



# Evaluating the Ability of Large Language Models to Classify Scientific Papers as Empirical or Theoretical using the NeurIPS Checklist

**Adam Wielinga**

**Supervisor(s): David M.J Tax, Hayley Hung, Nergis Tömen, Chenxu Hao**

<sup>1</sup>EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,  
In Partial Fulfilment of the Requirements  
For the Bachelor of Computer Science and Engineering  
June 18, 2026

Name of the student: <Adam Wielinga>

Final project course: CSE3000 Research Project

Thesis committee: David M.J Tax, Hayley Hung, Nergis Tömen, Chenxu Hao, Klaus Hildebrandt

## Abstract

As machine learning conferences such as NeurIPS expand rapidly, the manual classification and evaluation of responsible research checklists impose a significant burden on reviewers. This study investigates the ability of Large Language Models (LLMs) to automatically classify research papers as empirical, theoretical, or hybrid, and to extract checklist compliance data. Using a dataset of publicly available NeurIPS papers, we designed an automated pipeline and evaluated its outputs against a human-annotated ground truth. Our results demonstrate that the LLM achieves high accuracy in the core classification task, reliably distinguishing the papers core methodology by identifying clear structural indicators like mathematical proofs and benchmark datasets. Furthermore, the model excels at extracting objective checklist elements, performing well on close-ended extraction tasks that rely on clear structural indicators. However, performance noticeably decreased on structurally scattered or subjective criteria, such as broader impacts and the declaration of AI usage. This drop highlights a limitation in the model’s broader reading comprehension, as it struggles to merge contextual information without explicit headers. Notably, this automated failure closely mirrors human task ambiguity, as these exact subjective items also generated the lower inter-annotator agreement among human annotators. Conclusively, while LLMs provide a highly consistent baseline for classifying paper typologies and extracting explicit methodological data, their reliance on structural cues indicates they should serve as assistive screening tools rather than autonomous evaluators in academic peer review.

## 1 Introduction

The fast acceleration of research in machine learning and artificial intelligence has resulted in a high volume of paper submissions to major academic conferences. This exponential growth places a huge strain on the traditional peer-review process, often leading to reviewer fatigue and even possible inconsistent evaluation standards [8,9]. To address these challenges, the academic community has looked toward Large Language Models (LLMs) to assist in or partially automate the peer-review process. While these recent studies do demonstrate the potential of LLMs to summarize texts and identify basic technical errors, deploying these models as reliable automated reviewers requires a rigorous understanding of their analytical capabilities, specifically their ability to extract structured information and classify complex methodologies.

To maintain high standards of scientific integrity, conferences such as NeurIPS require authors to complete a 16 question checklist detailing matters like experimental setups, theoretical claims, and limitations. An important step in evaluating these papers is correctly identifying their core methodology specifically, distinguishing whether a paper is primarily empirical, theoretical, or a hybrid of both, since different evaluation criteria might apply to the different fundamental elements of each paper type. While prior exploratory studies show that LLMs perform great on isolated, closed-ended checklist extraction tasks, a critical research gap remains: it is currently unknown whether models can reliably use this structured data to classify the dominating methodological nature of a paper. Because classification requires moving beyond simple text extraction to broader logical reasoning an area where LLMs historically struggle this capability must be evaluated against human expert standards.

To address this gap, this study investigates the main research question: To what extent

can Large Language Models reliably extract structured checklist information to classify the methodological nature of scientific papers compared to human experts? We decompose this main objective into three targeted sub-questions:

- How accurately does the LLM classify scientific papers typologies compared to human annotations?
- How well does the LLM’s generated reasoning and granular checklist extraction align with human-annotated rationale?
- How consistent are human experts when annotating these papers based on the NeurIPS checklist?

To answer these questions, we designed a comparative, two-tier evaluation framework. We compiled a dataset of approximately 80 accepted papers from the NeurIPS 2024 proceedings. A team of five independent human annotators manually labeled these papers based on the NeurIPS checklist to establish a baseline for human consistency. Additionally, a subset of these papers were manually labeled (empirical, theoretical or hybrid) as the ground-truth for the classification task, since the paper type is not included in the NeurIPS Checklist. We subsequently deployed the Gemini 3.0 Flash model to perform the exact same checklist extraction and paper classification tasks. Our evaluation first compares the LLM’s high-level classification accuracy and granular reasoning against the human ground truth subset, and then scales up to an exploratory analysis of the LLM’s checklist and classification behavior across the complete dataset.

The remainder of this paper is structured as follows: Section 2 reviews related work concerning automated peer review and information extraction. Section 3 formalizes the problem definition and details the research methodology. Section 4 outlines the approach and design choices for the LLM pipeline. Section 5 presents the experimental setup and the results addressing each sub-question. Section 6 discusses the ethical implications and reproducibility of the research. Section 7 interprets the findings, acknowledging limitations and exploring implications for the peer-review process. Finally, Section 8 concludes the study and outlines directions for future work.

## 2 Related Work

This section reviews prior literature concerning LLM-based information extraction, peer review automation, the NeurIPS evaluation framework, and the specific research gap our study addresses.

### 2.1 LLM-Based Information Extraction and Summarization

Recent studies demonstrate the efficiency of LLMs in extracting structured information from complex documents. For example, research on the evaluation of the use of LLMs in systematic reviews shows that while models cannot fully automate paper selection autonomously yet, they do serve as effective assisting tools [1]. Semi-automated approaches utilizing LLMs have been shown to not only correctly include up to 82.7% of relevant papers but also correctly exclude 92.2% of irrelevant ones [1]. We build upon this concept by applying similar extraction methods to the machine learning domain, specifically targeting structured academic checklists.

## 2.2 Automation in Peer Review

The academic community has shown a big increase in the exploration of the automation of peer review tasks using LLMs. Comprehensive evaluations show that while LLMs are generally helpful, they are still weak at long paper processing and providing critical feedback comparable to those of human reviewers [2]. Furthermore, studies automatically evaluating the paper reviewing capability of LLMs reveal different limitations [3]. When compared against expert-generated reviews, LLMs show bias towards technical validity, overlooking novelty assessment, and produce poor acceptance decisions [3]. These findings highlight that LLMs do not yet accurately replicate expert human judgment, further pointing out the need for more targeted and objective evaluation tasks rather than long open review generation

## 2.3 NeurIPS Evaluation Framework Overview

To standardize the evaluation of scientific papers and improve publication quality and reproducibility of papers, major conferences like NeurIPS employed a structured checklist framework. Prior work has shown that releasing code alone is not always sufficient for determining true reproducibility, highlighting the need for structured checklists [7]. This program was introduced to improve the standards across the community for how machine learning research is conducted, communicated, and evaluated [4]. The NeurIPS checklist requires authors to explicitly declare specific methodological details, including experimental setups, computational resources, limitations, and theoretical assumptions, which serves to promote the use of their experimental workflows and reduce unintentional errors [4].

Prior studies have tested LLMs on reviewing tasks within this framework [5]. This research shows that LLMs perform well on specific closed-ended extraction tasks, but struggle with broader and more subjective comparisons [5]. Furthermore, the structured and categorical nature of the NeurIPS checklist establishes it as an ideal and useful framework for systematically testing LLM extraction capabilities.

## 2.4 Research Gap

Despite the increase of literature on automated peer reviews, a gap remains regarding the ability of LLMs to classify the fundamental nature of a paper based on extracted features. While existing work focuses heavily on generating full reviews [2] or identifying specific technical flaws [3], it remains rather unclear whether LLMs can correctly distinguish between empirical and theoretical research using a structured framework like the NeurIPS checklist. This study directly addresses this gap by evaluating the classification accuracy and reasoning consistency of the Gemini 3.0 Flash model against human-annotated ground truth.

# 3 Problem Definition & Methodology

This section formalizes the extraction task and outlines our research strategy.

## 3.1 Problem definition

The main problem this study addresses is the automated evaluation and classification of scientific texts. Specifically, we define the task as evaluating the ability of Large Language

Models (LLMs) to accurately classify scientific papers into three distinct categories: empirical, theoretical, or hybrid. This classification is performed by extracting information corresponding to the structured NeurIPS checklist.

### 3.2 Dataset Collection and Annotation

To execute this research, we utilize a dataset of around 80 papers selected from the NeurIPS 2024 proceedings. The annotation process is distributed among five independent annotators, with each individual labeling the NeurIPS checklist for roughly 20 papers. To better determine human consistency and determine how ambiguous the annotation task is, we established an overlapping subset of papers across all annotators, which we evaluate using Cohen’s kappa.

From this broader collection, we assigned manual classification labels (empirical, theoretical, or hybrid) to a specific core ground-truth subset of 20 papers. This manually verified subset provides the strict baseline benchmark required to measure the true classification performance of the model. Additionally each of these labeled papers were also provided with a justification for the classification choice to compare against the justifications generated by the LLM.

### 3.3 Evaluation Framework And Analysis

With the dataset established, we structure our analysis using a comparative, two-tier evaluation framework:

- **Tier 1 (Ground Truth Evaluation):** We measure the classification performance of the LLM against the 20-paper manually labeled ground-truth subset, relying on Accuracy and Macro-F1 score metrics.
- **Tier 2 (Exploratory Analysis):** We conduct a dataset-wide exploratory analysis on the complete set of 80 papers to examine the distribution of predicted paper types and any possible checklist response patterns made by the LLM. This tier investigates probable systematic tendencies in the model’s behavior such as class over representation or under representation and identifies any potential relationships between specific checklist items and predicted paper types.

To address the second sub-question regarding model reliability and reasoning, we analyze and compare the justification generated by the LLM against the human-annotated justification for the ground truth subset. Furthermore we analyze performance at checklist extraction to evaluate how well its logic and data extraction aligns with humans. Finally to address the third sub-question regarding task ambiguity, a subset of the annotated papers overlaps across all annotators. We will evaluate this overlapping subset using Cohen’s Kappa. By measuring the inter-annotator agreement, we can determine the complexity and subjectivity of the manual extraction task. This provides us with the necessary context for evaluating the LLM’s accuracy.

## 4 Approach & Design

This section details the design of the LLM pipeline and the justifications for our specific implementation choices.

## 4.1 Pipeline Components

The automated pipeline relies on an LLM to process and analyze the selected papers. The system is designed to execute three sequential tasks:

1. **Checklist Information Extraction:** The LLM processes the complete paper (provided as a PDF) to extract relevant checklist data. The extracted fields directly map to the NeurIPS checklist, including claims, limitations, theoretical assumptions, reproducibility, experimental setup, and ethical considerations. The model outputs strict categorical labels for each element: *yes*, *no*, or *n/a* (not applicable).
2. **Paper Classification:** Following the extraction phase, the LLM utilizes the processed information to classify the paper into one of the three predefined categories: empirical, theoretical, or hybrid, and provides a short justification for its response referring to the relevant parts of the paper.
3. **Automated Evaluation:** Finally, custom Python scripts parse the structured outputs generated by the LLM. These scripts calculate the core evaluation metrics (such as Accuracy and Macro-F1) and generate the comparative distribution patterns against the human-annotated ground truth.

We designed our approach with strict constraints to maximize reliability and reproducibility. For the core extraction and classification tasks, we selected the Gemini 3.0 Flash model. This decision was primarily driven by practical constraints and accessibility; the model’s API is available free of charge and provides an allowance of up to 20 requests per day, which we decided gave the best balance. As switching to a model with more RPD would risk losing necessary reasoning skills and complexity.

Furthermore, we supply the complete paper stripped of the author’s annotated checklist as a PDF to the model; this ensures comprehensive context and consistency across all analyzed samples. And prevents the LLM from directing its answers towards those of the author, which may be biased.

## 4.2 Prompt Engineering

To ensure the model executes the extraction and classification tasks with high precision, we implemented a structured prompt engineering strategy. The complete verbatim prompt utilized for these tasks is available in Appendix A. The prompt itself relies on a zero-shot framework, meaning the model is not provided with prior examples of correctly filled checklists, but rather must rely on its own reasoning capabilities and the provided instructions [10]. We start the prompt by utilizing role-prompting by instructing the model to act as an expert academic reviewer evaluating submissions for the NeurIPS conference.

We then give it the instructions on both the classification and extraction task.

- The instructions for the classification task consists of a definition for an empirical, theoretical and hybrid paper. The LLM has to adhere to these definitions when making its decision. Finally we command it to include a justification for its classification choice as well.
- For the extraction task, we pasted the full guidelines for all 16 questions of the NeurIPS checklist. The prompt then restricts the LLM’s response format to a strictly structured JSON object, to integrate the output into our python evaluation pipeline.

## 5 Experiments

### 5.1 Experimental Setup

This section details the computational environment, and the evaluation metrics we utilized to ensure the full reproducibility and validity of our findings.

#### 5.1.1 Hardware and Software Environment

We implemented and executed the pipeline using Python, to handle data parsing, evaluation metric computation, and exploratory analysis. Local scripts were all executed on a MacBook Pro configured with an Apple M2 Pro chip and 16 GB of memory. The automated checklist extraction and paper classification tasks were performed remotely using the Gemini 3.0 Flash model via its API. We selected this specific model since we believed this gave the best balance between Requests per Day and model complexity. To ensure consistency, we instructed the model to base its extractions strictly on the provided PDF texts without introducing external information. Finally, to guarantee full transparency and independent verification, we host all Python scripts, JSON annotation files, and evaluation datasets in a dedicated TU Delft GitLab repository. <sup>1</sup>

#### 5.1.2 Evaluation Metrics

We employed a broad range of statistical metrics to evaluate the model’s performance and address our core sub-questions. We quantified performance for the high-level paper classification and the granular checklist extraction tasks using Accuracy and Macro-F1 scores, which specifically accounts for the heavy class imbalances built in to certain checklist items (such as IRB approvals). The accuracy threshold chosen for the checklist extraction task is 85% , as this aligns with the upper-tier performance of LLM’s on complex text classification tasks [11]. We generated confusion matrices for the classification task for a detailed analysis. For the Tier 2 evaluation and checklist extraction across the entire dataset we utilized among others accuracy heatmaps and response distribution charts to analyze the LLM’s response tendencies in comparison to the human annotated ground truth. Finally, inter-annotator agreement was assessed using Cohen’s Kappa, with results visualized through a heatmap showing agreement scores for each annotator pair and the average score for each checklist question.

### 5.2 How accurately does the LLM classify scientific papers compared to human annotations?

Our first research question evaluates the high-level accuracy of the Gemini 3.0 Flash model in classifying scientific papers into empirical, theoretical, or hybrid methodologies against the human-annotated ground truth. This evaluation was conducted on the core subset of 20 manually verified papers.

The quantitative results demonstrate a perfect alignment between the LLM’s predictions and the human baseline. The model achieved an overall Accuracy of 100%, successfully classifying every paper in the ground-truth benchmark without a single misclassification.

---

<sup>1</sup>[https://gitlab.tudelft.nl/cse3000/analysisMLresults/-/blob/AdamsBranch/README.md?ref\\_type=heads](https://gitlab.tudelft.nl/cse3000/analysisMLresults/-/blob/AdamsBranch/README.md?ref_type=heads)

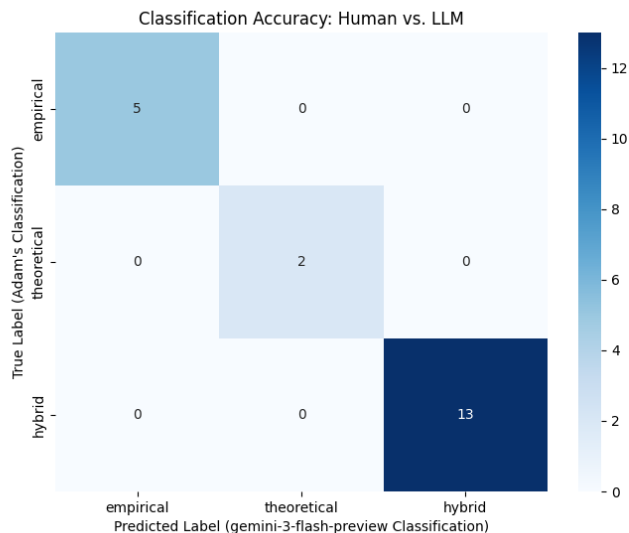


Figure 1: Classification Matrix

As shown in figure 1, the model correctly identified all 5 purely empirical papers, both of the 2 purely theoretical papers, and all 13 hybrid papers. This perfect distribution shows the model’s capability to blend broader paper structures and distinguish between the different scientific methodologies when evaluated on a categorical task. While these results show that the model is highly reliable for overall classification, a more in depth look at the checklist items and underlying reasoning is required to determine if the model arrives at these correct classifications for the right reasons, which we explore in RQ2.

### 5.3 How well does the LLM’s generated reasoning and granular checklist extraction align with human-annotated rationale?

#### 5.3.1 Granular Checklist Extraction Performance

While the classification task in RQ1 resulted a perfect score on the core subset, analyzing the model’s performance on the individual checklist fields across the dataset reveals a more realistic picture of automated information extraction using LLM’s.

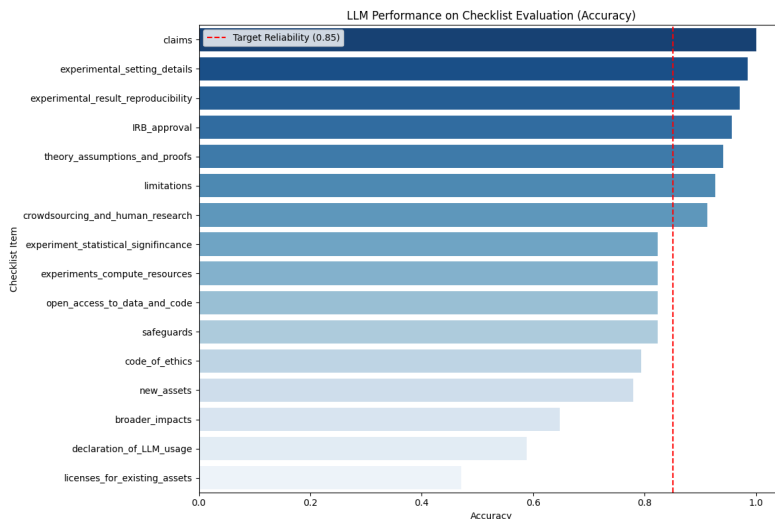


Figure 2: LLM Accuracy on NeurIPS Checklist - Complete Dataset

As illustrated in Figure 2, the model’s accuracy varies significantly depending on the nature of the checklist item. As shown, the items where the LLM achieves the highest accuracy, such as `claims`, `experimental_setting_details`, `experimental_result_reproducibility`, and `theory_assumptions_and_proofs`, are exactly the most important features necessary for the high-level classification task. These checklist questions are inherently easier for the model to spot because they are typically concentrated in dedicated, highly structured sections of the papers (e.g., an "Experiments" or "Theoretical Framework" section). As for certain checklist items such as `experiment_statistical_significance` and `experiments_compute_resources` we observed that the accuracy for these questions is affected by a limitation regarding the LLM. The specific section for these questions were occasionally found in the supplementary material, as this is a separate downloadable file the LLM does not have access to it, which results in the LLM answering with "no"

On the other hand, performance drops significantly below our Target Reliability threshold of 0.85 on compliance and more open-ended checklist items, falling to approximately 0.65 for `broader_impacts` and reaching its lowest point of 0.47 for `licenses_for_existing_assets`. When accounting for class imbalance using Macro-F1 scores as shown in Appendix C.2 this performance drop is even more pronounced.

A qualitative analysis of these errors shows a behavioral trend: the LLM heavily relies on clear structural indicators. For instance, on questions regarding `broader_impacts`, the model often generated a "no" response simply because the paper lacked a dedicated section header titled "Broader Impacts." Human annotators marked these items as "yes" because the authors discussed potential societal effects throughout the paper. This reveals a critical limitation in the model’s reading comprehension; while it excels at extracting questions containing dedicated structural indicators, it struggles to integrate meaning and apply deeper contextual reasoning when tracking themes that appear across different parts of a paper.

### 5.3.2 Model Behavioral and Response Trends

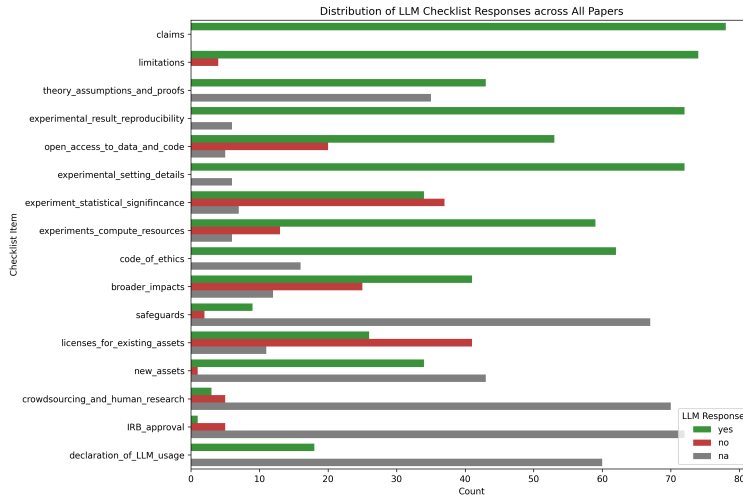


Figure 3: LLM Checklist Response Distribution

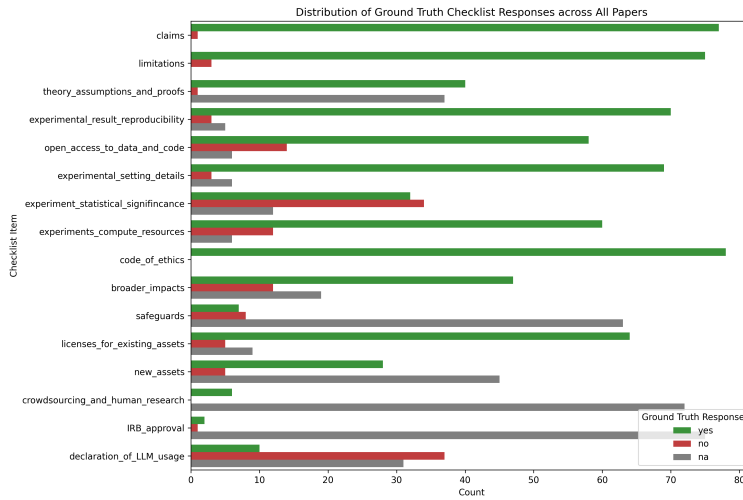


Figure 4: Human-annotated Checklist distribution

When moving beyond accuracy, and focusing on model response behavior we observe the following: A macro-level comparison between the LLM’s aggregate response patterns (figure 3) and the human-annotated ground truth (figure 4) reveals that the overall distributions are very similar. The model successfully mimics human frequencies for most of the checklist items. Although, a subtle behavioral divergence is shown in how negative classifications are applied: human annotators were more likely to mark "no" for missing core methodological elements (such as theoretical assumptions and proofs), whereas the LLM defaulted to "no" more frequently on lower-level compliance metrics. Such as `broader_impacts`, which further indicates the observation we stated regarding the LLM’s lack in reading comprehension.

However, the broader alignment does show promise in the LLM’s capacity to mimic human response patterns. To determine if these classifications are driven by logical feature groupings, the consistency matrix in figure 5 tracks the percentage of "Yes" responses mapped against the model’s predicted paper type per checklist item.

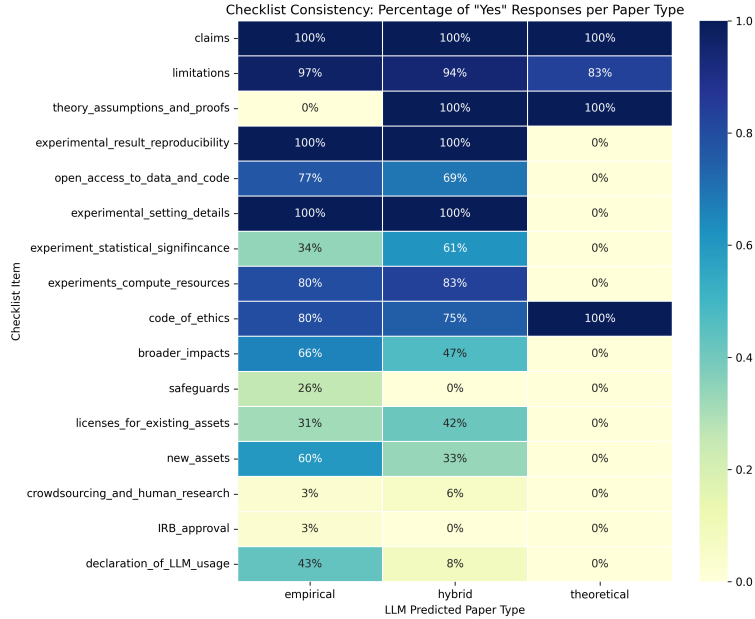


Figure 5: Consistency Heatmap

- Papers classified by the LLM as empirical exhibit a 100 percent confirmation rate for `experimental_result_reproducibility` and `experimental_setting_details`, alongside a strict 0% confirmation rate for `theory_assumptions_and_proofs`.
- On the other hand, papers classified as theoretical show a 100% confirmation rate for `theory_assumptions_and_proofs` and a strict 0% rate for all experimental and data-reproducibility metrics.
- Papers classified as hybrid correctly close this gap, displaying 100% frequencies for both core experimental fields and `theory_assumptions_and_proofs`

This distribution confirms that the model follows a highly consistent, rule-based classification logic. Rather than treating the extraction tasks as separate, disconnected prompts, the LLM maps its final methodological judgment directly to the presence or absence of these key technical indicators. This quantitative trend provides a useful basis for our qualitative evaluation, where we examine the verbatim textual justifications generated alongside these extractions.

### 5.3.3 Reasoning Alignment and Qualitative Analysis

To verify that the LLM’s perfect classification accuracy stems from actual text comprehension rather than statistical guessing, we qualitatively compared its generated justifications

against the human baseline (full transcripts available in Appendix B). The analysis reveals a strong alignment in the underlying evaluation logic. Both the human annotator and the model rely on identical foundational criteria: the presence of mathematical proofs drives a theoretical classification, whereas the use of benchmark datasets and results mainly drive an empirical one. However, the LLM exhibits a strict "auditor-like" extraction style that distinguishes it from the human justifications. We identified three primary behavioral patterns:

- **Precise Referencing:** While the human annotator summarizes broadly, the LLM supports its reasoning by referring to specific locations in the text, including theorem numbers (e.g., "Theorem 3.1") and exact dataset names.
- **Explicit Absences:** For purely empirical or theoretical classifications, the LLM consistently supports its decision by explicitly stating what the paper lacks (e.g., "contains no formal theoretical proofs").
- **Strict Binary Logic:** The human annotator occasionally notes if a hybrid paper "leans" more empirical or theoretical. However, the LLM processes hybrid papers with more symmetry, strictly verifying that *both* elements are present without weighing one as dominant.

This confirms the LLM’s high-level classifications are supported by accurate, context-aware textual findings.

#### 5.4 How consistent are human experts when annotating these papers based on the NeurIPS checklist?

Our final research question looks at human consistency to determine the inherent ambiguity of the checklist extraction task. Comparing this human baseline against the LLM’s accuracy reveals a solid correlation between task ambiguity and automated performance drop-offs.

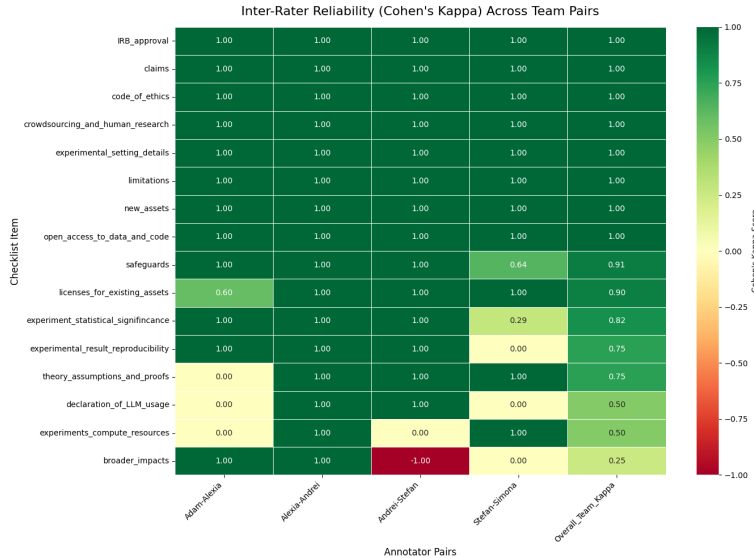


Figure 6: Inter-Annotator Agreement between each annotator pair

In figure 6 we observe a cluster of highly objective, structural items such as `claims`, `experimental_setting_details`, `IRB_approval`, and `limitations`, where human annotators achieved near-perfect consensus (Kappa scores of 1.00). These exact metrics correspond to the LLM’s highest performance tier for the checklist accuracy, consistently exceeding around 94% accuracy. Because these items are localized within dedicated sections and have clear structural indicators, they likely present minimal ambiguity, providing a clean target for both human annotators and the model. Conversely, we observe a correlation between the specific checklist items where the LLM scored lower and those that generated higher disagreement among human annotators. such as `broader_impacts`, `experiment_statistical_significance`, and `declaration_of_LLM_usage`. This trend provides important context when reviewing the LLM’s performance, which will be further explored in the discussion section.

While the heatmap reflects perfect consensus on core structural items, it also displays some anomalies, with specific pairs dropping to 0.00. Rather than indicating human disagreement, statistical analysis reveals these anomalies are manifestations of the **Kappa Paradox** [6]. For some checklist items where the true prevalence is heavily weighted toward a single class (e.g., nearly all papers correctly answering "n/a"), the expected agreement by chance is extremely high. Therefore, even a single divergent label between two annotators disproportionately penalizes the Kappa statistic, driving it to zero or negative values.

To resolve this paradox we calculated the raw extraction accuracy between the annotator pairs (provided in the Raw Accuracy Heatmap in Appendix C.1).

## 6 Responsible Research

In line with the evaluation standards of this study, we outline the ethical and practical considerations that inform our research methodology.

**Data Privacy:** The dataset used in this study consists entirely of publicly available papers published in the NeurIPS proceedings. No sensitive, or personally identifiable information was collected or analyzed during this research. Both the human annotation and the automated LLM pipeline phase were strictly confined to analyzing the publicly distributed, open-access PDF texts and their supplementary material, ensuring full compliance with academic data privacy standards.

**Compute Resources and Environmental Impact:** The automated extraction and classification pipeline relied on API calls to the Gemini 3.0 Flash model. While the environmental impact of querying pre-trained models is a lot lower than training them from scratch, we still recognize the energy consumption inherent in LLM use. To minimize redundant use of the LLM, prompt engineering and pipeline debugging were initially conducted on individual test papers before executing the final evaluation run on the broader dataset.

**Algorithmic Bias in Automated Evaluation:** Using an LLM to evaluate research papers can introduce algorithmic bias. As shown in our results, the model tends to be strict and often looks for clear structural indicators. This may put some authors at a disadvantage, particularly those who use different writing styles, non-traditional paper structures, or discuss ethical considerations throughout the text rather than in dedicated sections. For this reason, our pipeline is intended as a support tool for preliminary screening, not as a system that makes final decisions about academic publication.

**Declaration of AI Usage:** Finally, to comply with the guidelines regarding the declaration of LLM usage, we disclose that large language models were utilized during this project. Beyond the core of the paper where the LLM served as the primary subject of

evaluation, AI assistants were used to assist in localized code generation (e.g., Python matplotlib visualizations), and to assist in the formatting and writing of the paper. Example prompts used will be further specified in Appendix C.

## 7 Discussion

In this section, we interpret our findings within a broader context, addressing how the sub-questions answer our primary research objective while acknowledging key limitations and practical applications.

**Synthesis of Cohen’s Kappa:** While the LLM reliably classifies paper methodologies based on clear structural indicators, analyzing the Cohen’s Kappa scores reveals nuances about task ambiguity. We see a clear pattern: for items with high human agreement, the LLM performed well, and for subjective items with higher human disagreement, the LLM’s accuracy dropped. However, the ambiguity inherent to some of the checklist items only justifies part of the LLM’s lower accuracy. For other checklist items, such as `licenses_for_existing_assets` and `new_assets`, human consensus remained remarkably strong, yet the LLM still performed poorly. This demonstrates that while human disagreement provides useful context for some of the LLM’s lower scores, it cannot explain all of them. In these cases, the model failed either because the text lacked clear structural indicators, such as license details being spread throughout the paper rather than organized in a dedicated appendix, or due to document scoping where the required context existed entirely outside the primary PDF text.

**Comparison with Related Work:** These results strongly align with and extend prior literature regarding LLM-based information extraction. As noted in our review of existing work, models typically struggle with generating comprehensive, subjective open reviews but perform well on targeted, closed-ended questions. Our findings directly support this distinction. We confirm that while LLMs provide a highly consistent baseline for clearly structured items within the NeurIPS evaluation framework, they fall short on tasks requiring nuanced contextual reasoning. Additionally demonstrating that the model’s automated extraction failures for certain questions closely mirror the areas where human annotators experienced disagreement, this study bridges the gap between basic automated text extraction and complex methodological classification.

**Threats to Validity and Limitations:** Despite the promising classification accuracy, several threats to validity and experimental limitations must be acknowledged.

- **Sample Size:** The perfect classification accuracy (100%) was achieved on a smaller, manually verified ground-truth subset of 20 papers. Scaling this manual verification to a significantly larger, more diverse dataset may introduce edge cases or complex hybrid methodologies not captured in our current sample.
- **Contextual Blind spots:** Our extraction pipeline was strictly limited to parsing the PDF texts. As observed in the results for `licenses_for_existing_assets` and `experiments_compute_resources`, the model systematically failed when the required information was located in external supplementary materials, zip files, or GitHub repositories; areas easily accessed and verified by human annotators.
- **Model Dependency:** Our study relied exclusively on the Gemini 3.0 Flash model due to API constraints and the need for a balanced request limit. Evaluating multiple LLMs (such as GPT-4 or Claude 3) could give different reasoning capabilities or

extraction results, meaning our findings represent the capabilities of a specific model tier rather than an LLM baseline.

**Implications for the Peer Review Process:** Our findings suggest that LLMs are best suited as assistive screening tools rather than autonomous decision-makers. The model performs well at quickly verifying standard methodological components, which could help reduce reviewer workload. However, its reliance on clear structural indicators may disadvantage papers that use less conventional reporting styles. As a result, LLMs are most effective for initial screening and information extraction, while more complex and subjective assessments should remain the responsibility of human reviewers.

## 8 Conclusions and Future Work

This study addressed the primary research question: *To what extent can Large Language Models reliably extract structured checklist information to classify the methodological nature of scientific papers compared to human annotators?* We investigated this by evaluating the Gemini 3.0 Flash model against a human-annotated ground truth based on the NeurIPS 2024 proceedings checklist.

Our findings demonstrate that the LLM is highly accurate at classifying scientific papers as empirical, theoretical, or hybrid when relying on information containing clear structural indicators. The model successfully identified core methodological components such as mathematical proofs and explicit experimental settings mimicking the evaluation rationale of human annotators. However, the model’s extraction accuracy decreased on certain checklist items due to a combination of distinct factors. While inherent task ambiguity and minor human disagreement partially explained the performance drop on subjective criteria like broader impacts, it did not account for all failures. The LLM also struggled significantly with a lack of clear structural indicators failing when required information was spread throughout the text rather than neatly localized and when missing data like open-source licenses or the experiments’ computational specifications resided in external supplementary materials.

The main contribution of this study is bridging the gap between basic automated text extraction and complex methodological classification. We establish that while LLMs provide a highly consistent baseline for classifying paper typologies and extracting explicit methodological data, their inherent dependence on structural indicators and their limited contextual reasoning restrict their ability to be used independent. Consequently, we conclude that LLMs are currently best deployed as assistive screening tools to automate initial compliance checks, rather than as autonomous evaluators.

To build upon these findings, future research should expand the evaluation dataset to include a greater variety of conferences and edge-case methodologies to ensure these patterns hold at a larger scale. Additionally, future iterations of the automated pipeline should integrate multi-document processing to allow the model to assess supplementary zip files and external repositories. Finally, comparing these baseline results against other advanced models could help determine if these limitations in reading comprehension and document scoping represent a universal limitation in current LLM architectures.

## A LLM Prompt

The following prompt was used to generate the classification and checklist annotations:

LLM\_PROMPT = ""

You are an expert peer-reviewer for the NeurIPS conference. You are provided with a PDF of a research paper from the 2024 Main Conference Track. Based only on information explicitly the provided PDF, please complete the following tasks.

#### Task 1 â Checklist Annotation

For each item below, answer strictly with:

- yes
- no
- na

Base your answers only on information explicitly present in the paper. Do not assume missing details.

#### Task 2 â Paper Classification

Classify the paper as exactly one of:

- empirical (focus on experiments and data)
- theoretical (focus on formal analysis, proofs, or theory)
- hybrid (contains both theoretical and experimental contributions)

Here are the checklist guidelines:

##### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Guidelines:

- The answer "na" means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A "no" or "na" answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

##### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Guidelines:

- The answer "na" means that the paper has no limitation while the answer "no" means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic

approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Guidelines:

- The answer "na" means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and crossreferenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Guidelines:

- The answer "na" means that the paper does not include experiments.
- If the paper includes experiments, a "no" answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Guidelines:

- The answer "na" means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "no" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results

for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

#### 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Guidelines:

- The answer "na" means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

#### 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Guidelines:

- The answer "na" means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Guidelines:

- The answer "na" means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Guidelines:

- The answer "na" means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Guidelines:

- The answer "na" means that there is no societal impact of the work performed.
- If the authors answer "na" or "no", they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g

., pretrained language models, image generators, or scraped datasets)?

Guidelines:

- The answer "na" means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Guidelines:

- The answer "na" means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

#### 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Guidelines:

- The answer "na" means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Guidelines:

- The answer "na" means that the paper does not involve crowdsourcing nor

research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Guidelines:

- The answer "na" means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our MainTrackHandbook for what should or should not be described.

Output JSON Format (STRICT â follow exactly)

```
{
  "paper_title": "
    in_this_format_with_underscores_instead_of_spaces_everything_in_lowercase",
  "paper_type": "",
  "claims": "",
  "limitations": "",
  "theory_assumptions_and_proofs": "",
  "experimental_result_reproducibility": "",
  "open_access_to_data_and_code": "",
  "experimental_setting_details": "",
  "experiment_statistical_significance": ""
}
```

```

"experiments_compute_resources": "",
"code_of_ethics": "",
"broader_impacts": "",
"safeguards": "",
"licenses_for_existing_assets": "",
"new_assets": "",
"crowdsourcing_and_human_research": "",
"IRB_approval": "",
"declaration_of_LLM_usage": "",
"justification": "Provide a brief explanation for your classification choice,
referencing specific aspects of the paper that led to your decision."
"short justification for checklist": ""
}
"""

```

## B Justification Comparison

Manual Justification	LLM Justification
<b>Paper: einspace_searching_for_neural_architectures_from_fundamental_operations</b>	
The paper claims competitive results with their contribution, 'the einspace', and backs this with empirical results. However, their einspace is a theoretical contribution; they utilize and build on top of concepts from formal language theory. Therefore, the paper is a hybrid.	The paper is classified as hybrid because it introduces a formal methodological framework (a parameterised probabilistic context-free grammar for neural architecture search) and provides theoretical analysis of its properties (such as consistency via branching rate analysis in Section A.4 and search space size calculations in Section A.6), while also conducting extensive empirical benchmarks across eight diverse datasets to validate the expressivity and performance of the proposed search space.
<b>Paper: structured_flexibility_in_recurrent_neural_networks_via_neuromodulation</b>	
The paper provides a mathematical explanation for the intuition behind their model and generates predictions for what the signals should look like. They then verify these theories and assumptions using experiments, which represents the empirical side, so it is fair to say this paper is a hybrid.	The paper is classified as hybrid because it provides formal mathematical analyses and derivations (such as the connection to LSTMs in Section 3.2 and the analytical solution for rank-1 networks in Section 4.1) alongside extensive empirical evaluations across three distinct task types (timing, multitask learning, and long-term dependencies).
<b>Paper: du-shapley:_a_shapley_value_proxy_for_efficient_dataset_valuation</b>	

Manual Justification (Continued)	LLM Justification (Continued)
<p>The paper provides theoretical guarantees showing that the DU-Shapley almost surely converges to the Shapley value as the dataset grows, and it also includes theoretical proofs demonstrating an exponential speedup. They then back the benefits of these methods with experiments.</p>	<p>The paper is classified as hybrid because it contributes both a theoretical framework with formal proofs for a new Shapley value approximation method (DU-Shapley) and an extensive empirical evaluation using real-world and synthetic datasets to validate its performance and efficiency against existing benchmarks.</p>
<p><b>Paper:</b> <code>enhancing_zero-shot_vision_models_by_label-free_prompt_distribution_learning_and_bias_co</code></p>	
<p>The paper does include theoretical proofs; these are provided to give more credibility to the empirical results by justifying the key equations used. I will classify this as a hybrid, but it is leaning towards empirical, as the focus is on the experimental results across the 16 datasets. Therefore, I can understand if the LLM chooses to classify it as empirical.</p>	<p>The paper is classified as hybrid because it presents both significant theoretical contributions (formal proofs for prompt distribution estimation and unsupervised bias correction in Appendix A) and extensive empirical evaluations across 16 classification datasets to validate the proposed 'Frolic' framework.</p>
<p><b>Paper:</b> <code>diffusionpid_interpreting_diffusion_via_partial_information_decomposition</code></p>	
<p>They use empirical results to show that PID is a potent tool for evaluating and diagnosing text-to-image diffusion models, as well as to validate the theoretical analysis performed beforehand which included a theoretical framework based on information theory.</p>	<p>The paper is classified as hybrid because it introduces a novel theoretical framework based on information theory (Partial Information Decomposition) to interpret diffusion models, including formal mathematical derivations for its application at the pixel level (Section 3 and Appendix 6.2). Simultaneously, it presents extensive empirical evaluations of this framework across diverse tasks such as bias analysis, homonym/synonym characterization, and prompt intervention (Section 4).</p>
<p><b>Paper:</b> <code>test-time_dynamic_image_fusion</code></p>	
<p>The paper introduces a theoretically grounded image fusion framework with provable generalization guarantees, which is then validated through extensive experiments. Therefore, it is fair to say this paper is a hybrid.</p>	<p>The paper is classified as hybrid because it provides a formal theoretical proof (Theorem 3.1 and Appendix A.1) for the superiority of dynamic image fusion based on generalization error bounds, alongside extensive empirical evaluations across four different image fusion tasks and multiple datasets.</p>
<p><b>Paper:</b> <code>an_end-to-end_graph_attention_network_hashing_for_cross-modal_retrieval</code></p>	

Manual Justification (Continued)	LLM Justification (Continued)
<p>The paper is empirical because it mainly focuses on building a model (EGATH) and testing it on datasets like MIRFlickr25K, NUS-WIDE, and MS-COCO. Although the paper does include math and equations, these are design descriptions and operational formulas; they do not include any theorems, lemmas, or propositions stated anywhere in the text. In my opinion, this does not count as a theoretical contribution, but rather as a design choice. Thus, I would classify this paper as empirical, though I could understand if it gets classified as a hybrid.</p>	<p>The paper is classified as empirical because its primary focus is on the architectural design and experimental validation of a new cross-modal retrieval framework (EGATH). It provides comprehensive performance comparisons on three benchmark datasets (NUS-WIDE, MIRFlickr25K, and MS-COCO) using evaluation metrics like mAP and PR curves. The paper does not include formal theoretical proofs or new mathematical theorems.</p>
<p><b>Paper:</b> hamba_single-view_3d_hand_reconstruction_with_graph-guided_bi-scanning_mamba</p>	
<p>The paper does not contain any theoretical analysis. It is purely empirical, as it focuses on building a model (HAMBAs) and testing it on datasets like FreiHAND, RHD, STB, and HO3D.</p>	<p>The paper focuses on the development of a new neural network architecture, Hamba, for 3D hand reconstruction. Its primary contributions are empirical, demonstrated through extensive performance evaluations on several benchmark datasets (FreiHAND, HO3D, HInt) and detailed ablation studies. It does not provide formal theoretical proofs or analysis.</p>
<p><b>Paper:</b> omnigrasp_grasping_diverse_objects_with_simulated_humanoids</p>	
<p>The paper does not contain any theoretical analysis. It is purely empirical, as it focuses entirely on building a model (Omnigrasp) and testing it.</p>	<p>The paper is classified as empirical because its primary contributions are a novel humanoid controller architecture (Omnigrasp) and a dexterous motion representation (PULSE-X), which are validated through extensive experiments on multiple datasets (GRAB, OakInk, OMOMO). It contains no formal theoretical proofs or analysis.</p>
<p><b>Paper:</b> alchemistcoder_harmonizing_and_eliciting_code_capability_by_hindsight_tuning_on_multi_so</p>	

Manual Justification (Continued)	LLM Justification (Continued)
<p>This study is purely empirical; it verifies its claims by using extensive experiments and does not include any theoretical analysis.</p>	<p>The paper focuses on a novel data construction and fine-tuning methodology for Code LLMs. It presents the AlchemistCoder models and supports its claims through extensive experimental evaluations on established benchmarks (HumanEval, MBPP, etc.) and comprehensive ablation studies. No formal theoretical proofs or analysis are provided.</p>
<p><b>Paper:</b> <code>gliding_over_the_pareto_front_with_uniform_designs</code></p>	
<p>This paper contains extensive experiments but also includes theoretical analysis, as they introduce new concepts and prove their properties. Both aspects make up a great portion of the text and could not be removed without significantly impacting the paper, so it is fair to say this is a hybrid.</p>	<p>The paper is classified as hybrid because it provides a significant theoretical contribution, including the introduction of 'fill distance' as a metric and formal proofs for theorems (2-7) regarding its properties and bounds (Appendix A). Simultaneously, it presents a substantial empirical evaluation of the proposed UMOD method across various synthetic and real-world multi-objective optimization benchmarks, reporting performance on multiple indicators with statistical significance analysis.</p>
<p><b>Paper:</b> <code>socialgpt_prompting_llms_for_social_relation_reasoning_via_greedy_segment_optimization</code></p>	
<p>This paper is purely empirical. It contains extensive experiments to back up its claims and does not include any theoretical analysis.</p>	<p>The paper focuses on the development and empirical evaluation of a modular framework (SocialGPT) and an optimization algorithm (GSPO) for social relation reasoning. It demonstrates performance improvements on standard benchmarks (PIPA, PISC) and provides ablation studies, but does not present formal theoretical proofs or analysis.</p>
<p><b>Paper:</b> <code>online_learning_of_delayed_choices</code></p>	
<p>The paper provides regret bounds and a lower bound matching the upper bound. They accomplish this through theoretical analysis and proofs. The experiments help validate their assumptions, and while they are not extensive, they still contribute to the findings. Therefore, I would say it is a hybrid leaning towards theoretical.</p>	<p>The paper is classified as hybrid because it makes significant contributions in both formal theoretical analysis (providing regret upper and lower bounds for the proposed algorithms with complete proofs in the appendix) and empirical evaluation (conducting simulation-based experiments to validate the algorithms against benchmarks).</p>
<p><b>Paper:</b> <code>overcoming_common_flaws_in_the_evaluation_of_selective_classification_systems</code></p>	

Manual Justification (Continued)	LLM Justification (Continued)
<p>Their theoretical contribution includes introducing five formal requirements, proposing a new metric, and evaluating it using a framework. Their empirical study involves running a comprehensive benchmark, using six datasets, and comparing 13 scoring functions to analyze how rankings change. Both make a significant contribution to the paper, so it is fair to say this is a hybrid.</p>	<p>The paper is classified as hybrid because it presents both a formal theoretical contribution—defining a new metric (AUGRC), establishing mathematical requirements for SC metrics, and providing formal derivations and proofs (Section 3 and Appendix A.1)—and an extensive empirical study that benchmarks various scoring functions across multiple datasets to demonstrate the practical relevance of the proposed metric (Section 4 and Appendix A.4).</p>
<p><b>Paper:</b> <code>connecting_joint_embedding_predictive_architecture_with_contrastive_self_supervised_learning</code></p>	
<p>The paper contains significant empirical benchmarking and formal mathematical proofs of the dynamics; therefore, it is a hybrid.</p>	<p>The paper is classified as hybrid because it provides both a formal theoretical analysis of representational dynamics (Section 3.3 and Appendices A and B) and an extensive empirical evaluation across diverse computer vision benchmarks, including ImageNet-1K, MS-COCO, ADE20K, and DAVIS-2017 (Section 4).</p>
<p><b>Paper:</b> <code>multi_label_learning_with_stronger_consistency_guarantees</code></p>	
<p>This paper does not contain any empirical results. It is purely theoretical, as it focuses on providing stronger consistency guarantees for multi-label learning through theoretical analysis and proofs.</p>	<p>The paper is strictly theoretical, providing formal analysis through H-consistency bounds and proofs of Bayes-consistency for several new surrogate losses. It characterizes the computational complexity of the proposed algorithms but explicitly states in the conclusion that empirical validation is left for future work, containing no experimental sections or results.</p>
<p><b>Paper:</b> <code>data_distribution_valuation</code></p>	
<p>The paper is a hybrid, as it introduces formal theoretical contributions—including problem formulation, error guarantees, and theorems—while also validating the approach through experiments on real-world tasks.</p>	<p>The paper is classified as hybrid because it introduces a novel theoretical framework for data distribution valuation using MMD and the Huber model (including proofs for error bounds and game-theoretic optimality in Appendix A) and complements this with extensive empirical evaluations across multiple real-world datasets and tasks to demonstrate sample efficiency and ranking performance.</p>
<p><b>Paper:</b> <code>adjointdeis_efficient_gradients_for_diffusion_models</code></p>	

Manual Justification (Continued)	LLM Justification (Continued)
<p>The paper is a hybrid leaning heavily towards theoretical, as it provides theoretical analysis and proofs for the adjoint method for efficient gradients in diffusion models, while also including empirical results to validate the theoretical findings. However, I can understand if the LLM classifies it as purely theoretical.</p>	<p>The paper is classified as hybrid because it introduces a novel theoretical framework for simplifying continuous adjoint equations for both diffusion ODEs and SDEs (Sections 2 and 3, Proposition 2.1, Theorem 3.1), provides formal convergence order proofs for the proposed AdjointDEIS solvers (Theorem 2.1 and Appendix C), and validates the approach through extensive empirical experiments on a face morphing task (Section 5 and Table 1).</p>
<p><b>Paper:</b> <code>differentially_private_equivalence_testing_for_continuous_distributions_and_applications</code></p>	
<p>This paper is purely theoretical, as it does not contain any experiments. It focuses on providing theoretical analysis and proofs for differentially private equivalence testing for continuous distributions and their applications.</p>	<p>The paper is purely theoretical, focusing on the development and formal analysis of a differentially private algorithm for testing equivalence between continuous distributions. It provides mathematical definitions, algorithmic descriptions, and rigorous proofs for privacy and utility (sample complexity), without any empirical experiments or data evaluations.</p>
<p><b>Paper:</b> <code>reinforcement_learning_with_adaptive_regularization_for_safe_control_of_critical_systems</code></p>	
<p>The paper provides analytical guarantees for the safety and convergence of their proposed method, while also validating the method through experiments. Thus it is a hybrid.</p>	<p>The paper is classified as hybrid because it provides both a formal theoretical analysis (including proofs for Lemmas and Theorems regarding policy regularization and convergence in Appendix A) and an extensive empirical evaluation across four safety-critical control environments (Glucose, BiGlucose, CSTR, and Cart Pole) as detailed in Section 4 and Appendix B.</p>

## C Remaining visualizations

### C.1 Inter-Annotator Accuracy

Here we present the heatmap visualizing the total accuracy between annotator pairs. This is mainly used to give more context to the checklist items that fell victim to the Kappa’s paradox

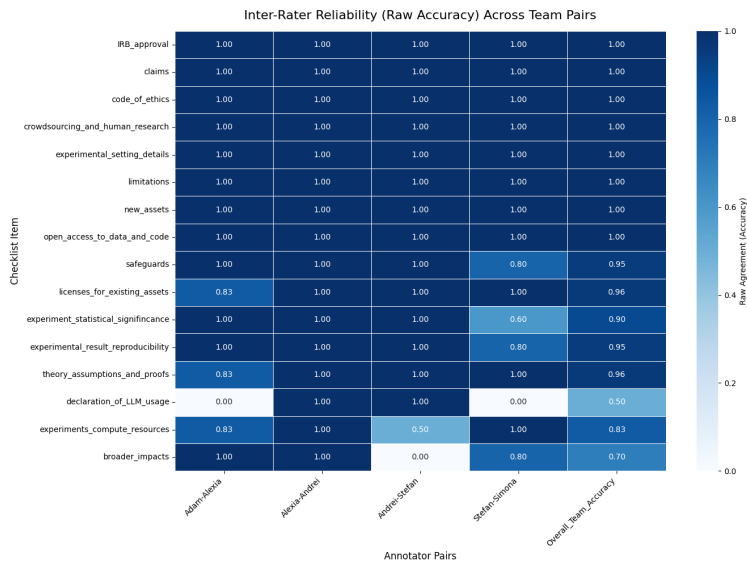


Figure 7: Raw Inter-Annotator Accuracy Heatmap

## C.2 Checklist Macro-F1 Score

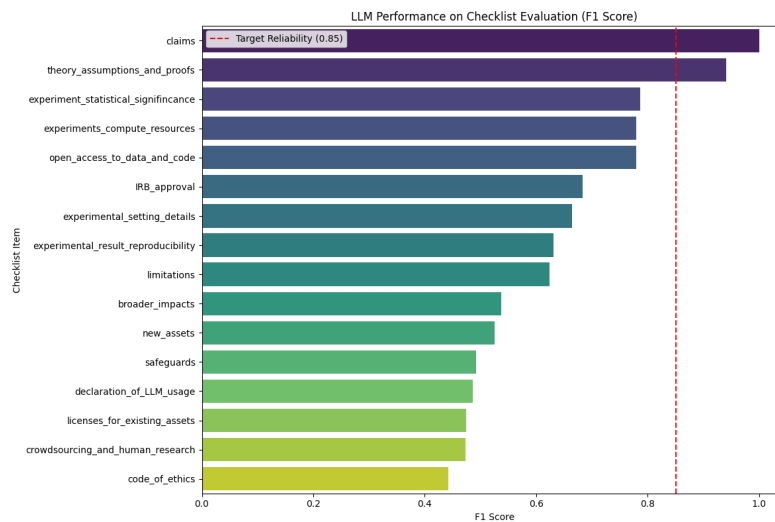


Figure 8: LLM Accuracy on NeurIPS Checklist - Complete Dataset

## D Declaration of AI usage

This part of the appendix declares any use of artificial intelligence that contributed to the project in a supporting capacity, such as assisting with coding, debugging, and writing. They will be demonstrated by showing some example prompts for each specific case

## D.1 Python visualizations and Debugging

The following prompts were used to help solve errors, or help create Matplotlib visualizations after the author had already created a draft for the code of the scripts

```
File "/Users/adam/Desktop/RP/analysisMLresults/venv/lib/python3.13/site-
packages/numpy/lib/_arraysetops_impl.py", line 296, in unique ret =
_uniqueid(ar, return_index, return_inverse, return_counts, equal_nan=
equal_nan, inverse_shape=ar.shape, axis=None, sorted=sorted) File "/
Users/adam/Desktop/RP/analysisMLresults/venv/lib/python3.13/site-
packages/numpy/lib/_arraysetops_impl.py", line 385, in _uniqueid ar.
sort() ~~~~~~ TypeError: '<' not supported between instances of '
str' and 'float'
```

*Why did this error take place in t1.py but not in t2.py Even though the format of the json objects remained the same*

```
--- Final Cohen's Kappa Matrix ---
Pair Adam-Alexia Alexia-Andrei Andrei-Stefan Stefan-Simona Overall_Team_Kappa
Checklist_Item
IRB_approval 1.0 1.0 1.0 1.000 1.000
claims 1.0 1.0 1.0 1.000 1.000
code_of_ethics 1.0 1.0 1.0 1.000 1.000
crowdsourcing_and_human_research 1.0 1.0 1.0 1.000 1.000
experimental_setting_details 1.0 1.0 1.0 1.000 1.000
limitations 1.0 1.0 1.0 1.000 1.000
new_assets 1.0 1.0 1.0 1.000 1.000
open_access_to_data_and_code 1.0 1.0 1.0 1.000 1.000
safeguards 1.0 1.0 1.0 0.643 0.911
licenses_for_existing_assets 0.6 1.0 1.0 1.000 0.900
experiment_statistical_signifincance 1.0 1.0 1.0 0.286 0.821
experimental_result_reproducibility 1.0 1.0 1.0 0.000 0.750
theory_assumptions_and_proofs 0.0 1.0 1.0 1.000 0.750
declaration_of_LLM_usage 0.0 1.0 1.0 0.000 0.500
experiments_compute_resources 0.0 1.0 0.0 1.000 0.500
broader_impacts 1.0 1.0 -1.0 0.000 0.250
(venv) adam@Adams-MacBook-Pro analysisMLresults %
```

*i want this but as a visual preferably a heatmap can you generate an example for me*

## D.2 Latex formatting

These prompts were used to assist in any problems caused by unfortunate latex formatting

- *perfect but in the code i placed the pictures before the sub-subsection: behavior. Why are they after it in the PDF. How do i get them to appear before, like in the code*
- *Is there a way to get the pictures you see in the right hand side beside each other instead of under, because i have more results to show and it takes up a lot of space*
- <https://www.bibtex.com/c/bibtex-to-ieee-converter/> - This website was utilized to transform the BibTeX files into IEEE format for all cited literature

### D.3 Writing clarity

These prompts were used to improve written sections after the author had already created a draft for it.

- Here is my discussion draft, I would like it to be made more formal, as well as improve the structure a bit

I want to go a little less in depth on the inter-annotator agreement, since it was only the third sub-question and mainly serves to provide context for the LLM’s performance. Make that section smaller and place more emphasis on the fact that the model performed well on the extraction task. “The LLM is generally strong at closed-ended extraction tasks”: this should be highlighted.

Then move to the checklist accuracy results, “where performance was somewhat lower on items that required identifying information scattered throughout the paper”: Use that to show the model’s limitations in reading comprehension. You can also briefly mention that Cohen’s kappa was somewhat lower for these items as well, but make that shorter than it is in the draft shown to you.

- Here is my draft for Responsible Research section. The content is good, only the flow feels off when reading. Can you please improve the flow of the draft without changing any of the structure or points made.

## References

- [1] H. Chen *et al.*, “Can large language models fully automate or partially assist paper selection in systematic reviews?,” *British Journal of Ophthalmology*, vol. 109, no. 8, pp. 962–966, 2025.
- [2] R. Zhou, L. Chen, and K. Yu, “Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 9340–9351.
- [3] H. Shin *et al.*, “Automatically Evaluating the Paper Reviewing Capability of Large Language Models,” 2025.
- [4] J. Pineau *et al.*, “Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program),” *J. Mach. Learn. Res.*, vol. 22, p. 164:1–164:20, 2020.
- [5] R. Liu and N. B. Shah, “ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing,” *ArXiv*, vol. abs/2306.00622, 2023.
- [6] S. Zec, N. Soriani, R. Comoretto, and I. Baldi, “High agreement and high prevalence: The paradox of Cohen’s Kappa,” *Open Nurs. J.*, vol. 11, no. 1, pp. 211–218, Oct. 2017.
- [7] E. Raff, “A step toward quantifying independently reproducible machine learning research,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [8] S. Jiang, “Tenure under pressure: Simulating the disruptive effects of AI on academic publishing,” *arXiv preprint arXiv:2509.16925*, 2025.

- [9] S. P. Mann *et al.*, “AI and the Future of Academic Peer Review”, *arXiv preprint arXiv:2509.14189*, 2025.
- [10] Y. Li, “A practical survey on zero-shot prompt design for in-context learning”, in *Proceedings of the 14th international conference on recent advances in natural language processing*, 2023, pp. 641-647.
- [11] F. M. Megahed, Y.-J. Chen, L. A. Jones-Farmer, Y. Lee, J. B. Wang, and I. M. Zwetsloot, “Reliable decision support with LLMs: A framework for evaluating consistency in binary text classification applications”, *Journal of Business Analytics*, pp. 1-19, 2026.