks access to the target corpus or cannot be retrained due to privacy, governance, compute, or other operational constraints. In such cases, rather than adjusting the decision boundary through parameter updates, we can inject *context* at inference time to condition the language model to better adapt to the target domain. Part II (Section 3.4) develops this *proactive* path: for context-aware embedding models, it replaces direct access to the target corpus with a compact *synthetic proxy* generated offline from a handful of exemplars and uses that proxy as contextual knowledge to compute improved embeddings without updating model parameters.

Viewed together, these two separate Parts instantiate a single design principle: whether using targeted synthetic data for additional model training when diagnostics expose blind spots, or using zero-shot contextual adaptation when the dominant deficit is contextual and the model must remain frozen, the goal is to inject the right knowledge to ensure the model is able to reason correctly and robustly.

3.4. PART II: ZERO-SHOT CONTEXTUAL ADAPTATION

3.4.1. INTRODUCTION

Effective neural information retrieval relies heavily on high-quality dense vector representations for documents and queries [48, 49]. However, standard embedding approaches typically generate these representations without dynamically incorporating information from the specific target corpus being searched [50]. This lack of dynamic context sensitivity limits their adaptability and retrieval performance, especially when deployed in domains that differ from their pretraining data [51], a limitation partially addressed by traditional methods leveraging corpus statistics [52]. However, such lexical methods based on surface counts have largely been superseded, as their bag-of-words nature fundamentally fails to capture semantic meaning.

To provide neural embeddings with sensitivity to the target corpus, context-aware architectures have recently emerged [53]. Such models utilize multi-stage processing, where the final embedding is conditioned on representations derived from neighboring documents within the target corpus. Although these methods significantly enhance retrieval by tailoring embeddings to domain-specific characteristics, their requirement to access the target corpus during inference is a critical limitation. Practical constraints related to data privacy or corpus scale often make such corpus access infeasible; for instance, sensitive medical documents may not be exposed during inference.

This critical limitation motivates the search for alternative ways to provide essential domain signals to context-aware models. While large language models (LLMs) possess good generative capabilities for retrieval data [54], their effective use in this zero-access environment is non-trivial. The core challenge is not merely generating domain-relevant text, but ensuring that any LLM-derived output possesses an appropriate degree of representational fidelity.

In this work, to overcome this critical barrier, we introduce zero-shot embeddings via synthetic context (ZEST), enabling effective contextual adaptation without accessing the target corpus. ZEST, depicted in Figure 3.4, operates in two phases. First, during *offline synthesis*, a LLM is employed to hierarchically generate a compact set of domain anchors from a minimal set of example documents randomly chosen from a domain-relevant source. The example documents typify the target domain's characteristics but remain distinct from the actual target corpus data. Using the domain anchors, a synthetic proxy corpus is generated that approximates the actual distribution of the target corpus. Second, during *online inference*, the generated synthetic corpus serves as the contextual input to a pretrained, unmodified context-aware embedding architecture. Thus, ZEST computes context-influenced embeddings, effectively simulating domain adaptation without requiring direct corpus access or any parameter updates.

Our thorough validation on relevant retrieval benchmarks demonstrates that ZEST – using only $k = 5$ example documents – significantly outperforms strong context-agnostic baselines and achieves performance comparable to context-aware models that utilize full corpus access. Although synthesizing a corpus of, in our case, a few hundred documents involves upfront computational resources, this process is performed once per domain, offline, making it practical and cost-effective for real-world deployments. ZEST thus provides a viable and efficient strategy for deploying adaptable document

embeddings in environments constrained by corpus access due to concerns regarding privacy or scale.

3.4.2. BACKGROUND

This section first revisits dense neural retrieval, then describes the class of context-aware embedding architectures most relevant to our work, and finally formalizes the zero-shot deployment setting that ZEST is designed to address.

Dense retrieval. Given a query q and a corpus \mathcal{D} , dense vector retrieval methods learn two neural encoders $\phi : D \rightarrow \mathbb{R}^n$ and $\psi : Q \rightarrow \mathbb{R}^n$ that map documents and queries into a shared vector space in which relevance is scored with the dot product $f(d, q) = \phi(d) \cdot \psi(q)$. Training typically relies on a contrastive loss that favours high similarity for relevant pairs (q, d^+) and low similarity for negatives (q, d^-) :

$$\mathcal{L} = -\log \frac{\exp(f(d^+, q)/\tau)}{\sum_{d' \in \mathcal{N}(q)} \exp(f(d', q)/\tau)}, \quad (3.1)$$

where $\mathcal{N}(q)$ is a set of negatives [55] and τ a temperature hyper-parameter. Although these biencoders are highly effective, their embeddings remain *context-agnostic* and therefore sensitive to corpus shift [51].

Context-aware embedding architectures. Recent retrievers mitigate the limitation of context-agnostic embeddings by incorporating information from the target corpus at inference time. These approaches vary in their mechanism. One prominent strategy, which our work adapts for zero-shot settings, involves a multi-stage architecture exemplified by Contextual Document Embeddings (CDE) [53]. Here, a first-stage encoder M_1 processes a sample of J context documents $\{d_1, \dots, d_J\} \subset D$ from the target corpus. A second-stage encoder M_2 then conditions the final embedding of a target text (document d or query q) on these context representations alongside the text’s own token sequence. Formally, the first-stage encoder M_1 maps an input document to a vector in \mathbb{R}^h . The second-stage encoder is $M_2 : \mathbb{R}^{J \times h} \times \mathbb{R}^{T \times e} \rightarrow \mathbb{R}^n$, so that, given a document d whose tokens are embedded as $E(d) = \{E(w_1), \dots, E(w_T)\}$, the contextual document representation is a single dense vector:

$$\phi(d; \mathcal{D}) = M_2(M_1(d_1), \dots, M_1(d_J), E(d)). \quad (3.2)$$

The query counterpart $\psi(q; D)$ is computed similarly. This approach directly modifies the final embedding vector based on corpus context before standard similarity search (e.g., dot product).

This differs from unsupervised adaptation methods such as GPL [56] and Boot&Switch [57], which require corpus access or additional tuning. It also contrasts with late-interaction systems such as ColBERT [58], which, while domain-robust, incur higher online computational costs due to their token-level interaction mechanisms at query time, unlike the single-vector representations produced by CDE.

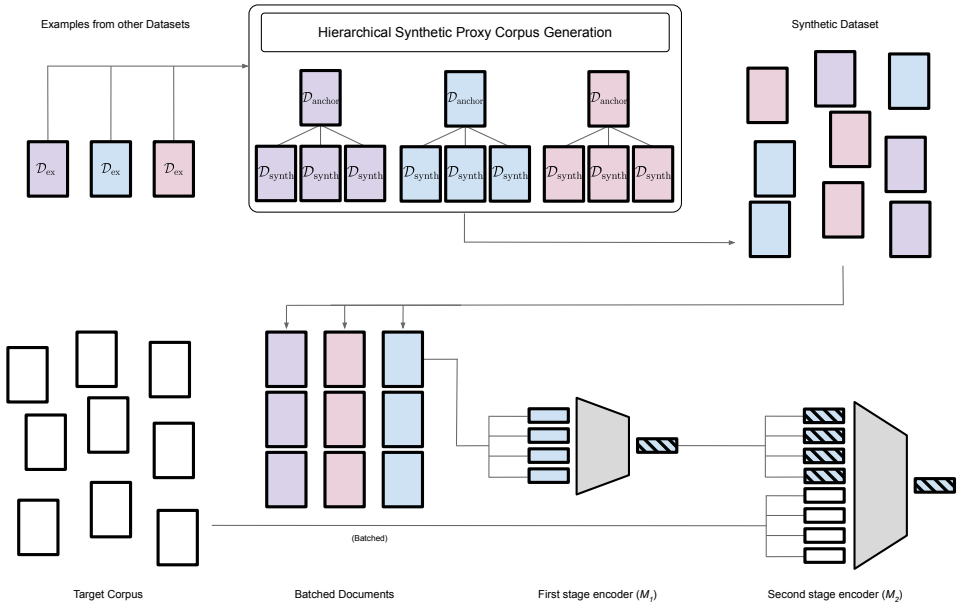


Figure 3.4.: Overview of the ZEST framework for zero-shot contextual adaptation.

Offline Synthesis (Top): From a few example documents (\mathcal{D}_{ex}), domain anchors ($\mathcal{D}_{\text{anchor}}$) are sequentially generated to ensure thematic diversity. These anchors then guide the parallel generation of a synthetic proxy corpus $\mathcal{D}_{\text{synth}}$. **Online Inference (Bottom):** A frozen, pretrained context-aware model (M_1, M_2) embeds new documents or queries (d/q) by conditioning on $\mathcal{D}_{\text{synth}}$ (with its document representations pre-processed via M_1) instead of the inaccessible target corpus \mathcal{D} . This provides domain-adapted embeddings without requiring model retraining or direct corpus access.

By explicitly exposing corpus statistics (such as term frequency or topical patterns via neighbor documents) to the embedding generation process, CDE-style models improve robustness across domains. However, their reliance on accessing the target corpus \mathcal{D} at inference time remains a significant practical hurdle. Our work focuses on overcoming the corpus access requirement.

Problem setting: zero-shot contextual adaptation. We consider deploying a *frozen* context-aware model (M_1, M_2) in a new domain where the full corpus \mathcal{D} is inaccessible. Instead, in lieu of supplying documents from the target corpus, the practitioner can supply only a small exemplar set $\mathcal{D}_{\text{ex}} = \{d_{\text{ex}}^1, \dots, d_{\text{ex}}^k\}$ that typifies the general domain. The challenge is to generate domain-sensitive embeddings $\phi(d; \cdot)$ and $\psi(q; \cdot)$ without accessing \mathcal{D} and without any parameter updates. The remainder of this chapter presents ZEST, a solution to this problem.

3.4.3. METHOD: ZERO-SHOT EMBEDDINGS VIA SYNTHETIC CONTEXT

ZEST enables corpus-aware adaptation without direct corpus access by substituting the real neighbor set required by context-aware models with a compact, LLM-generated proxy corpus $\mathcal{D}_{\text{synth}}$. This involves a one-time offline synthesis phase and an online inference phase using the fixed proxy.

RATIONALE: LEVERAGING SYNTHETIC CONTEXT

Context-aware retrievers rely on neighbor documents to capture domain statistics (Section 3.4.2). Our hypothesis is that an LLM, guided by few exemplars \mathcal{D}_{ex} , can generate a synthetic corpus $\mathcal{D}_{\text{synth}}$ whose statistics sufficiently approximate the target domain’s regularities. If the frozen model’s first-stage encoder M_1 is applied to the members of $\mathcal{D}_{\text{synth}}$ and the resulting vectors are fed into the second stage M_2 together with the query or document to be embedded, then the outputs $\phi(d; \mathcal{D}_{\text{synth}})$ and $\psi(q; \mathcal{D}_{\text{synth}})$ should exhibit much of the desired domain adaptation – achieving effective zero-shot contextualization without \mathcal{D} .

OFFLINE PHASE: FEW-SHOT SYNTHETIC-CONTEXT GENERATION

Input: Domain-specific examples. The offline pipeline begins with a small exemplar set of k documents, $\mathcal{D}_{\text{ex}} = \{d_{\text{ex}}^1, \dots, d_{\text{ex}}^k\}$, selected to typify the target domain (for instance, finance or healthcare) which are sourced separately from \mathcal{D} from domain-similar public corpora. These serve as concrete stylistic and topical anchors for the LLM. While these examples provide crucial domain signals, the specific selection is not expected to be overly sensitive, particularly because the target corpus \mathcal{D} is inaccessible by design, which makes fine-grained optimization of the exemplar set infeasible. As such, the primary goal is to provide the LLM with a general sense of the domain’s characteristics, rather than perfectly matching unknown corpus specifics.

Hierarchical Synthetic Corpus Generation via Domain Anchors. To enhance the representational fidelity and thematic coherence of the synthetic context, we introduce a hierarchical generation approach based on explicit *domain anchors*. We hypothesize that this intermediate anchor step is beneficial because it (i) explicitly encourages topical diversity across the final synthetic corpus, preventing fixation on only a few aspects of the initial exemplars, (ii) mitigates potential mode collapse where the LLM might over-produce content related to a single dominant theme, and (iii) grants finer-grained semantic control when expanding each focused anchor into multiple full documents. Concretely, this procedure unfolds in two steps.

Step 1: Domain Anchor Generation. To establish a diverse set of thematic seeds that broadly represent the target domain, the LLM is prompted to *sequentially* generate A domain anchor documents, $\mathcal{D}_{\text{anchor}} = \{a_1, \dots, a_A\}$, from the exemplar set \mathcal{D}_{ex} . Each anchor a_i is a concise text capturing a distinct topical or stylistic facet observed in \mathcal{D}_{ex} . By generating anchors one after another, the process can be guided to ensure each new anchor explores different characteristics of the exemplars, thereby constructing a varied foundation for the subsequent corpus expansion. Practically, these anchors are

generated by instructing the LLM to produce brief documents that explicitly highlight key concepts, terminology, and typical stylistic attributes of the domain as evidenced in \mathcal{D}_{ex} .

Step 2: Synthetic Corpus Expansion. Next, the complete synthetic corpus $\mathcal{D}_{\text{synth}}$ is created by expanding upon these domain anchors. For each anchor document $a_i \in \mathcal{D}_{\text{anchor}}$, the LLM generates a corresponding subset of synthetic documents. This generation for each anchor can proceed in *parallel*, with the LLM prompted to elaborate on and diversify the theme encapsulated by a_i . This “branching out” from each anchor aims to populate $\mathcal{D}_{\text{synth}}$ with a rich collection of J' novel documents that exhibit broad topical and stylistic coverage pertinent to the target domain. Formally, the final synthetic corpus is the union of these anchor-conditioned subsets:

$$\mathcal{D}_{\text{synth}} = \bigcup_{i=1}^A \{d'_{i,1}, \dots, d'_{i,J'}\}, \quad \text{where } \sum_{i=1}^A J'_i = J'.$$

This hierarchical approach ensures explicit semantic coherence and comprehensive topical coverage in the resulting synthetic corpus, potentially improving the effectiveness of downstream contextual embedding adaptation. Specific implementation details on this generation process and prompting strategies are given in [Section 3.4.4](#) and [Appendix Section D.1](#), respectively. Because $\mathcal{D}_{\text{synth}}$ is reused verbatim during deployment, this synthesis step must only be executed once per domain, making it computationally efficient for practical application.

3.4.4. EXPERIMENTAL SETUP

This section details the experimental protocol designed to evaluate the effectiveness of ZEST in realistic zero-shot domain adaptation scenarios using established retrieval benchmarks.

Datasets and Metrics. We evaluate our approach on the widely used MTEB [59] benchmark, which covers a diverse range of embedding tasks. Unless otherwise noted, we evaluate on the complete MTEB benchmark across all task categories listed in [Table 3.3](#). For each task, we source our \mathcal{D}_{ex} from domain-similar public corpora. For example, for retrieval tasks we sample \mathcal{D}_{ex} from the BEIR [51] benchmark; specifically, from those tasks that are the closest match to the domain of the target corpus (see [Appendix Section D.2](#) for details). Should the same task be present across both datasets, then we choose the next most relevant one instead. For example, for ArguAna [60], which is present in both benchmarks, we choose the most similar task from BEIR instead to sample \mathcal{D}_{ex} from. We randomly sample documents with ≥ 100 tokens to provide sufficient content for the LLM to capture domain characteristics. This simulates a realistic scenario where a user provides a few characteristic examples for domain adaptation. We ensure no leakage by replacing any document that has a 20-token span overlap between \mathcal{D}_{ex} and the corresponding MTEB evaluation datasets. Retrieval quality is evaluated using the standard NDCG@10 metric.

Baselines for Comparison. We compare ZEST against key baselines of similar size to contextualize its performance. We establish context-agnostic performance

using strong, standard biencoder models: `gte-base-en-v1.5` (GTE v1.5) [61] and `bge-base-en-v1.5` (BGE v1.5) [62]. For experiments involving context-aware embeddings, we utilize the publicly available `cde-small-v1` model [53]. This model comprises 137M parameters and was pretrained on a large, diverse corpus. We use its original frozen weights throughout all experiments, ensuring fair comparison and isolating the effect of the context source. For this baseline, context embeddings are computed from $J = 512$ real documents randomly sampled from the target corpus partition, serving as a practical upper bound using real context with a comparable context size. Additionally, we include the Generic Synthetic Context (GSC) baseline, which generates synthetic documents using a generic prompt applied to the same LLM as ZEST, but without its hierarchical approach. This baseline isolates the impact of ZEST’s use of domain anchors, providing a direct comparison to a simpler synthetic context generation method. Finally, to test the sensitivity of our approach to exemplar document selection, we include a random baseline. Here, we randomly sample 512 documents from the first 10k entries of the Colossal Clean Crawled Corpus (C4) dataset as contextual documents. Our baseline selection directly tests ZEST’s core hypothesis: using synthetic context as a drop-in replacement for real context with a frozen architecture. Consequently, methods requiring training-time adaptation (e.g., GPL [56]) or online corpus access (e.g., pseudo-relevance feedback [63, 64]) are considered orthogonal to this specific evaluation.

Synthetic Context Generation. We generate the synthetic context $\mathcal{D}_{\text{synth}}$ for ZEST using GPT-4o via its API, chosen for its strong instruction-following, ability to capture nuanced stylistic and topical patterns from limited examples across diverse domains, and cost-effectiveness. A carefully constructed prompt (see Appendix Section D.1) first provides the $k = 5$ curated domain examples and instructs the LLM to generate synthetic documents hierarchically per Section 3.4.3. This prompt is designed for outputs that are stylistically and topically aligned with these examples — building on domain anchors — and sufficiently diverse to form a rich context. For each of $A = 20$ anchor documents $a_i \in \mathcal{D}_{\text{anchor}}$, the LLM generates an equal fraction (J'/A) of the $J' = 512$ total synthetic documents forming $\mathcal{D}_{\text{synth}}$. This per-anchor generation, designed to elaborate on and diversify a_i ’s theme, proceeds in *parallel*. We use default API sampling parameters for reproducibility.

Additional Implementation Details. For ZEST, the synthetic context embeddings ($\{C_j\}$) were pre-computed from $\mathcal{D}_{\text{synth}}$ using a batch size of 16. During MTEB evaluation runs the models processed task queries and documents with a batch size of 512. Following the methodology of CDE, task-specific prefixes (see Appendix Section D.3) were applied to inputs before being processed. This ensures consistency in how the model receives data for both real-context and synthetic-context scenarios. Experiments were conducted using NVIDIA A100 GPUs.

3.4.5. RESULTS AND DISCUSSION

This section presents the empirical evaluation of ZEST, demonstrating its ability to achieve effective contextual adaptation in zero-shot scenarios. We analyze its performance against established baselines, investigate the impact of key hyperparameters through ablation studies, and discuss the implications of our findings.

MAIN RESULTS: ZERO-SHOT CONTEXTUAL ADAPTATION

The results presented in [Table 3.3](#) compellingly demonstrate the efficacy of ZEST. Across the MTEB benchmark, ZEST using its exemplar-guided synthetic context achieves performance strikingly close to the CDE model that leverages full target corpus access, with an average difference of merely 0.29 NDCG@10 points. This indicates that ZEST comes within 0.45% of the performance attainable with unrestricted access to the real corpus – a significant finding given its zero-shot nature. While the real-context baseline naturally sets a practical upper bound, ZEST closes a large portion of the gap between the no-context baseline and this upper bound. Additionally, the random baseline sets the lower bound, with a significant decrease in performance. This indicates that while our method is not particularly sensitive when choosing documents within the general domain of the target corpus, choosing exemplar documents from an unrelated domain hinders the contextual embeddings, as expected.

Notably, ZEST also outperforms the GSC synthetic baseline by 0.31 NDCG@10 points on average, highlighting the benefit of our domain-anchor-based synthesis approach over simpler synthetic generation from the k exemplars. Furthermore, ZEST establishes substantial gains over strong context-agnostic baselines. It surpasses GTE v1.5 by an average of 2.04 and BGE v1.5 by 2.76 as measured by NDCG@10. These improvements underscore the value of carefully generated synthetic context. Indeed, by recovering 87.6% of the performance gap between GTE v1.5 and the full-access CDE model, ZEST effectively emulates the benefits of real corpus statistics without requiring direct access.

ABLATION STUDIES

Effect of Number of Examples and Anchors. To better understand the behavior and robustness of ZEST, we perform ablation studies on the number of guiding examples $k \in \{1, 2, 5, 10\}$, holding the synthetic context size fixed at $J' = 512$. Results are shown in [Figure 3.5b](#). Performance increases when moving from $k = 1$ to $k = 5$, indicating that providing the LLM with a few diverse examples significantly helps it capture the target domain’s characteristics more accurately by not overfitting to a single example. Using $k = 10$ provides only marginal, if any, additional benefit over $k = 5$ in our experiments, suggesting that our hierarchical generation using $A = 20$ domain anchors (as described in [Section 3.4.3](#)) provides sufficient diversity with only $k = 5$ examples. Furthermore, we also investigate the number of A and find that performance is not sensitive to this parameter, as variations around our default yield negligible differences in overall MTEB scores.

Task Category	GTE v1.5	BGE v1.5	CDE	GSC	Random	ZEST
Classification	77.2	74.7	82.5	81.8	80.4	82.2
Clustering	46.8	45.3	49.3	48.7	46.9	49.1
Pair Classification	85.2	85.7	87.5	87.0	85.6	87.2
Reranking	57.7	58.3	60.0	59.4	57.3	59.7
Retrieval	54.1	52.8	55.2	54.6	52.8	55.0
STS	82.0	81.6	83.3	82.7	81.4	83.0
Summarization	31.2	30.8	32.7	32.1	30.4	32.3
Average	62.03	61.31	64.36	63.76	62.11	<u>64.07</u>

Table 3.3.: Retrieval performance on the MTEB benchmark, shown across its task categories. Baselines include context-agnostic models (GTE v1.5, BGE v1.5), CDE with real context ($J = 512$), ZEST with random documents, and our synthetic GSC baseline. ZEST uses $k = 5$ examples and $J' = 512$ synthetic documents, without accessing the target corpus. Best overall result in **bold**, best result without corpus access underlined.

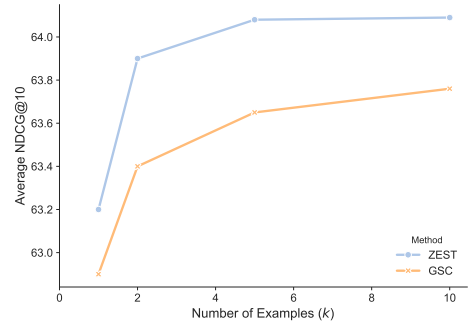
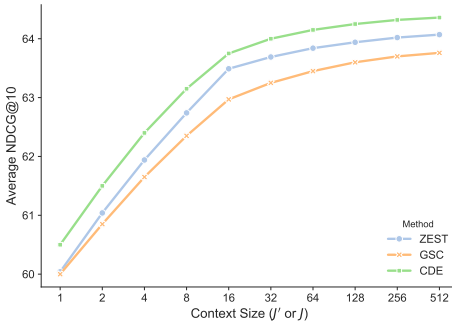
Effect of Synthetic Context Size. We further investigate changing the context size while keeping k constant. Figure 3.5a illustrates how retrieval performance varies with J' . Here, performance generally improves substantially as J' increases from small values (e.g., 2) towards $J' = 512$. While the most significant gains often occur with initial increases in context size (improvement slows after $J' = 16$), the results suggest that a larger synthetic context allows the LLM to generate a richer and more diverse set of documents, providing a more comprehensive contextual signal and leading to performance levels comparable to using real context. This justifies our choice of $J' = 512$ for the main results, although smaller values of J' still offer improvements.

Impact of LLM Choice. To assess the sensitivity of our synthetic context generation approach to the choice of LLM, we compared the performance when using GPT-4o versus the open-source Llama-3.3-70B-Instruct (Llama 70B) model. This comparison, the results of which are shown in Table 3.4, focuses on the two embedding methods directly employing LLM-generated context: our proposed ZEST framework and the synthetic GSC baseline.

As shown, the advanced capabilities of GPT-4o translate to more effective synthetic context for both GSC and ZEST, outperforming Llama 70B by 1.05 points for GSC and 1.11 points for ZEST. However, it is noteworthy that ZEST still demonstrates a clear advantage over GSC even when both utilize Llama 70B, with ZEST achieving a score of 62.96, maintaining a lead of 0.25 points over GSC's 62.71. This suggests that while the quality of the generator LLM is impactful, ZEST's exemplar-guided hierarchical synthesis strategy provides inherent benefits in creating more effective domain-specific context, regardless of the specific LLM employed.

LLM used for Synthesis	GSC	ZEST
GPT-4o	63.76	64.07
Llama-3.3-70B-Instruct	62.71	62.96

Table 3.4.: Influence of the generating LLM on retrieval performance via synthetic context. Average MTEB NDCG@10 scores for GSC and ZEST using $k = 5$ and $J' = 512$.



(a) Performance across benchmark datasets, comparing ZEST to synthetic baseline and CDE at equal context sizes, with constant $k = 5$ examples.

(b) Performance across benchmark datasets when altering the number of few-shot examples (k), with constant $J' = 512$ synthetic documents. We compare it to the performance of GSC, which does not use domain anchors.

DISCUSSION

The strong performance of ZEST, achieving comparability with CDE using real context ($J = 512$), underscores the capability of modern LLMs to act as effective simulators of domain-specific corpus characteristics based on minimal examples. Effectively, we treat the LLM as a giant database that we retrieve our synthetic documents from. The synthetic context $\mathcal{D}_{\text{synth}}$ captures not just topical relevance but also implicit statistical patterns (such as term co-occurrence and relative frequency) and stylistic elements that the pretrained CDE model leverages for adaptation. While not a perfect replacement for the real corpus, the synthetic context provides a remarkably effective proxy, enabling high-performance context-aware retrieval where it was previously infeasible.

Computational Considerations. ZEST introduces an offline cost for generating $\mathcal{D}_{\text{synth}}$. Importantly, this represents a one-time, offline process per target domain. Specifically, generating $\mathcal{D}_{\text{synth}}$ using GPT-4o via its API results in synthetic documents with an average length of 255.6 output tokens, compared to 225.1 tokens for the benchmark documents. The online inference cost for ZEST involves only the standard forward pass using the precomputed synthetic context embeddings $\{C_j\}$, similar to that of CDE using real context. In our experiments, the time for the hierarchical generation

process is negligible compared to inference. This compares favorably to the resources required for alternative adaptation strategies, which typically demand significant GPU hours for finetuning and/or assume the feasibility of corpus access for data acquisition and processing.

Qualitative Insights. Examining samples from $\mathcal{D}_{\text{synth}}$ — see Appendix Section D.4 — reveals the ability of LLMs to generate relevant and stylistically consistent documents for a specific domain. However, occasional generic or less relevant documents do occur, potentially limiting performance.

3.4.6. RELATED WORK

Neural Dense Retrieval. Neural retrievers learn to map documents and queries into a shared embedding space, optimizing a contrastive loss for efficient retrieval [48–50, 65, 66]. However, their *context-agnostic* design makes them vulnerable to domain shifts when the test corpus differs from pretraining data [51]. Our work builds on this foundation but injects domain signals at inference without requiring access to the full corpus.

Context-Aware Embeddings. To mitigate corpus shift, recent methods augment models with contextual information drawn from *neighbor* documents. Contextual Document Embeddings [53] and late-interaction models like ColBERT [58, 67] refine each embedding by reading a subset of the target corpus at inference. While effective, these approaches assume unrestricted access to the entire document collection—an assumption that fails in privacy-sensitive or large-scale environments. Semi-parametric language models such as k NN-LM [68] or non-parametric transformers [69] also incorporate external information, yet focus on embeddings for text generation rather than retrieval embeddings. In contrast, ZEST replaces the real neighbor set with a compact, LLM-generated proxy, enabling analogous context-aware gains under a zero-access constraint.

Test-time and unsupervised retrieval adaptation. A rich line of work explores adapting retrievers to new domains without labeled data. Unsupervised corpus-aware pretraining methods (e.g., GPL [56], LaPraDoR [70], SimLM [71]) fine-tune on the target documents themselves, while few-shot or parameter-efficient schemes (e.g., TSDAE [72], adapters [73], prompt tuning [74]) require some labeled examples. Test-time techniques such as pseudo-relevance feedback [63, 75] or Boot&Switch [57] adjust queries or model parameters on the fly but likewise need access or online optimization. Unlike these methods, ZEST performs *training-free* adaptation: it freezes all model weights and synthesizes a proxy context once offline from only a handful of example documents.

Synthetic data generation for retrieval. LLMs are now routinely used in retrieval-related tasks to fabricate documents via few-shot prompting [76], for

supervision—such as with Promptagator [74] and CRAFT [77]—or to curate hard negatives [78]. Prior work typically deploys the synthetic text for *training* [54]. In contrast, ZEST exploits LLMs at *inference time*: the generated mini-corpus acts as a stand-in reference that unlocks corpus-aware embeddings without accessing the target corpus.

3.4.7. CONCLUSION

We introduced ZEST, a novel method that enables context-aware document retrieval adaptation without requiring access to the target corpus. By employing LLMs guided by few-shot examples to synthesize a representative context corpus offline, ZEST allows a pretrained contextual embedding model to adapt effectively during online inference. Our findings show ZEST substantially improves zero-shot retrieval over context-agnostic methods, nearing the performance of models with full corpus access. ZEST addresses a critical practical limitation of context-aware models, offering a viable path toward more adaptable and effective document embeddings in scenarios constrained by corpus access, privacy, or scale. This data-centric adaptation strategy opens new possibilities for deploying sophisticated retrieval models in challenging real-world environments.

3.4.8. LIMITATIONS

Despite its effectiveness, ZEST has limitations. Its performance is inherently linked to the quality of the LLM-generated synthetic context, as it remains a proxy for the true corpus statistics. The reliance on LLMs introduces dependencies on external APIs and associated costs for the offline generation step. Furthermore, potential biases present in the LLM or \mathcal{D}_{ex} could be amplified in the synthetic context, requiring careful consideration in sensitive applications. Finally, the selection of k examples introduces variability; automating or guiding this selection could improve robustness. We leave this to future work.

Future work could further explore: (1) finetuning open-source LLMs specifically for high-fidelity context generation to reduce dependencies and (2) developing techniques for automated quality assessment of the generated $\mathcal{D}_{\text{synth}}$ and potentially filtering or refining it.

3.5. FROM SYNTHETIC CONTEXT TO MULTIMODAL CONTEXT

Part II (Section 3.4) demonstrated that effective, domain-relevant context can be supplied *without* direct access to the target corpus by synthesizing a compact proxy that can be used to condition a frozen context-aware encoder at inference time. Using a handful of exemplar documents to guide hierarchical anchor generation, we construct a synthetic corpus whose statistics approximate the target domain, yielding contextual embeddings that approach the performance of embedding models with full corpus access while respecting privacy and deployment constraints.

However, effective contextual knowledge is not purely textual, and is needed in domains beyond embeddings. Many linguistic ambiguities are resolved by

extra-linguistic cues such as layout, scene composition, or visual co-reference. In the spirit of targeted knowledge injection, we next examine *multimodal* context as an inference-time signal that can be injected without parameter updates: page- and panel-level visual information that disambiguates utterances, speaker intent, and local story state.

Part III (Section 3.6) of this chapter operationalizes this idea for multimodal translation, a setting where visual evidence is often essential for disambiguation. We compare text-only baselines to image-informed methods and long-context variants. Further, we evaluate when visual context helps while excessive textual context hurts. This multimodal case study complements previous parts by showing when and how visual signals act as effective targeted contextual knowledge at inference time.

3.6. PART III: CONTEXTUAL ADAPTATION IN MULTIMODAL SETTINGS

3.6.1. INTRODUCTION

A Japanese style of comics – referred to as *manga* – has been popular with audiences outside of Japan for decades. Handcrafting high quality translations, key to distributing manga world wide, is a difficult undertaking that takes significant time and effort. As such, most manga never leave the domestic Japanese market. Additionally, readers who do not speak a language into which manga is typically translated have limited or no access at all due to the high initial costs of translations.

The use of Neural Machine Translation (NMT) promises seamless translations from one language to another without involving a human translator [79, 80]. Still, successful applications of NMT to manga – or comics in general – remain limited, and automatic methods remain far from being able to reliably translate manga at a level comparable to humans [81]. This is in part due to the unique requirements of manga as a translation problem, which involves literary translation, handling split sentences across multiple speech bubbles, and especially resolving ambiguities using visual information. For example, in Figure 3.6, achieving an accurate translation requires integrating both textual and visual context from the current and preceding scenes.

Research into manga-specific NMT methods is limited, focusing mainly on Japanese-English translation due to a lack of parallel corpora for other language pairs [81, 82]. Of these, only one method has attempted incorporating visual context into a model via a limited number of descriptive tags, yielding inconclusive results [81]. Previously proposed models were trained on a private JA-EN data set, which is not shareable due to copyright [81, 82]. Although there exist several general purpose manga data sets, such as Manga109 [83, 84], so far only one manga translation data set has been published for research purposes: OpenMantra [81]. However, its limited size makes it effectively an evaluation data set only, making it challenging to train models.

Large language models (LLMs) have shown to be capable translators across languages [85, 86]. The release of multimodal LLMs – those that make use of visual information in addition to text – makes translation of media with visual nuance a possibility [87]. This potentially bypasses the need for large parallel manga corpora for

3.6.2. RELATED WORK

AUTOMATIC METHODS FOR MANGA

Up to this point, the development of automatic manga translation methods that incorporate multimodal context has been limited. Hinami *et al.* [81] first proposed an NMT system for manga that makes use of contextual information obtained from images to inform the translation. Their method is restricted to a single frame of context and the visual information obtained from the images is limited to 512 predefined labels. Further work has explored the use of an additional frame or manga metadata to improve translation quality [82], however, without visual context. Instead, we propose taking additional textual content of up to the entire manga volume into account to improve translations, as well as the full manga image without predefined labels. Outside of translations, Chen *et al.* [89] propose a sentiment analysis method on manga text and Guo *et al.* [90] propose an approach that makes use of both visual and textual modalities to complete empty speech bubbles in existing manga. There has been sparse early-stage research into automatic methods for similar media, such as graphical novels [91] and American comics [92].

LARGE LANGUAGE MODEL TRANSLATIONS

Translation using LLMs is appealing due to their ability to generate high-quality translations for various language pairs without the need for training on extensive parallel corpora or fine-tuning [93]. LLMs have previously been shown to be capable translators [94], as well as evaluators of translation quality [95]. Further, paragraph translations performed by LLMs have been shown to be effective when using basic English prompts at the sentence level [96]. We propose multiple translation approaches and evaluate the quality of our LLM manga translations compared to finetuned transformer models and explore a low-resource language pair, JA-PL, as well as contribute a data set for evaluation.

MULTIMODAL MACHINE TRANSLATIONS

Translating text embedded in images has been extensively explored in research [97, 98]. Multimodal machine translation (MMT) has so far mainly been applied to translating image captions, outperforming the text-only baseline by leveraging additional visual information [99]. MMT typically uses a single image with its corresponding text description as input [100]. We investigate to what extent an increased visual context length is effective. A further challenge comes from the discrepancy between the natural images and their description used to train vision encoders used for MMT and manga images, as manga has a unique hand-drawn art style with relevant text drawn into the image [101]. Additionally, little attention has been paid to low-resource languages, with the vast majority of MMT research focused on the most popular language translation pairs [102, 103]. More recent LLMs have additional multimodal capabilities [104, 105], enabling them to perform MMT, though this has not been explored for the manga use case.

3.6.3. METHODOLOGY

In this section, we first outline manga terminology, then present the problem, and finally introduce our LLM-based translation approaches that take advantage of multimodality and a longer context.

TERMINOLOGY & PROBLEM FORMULATION

Page-to-page manga translation involves three steps: (1) *page processing* to identify elements on the page, detect text, and estimate reading order; (2) *translating* the text into the target language; and (3) *typesetting* the translated text onto the page in stylized font. The focus of this chapter is on (2), but we will discuss (1) and (3) in [Section E.1](#).

A manga page consists of multiple story panels, referred to simply as *panels*, as shown in [Figure 3.7](#). Panels often contain text, which can be enclosed in a *speech bubble* for text spoken or thought by characters, or free-flowing for background noise or sound effects. The term *line* will always refer to the content of one speech bubble, narrative box, or cluster of free-flowing text.

For multimodal manga translation, we make use of the *image* of the drawings on a single manga page, such as [Figure 3.7](#), which contains lines of text. We make the assumption that the text contained on the page has already been recognized and is available. Our goal is to obtain the correctly translated text for each line.

TRANSLATION APPROACHES

We use a variety of translation approaches – summarized in [Table 3.5](#) – to assess the impact of multimodality, translation unit size, and context length and find the most performant configuration. To establish a baseline, our first approach is a simple line-by-line approach (LBL). This means that the model receives one single line to translate at a time, without any additional context about the manga it is translating. Previous research has shown that LLMs perform better on translation tasks when given the entirety of a text compared to snippets, as they are able to incorporate the broader context more effectively [106]. As such, the second approach we evaluate is page-by-page (PBP), where the model is given all lines from a given page in the correct reading order and outputs all the corresponding translations.

MULTIMODAL TRANSLATION

Ideally, we would provide the LLM with just the image and it would recognize the text, obtain the visual context, and perform the translation. However, current models are not capable of this. Instead, we investigate approaches where the model is given the lines on a page to translate, along with the corresponding image as additional visual context, enabling multimodal translation.

The first approach we investigate is the multimodal equivalent to LBL, referred to as LBL-VIS, where the model receives lines and the corresponding page image as visual context. The second approach utilizing visual context is PBP-VIS, which involves giving the model the entire text from one manga page and the page itself as an image.

The final approach aims to directly address the issues that multimodal LLMs have with reading non-Latin scripts. The setup is the same as PBP-VIS, but the image of the



Figure 3.7.: A manga page: panel borders (green), example lines in speech bubbles (purple), free flowing text (orange) and sound effects (red). Courtesy of Akamatsu Ken, ©Kodansha

manga page is modified to avoid the LLM performing any optical character recognition (OCR). The contents of the speech bubbles in the image are removed and replaced with numbers indicating the reading order and corresponding to the list of Japanese lines the model is given to translate (see Figure 3.8). We call this approach PBP-VIS-NUM and it enables the model to locate the speech bubble more easily and relate its content to the exact panel in which it was placed, without performing any OCR on the text itself.

LONG-CONTEXT TRANSLATION

Intuitively we want to make use of context lengths exceeding single lines or pages to adequately capture evolving story lines and character development and accurately translate entire stories in an internally consistent way. The remaining approaches we present are designed to address this.



Figure 3.8.: Fragment of a page annotated for the PBP-VIS-NUM method ©Mitsuki Kuchitaka.

The first of these multi-page approaches provides the model with the previous and next page as additional context to give more information to the LLM. We refer to this as VBP-VIS-3P, as it translates the volume sequentially one page at a time (VBP), using visual context (VIS), and using three pages' worth of context (3P). Going a step further,

Approach	Translation Unit	Textual Context	Visual Context
LBL	line	line	×
PBP	page	page	×
LBL-VIS	line	line	page
PBP-VIS	page	page	page
PBP-VIS-NUM	page	page	num. page
VBP-VIS-COD	page	page + sum.	page
VBP-VIS-3P	page	3 pages	3 pages
VBP-VIS-ALL	page	vol. + trans.	volume
VBV-VIS	volume	volume	volume

Table 3.5.: Overview of the studied translation approaches and their respective translation unit size and context. Abbreviations: “num.” is number of, “sum.” is summary, “vol.” is volume, and “trans.” is translation so far.

we explore VBP-VIS-ALL, where the model is provided with the images and lines from an entire manga volume, as well as all the translations so far, and queried to translate the next untranslated page. This process is repeated sequentially for every page in the volume.

As the input and output length increases, the limited context windows of LLMs are quickly exhausted and performance is diminished [107]. To overcome this, we introduce the scalable VBP-VIS-COD approach, where we extend chain of density summarization (COD) [108] to keep a rolling, fixed-length summary of the story’s developments as our context. Besides the image and its corresponding text, the model is given a summary of the story thus far in the target language as additional context. For a detailed overview of this approach, see Section E.2. The last evaluated approach, VBV-VIS translates an entire manga volume in a single call. Similar to VBP-VIS-ALL, we provide the LLM with the texts and images from an entire manga volume, but then instruct it to respond with the translations for the entire volume in a single query.

3.6.4. DATASETS

However, from a research perspective, the main issue with manga is that, due to its commercial nature, most manga is protected by Japanese and local copyright laws [109]. Previous manga-related works have addressed this problem in different ways. Some researchers resort to using private data sets [81, 82, 110], while others use the very few publicly available copyright-free manga, accepting the trade-off of unlabeled data [111].

To date, there has been only one manga translation data set made public for research purposes – the OpenMantra data set by Hinami *et al.* [81]. It consists of five independent Japanese-language manga volumes, totaling 214 pages (1593 speech bubbles). Details of this data set are shown in Table 3.6. Each volume in this data set has annotations for the locations of panels and text boxes on the page, as well as the contents of the text boxes, and the reading order, with professional translations into English and Chinese. We use this data set to evaluate JA-EN translations, splitting it into two parts: validation set (*Balloon Dream* and *Tojime no Siora*) and test set (*Boureisougi*, *Rasetugari*, and

Tencho Isoro).

NEW JAPANESE-POLISH MANGA DATA SET

In addition to JA-EN, we explore JA-PL translation; as English and Polish belong to different language families, diverge significantly in terms of morphology, and have different grammatical structures. We provide professional JA-PL translations of the slice-of-life manga *Love Hina* to create a data set for research purposes. We make volumes 1 and 14 available and our annotation process closely follows the existing annotations of the Japanese text. The newly contributed data set contains 400 pages and 3705 individual lines (*i.e.* speech bubbles, sound effects, etc.) split across the two volumes and is distributed as a set of images, corresponding to one image per page, and the corresponding metadata containing original and translated text, as well as their coordinates on the page. This exceeds the previously largest manga translation data set, OpenMantra [81], in size. We propose a 50:50 validation:test split for this data set, using the first volume (200 pages and 1810 lines) as the test set and the last volume (200 pages and 1895 lines) as the validation set. This decision is motivated primarily by the fact that the first volume establishes the story, providing a fairer benchmark for the long-context methods, as opposed to the last volume, which depends on unavailable context, being the 14th installment in the series.

Our annotation process closely follows the existing annotations of the Japanese text. The original lines were matched with the corresponding translated lines primarily based on location, and if impossible, based on content. However, in edge cases the Polish edition left very small text untranslated as a stylistic choice. The reading order was first estimated using the tool provided by Sachdeva and Zisserman [112] and then corrected by hand based on the actual speech bubbles. During the annotation process, we noticed several characteristics of this title and the unique challenges it presents for translation. Some characters in *Love Hina* speak the Kansai dialect of Japanese. According to the literature, there is no consensus on how to translate this dialect into Polish, with different translators choosing different Polish dialects [113]. Another challenge is that one of the secondary characters speaks in a manner resembling samurai speech – a common trope in manga [114]. Again, there is no consensus on how to convey this in Polish. As such, users of the data set should be aware that some “incorrect” translations may be just as valid in these cases.

3.6.5. BASELINES

We employ four baseline methods for JA-EN translations. The first two baselines, *Scene-NMT* and *Scene-NMT-VIS*, come from the original automatic manga translation work by Hinami *et al.* [81]. The first method uses a transformer-based model to translate the contents of entire panels at once without multimodal context, while the second method includes visual features as well. The third baseline method we use – and current state-of-the-art for automatic manga translation – is *Scene-aware-NMT* [82], which translates manga panel by panel as well, using a transformer-based model but uses the text from the previous panel as additional context. The translation outputs for all these previously listed methods were kindly provided to us by the authors of

Manga Title	Genre	# Pages	# Lines
<i>Balloon Dream</i>	Romance	38	314
<i>Boureisougi</i>	Mystery	36	274
<i>Rasetugari</i>	Fantasy	54	359
<i>Tencho Isoro</i>	Slice-of-life	40	311
<i>Tojime no Siora</i>	Battle	46	334

Table 3.6.: Overview of the OpenMantra dataset [81].

the respective works. This allowed us to use our own data splits and ensure that all methods were evaluated equally and comparably.

The last baseline method we use is Google Translate (GT) due to its support for a wide range of languages and availability. GT is our only baseline for JA-PL translations. All GT translations were carried out in April and May 2024, using the Google Translate API with the corresponding Python library.¹

3.6.6. EVALUATION PROTOCOL

AUTOMATIC EVALUATION

For evaluation, we use a range of automated metrics applied at the sentence level. We use a lexical n-gram matching heuristic metric in ChrF [115]. Although the reliability of this type of metric has been questioned over the years [116], they remain among the most widely used in machine translation [117, 118]. ChrF provides scores on a scale from 0 to 100, where higher scores indicate higher quality translations.

The first non-lexical machine translation evaluation metric we use is BERTScore [34], considered a good representative of the embedding-based metrics category [119]. Although not perfect, it has been shown to detect important content words and is well suited to score candidates from different languages [120]. Next, we report scores with a learned metric, BLEURT [121], specifically the top-performing BLEURT-20 model [122]. The last metric we report is the learned metric xCOMET [123], specifically xCOMET-XXL. xCOMET is an open-source learned metric that performs error span detection in addition to standard sentence-level evaluation. It is currently considered the best-performing publicly available metric [124]. Among all the metrics we employ, it is the only one that calculates its score based not only on the references and hypotheses but also on the source text. BERTScore, BLEURT, and xCOMET return a score on a scale of 0 to 1, with results closer to 1 being preferable.

HUMAN EVALUATION

In addition to our extensive automatic evaluation, we perform a human evaluation with a professional JA-EN manga translator using the Multidimensional Quality Metrics (MQM) translation evaluation framework [125, 126]. We use MQM with a manga-specific list of issue types that cover different types of errors, such as accuracy, fluency, and style. A

¹<https://pypi.org/project/googletrans/>

complete overview of our MQM process is shown in Appendix [Section E.3](#). Each error type is assigned a severity level, ranging from minor to critical, depending on the impact of the issue on overall quality. MQM provides a scoring system that allows for the calculation of overall quality scores based on the number of identified issues and their severity levels. These scores have an upper bound of 1 and no lower bound, with a higher score being preferable. We choose the *Tencho Isoro* manga for our MQM evaluation. We compare the official commercial translation of the manga, the GT baseline, and our best performing approach (PBP-VIS) to evaluate how a professional human translator would judge each.

PROMPTING

We follow the approach of previous works [[86](#), [87](#), [106](#)] and investigate the out-of-the-box translation performance of GPT-4 Turbo [[127](#)]. The specific version we use is `gpt-4-turbo-2024-04-09` at default hyperparameters with a temperature $T = 0.5$, accessed through the OpenAI Python library.² For all multimodal translations, we append the relevant image(s) of the page(s) as a *jpeg* file to the LLM query via its respective API. We run each configuration once due to the high costs involved in sending entire manga volumes to commercial multimodal LLMs. The complete prompts we use for every translation are shown in Appendix [Section E.4](#). Each approach described in [Section 3.6.3](#) is evaluated one-shot, i.e., with one given example in the prompt. We did not find a measurable difference between one-shot and five-shot prompting when evaluating on the validation data. The model is always prompted in English – regardless of the target language – as this yields the best results for LLM translations [[96](#)]. Based on experiments on the validation data, we ask the model to explain how the image influences the translation, ensuring that the visual context is taken into account.

MANGA TRANSLATION EVALUATION SUITE

We release our evaluation suite to advance research in automatic manga translation. It enables benchmarking of various LLMs by adjusting textual context, visual context, and translation unit size. The suite integrates all methods from [Section 3.6.3](#) for comprehensive evaluation and facilitates automatic assessment using the four previously outlined metrics. With plug-and-play functionality, researchers can easily utilize existing data sets, including OpenMantra and ours, while introducing new prompts and exploring alternative LLMs.

3.6.7. RESULTS

JA-EN TRANSLATION

We present our findings in [Table 3.7](#). Among the methods proposed in previous studies, *Scene-aware-NMT* demonstrates competitive performance, especially on BERTScore, surpassing other previous manga-focused translation methods, consistent with their reported findings. However, our proposed methods show improvements across multiple metrics. Our basic approach, LBL, performs slightly worse than GT in most

²<https://github.com/openai/openai-python>

Method	JA-EN				JA-PL			
	ChrF	BRTS	BLRT	xCMT	ChrF	BRTS	BLRT	xCMT
GT	34.2	0.895	0.525	0.729	22.3	0.826	0.446	0.457
Scene-NMT	34.2	0.897	0.512	0.651	-	-	-	-
Scene-NMT-VIS	34.5	0.895	0.507	0.664	-	-	-	-
Scene-aware-NMT	36.1	0.903	0.534	0.670	-	-	-	-
LBL	32.7	0.883	0.523	0.716	24.2	0.844	0.495	0.531
PBP	36.0	0.898	0.565	0.758	25.6	0.852	0.538	0.565
LBL-VIS	35.6	0.900	0.551	0.746	24.9	0.845	0.515	0.543
PBP-VIS	36.6	0.903	0.581	0.776	25.6	0.852	0.539	0.567
PBP-VIS-NUM	36.8	0.900	0.582	0.776	25.7	0.851	0.532	0.566
VBP-VIS-COD	35.9	0.900	0.566	0.769	25.1	0.846	0.523	0.550
VBP-VIS-3P	35.9	0.897	0.565	0.754	25.6	0.843	0.530	0.559
VBP-VIS-ALL	35.7	0.893	0.556	0.760	24.9	0.840	0.521	0.561
VBV-VIS	34.9	0.884	0.539	0.733	24.5	0.833	0.510	0.534

Table 3.7.: Performance metrics for all approaches for JA-EN and JA-PL translation. Best scores for each translation direction are in **bold**. BRTS refers to BERTScore, BLRT to BLEURT, and xCMT to xCOMET as automatic evaluation metric.

aspects. The PBP method shows substantial improvement over LBL, outperforming all baselines on BLEURT (0.565) and xCOMET (0.758), confirming the potential of LLMs as manga translators, even without visual context.

Visual Context The addition of visual context significantly improves scores across all metrics for both LBL and PBP methods. PBP-VIS and PBP-VIS-NUM achieve the best scores across most metrics, with PBP-VIS-NUM slightly outperforming on ChrF (36.8) and BLEURT (0.582). These results confirm that additional visual context significantly improves LLM translation quality, representing a novel approach in automatic manga translation. Additionally, we perform an ablation study to clarify the role of key visual features, the results of which are discussed in [Section E.5](#).

The results of our human evaluation are presented in [Table 3.8](#). PBP-VIS, clearly outperforms the GT baseline in overall score. However, according to the MQM evaluation conducted by a single professional translator, PBP-VIS is more prone to errors than the official human manga translation. While our method has fewer “minor” and “major” errors compared to the official translation, it exhibits a significantly higher number of “critical” errors. These findings indicate that although our method establishes the current state of the art for automatic manga translation, human translation remains superior in quality. Although these metrics provide context for assessing our method’s efficacy, translation quality is inherently subjective and challenging to measure. While our best translation scores lower than the official translation, we find it enjoyable to read and coherent – a standard the GT translation does not meet.

Context Length Interestingly, providing context beyond the page level does not enhance

JA-EN	Minor	Major	Critical	Score
Official	14	50	107	-1.31
GT	5	20	272	-4.25
PBP-VIS	8	18	160	-1.98

Table 3.8.: Human evaluation MQM results for JA-EN. Errors are listed number per category (lower preferable) with an overall score (higher preferable).

translation quality. VBP-VIS-COD, using only a short summary of previous events, performs better than other long-context methods across most metrics. Conversely, VBV-VIS, which translates the entire volume in one query, shows the lowest performance among our visual context methods. These findings suggest an inverse relationship between translation quality and input length beyond a single page for multimodal LLM translation. This counter intuitive result highlights the importance of optimizing input size for LLM-based translations.

JA-PL TRANSLATION

For the JA-PL data, we do not report the results of methods proposed by other authors, as these are not trained on Polish data and therefore perform poorly. For JA-PL translations, we observe that across methods, scores are generally lower compared to JA-EN translations. However, all our approaches significantly outperform the GT baseline. Further, we note that our top performing methods, PBP-VIS and PBP-VIS-NUM, perform similarly on JA-PL to JA-EN.

Visual Context and Context Length Again visual context improves performance, though to a much lesser extent than for JA-EN translation. For PBP, the impact of visual context is minimal, with PBP-VIS and PBP-VIS-NUM performing similarly to PBP. Long-context approaches show mixed results, again performing worse than PBP-VIS and PBP-VIS-NUM.

IMPLICATIONS AND BROADER IMPACT

PBP-VIS and PBP-VIS-NUM consistently achieve the best results for both JA-EN and JA-PL translations. The effectiveness of our methods across translations suggests broad applicability to different language pairs. The cross-lingual success of our methods indicates that the benefits of incorporating visual context in manga translation are language-independent. Moreover, our PBP-VIS and PBP-VIS-NUM methods achieve the highest scores across all metrics, setting the state of the art for automatic manga translation.

Notably, we observe that translation quality does not necessarily improve with longer context, challenging common assumptions in machine translation. This finding aligns with previous research, which indicates that the quality of output from LLMs tends to diminish as the length of the input increases [128]. This is contrary to the results we observe when additional visual context is taken into account. To optimize performance when using LLMs for multimodal translations, it is advisable to prioritize smaller input

sizes of a single page. Translation quality tends to deteriorate more significantly as the LLM processes longer text, even if it contains more information relevant to the story.

3.6.8. LIMITATIONS

The first limitation of this study is the amount of data used for testing. While we make meaningful contributions to addressing this issue, there is still a severe lack of evaluation data, making it difficult to determine how consistent our findings would be across different authors and genres. Additionally, we only investigate one language other than English, constrained by our ability to manually inspect outputs and analyze model mistakes in other languages.

Related to this is the fact that some manga series span multiple volumes. Translations of later volumes in a series would undoubtedly benefit from including earlier volumes in the available context. Due to the lack of suitable data, we limit ourselves to translations of single volumes, leaving multi-volume narratives to future work.

Finally, there are obvious limitations when using a commercial, closed-source LLM as we do in this chapter, such as potential data leakage issues and the unlikely scenario that some of the manga used might have been part of the training data. Still, the availability and quality of open-source multimodal multilingual LLMs is very limited at this time, and as such we leave a study using alternatives to future work.

3.6.9. CONCLUSION

Our investigation of multimodal LLMs for automatic manga translation reveals significant advancements in this emerging field. We evaluate various LLM-based translation approaches, considering text-only, image-informed, and volume-level contexts. Leveraging the vision component of multimodal LLMs, we enhance translation quality by incorporating visual elements to resolve ambiguities. However, we find that additional textual context does not consistently improve performance. Our methodology achieves state-of-the-art results for JA-EN translations and sets a new standard for JA-PL translations. We also introduce the first parallel JA-PL manga translation data set and an open-source benchmarking suite for LLMs.

These results support the knowledge injection view advanced in [Chapter 3](#): visual signals act as targeted contextual knowledge that can be injected at inference to disambiguate language without parameter updates (cf. [Chapter 3](#)). They also motivate downstream oversight: even with improved context, residual uncertainty and value-laden trade-offs remain, linking to [Chapter 4](#) on value-sensitive human collaboration.

3.7. CONCLUSION AND BRIDGE TO CHAPTER 4

Targeted knowledge injection improves robustness by supplying the right information at the right interface. [Section 3.2](#) showed that agent-in-the-loop (AIL) synthesis can improve language model robustness by mitigating blind spots (unknown unknowns) by translating diagnosed failure patterns into small, hypothesis-guided additions to the training data that reduce high-confidence errors while preserving overall accuracy. Further, [Section 3.4](#) demonstrated that when direct corpus access is infeasible, we

can recover most of the performance gains of context-aware embeddings by replacing real documents with a compact, hierarchically synthesized proxy corpus generated from a handful of exemplars, enabling zero-shot domain adaptation without parameter updates. Finally, [Section 3.6](#) motivated contextual knowledge beyond text: in multimodal translation, visual cues function as targeted context that improves disambiguation and clarifies when additional evidence is necessary.

Together, these results support a pragmatic view of knowledge injection to improve model performance and robustness: reactive AIL can locally adjust decision boundaries around observed blind spots, while proactive (and, where applicable, multimodal context) provide needed information at inference time. Across both mechanisms we observed diminishing returns and risks of overfitting to synthetic artifacts. We prioritize small, testable interventions with explicit guardrails over indiscriminate augmentation, reducing the risk of negative transfer. Yet targeted knowledge injection alone — while beneficial to improve model performance and robustness — cannot by itself determine when model outputs should be trusted. Residual uncertainty, distribution shift, and value-laden scenarios introduce trade-offs that accuracy or confidence do not capture. The following chapter develops value-sensitive AI alignment, formalizing how to elicit stakeholder priorities and translate them into principled escalation policies that decide when to defer to human experts instead of blindly trusting the model.

REFERENCES

- [1] P. Lippmann, M. T. J. Spaan and J. Yang. 'Illuminating Blind Spots of Language Models with Targeted Agent-in-the-Loop Synthetic Data'. In: *Proceedings of the ORIGen Workshop at the Conference on Language Modeling (COLM)*. Workshop on Optimal Reliance and Accountability in Interactions with Generative LMs. 2025. url: https://origen-workshop.github.io/assets/pdf/2025/papers/1_Illuminating_Blind_Spots_of_.pdf.
- [2] P. Lippmann and J. Yang. 'Zero-Shot Contextual Embeddings via Offline Synthetic Corpus Generation'. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose and V. Peng. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 2089–2104. isbn: 979-8-89176-335-7. doi: [10.18653/v1/2025.findings-emnlp.111](https://doi.org/10.18653/v1/2025.findings-emnlp.111). url: <https://aclanthology.org/2025.findings-emnlp.111/>.
- [3] P. Lippmann, K. Skublicki, J. Tanner, S. Ishiwatari and J. Yang. 'Context-Informed Machine Translation of Manga using Multimodal Large Language Models'. In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio and S. Schockaert. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Jan. 2025, pp. 3444–3464. url: <https://aclanthology.org/2025.coling-main.232/>.
- [4] J. Attenberg, P. Ipeirotis and F. Provost. 'Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns"'. In: *Journal of Data and Information Quality* 6.1 (4th Mar. 2015), pp. 1–17. issn: 1936-1955. doi: [10.1145/2700832](https://doi.org/10.1145/2700832). url: <https://doi.org/10.1145/2700832>.
- [5] G. Bansal and D. Weld. 'A Coverage-Based Utility Model for Identifying Unknown Unknowns'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (25th Apr. 2018). issn: 2374-3468. doi: [10.1609/aaai.v32i1.11493](https://doi.org/10.1609/aaai.v32i1.11493). url: <https://ojs.aaai.org/index.php/AAAI/article/view/11493>.
- [6] C. Vandenhof. 'A Hybrid Approach to Identifying Unknown Unknowns of Predictive Models'. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (28th Oct. 2019), pp. 180–187. doi: [10.1609/hcomp.v7i1.5274](https://doi.org/10.1609/hcomp.v7i1.5274). url: <https://ojs.aaai.org/index.php/HCOMP/article/view/5274>.
- [7] A. Liu, S. Guerra, I. Fung, G. Matute, E. Kamar and W. Lasecki. 'Towards Hybrid Human-AI Workflows for Unknown Unknown Detection'. In: *Proceedings of The Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, 20th Apr. 2020, pp. 2432–2442. isbn: 9781450370233. doi: [10.1145/3366423.3380306](https://doi.org/10.1145/3366423.3380306). url: <https://doi.org/10.1145/3366423.3380306>.

- [8] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami. 'The Limitations of Deep Learning in Adversarial Settings'. In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. Saarbrücken, Germany: IEEE, Mar. 2016, pp. 372–387. doi: [10.1109/EuroSP.2016.36](https://doi.org/10.1109/EuroSP.2016.36). url: <https://doi.org/10.1109/EuroSP.2016.36>.
- [9] W. Wang, R. Wang, L. Wang, Z. Wang and A. Ye. *Towards a Robust Deep Neural Network in Texts: A Survey*. 12th Feb. 2019. doi: [10.48550/arXiv.1902.07285](https://arxiv.org/abs/1902.07285). arXiv: [1902.07285](https://arxiv.org/abs/1902.07285) [cs.CL]. url: <https://arxiv.org/abs/1902.07285>.
- [10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei. 'Language Models are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan and H.-T. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. url: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [11] X. Wang, H. Wang and D. Yang. 'Measure and Improve Robustness in NLP Models: A Survey'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by M. Carpuat, M.-C. de Marneffe and I. V. Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 4569–4586. doi: [10.18653/v1/2022.naacl-main.339](https://aclanthology.org/2022.naacl-main.339/). url: <https://aclanthology.org/2022.naacl-main.339/>.
- [12] M. Du, S. Mukherjee, Y. Cheng, M. Shokouhi, X. Hu and A. H. Awadallah. 'Robustness Challenges in Model Distillation and Pruning for Natural Language Understanding'. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by A. Vlachos and I. Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1766–1778. doi: [10.18653/v1/2023.eacl-main.129](https://aclanthology.org/2023.eacl-main.129/). url: <https://aclanthology.org/2023.eacl-main.129/>.
- [13] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov and T. Scialom. *Llama 2: Open Foundation*

- and Fine-Tuned Chat Models*. 18th July 2023. arXiv: [2307.09288](https://arxiv.org/abs/2307.09288) [cs.CL]. url: <https://arxiv.org/abs/2307.09288>.
- [14] M. Large, C. Galletly, N. Myles, C. J. Ryan and H. Myles. ‘Known unknowns and unknown unknowns in suicide risk assessment: Evidence from meta-analyses of aleatory and epistemic uncertainty’. In: *BJPsych Bulletin* 41.3 (June 2017), pp. 160–163. issn: 2056-4708. doi: [10.1192/pb.bp.116.054940](https://doi.org/10.1192/pb.bp.116.054940). url: <https://doi.org/10.1192/pb.bp.116.054940>.
- [15] K. Crawford. ‘Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics’. In: *Science, Technology, & Human Values* 41.1 (Jan. 2016), pp. 77–92. issn: 0162-2439. doi: [10.1177/0162243915589635](https://doi.org/10.1177/0162243915589635). url: <https://doi.org/10.1177/0162243915589635>.
- [16] H. Lakkaraju, E. Kamar, R. Caruana and E. Horvitz. ‘Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (13th Feb. 2017), pp. 2124–2132. issn: 2374-3468. doi: [10.1609/aaai.v31i1.10821](https://ojs.aaai.org/index.php/AAAI/article/view/10821). url: <https://ojs.aaai.org/index.php/AAAI/article/view/10821>.
- [17] Á. A. Cabrera, A. J. Druck, J. I. Hong and A. Perer. ‘Discovering and Validating AI Errors With Crowdsourced Failure Reports’. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021). Article 425, pp. 1–22. issn: 2573-0142. doi: [10.1145/3479569](https://doi.org/10.1145/3479569). url: <https://doi.org/10.1145/3479569>.
- [18] L. Han, X. Dong and G. Demartini. ‘Iterative Human-in-the-Loop Discovery of Unknown Unknowns in Image Datasets’. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9.1 (Nov. 2021), pp. 72–83. doi: [10.1609/hcomp.v9i1.18941](https://ojs.aaai.org/index.php/HCOMP/article/view/18941). url: <https://ojs.aaai.org/index.php/HCOMP/article/view/18941>.
- [19] M. A. Gluck, E. Mercado and C. E. Myers. *Learning and Memory. From Brain to Behavior*. 2nd ed. Worth Publishers, 7th Jan. 2013. isbn: 9781429298582. url: <https://books.google.com/books?id=BDABYwAACAAJ>.
- [20] M. T. Banich and D. Caccamise, eds. *Generalization of Knowledge. Multidisciplinary Perspectives*. Psychology Press, 7th Jan. 2011. isbn: 9781136945465. doi: [10.4324/9780203848036](https://doi.org/10.4324/9780203848036). url: <https://doi.org/10.4324/9780203848036>.
- [21] E. Allaway and K. McKeown. ‘Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 8913–8931. doi: [10.18653/v1/2020.emnlp-main.717](https://aclanthology.org/2020.emnlp-main.717/). url: <https://aclanthology.org/2020.emnlp-main.717/>.

- [22] B. M. Lake, R. Salakhutdinov and J. B. Tenenbaum. ‘Human-Level Concept Learning Through Probabilistic Program Induction’. In: *Science* 350.6266 (11th Dec. 2015), pp. 1332–1338. issn: 0036-8075. doi: [10.1126/science.aab3050](https://doi.org/10.1126/science.aab3050). url: <https://doi.org/10.1126/science.aab3050>.
- [23] F. C. Bartlett. *Thinking: An Experimental and Social Study*. London: Allen & Unwin, 1958. 203 pp. url: <https://openlibrary.org/books/OL6242500M>.
- [24] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng and C. Potts. ‘Learning Word Vectors for Sentiment Analysis’. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. by D. Lin, Y. Matsumoto and R. Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. url: <https://aclanthology.org/P11-1015/>.
- [25] W. B. Dolan and C. Brockett. ‘Automatically Constructing a Corpus of Sentential Paraphrases’. In: *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. 2005. url: <https://aclanthology.org/I05-5002/>.
- [26] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang. ‘SQuAD: 100,000+ Questions for Machine Comprehension of Text’. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by J. Su, K. Duh and X. Carreras. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. doi: [10.18653/v1/D16-1264](https://doi.org/10.18653/v1/D16-1264). url: <https://aclanthology.org/D16-1264/>.
- [27] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova. ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). url: <https://aclanthology.org/N19-1423/>.
- [28] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin and Y. Qi. ‘TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Q. Liu and D. Schlangen. Online: Association for Computational Linguistics, Oct. 2020, pp. 119–126. doi: [10.18653/v1/2020.emnlp-demos.16](https://doi.org/10.18653/v1/2020.emnlp-demos.16). url: <https://aclanthology.org/2020.emnlp-demos.16/>.
- [29] D. Jin, Z. Jin, J. T. Zhou and P. Szolovits. ‘Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.5 (Apr. 2020), pp. 8018–8025. issn: 2159-5399. doi: [10.1609/aaai.v34i05.6311](https://doi.org/10.1609/aaai.v34i05.6311). url: <https://ojs.aaai.org/index.php/AAAI/article/view/6311>.

- [30] J. Gao, J. Lanchantin, M. L. Soffa and Y. Qi. ‘Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers’. In: *2018 IEEE Security and Privacy Workshops (SPW)*. San Francisco, CA, USA: IEEE, May 2018, pp. 50–56. doi: [10.1109/SPW.2018.00016](https://doi.org/10.1109/SPW.2018.00016). url: <https://doi.org/10.1109/SPW.2018.00016>.
- [31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen. ‘LoRA: Low-Rank Adaptation of Large Language Models’. In: *International Conference on Learning Representations*. 2022. url: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [32] I. Loshchilov and F. Hutter. ‘Decoupled Weight Decay Regularization’. In: *International Conference on Learning Representations*. New Orleans, Louisiana, USA: OpenReview.net, May 2019. url: <https://openreview.net/forum?id=Bkg6RiCqY7>.
- [33] D. M. Oppenheimer, T. Meyvis and N. Davidenko. ‘Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power’. In: *Journal of Experimental Social Psychology* 45.4 (July 2009), pp. 867–872. issn: 0022-1031. doi: [10.1016/j.jesp.2009.03.009](https://doi.org/10.1016/j.jesp.2009.03.009). url: <https://doi.org/10.1016/j.jesp.2009.03.009>.
- [34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi. ‘BERTScore: Evaluating Text Generation with BERT’. In: *International Conference on Learning Representations*. ICLR 2020 Poster. Virtual Event: OpenReview.net, Apr. 2020. url: <https://openreview.net/forum?id=SkeHuCVFDr>.
- [35] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner and A. Madry. ‘Robustness May Be at Odds with Accuracy’. In: *International Conference on Learning Representations*. 2019. url: <https://openreview.net/forum?id=SyxAb30cY7>.
- [36] H. Yu, Z. Yang, K. Pelrine, J. F. Godbout and R. Rabbany. *Open, Closed, or Small Language Models for Text Classification?* 19th Aug. 2023. doi: [10.48550/arXiv.2308.10092](https://arxiv.org/abs/2308.10092). arXiv: [2308.10092](https://arxiv.org/abs/2308.10092) [cs.CL]. url: <https://arxiv.org/abs/2308.10092>.
- [37] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger. ‘On Calibration of Modern Neural Networks’. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 1321–1330. url: <https://proceedings.mlr.press/v70/guo17a.html>.
- [38] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran and M. Lucic. ‘Revisiting the Calibration of Modern Neural Networks’. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 15682–15694. url: https://proceedings.neurips.cc/paper_files/paper/2021/file/8420d359404024567b5aefda1231af24-Paper.pdf.

- [39] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn and C. Manning. ‘Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback’. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5433–5442. doi: [10.18653/v1/2023.emnlp-main.330](https://doi.org/10.18653/v1/2023.emnlp-main.330). url: <https://aclanthology.org/2023.emnlp-main.330/>.
- [40] M. T. Ribeiro, S. Singh and C. Guestrin. ‘Semantically Equivalent Adversarial Rules for Debugging NLP Models’. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 856–865. doi: [10.18653/v1/P18-1079](https://doi.org/10.18653/v1/P18-1079). url: <https://aclanthology.org/P18-1079/>.
- [41] E. Wallace, S. Feng, N. Kandpal, M. Gardner and S. Singh. ‘Universal Adversarial Triggers for Attacking and Analyzing NLP’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2153–2162. doi: [10.18653/v1/D19-1221](https://doi.org/10.18653/v1/D19-1221). url: <https://aclanthology.org/D19-1221/>.
- [42] T. Wang, X. Wang, Y. Qin, B. Packer, K. Li, J. Chen, A. Beutel and E. Chi. ‘CAT-Gen: Improving Robustness in NLP Models via Controlled Adversarial Text Generation’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 5141–5146. doi: [10.18653/v1/2020.emnlp-main.417](https://doi.org/10.18653/v1/2020.emnlp-main.417). url: <https://aclanthology.org/2020.emnlp-main.417/>.
- [43] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu. ‘Towards Deep Learning Models Resistant to Adversarial Attacks’. In: *International Conference on Learning Representations*. OpenReview.net, 15th Feb. 2018. url: <https://openreview.net/forum?id=rJzIBfZAb>.
- [44] T. Pang, X. Yang, Y. Dong, H. Su and J. Zhu. ‘Bag of Tricks for Adversarial Training’. In: *International Conference on Learning Representations*. OpenReview.net, 2021. url: <https://openreview.net/forum?id=Xb8xvrtB8Ce>.
- [45] R. Puri, R. Spring, M. Patwary, M. Shoeybi and B. Catanzaro. *Training Question Answering Models from Synthetic Data*. 22nd Feb. 2020. arXiv: [2002.09599](https://arxiv.org/abs/2002.09599) [cs.CL]. url: <https://arxiv.org/abs/2002.09599>.
- [46] W. E. Zhang, Q. Z. Sheng and A. Alhazmi. *Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey*. 21st Jan. 2019. arXiv: [1901.06796](https://arxiv.org/abs/1901.06796) [cs.CL]. url: <https://arxiv.org/abs/1901.06796>.

- [47] X. He, I. Nassar, J. Kiros, G. Haffari and M. Norouzi. ‘Generate, Annotate, and Learn: NLP with Synthetic Text’. In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 826–842. issn: 2307-387X. doi: [10.1162/tacl_a_00492](https://doi.org/10.1162/tacl_a_00492). url: https://doi.org/10.1162/tacl_a_00492.
- [48] N. Reimers and I. Gurevych. ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. doi: [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410). url: <https://aclanthology.org/D19-1410/>.
- [49] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen and W.-t. Yih. ‘Dense Passage Retrieval for Open-Domain Question Answering’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. doi: [10.18653/v1/2020.emnlp-main.550](https://doi.org/10.18653/v1/2020.emnlp-main.550). url: <https://aclanthology.org/2020.emnlp-main.550/>.
- [50] J. Ni, C. Qu, J. Lu, Z. Dai, G. H. Ábrego, J. Ma, V. Y. Zhao, Y. Luan, K. B. Hall, M.-W. Chang and Y. Yang. *Large Dual Encoders Are Generalizable Retrievers*. 15th Dec. 2021. arXiv: [2112.07899](https://arxiv.org/abs/2112.07899) [cs.IR]. url: <https://arxiv.org/abs/2112.07899>.
- [51] N. Thakur, N. Reimers, A. Rüclé, A. Srivastava and I. Gurevych. *BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models*. 17th Apr. 2021. arXiv: [2104.08663](https://arxiv.org/abs/2104.08663) [cs.IR]. url: <https://arxiv.org/abs/2104.08663>.
- [52] S. Robertson and H. Zaragoza. ‘The Probabilistic Relevance Framework: BM25 and Beyond’. In: *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389. doi: [10.1561/1500000019](https://doi.org/10.1561/1500000019). url: <https://doi.org/10.1561/1500000019>.
- [53] J. X. Morris and A. M. Rush. *Contextual Document Embeddings*. 2024. arXiv: [2410.02525](https://arxiv.org/abs/2410.02525) [cs.CL]. url: <https://arxiv.org/abs/2410.02525>.
- [54] R. Shao, R. Qiao, V. Kishore, N. Muennighoff, X. V. Lin, D. Rus, B. K. H. Low, S. Min, W.-t. Yih, P. W. Koh and L. Zettlemoyer. *ReasonIR: Training Retrievers for Reasoning Tasks*. 29th Apr. 2025. arXiv: [2504.20595](https://arxiv.org/abs/2504.20595) [cs.AI]. url: <https://arxiv.org/abs/2504.20595>.
- [55] J. Robinson, C.-Y. Chuang, S. Sra and S. Jegelka. *Contrastive Learning with Hard Negative Samples*. 9th Oct. 2020. arXiv: [2010.04592](https://arxiv.org/abs/2010.04592) [cs.LG]. url: <https://arxiv.org/abs/2010.04592>.

- [56] K. Wang, N. Thakur, N. Reimers and I. Gurevych. ‘GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval’. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by M. Carpuat, M.-C. de Marneffe and I. V. Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 2345–2360. doi: [10.18653/v1/2022.naacl-main.168](https://doi.org/10.18653/v1/2022.naacl-main.168). url: <https://aclanthology.org/2022.naacl-main.168/>.
- [57] F. Jiang, Q. Xu, T. Drummond and T. Cohn. ‘Boot and Switch: Alternating Distillation for Zero-Shot Dense Retrieval’. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 912–931. doi: [10.18653/v1/2023.findings-emnlp.65](https://doi.org/10.18653/v1/2023.findings-emnlp.65). url: <https://doi.org/10.18653/v1/2023.findings-emnlp.65>.
- [58] O. Khattab and M. Zaharia. ‘ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT’. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2020, pp. 39–48. doi: [10.1145/3397271.3401075](https://doi.org/10.1145/3397271.3401075). url: <https://doi.org/10.1145/3397271.3401075>.
- [59] N. Muennighoff, N. Tazi, L. Magne and N. Reimers. *MTEB: Massive Text Embedding Benchmark*. 13th Oct. 2022. arXiv: [2210.07316](https://arxiv.org/abs/2210.07316) [cs.CL]. url: <https://arxiv.org/abs/2210.07316>.
- [60] V. Boteva, D. Gholipour, A. Sokolov and S. Riezler. ‘A Full-Text Learning to Rank Dataset for Medical Information Retrieval’. In: *Advances in Information Retrieval*. Ed. by N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff and G. Silvello. Vol. 9626. Lecture Notes in Computer Science. Padua, Italy: Springer International Publishing, Mar. 2016, pp. 716–722. isbn: 9783319306704. doi: [10.1007/978-3-319-30671-1_58](https://doi.org/10.1007/978-3-319-30671-1_58). url: https://doi.org/10.1007/978-3-319-30671-1_58.
- [61] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie and M. Zhang. *Towards General Text Embeddings with Multi-stage Contrastive Learning*. 7th Aug. 2023. arXiv: [2308.03281](https://arxiv.org/abs/2308.03281) [cs.CL]. url: <https://arxiv.org/abs/2308.03281>.
- [62] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian and J.-Y. Nie. ‘C-Pack: Packed Resources For General Chinese Embeddings’. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Washington DC, USA: ACM, 10th July 2024, pp. 641–649. doi: [10.1145/3626772.3657878](https://doi.org/10.1145/3626772.3657878). url: <https://doi.org/10.1145/3626772.3657878>.
- [63] J. J. Rocchio. ‘Relevance Feedback in Information Retrieval’. In: *The SMART Retrieval System: Experiments in Automatic Document Processing*. Ed. by G. Salton. Prentice-Hall, 1971, pp. 313–323. url: <https://openlibrary.org/books/OL46956334M>.

- [64] C. Li, Y. Sun, B. He, L. Wang, K. Hui, A. Yates, L. Sun and J. Xu. 'NPRF: A Neural Pseudo Relevance Feedback Framework for Ad-hoc Information Retrieval'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by E. Riloff, D. Chiang, J. Hockenmaier and J. Tsujii. Brussels, Belgium: Association for Computational Linguistics, Oct.–Nov. 2018, pp. 4482–4491. doi: [10.18653/v1/D18-1478](https://doi.org/10.18653/v1/D18-1478). url: <https://aclanthology.org/D18-1478/>.
- [65] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin and E. Grave. *Unsupervised Dense Information Retrieval with Contrastive Learning*. 16th Dec. 2021. arXiv: [2112.09118](https://arxiv.org/abs/2112.09118) [cs.IR]. url: <https://arxiv.org/abs/2112.09118>.
- [66] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed and A. Overwijk. 'Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval'. In: *International Conference on Learning Representations*. 2021. url: <https://openreview.net/forum?id=zeFrfgYZln>.
- [67] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts and M. Zaharia. 'ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by M. Carpuat, M.-C. de Marneffe and I. V. Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3715–3734. doi: [10.18653/v1/2022.naacl-main.272](https://doi.org/10.18653/v1/2022.naacl-main.272). url: <https://aclanthology.org/2022.naacl-main.272/>.
- [68] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer and M. Lewis. 'Generalization through Memorization: Nearest Neighbor Language Models'. In: *International Conference on Learning Representations*. 2020. url: <https://openreview.net/forum?id=HklBjCEKvH>.
- [69] J. Kossen, N. Band, C. Lyle, A. N. Gomez, T. Rainforth and Y. Gal. *Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning*. 4th June 2021. arXiv: [2106.02584](https://arxiv.org/abs/2106.02584) [cs.LG]. url: <https://arxiv.org/abs/2106.02584>.
- [70] C. Xu, D. Guo, N. Duan and J. McAuley. 'LaPraDoR: Unsupervised Pretrained Dense Retriever for Zero-Shot Text Retrieval'. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by S. Muresan, P. Nakov and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3557–3569. doi: [10.18653/v1/2022.findings-acl.281](https://doi.org/10.18653/v1/2022.findings-acl.281). url: <https://aclanthology.org/2022.findings-acl.281/>.
- [71] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder and F. Wei. *SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval*. 6th July 2022. arXiv: [2207.02578](https://arxiv.org/abs/2207.02578) [cs.IR]. url: <https://arxiv.org/abs/2207.02578> (visited on 02/01/2026).

- [72] K. Wang, N. Reimers and I. Gurevych. 'TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning'. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by M.-F. Moens, X. Huang, L. Specia and S. W.-t. Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 671–688. doi: [10.18653/v1/2021.findings-emnlp.59](https://doi.org/10.18653/v1/2021.findings-emnlp.59). url: <https://aclanthology.org/2021.findings-emnlp.59/>.
- [73] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan and S. Gelly. 'Parameter-Efficient Transfer Learning for NLP'. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 9th–15th June 2019, pp. 2790–2799. url: <https://proceedings.mlr.press/v97/houlsby19a.html>.
- [74] Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. B. Hall and M.-W. Chang. *Promptagator: Few-shot Dense Retrieval From 8 Examples*. 23rd Sept. 2022. arXiv: [2209.11755](https://arxiv.org/abs/2209.11755) [cs.CL]. url: <https://arxiv.org/abs/2209.11755>.
- [75] X. Wang, C. Macdonald, N. Tonello and I. Ounis. 'Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval'. In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. Virtual Event, Canada: ACM, 11th July 2021, pp. 297–306. doi: [10.1145/3471158.3472250](https://doi.org/10.1145/3471158.3472250). url: <https://doi.org/10.1145/3471158.3472250>.
- [76] L. Bonifacio, H. Abonizio, M. Fadaee and R. Nogueira. 'InPars: Unsupervised Dataset Generation for Information Retrieval'. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain: Association for Computing Machinery, July 2022, pp. 2387–2392. isbn: 9781450387323. doi: [10.1145/3477495.3531863](https://doi.org/10.1145/3477495.3531863). url: <https://doi.org/10.1145/3477495.3531863>.
- [77] I. Ziegler, A. Köksal, D. Elliott and H. Schütze. 'CRAFT Your Dataset: Task-Specific Synthetic Dataset Generation Through Corpus Retrieval and Augmentation'. In: *Transactions of the Association for Computational Linguistics* 13 (1st Dec. 2025), pp. 1693–1721. issn: 2307-387X. doi: [10.1162/tac1.a.56](https://doi.org/10.1162/tac1.a.56). url: <https://doi.org/10.1162/tac1.a.56>.
- [78] A. V. Solatorio. *GSTEmbed: Guided In-sample Selection of Training Negatives for Text Embedding Fine-tuning*. 26th Feb. 2024. arXiv: [2402.16829](https://arxiv.org/abs/2402.16829) [cs.LG]. url: <https://arxiv.org/abs/2402.16829>.
- [79] I. Sutskever, O. Vinyals and Q. V. Le. 'Sequence to Sequence Learning with Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger. Vol. 27. Curran Associates, Inc., 2014. url: https://proceedings.neurips.cc/paper_files/paper/2014/file/5a18e133cbf9f257297f410bb7eca942-Paper.pdf.

- [80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin. 'Attention is All You Need'. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett. 2017, pp. 5998–6008. url: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [81] R. Hinami, S. Ishiwatari, K. Yasuda and Y. Matsui. *Towards Fully Automated Manga Translation*. 28th Dec. 2020. arXiv: [2012.14271](https://arxiv.org/abs/2012.14271) [cs.CL]. url: <https://arxiv.org/abs/2012.14271>.
- [82] H. Kaino, S. Sugihara, T. Kajiwara, T. Ninomiya, J. B. Tanner and S. Ishiwatari. 'Utilizing Longer Context than Speech Bubbles in Automated Manga Translation'. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti and N. Xue. Torino, Italia: ELRA and ICCL, May 2024, pp. 17337–17342. url: <https://aclanthology.org/2024.lrec-main.1505/>.
- [83] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki and K. Aizawa. 'Sketch-based Manga Retrieval Using Manga109 Dataset'. In: *Multimedia Tools and Applications* 76.20 (Oct. 2017), pp. 21811–21838. issn: 1380-7501. doi: [10.1007/S11042-016-4020-Z](https://doi.org/10.1007/S11042-016-4020-Z). url: <https://doi.org/10.1007/s11042-016-4020-z>.
- [84] T. Ogawa, A. Otsubo, R. Narita, Y. Matsui, T. Yamasaki and K. Aizawa. *Object Detection for Comics using Manga109 Annotations*. 23rd Mar. 2018. arXiv: [1803.08670](https://arxiv.org/abs/1803.08670) [cs.CV]. url: <https://arxiv.org/abs/1803.08670>.
- [85] C. Lyu, Z. Du, J. Xu, Y. Duan, M. Wu, T. Lynn, A. F. Aji, D. F. Wong, S. Liu and L. Wang. *A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models*. 2nd May 2023. arXiv: [2305.01181](https://arxiv.org/abs/2305.01181) [cs.CL]. url: <https://arxiv.org/abs/2305.01181>.
- [86] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify and H. H. Awadalla. *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*. 18th Feb. 2023. arXiv: [2302.09210](https://arxiv.org/abs/2302.09210) [cs.CL]. url: <https://arxiv.org/abs/2302.09210>.
- [87] C. Lyu, Z. Du, J. Xu, Y. Duan, M. Wu, T. Lynn, A. F. Aji, D. F. Wong, S. Liu and L. Wang. *A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models*. 2nd May 2023. arXiv: [2305.01181](https://arxiv.org/abs/2305.01181) [cs.CL]. url: <https://arxiv.org/abs/2305.01181>.
- [88] P. Świczekowska. 'Towards a direct Japanese-Polish machine translation system'. In: *Proceedings of the 8th Language & Technology Conference*. 2017.

- [89] J. Chen, R. Iwasaki, N. Mori, M. Okada and M. Ueno. ‘Understanding Multilingual Four-Scene Comics with Deep Learning Methods’. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Sydney, Australia: IEEE, Sept. 2019, pp. 32–37. doi: [10.1109/ICDARW.2019.00011](https://doi.org/10.1109/ICDARW.2019.00011). url: <https://doi.org/10.1109/ICDARW.2019.00011>.
- [90] H. Guo, B. Wang, J. Bai, J. Liu, J. Yang and Z. Li. ‘M2C: Towards Automatic Multimodal Manga Complement’. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by H. Bouamor, J. Pino and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9876–9882. doi: [10.18653/v1/2023.findings-emnlp.661](https://doi.org/10.18653/v1/2023.findings-emnlp.661). url: <https://aclanthology.org/2023.findings-emnlp.661>.
- [91] G. Harshavardhan, S. S. Kang, G. K. Kaur and S. Singla. ‘The Future of Graphic Novel Translation: Fully Automated Systems’. In: *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*. Chikkaballapur, India: IEEE, 18th Apr. 2024, pp. 1–8. doi: [10.1109/ickecs61492.2024.10617012](https://doi.org/10.1109/ickecs61492.2024.10617012). url: <https://doi.org/10.1109/ickecs61492.2024.10617012>.
- [92] A. G. Hapsani, F. Utamingrum and H. Tolle. ‘Optical character recognition on English comic digital data for automated language translation’. In: *Int. J. Advance Soft Compu. Appl* 9.3 (2017).
- [93] Z. He, T. Liang, W. Jiao, Z. Zhang, Y. Yang, R. Wang, Z. Tu, S. Shi and X. Wang. *Exploring Human-Like Translation Strategy with Large Language Models*. 6th May 2023. arXiv: [2305.04118](https://arxiv.org/abs/2305.04118) [cs.CL]. url: <https://arxiv.org/abs/2305.04118>.
- [94] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi and Z. Tu. ‘Document-Level Machine Translation with Large Language Models’. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16646–16661. doi: [10.18653/v1/2023.emnlp-main.1036](https://doi.org/10.18653/v1/2023.emnlp-main.1036). url: <https://aclanthology.org/2023.emnlp-main.1036/>.
- [95] T. Kocmi and C. Federmann. ‘Large Language Models Are State-of-the-Art Evaluators of Translation Quality’. In: *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. Ed. by M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ransinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton and H. Moniz. Tampere, Finland: European Association for Machine Translation, June 2023, pp. 193–203. url: <https://aclanthology.org/2023.eamt-1.19/>.
- [96] B. Zhang, B. Haddow and A. Birch. ‘Prompting Large Language Model for Machine Translation: A Case Study’. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato and J. Scarlett. Vol. 202. Proceedings

- of Machine Learning Research. PMLR, July 2023, pp. 41092–41110. url: <https://proceedings.mlr.press/v202/zhang23c.html>.
- [97] S. Zhu, S. Li, Y. Lei and D. Xiong. ‘PEIT: Bridging the Modality Gap with Pre-trained Models for End-to-End Image Translation’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 13433–13447. doi: [10.18653/v1/2023.acl-long.751](https://doi.org/10.18653/v1/2023.acl-long.751). url: <https://aclanthology.org/2023.acl-long.751>.
- [98] Z. Lan, J. Yu, X. Li, W. Zhang, J. Luan, B. Wang, D. Huang and J. Su. *Exploring Better Text Image Translation with Multimodal Codebook*. 27th May 2023. arXiv: [2305.17415](https://arxiv.org/abs/2305.17415) [cs.CL]. url: <https://arxiv.org/abs/2305.17415>.
- [99] J. Gwinnup and K. Duh. *A Survey of Vision-Language Pre-training from the Lens of Multimodal Machine Translation*. 12th June 2023. arXiv: [2306.07198](https://arxiv.org/abs/2306.07198) [cs.CL]. url: <https://arxiv.org/abs/2306.07198>.
- [100] D. Elliott, S. Frank, K. Sima’an and L. Specia. ‘Multi30K: Multilingual English-German Image Descriptions’. In: *Proceedings of the 5th Workshop on Vision and Language*. Ed. by A. Belz, E. Erdem, K. Mikolajczyk and K. Pastra. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 70–74. doi: [10.18653/v1/W16-3210](https://doi.org/10.18653/v1/W16-3210). url: <https://aclanthology.org/W16-3210>.
- [101] W. Guo, Q. Fang, D. Yu and Y. Feng. ‘Bridging the Gap between Synthetic and Authentic Images for Multimodal Machine Translation’. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2863–2874. doi: [10.18653/v1/2023.emnlp-main.173](https://doi.org/10.18653/v1/2023.emnlp-main.173). url: <https://aclanthology.org/2023.emnlp-main.173>.
- [102] H. Guo, J. Liu, H. Huang, J. Yang, Z. Li, D. Zhang and Z. Cui. ‘LVP-M3: Language-aware Visual Prompt for Multilingual Multimodal Machine Translation’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2862–2872. doi: [10.18653/v1/2022.emnlp-main.184](https://doi.org/10.18653/v1/2022.emnlp-main.184). url: <https://aclanthology.org/2022.emnlp-main.184>.
- [103] X. Huang, J. Zhang and C. Zong. ‘Contrastive Adversarial Training for Multi-Modal Machine Translation’. In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 22.6 (16th June 2023), pp. 1–18. issn: 2375-4699. doi: [10.1145/3587267](https://doi.org/10.1145/3587267). url: <https://doi.org/10.1145/3587267>.

- [104] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, N. Bjorck, V. Chaudhary, S. Som, X. Song and F. Wei. 'Language Is Not All You Need: Aligning Perception with Language Models'. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 72096–72109. url: https://proceedings.neurips.cc/paper_files/paper/2023/file/e425b75bac5742a008d643826428787c-Paper-Conference.pdf.
- [105] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu and E. Chen. 'A Survey on Multimodal Large Language Models'. In: *National Science Review* 11.12 (12th Nov. 2024). issn: 2095-5138. doi: [10.1093/nsr/nwae403](https://doi.org/10.1093/nsr/nwae403). url: <https://doi.org/10.1093/nsr/nwae403>.
- [106] M. Karpinska and M. Iyyer. 'Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist'. In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by P. Koehn, B. Haddow, T. Kocmi and C. Monz. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 419–451. doi: [10.18653/v1/2023.wmt-1.41](https://doi.org/10.18653/v1/2023.wmt-1.41). url: <https://aclanthology.org/2023.wmt-1.41/>.
- [107] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni and P. Liang. 'Lost in the Middle: How Language Models Use Long Contexts'. In: *Transactions of the Association for Computational Linguistics* 12 (23rd Feb. 2024), pp. 157–173. issn: 2307-387X. doi: [10.1162/tacl_a_00638](https://doi.org/10.1162/tacl_a_00638). url: https://doi.org/10.1162/tacl_a_00638.
- [108] G. Adams, A. Fabbri, F. Ladhak, E. Lehman and N. Elhadad. *From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting*. arXiv:2309.04269v1. 8th Sept. 2023. arXiv: [2309.04269](https://arxiv.org/abs/2309.04269) [cs.CL]. url: <https://arxiv.org/abs/2309.04269>.
- [109] S. Schroff. 'An alternative universe? Authors as copyright owners- the case of the Japanese Manga Industry'. In: *Creative Industries Journal* 12.1 (2nd Jan. 2019), pp. 125–150. issn: 1751-0694. doi: [10.1080/17510694.2018.1563420](https://doi.org/10.1080/17510694.2018.1563420). url: <https://doi.org/10.1080/17510694.2018.1563420>.
- [110] C. Rigaud, N.-V. Nguyen and J.-C. Burie. 'Text Block Segmentation in Comic Speech Bubbles'. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Lecture Notes in Computer Science. Springer International Publishing, 2021, pp. 250–261. isbn: 9783030687793. doi: [10.1007/978-3-030-68780-9_22](https://doi.org/10.1007/978-3-030-68780-9_22). url: https://doi.org/10.1007/978-3-030-68780-9_22.
- [111] M. S. Sharif, B. A. Romo, H. Maltby and A. Al-Bayatti. 'An Effective Hybrid Approach Based on Machine Learning Techniques for Auto-Translation: Japanese to English'. In: *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. Zallaq, Bahrain: IEEE, 29th Sept. 2021, pp. 557–562. doi: [10.1109/3ICT53449.2021.9581629](https://doi.org/10.1109/3ICT53449.2021.9581629). url: <https://doi.org/10.1109/3ICT53449.2021.9581629>.

- [112] R. Sachdeva and A. Zisserman. 'The Manga Whisperer: Automatically Generating Transcriptions for Comics'. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 16th June 2024, pp. 12967–12976. doi: [10.1109/CVPR52733.2024.01232](https://doi.org/10.1109/CVPR52733.2024.01232). url: <https://doi.org/10.1109/CVPR52733.2024.01232>.
- [113] H. Jaśkiewicz. 'Reprezentacja dialektu bawarskiego i dialektu Kansai w literaturze współczesnej w kontekście ideologii językowych w Niemczech i Japonii'. In: *Forum Filologiczne Ateneum* 2(8)2020 (31st Dec. 2020), pp. 85–109. issn: 2353-2912. doi: [10.36575/2353-2912/2\(8\)2020.085](https://doi.org/10.36575/2353-2912/2(8)2020.085). url: [https://doi.org/10.36575/2353-2912/2\(8\)2020.085](https://doi.org/10.36575/2353-2912/2(8)2020.085).
- [114] P. Duc-Harada. 'Znaczenie i wpływ języka postaci (yakuwarigo) na kształtowanie kompetencji językowych studentów japonistyki w Polsce'. In: *Ogrody Nauk i Sztuk* 9 (15th Aug. 2019), pp. 301–319. issn: 2084-1426. doi: [10.15503/onis2019.301.319](https://doi.org/10.15503/onis2019.301.319). url: <https://doi.org/10.15503/onis2019.301.319>.
- [115] M. Popović. 'chrF: character n-gram F-score for automatic MT evaluation'. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva and P. Pecina. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 392–395. doi: [10.18653/v1/W15-3049](https://doi.org/10.18653/v1/W15-3049). url: <https://aclanthology.org/W15-3049/>.
- [116] K. Thai, M. Karpinska, K. Krishna, B. Ray, M. Inghilleri, J. Wieting and M. Iyyer. 'Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature'. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9882–9902. doi: [10.18653/v1/2022.emnlp-main.672](https://doi.org/10.18653/v1/2022.emnlp-main.672). url: <https://aclanthology.org/2022.emnlp-main.672/>.
- [117] N. Mathur, T. Baldwin and T. Cohn. *Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Metrics*. 11th June 2020. arXiv: [2006.06264](https://arxiv.org/abs/2006.06264) [cs.CL]. url: <https://arxiv.org/abs/2006.06264>.
- [118] T. Kocmi, V. Zouhar, C. Federmann and M. Post. *Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies*. 12th Jan. 2024. arXiv: [2401.06760](https://arxiv.org/abs/2401.06760) [cs.CL]. url: <https://arxiv.org/abs/2401.06760>.
- [119] H. Saadany and C. Orasan. 'BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text'. In: *Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*. INCOMA Ltd. Shoumen, Bulgaria, 2021, pp. 48–56. doi: [10.26615/978-954-452-071-7_006](https://doi.org/10.26615/978-954-452-071-7_006). url: https://doi.org/10.26615/978-954-452-071-7_006.

- [120] M. Hanna and O. Bojar. ‘A Fine-Grained Analysis of BERTScore’. In: *Proceedings of the Sixth Conference on Machine Translation*. Ed. by L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita and C. Monz. Online: Association for Computational Linguistics, Nov. 2021, pp. 507–517. url: <https://aclanthology.org/2021.wmt-1.59/>.
- [121] T. Sellam, D. Das and A. Parikh. ‘BLEURT: Learning Robust Metrics for Text Generation’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schlueter and J. Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 7881–7892. doi: [10.18653/v1/2020.acl-main.704](https://doi.org/10.18653/v1/2020.acl-main.704). url: <https://aclanthology.org/2020.acl-main.704/>.
- [122] A. Pu, H. W. Chung, A. Parikh, S. Gehrmann and T. Sellam. ‘Learning Compact Metrics for MT’. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by M.-F. Moens, X. Huang, L. Specia and S. W.-t. Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 751–762. doi: [10.18653/v1/2021.emnlp-main.58](https://doi.org/10.18653/v1/2021.emnlp-main.58). url: <https://aclanthology.org/2021.emnlp-main.58/>.
- [123] N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo and A. F. T. Martins. ‘xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection’. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 979–995. issn: 2307-387X. doi: [10.1162/TACL_A_00683](https://doi.org/10.1162/TACL_A_00683). url: https://doi.org/10.1162/tacl_a_00683.
- [124] M. Freitag, N. Mathur, C.-k. Lo, E. Avramidis, R. Rei, B. Thompson, T. Kocmi, F. Blain, D. Deutsch, C. Stewart, C. Zerva, S. Castilho, A. Lavie and G. Foster. ‘Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent’. In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by P. Koehn, B. Haddow, T. Kocmi and C. Monz. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 578–628. doi: [10.18653/v1/2023.wmt-1.51](https://doi.org/10.18653/v1/2023.wmt-1.51). url: <https://aclanthology.org/2023.wmt-1.51/>.
- [125] A. R. Lommel, A. Burchardt and H. Uszkoreit. ‘Multidimensional quality metrics: a flexible system for assessing translation quality’. In: *Proceedings of Translating and the Computer* 35. London, UK: Aslib, 28th–29th Nov. 2013. url: <https://aclanthology.org/2013.tc-1.6/>.
- [126] International Organization for Standardization. *ISO 5060:2024 – Translation services — Evaluation of translation output — General guidance*. International Standard. Geneva, Switzerland, Feb. 2024. url: <https://www.iso.org/obp/ui/#iso:std:iso:5060:ed-1:v1:en>.

- [127] OpenAI. *GPT-4 Technical Report*. arXiv:2303.08774v6. 4th Mar. 2024. arXiv: [2303.08774](https://arxiv.org/abs/2303.08774) [cs.CL]. url: <https://arxiv.org/abs/2303.08774> (visited on 29/12/2025).
- [128] M. Levy, A. Jacoby and Y. Goldberg. *Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models*. 19th Feb. 2024. arXiv: [2402.14848](https://arxiv.org/abs/2402.14848) [cs.CL]. url: <https://arxiv.org/abs/2402.14848>.

4

Beyond Correctness: Eliciting Human Values for Improved Alignment

Robust deployment in value-laden settings often fails not through factual error but through misaligned decisions made with misplaced confidence. Building on the thesis’s three-pillar approach laid out in [Chapter 1](#), this chapter formalizes Pillar 3: a value-sensitive handover framework that replaces confidence-only thresholds with a ratio-scale signal of user-perceived value elicited via Magnitude Estimation (ME). We map ME judgments to scenario-level utilities for correct/incorrect decisions and rejection, and optimize global or class-conditional thresholds to maximize expected total value, enabling model selection by value rather than accuracy. In a social media moderation study, our ME methodology exhibits strong reliability and validity and, when used to optimize reject policies, improves total value on both seen and unseen data compared to purely confidence-based baselines. This approach addresses the miscalibration of model confidence and steers escalation and deployment choices according to human values rather than proxy metrics. The result is a principled, auditable pathway for human–AI collaboration that prioritizes alignment and stakeholder priorities in high-stakes contexts.

Parts of this chapter are based on work published in:

- *How Do You Feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection* [1].

4.1. OVERVIEW AND MOTIVATION

Language models increasingly support high-stakes decisions across moderation, healthcare, and governance. Their fluency and breadth of knowledge motivate ambitious deployments, yet practical use reveals a persistent gap between apparent competence and reliable, value-aligned behavior [2]. Models can produce specific, confident outputs that are nevertheless wrong or misaligned with stakeholder priorities. This gap is exacerbated under distribution shift and novel inputs, where performance and calibration degrade [3, 4]. While benchmark results and impressive demonstrations suggest wide applicability [5, 6], robust deployment requires a principled way to decide when to trust model outputs and when to escalate to human experts.

As outlined in [Chapter 1](#), high-confidence errors under shift and miscalibration motivate escalation criteria grounded in stakeholder value rather than proxy metrics. Conventional escalation policies rely on model confidence thresholds or accuracy-oriented model selection. However, confidence is often miscalibrated — including for modern models trained with instruction tuning and human feedback — and can fail to track reliability in the wild [7, 8]. Miscalibrated certainty is particularly dangerous in value-laden contexts, where high-confidence errors bypass human oversight precisely when expert intervention is most needed. In addition to calibration issues, confidence-based policies implicitly optimize proxy objectives detached from the actual distribution of harms and benefits that stakeholders care about. This misalignment manifests as automation bias and over-reliance on model outputs when confidence is high [9], and as degraded human–AI collaboration performance when mental models are not aligned with the system’s limitations [10]. While selective prediction and reject-option methods learn confidence thresholds from empirical correctness [11], they typically do not encode stakeholder-specific utilities for different error types. A robustness program centered solely on accuracy, empirical correctness, or confidence is therefore insufficient for governing when models should act autonomously and when they should not.

We ground the chapter in the following research question from [chapter 1](#):

RQ3. Can a quantitative measure of user-perceived value provide a more reliable signal than model confidence for determining when to escalate decisions to human experts in value-laden contexts?

We develop Pillar 3: improving alignment by value-sensitive human–AI collaboration via principled escalation. We propose to replace purely confidence-based thresholds with a ratio-scale signal of user-perceived value elicited through Magnitude Estimation (ME) [12]. Rather than asking, “How sure is the model?”, we ask, “How much does this decision outcome matter to stakeholders?” The resulting per-scenario utilities quantify perceived gains for correct decisions and costs for incorrect or rejected decisions. These values induce thresholding policies that optimize expected total value, moving selection and escalation criteria closer to what stakeholders actually care about. In contrast to pairwise-preference approaches used in RLHF [13], ME provides a cardinal signal suitable for direct use in policy optimization.

Positioning within the thesis, Pillar 3 complements diagnostics from Pillar 1 ([Chapter 2](#)) and targeted data interventions from Pillar 2 ([Chapter 3](#)). Pillar 1 shows that distilled reasoning improvements can be driven by stylistic replication rather than genuine logical

competence [14], highlighting the need for explicit signals that reflect downstream stakes rather than surface form. Pillar 2 introduces agent-driven, targeted synthetic data to characterize and remediate blind spots [15]. While these diagnose and intervene on failure modes, residual disagreements about acceptable risk persist in deployment. Pillar 3 provides the governance mechanism: use value-sensitive thresholds to decide when to accept model outputs and when to escalate, aligning system behavior with stakeholder priorities even when confidence is unreliable.

We instantiate and validate this approach in social media moderation, a domain where the stakes of errors are asymmetric and socially consequential. Building on prior human-centered robustness perspectives [2] and evidence that modern models can exhibit strong capabilities yet remain brittle under shift [3, 4], we frame moderation as a decision with an explicit reject option (handover to humans). Using ME, we elicit scenario-level utilities for correctly accepted, incorrectly accepted, and rejected decisions, then learn thresholds that maximize expected total value on validation data. Empirically, and consistent with prior results, we find that value-optimized thresholds differ from accuracy-optimized ones and can better select appropriate models, improving total value in deployment [1].

The broader implication is a shift from proxy optimization to stakeholder-centered governance. Where calibration is imperfect [7, 8] and distribution shift common, confidence in isolation is an insufficient basis for escalation. ME provides a principled, ratio-scale alternative that we can use to define and optimize policies directly against stakeholder value. In doing so, Pillar 3 connects high-level human-centered robustness goals [2] to actionable decision rules and integrates with targeted knowledge injection (Chapter 3) and reasoning diagnostics (Chapter 2). The remainder of this chapter introduces background on human–AI collaboration and value elicitation, including the reject option and selective prediction, formalizes the value model and threshold optimization, and presents measurements and results from our hate speech study. We close with limitations and ethics, and discuss how value-sensitive escalation complements the thesis’s reasoning and knowledge pillars.

To conclude this introduction and make RQ3 empirically actionable for the remainder of the chapter, we articulate the hypothesis that guides our evaluation.

Hypothesis H3a Magnitude Estimation-based escalation policies achieve significantly higher human–AI agreement and outperform confidence-threshold baselines in the accuracy of escalation decisions and total value, across value-sensitive scenarios.

We test H3a by mapping ME judgments to scenario utilities, optimizing value-aware thresholds, and comparing against confidence-based selection and reject rules in our hate speech moderation study, reporting agreement and value metrics on both seen and unseen data.

4.2. INTRODUCTION TO MEASURING USER-PERCEIVED VALUE

Hateful content spread online through social media remains a significant problem. Ignoring its presence can lead to psychological harm and even result in violence

and other conflicts [16–19]. Governmental institutions and social media platforms are increasingly aware of these risks and are combating hate speech. For example, the European Union developed a Code of Conduct on countering hate speech [20], requesting large social media companies to moderate hate speech and report their progress yearly. However, results reported so far are not yet satisfactory, as, for example, less than 5% of hateful content has been removed from Facebook [21].

Hateful content moderation is either carried out manually or automatically by computational algorithms, where manual moderation may be more reliable but is not scalable to handle the deluge of user-generated content [22]. Further, continuous exposure to harmful content can be harmful to moderators as it can induce mental issues and potentially even lead to acts of self-harm [23]. Computational solutions are, therefore, urgently in demand by online platforms [24]. The methods considered best suited to this task are mainly based on machine learning, which has achieved reasonable performance at scale [25]. Yet, machine learning methods are far from being reliable, especially in dealing with hateful content previously unseen in the training data, which is often limited in size and biased [26]. Several recent studies on hate speech have shown a significant drop in machine learning performance when assessed on different data from those captured in the training phase [27, 28].

An approach that can combine the strengths of both previously mentioned approaches is human-AI collaboration, where humans are involved to solve AI-hard tasks, typically by taking over decisions where machines are unreliable [29, 30]. Such an approach is favorable in applications where decisions involve high-stakes and incorrect decisions can lead to damaging effects, as is the case for hate speech detection. Human-AI collaboration has been advocated in the human computation community [29, 31, 32] and, likely, is also an approach widely being used in enterprise applications such as search and conversational agents [33]. Despite this, methods for implementing human-AI collaboration so far are limited to predefined heuristics and have largely ignored the complexity of real-world problems, especially the cost of incorrect predictions being context-dependent.

Common heuristics of task handover from machines to humans are based on machine confidence: humans take over the task when the confidence of the machine in its decision is lower than a predefined threshold [30]. Such heuristics assume that machine confidence is well-calibrated, that is, a decision with high confidence should be more likely to be reliable and vice versa. This assumption however does not hold for many machine learning models, especially deep learning models, which may indicate high confidence when decisions are incorrect or vice versa [7, 34]. An improved approach is proposed by Geifman and El-Yaniv [35] which determines the appropriate confidence threshold based on empirical evidence of machine correctness, e.g., based on the accuracy-threshold curve obtained on an empirical dataset. Such an approach, however, does not take into account the implications of right or wrong decisions. Incorrect decisions in high-stakes domains have a larger impact that, in turn, should pose a stricter constraint on accepting machine decisions, e.g., via a higher confidence threshold. Similar ideas have recently been discussed in position papers that advocate the adoption of the notion of context-dependent *value* as a replacement of accuracy, the most common metric in machine decisions assessment [36, 37]. Value,

however, is an abstract term – it can be interpreted from social, ethical, or commercial perspectives [38–40] – yet the discussion on what creates value and how to measure it, specifically in a machine learning context, is limited due to it depending on the application.

In this chapter, we study the problem of operationalizing value perception of machine decisions and its integration into human-AI collaboration in the specific context of hate speech detection. We start by identifying several factors that may affect the value definition, namely the selection of a specific stakeholder’s standpoint and the relativity of value perception as affected by stakeholder expectation or regulation. We then operationalize user-perceived value in hate speech moderation scenarios, where a decision with a corresponding confidence has been made by a machine. To measure these perceptions, we explore several measurement scales and propose to select Magnitude Estimation (ME) [41] as the primary scale. ME allows the measurement of the magnitude of user (dis)agreement using an unbounded scale and makes it possible to obtain the relative ratios between the magnitudes of different machine decisions. These ratios are essential to determine the optimal confidence threshold for rejecting machine decisions (see [section 4.2.1](#)).

To validate ME in value operationalization, we designed a survey study where we recruited 160 participants. Each participant’s perception regarding a dataset of 40 selected hateful and non-hateful tweets and their (dis)agreement regarding the corresponding machine decisions were evaluated. Through a between-subject study, we show that Magnitude Estimation returns results with significantly higher inter-rater reliability compared to other scales, showing its suitability in measuring user perception. Our results show that the inter-rater reliability is significantly higher for incorrect decisions than for correct decisions, indicating a strong consensus among participants regarding the consequences of harm, as well as disagreements on what constitutes hate online. Further, users appear to be more negatively affected when a non-hateful post is subject to moderation than when an instance of hate speech is classified as non-hateful, implying that users would rather contend with an instance of hate speech than have an innocent user punished for a non-hateful post.

To demonstrate the utility of value integration in human-AI collaboration, we evaluate the effect of rejecting machine decisions made by three machine learning-based hate speech detection models – including traditional, deep learning, and BERT-based models [42] – in handling data from both seen and unseen sources. Our results show that for all three models, when evaluated on unseen data, the optimal confidence thresholds determined by the model-delivered value are much higher than the optimal thresholds on seen data. These results confirm the findings from previous studies on machine biases and demonstrate the effectiveness of using value as a target for optimally rejecting machine decisions. We further show that when selecting the optimal model, using value as the criterion returns different results compared to using accuracy. Note, that our approach to measuring value perception can be applied to different tasks and is model-agnostic.

In summary, we make the following key contributions:

- We introduce Magnitude Estimation as a scale for measuring user perception of machine decisions in scenarios where these decisions are correct and incorrect;

- We demonstrate the applicability of Magnitude Estimation through a between-subject survey study, as well as the utility of value for optimally rejecting machine decisions;
- We contribute a set of insights into user-perceived value of automated machine decisions, especially their attitudes towards different types of (mis)classifications.

We now introduce the background of value-sensitive rejection of machine decisions in a hybrid human-AI workflow, based on previous work [36, 43], and subsequently identifies factors that influence value perception in hate speech detection.

4.2.1. REJECT OPTION AND SELECTIVE PREDICTION

We consider the general case of human-AI collaboration as follows: the machine decision can either be accepted or rejected; if rejected, the decision will be taken over by a human decision maker. Formally, consider a binary classification problem for which we have a machine learning classifier, whose output on a data item x is confidence, \mathbf{c} , (e.g., the output from the softmax layer of a neural network). The rejection is dependent on a threshold denoted by $\tau \in [0, 1]$, which then modifies the final output of the machine as

$$\hat{y} = \begin{cases} y, & \mathbf{c}_y \geq \tau, \\ y_r, & \text{otherwise.} \end{cases} \quad (4.1)$$

where y denotes an accepted decision and y_r denotes the special decision of rejection, resulting in a human making the final decision.

We now discuss how the optimal confidence threshold for rejecting machine decisions is affected by the value formulation. We consider the binary classification case: when the machine decision is either positive (i.e., the content is deemed hateful) or negative (i.e., non-hateful). There is a value, V , attached to each of these, depending on whether this positive or negative decision is correct or not. This results in true positive (TP), true negative (TN), false positive (FP), false negative (FN), and rejected predictions as possible outcomes. V_{TP} and V_{TN} are positive, while V_{FP} , V_{FN} , and rejected predictions, V_r , are negative (i.e., costs). The optimal threshold for positive classifications is:

$$\tau_O^p = \frac{V_{FP}}{V_{FP} - V_{TP}} = \frac{\gamma^p}{\gamma^p + 1} \quad (4.2)$$

if we assume $V_{FP} = -\gamma^p \cdot V_{TP}$, that is, the cost of a false positive is γ^p times worse than the value of a true positive. Similarly, in the case of negative classifications, the optimal threshold would be $\tau_O^n = \frac{\gamma^n}{\gamma^n + 1}$ where $V_{FN} = -\gamma^n \cdot V_{TN}$, i.e., the cost of false negative is γ^n times worse than the value of a true negative.

When the cost of incorrect decisions is very high, i.e., $\gamma \gg 1$, the optimal confidence threshold would tend close to 1, meaning almost all machine decisions are rejected. When the cost of an incorrect decision is very low, i.e., $\gamma \approx 0$, the optimal threshold would be close to 0, and virtually all machine decisions are accepted. These results, therefore, follow our intuition. An important conclusion we can draw from [equation \(4.2\)](#) is that the optimal threshold is dependent *only on the ratio* of the value (or cost) between

an incorrect decision and that of a correct one (per class). Threshold optimization is the process of finding the threshold that maximizes value empirically. If a system is calibrated before use, simulations can be used to find the optimal theoretical threshold, which is the optimal τ that maximizes value. In this chapter, τ is determined by means of calibration, done by means of temperature scaling [44], followed by a calculation of the theoretical threshold based on the crowdsourced survey data, as it allows us to quantify and compare the opinions of participants on the value of true and false predictions and thus compute the ratios for our use case.

4.2.2. VALUE MEASUREMENT

We denote the value of classifying a data item correctly, or incorrectly, and that of rejecting a classification as V_c , V_w , and V_r , respectively. We make the following observations when considering value for hate speech detection: 1) Value is dependent not only on the machine learning model but also on the specific context to which the model is applied. For example, an incorrect prediction in the medical domain potentially has a bigger impact than one in e-commerce. In a high-stakes domain, generally, we would assume $V_c > V_r > V_w$ and thus a correct machine decision saves the cost of human moderation and accelerates the decision-making process, while a rejection requires additional human intervention. 2) Value interpretations from different stakeholders can vary. In hate speech detection, for example, a rejection of a machine decision induces the cost of human moderation from the business perspective, while from the user perspective what is more important is the exposure to hateful content. In our study, we choose to take the user's standpoint, and, as such, view V_r to come with an inherent cost since human moderation will be pending and the potentially hateful content will remain visible. 3) Value is affected by both stakeholder expectations and regulation. For example, in the hate speech detection case, when hateful content is posted, from the user's perspective, the value derived from a correct machine decision depends on the user's general expectation of how hateful content should be handled. Similarly, the legality of hate speech in certain jurisdictions may influence stakeholder perception.

Given the above observations, we now introduce the function to determine the total value, $V(\tau)$, of a given model with a reject option at the rejection threshold τ on a given dataset. Assuming that when accepted, correct decisions increase the overall value and when rejected, they decrease the overall value and vice versa, then, $V(\tau)$ may be formalized as:

$$V(\tau) = \sum_p (V_p - V_r)N_p + \sum_q (V_r - V_q)N_q, \quad (4.3)$$

where $p \in [TP, TN, FP, FN]$, $q \in [TP, TN, FP, FN]$, and N_p and N_q are the number of accepted and rejected data items for the difference scenarios, respectively. Note, that we assume that rejected decisions have a cost that decreases the overall value, i.e., V_r is negative, as users have to wait on a moderation decision. Thus, [equation \(4.3\)](#) allows us to summarize the value gained and the cost subtracted into a single value for the model by considering the value or cost of each scenario and how often it occurs, while also taking the cost of rejection into account.

4.3. SURVEY DESIGN

To define the relative value of scenarios, we design a survey to ask participants the degree to which they agree or disagree with the decisions of a fictional social media platform, SocialNet. These scenarios represent TP, TN, FP, FN, and rejected predictions. The TP and TN scenarios imply that SocialNet successfully detects whether a post is hateful or not hateful, respectively. The FP scenario means that SocialNet incorrectly predicts a non-hateful post as hateful, and conversely for the FN scenario. For example, in the FN scenario, the survey shows a hateful post to the subject and explains that SocialNet did not identify the post as hate speech.

4.3.1. CHOICE OF THE SCALE

We use ME as the primary scale. A Likert scale was initially considered, as it is widely used in research for retrieving participant opinions and is perhaps more intuitive for participants [45]. However, a Likert scale is not suitable in our case, as Likert-type items are ordinal, meaning that we only know the ranks but not the exact distances between the items [46]. In our case, computing the relative values (i.e., ratios) of our scenarios requires measuring the distances between different items, which cannot be provided by a Likert scale. On the contrary, the ME scale allows us to measure ratios by asking participants to provide numerical ratings. ME originated from psychophysics, where participants gave quantitative estimates of sensory magnitudes [41]. For sound loudness, a sound twice as loud as the previous one, should ideally receive a rating twice as large.

Researchers have previously applied the ME scale to different physical stimuli (e.g. line length, brightness, or duration) and proved that the results are reproducible, as well as that the data has ratio properties [47]. Other works have shown that the ME technique is also helpful for rating abstract types of stimuli, such as judging the relevance of documents [48], the linguistic acceptability of sentences [49], and the usability of system interfaces [50]. Thus, we conclude that ME is a promising method for judging hate speech.

4.3.2. NORMALIZATION AND VALIDATION OF THE SCALE

The ME scale is unbounded. For example, suppose we first show a scenario and the participant provides a value (e.g., 100) to indicate the degree of agreement. Suppose we next present a scenario that the participant agrees with more. The participant can always provide a higher value (e.g., 125) and not be restricted within a fixed range. The results need to be normalized as different participants rate the agreement/disagreement degree differently.

Multiple solutions exist for normalizing the ME scale, such as modulus normalization, which uses geometric averaging to preserve the ratio information [47, 50]. Unlike the unipolar ME scales used in previous research [49, 50], we use bipolar scales. Using arithmetic averaging is inappropriate since it uses logarithmic calculations and would disrupt the ratio scale properties [47]. Therefore, we normalize the results by dividing the magnitude estimates of each subject by their maximum estimate. We multiply the

normalized magnitude estimates by 100 for the sake of clarity. This way, all magnitudes estimates are in the range $[-100, 100]$ while maintaining their ratio properties.

Most previous research using the ME scale applies validation, such as cross-modality validation, where estimated magnitudes are compared to the physical stimuli using correlation analysis [49]. Cross-modality validation is difficult in domains that do not have exact measures of stimuli, such as hate speech. Some previous work compared ME with other validated scales [48]. In our case, we use the 100-level scale to validate the ME scale by analyzing their correlation [51], which is a form of convergent validation [52].

4.3.3. PARTICIPANTS AND DATA

We use Prolific to recruit crowd workers for the study.¹ Participants need to be at least 18 years of age, be fluent in English, and have an approval rating of over 90%. Participants also need to have experience using a social media platform regularly (at least once a month). Every participant is paid an hourly wage of 9 GBP, exceeding the UK minimum wage at the time of the study. Regarding sample size, we recruit 24 participants for the pilot study and 136 participants for the official study. Of the recruited participants, 50% identified as female, though Gold [53] showed that there is no significant difference when perceiving hate between genders. Half of the participants are assigned the ME scale and the other half the 100-level scale. We choose a 90% Confidence Interval (CI) and 10% Margin of Error (MoE) for this study due to budget limitations. There are billions of social media users, and according to Müller, Sedley and Ferrall-Nunge [54], we need a sample size of 68 participants per measurement scale, i.e., 136 participants, to reach the desired CI and MoE.

The final dataset consists of 20 hateful and 20 non-hateful social media posts from a public dataset [55] to build the machine decision scenarios (TP, TN, FP, FN, and rejection). The dataset contains 13,000 English tweets, and each tweet is annotated with three categories: hate speech (yes/no), target (group/individual), and aggressiveness (yes/no). We first exclude tweets that are replies or contained mentions or URLs since they have unclear contexts. Finally, we use clustering analysis to select 40 tweets for our study. We use a cluster size of 20 for the non-hateful tweets and sample one tweet per cluster by taking the nearest sample to each cluster centroid to obtain each cluster's most representative tweets. For the hateful tweets, we first divide them into four groups using the target and aggressiveness categories. Similarly, for each hateful tweet group, we use a cluster size of 5 and sample one tweet per cluster. We perform latent semantic analysis (LSA), which is a combination of term frequency-inverse document frequency (TF-IDF) and Singular Value Decomposition (SVD), and k-means clustering on each group of tweets. We calculate the silhouette coefficient to determine the optimal cluster size (k value) for the neutral tweets and the four groups of hateful tweets. We manually select one tweet per cluster using a majority vote from three members of our group to choose representative tweets and create the final set of 40 tweets.

Additional information on the study's variables, pilot study, demographics, as well as example tasks may be found in [chapter F](#).

¹Approved by the ethics committee of our organization.

4.3.4. PROCEDURE AND DATA QUALITY CONTROL

The survey first presents the informed consent policy and excludes participants that do not agree with it. Next, introductory texts are shown to explain the possible machine decisions. In the case of using the ME scale, participants are presented with a warm-up task to estimate different line lengths. Then, the survey asks 40 randomly shuffled question sets regarding the TP, TN, FP, FN, and rejection scenarios (with 8 question sets per scenario). The first question is about whether participants think the post is hateful (yes/no). The second question is whether participants agree or disagree with the decision made by the machine, which may be correct or incorrect, or are neutral towards it. In the case of a non-neutral decision, the survey asks the third question about the degree to which participants agree or disagree with the machine's decisions, using either the ME or 100-level scale, depending on their group. There is no time limit for the survey.

In the middle of the question sets, we use two Instructional Manipulation Checks to determine if the user is paying attention². These attention checks ask participants to select a specific option from multiple choices (e.g., "You must select Orange"). We exclude responses from the participants who fail the attention checks or do not complete all questions. For the ME scale, we discard responses that do not perform well in the line length warm-up task.

4.3.5. ANALYSIS

We first compute the values for the TP, TN, FP, FN, and rejection scenarios using the survey study data. For both scales, we convert disagreement (with the machine decision) ratings to negative values, neutral stances to 0, and agreement ratings to positive values. We apply convergent validity, in which a correlation analysis between different scales (i.e., the ME and 100-level scales) is conducted to determine if they measure the same phenomenon [52]. We expect a medium-large correlation between both scales, meaning that ME responses small in magnitude should correspond to 100-level scale responses small in magnitude and vice versa. Finally, we analyze reliability, which determines whether we can trust our results and achieve consistent outcomes [52]. In our case, we use inter-rater reliability to investigate whether different subjects give approximately the same judgments to the same scenarios and, thus, whether the degree to which hate speech is subjective. It is measured using Krippendorff's alpha, which we calculate using the normalized ME and 100-level values for all scenarios.

4.3.6. RELIABILITY AND VALIDITY

First, for each survey question set, we calculate the median of all responses. This step yields 40 values (eight values per scenario). We use the median since data from both scales are highly skewed. Then, we calculate the mean of the values (V_{TP} , V_{TN} , V_{FP} , V_{FN} , V_r) within each scenario, giving us the final five values for the TP, TN,

²Prolific's Attention and Comprehension Check Policy

	Magnitude Estimation		Bounded Scale	
	α	v	α	v
True Positive	0.07	18.15	0.04	77.00
True Negative	0.10	36.32	0.11	86.31
alse Positive	0.39	-16.69	0.07	-51.00
False Negative	0.92	-28.08	0.14	-62.43
Rejection	-0.31	-4.82	0.07	-16.37
All	0.78	—	0.44	—

Table 4.1.: Krippendorff’s alpha (α) and scenario utilities (v) by scenario for ME and the bounded 0–100 scale (S100).

FP, FN, and rejection cases. The results for both scales can be seen in [table 4.1](#). The total value, V , is calculated at a later point in this section using the different values.

We calculate Krippendorff’s alpha to measure the inter-rater reliability of all scenarios for each scale, as shown in [table 4.1](#). The last row of the table contains the α values for the entire scale, measuring the inter-rater reliability for all answers. We observe that the ME scale has high inter-rater reliability while the 100-level scale is less reliable. Also, participants using the ME scale tend to exhibit higher agreement regarding the FP and FN cases and systematically disagree on the rejected cases. For the 100-level scale, we observe that participants have low agreement on all scenarios.

We analyze the validity of the ME scale by comparing the median normalized magnitude estimates with the median 100-level scores for each question set. [Figure 4.1](#) presents the correlation plot between the two scales. A Shapiro-Wilk test indicates that the data of both scales do not follow a normal distribution ($p < 0.05$). Thus, we use the Spearman and Kendall rank correlation coefficients since these are non-parametric tests. Spearman returned a 0.98 and Kendall a 0.89 correlation between the ME and the 100-level scales ($p < 0.05$). Finally, a Mann-Whitney U test between the ME and 100-level scales gives a large p-value, indicating no statistically significant difference between the two scales.

4.3.7. TOTAL MODEL VALUE DUE TO THRESHOLD

We evaluate the $V(\tau)$ function (i.e., the value at different rejection thresholds) using the values from the survey study obtained using the ME scale. We train three different binary hate speech classification models on the Waseem and Hovy [56] dataset. The used models are Logistic Regression (LR) with Character N-gram [56], a Convolutional Neural Network (CNN) based on Agrawal and Awekar [57], and a DistilBERT transformer [58]. We use Temperature Scaling to calibrate the CNN and the DistilBERT models following the approach from Guo *et al.* [7]. The model predictions are based on two different test datasets: the *seen* dataset and the *unseen* dataset. The *seen* dataset is the test set of Waseem and Hovy [56] and the *unseen* dataset is a test set from a separate but similar source [55]. We use the *unseen* dataset to simulate how the models would perform in a more challenging, realistic use case. Using unseen data that is similar

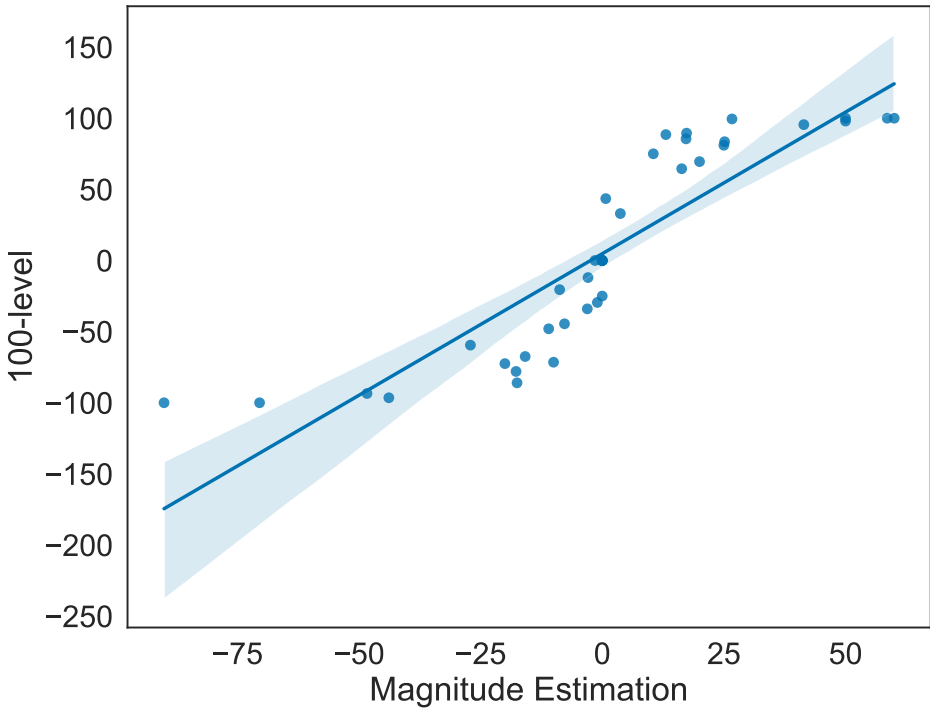
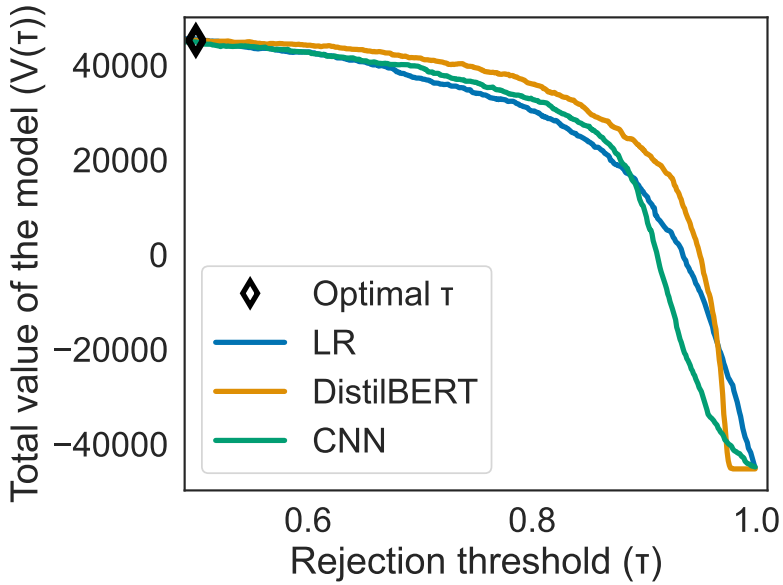


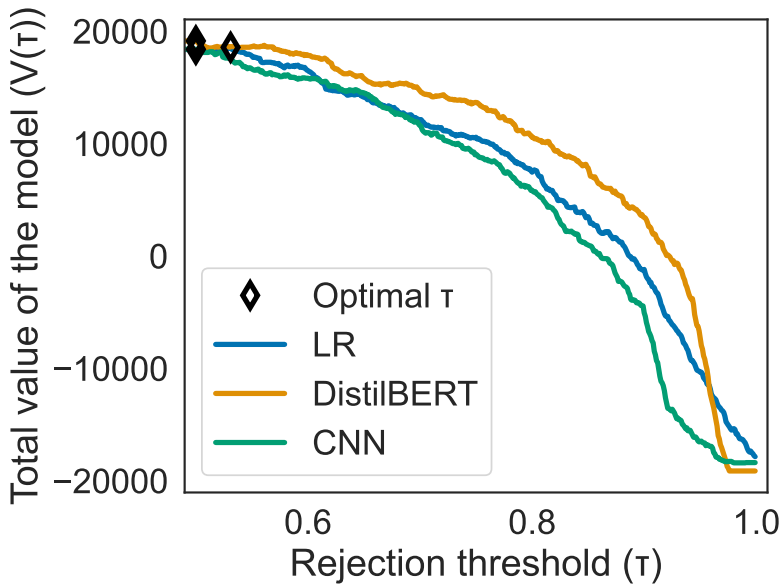
Figure 4.1.: Plot of the convergent validity, showing the correlation between per-item medians from ME and the bounded 0–100 scale. The S-shape reflects floor/ceiling effects in the bounded scale compared to ME.

but separate from the training set, we also investigate the impact of bias. Finally, we calculate the total value as a function of the threshold, $V(\tau)$, for all models with the reject option at all possible rejection thresholds (τ). When $\tau \in [0.0, 0.5]$, all predictions are accepted since the confidence of all predictions is above 0.5 in the case of binary classification. On the other hand, $\tau = 1.0$ implies that all predictions are rejected. We use the v values of the ME scale from [table 4.1](#) to plot the results of all three models in [figures 4.2a](#) and [4.2b](#) using [equation \(4.3\)](#). The diamond-shaped markers indicate the optimal confidence thresholds for rejection at which the model achieves the highest total value.

Participants ascribe higher absolute values to TP and TN scenarios compared to FP and FN ones (see [table 4.1](#)), which results in all but one model having the highest value when all predictions are accepted (see [figures 4.2a](#) and [4.2b](#)). The rejection rates (i.e., the percentage of rejected predictions) and accuracies of accepted predictions at the optimal threshold across the three classifiers can be seen in the first two rows of [table 4.2](#). If we were to take the view that the users' baseline expectation is correct machine decisions, then we can set the value of TP and TN to 0.0 and repeat our analysis to examine how $V(\tau)$ behaves as we consider only punishing incorrect



(a) Total value versus threshold τ on seen data under the full value model. The maximizing τ defines the value-optimized operating point.



(b) Total value versus threshold τ on unseen data under the full value model. Curves inform deployment-time selection without relying on accuracy alone.

Figure 4.2.: Value-Rejection curves across models for seen and unseen data with optimal rejection thresholds indicated by diamonds. Curves inform deployment-time selection without relying on accuracy alone.

	LR			DistilBERT			CNN		
	τ	Acc	RR	τ	Acc	RR	τ	Acc	RR
Seen data	0.500	0.853	0.000	0.500	0.853	0.000	0.500	0.845	0.000
Unseen data	0.531	0.646	0.043	0.500	0.643	0.000	0.500	0.624	0.000
Seen ($V_{TP} = 0, V_{TN} = 0$)	0.829	0.925	0.316	0.786	0.923	0.202	0.815	0.934	0.299
Unseen ($V_{TP} = 0, V_{TN} = 0$)	0.999	0.818	0.991	0.974	1.000	0.996	0.961	0.833	0.980

Table 4.2.: The optimal rejection thresholds (τ), the accuracy of the accepted predictions (Acc), and rejection rates (RR) of all models for both datasets using the values from the survey across different models.

predictions without rewarding correct predictions made by the model (considering the regulation effect discussed in [section 4.2.1](#)). Figures [4.3a](#) and [4.3b](#) demonstrate that the optimal values are achieved at increased rejection thresholds (τ). The last two rows of [table 4.2](#) show that the optimal τ values result in higher accuracies for the *seen* data while rejecting 31.6% of predictions. For the *unseen* data, we achieve high accuracies but reject a large fraction of the predictions.

We also compare the effect of using value and the widely-used accuracy metric in selecting the best model, shown in [table 4.3](#). We observe that both metrics return the same optimal model when correct predictions are rewarded, though there is a difference between *seen* and *unseen* cases. When only incorrect predictions are punished, the optimal models are different as measured by the two metrics: in the case of *seen* data, both LR and DistilBERT perform better than CNN when measured by accuracy, while CNN delivers the highest value; the same observation holds true in the case of *unseen* data – where the optimal model switches from DistilBERT to CNN when we consider the value they deliver instead of accuracy.

4.4. DISCUSSION

4.4.1. VALUE RATIOS, RELIABILITY, AND VALIDITY

Our results show that TP and TN scenarios are highly valued. Participants seem to value correct predictions more than incorrect predictions across all scenarios, regardless of whether they are positive or negative. The value of rejected predictions is the closest to 0 (neutral), as expected, due to them not contributing any benefit or harm, but just delaying the publishing of the post due to the additional human moderation effort. For both scales, we observe the same relation of scenarios in terms of values ($FN < FP < Rejection < TP < TN$). The fact that correct decisions receive higher value ratings indicates strong user appreciation of correct machine decisions. The value of FN having a larger magnitude than the value of FP is noteworthy, as users appear to be more negatively affected when a non-hateful post is subject to moderation than when an instance of hate speech is classified as non-hateful. This implies that users would rather contend with an instance of hate speech than have an innocent user punished for a non-hateful post. This phenomenon may be explained by the Blackstone principle

from the domain of criminal law: “Better that ten guilty persons escape, than that one innocent suffer” [59]. However, we do consider it surprising that the value of TN is greater than the value of TP. One possible reason could be that people disagree more on what is considered hateful among the TP scenarios. We also encountered this phenomenon in the survey results where most people rated TN cases as non-hateful, while for the TP cases there were more disagreements.

Regarding reliability, Krippendorff’s alpha, α , for the 100-level scale being lower than the one for the ME scale is unexpected, as the 100-level scale is bounded with fewer possible options. The stronger agreement for the ME scale indicates that it is indeed suitable for this task. Since α compares the expected difference with the observed difference, it follows that the alpha values for the entire scale should be greater than for the individual scenarios. Generally, participants tend to have low agreement on TP, TN, and rejection cases while they have a high agreement regarding the FP and FN cases. Users tend to agree more regarding what constitutes a misclassified instance than what constitutes a correctly classified instance. For the ME scale, we even observe systematic disagreement for the rejection case, as can be seen by its negative α value. This indicates that users are lower in agreement than one would expect by chance, showing the wide variety of opinions regarding rejection cases by users. By considering all answers, instead of answers for certain scenarios, we observe a greatly increased α , as the observed difference between ratings is closer to the difference expected by chance. For example, participants tend to agree on the classification of a single scenario, e.g. TP, but may give different values on both scales, resulting in lower α for the scenario but greater α across all scenarios. Beyond this, the low reliability for the positive compared to negative predictions indicates that participants disagree on what constitutes hate speech in the first place.

Regarding validity, we observe a strong correlation between scales, demonstrating that the ME scale is validated for measuring people’s opinions about different hate speech detection scenarios. The almost S-shaped curve for the data points in [figure 4.1](#) is due to the lower and upper bounds of the 100-level scale that restrict the participants’ choices, making them more likely to assign the lowest or highest value. Meanwhile, the data points corresponding to the ME scale are skewed towards 0 because of the normalization.

4.4.2. VALUE FUNCTION FOR REJECTION

The purpose of the reject option is to reject predictions where the risk of an incorrect prediction is too high. However, when we use all values obtained from the survey to measure the value function $V(\tau)$, the total value of a model with a reject option is maximized by accepting all predictions. As shown in [figures 4.2a](#) and [4.2b](#), values are positive at the beginning, decline steadily as the rejection threshold increases, and eventually become negative as more predictions are rejected. This observation is not surprising, as the absolute values of correct predictions are greater than the absolute values of incorrect predictions (see [table 4.1](#)).

However, instead of rewarding correct predictions, we believe it is more critical to emphasize penalizing incorrect predictions, as hate speech should be moderated effectively to minimize harm. To study the effects of this we also analyze the behavior of

	LR		DistilBERT		CNN	
	$V(\tau_0)$	Acc	$V(\tau_0)$	Acc	$V(\tau_0)$	Acc
Seen data	45534	0.853	45250	0.853	44893	0.845
Unseen data	18563	0.631	19132	0.643	18385	0.624
Seen data ($V_{TP} = 0, V_{TN} = 0$)	4325	0.853	5172	0.853	5460	0.845
Unseen data ($V_{TP} = 0, V_{TN} = 0$)	4404	0.631	4213	0.643	5291	0.624

Table 4.3.: The total values $V(\tau_0)$ and the accuracies (Acc) of all models. Here, τ_0 is the optimal rejection threshold.

$V(\tau)$ when users do not experience an increase in value through correct classifications, i.e. TP and TN. To achieve this, we set the scenario values v of TP and TN equal to zero. This results in correct predictions effectively only increasing the total value by the v of rejection when accepted and decreasing when rejected, as can be seen in [equation \(4.3\)](#). The result in [figure 4.3a](#) shows a steady increase in value before it peaks for each of the three models, eventually falling again and becoming negative as almost all predictions are rejected. Hence, there is a strong incentive to reject some (but not all) predictions for the *seen* data. At the points where values are maximized, we found an optimal balance between accepting and rejecting predictions. [Figure 4.3b](#) shows that the values continually rise for all three models, only peaking as the rejection threshold approaches 1. This indicates that the model is very uncertain regarding its predictions for the *unseen* data, which may be expected. Initially, at the 0.5 rejection threshold, the value is negative as all predictions are accepted. When the rejection threshold increases, the value rises steadily since too many incorrect predictions are made. This indicates that the model is not performing well at the task (i.e., high confidence false predictions), and thus the optimal condition to reject most predictions makes the unviable model.

The results show that by penalizing incorrect predictions without rewarding correct predictions, a significant fraction of the predictions can be accepted from all three models. For unseen data, however, very few predictions from these models can be accepted and the majority are rejected. Such a result confirms the bias in the dataset as also found in previous studies [27, 28]. The results also show the utility of value as a metric in guiding the decision on when to reject machine predictions. Value utility is further confirmed in the results in [table 4.3](#) from our experiment on optimal model selection: the best model selected by value is different compared to using accuracy as the metric.

4.4.3. FINDINGS, IMPLICATIONS, AND LIMITATIONS

Our survey study uncovers several interesting findings. First, social media users are more appreciative of correct decisions made by the platform, with an absolute magnitude higher than the (negative) perception of incorrect decisions. Among the correct decisions, users especially appreciate that non-hateful content is correctly identified and not banned. On the other hand, users show a much higher agreement on the

negative value of incorrect decisions than correct ones, indicating a strong consensus over the harm (from both identifying hateful content to be non-hateful, and vice versa). These results indicate that while users appreciate correct decisions, minimizing incorrect decisions remains an important task for social media platforms. On the methodological side, we also believe our proposal of using ME for rating human perception can be particularly relevant for research that aims to tackle social science problems through quantitative approaches, like machine learning.

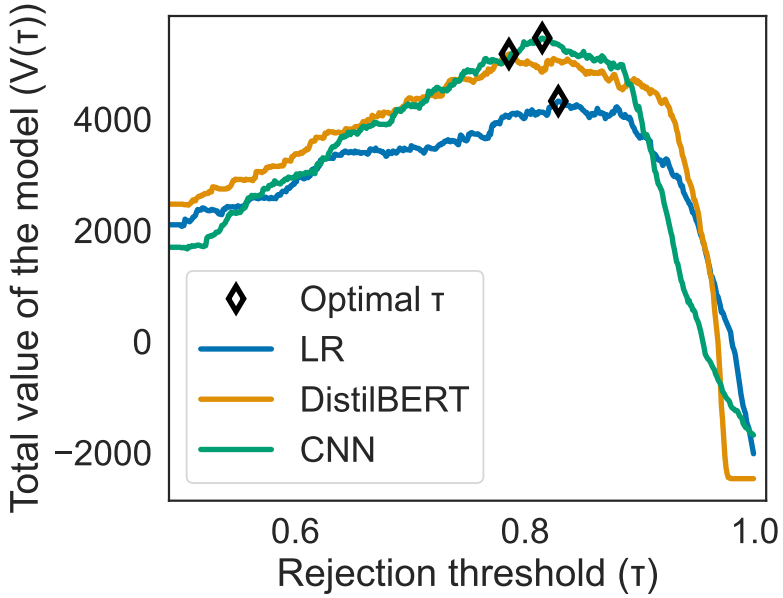
By integrating value as a parameter into the human-AI collaboration framework for rejecting machine decisions, we show that value can help guide the decision of when to accept machine decisions to reach the optimal value a model can deliver. By showing how the number of acceptable machine decisions changes when the model is applied to a dataset different from the training data, our results confirm findings from previous research that such datasets are biased and hence the trained models are as well. Our results also show that when considering value as an optimization target, the best model selected can be different compared with using accuracy as the metric. We believe these findings can benefit the research community and industry alike, as they present a novel way of using a value-sensitive reject option to increase the utility of human-AI collaboration across domains.

Our work is limited to a relatively small sample size (68 subjects per scale). We expect the results to be more reliable at a larger sample size. Besides, optimal confidence threshold determination relies heavily on empirical data, which may not be available in real applications. An easier way for selecting the optimal threshold would be using well-calibrated models, for which the optimal threshold is only dependent on the human-perceived value. Although techniques such as Temperature Scaling can help improve the calibration of existing neural networks or transformer models such as DistilBERT, we still observe that all models are predisposed to producing high-confidence errors. Finally, due to taking the users' standpoint, we do not fully capture the cost of the moderation team being exposed to hate speech. We leave this as possible future work.

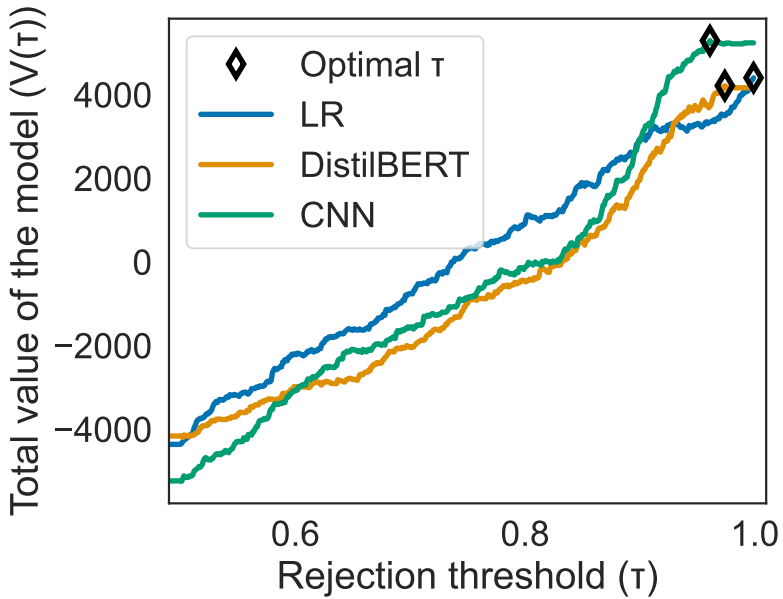
4.5. RELATED WORK

4.5.1. HATE SPEECH DETECTION

Online hate speech content refers to “online messages demeaning people on the basis of their race/ethnicity, gender, national origin, or sexual preference” [60]. Its characterizing features are properties of the target of the language, as compared to other types of online conflictual languages, which are defined by the intention of the author such as cyberbullying or flaming [61, 62]. A large body of discussion can be found on conflictual languages from social sciences, political science, and computer science [63–65]. Hate speech-related research in computer science has identified mismatches between the formalization of hateful content and how people perceive such languages [26]. These mismatches conceptually are further reflected in the technical biases of the machine learning systems used for filtering hateful content. For instance, Gröndahl *et al.* [28] found that F1 scores were reduced by up to 69% when training a hate speech detection model on one dataset and evaluating it using another



(a) Total value versus threshold τ on seen data under the TP/TN=0 variant.



(b) Total value versus threshold τ on unseen data under the TP/TN=0 variant.

Figure 4.3.: Value-Rejection curves across models for the TP/TN=0 variant.

dataset from a similar source. Similarly, Arango, Pérez and Poblete [27] found that most research in hate speech detection overestimates the performance of the automated methods due to dataset bias. In response to these findings, our work aims to explore a human-AI collaborative approach for effective hate speech detection.

4.5.2. HUMAN-AI COLLABORATION AND REJECTION

Human-AI collaboration aims to exploit the complementarity between the cognitive ability of humans and the scalability of machines to solve complex tasks at scale [66, 67]. Some work proposed new ways of collaboration, such as learning crowd vote aggregation models from features of the crowd task [68] and leveraging crowds to learn features of ML models [69, 70]. Recent work has shifted attention to human involvement in providing interpretations of model decisions and evaluating these interpretations [71, 72]. A notable idea for hybrid human-AI decision-making was recently proposed by Law *et al.* [30]: humans are involved after a machine decision is observed to have low confidence. Following works can be categorized in several dimensions, namely *when* rejection happens, on *what models*, and based on *what criteria* [73]. Regarding the “when”, rejection can be implemented in three ways: the preemptive way where whether a data item needs to be handled by a human is decided beforehand [74]; the integrated way which uses a rejector inside the machine learning model (e.g., a rejection layer in a neural network) to decide whether a decision should be rejected [75]; and the dependent way, which is also the most common, which analyzes the rejection option after model decisions [35, 76, 77]. In terms of “what models”, work has been done on rejecting decisions made by a range of models, such as SVMs [74, 77] and different neural networks [75, 76]. In our case, we apply the dependent way to reject models that are based on neural networks. In terms of “what criteria”, Geifman and El-Yaniv [35] proposed a rejection function based on a predefined risk value, an idea also explored in [78]. But unlike ours, their proposals do not consider the impact of machine decisions in a specific context. The most relevant proposal to our work is from Herbei and Wegkamp [76], who studied a confidence metric for determining the optimal rejection threshold. In their work, the threshold is calculated with simulations based on a set of predictions. Going beyond defining cost values from simulations, our approach determines cost values based on users’ perception of machine decisions using a survey study with crowd workers.

4.5.3. VALUE ASSESSMENT AND MEASUREMENT

Value is generally defined as desirable properties of an entity [79]. Specifically for machine learning systems Yurrita *et al.* [80] have identified relevant properties, including individual empowerment, conservation, universalism, and openness. Examples include outlining ethical principles of algorithmic systems [81], developing value-based assessment frameworks [80], and proposing new metrics for evaluating machine learning systems that incorporate value parameters [37]. However, a research gap in measuring value in social contexts has been identified by Olteanu, Talamadupula and Varshney [82], who investigated human-centered metrics for machine learning evaluation in hate speech detection. Their work highlights the gap between accuracy-based

evaluation metrics and user perception. Our work represents a first step towards filling the gap in the context of hate speech detection using ME with a crowdsourced survey.

4.6. CONCLUSIONS

This chapter studies the operationalization and integration of value into human-AI collaboration for hate speech detection. We introduce a value-sensitive rejection mechanism for machine decisions that takes into account the implications of decisions from a user-centered standpoint. We propose ME to measure users' value perception regarding different hate speech detection scenarios. To validate ME, we design a survey study, showing that it can provide a reliable, human-centered assessment of the value a machine learning model delivers. Our survey study uncovers a series of interesting findings on user perception. In particular, participants appreciate correct decisions made by the platform, while they show a strong consensus over the harm of incorrect decisions. Our results show that value assessment performed by means of ME can guide us to select the best confidence threshold for rejecting machine decisions, thereby maximizing model value and potentially leading to a different best model than when using accuracy.

4.6.1. INTEGRATION WITH PILLARS 1 AND 2

The proposed handover mechanism interlocks with the other pillars to form a coordinated robustness pipeline. Pillar 1 (Chapter 2) diagnoses reasoning behavior and failure modes, surfacing patterns (e.g., stylistic replication or brittle chains-of-thought) that predict costly errors under shift. These diagnostics inform *where* value-sensitive thresholds are likely to bind and can prioritize categories for human review even before they surface in incident logs. Pillar 2 (Chapter 3) then targets those recurrent, high-value failure modes via agent-driven synthesis, reducing unknown-unknowns and shifting the error distribution. After any material model change (finetuning or targeted data injection), the value landscape should be re-estimated (or at least thresholds τ re-tuned) on an evaluation set to reflect the new operating regime. This closed loop—diagnose, synthesize, re-optimize thresholds—keeps escalation policies aligned with both stakeholder priorities and the model's evolving capabilities. In this way, Pillars 1 and 2 reduce the *frequency* of costly errors, while Pillar 3 minimizes their *impact* through principled escalation. Together, they instantiate the human-centered robustness perspective articulated in Chapter 1 and surveyed in [2].

4.7. CONCLUSION

This chapter operationalized Pillar 3 of the thesis by replacing confidence-only handover rules with a ratio-scale, stakeholder-centered value signal elicited via Magnitude Estimation (ME). We showed how ME produces reliable and valid judgments that can be normalized and aggregated to obtain per-scenario utilities for correct and incorrect decisions and for rejection. Optimizing accept-reject thresholds against these utilities selects operating points and models that maximize expected total value rather than

accuracy, particularly where confidence alone is a poor proxy for robustness [3, 7]. In the social media moderation setting, value-optimized policies improved total value on both seen and unseen data and recommended stricter acceptance in regimes where costly errors dominate, aligning automation with stakeholder priorities and governance constraints. These findings provide an auditable mechanism for principled escalation that complements model-centric approaches.

Conceptually, the contribution is twofold. First, it decouples *prediction* from *selection*: we retain the base model as a scoring device, but we govern acceptance by explicit utilities reflecting human stakes, not by raw or calibrated confidence. Second, it grounds escalation in a ratio-scale measurement that supports meaningful trade-offs between harms, benefits, and review costs. While post-hoc calibration can mitigate systematic overconfidence, it does not by itself encode stakeholder utilities, and it can degrade under shift; value-sensitive thresholding addresses this gap by optimizing directly for the objective we care about. Complementing the diagnostics from Pillar 1 and the targeted synthesis from Pillar 2, this handover mechanism reduces the *impact* of residual failures by routing high-stakes, ambiguous, or distribution-shifted cases to human experts even when model confidence is high.

REFERENCES

- [1] P. Lammerts, P. Lippmann, Y.-C. Hsu, F. Casati and J. Yang. ‘How Do You Feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection’. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23 (Montréal, QC, Canada). New York, NY, USA: Association for Computing Machinery, 8th Aug. 2023, pp. 834–844. doi: [10.1145/3600211.3604655](https://doi.org/10.1145/3600211.3604655). url: <https://dl.acm.org/doi/10.1145/3600211.3604655>.
- [2] A. Tocchetti, L. Corti, A. Balayn, M. Yurrita, P. Lippmann, M. Brambilla and J. Yang. ‘A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities’. In: *ACM Computing Surveys* 57.6 (10th Feb. 2025), pp. 1–38. issn: 0360-0300. doi: [10.1145/3665926](https://doi.org/10.1145/3665926). url: <https://doi.org/10.1145/3665926>.
- [3] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan and J. Snoek. ‘Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift’. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. url: <https://proceedings.neurips.cc/paper/2019/file/1e0776d291463a1ec8926f3b3f60e8ea-Paper.pdf>.
- [4] S. Sagawa, P. W. Koh, T. B. Hashimoto and P. Liang. ‘Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization’. In: *Proceedings of the Eighth International Conference on Learning Representations*. Online: OpenReview.net, 26th Apr.

2020. url: <https://openreview.net/forum?id=r1xXgCVYr8> (visited on 23/02/2025).
- [5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro and Y. Zhang. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 22nd Mar. 2023. arXiv: [2303.12712](https://arxiv.org/abs/2303.12712) [cs.CL]. url: <https://arxiv.org/abs/2303.12712>.
- [6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei. ‘Language Models are Few-Shot Learners’. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan and H.-T. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. url: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [7] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger. ‘On Calibration of Modern Neural Networks’. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia: PMLR, Aug. 2017, pp. 1321–1330. url: <https://proceedings.mlr.press/v70/guo17a.html>.
- [8] A. Kumar, P. Liang and T. Ma. *Verified Uncertainty Calibration*. NeurIPS 2019. 22nd Sept. 2019. arXiv: [1909.10155](https://arxiv.org/abs/1909.10155) [cs.LG]. url: <https://arxiv.org/abs/1909.10155>.
- [9] R. Parasuraman and V. Riley. ‘Humans and Automation: Use, Misuse, Disuse, Abuse’. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39.2 (June 1997), pp. 230–253. issn: 0018-7208. doi: [10.1518/001872097778543886](https://doi.org/10.1518/001872097778543886). url: <https://doi.org/10.1518/001872097778543886>.
- [10] Z. Buçinca, M. B. Malaya and K. Z. Gajos. ‘To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making’. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (13th Apr. 2021), pp. 1–21. issn: 2573-0142. doi: [10.1145/3449287](https://doi.org/10.1145/3449287). url: <https://doi.org/10.1145/3449287>.
- [11] Y. Geifman and R. El-Yaniv. ‘Selective Classification for Deep Neural Networks’. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett. Vol. 30. Curran Associates, Inc., 2017, pp. 4878–4887. url: https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

- [12] S. S. Stevens. 'The Direct Estimation of Sensory Magnitudes: Loudness'. In: *The American Journal of Psychology* 69.1 (Mar. 1956), pp. 1–25. issn: 0002-9556. doi: [10.2307/1418112](https://doi.org/10.2307/1418112). url: <https://doi.org/10.2307/1418112>.
- [13] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg and D. Amodei. 'Deep Reinforcement Learning from Human Preferences'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett. Vol. 30. Paper URL includes links to PDF and supplemental material. Curran Associates, Inc., 2017, pp. 4299–4307. url: <https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html>.
- [14] P. Lippmann and J. Yang. 'Style over Substance: Distilled Language Models Reason Via Stylistic Replication'. In: *Proceedings of the Second Conference on Language Modeling*. 8th July 2025. url: <https://openreview.net/forum?id=5wAfbEs34A>.
- [15] P. Lippmann, M. T. J. Spaan and J. Yang. 'Illuminating Blind Spots of Language Models with Targeted Agent-in-the-Loop Synthetic Data'. In: *Proceedings of the ORIGen Workshop at the Conference on Language Modeling (COLM)*. Workshop on Optimal Reliance and Accountability in Interactions with Generative LMs. 2025. url: https://origen-workshop.github.io/assets/pdf/2025/papers/1_Illuminating_Blind_Spots_of_.pdf.
- [16] M. Ingram. *Facebook now linked to violence in the Philippines, Libya, Germany, Myanmar, and India*. Columbia Journalism Review. 5th Sept. 2018. url: https://www.cjr.org/the_media_today/facebook-linked-to-violence.php (visited on 23/12/2025).
- [17] M. Mashal. *India's Hate Speech Problem*. New York Times. 2018.
- [18] P. Mozur. *A Genocide Incited on Facebook, With Posts From Myanmar's Military*. The New York Times. 15th Oct. 2018. url: <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html> (visited on 29/12/2025).
- [19] J. Muller. 'Fanning the Flames? Recommender Systems and Hate Speech'. In: *Journal of Online Behavior* (2021).
- [20] E. Commission. *The EU Code of conduct on countering illegal hate speech online*. May 2016. url: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (visited on 20/12/2025).
- [21] N. Giansiracusa. *How Algorithms Create and Prevent Hate Speech*. 2021.
- [22] K. Klonick. 'The New Governors: The People, Rules, and Processes Governing Online Speech'. In: *Harvard Law Review* 131.6 (10th Apr. 2018), pp. 1598–1670. url: <https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/>.

- [23] O. Solon. *Facebook Is Hiring Moderators. But Is the Job Too Gruesome to Handle?* Online; accessed 2025-12-29. The Guardian, 4th May 2017. url: <https://www.theguardian.com/technology/2017/may/04/facebook-content-moderators-ptsd-psychological-dangers> (visited on 29/12/2025).
- [24] P. Fortuna and S. Nunes. 'A Survey on Automatic Detection of Hate Speech in Text'. In: *ACM Computing Surveys* 51.4 (31st July 2018), pp. 1–30. issn: 0360-0300. doi: [10.1145/3232676](https://doi.org/10.1145/3232676). url: <https://doi.org/10.1145/3232676>.
- [25] I. Gagliardone, D. Gal, T. Alves and G. Martinez. *Countering Online Hate Speech*. UNESCO series on internet freedom. UNESCO, 2015. 71 pp. isbn: 978-92-3-100105-5. url: <https://unesdoc.unesco.org/ark:/48223/pf0000233231>.
- [26] A. Balayn, J. Yang, Z. Szlavik and A. Bozzon. 'Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature'. In: *ACM Transactions on Social Computing* 4.3 (30th Sept. 2021), pp. 1–56. issn: 2469-7818. doi: [10.1145/3479158](https://doi.org/10.1145/3479158). url: <https://doi.org/10.1145/3479158>.
- [27] A. Arango, J. Pérez and B. Poblete. 'Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation'. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France: ACM, 18th July 2019, pp. 45–54. doi: [10.1145/3331184.3331262](https://doi.org/10.1145/3331184.3331262). url: <https://doi.org/10.1145/3331184.3331262>.
- [28] T. Gröndahl, L. Pajola, M. Juuti, M. Conti and N. Asokan. 'All You Need is "Love". Evading Hate Speech Detection'. In: *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. Toronto, Canada: ACM, 2018, pp. 2–12. doi: [10.1145/3270101.3270103](https://doi.org/10.1145/3270101.3270103). url: <https://doi.org/10.1145/3270101.3270103>.
- [29] J. Cheng and M. S. Bernstein. 'Flock: Hybrid Crowd-Machine Learning Classifiers'. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. Vancouver, BC, Canada: ACM, 28th Feb. 2015, pp. 600–611. doi: [10.1145/2675133.2675214](https://doi.org/10.1145/2675133.2675214). url: <https://doi.org/10.1145/2675133.2675214>.
- [30] E. H. Law, K. Z. Gajos, J. O. Wobbrock and M. A. Hearst. 'Harmonious crowdsourcing'. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–23.
- [31] M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer and S. Mullainathan. *The Algorithmic Automation Problem: Prediction, Triage, and Human Effort*. 28th Mar. 2019. arXiv: [1903.12220](https://arxiv.org/abs/1903.12220) [cs.CV]. url: <https://arxiv.org/abs/1903.12220>.

- [32] B. Wilder, E. Horvitz and E. Kamar. ‘Learning to Complement Humans’. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 1526–1533. doi: [10.24963/ijcai.2020/212](https://doi.org/10.24963/ijcai.2020/212). url: <https://doi.org/10.24963/ijcai.2020/212>.
- [33] S. Ransbotham, S. Khodabandeh, D. Kiron, F. Candelon, M. Chu and B. LaFountain. *Expanding AI’s Impact With Organizational Learning*. Research report. Findings from the 2020 Artificial Intelligence Global Executive Study and Research Project. MIT Sloan Management Review and Boston Consulting Group, Oct. 2020. url: <https://sloanreview.mit.edu/projects/expanding-ais-impact-with-organizational-learning/>.
- [34] E. R. Balda, A. Behboodi and R. Mathar. ‘Adversarial Examples in Deep Neural Networks: An Overview’. In: *Deep Learning: Algorithms and Applications*. Ed. by W. Pedrycz and S.-M. Chen. Studies in Computational Intelligence. Cham: Springer International Publishing, 24th Oct. 2019, pp. 31–65. isbn: 978-3-030-31760-7. doi: [10.1007/978-3-030-31760-7_2](https://doi.org/10.1007/978-3-030-31760-7_2). url: https://doi.org/10.1007/978-3-030-31760-7_2.
- [35] Y. Geifman and R. El-Yaniv. ‘Selective Classification for Deep Neural Networks’. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett. Vol. 30. Curran Associates, Inc., 2017, pp. 4878–4887. url: https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.
- [36] B. Sayin, J. Yang, A. Passerini and F. Casati. *The Science of Rejection: A Research Area for Human Computation*. 12th Nov. 2021. arXiv: [2111.06736](https://arxiv.org/abs/2111.06736) [cs.LG]. url: <https://arxiv.org/abs/2111.06736> (visited on 23/02/2025).
- [37] F. Casati, P.-A. Noël and J. Yang. *On the Value of ML Models*. 13th Dec. 2021. arXiv: [2112.06775](https://arxiv.org/abs/2112.06775) [cs.LG]. url: <https://arxiv.org/abs/2112.06775>.
- [38] M. L. Cummings. ‘Integrating Ethics in Design through the Value-Sensitive Design Approach’. In: *Science and Engineering Ethics* 12.4 (Dec. 2006), pp. 701–715. issn: 1353-3452. doi: [10.1007/s11948-006-0065-0](https://doi.org/10.1007/s11948-006-0065-0). url: <https://doi.org/10.1007/s11948-006-0065-0>.
- [39] H. Zhu, B. Yu, A. Halfaker and L. Terveen. ‘Value-Sensitive Algorithm Design: Method, Case Study, and Lessons’. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (Nov. 2018). Article 194, pp. 1–23. doi: [10.1145/3274463](https://doi.org/10.1145/3274463). url: <https://doi.org/10.1145/3274463>.
- [40] M. Gilliland. ‘The value added by machine learning approaches in forecasting’. In: *International Journal of Forecasting* 36.1 (Jan. 2020). M4 Competition, pp. 161–166. issn: 0169-2070. doi: [10.1016/j.ijforecast.2019.04.016](https://doi.org/10.1016/j.ijforecast.2019.04.016). url: <https://doi.org/10.1016/j.ijforecast.2019.04.016>.

- [41] S. S. Stevens. 'The Direct Estimation of Sensory Magnitudes: Loudness'. In: *The American Journal of Psychology* 69.1 (Mar. 1956), pp. 1–25. issn: 0002-9556. doi: [10.2307/1418112](https://doi.org/10.2307/1418112). url: <https://www.jstor.org/stable/1418112>.
- [42] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). url: <https://aclanthology.org/N19-1423/>.
- [43] B. Sayin, J. Yang, X. Chen, A. Passerini and F. Casati. 'Rethinking and Recomputing the Value of Machine Learning Models'. In: *Artificial Intelligence Review* 58.8 (8th May 2025), p. 238. issn: 1573-7462. doi: [10.1007/s10462-025-11242-6](https://doi.org/10.1007/s10462-025-11242-6). url: <https://doi.org/10.1007/s10462-025-11242-6> (visited on 19/02/2025).
- [44] M. Mozafari, R. Farahbakhsh and N. Crespi. 'Scaling to very very large corpora for early stage sentiment analysis'. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2018), pp. 532–535.
- [45] J. Boone Harry N. and D. A. Boone. 'Analyzing Likert Data'. In: *Journal of Extension* 50.2, 48 (2012). doi: [10.34068/joe.50.02.48](https://doi.org/10.34068/joe.50.02.48). url: <https://doi.org/10.34068/joe.50.02.48>.
- [46] I. E. Allen and C. A. Seaman. 'Likert Scales and Data Analyses'. In: *Quality Progress* 40.7 (July 2007), pp. 64–65. url: <https://asq.org/quality-progress/articles/statistics-roundtable-likert-scales-and-data-analyses?id=617a8521fc0a4c9aa366fcbb0486bbd5> (visited on 22/12/2025).
- [47] H. R. Moskowitz. 'Magnitude Estimation: Notes on What, How, When, and Why to Use It'. In: *Journal of Food Quality* 1.3 (1977), pp. 195–227. doi: [10.1111/j.1745-4557.1977.tb00942.x](https://doi.org/10.1111/j.1745-4557.1977.tb00942.x). url: <https://doi.org/10.1111/j.1745-4557.1977.tb00942.x>.
- [48] E. Maddalena, C. Tasso and F. Monti. 'Crowdsourcing for sentiment and stance analysis'. In: *IEEE Transactions on Affective Computing* (2017).
- [49] E. G. Bard, D. Robertson and A. Sorace. 'Magnitude Estimation of Linguistic Acceptability'. In: *Language* 72.1 (Mar. 1996), p. 32. issn: 0097-8507. doi: [10.2307/416793](https://doi.org/10.2307/416793). url: <https://doi.org/10.2307/416793>.
- [50] M. McGee. 'Master usability scaling'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2004, pp. 335–342. doi: [10.1145/985692.985735](https://doi.org/10.1145/985692.985735). url: <https://doi.org/10.1145/985692.985735>.

- [51] K. Roitero, E. Maddalena, G. Demartini and S. Mizzaro. ‘On Fine-Grained Relevance Scales’. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. Ann Arbor, Michigan, USA: ACM, 27th June 2018, pp. 675–684. doi: [10.1145/3209978.3210052](https://doi.org/10.1145/3209978.3210052). url: <https://doi.org/10.1145/3209978.3210052>.
- [52] K. Fitzner. ‘Reliability and Validity A Quick Review’. In: *The Diabetes Educator* 33.5 (2007), pp. 775–780. doi: [10.1177/0145721707308172](https://doi.org/10.1177/0145721707308172). url: <https://doi.org/10.1177/0145721707308172>.
- [53] K. Gold. ‘Women and online harassment’. In: *Online Media Journal* (2018).
- [54] H. Müller, A. Sedley and E. Ferrall-Nunge. ‘Survey Research in HCI’. In: *Ways of Knowing in HCI*. Ed. by J. S. Olson and W. A. Kellogg. Springer New York, 2014, pp. 229–266. doi: [10.1007/978-1-4939-0378-8_10](https://doi.org/10.1007/978-1-4939-0378-8_10). url: https://doi.org/10.1007/978-1-4939-0378-8_10.
- [55] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso and M. Sanguinetti. ‘SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter’. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Ed. by J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki and S. M. Mohammad. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. doi: [10.18653/v1/S19-2007](https://aclanthology.org/S19-2007/). url: <https://aclanthology.org/S19-2007/>.
- [56] Z. Waseem and D. Hovy. ‘Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter’. In: *Proceedings of the NAACL Student Research Workshop*. Ed. by J. Andreas, E. Choi and A. Lazaridou. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88–93. doi: [10.18653/v1/N16-2013](https://aclanthology.org/N16-2013/). url: <https://aclanthology.org/N16-2013/>.
- [57] S. Agrawal and A. Awekar. ‘Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms’. In: *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26–29, 2018, Proceedings*. Ed. by G. Pasi, B. Piwowarski, L. Azzopardi and A. Hanbury. Vol. 10772. Lecture Notes in Computer Science. Grenoble, France: Springer, Mar. 2018, pp. 141–153. doi: [10.1007/978-3-319-76941-7_11](https://doi.org/10.1007/978-3-319-76941-7_11). url: https://doi.org/10.1007/978-3-319-76941-7_11.
- [58] V. Sanh, L. Debut, J. Chaumond and T. Wolf. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2nd Oct. 2019. arXiv: [1910.01108](https://arxiv.org/abs/1910.01108) [cs.CL]. url: <https://arxiv.org/abs/1910.01108>.
- [59] D. Epps. ‘The Consequences of Error in Criminal Justice’. In: *Harvard Law Review* 128.4 (2015), pp. 1065–1151. url: <https://harvardlawreview.org/print/vol-128/the-consequences-of-error-in-criminal-justice/>.

- [60] R. J. Lee-Won, T. N. White, H. Song, J. Y. Lee and M. R. Smith. ‘Source magnification of cyberhate: affective and cognitive effects of multiple-source hate messages on target group members’. In: *Media Psychology* 23.5 (17th May 2019), pp. 603–624. issn: 1521-3269. doi: [10.1080/15213269.2019.1612760](https://doi.org/10.1080/15213269.2019.1612760). url: <https://doi.org/10.1080/15213269.2019.1612760>.
- [61] P. M. Burbank *et al.* ‘Cross-cultural comparison of magnitude estimation and category rating scales of lightness’. In: *Journal of Experimental Psychology: Human Perception and Performance* (1994).
- [62] C. Rayner and H. Hoel. ‘A Summary Review of Literature Relating to Workplace Bullying’. In: *Journal of Community & Applied Social Psychology* 7.3 (June 1997), pp. 181–191. doi: [10.1002/\(sici\)1099-1298\(199706\)7:3<181::aid-casp416>3.0.co;2-y](https://doi.org/10.1002/(sici)1099-1298(199706)7:3<181::aid-casp416>3.0.co;2-y). url: [https://doi.org/10.1002/\(sici\)1099-1298\(199706\)7:3%3C181::aid-casp416%3E3.0.co;2-y](https://doi.org/10.1002/(sici)1099-1298(199706)7:3%3C181::aid-casp416%3E3.0.co;2-y).
- [63] J. Matias. ‘Preventing Data Harms’. In: *MIT Press* (2019).
- [64] A. Tsesis. ‘Hate in Cyberspace: Regulating Hate Speech on the Internet’. In: *San Diego Law Review* 38.3 (2001), pp. 817–874. issn: 0036-4037. url: <https://digital.sandiego.edu/sdlr/vol38/iss3/4>.
- [65] J. Waldron. *The Harm in Hate Speech*. Cambridge, Mass: Harvard University Press, 2012. isbn: 9780674065895. url: <https://openlibrary.org/books/OL25110007M>.
- [66] J. W. Vaughan. ‘Tutorial: Making Better Use of the Crowd’. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Ed. by M. Popović and J. Boyd-Graber. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 17–18. url: <https://aclanthology.org/P17-5006/>.
- [67] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro and D. Weld. ‘Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance’. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 6th May 2021, pp. 1–16. doi: [10.1145/3411764.3445717](https://doi.org/10.1145/3411764.3445717). url: <https://doi.org/10.1145/3411764.3445717>.
- [68] E. Kamar, S. Hacker and E. Horvitz. ‘Combining Human and Machine Intelligence in Large-Scale Crowdsourcing’. In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS ’12)*. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 467–474. url: <https://dl.acm.org/doi/10.5555/2343776.2343833>.

- [69] J. Cheng and M. S. Bernstein. ‘Flock: Hybrid Crowd-Machine Learning Classifiers’. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. New York, NY, USA: ACM, 28th Feb. 2015, pp. 600–611. doi: [10.1145/2675133.2675214](https://doi.org/10.1145/2675133.2675214). url: <https://doi.org/10.1145/2675133.2675214>.
- [70] C. Rodríguez, F. Daniel and F. Casati. ‘Crowd-Based Mining of Reusable Process Model Patterns’. In: *Business Process Management. 12th International Conference, BPM 2014, Haifa, Israel, September 7–11, 2014. Proceedings*. Ed. by S. Sadiq, P. Soffer and H. Völzer. Vol. 8659. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 51–66. isbn: 978-3-319-10172-9. doi: [10.1007/978-3-319-10172-9_4](https://doi.org/10.1007/978-3-319-10172-9_4). url: https://doi.org/10.1007/978-3-319-10172-9_4.
- [71] M. K. Lee. ‘Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management’. In: *Big Data & Society* 5.1 (8th Mar. 2018). issn: 2053-9517. doi: [10.1177/2053951718756684](https://doi.org/10.1177/2053951718756684). url: <https://doi.org/10.1177/2053951718756684>.
- [72] M. T. Ribeiro, S. Singh and C. Guestrin. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: ACM, 13th Aug. 2016, pp. 1135–1144. doi: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778). url: <https://doi.org/10.1145/2939672.2939778>.
- [73] K. Hendrickx, L. Perini, D. van der Plas, W. Meert and J. Davis. ‘Machine learning with a reject option: a survey’. In: *Machine Learning* 113.5 (29th Mar. 2024), pp. 3073–3110. issn: 0885-6125. doi: [10.1007/s10994-024-06534-x](https://doi.org/10.1007/s10994-024-06534-x). url: <https://doi.org/10.1007/s10994-024-06534-x>.
- [74] L. Coenen, A. K. A. Abdullah and T. Guns. ‘Probability of Default Estimation, with a Reject Option’. In: *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. Sydney, NSW, Australia: IEEE, Oct. 2020, pp. 439–448. doi: [10.1109/DSAA49011.2020.00058](https://doi.org/10.1109/DSAA49011.2020.00058). url: <https://doi.org/10.1109/DSAA49011.2020.00058>.
- [75] Y. Geifman and R. El-Yaniv. ‘SelectiveNet: A Deep Neural Network with an Integrated Reject Option’. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 2151–2159. url: <https://proceedings.mlr.press/v97/geifman19a.html>.
- [76] R. Herbei and M. Wegkamp. ‘Classification with reject option’. In: *Canadian Journal of Statistics* 34.4 (Dec. 2006), pp. 709–721. issn: 0319-5724. doi: [10.1002/cjs.5550340410](https://doi.org/10.1002/cjs.5550340410). url: <https://doi.org/10.1002/cjs.5550340410>.

- [77] Y. Grandvalet, A. Rakotomamonjy, J. Keshet and S. Canu. ‘Support Vector Machines with a Reject Option’. In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio and L. Bottou. Vol. 21. Curran Associates, Inc., 2008, pp. 537–544. url: https://papers.nips.cc/paper_files/paper/2008/hash/3df1d4b96d8976fff5986393e8767f5b2-Paper.pdf.
- [78] M. Nadeem *et al.* ‘A reject option classifier’. In: *Proceedings of the 26th International Conference on Machine Learning*. 2009.
- [79] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan and M. Bao. ‘The Values Encoded in Machine Learning Research’. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, Republic of Korea: Association for Computing Machinery, June 2022, pp. 173–184. doi: [10.1145/3531146.3533083](https://doi.org/10.1145/3531146.3533083). url: <https://doi.org/10.1145/3531146.3533083>.
- [80] M. Yurrita, D. Murray-Rust, A. Balayn and A. Bozzon. ‘Towards a Multi-Stakeholder Value-Based Assessment Framework for Algorithmic Systems’. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul, Republic of Korea: Association for Computing Machinery, 21st June 2022, pp. 535–547. doi: [10.1145/3531146.3533118](https://doi.org/10.1145/3531146.3533118). url: <https://doi.org/10.1145/3531146.3533118>.
- [81] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy and M. Srikumar. ‘Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI’. In: *SSRN Electronic Journal* (2020). doi: [10.2139/ssrn.3518482](https://www.ssrn.com/abstract=3518482). url: <https://www.ssrn.com/abstract=3518482>.
- [82] A. Olteanu, K. Talamadupula and K. R. Varshney. ‘The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection’. In: *Proceedings of the 2017 ACM on Web Science Conference*. Troy, NY, USA: ACM, 25th June 2017, pp. 405–406. doi: [10.1145/3091478.3098871](https://doi.org/10.1145/3091478.3098871). url: <https://doi.org/10.1145/3091478.3098871>.

5

Conclusion and Future Directions

This chapter consolidates the thesis’s answer on how to improve language model robustness: improved understanding of model behavior, targeted, diagnosis-driven synthetic data that improves reliability under shift, and value-sensitive escalation that governs deployment when uncertainty or unequal error costs remain. We outline our three pillars — structure-aware diagnostics that separate style from substance, targeted knowledge injection via agent-in-the-loop synthesis and synthetic context, and Magnitude Estimation-based decision policies. We revisit the research questions and consolidate evidence across domains and tasks, then derive broader implications for system design, evaluation, and governance. From these implications, we identify concrete gaps that define future research and close with a brief, overarching conclusion.

5.1. SUMMARY OF FINDINGS AND ANSWERS TO RESEARCH QUESTIONS

This thesis contributes to understanding how and why language models can be made robust in practice, and how targeted synthetic data, structure-aware diagnostics, and value-sensitive escalation can be utilized for improved robustness and dependable behavior under distribution shift and uncertainty. The insights from this research can guide both the design of language model training and deployment policies, clarifying when to prioritize stylistic versus substantive reasoning, when to invest in targeted data generation, and when to defer decisions to human expertise based on stakeholder value. Improving robustness of language models allows researchers and practitioners to reduce high-confidence failures, sustain performance across domains and modalities, and operate systems with greater understanding and control. In this concluding chapter, we revisit the central research question and the three specific research questions (Section 1.3) through the lens of the three pillars, synthesize cross-cutting lessons, and outline concrete avenues for future work.

5.1.1. OVERARCHING RESEARCH QUESTION

How can targeted synthetic data generation systematically improve language model robustness?

We advance a methodologically grounded answer to the central robustness question by developing and validating a three-pillar methodology that integrates diagnostics, intervention design, and governance. We find that robustness improves when synthetic data are *diagnosis-driven* and *value-aware*. Pillar 1 provides structure-aware diagnostics that distinguish stylistic replication from substantive reasoning and localize high-confidence failure patterns. Pillar 2 focuses on the bridge between diagnoses and intervention by creating *targeted* synthetic example corpora that reduce blind spots while preserving overall accuracy, as well as compact synthetic proxy contexts. Finally, Pillar 3 replaces confidence-only thresholds with ratio-scale utilities that optimize accept-reject decisions for stakeholder value. Together, these mechanisms provide a holistic approach to improving different facets of language model robustness.

5.1.2. REASONING AND STYLISTIC REPLICATION

RQ1. To what extent do common techniques for transferring reasoning capabilities, such as knowledge distillation, rely on replicating superficial stylistic patterns rather than conveying substantive logical processes?

Pillar 1 establishes what language model reasoning actually transfers in practice and how to diagnose it. We show that much of the benefit from distilling on reasoning traces comes from the target model adopting a particular *stylistic structure*, including metacognitive pivot markers and consistent trace scaffolding, rather than by learning a particular reasoning process. Style-consistent synthetic traces match emergent traces across benchmarks, and even style-consistent traces that lead to incorrect answers yield significant gains over base models. Controlled ablations isolate pivot markers, staging,

and trace length as primary drivers of improvement; correctness still is beneficial, but the supervision signal is dominated by style. These effects generalize across model families and tasks, clarifying that apparent reasoning gains can arise from stylistic replication.

Beyond model post-training diagnostics, we analyze LLM-based agentic behavior and reasoning in interactive settings where memory and interface act as external scaffolds. Positive experience reflection with managed short- and long-term memory improves reliability without any model parameter updates, underscoring the role of interface and memory design for improving language model reasoning. Together, these findings separate internal transfer (style versus substance) from external scaffolding (interfaces and memory). The localized weaknesses they reveal feed directly into Pillar 2’s targeted synthesis.

We show that the transfer of reasoning capabilities is driven primarily by stylistic regularities — such as distinct thinking stages of metacognitive behavior such as pivots, and trace scaffolding — rather than factual correctness of the content. Models trained on synthetic traces, which aim to approximate the distribution of emergent traces and thus replicate these stylistic regularities, achieve accuracy comparable to those trained on the emergent traces they try to emulate across math and logic benchmarks. Furthermore, style-consistent traces that culminate in incorrect answers still lead to measurable improvements over base models. Targeted ablations attribute the majority of gains to pivots, staging, and trace length, with correctness still contributing but not dominating the supervision signal; the pattern generalizes across model families and tasks.

Complementing these findings, agentic experiments show that positive experience reflection with managed memory improves reliability without updating model parameters, underscoring the importance of interface and memory as external scaffolds for reasoning. Taken together, these results argue for structure-aware diagnostics and verification: distinguish style from substance when interpreting gains, preserve helpful structure when it supports generalization, and rely on verifier signals and calibration checks where logical faithfulness matters most. Importantly, these effects do not imply that logical content is unimportant, and gains can vary by task and dataset and depend on teacher and data quality, so calibration and verification evaluations are necessary to avoid overfitting to just stylistic cues.

5.1.3. TARGETED KNOWLEDGE INJECTION

RQ2. How effectively can an agent-driven approach that identifies blind spots and generates targeted synthetic data mitigate specific failure modes compared to untargeted augmentation, and can adaptation be achieved without target-corpus access?

Pillar 2 develops *targeted*, agent-guided knowledge injection that is conditioned on diagnosed weaknesses. An agent-in-the-loop (AIL) workflow uses abstraction and extrapolation over discovered high-confidence errors to hypothesize failure patterns, then generates minimal, pattern-consistent examples that improve model robustness in a targeted manner. Across IMDB, MRPC, and QNLI with BERT-base and Llama 2 7B, our AIL workflow reduces unknown unknowns substantially with overall accuracy changes within only $\pm 1\%$, comparing favorably to reactive relabeling while preserving a

similar level of calibration. The method is agent-agnostic, working with either human or language model agents, and explains model behavior in out-of-distribution settings.

Complementarily, we perform zero-shot contextual adaptation of embeddings by constructing a compact synthetic proxy corpus from only a handful of exemplar documents, enabling context-aware retrieval and inference *without* accessing the target corpus. On the MTEB retrieval benchmark, our approach achieves near-parity with methods with full target-corpus access using ≤ 5 separate exemplars, making it particularly relevant for privacy- and resource-constrained settings. Finally, a multimodal manga translation study shows how visual context disambiguates language and strengthens performance if properly integrated, improving automatic translation evaluation metrics and human evaluation scores. Together, these reactive (AIL) and proactive (contextual adaptation, multimodal context) mechanisms demonstrate that injecting targeted knowledge measurably improves performance across domains and tasks, especially in out-of-distribution settings.

Pillar 2 demonstrates that targeted synthetic data is both effective and sample-efficient: an agent-in-the-loop workflow abstracts patterns from high-confidence errors and synthesizes minimal, targeted examples that locally adjust decision boundaries, reducing unknown unknowns while holding overall accuracy constant and improving model calibration. Under equal data budgets, this targeted approach outperforms reactive relabeling and improves robustness to adversarial perturbations. Complementing reactive mitigation, we show that effective zero-shot contextual adaptation by constructing compact synthetic proxy corpora from a handful of exemplars is possible, reaching near-parity with corpus-conditioned retrieval methods. A multimodal case study in translation further shows that visual context disambiguates language and mitigates context-dependent failure modes, with gains across automatic metrics and human evaluation. Together these results address the research question and support a practical recommendation: to prioritize diagnosis-conditioned synthesis over general augmentation and employ synthetic data to increase available context and thus improve adaptation.

5.1.4. VALUE-SENSITIVE HUMAN ALIGNMENT

RQ3. Can a quantitative measure of user-perceived value provide a more reliable signal than model confidence for determining when to escalate decisions to humans in value-laden contexts?

Pillar 3 operationalizes value-sensitive escalation to align system behavior with stakeholder priorities to better align AI systems. We replace confidence-only thresholds with ratio-scale utilities of user-perceived value elicited via Magnitude Estimation, enabling accept-reject policies that maximize expected total value. Value-optimized policies produce different operating points — and sometimes different model selections — than accuracy- or confidence-based baselines, and improve total value on seen and unseen data. We further show strong inter-rater reliability and convergent validity for the elicited utilities, supporting stable threshold optimization and governance.

Practically, value-sensitive policies complement the model-centric pillars by deferring borderline, high-cost cases while allowing confident, low-cost cases to flow through,

reducing harm without unnecessary rejection. They also provide an auditable mechanism for handover that makes trade-offs explicit and adjustable as contexts change.

Pillar 3 demonstrates that eliciting ratio-scale utilities via magnitude estimation enables accept-reject policies that optimize expected total value and outperform confidence- or accuracy-based baselines on both seen and unseen data. Under a value-based objective, optimal operating points shift—and in some cases the preferred model changes — when harms are asymmetric, this provides a governance-aligned basis for making trade-offs explicit and auditable. Empirically, the value signal exhibits strong inter-rater reliability and convergent validity relative to a bounded rating scale, supporting stable threshold optimization. Class-conditional policies defer borderline, high-cost cases while preserving accuracy where appropriate. Because values are context- and population-dependent, value models should be versioned, periodically re-estimated, and reported alongside calibration metrics so that policy choices remain transparent and accountable over time. Careful study design, normalization procedures, and diverse stakeholder participation are needed to mitigate elicitation bias and avoid entrenching unfair operating points.

5.2. BROADER IMPLICATIONS

This thesis advances a practical answer to the central robustness question by showing that diagnosis-driven synthetic data improve reliability under shift, while value-sensitive escalation effectively governs deployment under uncertainty and asymmetric harms. Three pillars operationalize this answer: (i) structure-aware diagnostics that separate stylistic replication from substantive reasoning, (ii) targeted knowledge injection via agent-in-the-loop synthesis and synthetic context, and (iii) Magnitude Estimation decision policies that optimize expected total value rather than confidence alone.

Across RQ1-RQ3, the evidence is consistent with this view. Pillar 1 shows that much of the apparent transfer of reasoning arises from stylistic replication, and that interface and memory design can improve reliability without updating model parameters. Pillar 2 demonstrates that agent-guided synthesis reduces unknown unknowns with minimal impact on overall accuracy and calibration, while synthetic proxy corpora can approach corpus-aware retrieval without accessing the target corpus, and, in a multimodal case study, visual context strengthens performance when properly integrated. Pillar 3 shows that Magnitude Estimation enables value-sensitive deferral that improves total value over accuracy- or confidence-based baselines when harms are asymmetric.

These findings have broader implications for how language-model systems are designed, evaluated, and governed. Below, we group key implications and relate them to prior work, with short bridges to the future-work agenda in Section 5.3.

System design as control: diagnose → synthesize → defer. Robustness can be operationalized as a control loop: structure-aware diagnostics under distribution shift [1, 2] and uncertainty [3] guide targeted data interventions, while calibrated scores [4, 5] and an explicit reject option [6, 7] or coverage-controlled decisions via conformal prediction [8] set the operating point. The focus on high-confidence blind spots aligns with broader work on unknown unknowns and human-in-the-loop discovery [9–11].

Looking ahead, this motivates the monitored workflow and guarded rollout procedures in Section 5.3.

Evaluation and auditing beyond accuracy. Impacts extend beyond average accuracy: style-sensitivity audits and calibration checks should accompany behavioral test suites and dynamic benchmarks [12–16]. Under shift [1] and uncertainty [3], reporting should include calibration summaries and accept-reject value curves alongside accuracy [4, 5, 17, 18]. Looking ahead, these diagnostics inform where to apply targeted synthesis and how to set value-aware operating points in Section 5.3.

Governance, risk, and value transparency. Value-sensitive deferral clarifies trade-offs in risk-bearing applications and complements broader analyses of LM risks and socio-technical impacts [19, 20] and human-AI decision-making and trust calibration [21, 22]. Eliciting ratio-scale utilities via magnitude estimation [23] and optimizing thresholds using those utilities (Section 5.1.4) improves total value in our experiments [24]. Looking ahead, versioned value models and auditable policy rollouts (Section 5.3) make these choices transparent and accountable.

Unknown unknowns and blind spots. Our agent-in-the-loop synthesis connects to a broader literature on discovering model blind spots and unknown unknowns [9–11, 25–27]. In practice, this favors diagnosis-conditioned, sample-efficient synthesis over untargeted augmentation. Looking ahead, Section 5.3 details holdout-based validation, generator audits, and drift monitors for synthetic proxy corpora.

Tool-augmented and agentic systems. When models reason and act with tools, failures often occur at interfaces among models, memory, and tools. Recent work on reasoning-action traces, tool use, and augmented language models [28–31] motivates logging structured traces and attributing faults across layers, which in turn guides targeted synthesis and tool-grounded verifiers. Looking ahead, we outline fault-aware diagnostics and value-of-information routing in Section 5.3.

Efficiency, privacy, and sustainability. Targeted examples can reduce retraining by addressing localized failures; synthetic proxy corpora enable contextual adaptation without direct access to sensitive corpora; and value-aware thresholds focus human effort where it yields the most benefit. These choices support robustness under shift [1] and uncertainty [3] while containing cost. Looking ahead, the control-loop integration in Section 5.3 operationalizes these benefits.

Finally, there are limitations that researchers and practitioners alike should be mindful of, and that follow directly from the results. Pillar 1 shows that measured reasoning gains can be driven by stylistic cues; thus any trace-based improvement should be paired with verifier-based faithfulness checks and a style-sensitivity audit reported alongside accuracy and calibration. Pillar 2 is bounded by the fidelity of failure hypotheses, generator quality, and exemplar representativeness; synthetic proxy corpora help only insofar as their synthetic statistics approximate the target domain. The same is true for additional multimodal context. These constraints argue for A/B holdouts, generator audits, and drift monitors on proxy corpora. Pillar 3 depends on context- and population-specific utilities and to ensure robustness and responsible use, value models should be versioned, periodically re-elicited, and backtested, with thresholds recalibrated when calibration drifts.

5.3. FUTURE RESEARCH DIRECTIONS

Looking ahead, we outline future research directions that researchers may take to build on this work.

Operational control loop integration. We place the emphasis on consolidating our distinct pillars into a unified approach, threading Pillar 1’s diagnostics, Pillar 2’s targeted knowledge injection, and Pillar 3’s value-sensitive escalation together into a monitored workflow that runs alongside training and deployment. Future work may use structure-aware diagnostics from Pillar 1 to inform model selection, target areas for training, and decide when and how to inject knowledge via Pillar 2, while incorporating Pillar 3 by optimizing value-sensitive accept-reject policies and deferring borderline, high-cost cases. This could be further supported by maintaining versioned models and synthetic context with telemetry and defining guarded rollout procedures, such as: staged deployments with A/B holdouts, pre-defined rollback triggers when behavior deviates from expectation, and automatic diagnosis before attempting an intervention. Closing this loop operationalizes the central thesis: diagnosis-conditioned synthetic knowledge improves reliability, while value-sensitive escalation governs deployment under distribution shift.

From Style Replication to Style-Invariant Reasoning in Distilled Models. Future work should rigorously measure and, where appropriate, reduce style dependence in distilled models. First, it should develop a standardized measurement suite that isolates stylistic cues from reasoning content by constructing counter-styled traces that preserve logical content while randomizing the metacognitive behaviors identified as part of this thesis, and evaluate cross-style generalization and verifier-based faithfulness under controlled perturbations (pivot insertion or removal, trace reordering, template swaps). Second, future work may study mitigation strategies that encourage style invariance, including SFT curricula that gradually remove stylistic hints and contrastive alignment of logically equivalent but stylistically diverse traces. Finally, future work could integrate Pillar 2 by synthesizing content-preserving, style-switching examples targeted at diagnosed high-confidence failures and by stress-testing robustness under distribution shift via a combination of style perturbations, adversarial edits, and out-of-distribution inputs.

Modalities and settings expansion. Future work could extend the framework beyond text while keeping the diagnose, synthesize, escalate loop stable. In diagnostics (Pillar 1), explicitly disentangle modality-specific style from semantic content: for vision-language, separate caption length/templates, punctuation, and composition cues from visual semantics; for speech, distinguish prosody (pitch, speaking rate, pause structure) and transcript formatting from lexical content; for code, separate docstring and formatting conventions from program behavior. In mitigation (Pillar 2), target blind spots via multimodal knowledge injection (e.g., visual disambiguation cues; synthetic context for retrieval) and privacy-preserving proxies (synthetic images, captions, or audio) when direct access to sensitive corpora is not possible or limited. Further, future work may optimize value-aware operating points (Pillar 3), so the benefit of selective deferral is visible and measurable. This expansion tests the generality of the style versus substance distinction, broadens real-world coverage, and sharpens interfaces and safeguards for heterogeneous pipelines.

Tool-augmented and agentic settings. When models act via tools — such as retrieval, calculators, code execution, or external APIs — failures often arise at the interfaces among the model, memory, and tools. Future work could extend diagnostics to attribute errors across layers by logging structured traces — including planned actions, function-call schemas and arguments, tool latency and status, and post-tool reasoning — and by categorizing faults such as retrieval misses, stale or poisoned context, schema violations, timeouts, and misinterpretation of tool outputs. Mitigation can then proceed via Pillar 2 by synthesizing counterfactual retrieval contexts and corrected tool traces to address recurrent faults, injecting adversarial tool outcomes, and training with tool-grounded verifiers. To maximize expected value, deployment can be governed with Pillar 3 by budgeting tool invocations and applying value-of-information routing (where invoking a tool or escalating to a human is done only when the expected value exceeds its cost), escalating high-risk or injection-susceptible actions to human oversight.

5.4. CONCLUDING REMARK

Taken together, this thesis advances a grounded approach to robustness that clarifies how to improve reliability in practice. We show that gains from distillation on reasoning traces are largely driven by stylistic structure (e.g., staged organization and pivot markers) and recommend pairing such methods with faithfulness checks and style-sensitivity audits. We demonstrate that agent-guided, diagnosis-conditioned synthesis reduces high-confidence blind spots while preserving accuracy and calibration, and, separately, that a zero-shot contextual adaptation method can build compact synthetic proxy corpora to provide context when target data are inaccessible. Finally, we elicit ratio-scale utilities through Magnitude Estimation to support value-sensitive deferral, improving total value over confidence-based baselines under asymmetric harms.

Practically, data and policy should stand alongside models: report style sensitivity, calibration, and value curves; audit generators and proxy corpora; version and re-estimate value models; and backtest decisions under drift. In doing so, the work turns robustness into a cumulative, auditable practice—one that explains, adapts, and defers when it should.

REFERENCES

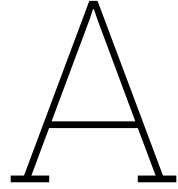
- [1] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn and P. Liang. ‘WILDS: A Benchmark of in-the-Wild Distribution Shifts’. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18th–24th July 2021, pp. 5637–5664. url: <https://proceedings.mlr.press/v139/koh21a.html>.

- [2] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt and J. Gilmer. ‘The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, 2021, pp. 8340–8349. doi: [10.1109/ICCV48922.2021.00823](https://doi.org/10.1109/ICCV48922.2021.00823). url: https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.
- [3] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan and J. Snoek. ‘Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift’. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. url: <https://proceedings.neurips.cc/paper/2019/file/1e0776d291463a1ec8926f3b3f60e8ea-Paper.pdf>.
- [4] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger. ‘On Calibration of Modern Neural Networks’. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia: PMLR, Aug. 2017, pp. 1321–1330. url: <https://proceedings.mlr.press/v70/guo17a.html>.
- [5] S. Desai and G. Durrett. ‘Calibration of Pre-trained Transformers’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 295–302. doi: [10.18653/v1/2020.emnlp-main.21](https://doi.org/10.18653/v1/2020.emnlp-main.21). url: <https://aclanthology.org/2020.emnlp-main.21/>.
- [6] C. K. Chow. ‘On Optimum Recognition Error and Reject Tradeoff’. In: *IEEE Transactions on Information Theory* 16.1 (Jan. 1970), pp. 41–46. issn: 0018-9448. doi: [10.1109/TIT.1970.1054406](https://doi.org/10.1109/TIT.1970.1054406). url: <https://doi.org/10.1109/tit.1970.1054406>.
- [7] Y. Geifman and R. El-Yaniv. ‘SelectiveNet: A Deep Neural Network with an Integrated Reject Option’. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, CA, USA: PMLR, June 2019, pp. 2151–2159. url: <https://proceedings.mlr.press/v97/geifman19a.html>.
- [8] A. N. Angelopoulos and S. Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. July 2021. arXiv: [2107.07511](https://arxiv.org/abs/2107.07511) [cs.LG]. url: <https://arxiv.org/abs/2107.07511>.

- [9] J. Attenberg, P. Ipeirotis and F. Provost. ‘Beat the Machine: Challenging Humans to Find a Predictive Model’s “Unknown Unknowns”’. In: *Journal of Data and Information Quality* 6.1 (4th Mar. 2015), pp. 1–17. issn: 1936-1955. doi: [10.1145/2700832](https://doi.org/10.1145/2700832). url: <https://doi.org/10.1145/2700832>.
- [10] H. Lakkaraju, E. Kamar, R. Caruana and E. Horvitz. ‘Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (13th Feb. 2017), pp. 2124–2132. issn: 2374-3468. doi: [10.1609/aaai.v31i1.10821](https://ojs.aaai.org/index.php/AAAI/article/view/10821). url: <https://ojs.aaai.org/index.php/AAAI/article/view/10821>.
- [11] G. Bansal and D. Weld. ‘A Coverage-Based Utility Model for Identifying Unknown Unknowns’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (25th Apr. 2018). issn: 2374-3468. doi: [10.1609/aaai.v32i1.11493](https://ojs.aaai.org/index.php/AAAI/article/view/11493). url: <https://ojs.aaai.org/index.php/AAAI/article/view/11493>.
- [12] M. T. Ribeiro, T. Wu, C. Guestrin and S. Singh. ‘Beyond Accuracy: Behavioral Testing of NLP Models with CheckList’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter and J. Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4902–4912. doi: [10.18653/v1/2020.acl-main.442](https://aclanthology.org/2020.acl-main.442/). url: <https://aclanthology.org/2020.acl-main.442/>.
- [13] M. T. Ribeiro, S. Singh and C. Guestrin. ‘Semantically Equivalent Adversarial Rules for Debugging NLP Models’. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 856–865. doi: [10.18653/v1/P18-1079](https://aclanthology.org/P18-1079/). url: <https://aclanthology.org/P18-1079/>.
- [14] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi and Y. Choi. ‘HellaSwag: Can a Machine Really Finish Your Sentence?’ In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum and L. Márquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4791–4800. doi: [10.18653/v1/P19-1472](https://aclanthology.org/P19-1472/). url: <https://aclanthology.org/P19-1472/>.
- [15] E. Wallace, S. Feng, N. Kandpal, M. Gardner and S. Singh. ‘Universal Adversarial Triggers for Attacking and Analyzing NLP’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng and X. Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2153–2162. doi: [10.18653/v1/D19-1221](https://aclanthology.org/D19-1221/). url: <https://aclanthology.org/D19-1221/>.

- [16] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts and A. Williams. ‘Dynabench: Rethinking Benchmarking in NLP’. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty and Y. Zhou. Online: Association for Computational Linguistics, June 2021, pp. 4110–4124. doi: [10.18653/v1/2021.naacl-main.324](https://doi.org/10.18653/v1/2021.naacl-main.324). url: <https://aclanthology.org/2021.naacl-main.324/>.
- [17] J. Nixon, M. Dusenberry, G. Jerfel, T. Nguyen, J. Liu, L. Zhang and D. Tran. *Measuring Calibration in Deep Learning*. arXiv:1904.01685v2, 2020-08-07. 2nd Apr. 2019. arXiv: [1904.01685](https://arxiv.org/abs/1904.01685) [cs.LG]. url: <https://arxiv.org/abs/1904.01685>.
- [18] A. Kumar, P. Liang and T. Ma. *Verified Uncertainty Calibration*. NeurIPS 2019. 22nd Sept. 2019. arXiv: [1909.10155](https://arxiv.org/abs/1909.10155) [cs.LG]. url: <https://arxiv.org/abs/1909.10155>.
- [19] E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell. ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?’ In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, Mar. 2021, pp. 610–623. doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922). url: <https://doi.org/10.1145/3442188.3445922>.
- [20] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving and I. Gabriel. *Ethical and Social Risks of Harm from Language Models*. 8th Dec. 2021. arXiv: [2112.04359](https://arxiv.org/abs/2112.04359) [cs.CL]. url: <https://arxiv.org/abs/2112.04359> (visited on 25/02/2025).
- [21] Y. Zhang, Q. V. Liao and R. K. E. Bellamy. ‘Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making’. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain: Association for Computing Machinery, 27th Jan. 2020, pp. 295–305. doi: [10.1145/3351095.3372852](https://doi.org/10.1145/3351095.3372852). url: <https://doi.org/10.1145/3351095.3372852>.
- [22] Z. Buçinca, M. B. Malaya and K. Z. Gajos. ‘To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making’. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (13th Apr. 2021), pp. 1–21. issn: 2573-0142. doi: [10.1145/3449287](https://doi.org/10.1145/3449287). url: <https://doi.org/10.1145/3449287>.
- [23] S. S. Stevens. ‘The Direct Estimation of Sensory Magnitudes: Loudness’. In: *The American Journal of Psychology* 69.1 (Mar. 1956), pp. 1–25. issn: 0002-9556. doi: [10.2307/1418112](https://doi.org/10.2307/1418112). url: <https://doi.org/10.2307/1418112>.

- [24] P. Lammerts, P. Lippmann, Y.-C. Hsu, F. Casati and J. Yang. ‘How Do You Feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection’. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23 (Montréal, QC, Canada). New York, NY, USA: Association for Computing Machinery, 8th Aug. 2023, pp. 834–844. doi: [10.1145/3600211.3604655](https://doi.org/10.1145/3600211.3604655). url: <https://dl.acm.org/doi/10.1145/3600211.3604655>.
- [25] Á. A. Cabrera, A. J. Druck, J. I. Hong and A. Perer. ‘Discovering and Validating AI Errors With Crowdsourced Failure Reports’. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021). Article 425, pp. 1–22. issn: 2573-0142. doi: [10.1145/3479569](https://doi.org/10.1145/3479569). url: <https://doi.org/10.1145/3479569>.
- [26] L. Han, X. Dong and G. Demartini. ‘Iterative Human-in-the-Loop Discovery of Unknown Unknowns in Image Datasets’. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9.1 (Nov. 2021), pp. 72–83. doi: [10.1609/hcomp.v9i1.18941](https://doi.org/10.1609/hcomp.v9i1.18941). url: <https://ojs.aaai.org/index.php/HCOMP/article/view/18941>.
- [27] C. Vandenhof. ‘A Hybrid Approach to Identifying Unknown Unknowns of Predictive Models’. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (28th Oct. 2019), pp. 180–187. doi: [10.1609/hcomp.v7i1.5274](https://doi.org/10.1609/hcomp.v7i1.5274). url: <https://ojs.aaai.org/index.php/HCOMP/article/view/5274>.
- [28] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan and Y. Cao. ‘ReAct: Synergizing Reasoning and Acting in Language Models’. In: *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview.net, 1st May 2023. url: https://openreview.net/forum?id=WE_vluYUL-X (visited on 22/12/2025).
- [29] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda and T. Scialom. ‘Toolformer: Language Models Can Teach Themselves to Use Tools’. In: *Advances in Neural Information Processing Systems*. 9th Feb. 2023. url: <https://arxiv.org/abs/2302.04761> (visited on 23/02/2025).
- [30] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan and S. Yao. *Reflexion: Language Agents with Verbal Reinforcement Learning*. 20th Mar. 2023. arXiv: [2303.11366](https://arxiv.org/abs/2303.11366) [cs.AI]. url: <https://arxiv.org/abs/2303.11366> (visited on 23/02/2025).
- [31] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun and T. Scialom. *Augmented Language Models: a Survey*. arXiv preprint. 15th Feb. 2023. arXiv: [2302.07842](https://arxiv.org/abs/2302.07842) [cs.CL]. url: <https://arxiv.org/abs/2302.07842>.



Appendix for Chapter 2: Reasoning Traces and Analysis

This appendix complements Chapter 2 with expanded trace examples, analysis details, and dataset statistics.

A.1. EMERGENT REASONING TRACE EXAMPLE

We include an illustrative reasoning trace (see [figure A.1](#)) that demonstrates the staged structure and metacognitive pivots discussed in Chapter 2.

A.2. EMERGENT TRACE ANALYSIS

A.2.1. DETAILS OF CATEGORIZING THE REASONING TRACE CONTENTS

To analyze reasoning traces, we developed a systematic approach for identifying pivots and reasoning stages using regular expression pattern matching. For each of the four pivot categories (Realization, Verification, Exploration, and Integration), we created regex patterns capturing lexical markers that signal metacognitive transitions. Similarly, we defined patterns for the four reasoning stages (Problem Framing, Exploration, Verification, and Synthesis) based on characteristic expressions and structural elements. While we experimented with using an LM for this tagging task, we found it offered no clear advantage over regex for these standardized traces.

This framework enabled automated extraction and quantification of reasoning elements across all traces. Although regex matching is an approximation, it performed well given the standardized nature of the traces. Patterns were refined iteratively after manual inspection to ensure balanced detection across categories. Each trace was analyzed for both pivot frequency and the presence of reasoning stages, allowing quantification of metacognitive transitions and structural patterns.

Reasoning Trace Example

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Reasoning: Okay, let's see. The problem says a juggler can juggle 16 balls. Half of them are golf balls, and half of the golf balls are blue. I need to find out how many blue golf balls there are. Alright, let me break this down step by step. First, the total number of balls is 16. Half of those are golf balls. So, half of 16 would be... 16 divided by 2, which is 8. So, there are 8 golf balls. Got that down. Now, out of these 8 golf balls, half are blue. Hmm, so again, half of 8 would be 8 divided by 2. Let me calculate that. 8 divided by 2 is 4. So, does that mean there are 4 blue golf balls? That seems right, but let me double-check to make sure I didn't skip any steps or misunderstand the question.

Wait, the problem says half of the balls are golf balls. So total balls: 16. Half are golf balls: 8. Then half of the golf balls are blue. So half of 8 is indeed 4. So, 4 blue golf balls. Let me just visualize this. If there are 16 balls in total, and half (so 8) are golf balls, and of those 8 golf balls, half are blue and half are another color. So splitting 8 into two equal parts gives 4 and 4. So yes, 4 blue golf balls. That makes sense.

Is there another way to approach this problem? Maybe by using fractions. Let's see. Total balls = 16. Half are golf balls: so $1/2$ of 16. Then half of those are blue: so $1/2$ of $1/2$, which is $1/4$. So $1/4$ of the total balls are blue golf balls. $1/4$ of 16 is 4. Yep, same answer. So whether I do it step by step or use fractions, I still get 4. That seems consistent.

Alternatively, if I were to write equations, let's denote total balls as $T = 16$. Number of golf balls $G = T / 2$. Then number of blue golf balls $B = G / 2$. Substituting, $B = (16 / 2) / 2 = 8 / 2 = 4$. Same result. So all methods lead to 4.

I think that's solid. Let me just check if I interpreted "half of the golf balls are blue" correctly. The problem didn't mention anything about the other half, but since half of the golf balls are blue, the other half could be any other color, but that's irrelevant to the question. The question only asks for the number of blue golf balls, so 4 is the answer.

Wait, just to be thorough, the problem doesn't have any trick wording, right? Like, when it says half of the balls are golf balls, that's straightforward. Then half of those golf balls are blue. So no, no tricky parts. If it had said half of the blue balls are golf balls, that would be different, but it's the other way around. Half of the golf balls are blue. So straightforward.

Another check: total golf balls are 8. If 4 are blue, then 4 are another color. That still keeps the total at 8 golf balls out of 16. Everything stays consistent. So the final answer is 4 blue golf balls.

Figure A.1.: An example emergent reasoning trace illustrating staged structure and metacognitive pivots.

A.2.2. PIVOT ANALYSIS

Pivot Diversity A key indicator of complex reasoning is the variety of metacognitive shifts employed. The analyzed traces show an average diversity of 3.51 distinct pivot types per trace, and 96.1% of traces contain at least three different pivot categories.

Pivot Type Frequencies The frequency and prevalence of each pivot type are summarized in Table A.1. Integration and Realization pivots are nearly universal; Exploration pivots are least prevalent at 87.0% of traces.

A.2.3. REASONING STAGE ANALYSIS

We also quantify segments corresponding to the four cognitive reasoning stages discussed in section 2.2.3. A single reasoning stage can contain multiple pivots of different types; we ascribe the trace to the stage found at the beginning of the segment. The average occurrences and prevalence of these stages are shown in Table A.2. Synthesis is universally present; Problem Framing, Verification, and Exploration are all highly prevalent.

A.2.4. KEY INSIGHTS FROM ANALYSIS

The quantitative analysis reinforces the qualitative observations in Chapter 2: successful emergent reasoning traces utilize diverse pivot types (avg. 3.51), and all four reasoning stages are highly prevalent, underscoring the iterative, staged nature of effective traces.

Pivot Type	Avg. Occurrences per Trace	% Traces Present
Realization	18.96	98.6%
Exploration	16.11	87.0%
Verification	1.37	89.6%
Integration	67.64	100.0%

Table A.1.: Frequency and prevalence of identified pivot types within analyzed emergent reasoning traces.

Reasoning Stage	Avg. Occurrences per Trace	% Traces Present
Problem Framing	3.13	79.2%
Exploration	6.81	87.0%
Verification	3.34	89.6%
Synthesis	84.08	100.0%

Table A.2.: Frequency and prevalence of identified reasoning stages within analyzed emergent reasoning traces.

A.3. DATASET STATISTICS

This section details the datasets used for Part I: the seed data pool and the generated reasoning trace datasets SmolTraces (ST) and SmolTraces-HardCoded (ST-HC).

A.3.1. SEED DATA COMPILATION

We curated question-answer pairs from multiple domains to build seed data, selecting problems from OlympicArena [1], AGIEval [2], LiveCodeBench v4 [3], NuminaMATH [4], and OmniMath [5]. This yielded 31,586 unique QA pairs after decontaminating overlaps with evaluation benchmarks (MATH500, AIME2024, GPQA Diamond).

A.3.2. GENERATED REASONING TRACE DATASETS

Using the seed data, we generated two parallel datasets with detailed reasoning traces as described in section 2.2.3. ST contains emergent traces produced by R1 [6]. ST-HC contains synthetic traces created with GPT-4o guided by the structured prompt described in the text. Both used up to five zero-shot attempts per seed question, keeping the first correct trace.

For a fair comparison, we balanced dataset sizes by downsampling the larger set to match the smaller one, yielding two equal-sized datasets (N=18,242 each). Summary statistics for the final datasets are provided in Table A.4.

A.4. SYNTHETIC TRACES WITH WRONG ANSWER DETAILS

We construct ST-HC-W to isolate stylistic effects independent of final answer correctness for the ablation in section 2.2.4. Starting from ST-HC, we use the known correct answer and prompt a weaker LM (GPT-4o-mini) to produce a different, incorrect answer with a similar format (e.g., “9.11” instead of “9.9”). If the generated answer matches the correct one, we resample. The resulting ST-HC-W mirrors the size of ST-HC (N=18,242) and retains trace style while ending with an incorrect final answer.

Data Source	Number of Samples
OlympicArena	4,250
AGIEval	2,385
LiveCodeBench v4	713
NuminaMATH (subset)	20,000
OmniMath	4,238
Total	31,586

Table A.3.: Composition of the seed data pool used for generating reasoning traces.

Statistic	Dataset Values
Trace Origin	ST: Emergent (R1); ST-HC: Synthetic (GPT-4o + structured prompt)
Seed Questions Source	Shared pool (Table A.3, N=31,586)
Final Number of Samples	ST: 18,242; ST-HC: 18,242
Filtering Applied	Correct answer (up to 5 attempts); min. trace length (50 tokens)
Avg. Trace Length (Tokens)	ST: 2,521; ST-HC: 2,101
Avg. Pivots per Trace	ST: 93.4; ST-HC: 89.1

Table A.4.: Summary statistics for the final generated reasoning trace datasets used in fine-tuning.

REFERENCES

- [1] Z. Huang, Z. Wang, S. Xia, X. Li, H. Zou, R. Xu, R.-Z. Fan, L. Ye, E. Chern, Y. Ye, Y. Zhang, Y. Yang, T. Wu, B. Wang, S. Sun, Y. Xiao, Y. Li, F. Zhou, S. Chern, Y. Qin, Y. Ma, J. Su, Y. Liu, Y. Zheng, S. Zhang, D. Lin, Y. Qiao and P. Liu. *OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI*. 18th June 2024. arXiv: [2406.12753](https://arxiv.org/abs/2406.12753) [cs.CL]. url: <https://arxiv.org/abs/2406.12753>.
- [2] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen and N. Duan. *AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models*. arXiv:2304.06364v2. 13th Apr. 2023. arXiv: [2304.06364](https://arxiv.org/abs/2304.06364) [cs.CL]. url: <https://arxiv.org/abs/2304.06364>.
- [3] N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen and I. Stoica. *LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code*. 6th June 2024. arXiv: [2403.07974](https://arxiv.org/abs/2403.07974) [cs.SE]. url: <https://arxiv.org/abs/2403.07974>.
- [4] J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. C. Huang, K. Rasul, L. Yu, A. Jiang, Z. Shen, Z. Qin, B. Dong, L. Zhou, Y. Fleureau, G. Lample and S. Polu. *NuminaMath-CoT*. Hugging Face dataset card describing the 860k-problem NuminaMath CoT corpus. AI-MO / Numina. 15th July 2024. url: <https://huggingface.co/datasets/AI-MO/NuminaMath-CoT> (visited on 11/12/2025).
- [5] B. Gao, F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, Z. Tang, B. Wang, D. Zan, S. Quan, G. Zhang, L. Sha, Y. Zhang, X. Ren, T. Liu and B. Chang. *Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models*. 10th Oct. 2024. arXiv: [2410.07985](https://arxiv.org/abs/2410.07985) [cs.CL]. url: <https://arxiv.org/abs/2410.07985>.

- [6] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 22nd Jan. 2025. arXiv: 2501.12948 [cs.CL]. url: <https://arxiv.org/abs/2501.12948>.

B

Appendix for Chapter 2: Sweet&Sour Formal Background and Examples

B.1. FORMAL BACKGROUND FOR PART II: POMDP FRAMING OF TEXT-BASED GAMES

We formalize interactive text environments such as ScienceWorld and ALFWorld as partially observable Markov decision processes (POMDPs) [1] with tuple $\langle S, T, A, \Omega, R, \gamma \rangle$. Here, S denotes latent states, A the (textual) action space, $T(s' | s, a)$ the transition function, $\Omega(o | s', a)$ the observation function, $R(s, a)$ the reward function, and $\gamma \in [0, 1]$ a discount factor. At time t , the agent receives observation $o_t \sim \Omega(\cdot | s_t, a_{t-1})$, selects an action $a_t \in A$, and receives a reward update r_t with next state $s_{t+1} \sim T(\cdot | s_t, a_t)$. Both observations and actions are sequences of tokens. Episodes terminate on goal satisfaction or a step cap ($T = 150$ in our experiments). In ScienceWorld, the agent additionally maintains an inventory i_t for manipulable objects.

This formalization is the basis for the action–observation–reflection loop in [section 2.4.3](#), where the policy conditions not only on the current observation but also on managed memory.

B.2. QUALITATIVE EXAMPLES FOR SWEET&SOUR

We include compact transcripts illustrating the reflection loop and managed memory effects. These examples show how positive reflections enter STM immediately and how managed memory maintains succinct takeaways across episodes, stabilizing behavior after early wins and mitigating tilt.

ScienceWorld Task 8-1**Task Description:**

Your task is to find a(n) animal. First, focus on the thing. Then, move it to the red box in the kitchen.

> look around

This room is called the hallway. In it, you see:

- a substance called air
- a picture
- the agent

You also see:

- A door to the greenhouse (open)
- A door to the living room (open)
- A door to the art studio (open)
- A door to the kitchen (open)
- A door to the bedroom (open)
- A door to the workshop (open)

> open door to greenhouse

The door is already open.

> go to greenhouse

You move to the greenhouse.

> open door to outside

The door is already open.

> go to outside

You move to the outside.

> look around

This outside location is called the outside. Here, you see:

- a substance called air
- an axe
- a substance called wood
- the ground
- a fire pit (containing nothing)
- a fountain (containing a substance called water)

- a butterfly egg
- a blue jay egg
- a dove egg
- the agent

You also see:

- A door to the kitchen (open)
- A door to the foundry (open)
- A door to the greenhouse (open)

Case #1: Failure

> focus on axe

You focus on the axe.

> pick up axe

You move the axe to the inventory.

Reflection (I failed because I assumed eggs are not animals. My plan was to go outside to find animals. I tried to pick up an axe. An alternative course of action would have been to go outside and pick up an egg, as it might be considered an animal.)

→ Managed Memory

Case #2: Success

> focus on dove egg

You focus on the dove egg.

> pick up dove egg

You move the dove egg to the inventory.

Reflection (Picking up an egg was the correct choice. Even though the animal was in an egg, it is still considered an animal. My plan was to go outside and find an animal in the wild. Items such as an axe, wood, or landscape elements are not related to animals.)

→ Managed Memory

Figure B.1.: Example of reflection in a find-animal task (ScienceWorld).

ScienceWorld Task 1-1**Task Description:**

Your task is to boil water. For compounds without a boiling point, combusting the substance is also acceptable. First, focus on the substance. Then, take actions that will cause it to change its state of matter.

> look around

This room is called the hallway. In it, you see:

- a picture
- a substance called air
- the agent
- A door to the green house (open)
- A door to the living room (open)
- A door to the art studio (open)
- A door to the kitchen (open)
- A door to the bedroom (open)
- A door to the workshop (open)

> open door to kitchen

The door is already open.

> go to kitchen

You move to the kitchen.

> look around

This room is called the kitchen. In it, you see:

- a substance called soap
- a painting
- a thermometer reading 10°C
- a counter with various items (banana, potato, red apple, orange)
- a sink (off, empty)
- a table with a glass cup (empty)
- a chair (empty)
- a freezer (closed)
- a lighter

- a stopwatch (deactivated)
- a fridge (closed)
- a cupboard (closed)
- a stove (turned off)
- a glass jar containing sodium chloride
- various other items (air, the agent, etc.)

> pick up thermometer

You move the thermometer to the inventory.

> open cupboard

The cupboard is now open.

> pick up metal pot

You move the metal pot to the inventory.

> move metal pot to sink

You move the metal pot to the sink.

> activate sink

The sink is now activated.

> deactivate sink

The sink is now deactivated.

> pick up metal pot

You move the metal pot to the inventory.

> focus on substance in metal pot

You focus on the water.

> move metal pot to stove

You move the metal pot to the stove.

> activate stove

The stove is now activated.

> use thermometer in inventory on substance in metal pot

The thermometer measures a temperature of 13 degrees Celsius.

> examine substance in metal pot

A substance called water.

> use thermometer in inventory on substance in metal pot

The thermometer measures a temperature of 28 degrees Celsius.

> examine substance in metal pot

A substance called water.

> use thermometer in inventory on substance in metal pot

The thermometer measures a temperature of 58 degrees Celsius.

> examine substance in metal pot

A substance called water.

```

> use thermometer in inventory on substance in metal
pot
The thermometer measures a temperature of 78 degrees Celsius.
> examine substance in metal pot
A substance called water.
> use thermometer in inventory on substance in metal
pot
The thermometer measures a temperature of 98 degrees Celsius.
> examine steam
A substance called steam. Task completed.

```

Figure B.2.: Successful sequence of events in a boil-water task (ScienceWorld).

B.3. ADDITIONAL NOTES: ALFWORLD

For completeness, we note that ALFWORLD follows similar agent and evaluation settings to ScienceWorld but features shorter horizons and simpler object-interaction graphs. We report compact success rates in Table B.1; hyperparameters, memory budgets, and reflection triggers mirror those used in ScienceWorld unless stated otherwise.

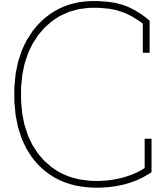
Method / Model	Llama 8B	Mistral Large 2	GPT-4o
ReAct	8.0	41.0	77.0
Reflexion	13.0	61.0	91.0
Sweet&Sour (ours)	20.0	78.0	97.0

Table B.1.: Performance on ALFWORLD (success rates).

ALFWORLD results table. Detailed success rates across models are shown in Table B.1.

REFERENCES

- [1] M. T. J. Spaan. ‘Partially Observable Markov Decision Processes’. In: *Reinforcement Learning*. Ed. by M. Wiering and M. van Otterlo. Adaptation, Learning, and Optimization. Springer Berlin Heidelberg, 2012, pp. 387–414. isbn: 978-3-642-27645-3. doi: [10.1007/978-3-642-27645-3_12](https://doi.org/10.1007/978-3-642-27645-3_12). url: https://doi.org/10.1007/978-3-642-27645-3_12.



Appendix for Chapter 3 (Part I): Agent-in-the-Loop (AIL) Artifacts

C.1. SYNTHETIC BLIND SPOTS

We use the synthetic blind spot study akin to a sanity check for our approach. As such, compared to the full natural blind spot study, we use a only a single task, a simpler model architecture, and make other simplifications to our mitigation process. We select an LSTM [1] as our model of choice due to the absence of pretraining and apply the TF perturbation method on the SA task. The LSTM used is the standard version of the Bi-LSTM provided by Morris *et al.* [2].

	Original			Retrain		
	Accuracy (%)	Perturbation (%)	UUs (#)	Accuracy (%)	Perturbation (%)	UUs (#)
Clean	88.03	82.22	1725	88.03	82.21	784
Biased R	78.55	78.56	3785	78.61	78.58	2593
Biased P	75.10	75.02	4607	74.25	73.12	1201
Biased N	76.64	76.64	4394	77.38	77.35	845
Biased PN	74.17	73.94	9231	74.81	74.01	2331

Table C.1.: Results of synthetic blind spot study for accuracy, perturbation success rate, and number of UUs before and after retraining for all LSTM model variants. The used perturbation method is TF and the dataset is IMDB.

C.1.1. BLINDSPOT CREATION AND MITIGATION

To assess whether our method can tackle existing synthetic blind spots we perform a type of Controlled Synthetic Data Check [3]. We create synthetic blind spots by systematically

excluding some data from training that have commonalities, namely containing a positive or negative term according to lexica by Liu, Hu and Cheng [4]. Here, we randomly subsample 600 of each as our selection of positive and negative terms, due to the extensive nature of the lexica.

We create a false positive blind spot by removing samples from the train set using our selection of negative terms, resulting in a *negatively biased* LSTM (N). Similarly, we create a false negative blind spot, resulting in a *positively biased* LSTM (P), as well as a blind spot resulting from a selection of 50% randomly chosen terms from each, leading to a *positive/negative biased* LSTM (PN). For comparison, we also include a *randomly biased* LSTM (R), where samples were removed from the train set randomly to obtain a size comparable to the P, N, and PN ones.¹

After creating the synthetic blind spots through biasing, the authors perform the generalization procedure and provide handcrafted hypotheses that precisely describe these, similar to golden labels. To generate the new samples from our handcrafted hypotheses, we prompt ChatGPT to generate movie review-related sentences (to fit the chosen task) that follow a given hypothesis. This was done in an attempt to simplify the procedure by taking advantage of human strengths, generalization and extrapolative thinking, and LLM strengths, low-cost text generation, simultaneously.

C.1.2. SYNTHETIC BLIND SPOT STUDY RESULTS

The mitigation results of this human-LLM approach for our Controlled Synthetic Data Check can be seen in [table C.1](#). As can be seen in the first column of [table C.1](#), before retraining, the overall test accuracy declines in line with the degree to which the train set is biased. Interestingly, the percentage of successful perturbations by TF, i.e., the percentage of successful label flips, closely follows the overall accuracy. This mirrors the findings of Tsipras *et al.* [5], that there is a strong relationship between high accuracy and brittleness – or a lack of robustness. The number of occurring UUs as a result of the perturbation does not follow this trend, instead increasing as the training data becomes more biased, as expected. This poses an interesting optimization problem since the model becomes most robust in general terms, i.e., the successful perturbation percentage falls, but simultaneously there is a significant uptick in blind spots as the training sets become more biased.

The effect of retraining on the overall accuracy and perturbation success rate is minimal, with accuracy changing by no more than $\pm 1\%$ and perturbation success rate changing no more than $\pm 2\%$. However, the number of found UUs decreases drastically due to the retraining, with reductions of 73.93%, 80.77%, and 74.75% for the biased P, N, and PN models, respectively. The clean and randomly biased models also show a reduction, though less significant at 54.55% and 31.49%, respectively. These results confirm that our method can be used to target synthetic blind spots found in biased models through the use of hypotheses and generated instances, without significantly affecting the performance or general robustness of the model.

¹Size of training sets: $N_{clean} = 25,000$, $N_R = 2,500$, $N_P = 2,439$, $N_N = 3,138$, and $N_{PN} = 2,438$.

C.2. USER STUDY FOR HUMAN COMPUTATION

We use Prolific as a crowdsourcing platform for all our participants. Below, we present the structure followed by all survey participants for the generalization user study, consisting of an initial disclaimer, an instruction set, examples, and finally the questions. Here, we use the abstraction and extrapolation assignments on the IMDB dataset as an example. The workflow is very similar between the different generalization assignments and datasets (MRPC, IMDB, or QNLI), with only slight differences in the wording between the surveys to fit the task and dataset used, as they all present the crowd worker with some input and result in plain text output. For the generation assignment, crowdworkers are asked to perform the same steps, with relevant examples related to the structure of the dataset being shown, before finally contributing usable samples based on shown hypotheses.

C.2.1. ABSTRACTION ON IMDB

Abstraction on IMDB

Disclaimer Crowdworkers were shown an initial disclaimer to inform them that our governing ethics body sanctions this survey and to remind them not to share personal information:

- “Welcome to the Hypothesis Extrapolation Survey! Please carefully read the following: You are invited to participate in our research study. This study is fully sanctioned by our governing ethics body, as is the handling and storing of the resulting data. This research study aims to use your creativity and generalization ability to come up with new abstractions. It will take you approximately 25 minutes to complete. As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by making this survey completely anonymous. Therefore, please do not provide any personal information anywhere. The anonymous results might be shared publicly in the future. Participation in this study is entirely voluntary, and you can withdraw anytime. Feel free to contact us with any questions or feedback you might have.”

Instructions Crowdworkers were then introduced to the specific task (SE, SA, or NLI) as follows:

- “Please read the following examples carefully. All tasks in this survey are related to a single task, sentiment analysis, which tests the sentiment of a sentence is either positive or negative, applied to movie reviews. The goal here is to use your creativity and ability to generalize to spot patterns and come up with new possible samples. A fully worked-out example can be found below, with user-generated text, similar to what you are expected to write, in *italic* and instructions **bold**. You will receive all relevant instructions again when for each question.”

Examples Then, they were presented with two examples that match the dataset

used, as well as the task (abstraction, expansion, or generation), before being asked if they understood the examples:

- **“There is a sentence pair below, with one original sample (O) and a perturbed one (P), which is similar but had some things changed (shown in double square brackets). These changes may relate to a pattern, related to semantics, syntax, specific words, or something else in the samples, that leads to the wrong True or False label being predicted for semantic similarity.**

- Example 1 – The two samples are:

O: There was an overarching [[story]] that was [[refusing]] to reveal itself to me. P: There was an overarching [[narrative]] that was [[unable]] to reveal itself to me.

Formulate a hypothesis on what this pattern for O and P might be and enter it below. Try to be specific when formulating a hypothesis.

The pattern that caused the wrong prediction may be related to the substitution of the word “story” with its synonym “narrative”.

- Example 2 – The two samples are:

O: Overall, I [[loved]] the cinematography of this through and [[through]]. P: Overall, I [[looved]] the cinematography of this through and [[thrOugh]].

Formulate a hypothesis on what this pattern for O and P might be and enter it below. Try to be specific when formulating a hypothesis.

Several words have been misspelled in the samples, all related to the letter “o”. Either more letters are added “oo” or the letter is substituted with a number “0” that looks similar, making it easy to misread.”

Main Questions Finally, the actual questions preceding the text entry field used for data collection all have the same structure with the unique O and P sentences substituted in for each question:

- “The two samples are:

O: {original sentence} P: {perturbed sentence}

Formulate a hypothesis on what this pattern might be and enter it below. Try to be specific when formulating a hypothesis.”

Figure C.1.: Survey instructions, examples, and question format for the IMDB abstraction task.

C.3. USED LLM PROMPTS

We specifically instruct the LLM to split its hypothesis from its reasoning because, in our experience, this leads to a clearer and more useful answer for further steps.

C.3.1. ABSTRACTION PROMPT

Abstraction Prompt

There is a sentence pair below, with one original sample (O) and a perturbed one (P), which is similar but had some things changed. These changes may relate to a pattern, related to semantics, syntax, specific words, or something else in the samples, that leads to them being the reason the sample is misclassified by a classification algorithm. This misclassification is made at a high level of confidence.

The model is not trained on the two samples. The two samples relate to {task} and are:

O: {sentence[0]}

P: {sentence[1]}

Formulate a hypothesis on what this pattern might be. Try to be specific when formulating a hypothesis. Your response should always follow the format:

Hypothesis: {hypothesis}

Reasoning: {reasoning}

Figure C.2.: Prompt used for the abstraction step in ALL blind-spot mitigation.

C.3.2. EXTRAPOLATION PROMPT

Extrapolation Prompt

There is a sentence pair, with one original sample (O) and a perturbed one (P), which is similar but had some things changed. These changes may relate to a pattern, related to semantics, syntax, specific words, or something else, that leads to them being the reason the sample is misclassified by a classification algorithm. This misclassification is made at a high level of confidence.

The model is not trained on the two samples. The two samples relate to {task}

There is an existing hypothesis regarding the samples, that may capture a pattern related to semantics, syntax, specific words, or something else in the sample pair. This pattern leads to a misclassification of the sample.

The hypothesis is: {hypothesis}

Formulate a new hypothesis regarding those sentence samples that is concerned with the same topic but is applied to a different possible pattern that could also lead to a misclassification. Try to be specific when formulating a new hypothesis. Your response should always follow the format:

Hypothesis: {hypothesis}
Reasoning: {reasoning}

Figure C.3.: Prompt used for the extrapolation step in AIL blind-spot mitigation.

C.3.3. GENERATION PROMPT

Generation Prompt

There is a sentence pair, with one original sample (O) and a perturbed one (P), which is similar but had some things changed. These changes may be related to a pattern related to semantics, syntax, specific words, or something else that leads to them being the reason the sample is misclassified by a classification algorithm. This misclassification is made at a high level of confidence.

The model is not trained on the two samples.

A hypothesis has been formulated regarding the samples, that may capture a pattern related to semantics, syntax, specific words, or something else in the sample pair. These samples led to a classification algorithm misclassifying them at a high level of confidence.

Given the samples and a previously generalized hypothesis, generate one new sample made up of one or more sentences that relate to {task} and could have a similar effect on the classification algorithm.

The new sample should be varied and detailed. Follow the logic laid out in the given hypothesis and follow the format of the sample pair (O and P) exactly. Also include whether the new sample should be given a (positive) or (negative) label for the task: {task}.

The hypothesis is: {hypothesis}

Your response should always follow the format:

Sample: {sample}

Label: {label}

Reasoning: {reasoning}

Figure C.4.: Prompt used for the generation step in AIL blind-spot mitigation.

C.4. PERTURBATION STATISTICS AND VISUALIZATION

To add additional context to the perturbation performed, we supply the detailed attack statistics across all performed perturbations. Specifically, we report *Original Accuracy* and *Accuracy Under Attack* are reported, which are the classifier accuracy on its own and while under attack. Further, *Attack Success Rate* is shown, which is the percentage of successful perturbation attempts to failed ones. Finally, we report the number of

	MRPC _O	MRPC _L	MRPC _H	MRPC _R	IMDB _O	IMDB _L	IMDB _H	IMDB _R	QNLI _O	QNLI _L	QNLI _H	QNLI _R
Original Accuracy (%)	82.38	81.57	81.58	82.49	94.84	95.40	94.43	93.94	89.88	89.31	89.42	88.24
Accuracy Under Attack (%)	9.80	17.40	12.99	10.42	10.18	10.44	19.21	10.22	8.91	11.67	14.89	9.97
Attack Success Rate (%)	71.83	64.87	68.29	69.65	88.46	93.18	63.85	85.34	87.35	86.80	78.84	84.92
Perturbed Words (%)	7.70	9.9	8.51	7.98	4.59	7.62	9.02	5.50	6.12	8.80	9.57	7.33
Words per Input	39.3	39.3	39.3	39.3	230.0	230.0	230.0	230.0	37.9	37.9	37.9	37.9
Avg. Number of Queries	51.40	68.62	55.17	57.86	185.24	184.94	198.31	186.37	49.38	51.27	56.11	53.27

Table C.2.: Perturbation statistics across datasets and models for attacks with TF using BERT. Subscripts O, L, H, R denote the original, LM-retrained, human-retrained, and relabeled models, respectively.

	MRPC _O	MRPC _L	MRPC _H	MRPC _R	IMDB _O	IMDB _L	IMDB _H	IMDB _R	QNLI _O	QNLI _L	QNLI _H	QNLI _R
Original Accuracy (%)	82.38	82.23	82.10	82.55	95.40	95.41	95.74	94.26	89.88	89.38	89.38	88.98
Accuracy Under Attack (%)	7.78	13.73	11.94	10.42	9.54	21.43	15.32	12.51	8.21	9.90	7.30	8.67
Attack Success Rate (%)	72.00	70.38	72.64	72.35	59.41	50.59	79.70	56.87	77.54	79.74	82.08	79.27
Perturbed Words (%)	8.47	9.18	9.03	8.91	6.43	8.11	13.09	9.37	7.99	8.32	11.03	8.31
Words per Input	39.3	39.3	39.3	39.3	230.0	230.0	230.0	230.0	37.9	37.9	37.9	37.9
Avg. Number of Queries	56.92	64.37	58.61	58.23	199.32	211.65	201.44	204.12	34.91	33.53	49.09	35.75

Table C.3.: Perturbation statistics across datasets and models for attacks with DWB using BERT. Subscripts O, L, H, R denote the original, LM-retrained, human-retrained, and relabeled models, respectively.

	MRPC _O	MRPC _L	MRPC _H	MRPC _R	IMDB _O	IMDB _L	IMDB _H	IMDB _R	QNLI _O	QNLI _L	QNLI _H	QNLI _R
Original Accuracy (%)	90.84	89.86	90.20	90.61	95.20	94.96	94.67	94.86	90.08	89.58	89.16	89.90
Accuracy Under Attack (%)	13.85	18.31	12.43	14.09	20.97	18.22	15.09	17.55	12.64	15.29	14.53	13.67
Attack Success Rate (%)	68.70	65.24	69.54	66.89	71.32	75.64	78.31	70.55	83.42	79.12	75.87	81.34
Perturbed Words (%)	9.23	8.12	9.68	8.97	6.45	7.54	10.88	8.36	7.34	8.69	9.11	7.92
Words per Input	39.3	39.3	39.3	39.3	230.0	230.0	230.0	230.0	37.9	37.9	37.9	37.9
Avg. Number of Queries	53.92	62.34	57.92	55.76	191.34	192.85	198.21	194.43	48.22	49.98	52.89	50.76

Table C.4.: Perturbation statistics across datasets and models for attacks with TF using Llama 2. Subscripts O, L, H, R denote the original, LM-retrained, human-retrained, and relabeled models, respectively.

	MRPC _O	MRPC _L	MRPC _H	MRPC _R	IMDB _O	IMDB _L	IMDB _H	IMDB _R	QNLI _O	QNLI _L	QNLI _H	QNLI _R
Original Accuracy (%)	90.66	89.73	89.91	90.73	95.33	95.13	94.90	95.10	90.72	90.10	89.73	90.60
Accuracy Under Attack (%)	16.35	14.79	13.87	15.68	21.78	20.32	19.12	22.19	11.78	10.95	14.28	12.44
Attack Success Rate (%)	66.40	63.89	67.56	65.78	70.42	68.55	71.32	74.65	79.78	77.24	82.43	80.34
Perturbed Words (%)	9.11	8.76	9.02	8.86	7.18	6.92	11.54	9.29	8.06	9.11	10.24	8.76
Words per Input	39.3	39.3	39.3	39.3	230.0	230.0	230.0	230.0	37.9	37.9	37.9	37.9
Avg. Number of Queries	60.22	65.14	62.03	61.76	203.56	199.42	204.29	208.23	45.29	43.87	50.77	47.83

Table C.5.: Perturbation statistics across datasets and models for attacks with DWB using Llama 2. Subscripts O, L, H, R denote the original, LM-retrained, human-retrained, and relabeled models, respectively.

Perturbed Words, the percentage of words that are perturbed, the *Words per Input*, the average number of words per input, and the *Average Number of Queries*, which is how many tries it took the perturbation method to find the best attack. For BERT, the attack statistics for TF attacks are shown in [table C.2](#) while the ones for DWB attacks are shown in [table C.3](#). For Llama 2 7B, the attack statistics for TF attacks are shown in [table C.4](#)

and for DWB in [table C.5](#).

To visualize the effect of the retraining across the perturbed samples, we show the confidence at which predictions are made in [figures C.5 to C.10](#). We observe a reduction in high confidence misclassifications; specifically, predictions with confidence greater than 90% are reduced as a result of our method. Additionally, we observe similar results across perturbation methods, indicating that the performance of our approach is not dependent on the type of perturbation method employed.

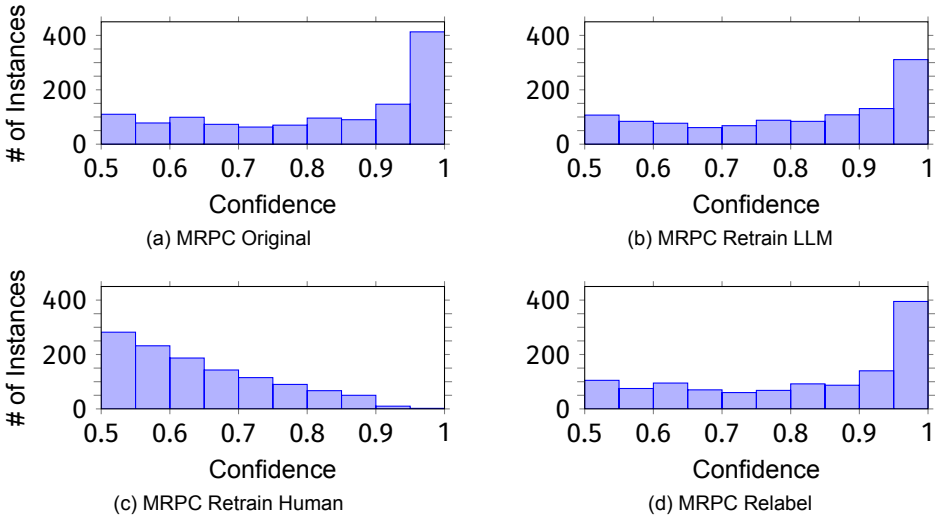


Figure C.5.: Plots of successful perturbations for MRPC when using TF, showing the distribution of the number of instances across confidence bins.

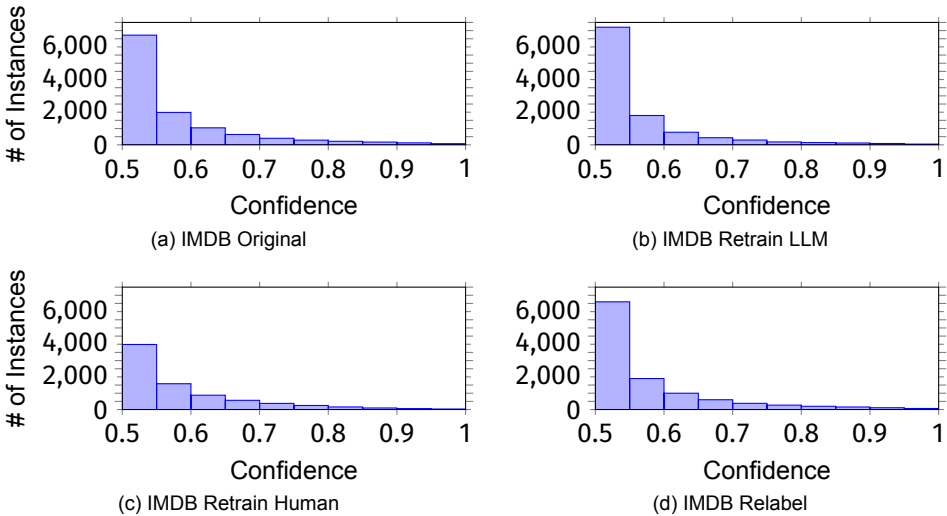


Figure C.6.: Plots of successful perturbations for IMDB when using TF, showing the distribution of the number of instances across confidence bins.

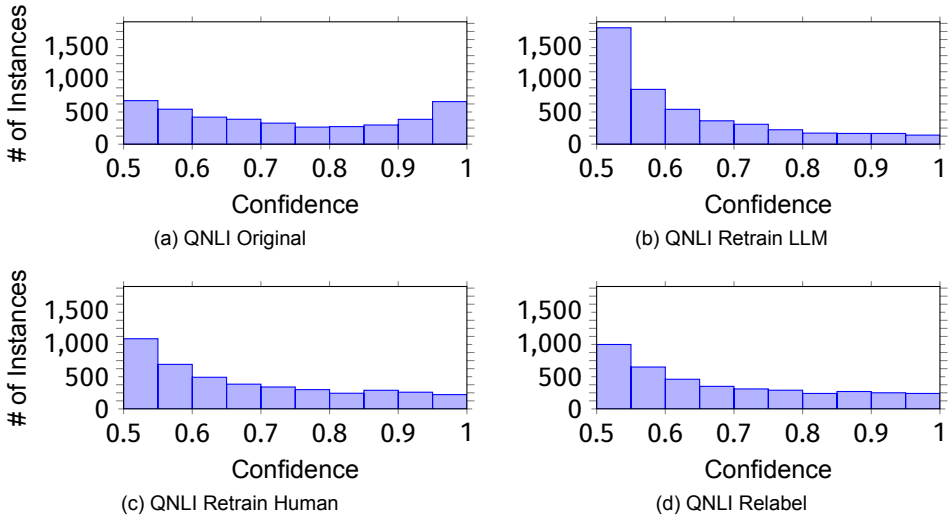


Figure C.7.: Plots of successful perturbations for QNLI when using TF, showing the distribution of the number of instances across confidence bins.

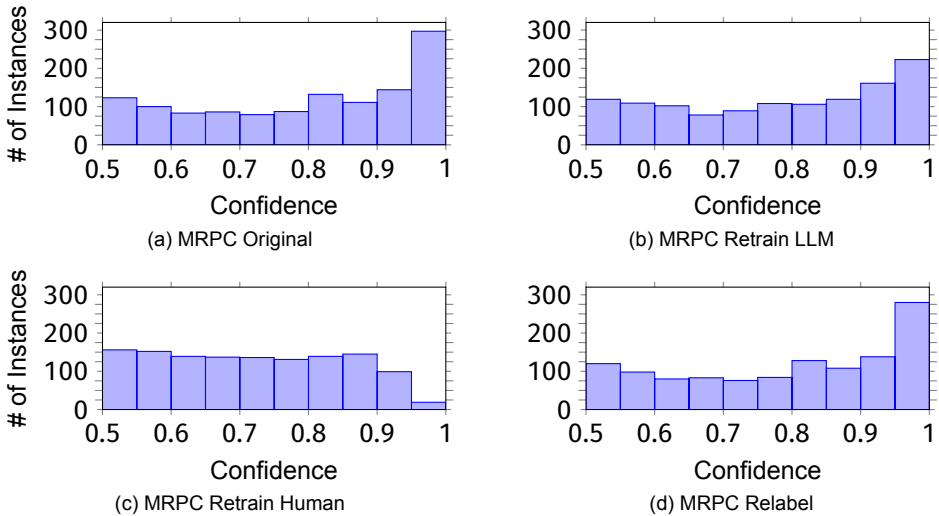


Figure C.8.: Plots of successful perturbations for MRPC when using DWB, showing the distribution of the number of instances across confidence bins.

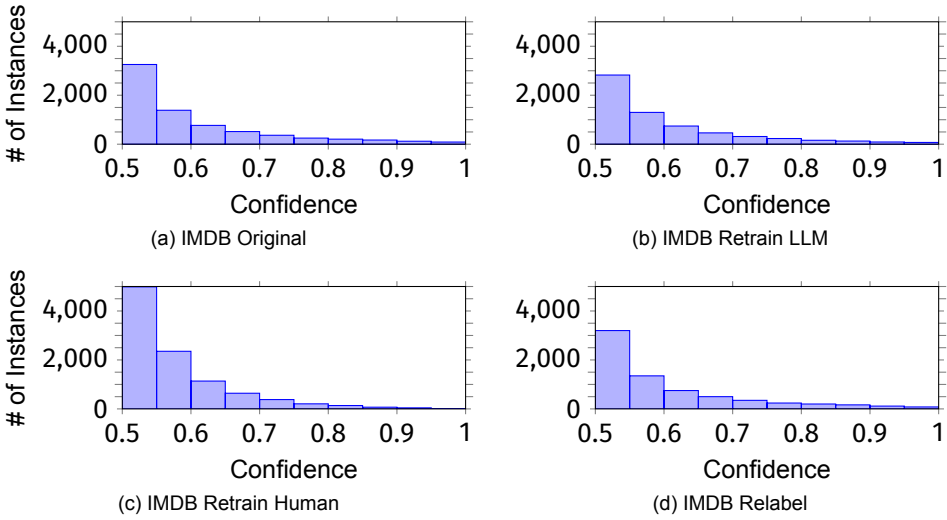


Figure C.9.: Plots of successful perturbations for IMDB when using DWB, showing the distribution of the number of instances across confidence bins.

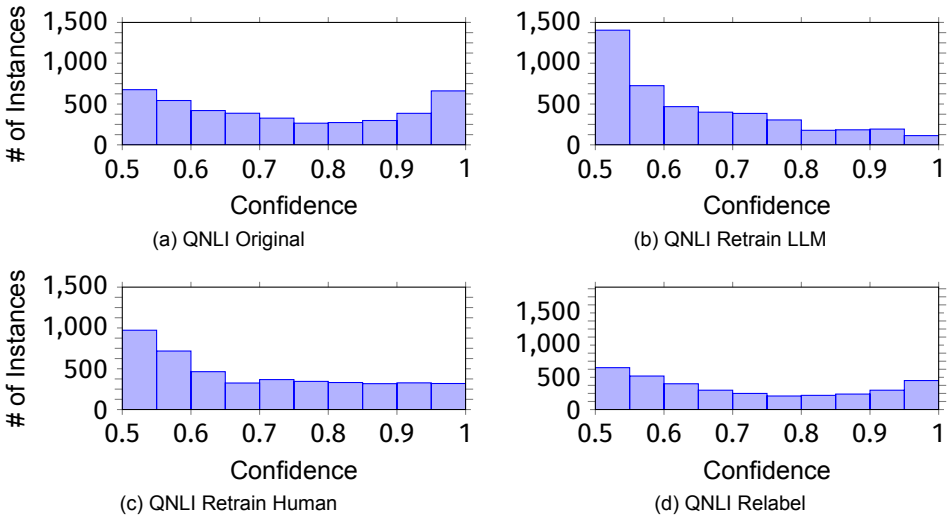


Figure C.10.: Plots of successful perturbations for QNLI when using DWB, showing the distribution of the number of instances across confidence bins.

REFERENCES

- [1] S. Hochreiter and J. Schmidhuber. 'Long Short-Term Memory'. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. issn: 0899-7667. doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735). url: <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [2] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin and Y. Qi. 'TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Q. Liu and D. Schlangen. Online: Association for Computational Linguistics, Oct. 2020, pp. 119–126. doi: [10.18653/v1/2020.emnlp-demos.16](https://aclanthology.org/2020.emnlp-demos.16/). url: <https://aclanthology.org/2020.emnlp-demos.16/>.
- [3] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen and C. Seifert. 'From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI'. In: *ACM Computing Surveys* 55.13s (13th July 2023), pp. 1–42. issn: 0360-0300. doi: [10.1145/3583558](https://doi.org/10.1145/3583558). url: <https://doi.org/10.1145/3583558>.
- [4] B. Liu, M. Hu and J. Cheng. 'Opinion Observer: Analyzing and Comparing Opinions on the Web'. In: *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan, 10th–14th May 2005). WWW '05. New York, NY, USA: Association for Computing Machinery, May 2005, pp. 342–351. isbn: 1-59593-046-9. doi: [10.1145/1060745.1060797](https://doi.org/10.1145/1060745.1060797). url: <https://doi.org/10.1145/1060745.1060797>.
- [5] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner and A. Madry. 'Robustness May Be at Odds with Accuracy'. In: *International Conference on Learning Representations*. 2019. url: <https://openreview.net/forum?id=SyxAb30cY7>.

D

Appendix for Chapter 3 (Part II): ZEST Artifacts

D.1. PROMPTING STRATEGIES FOR SYNTHETIC CORPUS GENERATION

This section details the prompting strategies we employ to generate the synthetic context corpus $\mathcal{D}_{\text{synth}}$. The process, as described in [Section 3.4.3](#), is hierarchical, involving two main steps: (1) Domain Anchor Generation and (2) Synthetic Corpus Expansion. We used the default API sampling parameters (e.g., temperature) for all generations to ensure reproducibility. The $k = 5$ exemplar documents \mathcal{D}_{ex} are assumed to be provided as part of the input to the LLM for the first step.

Step 1: Domain Anchor Generation The objective of this step is to generate A diverse domain anchor documents from the exemplar set \mathcal{D}_{ex} to create $\mathcal{D}_{\text{anchors}}$. These anchors serve as thematic seeds. They are generated sequentially to encourage diversity and avoid thematic repetition. For each anchor a_i , the LLM is instructed to produce a concise document that captures a distinct topical or stylistic facet present in the provided exemplars, while also being mindful of previously generated anchors in the sequence (if applicable). The generalized prompt structure for generating a single domain anchor a_i is shown in [Figure D.1](#).

Step 2: Synthetic Corpus Expansion Once the domain anchors are generated, the synthetic corpus $\mathcal{D}_{\text{synth}}$ (of size J') is created by expanding upon these anchors. For each anchor document $a_i \in \mathcal{D}_{\text{anchors}}$, the LLM is prompted to generate (J'/A) of synthetic documents (we assign one extra document per anchor until the remainder is exhausted). This step is performed in parallel for each anchor. The goal is for the LLM to elaborate on and diversify the theme encapsulated by the specific anchor a_i , producing a set of full-length, representative documents. The generalized prompt structure for expanding a single domain anchor is shown in [Figure D.2](#).

D.2. DETAILS ON EXEMPLAR SET SAMPLING

To enable zero-shot contextual adaptation in ZEST, we rely on a small exemplar set $\mathcal{D}_{\text{ex}} = \{d_{\text{ex}}^1, \dots, d_{\text{ex}}^k\}$ to guide the generation of the synthetic context corpus $\mathcal{D}_{\text{synth}}$. As described in [Section 3.4.4](#), these exemplars are sourced from the BEIR benchmark. This section details the process of selecting \mathcal{D}_{ex} , the mapping of BEIR tasks to unique domain keywords, and the measures taken to ensure no information leakage between \mathcal{D}_{ex} and the MTEB evaluation datasets.

Mapping BEIR Tasks to Domain Keywords. To systematically select exemplars that typify the domain of an MTEB target task, we first assign each of the 18 BEIR tasks a unique keyword that encapsulates its primary domain or task characteristic. These keywords serve as an intermediary representation, allowing us to later align MTEB tasks with the most relevant BEIR-derived exemplars based on domain similarity. [Table D.1](#) presents this mapping, with each keyword chosen to be distinct and representative of the task’s content.

The keyword assignment prioritizes the dominant domain or retrieval objective of each BEIR task. For instance, biomedical tasks like BioASQ and TREC-COVID are assigned keywords like “BiomedQA” and “COVIDResearch,” respectively, to distinguish their focus within the broader biomedical domain. Similarly, tasks like Quora and CQADupStack, both involving question answering, are differentiated by keywords “DuplicateQA” and “ForumQA,” reflecting their specific contexts (duplicate question detection versus forum-based Q&A). This approach ensures that the keywords are sufficiently granular to avoid overlap while remaining general enough to facilitate alignment with MTEB tasks. For MTEB tasks, we select the BEIR task whose keyword best matches the MTEB task’s domain or retrieval goal, determined by manual inspection of task descriptions and data characteristics. If an MTEB task corresponds to a BEIR task (e.g., ArguAna), we select the next closest task to avoid direct overlap, as noted in [Section 3.4.4](#).

Practical Considerations. The use of BEIR tasks as a source for \mathcal{D}_{ex} leverages their diversity and public availability, making the approach reproducible and scalable. The keyword-based mapping simplifies the alignment of MTEB tasks to appropriate exemplars while avoiding direct task-to-task dependencies, which could risk evaluation bias. By sampling only a small number of documents, we simulate a realistic scenario where practitioners provide minimal domain examples, aligning with ZEST’s goal of minimal input requirements. The ablation studies in [Section 3.4.5](#) confirm that our approach provides sufficient diversity for effective synthetic context generation, supporting the robustness of this sampling strategy.

D.3. TASK-SPECIFIC PREFIXES

We use standard prefixes, hand-written for each MTEB evaluation dataset, across all our evaluations. The prefix selection procedure follows the methodology outlined in [1]. The specific prefix categories are:

- Search query
- Search document
- Classification
- Clustering

Using these prefixes helps the model identify the task at hand and ensures consistency in how the model receives data for both real-context and synthetic-context scenarios. The specific prefixes used are given in [Table D.1](#).

D.4. EXAMPLES OF GENERATED SYNTHETIC DOCUMENTS: FULL PIPELINE

This section provides examples illustrating the full pipeline used by ZEST to generate synthetic documents for the $\mathcal{D}_{\text{synth}}$ corpus. These examples demonstrate how an initial exemplar document (\mathcal{D}_{ex}) from a specific domain guides the generation of a domain anchor, which in turn seeds the creation of a final synthetic document. This hierarchical process, as described in [Section 3.4.3](#), is based on k exemplar documents (for clarity, we show the pipeline for $k = 1$ exemplar in each domain example below, referenced as examples in [Figure D.3](#) through [Figure D.8](#)).

Example 1: Biomedical Domain An exemplar document focusing on genetic recoding in Archaea (see [Figure D.3](#)) was provided to ground the generation process in the biomedical domain. This initial document serves as the primary input for the LLM to understand the target domain’s characteristics.

Based on the provided biomedical exemplar, the LLM generated a domain anchor (see [Figure D.4](#)). This anchor encapsulates a core theme derived from the exemplar – in this case, programmed ribosomal frameshifting in Archaea – and serves as a more focused seed for subsequent document generation.

Expanding upon the biomedical domain anchor, the LLM then produced a full synthetic document (see [Figure D.5](#)). This final document elaborates on the mechanisms and implications of frameshifting, demonstrating how the anchor guides the creation of a more detailed and contextually relevant piece of text for the synthetic corpus.

Example 2: Financial Domain For the financial domain, an exemplar document discussing interest rates and loan types (see [Figure D.6](#)) was used, which leads to a corresponding domain anchor (see [Figure D.7](#)), and, finally, a synthetic document (see [Figure D.8](#)). This document also explores lender strategies and borrower behavior in response to varying interest rate environments.

BEIR Task	Unique Keyword
MS MARCO	WebSearch
TREC-COVID	COVIDResearch
NFCorpus	Nutrition
BioASQ	BiomedQA
HotpotQA	MultiHopQA
FiQA-2018	FinanceQA
Signal-1M (RT)	SocialMedia
TREC-NEWS	NewsSearch
Robust04	NewsArchive
ArguAna	Argumentation
Touché-2020	Debate
CQADupStack	ForumQA
Quora	DuplicateQA
DBPedia-Entity	EntityRetrieval
SCIDOCS	Citation
SciFact	SciFactCheck
Climate-FEVER	ClimateClaims
FEVER	FactCheck

Table D.1.: Mapping of BEIR tasks to unique domain keywords. Each keyword encapsulates the primary domain or task characteristic, enabling alignment with MTEB tasks based on domain similarity.

Prompt for Domain Anchor Generation

Systematically examine the {k} exemplar documents provided below to extract and synthesize their core themes, stylistic patterns, and domain-specific terminology. Leverage this analysis to craft a new domain anchor document that encapsulates these elements.

Here are the exemplar documents:

Exemplar 1: {exemplar_document_1_text}

Exemplar 2: {exemplar_document_2_text}

Exemplar k: {exemplar_document_k_text}

Previously generated anchor documents (if any):

- {anchor_1}

- {anchor_2}

- {anchor_i-1}

Your task is to generate a new, concise domain anchor document. This document should:

1. Be approximately as long as the exemplar documents.
2. Capture a distinct and specific topical theme, concept, or stylistic characteristic evident in the exemplar documents.
3. Cover key terminology, entities, and typical writing style of the domain as represented by the exemplars.
4. If previous anchors were mentioned, ensure this new anchor explores a DIFFERENT facet or theme than those already covered to maximize diversity.
5. The anchor should be a coherent piece of text, similar to the exemplar documents, not just a list of keywords.

Generate only the domain anchor document itself.

Figure D.1.: Prompt for Domain Anchor Generation

Prompt for Synthetic Corpus Expansion

You are tasked with generating a document that is representative of a specific domain and theme.

You are given the following domain anchor document to build on, which encapsulates a key theme or stylistic element of the target domain:

Domain Anchor:

{domain_anchor_document_text}

Your task is to generate another full synthetic document that elaborates on,

exemplifies, and diversifies the core theme and style presented in the domain anchor. This new document should:

1. Be topically coherent with the provided domain anchor.
2. Be a complete, well-structured document (e.g., an article, a report excerpt, a descriptive passage) of similar length.
3. Should explore various sub-topics, perspectives, or aspects related to the main theme of the anchor, ensuring diversity among them.
4. Maintain a style (e.g., tone, vocabulary, sentence structure) consistent with the domain anchor and typical of the implied domain.
5. Be factually plausible and internally consistent, even if entirely synthetic.

Respond only with your generated document. Ensure the document is clearly separated by placing "—DOCUMENT END—" at the end of the document you generate.

Figure D.2.: Prompt for Synthetic Corpus Expansion

Biomedical Exemplar Document

The standard rules of genetic translational decoding are altered in specific genes by different events that are globally termed recoding. In Archaea recoding has been unequivocally determined so far only for termination codon readthrough events. We study here the mechanism of expression of a gene encoding for a α -l-fucosidase from the archaeon *Sulfolobus solfataricus* (*fucA1*), which is split in two open reading frames separated by a -1 frameshifting. The expression in *Escherichia coli* of the wild-type split gene led to the production by frameshifting of full-length polypeptides with an efficiency of 5%. Mutations in the regulatory site where the shift takes place demonstrate that the expression in vivo occurs in a programmed way. Further, we identify a full-length product of *fucA1* in *S.solfataricus* extracts, which translate this gene in vitro by following programmed -1 frameshifting. This is the first experimental demonstration that this kind of recoding is present in Archaea.

Figure D.3.: Example of an input exemplar document (\mathcal{D}_{ex}) for the biomedical domain, detailing genetic recoding in Archaea. This document serves as the initial seed for the LLM.

Biomedical Domain Anchor

Programmed ribosomal frameshifting represents a crucial gene expression regulation mechanism in Archaea, enabling the synthesis of alternative protein products from a single mRNA. This recoding event, often involving a -1 shift, allows organisms like *Sulfolobus solfataricus* to expand their proteomic diversity and adapt to specific environmental conditions. Understanding the cis-acting RNA elements and trans-acting factors that govern frameshift efficiency in these extremophiles is key to deciphering their unique biology and potential biotechnological applications in areas such as novel enzyme discovery or controlled protein synthesis. The study of such mechanisms also sheds light on the evolutionary plasticity of the genetic code.

Figure D.4.: Domain anchor generated from the biomedical exemplar (see Figure D.3). It focuses on the theme of programmed ribosomal frameshifting in Archaea.

Biomedical Synthetic Document

The efficiency of programmed -1 ribosomal frameshifting in archaeal organisms, such as the crenarchaeon *Sulfolobus solfataricus*, is intricately regulated by specific mRNA sequences and structural motifs located proximal to the frameshift site. These regulatory elements typically include a 'slippery sequence' (often a heptanucleotide like A AAA AAC) where the ribosome can lose its grip on the mRNA, and a downstream RNA secondary structure, commonly a pseudoknot or a stable hairpin. This structure is thought to induce ribosomal pausing, providing a kinetic window for the realignment of the tRNA in the P-site to the -1 frame. The interplay between the ribosome's own components, the mRNA transcript, various translation factors, and the nascent peptide chain within the exit tunnel collectively influences the probability and fidelity of a successful frameshift event. Current research endeavors focus on genome-wide mapping of these frameshift-prone sites in diverse Archaea, aiming to uncover novel regulatory networks and identify potential targets for antimicrobial agents or innovative tools for synthetic biology, such as precisely controlling the stoichiometric production of multiple proteins from a single engineered transcript.

Figure D.5.: Final synthetic document ($\mathcal{D}_{\text{synth}}$) for the biomedical domain, expanded from the domain anchor (see Figure D.4), discussing the regulation and implications of ribosomal frameshifting.

Financial Exemplar Document

A lot of loans are taken out on a fixed rate basis, so the rate is part of the contract and is therefore covered by contract law. If the loan is taken out on a variable basis then in principle the rate can rise within the terms of the contract. If a particular lender tries to raise its rates out of line with the market then its customers will seek alternative, cheaper, loans and pay off their expensive loan if they can. If rates rise sharply in general due to unusual politico-economic circumstances then those with variable rate loans can find themselves in severe trouble. For example the base rate in the UK (and therefore variable mortgage rates closely tied to it) spiked sharply in the late 80s which caused severe stress to a lot of borrowers and undoubtedly pushed some into financial difficulties.

Figure D.6.: Example of an input exemplar document (\mathcal{D}_{ex}) for the financial domain, discussing factors influencing interest rates and loan types.

Financial Domain Anchor

The interplay between fixed and variable interest rates significantly impacts borrower stability and broader market dynamics. Fixed-rate loans offer predictability and insulate borrowers from sudden, adverse rate hikes, thereby fostering financial planning certainty. Conversely, variable-rate loans, while potentially offering lower initial rates, expose borrowers to market fluctuations, which can lead to payment shocks and financial distress during periods of rising interest rates. Central bank monetary policies, inflation expectations, and overall macroeconomic health are primary drivers influencing these rates, creating a complex and evolving landscape for both lenders managing risk and borrowers making critical financial decisions.

Figure D.7.: Domain anchor generated from the financial exemplar (see Figure D.6). It centers on the impact of fixed versus variable interest rates and market conditions.

Financial Synthetic Document

Financial institutions meticulously navigate the complexities of fluctuating interest rate environments by strategically managing their portfolios of fixed and variable rate loan products. This often involves sophisticated hedging strategies using interest rate swaps or other derivatives to mitigate the risk associated with rate

volatility. For consumers and businesses, a thorough understanding of their loan agreements' terms – particularly clauses related to rate adjustments – is paramount. During periods when monetary policy tightens and rate increases are anticipated, there's typically a discernible surge in demand for refinancing variable-rate debt (like adjustable-rate mortgages or ARMs) into fixed-rate alternatives, although eligibility for such refinancing heavily depends on the borrower's creditworthiness and prevailing market conditions. Conversely, when central banks signal an easing cycle and rates are expected to decline, variable-rate loans might appear more attractive due to potentially lower initial payments. However, these carry the inherent risk of future increases should economic conditions shift unexpectedly. Lenders also adjust their credit scoring models and underwriting standards in response to the perceived risk in the interest rate cycle.

Figure D.8.: Final synthetic document ($\mathcal{D}_{\text{synth}}$) for the financial domain, expanded from the domain anchor (see Figure D.7), detailing lender strategies and borrower behavior in different interest rate environments.

REFERENCES

- [1] Z. Nussbaum, J. X. Morris, A. Mulyar and B. Duderstadt. 'Nomic Embed: Training a Reproducible Long Context Text Embedder'. In: *Transactions on Machine Learning Research* (26th Feb. 2025). Reproducibility Certification. issn: 2835-8856. url: <https://openreview.net/forum?id=IPmzyQSiQE>.

E

Appendix for Chapter 3 (Part III): Multimodal Manga Translation

E.1. PAGE PROCESSING AND TYPESETTING

Page-to-page manga translation involves three distinct steps: (1) *page processing* to identify elements on the page, detect text, and estimate reading order; (2) *translating* the text into the target language; and (3) *typesetting* by removing the source text from the page and inserting the translated text in a stylized font. We will discuss (1) and (3) in this section.

Page Processing The first step in manga translation is identifying the elements on the page. Here, we will present an example of a manga page processing pipeline composed of methods proposed by previous research and publicly available manga tools. For text detection, we employ the unconstrained method proposed by Del Gobbo and Matuk Herrera [4] to account for text that is not contained within speech bubbles – see the top right of figure E.1. However, before applying Optical Character Recognition (OCR) to the detected text fields, we need to group it into clusters belonging to the same utterance. To accomplish this, we apply a method inspired by Rigaud, Nguyen and Burie [5] – we utilize the OPTICS algorithm [6], specifically the Python pyclustering library implementation [7], to cluster the text – see the bottom left of figure E.1. We then compute the bounding boxes of these obtained text clusters and discard those that are too small to contain text – see the bottom right of figure E.1. Finally, we apply Manga OCR¹ for text recognition – see figure E.2.

For panel detection and estimating the reading order, we utilize the Magi system [8]. In theory, Magi is capable of creating a transcript of a manga page independently, but it was trained on translations of manga and is not well-suited for Japanese text detection. As such, we only utilize some of its functionalities. A visualization of the page processing pipeline can be seen in Figure E.2. First, we use the process described previously to

¹<https://github.com/kha-white/manga-ocr>

detect text boxes. Then, we employ Magi to detect panels and estimate the reading order based on the relative locations of text boxes and panels. Lastly, we utilize Manga OCR for text extraction.

Typesetting The final step of a page-to-page manga translation – inserting the translated text back into the image – involves two steps: cleaning and lettering. We did not perform it as part of our study but, for the sake of giving potential future works a comprehensive guide to follow, we will still outline the procedure here. Cleaning refers to the removal step of the text, in which the original Japanese text is removed from the image used in the translation process. One could utilize an image inpainting model for this task [9]; the regions containing text lines are replaced by the inpainting model, which effectively removes the text even when it is overlaid on textured images or drawings. Alternatively, a method that performs this step specifically for manga has been proposed [10].

Lettering is the final step, where the translated text is rendered onto the cleaned image, with an optimized font size and placement that fits the manga aesthetic. The location of the rendered text is chosen to maximize the font size while ensuring that all text remains within the designated text region. This step ensures that the translated text is legible and properly integrated into the image. To the best of our knowledge, there exists no prior work that proposes to do this automatically at this time. Though there are plenty of manual tools (both free and commercial) that make it possible to come up with a semi-automated approach if the coordinates of the original text are known.

E.2. CHAIN OF DENSITY SUMMARIZATION

Chain-of-Density (COD) Prompt

Existing Summary from the previous Translation: {self.prev_context}

The most recent Observation from the {self.lang} translation was: {self.observation}

You will generate new increasingly concise, entity-dense summaries based on the above Existing Summary and most recent Observation.

Keep the summaries in {self.lang}.

You will create 3 summaries. You will create each of them by following the following two steps:

-Step 1. If possible, identify 1-3 Informative Entities (“;” delimited) from the most recent Observation which are missing from the Existing Summary.

-Step 2. Write a new, denser summary of identical length which covers every entity, action, and detail from the previous Existing Summary plus the Informative Entities from the Observation.

An Informative Entity is:

-Relevant: to the translation’s unfolding narrative.

-Specific: descriptive yet concise (10 words or fewer).

- Novel: not in the previous summary.
- Faithful: an accurate, detailed reflection of the translation.

Guidelines:

- The first of the three summaries must be long (but less than {lmax} words) yet highly non-specific, containing little information beyond the entities marked as missing. Use verbose language and fillers (e.g., "In this part of the translation, the main character encounters ...") to reach {lmax} words.
- Make every word count: rewrite the previous summary to improve flow and make space for additional Informative Entities.
- Make every word count: rewrite the previous summary to improve flow and make space for additional Informative Entities.
- Make space with fusion, compression, and removal of uninformative phrases like "the scenario presents".
- The summaries should become highly dense and concise yet self-contained, e.g., easily understood without referencing the fact that a translation is being performed, and contain all information of the narrative thus far.
- Informative Entities can appear anywhere in the new summary. -Only drop the least relevant Informative Entities from the previous summary if the summary length exceeds {lmax} words. Otherwise carry all previous Informative Entities to the new summary.

Answer in JSON. The JSON should be a list (length 3) of dictionaries under the key "summaries". Each dictionary should contain keys "Informative_Entities" (storing the Informative Entities included in the corresponding summary) and "Denser_Summary" (containing the summary).

Figure E.3.: Prompt used for Chain of Density summarization.

In the context of text summarization, Chain of Density (COD) [11] prompting can be used to generate high-quality summaries by breaking down the summarization task into a series of smaller, more manageable sub-tasks. COD prompting can be applied to summarization as follows:

Initial Prompt: The process starts with a prompt that instructs the LLM to read and understand the source text that needs to be summarized.

First Generation: The LLM generates a brief summary or highlights the key points from the source text based on the initial prompt.

Iterative Prompting: The generated summary from the previous step is used as the prompt for the next step. The LLM is then prompted to expand or refine the summary by adding more details, rephrasing certain parts, or reorganizing the information. This step can be repeated multiple times, with each subsequent prompt building upon the previous summary.

Final Summary: After several iterations, the final summary should be a coherent, concise, and informative representation of the source text.

By breaking down the summarization process into smaller steps, COD helps the LLM maintain focus and context throughout the summary generation process. This can lead to more coherent and accurate summaries, as the model can incrementally refine and improve the summary at each step. It's important to note that the effectiveness of COD prompting for summarization may depend on the quality of the initial prompt, the complexity of the source text, and the LLM's capabilities.

The VBP-VIS-COD approach we propose does not try to make use of the large context window size of GPT-4 Turbo and instead uses COD to maintain a rolling summary of the developments in the story so far. In addition to the image, the model is also given a summary of the story so far in the target language as additional context. It is then asked to, in addition to the translation, return the description of the events taking place on the page being translated, both in Japanese and the target language. A separate COD module then prompts the LLM to update the previous summary with the new developments, to achieve an even denser and updated summary through the process detailed above, with each summary within the same call being a more concise version of the previous one. We call this method VBP-VIS-COD, as it translates the volume one page at a time (VBP), using visual context (VIS) and chain of density prompting (COD).

The prompt used for COD is shown in [figure E.3](#).

E.3. DETAILS OF MQM HUMAN EVALUATION

The goal of using MQM is to produce a method of human evaluation that is consistent, efficient, and sufficiently granular. We use MQM with a manga-specific list of issue types, covering different error types:

MQM Issue Types

- Fluency
 - Punctuation
 - Orthography (spelling, punctuation)
 - Grammar (is it ungrammatical or not)
- Accuracy
 - Addition or omission
 - Mistranslation
 - Untranslated text
- Proper Nouns / Terminology
 - Orthography
 - Failed to recognize as proper noun
- Style

- Formality
- Awkward
- Boring
- Tone (emotional tone is miscalibrated)
- Other
 - Other

These types are not used in actual score computation, but they are useful for helping us understand the problems of a given piece of translated text. Each error type is assigned a severity level by the evaluating translator, ranging from minor to critical, depending on the impact of the issue on overall quality. MQM provides a scoring system that allows for the calculation of overall quality scores based on the number of identified issues and their severity levels. The MQM score is computed using the following equation

$$S = 1 - \frac{5 \times C_{Min} + 10 \times C_{Maj} + 25 \times C_{Crit}}{\text{Total Word Count}} \quad (\text{E.1})$$

where C_{Min} , C_{Maj} , and C_{Crit} are the number of errors with a severity of minor, major, and critical, respectively. The evaluating translator decides for each error what the most appropriate severity would be.

E.4. FULL PROMPTS

This section includes all the prompts used as part of our experiments. Only the JA-EN prompts are shown, as the only difference between them and the JA-PL prompts is that the target language needs to be explicitly specified in the prompt if it is not English and that the given example has Polish as its target language instead of English. The shown prompts can therefore be used with any target language with only very slight alterations.

Below, [figure E.5](#) to [figure E.13](#) show the prompts used for all of our approaches.

Line-by-Line (LBL) Prompt

You will act as a Japanese manga translator. You will be working with copyright-free manga exclusively. I will give you one line spoken by a character from a manga.

Here is the line: {self.line}

Your task is to translate the line to {self.lang}. Return the translated line in {self.lang} in square brackets [].

Example: {self.jp_example} Return: [{self.lang_example}]

Figure E.5.: Prompt used for LBL approach.

Page-by-Page (PBP) Prompt

You are a manga translator. You are working with copyright-free manga exclusively. I will provide the lines spoken by the characters on a page.

Here are lines spoken by the characters in order of appearance: {self.line}.

Provide the translated lines in square brackets [], without any additional words or characters. Provide only one translation for each line.

Example: {self.jp_example} Return: [{self.lang_example}]

Figure E.6.: Prompt used for PBP approach.

Line-by-Line with Visual Context (LBL-VIS) Prompt

You will act as a Japanese manga translator. You will be working with copyright-free manga exclusively. I will give you one line spoken by a character from a manga. I will also give you a manga page this manga comes from.

Here is the line: {self.line}

Your task is to translate the line to {self.lang} and to explain how the image informs your translation. Return the translated line in {self.lang} in square brackets and the explanation for how the image informs the translation in parentheses.

Example: {self.jp_example} Return: [{self.lang_example}] ({self.img_explanation_example}).

Figure E.7.: Prompt used for LBL-VIS approach.

Page-by-Page with Visual Context (PBP-VIS) Prompt

You are a manga translator. You are working with copyright-free manga exclusively. I have given you a manga page, and will provide the lines spoken by the characters.

Here is the page and the lines spoken by the characters in order of appearance: {self.page}

For each of the lines, provide a translation in square brackets and explanation for how the image informs the translation in parentheses. Provide only one translation for each line.

Example: {self.jp_example} Return: [{self.lang_example}] ({self.img_explanation_example}).

Figure E.8.: Prompt used for PBP-VIS approach.

PBP-VIS with Numbers (PBP-VIS-NUM) Prompt

You are a manga translator. You are working with copyright-free manga exclusively. I have given you a manga page, and will provide the lines spoken by the characters. The lines are taken from the speech bubbles with corresponding numbers.

Here is the page and the lines spoken by the characters in order of appearance: {self.page}

For each of the lines, provide a translation in square brackets and explanation for how the image informs the translation in parentheses. Provide only one translation for each line.

Example: Line 1: {self.jp_example} Return: Translation 1: [{self.lang_example}] ({self.img_explanation_example}).

Figure E.9.: Prompt used for PBP-VIS-NUM approach.

VBP-VIS with Chain-of-Density (VBP-VIS-COD) Prompt

You are a manga translator. You are working with copyright-free manga exclusively.

Here is a summary of the story so far: {self.lang_summary}

I have given you the next manga page, and will provide the lines spoken by the characters.

Here is the page and the lines spoken by the characters in order of appearance: {self.page}

Your task is to translate the lines I gave you. For each of the lines I want you to give the translation, and the reasoning behind choosing this particular translation. The reasoning has to relate the line to the relevant part of the page and explain how it makes sense. The translation should be consistent with the story so far.

Answer in JSON. The JSON should contain three keys.

The first key, "story_jp", should contain a string describing the events taking place on the manga page I provided. This story has to be in Japanese and incorporate the lines I gave you verbatim.

The second key, "story_en", should contain a translation of the Japanese story to English. Incorporate your translations of the character lines into that story and

make sure they fit.

The third key, "lines", should contain a list of dictionaries. The dictionary at position n, should contain information relevant to the n-th line. Each dictionary should contain five keys:

- "line" - containing the original Japanese line,
- "speaker" - information about the person speaking, such as age, gender etc.,
- "situation" - information about the place and social situation,
- "translation" - containing the translation of the line,
- "reasoning" - containing the explanation for the translation.

Example: Line 1: {self.jp_example}

Return:

```
(
  "story_jp": "{self.jp_story}",
  "story_en": "{self.lang_story}",
  "lines": [
    (
      "line": "{self.jp_example}",
      "speaker": "{self.lang_speaker}",
      "situation": "{self.lang_situation}",
      "translation": "{self.lang_example}",
      "explanation": "{self.lang_explanation}",
    ),
  ]
)
```

Figure E.10.: Prompt used for VBP-VIS-COD approach.

VBP-VIS, 3 Pages (VBP-VIS-3P) Prompt

You are a manga translator. You are working with copyright-free manga exclusively. I have given you a couple of consecutive manga pages, and will provide the lines spoken by the characters. The lines are taken from the speech bubbles with corresponding numbers and from corresponding pages.

Here is the page and the lines spoken by the characters in order of appearance: {self.page}

Your task is to translate the lines I gave you. For each page, for each of the lines I want you to give the translation, and the reasoning behind choosing this particular translation. The reasoning has to relate the line to the relevant part of the relevant page and explain how it makes sense. Make sure all the lines make

sense in context of all the pages.

Answer in JSON. The JSON should contain a list of lists under the key "pages". The list at position n, should contain information relevant to the n-th page. The n-th list, should be a list of dictionaries. The dictionary at position i, should contain information relevant to the t-th line. Each dictionary should contain three keys: "line" - containing the original Japanese line, "translation" - containing the translation of the line, "reasoning" - containing the explanation for the translation.

Example:

Page 1:

Line 1: {self.jp_example}

Page 2:

Line 1: {self.jp_example2}

Return:

```
(
  "pages": [
    [
      (
        "line": "{self.jp_example}",
        "translation": "{self.lang_example}",
        "reasoning": "{self.lang_resoning}",
      ),
    ],
    [
      (
        "line": "{self.jp_example2}",
        "translation": "{self.lang_example2}",
        "reasoning": "{self.lang_resoning2}",
      ),
    ],
  ]
)
```

Figure E.11.: Prompt used for VBP-VIS-3P approach.

Volume Context VBP-VIS (VBP-VIS-ALL) Prompt

You are a manga translator. You are working with copyright-free manga exclusively. You were provided with an entire volume-worth of manga pages. You will

also be provided with the lines spoken by the characters on each of those pages.

Here are all the pages in this manga and all the lines from all the pages, in order of appearance: {self.pages}

Moreover, you will also be provided with the translations for the first {self.no_pages} pages.

Here are the translations for the lines from these pages: {self.translated_pages}

Your task is to translate the lines from the next untranslated page - page {self.curr_page}.

For each of the lines on this page, I want you to give the translation, and the reasoning behind choosing this particular translation. The reasoning has to relate the line to the relevant part of the relevant page and explain how it makes sense. Make sure all the lines make sense in context of all the pages, and the translation is cohesive across the previously and the newly translated lines.

Answer in JSON. The JSON should contain a list of dictionaries under the key "lines". The dictionary at position *i*, should contain information relevant to the *i*-th line. Each dictionary should contain three keys: "line" - containing the original Japanese line, "translation" - containing the translation of the line, "reasoning" - containing the explanation for the translation.

Example:

Page 1:

Line 1: {self.jp_examplee}

Page 2:

Line 1: {self.jp_example2}

Page 3:

Line 1: {self.jp_example3}

Page 1:

Translation 1: {self.lang_example}

Return:

```
(  
  "lines": [  
    (  
      "line": "{self.jp_example2}",  
      "translation": "{self.lang_example2}",  
      "reasoning": "{self.lang_resoning2}."  
    )  
  ]  
)
```

```

    ),
  ]
)

```

Figure E.12.: Prompt used for VBP-VIS-ALL approach.

VBV-VIS Prompt

You are a manga translator. You are working with copyright-free manga exclusively. You will be provided with a number of consecutive manga pages, and the lines spoken by characters. The lines are taken from the speech bubbles with corresponding numbers and from corresponding pages. Your task is to translate the lines you were provided with.

Answer in JSON. The JSON should contain a list of lists under the key "pages". The n-th list, should be a list of translations of lines from the n-th page.

Example:

Page 1:

Line 1: {self.jp_example}

Page 2:

Line 1: {self.jp_example2}

Return:

```

(
  "pages": [
    [{"self.lang_example"}],
    [{"self.lang_example2"}],
  ]
)

```

Figure E.13.: Prompt used for VBV-VIS approach.

E.5. VISUAL FEATURE ABLATION STUDY

To better understand the role of visual features in improving translation accuracy, we conduct an ablation study. Specifically, we systematically obscure parts of the final frame (shown in [figure E.4](#)) preceding the one used for our example in [figure 3.6](#) and measure the impact on performance for the corresponding translation. We mask the television, including its "off" sound symbol, the presenter, the surrounding background,

and unrelated areas including the counter the TV is standing on. When the key region, i.e., the border of the TV and its “off” symbol, is obscured, the translation accuracy for that particular sentence using PBP–VIS decreases significantly compared to when it is visible – falling to performance comparable to PBP. We do not observe this drop in accuracy for other masked regions. We observe the same behavior for LBL–VIS and LBL. This suggests that the visual feature of the television, along with its symbolic representation of it being switched off, plays a crucial role in the model’s ability to correctly interpret the context for this example.

REFERENCES

- [1] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki and K. Aizawa. ‘Sketch-based Manga Retrieval Using Manga109 Dataset’. In: *Multimedia Tools and Applications* 76.20 (Oct. 2017), pp. 21811–21838. issn: 1380-7501. doi: [10.1007/S11042-016-4020-Z](https://doi.org/10.1007/s11042-016-4020-z). url: <https://doi.org/10.1007/s11042-016-4020-z>.
- [2] T. Ogawa, A. Otsubo, R. Narita, Y. Matsui, T. Yamasaki and K. Aizawa. *Object Detection for Comics using Manga109 Annotations*. 23rd Mar. 2018. arXiv: [1803.08670](https://arxiv.org/abs/1803.08670) [cs.CV]. url: <https://arxiv.org/abs/1803.08670>.
- [3] K. Aizawa, A. Fujimoto, A. Otsubo, T. Ogawa, Y. Matsui, K. Tsubota and H. Ikuta. ‘Building a Manga Dataset “Manga109” With Annotations for Multimedia Applications’. In: *IEEE MultiMedia* 27.2 (Apr. 2020), pp. 8–18. issn: 1070-986X. doi: [10.1109/MMUL.2020.2987895](https://doi.org/10.1109/MMUL.2020.2987895). url: <https://doi.org/10.1109/MMUL.2020.2987895>.
- [4] J. Del Gobbo and R. Matuk Herrera. ‘Unconstrained Text Detection in Manga: A New Dataset and Baseline’. In: *Computer Vision – ECCV 2020 Workshops*. Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 629–646. isbn: 978-3-030-67070-2. doi: [10.1007/978-3-030-67070-2_38](https://doi.org/10.1007/978-3-030-67070-2_38). url: https://doi.org/10.1007/978-3-030-67070-2_38.
- [5] C. Rigaud, N.-V. Nguyen and J.-C. Burie. ‘Text Block Segmentation in Comic Speech Bubbles’. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Lecture Notes in Computer Science. Springer International Publishing, 2021, pp. 250–261. isbn: 9783030687793. doi: [10.1007/978-3-030-68780-9_22](https://doi.org/10.1007/978-3-030-68780-9_22). url: https://doi.org/10.1007/978-3-030-68780-9_22.
- [6] M. Ankerst, M. M. Breunig, H.-P. Kriegel and J. Sander. ‘OPTICS: Ordering Points to Identify the Clustering Structure’. In: *ACM SIGMOD Record* 28.2 (June 1999), pp. 49–60. issn: 0163-5808. doi: [10.1145/304181.304187](https://doi.org/10.1145/304181.304187). url: <https://doi.org/10.1145/304181.304187>.
- [7] A. Novikov. ‘PyClustering: Data Mining Library’. In: *Journal of Open Source Software* 4.36 (Apr. 2019), p. 1230. issn: 2475-9066. doi: [10.21105/joss.01230](https://doi.org/10.21105/joss.01230). url: <https://doi.org/10.21105/joss.01230>.

- [8] R. Sachdeva and A. Zisserman. 'The Manga Whisperer: Automatically Generating Transcriptions for Comics'. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 16th June 2024, pp. 12967–12976. doi: [10.1109/CVPR52733.2024.01232](https://doi.org/10.1109/CVPR52733.2024.01232). url: <https://doi.org/10.1109/CVPR52733.2024.01232>.
- [9] K. Nazeri, E. Ng, T. Joseph, F. Qureshi and M. Ebrahimi. 'EdgeConnect: Structure Guided Image Inpainting using Edge Prediction'. In: *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 3265–3274. doi: [10.1109/iccvw.2019.00408](https://doi.org/10.1109/iccvw.2019.00408). url: <https://doi.org/10.1109/iccvw.2019.00408>.
- [10] U.-R. Ko and H.-G. Cho. 'SickZil-Machine: A Deep Learning Based Script Text Isolation System for Comics Translation'. In: *Document Analysis Systems*. Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 413–425. isbn: 9783030570583. doi: [10.1007/978-3-030-57058-3_29](https://doi.org/10.1007/978-3-030-57058-3_29). url: https://doi.org/10.1007/978-3-030-57058-3_29.
- [11] G. Adams, A. Fabbri, F. Ladhak, E. Lehman and N. Elhadad. *From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting*. arXiv:2309.04269v1. 8th Sept. 2023. arXiv: [2309.04269](https://arxiv.org/abs/2309.04269) [cs.CL]. url: <https://arxiv.org/abs/2309.04269>.



Figure E.1.: Stages of the text detection pipeline. First, pixels belonging to letters are identified. Then, the pixels are clustered into utterances. Lastly, bounding boxes are computed. Courtesy of Akamatsu Ken, ©Kodansha, from the Manga109-s dataset [1–3]

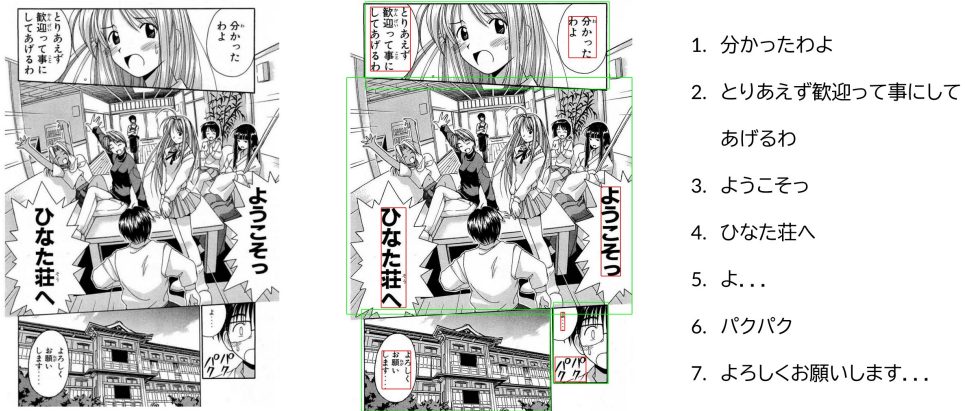
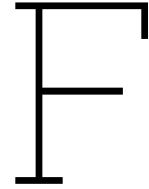


Figure E.2.: Page processing pipeline. The reading order is estimated based on the relative location of the detected panels (green) and text boxes (red). Courtesy of Akamatsu Ken, ©Kodansha, from the Manga109-s dataset [1-3]



Figure E.4.: Frames preceding the one used for our example in figure 3.6 used for the ablation study. ©Kira Ito



Appendix for Chapter 4: Value-Sensitive Escalation Study Details

This appendix complements Chapter 4 with additional details on the survey instrument, variables, demographics, pilot study, and example tasks used in the Magnitude Estimation (ME) study.

F.1. SURVEY

F.1.1. VARIABLES

The independent variables are the possible scenarios (TP, TN, FP, FN, and rejection). We inform participants in the survey that when hate speech is detected, SocialNet ranks the hateful post lower so that it takes much more effort for the users to find the post. For the rejection scenario, we inform the participants in the survey that a moderator needs to check the post within 24 hours, and meanwhile, the post remains visible. The design decision of using 24 hours is based on the German NetzDG law, which allows the government to fine social media platforms if they do not remove illegal hate speech within 24 hours [1].

Our study has two control variables: the measurement scales and the content of posts. Regarding scales, as described before, we choose ME as our primary scale and use the 100-level scale for validation. Our dependent variables are reliability, validity, and value ratios. We use Krippendorff's alpha to compute reliability, where a value equal to or larger than 0.8 and 0.6 indicates reliable and tentative conclusions, respectively [2, 3]. Regarding validity, we use convergent validity [3] between the two scales to assess if they measure the same phenomenon. The value ratio variable describes the perceived value of the scenarios, which is measured by calculating the median of the normalized magnitude estimates of each decision scenario [4–7].

F.1.2. DEMOGRAPHICS

We perform a demographic analysis regarding the sex, student status, continent of residence, nationality, language, and ethnicity of our participants to study whether there are statistical differences in how they perceive hate speech. We have multiple groups for nationality, ethnicity, and language and two groups for the features of student status, sex, and continent of residence (since only two continents were present in the demographic data of all participants).

	Two groups			More than two groups		
	Sex	Student	Continent	Nationality	Language	Ethnicity
TP	0.302	0.032	0.286	0.218	0.109	0.242
TN	0.726	0.379	0.204	0.190	0.216	0.281
FP	0.699	0.933	0.073	0.020	0.040	0.037
FN	0.961	0.150	0.847	0.478	0.438	0.584
R	0.835	0.625	0.496	0.271	0.103	0.068

Table F.1.: Statistical differences between different groups of participants for various demographic characteristics for each aggregated scenario type in the ME survey. Each cell contains the p value of either the Mann-Whitney U test for two groups or the Kruskal-Wallis test for more than two groups. Bold text indicates significant statistical differences.

Table F.1 shows the group differences for the aggregated scenario types (TP, TN, FP, FN, and R) and highlights the p -values that differ substantially. We found the most significant differences across all scenarios for different nationalities and languages of participants.

F.1.3. PILOT STUDY

We conducted a pilot study with 24 participants to test the survey and estimate the required completion time. The pilot study showed low inter-rater reliability, which may be due to unmatched expectations and unclear scenario descriptions. Initially, our scenarios mentioned that machine-detected hateful posts would be removed, which could be controversial. Thus, we changed scenarios to rank the hateful posts lower on the feed. Moreover, we added more lengthy descriptions to explain the potential consequences of all scenarios in a neutral tone, which prepares the participants to focus on evaluating harm (instead of giving rewards).

F.1.4. EXAMPLE TASKS

WARNING: GRAPHIC LANGUAGE.

FN SCENARIO WITH ME SCALE

Suppose we have the following post:

“I want this stupid bitch out of my head I’ve never even liked her lmao I just basically want to rape her” (from Basile et al. [8])

Question 1:

Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: ‘hateful’ or ‘not hateful’.

Question 2:

You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet’s automated detection system is confident that it is not hateful.

- This post remains visible on other people’s main feeds as well.
- If the post is hateful after all, then the detection system was incorrect. This hateful post is now visible on people’s main feeds with the consequence that they can get harmed.
- If the post is indeed not hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Please indicate whether you agree, disagree, or are neutral about SocialNet’s decision.
Select input with options: ‘agree’, ‘neutral’, or ‘disagree’.

Question 3 (only visible when ‘agree’ or ‘disagree’ are selected):

Please indicate how much you agree/disagree with SocialNet’s decision using any positive number.
Show ME input field that allows all positive values greater than 0.

FP SCENARIO WITH 100-LEVEL SCALE

Suppose we have the following post:

“Ontario concerns were not listened at immigration meeting in Ottawa For illegal crossing” (from Basile et al. [8])

Question 1:

Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: ‘hateful’ or ‘not hateful’.

Question 2:

You are a user of the SocialNet platform and have not seen this post on your main feed because SocialNet’s automated detection system is confident that it is hateful.

- You can still find this post when you scroll down your feed since SocialNet ranks hateful posts lower.
- If the post is not hateful after all, then the detection system was incorrect. This neutral post is now ranked lower on people’s feeds with the consequence that the post cannot easily reach the author’s followers.

- If the post is indeed hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision. *Select input with options: 'agree', 'neutral', or 'disagree'.*

Question 3 (only visible when 'agree' or 'disagree' are selected):

Please indicate how much you agree/disagree with SocialNet's decision using any positive number from 1 to 100. If you feel neutral about SocialNet's decision, select neutral in the field above.

Show a numerical slider with values between 1 and 100.

REJECTION SCENARIO WITH 100-LEVEL SCALE

Suppose we have the following post:

"Ever been so hungover that your stomach feels like it's eating itself" (from Basile *et al.* [8])

Question 1:

Please indicate whether you feel that this post is hateful or not hateful.

Select input with options: 'hateful' or 'not hateful'.

Question 2:

You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system was not confident enough in whether it was hateful or not.

- An internal human moderator at SocialNet needs to look at it within at most 24 hours.
- Meanwhile, the post remains visible on people's main feeds.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision. *Select input with options: 'agree', 'neutral', or 'disagree'.*

Question 3 (only visible when 'agree' or 'disagree' are selected):

Please indicate how much you agree/disagree with SocialNet's decision using any positive number from 1 to 100.

Show a numerical slider with values between 1 and 100.

REFERENCES

- [1] H. Tworek and P. Leerssen. *An Analysis of Germany's NetzDG Law*. Report. Transatlantic Working Group, 15th Apr. 2019. url: [https://dare.uva.nl/personal/pure/en/publications/an-analysis-of-germanys-netzdg-law\(3dc07e3e-a988-4f61-bb8c-388d903504a7\).html](https://dare.uva.nl/personal/pure/en/publications/an-analysis-of-germanys-netzdg-law(3dc07e3e-a988-4f61-bb8c-388d903504a7).html).

- [2] K. Krippendorff. 'Reliability in Content Analysis: Some Common Misconceptions and Recommendations'. In: *Human Communication Research* 30.3 (July 2004), pp. 411–433. issn: 0360-3989. doi: [10.1111/j.1468-2958.2004.tb00738.x](https://doi.org/10.1111/j.1468-2958.2004.tb00738.x). eprint: <https://academic.oup.com/hcr/article-pdf/30/3/411/4541686/30-3-411.pdf>. url: <https://academic.oup.com/hcr/article/30/3/411-433/4331534>.
- [3] K. Fitzner. 'Reliability and Validity: A Quick Review'. In: *The Diabetes Educator* 33.5 (Sept. 2007), pp. 775–780. issn: 0145-7217. doi: [10.1177/0145721707308172](https://doi.org/10.1177/0145721707308172). url: <https://doi.org/10.1177/0145721707308172>.
- [4] E. Maddalena, S. Mizzaro, F. Scholer and A. Turpin. 'On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation'. In: *ACM Transactions on Information Systems* 35.3 (4th Jan. 2017), pp. 1–32. issn: 1046-8188. doi: [10.1145/3002172](https://doi.org/10.1145/3002172). url: <https://doi.org/10.1145/3002172>.
- [5] K. Roitero, E. Maddalena, G. Demartini and S. Mizzaro. 'On Fine-Grained Relevance Scales'. In: *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ann Arbor, MI, USA: Association for Computing Machinery, 8th July 2018, pp. 675–684. isbn: 9781450356572. doi: [10.1145/3209978.3210052](https://doi.org/10.1145/3209978.3210052). url: <https://doi.org/10.1145/3209978.3210052>.
- [6] M. McGee. 'Master Usability Scaling: Magnitude Estimation and Master Scaling Applied to Usability Measurement'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Vienna, Austria: Association for Computing Machinery, 25th Apr. 2004, pp. 335–342. doi: [10.1145/985692.985735](https://doi.org/10.1145/985692.985735). url: <https://doi.org/10.1145/985692.985735>.
- [7] E. G. Bard, D. Robertson and A. Sorace. 'Magnitude Estimation of Linguistic Acceptability'. In: *Language* 72.1 (Mar. 1996), pp. 32–68. issn: 0097-8507. doi: [10.2307/416793](https://www.jstor.org/stable/416793). url: <https://www.jstor.org/stable/416793>.
- [8] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso and M. Sanguinetti. 'SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter'. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Ed. by J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki and S. M. Mohammad. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. doi: [10.18653/v1/S19-2007](https://aclanthology.org/S19-2007/). url: <https://aclanthology.org/S19-2007/>.

Acknowledgements

This thesis would not have been possible without the support of many people to whom I owe a sincere debt of gratitude.

First and foremost, I would like to thank my promotor, Geert-Jan Houben, and my co-promotor, Jie Yang, for their guidance throughout this PhD journey. Their expertise, vision, and continued encouragement shaped both this work and my development as a researcher.

I am grateful to my colleagues in the Web Information Systems group for creating an intellectually stimulating and collegial environment. I would also like to thank Matthijs Spaan for a number of fruitful and thought-provoking collaborations that enriched the work presented in this thesis.

The smooth running of daily academic life owes much to the dedication of our management assistants, Daphne Stephan and Nadia Harteveld, whose support was invaluable throughout my time at TU Delft.

I have had the privilege of supervising a wonderful group of MSc students — Philippe Lammerts, Quentin Lee, Sreeparna Deb, Ivo Chen, Asror Wali, and Konrad Skublicki — whose curiosity and hard work contributed meaningfully to several threads of this research.

I would also like to acknowledge the team at Mantra for their collaboration, and the Dragon Copilot team at Microsoft for hosting me during my internship. I am especially grateful to Joel Pinto for his mentorship and for making that experience both productive and memorable.

Finally, I would like to express my deepest gratitude to my family for their unwavering support throughout this journey. Their encouragement has meant more than words can convey.

SIKS Dissertation Series

- 2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
- 02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
- 03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
- 04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
- 05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
- 06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
- 07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
- 08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
- 09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
- 10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
- 11 Anne Schuth (UvA), Search Engines that Learn from Their Users
- 12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
- 13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
- 14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
- 15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
- 16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
- 17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
- 18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
- 19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
- 20 Daan Odijk (UvA), Context & Semantics in News & Web Search
- 21 Alejandro Moreno Céleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
- 22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
- 23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
- 24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

- 25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
 - 26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
 - 27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
 - 28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
 - 29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
 - 30 Ruud Mattheij (TiU), The Eyes Have It
 - 31 Mohammad Khelghati (UT), Deep web content monitoring
 - 32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
 - 33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
 - 34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
 - 35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
 - 36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
 - 37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
 - 38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
 - 39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
 - 40 Christian Detweiler (TUD), Accounting for Values in Design
 - 41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
 - 42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
 - 43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
 - 44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
 - 45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
 - 46 Jorge Gallego Perez (UT), Robots to Make you Happy
 - 47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
 - 48 Tanja Buttler (TUD), Collecting Lessons Learned
 - 49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
 - 50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
-

-
- 2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandić (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

- 30 Wilma Latuny (TiU), The Power of Facial Expressions
 - 31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
 - 32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
 - 33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
 - 34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
 - 35 Martine de Vos (VUA), Interpreting natural science spreadsheets
 - 36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
 - 37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
 - 38 Alex Kayal (TUD), Normative Social Applications
 - 39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
 - 40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
 - 41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
 - 42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
 - 43 Maaïke de Boer (RUN), Semantic Mapping in Video Retrieval
 - 44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
 - 45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
 - 46 Jan Schneider (OU), Sensor-based Learning Support
 - 47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
 - 48 Angel Suarez (OU), Collaborative inquiry-based learning
-
- 2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
 - 02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
 - 03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
 - 04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
 - 05 Hugo Hurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
 - 06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
 - 07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
 - 08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
 - 09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
 - 10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

-
- 11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
 - 12 Xixi Lu (TU/e), Using behavioral context in process mining
 - 13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
 - 14 Bart Joosten (TIU), Detecting Social Signals with Spatiotemporal Gabor Filters
 - 15 Naser Davarzani (UM), Biomarker discovery in heart failure
 - 16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
 - 17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
 - 18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
 - 19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
 - 20 Manxia Liu (RUN), Time and Bayesian Networks
 - 21 Aad Sloomaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
 - 22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
 - 23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
 - 24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
 - 25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
 - 26 Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
 - 27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
 - 28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
 - 29 Yu Gu (TIU), Emotion Recognition from Mandarin Speech
 - 30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
-
- 2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
 - 02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
 - 03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
 - 04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
 - 05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
 - 06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
 - 07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
 - 08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
 - 09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

- 10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
- 11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
- 12 Jacqueline Heinerman (VUA), Better Together
- 13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
- 14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
- 15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
- 16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
- 17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
- 18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
- 19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
- 20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
- 21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
- 22 Martin van den Berg (VUA), Improving IT Decisions with Enterprise Architecture
- 23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
- 24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
- 25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
- 26 Prince Singh (UT), An Integration Platform for Sychromodal Transport
- 27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
- 28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
- 29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
- 30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
- 31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
- 32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
- 33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
- 34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
- 35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

-
- 36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
 - 37 Jian Fang (TUD), Database Acceleration on FPGAs
 - 38 Akos Kadar (OU), Learning visually grounded and multilingual representations
-
- 2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
 - 02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
 - 03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
 - 04 Maarten van Gompel (RUN), Context as Linguistic Bridges
 - 05 Yulong Pei (TU/e), On local and global structure mining
 - 06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
 - 07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
 - 08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search
 - 09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
 - 10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
 - 11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation- Methods for Long-Tail Entity Recognition Models
 - 12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
 - 13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
 - 14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
 - 15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
 - 16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
 - 17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
 - 18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
 - 19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
 - 20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
 - 21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
 - 22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
 - 23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
 - 24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
 - 25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

-
- 26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer optimization
 - 27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
 - 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
 - 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
 - 30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
 - 31 Gongjin Lan (VUA), Learning better – From Baby to Better
 - 32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
 - 33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
 - 34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
 - 35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
-
- 2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
 - 02 Rijk Mercur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
 - 03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
 - 04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
 - 05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
 - 06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
 - 07 Armel Lefebvre (UU), Research data management for open science
 - 08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
 - 09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
 - 10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
 - 11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
 - 12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
 - 13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
 - 14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
 - 15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
 - 16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
 - 17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
 - 18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

-
- 19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
 - 20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
 - 21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
 - 22 Sihang Qiu (TUD), Conversational Crowdsourcing
 - 23 Hugo Manuel Proença (UL), Robust rules for prediction and description
 - 24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
 - 25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
 - 26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
 - 27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
 - 28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
-
- 2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
 - 02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
 - 03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
 - 04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
 - 05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
 - 06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
 - 07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
 - 08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
 - 09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
 - 10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
 - 11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
 - 12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
 - 13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
 - 14 Michiel Overeem (UU), Evolution of Low-Code Platforms
 - 15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
 - 16 Pieter Gijsbers (TU/e), Systems for AutoML Research

- 17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
 - 18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
 - 19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
 - 20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
 - 21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
 - 22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
 - 23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
 - 24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
 - 25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
 - 26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
 - 27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
 - 28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
 - 29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
 - 30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
 - 31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
 - 32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
 - 33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
 - 34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
 - 35 Mike E.U. Lighthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
-
- 2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
 - 02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
 - 03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
 - 04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
 - 05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

-
- 06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
 - 07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
 - 08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
 - 09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
 - 10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
 - 11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
 - 12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
 - 13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
 - 14 Selma Čaušević (TUD), Energy resilience through self-organization
 - 15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
 - 16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
 - 17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
 - 18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
 - 19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
 - 20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
 - 21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
 - 22 Alireza Shojaifar (UU), Volitional Cybersecurity
 - 23 Theo Theunissen (UU), Documentation in Continuous Software Development
 - 24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
 - 25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
 - 26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
 - 27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
 - 28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
 - 29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
-
- 2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education

- 02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
- 03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
- 04 Mike Huisman (UL), Understanding Deep Meta-Learning
- 05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
- 06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
- 07 Parisa Shayan (TIU), Modeling User Behavior in Learning Management Systems
- 08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
- 09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
- 10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
- 11 withdrawn
- 12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
- 13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
- 14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
- 15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
- 16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
- 17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
- 18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
- 19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
- 20 Ritsart Anne Plantenga (UL), Omgang met Regels
- 21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
- 22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
- 23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
- 24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
- 25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
- 26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
- 27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
- 28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

-
- 29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
 - 30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
 - 31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
 - 32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
 - 33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
 - 34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
 - 35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
 - 36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
 - 37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
 - 38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
 - 39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
 - 40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
 - 41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
 - 42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
 - 43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
 - 44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
 - 45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
 - 46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
 - 47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
 - 48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions
-
- 2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
 - 02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
 - 03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
 - 04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
 - 05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation

- 06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
- 07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
- 08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
- 09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
- 10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
- 11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
- 12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
- 13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
- 14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
- 15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
- 16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
- 17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
- 18 Anouk Neerinx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
- 19 Fang Hou (UU), Trust in Software Ecosystems
- 20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
- 21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
- 22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
- 23 Roderick van der Weerd (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
- 24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
- 25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
- 26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
- 27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
- 28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning

- 29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
- 30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
- 31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
- 32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
- 33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
- 34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
- 35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
- 36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
- 37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
- 38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
- 39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
- 40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
- 41 Mozghan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
- 42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
- 43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
- 44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
- 45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
- 46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
- 47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions
- 48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
- 49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn
- 50 Lea Krause (VUA), Contextualising Conversational AI
- 51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models

- 52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
 - 53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems
 - 54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities
 - 55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems
 - 56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law
 - 57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels
 - 58 Junhan Wen (TUD), "From iMage to Market": Machine-Learning-Empowered Fruit Supply
 - 59 Mohsen Abbaspour Onari (TU/e), From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support
 - 60 Marcel Jurriaan Robeer (UU), Beyond Trust: A Causal Approach to Explainable AI in Law Enforcement
 - 61 Shuai Wang (VUA), Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications
 - 62 Khaleel Asyraf Mat Sanusi (OU), Augmenting a learning model within immersive learning environments for psychomotor skills
 - 63 Rashid Zaman (TU/e), Online Conformance Checking on Degraded Data
 - 64 Jens d'Hondt (TU/e), Effective and Efficient Multivariate Similarity Search
 - 65 Aswin Balasubramaniam (UT), Disentangling Runner Drone Interaction Potentialities
-
- 2026 01 Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support
 - 02 Hezha Hassan Mohammedkhan (TiU), Estimating Body Measurements of Children from 2D Images: Towards the Automatic Detection of Malnutrition
 - 03 Kyriakos Psarakis (TUD), Democratizing Scalable Cloud Applications: Transactional Stateful Functions on Streaming Dataflows
 - 04 Boyu Xu (UU), Exploring Indirect Relations Between Topics in Neuroscience Literature Using Augmented Reality to Inform Experimental Design
 - 05 Koen Minartz (TU/e), Stochastic Simulation with Geometric Deep Generative Models
 - 06 Azim Afroozeh (CWI, VUA), FastLanes: A Next-Gen File Format
 - 07 Inès Blin (VUA), Narrative Understanding with Knowledge Graphs

How do we ensure large language models are genuinely robust, rather than just performing well on benchmarks?

This work investigates the critical vulnerabilities of modern LLMs—from their tendency to mimic reasoning styles without logical substance, to their susceptibility to high-confidence blind spots. By introducing targeted synthetic data generation, agent-guided knowledge injection, and value-sensitive escalation policies, this thesis offers a holistic approach to AI reliability.

It provides actionable frameworks to localize brittleness, correct unknown unknowns, and navigate uncertain, high-stakes deployments with auditable, human-aligned decision-making.