How Does the Representation of an Annotator's Moral Value Profile Influence an LLM's Offensiveness Judgements?

MSc Thesis in Data Science and Artificial Intelligence Technology

Yiming Chen

TU Delft

How Does the Representation of an Annotator's Moral Value Profile Influence an LLM's Offensiveness Judgements?

by

Yiming Chen

To obtain the degree of Master of Science

in Data Science and Artificial Intelligence Technology

at Delft University of Technology,

to be defended publicly on Monday 24 August 2026 at 10:00 AM.

Student number:	5541786	
Project duration:	24 November 2025 – 24 August 2026	
Thesis committee:	Luciano Cavalcante Siebert	TU Delft, Thesis Advisor
	Odette Scharenborg	TU Delft, Committee Member
	Amir Homayounirad	TU Delft, Daily Co-Supervisor

Cover image: generated with OpenAI ChatGPT image generation, directed and iteratively refined by the author. Further disclosure is provided in the statement on generative AI use.

An electronic version of this thesis is available at https://repository.tudelft.nl/.

TU Delft

Abstract

Offensive-language detection is often evaluated as a single-label classification task, although research on subjective annotation shows that people can make systematically different judgements about the same text. This thesis focuses on predicting annotator-specific offensiveness judgements rather than establishing an objective offensiveness label. Moral values are examined as one possible personalization signal because they provide an individual-level description of normative concerns that may be relevant when people interpret potential violations of social and moral norms. The study asks whether different representations of the same measured moral-value profile change an LLM's approximation of an annotator's observed binary judgement.

The experiments use D3CODE, which combines repeated individual offensiveness annotations with annotator-level scores on six moral foundations. Five deterministic renderings of each profile—natural language, categorical, numeric, ordinal, and top-k—are compared with a no-profile, unpersonalized reference. Qwen2.5-72B-Instruct is used as the primary model; Qwen2.5-7B-Instruct and Llama-3.3-70B-Instruct provide within-family and cross-family robustness comparisons. The models are evaluated in zero-shot prompting and in three same-annotator few-shot settings with $k = 3$ examples selected randomly, by semantic retrieval, or by diversity. Evaluation combines Accuracy, Macro-F1, paired McNemar comparisons, and an instance-level cross-representation decision-disagreement diagnostic.

The zero-shot condition exhibits the largest spread in aggregate performance across profile representations and cross-representation disagreement above 20% for each selected model. Same-annotator few-shot context is associated with lower disagreement, although profile representations continue to change individual predictions. Retrieval-based prompting produces the strongest Macro-F1 across the selected models and the lowest disagreement in most sub-categories for the primary model. No representation performs consistently best across every tested model, setting, and metric. Instead, the results show that profile representation and demonstration selection are consequential parts of the experimental treatment and should be reported, controlled, and evaluated in personalized, value-conditioned judgement prediction.

Contents

List of Figures

List of Tables

1

Introduction

Online platforms and researchers use offensive-language detection to support safety workflows and to study harmful discourse in large collections of user-generated text [1]. Much of the NLP literature formulates this task as supervised classification: given a post or comment, a system predicts whether it is hateful, offensive, or non-offensive [2, 3]. A single-reference benchmark deliberately evaluates models against one operational label. It therefore treats one interpretation as the target for evaluation, even though that label does not establish that all people would interpret the same text in the same way. In subjective language tasks, annotations can record social and normative judgements shaped by an annotator's experiences, beliefs, identities, and evaluative standards [4, 5].

This distinction matters for both data construction and model evaluation. A common annotation pipeline collects several labels and collapses them into one reference label, often by majority vote. Aggregation can be useful when the intended output is an institutional rule or a single deployable decision, but it removes information about who disagreed and whether the disagreement follows a systematic pattern. Several research directions address this loss at different stages. CrowdTruth treats disagreement as a signal about ambiguity and interpretation rather than automatically as annotation noise [6]. Perspectivist NLP is a broader family of approaches that retains, models, or explicitly evaluates multiple viewpoints [7]. Multi-annotator models go further by learning the label patterns of particular annotators; Davani et al., for example, show that individual subjective labels can be modelled instead of being discarded through majority aggregation [8]. Jury learning similarly retains dissenting annotations as part of the learning target [9]. Together, these approaches establish that human disagreement can be structured information and that the intended target—one policy label, a distribution of views, or a specified person's judgement—must be stated explicitly.

Once the target is a particular person's judgement, the next question is what information can plausibly condition a model on that individual. Sociodemographic and regional attributes can describe population-level patterns, but they are incomplete descriptions of a person's normative orientation. Inferring individual behaviour from group averages also risks an ecological fallacy. In individual toxic-content annotation prediction, Orlikowski et al. find that group-specific sociodemographic modelling does not significantly improve performance in their setting, illustrating that individual annotation behaviour depends on more than group membership alone [10]. This motivates examining individual-level signals that are closer to the considerations involved in the judgement, while remaining cautious about reducing a person to any single profile.

Moral values provide one candidate signal for two reasons. First, they describe normative priorities at the individual level rather than assigning a person the average attributes of a demographic group. Second, they are multidimensional: two people can differ in the relative importance they place on care, fairness, loyalty, authority, or purity even when they belong to the same social group. Moral Foundations Theory (MFT) is not the only framework for studying values; Schwartz's theory of basic human values, for example, provides a broader account of motivational values [11]. This thesis focuses on MFT because it was designed around recurring moral concerns that are plausibly relevant when people evaluate violations of interpersonal and social norms, and because the dataset used in this work measures those concerns for the same annotators whose offensiveness judgements are predicted [12, 13].

The revised Moral Foundations Questionnaire (MFQ-2) measures six dimensions—care, equality, proportionality, loyalty, authority, and purity—and was developed and evaluated across multiple popu-

lations [14]. These dimensions do not determine whether a person will find a particular text offensive. They do, however, summarize differences in which moral considerations a person tends to emphasize. A measured moral-value profile is therefore a theoretically motivated but deliberately partial personalization signal.

D3CODE makes this question empirically testable. Introduced at EMNLP 2024, D3CODE is a cross-cultural dataset with more than 170,000 offensive-language annotation entries from more than 4,000 annotators, together with MFQ-2 measurements for those annotators [15]. It retains repeated binary judgements at the individual level and reports that people within the same demographic groups can hold different value profiles and perceive offensiveness differently. D3CODE therefore supports a controlled study in which the target is an annotator's observed judgement and the conditioning signal is that annotator's measured moral-value profile. The profile is treated as partial information about normative orientation, not as a demographic proxy, causal explanation, or complete psychological model.

A further design problem arises before the profile reaches an LLM. A measured profile is structured data, whereas an LLM receives a sequence of tokens. The same six scores can be supplied as exact numbers, categories, a rank ordering, a short list of the highest-scoring dimensions, or a natural-language description. These forms do not preserve and emphasize the same information. A numeric list retains magnitude; an ordinal list retains order but discards distance; and a top-k list highlights selected dimensions while omitting the rest. Moreover, prompt wording, ordering, and serialization can change in-context model behaviour even when the intended task meaning remains stable [16, 17, 18]. These considerations motivate treating profile representation as an experimental variable rather than as a neutral presentation choice.

Existing work therefore leaves a specific intersection unresolved. Research on subjective annotation shows that individual judgements should not always be collapsed into one label; value-based personalization shows that explicit values can carry information about individual ratings; and prompt-sensitivity research shows that LLM outputs depend on how information is expressed. However, the reviewed work does not establish what happens when one annotator's *measured* moral-value vector is held fixed while only its deterministic representation changes. It also does not establish whether direct behavioural evidence from that same annotator reduces this dependence on representation. This is the research gap addressed by the thesis.

The study follows a descriptive rather than prescriptive task formulation [19]. It evaluates whether an LLM approximates the recorded judgement of a specified annotator; it does not claim that the annotator's judgement is objectively correct or that it should determine a platform's moderation policy. The central research question is:

> **How does the representation of an annotator's moral-value profile affect an LLM's prediction of that annotator's offensiveness judgement?**

The question is examined through two sub-questions:

- **RQ1: Zero-shot representation sensitivity.** Under zero-shot prompting, how do alternative deterministic renderings of the same annotator moral-value profile affect (a) agreement with the annotator's observed judgement and (b) instance-level disagreement across profile representations?
- **RQ2: Prompting context and representation sensitivity.** How does representation sensitivity differ between profile-only prompting and prompting augmented with same-annotator examples, and how does this pattern vary across random, retrieval-based, and diversity-based example selection?

The experiment compares five deterministic renderings of the same moral-value profile—natural language, categorical, numeric, ordinal, and top-k—together with a no-profile reference condition. It evaluates zero-shot prompting and three same-annotator few-shot strategies under a fixed $k = 3$ budget. Qwen2.5-72B-Instruct is the primary model, while Qwen2.5-7B-Instruct and Llama-3.3-70B-Instruct provide within-family and cross-family robustness comparisons. The exact rendering rules, sampling procedure, prompt construction, model screening, and evaluation measures are specified in Section 3.

The thesis makes three contributions. First, it isolates profile representation as an experimental variable by holding the annotator, item, measured profile, model, and prompting setting fixed while changing only the profile rendering. Second, it examines whether same-annotator demonstrations are associated with lower representation-induced variation and whether that pattern depends on demonstration selection. Third, it separates aggregate agreement from instance-level behavioural stability

by combining Accuracy and Macro-F1 with paired McNemar comparisons and cross-representation decision disagreement. Together, these contributions show that profile representation and example selection are consequential experimental choices under the tested conditions.

The remainder of the thesis is organized as follows. Section 2 distinguishes conceptual background from the most closely related research and ends by positioning the thesis within that literature. Section 3 formalizes the task and describes the D3CODE subset, profile renderings, prompting conditions, models, and evaluation procedure. Section 4 reports the empirical findings. Section 5 interprets the findings in relation to the research gap and research questions. Section 6 discusses applicability, societal implications, and the relation to Data Science and Artificial Intelligence Technology. Section 7 presents limitations and corresponding future work, and Section 8 concludes the thesis.

2

Background and related work

This chapter separates the concepts needed to understand the study from the research most closely related to its contribution. Sections 2.1–2.3 introduce subjective offensiveness judgements, moral-value measurement, and profile conditioning in LLMs. Sections 2.4 and 2.5 then compare approaches that model annotator disagreement, values, and in-context examples. Section 2.6 makes the remaining research gap explicit and positions the present experiment relative to those strands.

2.1. Background: offensive-language judgements and task targets

Offensive-language detection has commonly been formulated as supervised classification. Davidson et al. distinguish hate speech from other offensive language and show that the boundary between these categories is difficult to operationalize reliably in social-media text [2]. OffensEval subsequently provided a standardized shared-task formulation for identifying and categorizing offensive language [3]. Such benchmarks are valuable because they establish common datasets, labels, and evaluation measures. Their usual single-reference-label formulation nevertheless answers a different question from an annotator-conditioned task. A benchmark asks whether a model agrees with one operational target; an annotator-conditioned task asks whether a model agrees with the judgement recorded from a specified person.

The distinction is especially important for subjective tasks. An offensiveness judgement can depend on perceived intent, social context, group membership, personal experience, and normative expectations. Aroyo and Welty argue that disagreement in semantic annotation can reflect a range of reasonable interpretations rather than merely low-quality annotation [6]. Plank similarly describes human label variation as both a challenge for data creation and evaluation and a source of information that may be lost when one label is treated as the only ground truth [5].

Röttger et al. clarify the task-level consequence by distinguishing descriptive and prescriptive annotation paradigms [19]. A descriptive paradigm aims to record how people actually judge an item and therefore expects disagreement. A prescriptive paradigm trains annotators to apply a selected policy, so disagreement can indicate ambiguity or inconsistent application of that policy. Platform moderation commonly requires an accountable prescriptive rule, whereas the experiment in this thesis is descriptive: the prediction target is an observed individual judgement. Keeping these targets separate prevents agreement with one person from being interpreted as evidence about what a platform should remove.

2.2. Background: moral values, MFT, and D3CODE

Individual differences in offensiveness judgements may reflect identity, experience, social context, political or religious commitments, and broader normative orientation. Sap et al. show that annotator attitudes and identities are associated with differences in toxic-language annotation, demonstrating that labels in this domain are not independent of the people who provide them [4]. Sociodemographic variables can describe some variation, but they are not complete representations of individuals. Orlikowski et al. explicitly model sociodemographic information in individual toxic-content annotation prediction

and find no significant improvement in their setting, highlighting the ecological-fallacy risk of extrapolating an individual's behaviour from group averages [10].

Values provide a different type of individual-level signal. Schwartz's theory of basic human values offers a broad framework for motivational values and their relations [11]. Moral Foundations Theory (MFT) focuses more narrowly on recurring moral concerns involved in evaluations of harm, fairness, group relations, authority, and purity [12, 13]. MFT is therefore not treated as the only valid theory of values. It is selected here because the task concerns judgements about possible violations of social and moral norms, and because D3CODE measured an MFT-based instrument for the exact annotators whose labels are modelled.

MFT was developed from cultural psychology, evolutionary accounts of recurrent social problems, and the proposal that moral judgement is multidimensional and culturally variable [12]. The original framework distinguished concerns related to care and harm, fairness, loyalty, authority, and purity. Graham et al. developed the Moral Foundations Questionnaire and provided evidence that these concerns form distinguishable patterns rather than one undifferentiated moral dimension [20]. Later work presents MFT as a form of moral pluralism: it supplies a vocabulary for recurring moral concerns without claiming to exhaust every possible moral consideration [13].

The Moral Foundations Questionnaire-2 (MFQ-2) revises the instrument and part of the theoretical structure. Atari et al. distinguish equality from proportionality and retain care, loyalty, authority, and purity, yielding six measured dimensions [14]. The 36-item instrument was developed across 25 populations and evaluated for its factor structure and measurement properties. This matters for the thesis because the resulting profile is both individual-level and multidimensional. Two annotators can emphasize different concerns even when their demographic attributes or overall tendency to make moral judgements appear similar.

The link between moral values and offensiveness is a research hypothesis rather than a definitional identity. Offensiveness judgements often involve perceived violations of interpersonal treatment, fairness, group norms, authority, or purity, so variation in these concerns may be informative. At the same time, an MFQ-2 profile does not directly represent personal experience, knowledge of a speaker, pragmatic intent, humour, reclaimed language, or the immediate conversational setting. The profile is consequently treated as a theoretically grounded but partial signal.

D3CODE supplies the empirical bridge between this signal and repeated individual judgements. Published at EMNLP 2024, the cross-cultural dataset contains more than 170,000 offensive-language annotation entries from more than 4,000 annotators, together with self-reported MFQ-2 measurements for those annotators [15]. Participants completed the questionnaire after the annotation task, and responses were aggregated into scores from 1 to 5 for care, equality, proportionality, loyalty, authority, and purity. D3CODE reports that individuals within demographic groups can hold different value profiles and that moral values are associated with variation in perceived offensiveness. The dataset was designed to study disagreement and its correlates; it did not systematically test how alternative prompt representations of one fixed profile affect an LLM's prediction.

2.3. Background: LLM conditioning, profiles, and representations

An unconditioned LLM produces a response from patterns acquired through pre-training and instruction tuning; it does not automatically represent the judgement of a particular person. Santurkar et al. compare language-model opinions with survey responses and find systematic misalignment with the views of multiple demographic groups [21]. This motivates explicit conditioning when the target is a particular perspective, but it also limits what can be claimed: a short description does not turn a model into a faithful simulation of a whole person.

This thesis distinguishes a *profile* from a broader *persona*. A profile is a specified set of measured or supplied attributes, here six MFQ-2 scores. A persona is often a richer textual characterization that can include identity, background, preferences, or behavioural instructions. Profile conditioning is useful experimentally because the input can be held fixed and transformed deterministically. Persona prompting can be less controlled and may encourage stereotypes or caricatures. Cheng et al. show that persona-based LLM simulations can exaggerate simplified group attributes, which is why the present work treats the profile as limited predictive information rather than as a reconstruction of identity [22].

Before a profile reaches an LLM, it must be converted into text. Different conversions preserve

different information. Measurement theory distinguishes representations according to the comparisons and operations they support [23]. Exact numeric scores retain magnitude and distance; categories retain approximate levels; rankings retain order but not distance; and sparse lists omit part of the source vector. Discretization can improve interpretability while removing within-bin variation [24], and feature selection changes the available signal by retaining only a subset of variables [25].

Abstractly, a representation is a mapping

$$R_r : [1, 5]^6 \longrightarrow \mathcal{T},$$

where $[1, 5]^6$ is the measured profile space and $\mathcal{T}$ is the space of prompt-text sequences. Different choices of R_r preserve different relations from the source vector: a numeric mapping retains reported distances, an ordinal mapping retains only order, a categorical mapping collapses values into bins, and a top-k mapping omits lower-ranked dimensions. The mapping is therefore part of the information supplied to the model rather than a neutral display layer.

LLM research provides a second reason to treat this conversion as consequential. Zhao et al. show that in-context performance can vary with prompt-specific label and ordering effects [16]. Sclar et al. systematically quantify sensitivity to prompt-format features and find that changes not intended to alter task meaning can materially change outputs, with no single format preferred consistently across models [17]. TabLLM likewise treats serialization of structured variables as a model-design choice because the LLM receives only the resulting token sequence [18]. Numeric magnitude processing is also an empirical model capability rather than a guaranteed operation [26]. These findings motivate testing profile representation rather than assuming that semantically related formats are behaviourally interchangeable.

2.4. Related work: retaining and modelling annotator disagreement

The survey by Frenda et al. organizes perspectivist NLP around how multiple viewpoints are represented in data, incorporated into models, and evaluated [7]. It shows that disagreement can be retained as distributions, annotator-conditioned targets, or multiple valid outputs rather than collapsed automatically. The survey also makes clear that perspectivist methods vary in what they predict and whose perspective they represent. It does not address the narrower question of whether different textual representations of one measured annotator profile change an LLM's individual prediction.

CrowdTruth is primarily a data-quality and annotation framework: it interprets disagreement jointly across workers, units, and annotations and uses that disagreement as information about ambiguity [6]. Multi-annotator modelling moves from retaining disagreement to predicting it. Davani et al. provide the most direct connection to the present target by training models to predict annotator-specific subjective labels; their results show that individual annotation patterns can be learned rather than discarded through majority vote [8]. Gordon et al.'s jury learning framework instead constructs and models selected groups of annotators so that dissenting voices remain visible in model outcomes [9]. These works establish the value of preserving the annotator dimension, but they do not study deterministic representations of a measured moral profile as an LLM prompt variable.

2.5. Related work: value-based personalization and demonstrations

Recent work studies values as representations of human variation. Sorensen et al. compare value profiles, demographic information, and examples as ways to encode individual ratings, showing that value-based descriptions can contain information beyond demographics [27]. Their profiles are generated as natural-language summaries of observed behaviour, whereas the present study starts from independently measured MFQ-2 scores and changes only their deterministic rendering. Lee and Goldwasser's SOLAR models individual subjectivity through value conflicts and trade-offs inferred from user-generated text [28]. SOLAR demonstrates the relevance of values to individual moral judgements, but it does not isolate the representation of one fixed measured profile.

Few-shot prompting provides another form of personalization signal. Labelled examples can reveal the task format, label space, input distribution, or a decision boundary even when their content is not a complete model of the annotator [29, 30]. In an annotator-conditioned task, same-annotator examples additionally provide direct observations of prior behaviour, whereas a moral-value profile is an abstract measurement.

The choice of examples can change in-context learning. Liu et al. show that semantically relevant examples can outperform random examples in GPT-3 prompting [31]. Later work studies active selection, model- and data-dependent strategies, and the trade-off between local similarity and broader coverage [32, 33, 34]. These findings motivate comparing random, retrieval-based, and diversity-based same-annotator examples. They do not determine whether those examples also make a model less sensitive to how an accompanying value profile is represented.

2.6. Research gap and thesis positioning

The reviewed literature establishes three relevant points separately. First, subjective disagreement can contain structured information, and models can target individual annotator judgements rather than only an aggregate label [8, 7]. Second, moral or broader value information can provide an interpretable individual-level signal beyond demographic grouping [15, 27, 28]. Third, LLM behaviour can change when semantically related information is expressed differently in prompts [17, 18].

What remains unclear is how these three issues interact. Prior annotator-specific and value-based work mainly asks *which information* can represent a person; prompt-sensitivity work mainly asks how prompt design affects general task performance. The present thesis asks whether alternative deterministic representations of a fixed measured profile change the judgement attributed to a specified person. It then tests whether same-annotator demonstrations—a direct behavioural signal—are associated with lower dependence on that representation.

The experiment differs from adjacent work in four controlled respects. It uses observed individual offensiveness labels rather than a majority target; it uses independently measured MFQ-2 scores rather than an inferred persona; it compares five deterministic renderings while holding the source vector and target item fixed; and it evaluates both aggregate agreement and instance-level changes in the final binary decision. Together, these controls isolate profile representation as an experimental variable in annotator-conditioned judgement prediction.

3

Methodology and experimental setup

This chapter translates the research gap into a controlled prediction experiment. It first formalizes the annotator-conditioned task and summarizes the complete pipeline. It then describes the D3CODE subset, the five profile renderings and their conversion rules, the zero-shot and few-shot prompts, the selected models, and the evaluation measures. The design separates three comparisons that should not be conflated: whether an explicit profile changes agreement relative to an unpersonalized reference, whether alternative renderings of the same profile change predictions, and whether same-annotator demonstrations alter that representation sensitivity.

3.1. Task formulation

Let $a \in \mathcal{A}$ denote an annotator and $i \in \mathcal{I}$ an annotated text item. The item text is x_i, and the annotator's observed binary judgement is

$$y_{a,i} \in \{0, 1\},$$

where 1 denotes *offensive* and 0 denotes *not offensive*. The target is the judgement recorded from annotator a for item i; it is not a majority-vote label and does not establish whether the text is objectively offensive.

Each annotator has a measured moral-value vector

$$\mathbf{v}_a \in [1, 5]^6.$$

The six components are care, equality, proportionality, loyalty, authority, and purity. They are not six dimensions chosen by the present experiment: they are the six foundations measured by the MFQ-2 and provided by D3CODE for the same annotators whose labels form the prediction target [14, 15]. MFT is used because these dimensions describe recurring moral concerns plausibly related to perceived social and interpersonal violations, while the measured D3CODE profiles make the signal available at the individual level.

A deterministic rendering function R_r converts the same vector into one of five profile representations:

$$\mathcal{R} = \{\text{natural language},\ \text{categorical},\ \text{numeric},\ \text{ordinal},\ \text{top-}k\}.$$

For a particular run, $r \in \mathcal{R}$ identifies the selected representation. These representations were selected to create interpretable contrasts already motivated in Section 2.3: exact magnitude, approximate score levels, rank order, sparse selection, and prose versus list presentation. They do not exhaust the representation design space and are not assumed to be equally informative.

For a few-shot selection strategy s, let $E_{a,i}^{s,k}$ denote up to $k = 3$ labelled examples drawn from annotator a's previous annotations, excluding the target item. The model prediction can then be written as

$$\hat{y}_{a,i}^{(m,r,s)} = f_m\left(x_i, R_r(\mathbf{v}_a), E_{a,i}^{s,k}\right),$$

where m identifies the language model. In zero-shot prompting, $E_{a,i}^{s,k} = \emptyset$. In the no-profile baseline, the term $R_r(\mathbf{v}_a)$ is omitted. For the representation comparison, the annotator, item, measured source profile, model, and prompting setting are held fixed while only R_r changes.

The experiment therefore contains two related comparisons. Comparing the no-profile reference with profile-conditioned prompts asks whether explicit annotator information changes agreement with observed individual labels. Comparing the five rendering functions asks whether the representation of the same measured information changes aggregate performance or the final instance-level decision.

3.2. Methodology overview

Figure 3.1 summarizes the end-to-end pipeline. D3CODE supplies individual labels and MFQ-2 measurements. Annotators are sampled at the annotator level to retain repeated judgements and cover different regions of the observed profile space. For each target annotation, the same source profile is converted into alternative renderings; zero-shot or same-annotator few-shot context is then added; each selected model is evaluated over three seeds; and parsed binary outputs are analysed with aggregate, paired, and cross-representation measures.

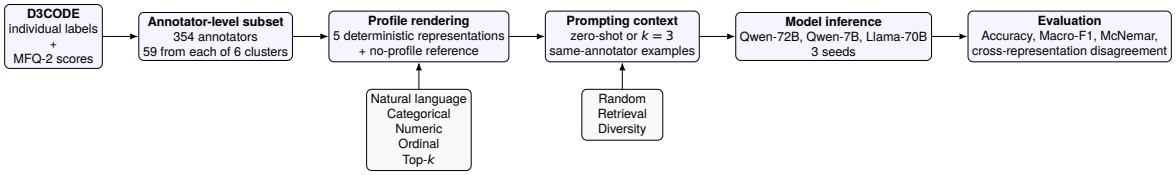

Figure 3.1: Overall experimental pipeline. The representation comparison holds the annotator, item, source profile, model, and prompting setting fixed while changing only the deterministic rendering of the profile.

Table 3.1 lists the principal design axes and distinguishes the variables central to the research questions from repetition and robustness factors.

Table 3.1: Overview of the experimental design.

Design component	Conditions or values	Role in the study
Prediction target	Annotator-specific binary offensiveness judgement	Outcome to be approximated; no majority aggregation
Profile information	No profile or the annotator's six MFQ-2 scores	Tests explicit personalization relative to an unpersonalized reference
Profile rendering	Natural language, categorical, numeric, ordinal, top-k	Primary representation variable for RQ1 and RQ2
Prompting context	Zero-shot, few-shot random, few-shot retrieval, few-shot diversity	Tests whether same-annotator examples change representation sensitivity
Models	Qwen-72B, Qwen-7B, Llama-70B	Primary analysis plus within-family and cross-family robustness checks
Repetition	Seeds 42, 521, and 1314 at temperature $T = 0.2$	Estimates run-to-run variation under stochastic decoding
Evaluation	Accuracy, Macro-F1, McNemar comparisons, decision disagreement	Separates aggregate agreement from paired and instance-level representation effects

3.3. Dataset and subset construction

The experiments use D3CODE, a cross-cultural dataset containing more than 170,000 offensive-language annotation entries from more than 4,000 annotators, together with annotator-level moral-value information [15]. D3CODE is appropriate for the present question because it records both components required by the task: repeated individual judgements as prediction targets and measured MFQ-2 profiles as a possible personalization signal.

Evaluating every combination of profile representation, prompting setting, model, and repeated seed over the full dataset was computationally infeasible. The experiment therefore uses a subset constructed at the annotator level. Sampling complete annotators rather than isolated annotation rows preserves each selected person's repeated judgements, which are required for same-annotator few-shot selection and per-annotator analyses.

The procedure first constructs one record per annotator using `rater_id`. Annotators are partitioned into six K-means clusters over the six MFQ-2 dimensions and sampled equally from each cluster to improve coverage of the observed profile space. The clusters are used only as sampling groups; they are not interpreted as natural psychological types or as an independent variable in the main experiment.

The target annotator count was determined with the finite-population sample-size formula for estimating a population proportion:

$$n = \frac{Nz^2p(1-p)}{e^2(N-1) + z^2p(1-p)}.$$

The calculation uses:

$N = 4309$	D3CODE participants
$z = 1.96$	95% confidence level
$e = 0.05$	5% margin of error
$p = 0.5$	maximum-variance assumption

Substitution gives approximately 353 annotators. The target was rounded to 354 so that 59 annotators could be sampled from each of the six clusters. All annotations from the selected annotators are retained, producing approximately 12,000 annotator–item pairs. The equal allocation is exactly balanced across the six clusters; Appendix A defines Shannon entropy and Simpson diversity and shows why both reach their maxima under this allocation. These measures check cluster balance only. They do not show that the clusters are well separated or that the subset reproduces the full D3CODE population distribution.

The sample-size calculation is a transparent rule for choosing a feasible subset with broad profile coverage; it is not a statistical-power guarantee for every pairwise comparison. The prediction target is `rating_binary`, where 1 corresponds to offensive and 0 to not offensive. Appendix A provides further construction details and a worked profile example.

3.4. Profile representations

Every value-conditioned prompt begins from the same measured numerical source vector. The five rendering functions are deterministic, so the representation conditions differ in which relations from the source profile they preserve or emphasize and, for some pairs, in surface realization. The representations were selected as a small set of interpretable contrasts, not as an exhaustive search over the representation design space:

- **Numeric** preserves the reported score magnitude and distance between dimensions.
- **Categorical** preserves an approximate score level in an explicit list.
- **Natural language** uses the same approximate levels as the categorical representation but changes the surface form from labels to prose.
- **Ordinal** preserves relative ordering while removing exact magnitude and distance.
- **Top-*k*** supplies only the highest-scoring dimensions and therefore tests sparse emphasis and omission.

The source scores range from 1 to 5. The categorical and natural-language representations use cut-points at 1.5, 2.5, 3.5, and 4.5. These values are the midpoints between the five source-scale anchors, so they implement a transparent equal-width mapping to the nearest anchor-level category. They are not presented as psychologically validated boundaries between, for example, "high" and "neutral" moral concern. Their purpose is to make the loss of numerical precision deterministic and reproducible.

For ordinal and top-*k* renderings, scores with an absolute difference of at most 0.05 are represented as tied. The same tolerance is applied at the top-three boundary so that a dimension effectively tied

with the third-ranked score is retained. The value 0.05 is a pragmatic prompt-construction convention at the reported score precision: it avoids turning a difference of five hundredths on a five-point aggregate scale into a strict claim that one concern outranks another. It is not derived from MFT, MFQ-2 validation, or a psychometric minimum-difference threshold. Because alternative tolerances were not evaluated, conclusions involving ordinal and top-k representations are conditional on this rule; Section 7 returns to this limitation.

Table 3.2 summarizes the complete deterministic rendering rules and the information each representation retains or highlights.

Table 3.2: Experimental conditions and deterministic rendering rules for moral-value profile representations. The baseline contains no explicit profile; all other conditions originate from the same six-dimensional source vector.

Condition	**Rendering rule**	**What the representation keeps or highlights**
Baseline	General task instruction and target text, without an annotator moral-value profile	No explicit annotator-specific value information; unpersonalized reference
Natural language	Each score becomes a phrase: strongly prioritizes (≥ 4.5), moderately values (≥ 3.5), is neutral toward (≥ 2.5), places little emphasis on (≥ 1.5), or largely disregards	Approximate score level expressed in short prose
Categorical	Each score becomes Very High (≥ 4.5), High (≥ 3.5), Neutral (≥ 2.5), Low (≥ 1.5), or Very Low	Approximate score level shown in an explicit list
Numeric	Each dimension is reported as a score out of 5, rounded to two decimals	Exact reported scores and score differences between dimensions
Ordinal	Dimensions are ranked without numeric values; scores within 0.05 are represented as tied	Rank order without exact scores or score differences
Top-k	The top three dimensions are reported, including additional dimensions within 0.05 of the third-ranked score	Only the highest-scoring dimensions; lower-ranked dimensions are omitted

The no-profile baseline is not treated as a sixth representation because it does not render the moral-value vector. It provides a reference for the model's response under the same general task without explicit annotator-specific value information. Representation spans and cross-representation disagreement are therefore computed over the five profile-conditioned representations only.

3.5. Prompting conditions

The experiment compares profile-only zero-shot prompting with three same-annotator few-shot variants. A zero-shot profile-conditioned prompt contains the task instruction, target text, and one rendering of the target annotator's moral-value profile. A few-shot prompt additionally contains up to $k = 3$ labelled examples from the same annotator. The target item is always excluded from the candidate pool.

All prompts request a constrained response: `Y` for offensive or `N` for not offensive. Responses that could not be parsed as one of these two labels were marked invalid and excluded only from the analyses that require a valid binary prediction. Unparseable outputs were assigned an out-of-domain sentinel value for processing. This value was not treated as a third class and did not enter any metric as a numerical prediction; Appendix A states the exact convention.

The fixed budget of $k = 3$ creates a controlled comparison among selection strategies while keeping prompts small enough to apply consistently across annotators and models. It is not claimed to be the optimal number of demonstrations. No explicit label-balancing constraint is imposed, so selected examples reflect the available same-annotator pool and the selection rule.

Table 3.3 summarizes the three selection strategies.

Table 3.3: Same-annotator few-shot example-selection strategies.

Strategy	Selection rule
Random	Uniformly samples valid same-annotator historical examples using a fixed seed.
Retrieval	Encodes texts with `all-MiniLM-L6-v2` and selects the same-annotator examples with the highest cosine similarity to the target text [35].
Diversity	Uses the same embedding space and farthest-point sampling to select examples distant from those already selected under $1 - \cos(\cdot, \cdot)$.

Random selection provides a non-semantic reference. Retrieval tests whether locally similar prior judgements help for the target item. Diversity tests whether broader coverage of an annotator's available examples is more informative than local similarity under the same budget. Appendix A gives the full prompt templates and operational details.

3.6. Models and inference settings

Three instruction-tuned LLMs are evaluated. The set is not presented as representative of all language models; it was selected to provide three interpretable contrasts under the available compute budget. Qwen2.5-72B-Instruct is the primary large model. Qwen2.5-7B-Instruct provides a within-family size comparison in which the model family and instruction-tuning lineage are held more nearly constant. Llama-3.3-70B-Instruct provides a comparably sized cross-family comparison. The Qwen2.5 and Llama 3 technical reports document the corresponding model families [36, 37].

- **Qwen2.5-72B-Instruct (Qwen-72B):** primary model for the main analysis.
- **Qwen2.5-7B-Instruct (Qwen-7B):** smaller checkpoint for the within-Qwen comparison.
- **Llama-3.3-70B-Instruct (Llama-70B):** comparably large checkpoint for a cross-family robustness comparison.

A strict binary-output screening was performed before the non-Qwen comparison model was fixed. The screening covered 1,240 items for GPT-4o-mini, GPT-OSS-120B, Grok-4.1-Fast, and Llama-3.3-70B-Instruct. A response was valid only if it could be read as the required `Y/N` output. Llama-3.3-70B-Instruct produced no invalid responses on this subset and was retained. GPT-OSS-120B produced invalid responses for all 1,240 items and was excluded. Under the three-model budget, the final set was chosen to preserve the planned within-Qwen size comparison and a comparably sized cross-family comparison. The screening is a practical selection procedure, not evidence that the retained checkpoints are generally superior.

A preliminary temperature pilot evaluated Qwen-72B at $T = 0$, $T = 0.2$, and $T = 0.7$, each over three runs. The full experiments use $T = 0.2$ based on the observed balance between agreement and run-to-run stability in that pilot. This is a practical setting choice rather than a claim of global optimality. Final experiments use seeds 42, 521, and 1314, and reported means and standard deviations are calculated across those runs.

3.7. Evaluation

Evaluation uses complementary measures because no single metric captures both agreement with the observed label and sensitivity to profile representation.

Accuracy and Macro-F1.

Accuracy is the proportion of predictions equal to $y_{a,i}$. Macro-F1 computes F1 separately for the two classes and averages them, giving equal class-level weight when the label distribution is imbalanced. These metrics summarize aggregate agreement with observed annotator judgements.

McNemar tests.

McNemar's test compares two conditions on the same instances and assesses whether their discordant error counts differ [38]. The reported matrices encode the test outcome at $\alpha = 0.05$: 1 denotes a

significant paired difference and 0 denotes no significant difference. The matrices do not report exact p-values, effect sizes, or direction. Because no multiple-comparison correction was applied, they are interpreted as descriptive paired evidence rather than as a confirmatory ranking of all condition pairs.

Cross-representation decision disagreement.

For each annotator–item pair, the five profile-conditioned binary predictions are compared while the annotator, item, source profile, model, and prompting setting remain fixed. Define

$$D_{a,i}^{(m,s)} = \begin{cases} 1, & \text{if the five predictions are not identical,} \\ 0, & \text{otherwise.} \end{cases}$$

The disagreement rate is the mean of $D_{a,i}^{(m,s)}$ over rows for which all five profile-conditioned outputs are valid binary labels. The measure does not indicate which prediction is correct; it answers whether changing only the representation changes the final decision.

3.8. Analysis plan

RQ1 is addressed using zero-shot Accuracy and Macro-F1, representation spans, pairwise McNemar comparisons, and cross-representation disagreement. The representation span is the difference between the best and worst profile-conditioned result within a model and setting. The no-profile reference is excluded because it does not encode the profile.

RQ2 compares zero-shot representation sensitivity with the three few-shot settings. Lower few-shot disagreement or a narrower metric span is interpreted as lower observed sensitivity to profile representation under the tested prompt context; the analysis does not identify the mechanism responsible for the change.

All analyses are repeated for the three selected models. These comparisons are robustness checks over the specified checkpoints. They do not support population-level claims about all LLMs or isolate the effects of parameter count, model family, tokenizer, training data, or instruction tuning.

4

Results

This chapter reports the results in the order of the research questions. Section 4.1 first evaluates zero-shot representation sensitivity. Section 4.2 then compares the three same-annotator few-shot strategies. Sections 4.3–4.5 examine representation sensitivity using aggregate spreads, instance-level decision disagreement, and paired McNemar comparisons. The chapter closes with cross-model robustness and a direct summary of RQ1 and RQ2.

4.1. Zero-shot representation effects

Table 4.1 reports zero-shot agreement with observed annotator judgements across the unpersonalized reference and the five profile representations.

Table 4.1: Zero-shot performance across models and profile representations. Results are mean ± standard deviation. Bold marks the best profile-conditioned value within each model and metric.

Model	Prompt condition	Macro-F1	Accuracy
Llama-70B	Baseline	0.6100 ± 0.0003	0.6435 ± 0.0004
	Natural Language	0.6226 ± 0.0007	0.6747 ± 0.0012
	Categorical	**0.6234 ± 0.0009**	0.6728 ± 0.0012
	Numeric	0.6194 ± 0.0019	**0.6789 ± 0.0002**
	Ordinal	0.6124 ± 0.0016	0.6573 ± 0.0034
	Top-*k*	0.6059 ± 0.0010	0.6297 ± 0.0018
Qwen-72B	Baseline	0.5892 ± 0.0007	0.6024 ± 0.0019
	Natural Language	**0.6047 ± 0.0035**	**0.6218 ± 0.0051**
	Categorical	0.5956 ± 0.0039	0.6054 ± 0.0050
	Numeric	0.5844 ± 0.0035	0.5930 ± 0.0047
	Ordinal	0.5791 ± 0.0025	0.5889 ± 0.0038
	Top-*k*	0.5663 ± 0.0021	0.5717 ± 0.0031
Qwen-7B	Baseline	0.5750 ± 0.0015	0.5856 ± 0.0021
	Natural Language	**0.5775 ± 0.0021**	**0.5911 ± 0.0035**
	Categorical	0.5577 ± 0.0034	0.5617 ± 0.0040
	Numeric	0.5262 ± 0.0031	0.5267 ± 0.0033
	Ordinal	0.5469 ± 0.0034	0.5499 ± 0.0040
	Top-*k*	0.5595 ± 0.0020	0.5659 ± 0.0028

The zero-shot results show representation sensitivity, but the strongest representation is model-conditional. Natural-language profiles produce the best Macro-F1 and Accuracy for both selected Qwen models. Llama-70B instead favours the categorical representation for Macro-F1 and the numeric representation for Accuracy. The same measured signal is therefore not used identically by the three selected models when represented in different ways.

The comparison with the no-profile reference also shows that explicit profile information is not uniformly associated with higher aggregate agreement. Some profile-conditioned prompts outperform the baseline, while others fall below it. The largest descriptive gap in Table 4.1 occurs for Qwen-7B with the numeric profile: its zero-shot Macro-F1 is 0.5262, compared with 0.5775 for natural language and 0.5750 for the baseline. The corresponding Accuracy is 0.5267, compared with 0.5911 and 0.5856. Pairwise McNemar tests indicate significant differences for both natural language versus numeric and the no-profile reference versus numeric at $\alpha = 0.05$ ($p < 0.05$). These comparisons support the conclusion that the observed Qwen-7B difference is representation-dependent.

4.2. Few-shot prompting and example selection

Table 4.2 summarizes the strongest profile-conditioned result in each model and prompting setting; complete tables are reported in Appendix D.

Table 4.2: Best value-profile representation by model and prompting setting. The no-profile baseline is excluded from the ranking.

Model	Setting	Best Macro-F1 profile	Macro-F1	Best Accuracy profile	Accuracy
Llama-70B	Zero-shot	Categorical	0.6234	Numeric	0.6789
	Few-shot random	Top-k	0.6534	Categorical	0.7178
	Few-shot retrieval	Categorical	0.6654	Categorical	0.7287
	Few-shot diversity	Categorical	0.6582	Categorical	0.7166
Qwen-72B	Zero-shot	Natural Language	0.6047	Natural Language	0.6218
	Few-shot random	Categorical	0.6639	Natural Language	0.7010
	Few-shot retrieval	Categorical	0.6759	Natural Language	0.7105
	Few-shot diversity	Categorical	0.6601	Natural Language	0.6962
Qwen-7B	Zero-shot	Natural Language	0.5775	Natural Language	0.5911
	Few-shot random	Categorical	0.6405	Categorical	0.6999
	Few-shot retrieval	Categorical	0.6467	Categorical	0.6913
	Few-shot diversity	Categorical	0.6455	Categorical	0.7008

Few-shot prompting produces higher best-case aggregate agreement than zero-shot prompting for all three selected models. Retrieval-based few-shot prompting yields the highest Macro-F1 in each model: 0.6654 for Llama-70B, 0.6759 for Qwen-72B, and 0.6467 for Qwen-7B. Relative to each model's strongest zero-shot profile, these are increases of 0.0420, 0.0712, and 0.0692 respectively.

Accuracy is less uniform. Retrieval gives the highest Accuracy for Llama-70B and Qwen-72B, while diversity gives the highest Qwen-7B Accuracy by a small margin over random. Retrieval therefore provides the most consistent advantage for Macro-F1 under the fixed $k = 3$ budget, while random and diversity remain competitive in model-specific Accuracy results.

4.3. Representation sensitivity across settings

Figure 4.1 visualizes the distribution of performance across the five profile-conditioned representations. The no-profile baseline is excluded because it does not render the profile. Each box contains the five representation scores for each of the three models.

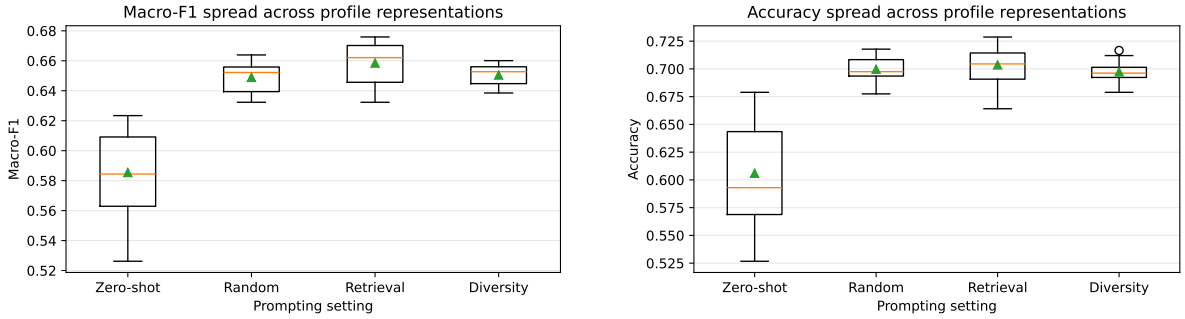

Figure 4.1: Distribution of performance across value-profile representations by prompting setting. Each box contains 15 values: five profile representations across three models. Left: Macro-F1. Right: Accuracy.

For both Macro-F1 and Accuracy, zero-shot prompting has the widest aggregate spread across representations, whereas the three few-shot settings are more compressed. This pattern is consistent with the profile carrying a larger share of the explicit annotator-specific information in zero-shot prompting. In few-shot settings, previous same-annotator judgements provide an additional behavioural signal, so changes in the abstract profile representation have a smaller aggregate association with performance.

The compression should not be interpreted as equivalence. A narrow score range can occur even when conditions make different predictions on individual items. The following analyses therefore examine the final binary decisions directly rather than relying only on aggregate performance.

4.4. Few-shot context and cross-representation disagreement

Table 4.3 reports the proportion of annotator–item pairs for which at least one of the five profile representations changes the parsed binary prediction. This diagnostic holds the text, annotator, source profile, model, and prompting setting fixed. It does not measure correctness relative to the observed label.

Table 4.3: Reduction in cross-representation decision disagreement from zero-shot to few-shot prompting. Rates are unweighted averages across the three experimental seeds; invalid non-binary predictions are excluded.

Model	Setting	Rate (%)	Drop (pp)	Red. (%)
Qwen-72B	Zero-shot	20.91	–	–
	Random	14.36	6.55	31.3
	Retrieval	12.29	8.62	41.2
	Diversity	15.12	5.79	27.7
	Few-shot avg.	13.92	6.99	33.4
Qwen-7B	Zero-shot	23.88	–	–
	Random	15.62	8.26	34.6
	Retrieval	16.74	7.14	29.9
	Diversity	16.02	7.86	32.9
	Few-shot avg.	16.13	7.75	32.5
Llama-70B	Zero-shot	22.52	–	–
	Random	11.24	11.28	50.1
	Retrieval	10.83	11.69	51.9
	Diversity	11.50	11.02	48.9
	Few-shot avg.	11.19	11.33	50.3

Zero-shot disagreement exceeds 20% for every model: 20.91% for Qwen-72B, 23.88% for Qwen-7B, and 22.52% for Llama-70B. Averaged over the three few-shot strategies, the rates fall to 13.92%, 16.13%, and 11.19%, corresponding to relative reductions of 33.4%, 32.5%, and 50.3%. Under the

tested prompts, same-annotator examples are therefore associated with a lower observed rate of representation-dependent decision changes.

The reduction is substantial but incomplete. Disagreement remains above 10% for every model in the few-shot average. A system could thus have similar aggregate Accuracy or Macro-F1 across profile representations while assigning different labels to a non-trivial set of individual texts. Illustrative Qwen-72B zero-shot cases are reported in Appendix B.

The exploratory sub-category analysis in Appendix C shows that disagreement is not uniform across item types. For Qwen-72B, zero-shot disagreement is highest for random items, fairness-related items, and several social-group sub-categories, while care, purity, and loyalty have lower rates. Retrieval-based few-shot prompting produces the lowest disagreement in 9 of the 11 Qwen-72B sub-categories. This breakdown is descriptive: it identifies where sensitivity is concentrated but does not establish that category membership causes the effect.

4.5. Pairwise McNemar significance

The McNemar matrices provide a paired-error view. In zero-shot prompting, many representation pairs make significantly different errors on the same instances. Significant differences also remain in few-shot prompting even where aggregate scores are close. The matrices therefore support the same qualification as the disagreement diagnostic: demonstrations are associated with lower representation sensitivity, but they do not make the representations behaviourally interchangeable. Complete matrices are reported in Appendix F.

Because the matrices encode only significance at $\alpha = 0.05$, they are not used to rank representations. They do not show the direction of improvement, the magnitude of the difference, or a multiple-comparison-adjusted inferential conclusion. Their role is to verify that aggregate similarity can conceal paired prediction differences.

4.6. Cross-model robustness

The two central empirical patterns recur across all three selected models. First, each model exhibits higher cross-representation disagreement in zero-shot prompting than in its average few-shot condition. Second, no profile representation is strongest for every model, setting, and metric. These recurrences provide robustness over the specified checkpoints.

The magnitude and direction of individual representation effects remain model-conditional. Qwen-7B is particularly weak with numeric zero-shot profiles, whereas Llama-70B obtains its highest zero-shot Accuracy with the numeric representation. Llama-70B also shows a larger relative reduction in disagreement under few-shot context than either Qwen model. The selected-model comparison cannot isolate whether these differences arise from scale, model family, tokenizer, pre-training data, or instruction tuning; those factors are entangled.

4.7. Result summary

For RQ1, alternative deterministic renderings of the same moral-value profile change both aggregate agreement and instance-level predictions in zero-shot prompting. The strongest representation depends on the model and metric. Explicit profile information is not uniformly better than the no-profile reference, demonstrating that the value of the signal is representation-dependent.

For RQ2, every selected model has lower cross-representation disagreement under each few-shot strategy than under zero-shot prompting. Same-annotator demonstrations are therefore associated with more representation-stable predictions, although disagreement and significant paired differences remain. Retrieval provides the strongest Macro-F1 for all three models and the lowest Qwen-72B disagreement in most sub-categories. No single representation or selection strategy performs consistently best across every tested model, setting, and metric. The more general conclusion is that profile representation and example selection are consequential design variables in annotator-conditioned judgement prediction.

5

Discussion

The results support a conclusion that is broader than selecting the best-performing prompt. When structured information is used to personalize an LLM, the representation of that information becomes part of the intervention. The same moral-value scores produced different aggregate results, paired error patterns, and final binary decisions when they were expressed as numbers, categories, rankings, sparse top-k lists, or natural language. Same-annotator examples reduced this sensitivity but did not remove it. This section interprets those findings in relation to the two research questions and the literature reviewed in Section 2.

5.1. Positioning the findings relative to adjacent work

The contribution is clearest when the findings are placed at the intersection identified in Section 2.6. D3CODE establishes that repeated individual offensiveness judgements and measured moral-value profiles can be studied together, and it reports associations between values and disagreement [15]. The present experiment adds a different result: even when the measured profile, annotator, and item are held fixed, an LLM's use of that profile is not representation-invariant. A useful personalization signal at the dataset level is therefore not automatically a stable signal once it is converted into a prompt.

Value Profiles and SOLAR likewise show that value-based abstractions can encode or model individual variation [27, 28]. Those studies primarily address whether value information can represent human variation or value conflicts. This thesis complements them by isolating *representation robustness*: a profile can contain predictive information while still producing different attributed judgements when its magnitude, ordering, sparsity, or wording is changed. This identifies profile representation as an additional design variable that value-conditioned studies need to document and test.

Prompt-sensitivity research shows that model outputs can change under variations in prompt format [17, 16]. The present findings extend that concern to an annotator-conditioned setting. Here, representation sensitivity is not only reflected in aggregate benchmark scores: it can change whether the model predicts that a specified person would judge a particular text as offensive. Profile representation is therefore consequential for the substantive attribution being made about an individual.

Finally, annotator-specific modelling demonstrates that individual labels can be retained and predicted rather than removed by majority aggregation [8]. The current study retains that target but asks how the personalizing information is supplied to an instruction-tuned LLM. Same-annotator examples reduce but do not eliminate dependence on the abstract profile representation, narrowing the research gap to how reliably a measured profile is operationalized through prompt representations.

5.2. Profile representation is part of the personalization treatment

RQ1 asks whether deterministic renderings of a fixed moral-value profile affect zero-shot prediction. The answer is yes under each of the selected models. The effect appears in several complementary forms: the profile-conditioned Accuracy and Macro-F1 values span non-trivial ranges, many representation pairs produce significantly different errors, and more than one fifth of annotator–item pairs receive different binary predictions across the five representations in zero-shot prompting.

This matters because personalization is often described primarily in terms of *which* information is supplied—values, preferences, demographics, history, or a persona description. The experiment shows that *how* the information is supplied can also determine model behaviour. A numerical vector and an ordinal ranking may refer to the same person, but they do not expose the same relations to the model. The numeric representation retains distance; the ordinal representation converts small and large differences alike into ordering relations; the top-k representation removes part of the profile; and natural language introduces lexical and interpretive framing. Consequently, representation should be reported as part of the treatment rather than as an implementation detail hidden behind the phrase "the model was conditioned on a profile."

This interpretation is consistent with prior findings that prompt-format features can alter LLM outputs even when task meaning is intended to remain stable [17, 16]. The present study extends that concern to value-conditioned individual judgement prediction. Here, representation sensitivity is not merely a benchmark nuisance: it changes which judgement is attributed to a specified annotator. This increases the methodological importance of profile representation because a seemingly minor representation choice can affect the substantive conclusion drawn about a person's perspective.

5.3. Information preservation helps explain, but does not fully determine, the pattern

The five representations differ along more than one dimension. Numeric profiles preserve the reported scores, categorical and natural-language profiles discretize them, ordinal profiles preserve only rank, and top-k profiles intentionally omit lower-ranked values. The results therefore cannot be reduced to a single contrast between "structured" and "unstructured" prompting. They reflect an interaction between information preservation, lexical realization, ordering, and model-specific processing.

Natural-language and categorical profiles provide the cleanest partial control because they use the same score thresholds. Their difference is primarily surface realization: prose phrases such as "moderately values" versus explicit category labels such as "High." Different predictions under these two conditions show that lexical form can matter even when the coarse-grained value information is approximately held constant. This supports the view that surface realization can itself be an active modelling choice rather than a neutral presentation layer.

Numeric profiles illustrate the limits of assuming that greater information retention is always better. Qwen-7B performs substantially worse with numbers in zero-shot prompting, whereas Llama-70B obtains its highest zero-shot Accuracy with the numeric representation. A possible interpretation is that the models differ in how reliably they compare scalar values and map those values to the task. Numeric magnitude effects have been observed in LLMs more generally [26], but controlled ablations would be needed to separate this mechanism from differences in model family, scale, tokenizer, training data, and instruction tuning.

The top-k and ordinal conditions test design hypotheses about priority and salience. Their mixed results do not support a general recommendation to simplify profiles by retaining only the highest values or only their rank order. In some contexts, emphasizing the most salient dimensions may reduce prompt complexity; in others, omitted or distance information may be necessary. The empirical lesson is to evaluate the transformation rather than infer its value from intuitiveness alone.

5.4. Moral values are informative but incomplete

The experiments show that a moral-value profile can change the model's approximation of individual judgements, but they do not establish that the profile fully explains those judgements. This distinction is central to the motivation and interpretation of the thesis. MFT and the MFQ-2 provide a structured account of recurring moral concerns [13, 14]; D3CODE links those measurements to repeated offensiveness annotations [15]. The profile is therefore more theoretically grounded than an arbitrary persona description. It remains a compressed measurement.

Many variables that may influence offensiveness are absent: direct experience with targeted language, membership in a referenced group, knowledge of the speaker, local norms, pragmatic intent, irony, humour, and contextual information beyond the item text. Two annotators with similar moral

scores can therefore make different judgements, and one annotator can judge two superficially similar texts differently. The best observed results, which remain below 0.68 Macro-F1 and 0.73 Accuracy, are consistent with this limited role. The profile contains some useful signal but does not make individual judgement prediction close to deterministic.

The comparison with the no-profile condition reinforces this point. Some profile renderings improve agreement, while others reduce it. Supplying measured information is not automatically beneficial if the model cannot use the representation appropriately or if the profile is weakly informative for the specific item. The baseline should likewise not be interpreted as objective or neutral. It reflects the model's learned and aligned normative regularities in the absence of explicit annotator information. The experiment compares two kinds of input context, not an individual bias against an unbiased ground truth.

5.5. Same-annotator examples reduce observed representation sensitivity

RQ2 asks whether representation sensitivity changes when the prompt includes direct evidence of the annotator's previous behaviour. The cross-representation diagnostic gives a consistent answer: every few-shot condition has lower disagreement than zero-shot prompting for each model. The relative reduction ranges from approximately one third for the Qwen models to one half for Llama-70B when the three strategies are averaged.

A plausible interpretation is that demonstrations reduce the profile's informational burden. In zero-shot prompting, the moral-value description is the only explicit annotator-specific signal. In few-shot prompting, the model can also infer local decision patterns from previous labelled texts. This behavioural evidence may constrain the prediction even when the abstract profile is expressed differently. The narrower aggregate spreads are consistent with the same interpretation.

Demonstrations can also convey task format, label semantics, and properties of the input distribution in addition to annotator-specific behaviour [30]. Because the current design does not separate these effects, the observed reduction is interpreted associatively rather than mechanistically.

The remaining disagreement is equally important. Few-shot averages still exceed 10% for all models, and McNemar comparisons remain significant for many representation pairs. Demonstrations make representations more stable, not interchangeable. A study that reports only the highest few-shot Accuracy could therefore overlook the fact that alternative profile representations continue to change individual decisions.

5.6. Why retrieval is strong under the tested budget

Retrieval-based selection gives the strongest Macro-F1 for all three selected models and the lowest disagreement in 9 of 11 Qwen-72B sub-categories. This suggests that local semantic relevance is particularly useful under a small $k = 3$ budget. A retrieved example can show how the annotator judged a text resembling the target, thereby providing a comparatively direct behavioural analogy.

Diversity selection pursues a different objective: representing a broader region of the annotator's historical examples. That coverage may be valuable when more demonstrations can be included or when the target requires a global model of the annotator. Under the three-example budget, broad coverage can select examples that are less directly relevant to the target. Retrieval is therefore strongest within the tested setting, where local relevance may be particularly valuable.

The observation aligns with prior in-context learning research showing that demonstration selection affects performance and that semantically relevant examples can outperform random selection [31, 33]. The annotator-conditioned setting adds an important qualification: relevance must be sought within the same annotator's history. Retrieval is useful here because it jointly preserves annotator identity and increases textual similarity.

5.7. Cross-model robustness and the limits of the comparison

The central patterns recur across Qwen-72B, Qwen-7B, and Llama-70B: zero-shot predictions are more representation-sensitive than few-shot predictions, and no single representation dominates every setting. This recurrence reduces the likelihood that the main conclusion is an idiosyncrasy of one checkpoint.

At the same time, the models differ substantially in which representation performs best and in the magnitude of few-shot stabilization. These differences prevent a model-independent recommendation such as “always use natural language” or “always preserve exact numbers.” The Qwen within-family comparison suggests that scale may be relevant, particularly for numeric profiles, but it does not identify scale as the cause. The cross-family comparison introduces additional uncontrolled differences. The three-model analysis should therefore be read as a robustness axis over selected checkpoints, not as a factorial experiment on model architecture or size.

5.8. Implications for experimental practice

The findings imply several reporting requirements for research on personalized or value-conditioned LLMs. First, studies should state the exact representation used for profiles and justify transformations that discard precision, impose order, or select only salient dimensions. Second, when feasible, multiple plausible representations should be evaluated rather than assuming semantic equivalence. Third, aggregate metrics should be supplemented with paired or instance-level analyses when the application concerns individual decisions. Similar Macro-F1 values do not guarantee that the same items receive the same labels.

Fourth, profile information and behavioural demonstrations should be treated as distinct personalization sources. A value profile is an abstract measurement; examples are direct observations of prior behaviour. Their interaction can change both accuracy and stability. Finally, the unpersonalized baseline should be described accurately: it is a model-specific generic response under a chosen instruction, not an objective moral standard.

These practices follow from the broader perspectivist motivation of the thesis. Retaining individual labels makes disagreement visible, but it also creates responsibility to distinguish descriptive prediction from prescriptive decision-making. The next section examines how the results can be used responsibly in applied systems and explains how the work fits within the DSAIT programme.

6

Applicability, societal impact, and DSAIT relevance

6.1. Applicability to personalized and pluralistic language systems

The practical relevance of this thesis lies in a possible class of systems that must account for heterogeneous human judgements rather than produce one undifferentiated response. General-purpose LLMs reflect normative regularities acquired from training and alignment data, but those regularities need not match the judgement of a particular individual or community. Santurkar et al. document systematic differences between language-model opinions and the responses of multiple demographic groups [21]. Research on pluralistic alignment consequently asks how AI systems might represent or negotiate multiple legitimate values without silently collapsing them into one assumed consensus [39].

The present work contributes a limited empirical result to that broader problem. It evaluates whether a measured moral-value profile changes an LLM's approximation of one recorded annotator judgement and whether the representation of that profile changes the output. It does not provide a complete architecture for pluralistic alignment, nor does it imply that moral values should determine what speech is permitted. The strongest observed profile-conditioned setting remains below 0.68 Macro-F1 and 0.73 Accuracy, so the results should be interpreted as benchmark evidence rather than deployment-level validation.

A plausible research or product use is preference-sensitive assistance. For example, a tool could help researchers examine how different users might interpret a text, identify items with high predicted disagreement, or prioritize cases for additional human review. The cross-representation diagnostic is especially relevant because it can flag cases where the model's prediction is unstable under alternative representations of the same profile. Such instability is itself useful information: it indicates that the personalized output depends heavily on a design choice and should not be treated as robust.

6.2. Assisted assessment rather than autonomous enforcement

The results support an assisted-assessment role more strongly than autonomous moderation. In an assisted design, the model could surface a predicted annotator-specific judgement, its uncertainty, and whether alternative profile renderings change the decision. A human reviewer or accountable policy process would then decide how the information should be used. This preserves the descriptive value of modelling individual perspectives without allowing a noisy prediction to become an unreviewed enforcement action.

An autonomous system would require evidence that is absent from the present study. It would need validation against an explicit application policy, calibration on held-out users, robust uncertainty estimates, monitoring under distribution shift, and evaluation of downstream harms from false positives and false negatives. It would also need to separate a user's personal preference from platform-level obligations, legal constraints, and protections for other users. Agreement with a historical annotation is not sufficient evidence for any of these requirements.

6.3. Societal risks and governance requirements

Personalization through moral profiles introduces risks in addition to possible benefits. The first is *misrepresentation*. A six-dimensional profile can be interpreted too strongly, giving the impression that a person has a fixed and predictable moral identity. Cheng et al.'s analysis of caricature in persona-based LLM simulations shows why simplified descriptions should not be treated as faithful models of whole people [22]. The present results reinforce this caution because different renderings of the same profile can lead to different predicted judgements.

The second risk is *privacy and profiling*. Moral-value information concerns beliefs and normative orientation. Any collection or use of such data should be voluntary, purpose-limited, transparent, and protected against secondary use. A system should not infer moral profiles from unrelated user behaviour and then apply them without consent. Even self-reported profiles can become sensitive when linked to content preferences, political views, religious commitments, or identity.

The third risk is *normative fragmentation*. Personalization can improve relevance for individual users, but a platform cannot resolve every conflict by presenting each person with a separate moral standard. Content moderation affects speakers, targets, bystanders, and communities, not only the user receiving the prediction. A responsible system therefore needs an explicit distinction between personal filtering or assistance and collective governance. Profile-conditioned outputs may inform the former; they should not silently replace the latter.

The fourth risk is *design-induced arbitrariness*. If two semantically plausible profile representations lead to different decisions, a developer can unintentionally or strategically select the representation that produces a preferred outcome. Auditing should therefore include representation sensitivity tests, documentation of conversion rules, and paired analyses of changed decisions. The main empirical contribution of this thesis is directly relevant to this governance requirement.

6.4. Relation to Data Science and Artificial Intelligence Technology

The study fits within Data Science and Artificial Intelligence Technology (DSAIT) through the integration of data measurement, machine-learning experimentation, statistical evaluation, and responsible system design.

Data science.

The project treats annotation disagreement as structured data rather than automatically removing it through aggregation. It requires linking repeated observations to annotator-level measurements, constructing a coverage-oriented sample, transforming a multidimensional measurement into alternative representations, and evaluating how those transformations affect downstream conclusions. The work therefore concerns the full data-science pipeline: data-generating assumptions, measurement, sampling, feature representation, evaluation, and interpretation. The distinction between an observed individual label and a majority-vote target is a data-modelling choice with direct consequences for the research question.

Artificial intelligence.

The project evaluates instruction-tuned LLMs as conditional predictors and studies two central AI mechanisms: prompt-based personalization and in-context learning. It compares model families and sizes as robustness checks, examines sensitivity to input representation, and distinguishes aggregate predictive quality from instance-level behavioural stability. These are core AI reliability questions because a system can achieve a similar average score while responding differently to individual inputs under minor prompt changes.

Technology and engineering.

The experiment operationalizes an end-to-end inference pipeline with deterministic profile rendering, example selection, repeated model calls, constrained output parsing, and reproducible evaluation. The results translate into engineering requirements for deployed systems: exact profile-rendering rules

should be versioned; invalid outputs should be monitored; representation changes should trigger regression tests; and personalized outputs should expose uncertainty and provenance. Retrieval-based demonstration selection also represents a concrete system component whose benefits and limitations can be tested independently.

Responsible innovation.

DSAIT systems are not evaluated only by predictive performance. They must also be assessed in relation to the people represented by the data and affected by the output. This thesis addresses that requirement by distinguishing descriptive modelling from prescriptive moderation, limiting claims about psychological representation, and identifying privacy, accountability, and pluralism concerns. The technical conclusion—that representation changes decisions—is therefore also a responsible-AI conclusion: data and prompt design choices shape whose judgement a system appears to reproduce.

7

Limitations and future work

The conclusions of this thesis are deliberately bounded by the dataset, sampling design, representation rules, prompting budget, model set, and evaluation procedure. This section explains why each limitation arises, how it constrains interpretation, and which follow-up would address it.

7.1. Limitations

One dataset and one task formulation.

D3CODE was selected because it uniquely combines repeated individual offensiveness judgements with annotator-level moral-value measurements. This makes it well matched to the research question, but it also limits external validity. The results may depend on D3CODE's item distribution, annotation protocol, cultural composition, English-language content, and binary label definition. They do not automatically generalize to other harmful-language datasets, multilingual inputs, continuous rating scales, or prescriptive moderation policies.

Coverage-oriented rather than fully representative sampling.

The 354-annotator subset was constructed to cover six regions of the observed moral-profile space while retaining all annotations from selected annotators. Equal allocation improves profile coverage and supports same-annotator analyses, but it is not a probability sample designed to reproduce the exact D3CODE population distribution. The finite-population formula is used as a transparent heuristic rather than a power analysis for every condition comparison. Population-level prevalence estimates should therefore not be inferred from the balanced subset.

Moral values are only one personalization signal.

The MFQ-2 profile is measured and theoretically grounded, but it does not represent an annotator's complete identity, experience, or contextual knowledge. The study cannot determine how much unobserved personal information contributes to judgement variation. It also cannot conclude that moral values cause the observed labels. The results establish conditional predictive associations under the prompt design, not psychological mechanisms.

Prompt components are not fully disentangled.

Profile-conditioned prompts differ from the no-profile reference in both information and perspective framing. Few-shot prompts add labelled examples, but the experiment does not include examples-only prompts, matched framing without profile values, shuffled-annotator examples, or examples from another annotator. Consequently, a reduction in representation sensitivity cannot be attributed exclusively to direct evidence of the target annotator's latent perspective. Demonstrations may also convey label semantics, task format, or general input-distribution information.

Five representations do not exhaust the design space.

The selected representations create interpretable contrasts, but many alternatives remain: JSON-like structures, natural language paired with numbers, explanations of score ranges, calibrated verbal probabilities, dimension-specific definitions, or learned profile embeddings. The findings show that

representation matters among the tested representations; they do not identify the global optimum over all possible encodings.

Rendering thresholds are pragmatic.

The category boundaries are midpoint mappings of the 1–5 source scale, and the 0.05 tie tolerance is an operational convention. These rules are reproducible but not psychologically validated cut-offs. No tolerance sensitivity analysis was run. Results for ordinal and top-k profiles are therefore conditional on the chosen tie rule, and small changes near the boundaries could alter which dimensions appear tied or selected.

A fixed few-shot budget.

The use of $k = 3$ keeps the comparison controlled and computationally feasible, but it cannot show whether disagreement decreases monotonically as more same-annotator examples are provided. Diversity selection in particular may require a larger budget before broader coverage becomes advantageous. The ranking of selection strategies is therefore specific to the tested prompt budget.

Three selected models.

Qwen-72B, Qwen-7B, and Llama-70B provide meaningful within-family and cross-family contrasts, but they are not a representative sample of LLMs. Model family, scale, tokenizer, pre-training data, and instruction tuning remain entangled. The output-compliance screening also introduces a practical selection filter. The study supports robustness over the selected checkpoints, not a population-level conclusion about all language models.

Limited statistical characterization.

The analysis reports means and standard deviations over three seeds, binary McNemar significance matrices, and disagreement rates. Three seeds provide only a coarse estimate of stochastic variation. The McNemar matrices omit p-values, effect sizes, and direction, and no multiple-comparison correction is applied. The disagreement reductions are descriptive point estimates without bootstrap or analytic confidence intervals. Statistical significance and practical importance should therefore not be conflated.

No deployment evaluation.

The experiment measures agreement with historical annotator labels. It does not evaluate calibration on new users, distribution shift, adversarial prompting, privacy-preserving profile collection, human factors, or the consequences of incorrect decisions. The results are not sufficient to justify autonomous moderation or high-stakes profiling.

7.2. Future work

Replicate across datasets, languages, and subjective tasks.

The strongest test of generality is replication on datasets that retain repeated individual labels and include value or preference measurements. Multilingual harmful-language datasets, moral acceptability tasks, political opinion judgements, and other subjective classification settings could reveal whether representation sensitivity is specific to D3CODE or a broader property of personalized prompting.

Expand the model set through a pre-specified design.

Future work should evaluate more model families and several sizes within each family. A factorial or otherwise pre-specified model sample would help separate scale from family and instruction-tuning effects. Open-weight checkpoints would also enable controlled tokenizer, hidden-state, and attention analyses that are not possible through output-only comparison.

Run representation and threshold sensitivity analyses.

Additional representations should include hybrid natural-language-plus-numeric profiles, explicit definitions of each dimension, structured JSON-like encodings, and calibrated descriptions of uncertainty. For the existing ordinal and top-k representations, the experiment should be repeated with

several tie tolerances, including no tolerance, 0.01, 0.05, and 0.10. This requires regenerating the affected prompts and rerunning the corresponding inference conditions.

Disentangle profile information from framing and examples.

A stronger ablation design should add: (1) perspective wording without profile values, (2) profile values without perspective wording, (3) examples-only prompts, (4) shuffled same-annotator labels, (5) examples from another annotator, and (6) profile–example mismatch conditions. These controls would estimate how much improvement or stabilization comes from the measured profile, direct behavioural evidence, generic additional context, or instruction framing.

Vary the number and composition of demonstrations.

Evaluating several values of k would test whether representation disagreement decreases smoothly or reaches a plateau. Selection could also impose label balance, recency, difficulty, moral-subcategory coverage, or a joint similarity–diversity objective. Such experiments would clarify whether retrieval's current advantage follows from the small budget rather than from an intrinsic superiority.

Analyse representation disagreement by annotator and moral-profile region.

The six sampling clusters were not used as an independent variable in the main study. A follow-up could compare disagreement rates across clusters or continuous regions of the profile space, with suitable uncertainty estimates. Per-annotator analyses could identify whether a small subset of annotators accounts for a disproportionate share of representation sensitivity. These analyses should remain descriptive unless supported by an appropriate inferential design.

Investigate mechanisms with controlled profiles.

Synthetic profiles that vary one dimension at a time could test whether models respond monotonically to score changes and whether the response depends on the rendering. Prompt ablations, counterfactual orderings, and token-level analyses could distinguish numerical-comparison failures from lexical salience effects. Calibration and error analyses by moral dimension could further identify whether a model ignores, overweights, or inconsistently maps specific profile components.

Strengthen statistical reporting.

Future analyses should report full McNemar contingency tables or exact p-values, effect sizes, and correction procedures for families of comparisons. Bootstrap confidence intervals could be added for Accuracy, Macro-F1, representation spans, and disagreement-rate reductions, with resampling designed to respect repeated observations within annotators. More seeds would improve estimates of stochastic variability.

Conduct qualitative and human-centred evaluation.

Cases in which representations disagree should be reviewed systematically to determine whether changes follow interpretable profile cues or unstable prompt behaviour. Human studies could test whether annotators recognize the model's prediction as representative of their judgement and whether explanations based on moral profiles feel accurate or reductive. Any deployment-oriented work should include consent, privacy, contestability, and harm assessment from the outset.

8

Conclusion

This thesis examined how alternative representations of an annotator's measured moral-value profile influence LLM predictions of that annotator's offensiveness judgement. The task was descriptive and annotator-conditioned: the target was the individual's recorded binary label, not an objective offensiveness standard or majority-vote ground truth. D3CODE enabled this design by combining repeated individual judgements with six-dimensional MFQ-2 profiles.

The research gap concerned the interaction among three observations established separately in prior work: individual annotator variation matters, value-based information can represent part of that variation, and LLMs can be sensitive to prompt representation. The thesis narrows that gap by holding the annotator, item, measured profile, model, and prompting setting fixed while changing only the deterministic rendering of the profile. It further tests whether direct behavioural evidence from the same annotator is associated with lower dependence on that rendering.

The first research question concerned zero-shot representation sensitivity. The results show that changing only the deterministic rendering of a fixed profile changes aggregate agreement, paired error behaviour, and final binary decisions. The strongest representation depends on the model and metric. Natural language is strongest for the selected Qwen models in zero-shot prompting, while categorical and numeric representations are stronger for different Llama-70B metrics, and some representations perform worse than the no-profile reference. Cross-representation disagreement above 20% for each model demonstrates that the effect is not limited to small differences in average scores.

The second research question concerned prompting context. Same-annotator few-shot examples are associated with lower representation sensitivity under every tested model and selection strategy. Averaged across the three strategies, disagreement falls by 33.4% for Qwen-72B, 32.5% for Qwen-7B, and 50.3% for Llama-70B relative to zero-shot prompting. Retrieval-based selection produces the strongest Macro-F1 for all three models and the lowest disagreement in most Qwen-72B sub-categories. Nevertheless, few-shot disagreement remains above 10%, and paired differences persist. Demonstrations reduce the influence of profile representation without making the representations interchangeable.

The contribution is the controlled finding that a fixed, independently measured personalization signal can yield different judgements attributed to the same annotator depending on how that signal is represented, and that same-annotator examples reduce but do not eliminate this dependence.

This supports a methodological conclusion: profile representation and behavioural-example selection are consequential parts of the experimental treatment. They should therefore be documented, justified, and evaluated using both aggregate metrics and instance-level comparisons.

For data science and AI practice, this result connects measurement, representation, in-context learning, robustness, and responsible deployment. A personalized system should not report only which information it uses; it should also make visible how that information is encoded and how stable the resulting decision is under plausible alternatives. That requirement is essential when the output is presented as an approximation of an individual's normative perspective.

A

Methodological details

This appendix provides methodological detail needed to reproduce the dataset construction, profile rendering, prompt templates, and example-selection procedures.

A.1. Dataset subset construction

To construct an annotator subset that covers different regions of the moral-profile space, we first created one record per annotator using the identifier `rater_id`. Annotators were then partitioned into six clusters using K-means over the six moral-value dimensions: care, equality, proportionality, authority, loyalty, and purity. We used $k = 6$, 10 random initializations, and a fixed random seed of 42 for reproducibility.

As described in Section 3, the finite-population sample-size calculation gives a target of approximately 353 annotators. The final sample uses 354 annotators so that 59 annotators can be sampled from each of the six clusters. We used a fixed random seed and retained all annotation records belonging to the sampled annotators.

Table A.1: Dataset-subset construction used in the experiments.

Component	Specification
Sampling unit	Annotator, identified by `rater_id`
Cluster construction	K-means over six moral-value dimensions, $k = 6$
K-means initializations	10
Random seed	42
Annotators per cluster	59
Total annotators	354
Text-annotation records	Approximately 12,000
Retained records	All annotations of sampled annotators
Shannon entropy	1.7918 / 1.7918 maximum
Simpson diversity	0.8333 / 0.8333 maximum

Shannon entropy and Simpson diversity summarize how evenly the 354 sampled annotators are distributed across the six clusters. Let n_g be the number of sampled annotators in cluster g, let $n = \sum_{g=1}^{G} n_g$, and let $p_g = n_g/n$ be the proportion assigned to that cluster. Shannon entropy is defined as [40]

$$H = -\sum_{g=1}^{G} p_g \ln p_g.$$

A larger value indicates a more even distribution across clusters. For a fixed number of clusters, the maximum is $\ln G$, reached when every cluster has the same proportion.

The Simpson diversity index used here is defined as [41]

$$S = 1 - \sum_{g=1}^{G} p_g^2.$$

Under this proportion-based definition, S can be interpreted as the probability that two independently selected annotators belong to different clusters. With six clusters and 59 annotators in each cluster, $p_g = 1/6$ for every g. Therefore,

$$H = \ln 6 = 1.7918, \qquad S = 1 - 6\left(\frac{1}{6}\right)^2 = \frac{5}{6} = 0.8333.$$

Both values are the maxima possible for six clusters. They provide a compact numerical check that the sampling procedure did not favour one cluster over another. Because the counts are exactly equal, however, the two indices summarize the same balance already visible from the count of 59 annotators per cluster. They do not show that the K-means clusters are internally varied, well separated, or representative of the full D3CODE population.

The prediction target was taken from the D3CODE field `rating_binary`. A value of 1 denotes *Offensive* and a value of 0 denotes *Not Offensive*. No majority-vote aggregation is used.

A.2. Worked example of moral-profile rendering

All profile representations are generated deterministically from the same source profile. Moral dimensions are ordered by descending score; ties are ordered alphabetically by dimension name. Table 3.2 in the main text gives the rendering rules. For the source profile `loyalty=4.67`, `proportionality=4.17`, `authority=4.17`, `care=3.83`, `equality=3.00`, and `purity=3.00`, the five representations are as follows.

```
Natural language:
When evaluating situations, this annotator strongly prioritizes
   loyalty, moderately values authority, moderately values
   proportionality, moderately values care, is neutral toward
   equality, and is neutral toward purity.

Categorical:
Annotator Moral Profile Categories:
- Loyalty: Very High
- Authority: High
- Proportionality: High
- Care: High
- Equality: Neutral
- Purity: Neutral

Numeric:
Annotator Moral Profile Scores:
- Loyalty: 4.67/5
- Authority: 4.17/5
- Proportionality: 4.17/5
- Care: 3.83/5
- Equality: 3.00/5
- Purity: 3.00/5

Ordinal:
Annotator Moral Profile Ranking:
Loyalty > Authority = Proportionality > Care > Equality = Purity

Top-k:
Annotator Top-3 Moral Values:
```

```
- Loyalty
- Authority
- Proportionality
```

A.3. Prompt templates

The prompt templates used in the experiments are reproduced below. Their use of the phrase “detecting offensive language” is an operational task instruction; throughout the thesis, the research object is the model’s prediction of an annotator-specific offensiveness judgement.

The zero-shot baseline is the generic, unpersonalized reference prompt: it omits an annotator-specific moral-value signal and asks for a general binary judgement.

```
You are an AI assistant tasked with detecting offensive language.
Read the following text and determine if it is offensive.

Text: "{text_item}"

Is this text offensive? Answer strictly with a single letter: 'Y' for
    Yes, or 'N' for No. Do not output any other text.
```

The zero-shot value-conditioned prompt adds an annotator-specific moral-value profile and asks the model to approximate that annotator’s judgement.

```
You are an AI assistant tasked with detecting offensive language from
    the perspective of a specific human annotator.

Here is the annotator's moral value profile:
{formatted_profile}

Based on this moral profile, how would this specific annotator judge
   the following text?

Text: "{text_item}"

Would this annotator consider this text offensive? Answer strictly
   with a single letter: 'Y' for Yes, or 'N' for No. Do not output
   any other text.
```

The few-shot value-conditioned prompt places the formatted profile first, then three previous same-annotator judgements, and finally the target text.

```
{profile_text}

Here are some previous examples of how this annotator evaluated other
    texts:
Example 1:
Text: "{example_1_text}"
Annotator Judgment: Yes (Offensive)

Example 2:
Text: "{example_2_text}"
Annotator Judgment: No (Not Offensive)

Example 3:
Text: "{example_3_text}"
Annotator Judgment: Yes (Offensive)

Using the annotator's moral profile and previous judgments above,
   would this annotator consider the following text offensive?
Text: "{current_text}"
Answer strictly with 'Y' for Yes (Offensive) or 'N' for No (Not
   Offensive).
```

A.4. Output parsing and invalid-response convention

Each model response was normalized to uppercase, with full stops and colons removed before parsing. A response containing `Y` but not `N` was assigned label `1`; a response containing `N` but not `Y` was assigned label `0`. If both letters occur, the parser uses whichever appears last in the normalized response. If neither occurred, the output was assigned the sentinel value -99.

The value -99 is an out-of-domain sentinel: the valid prediction set is $\{0, 1\}$, so it is not a third label and does not enter Accuracy, Macro-F1, or McNemar calculations as a numerical prediction. Condition-level metrics exclude rows that are invalid for that condition. Paired analyses retain only rows with valid predictions in every condition being compared; the cross-representation disagreement diagnostic likewise requires valid binary outputs for all five profile-conditioned representations. Invalid counts should therefore be monitored because systematic parsing failures can change the effective sample size, even though the sentinel itself has no substantive interpretation.

A.5. Few-shot example selection

For each target item, candidate demonstrations were drawn from the target annotator's historical labels after excluding the target item itself. Each prompt used at most three examples from this candidate set.

Table A.2: Few-shot example-selection procedures.

Strategy	Selection criterion	Operationalization
Random	Uniform random sample from the valid same-annotator pool	Uniform sampling with a fixed random seed of 42.
Retrieval	Nearest examples to the target text	Texts were encoded with `all-MiniLM-L6-v2`; candidates were ranked by cosine similarity to the target text, and the top k were selected.
Diversity	Farthest-point coverage of the annotator's examples	Starting from a seeded initial candidate, examples were selected iteratively to maximize their minimum distance from those already selected, where distance is $1 - \cos(\cdot, \cdot)$.

Retrieval and diversity selection use sentence embeddings from a SentenceTransformer model. No explicit label-balancing constraint was imposed.

B

Qualitative cross-representation disagreement examples

Table B.1 presents illustrative zero-shot cases from Qwen2.5-72B-Instruct. These examples are selected from annotator–item pairs for which the text, annotator, moral-value scores, and human binary label are fixed, while the moral-profile representation changes. The purpose of this table is illustrative rather than evidential: it shows what cross-representation disagreement can look like at the instance level, whereas the main evidence comes from the aggregate disagreement-rate analysis in Table 4.3.

The examples focus on socially or morally charged content, including religion, gender, sexuality, ethnicity, and political identity. Such cases are useful for qualitative inspection because they make representation-dependent changes easier to examine. However, they should not be read as showing that one representation is universally superior, nor as causal proof that a particular word or value dimension caused the change. They show only that alternative representations of the same annotator profile can lead the model to produce different binary judgments.

Table B.1: Illustrative Qwen2.5-72B-Instruct zero-shot cases where changing only the moral-profile representation changes the binary prediction. Prediction order: ordinal, natural language, numeric, categorical, top-k. The mixed-seeds column reports how many of the three zero-shot seeds produced mixed predictions for the same annotator–item pair.

Case	Annotator–item	Human label	Predictions	Mixed seeds	Text excerpt
Religion / gender	R_3a6lKrVJ Co42exr / 277	0	[1, 0, 1, 0, 1]	2/3	Yah and while he is at it let's stop the Catholic Church from discriminating against women once and for all!
Religion / ethnicity	R_1ikVcVRB IFqMAo9 / 1233	0	[1, 0, 1, 0, 1]	3/3	Churchill never mentioned even once the Holocaust of the Jews in his books on WW2 published after the war. Neithet did US generals Eisehhower & others. Why? Were they...
Sexuality / religion	R_D16mc67R YMHRYJz / 1489	1	[0, 0, 1, 1, 1]	3/3	Forcing your doctrine down the throats of ... Why do opponents of gay rights so often use this metaphor?
Transgender / politics	R_W1cpK8ln L8eeXN7 / 1571	0	[1, 0, 1, 0, 1]	1/3	If I were transgender I'd say bah-bye and be on my way, middle finger extended. Trump is such a nothing. And so are the people who support him. Just more to add to...
Muslim / policy	R_W1cpK8ln L8eeXN7 / 1590	0	[1, 0, 1, 0, 1]	3/3	It's not a ban of all Muslims in the Universe. I know, hard to believe, but it is a ban on Muslims from certain countries. Are you simple?
Gay marriage	R_RUeSQmv1 HCa2PQd / 1782	1	[0, 1, 1, 0, 1]	3/3	Then stop firing gays who marry. That is malice in action.

C

Sub-category diagnostic of representation disagreement

This appendix reports an exploratory sub-category breakdown of the cross-representation decision-disagreement diagnostic. The calculation uses the same definition as Section 3.7: for each annotator–item pair, the five profile-conditioned predictions are compared, and a pair is counted as disagreement if at least one representation changes the binary prediction. The no-profile baseline is excluded. Items with multiple D3CODE sub-category labels are expanded into each relevant sub-category bin, so the sub-category totals should be read as grouped diagnostic counts rather than mutually exclusive partitions.

Table C.1 summarizes the main model, Qwen-72B. Figures C.2–C.4 provide the full appendix plots for all models and settings. Bars show the seed-mean disagreement rate, error bars show seed standard deviation, and dots show individual seed values. The plots are descriptive and should not be interpreted as causal evidence that a sub-category itself causes representation sensitivity.

Table C.1: Qwen-72B sub-category disagreement rates (%). Values are mean disagreement rates across the three seeds. The lowest few-shot rate in each row is shown in the final column.

Sub-category	Zero	Rand.	Retr.	Div.	Lowest FS
random	23.49	13.76	12.79	14.70	Retrieval
fairness	22.47	13.90	11.97	15.96	Retrieval
christian	21.75	17.18	14.72	21.47	Retrieval
LGB	19.61	17.56	13.55	18.00	Retrieval
transgender	18.68	16.24	11.80	15.56	Retrieval
muslim	16.88	14.86	11.38	17.29	Retrieval
jewish	16.58	17.35	12.61	15.74	Retrieval
authority	10.45	6.88	5.38	7.03	Retrieval
loyalty	8.29	6.39	5.56	5.24	Diverse
purity	7.67	4.43	6.43	5.27	Random
care	5.87	6.65	6.03	7.36	Retrieval

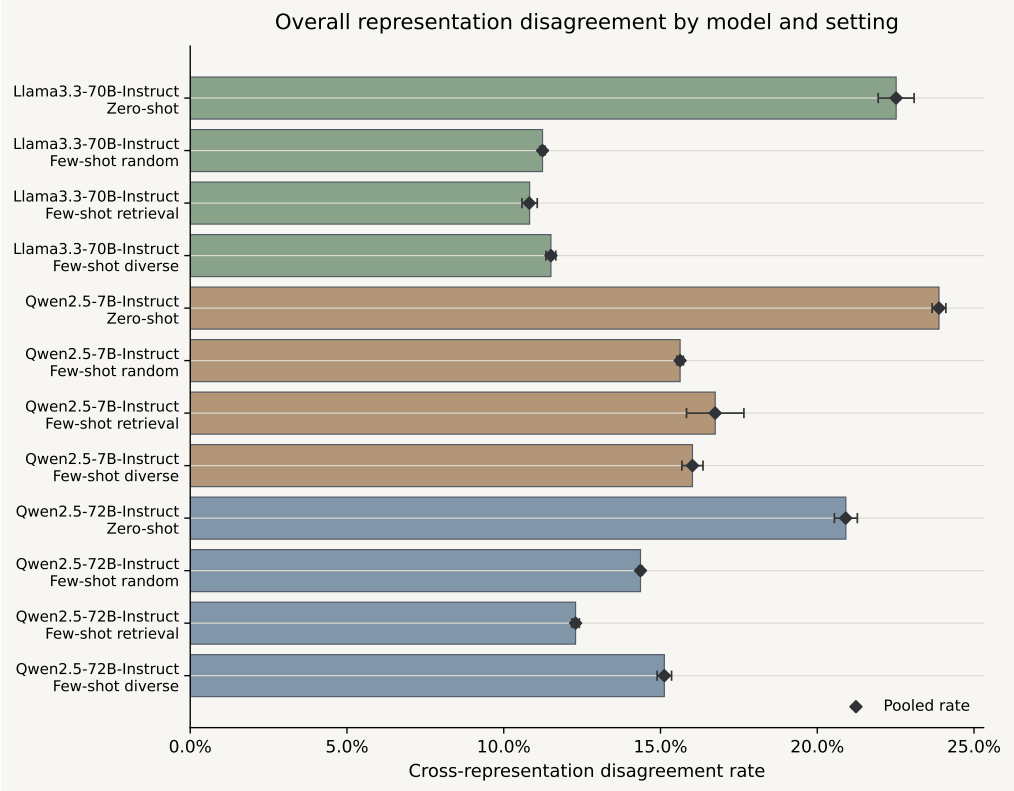

Figure C.1: Overall cross-representation decision-disagreement rates by model and prompting setting in the sub-category-expanded analysis.

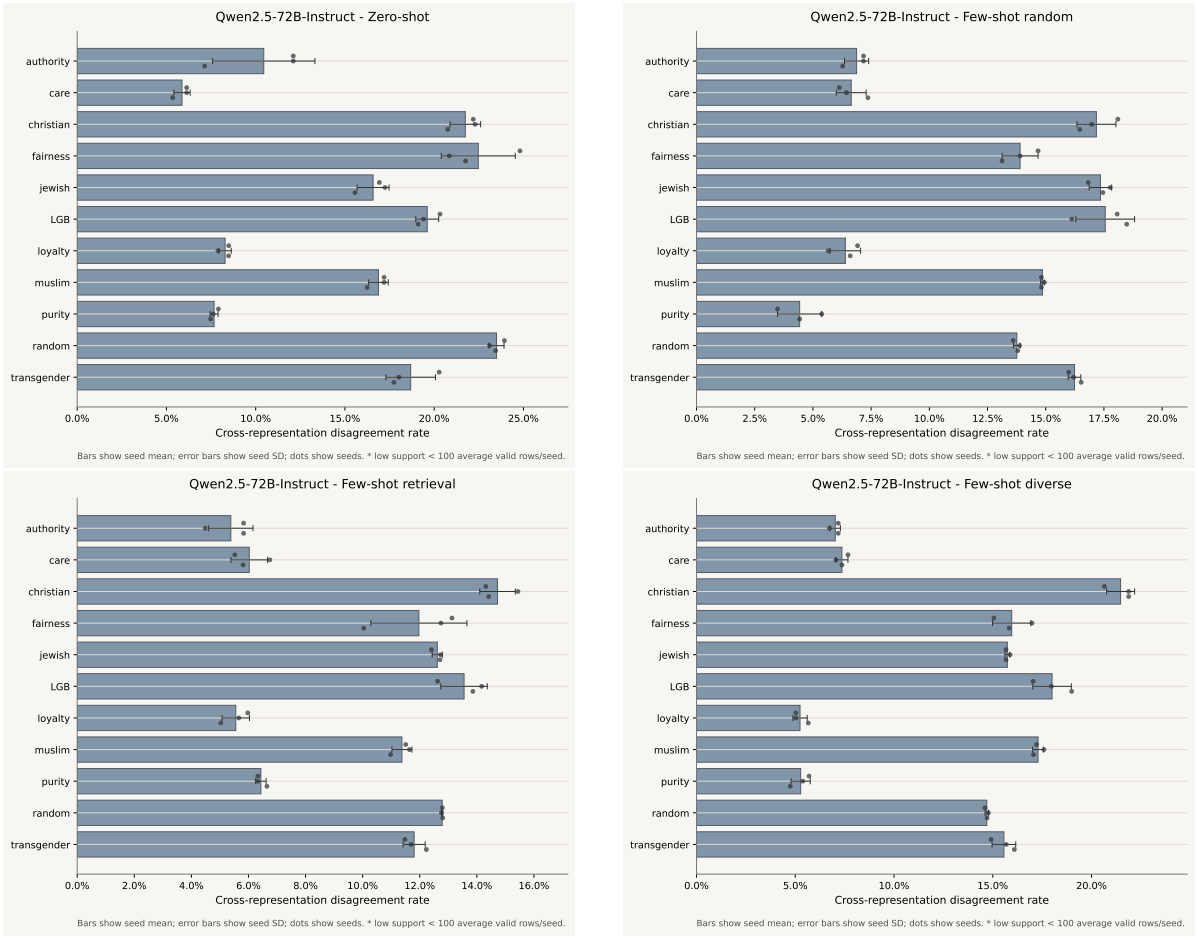

Figure C.2: Sub-category cross-representation decision disagreement for Qwen2.5-72B-Instruct across zero-shot and few-shot settings.

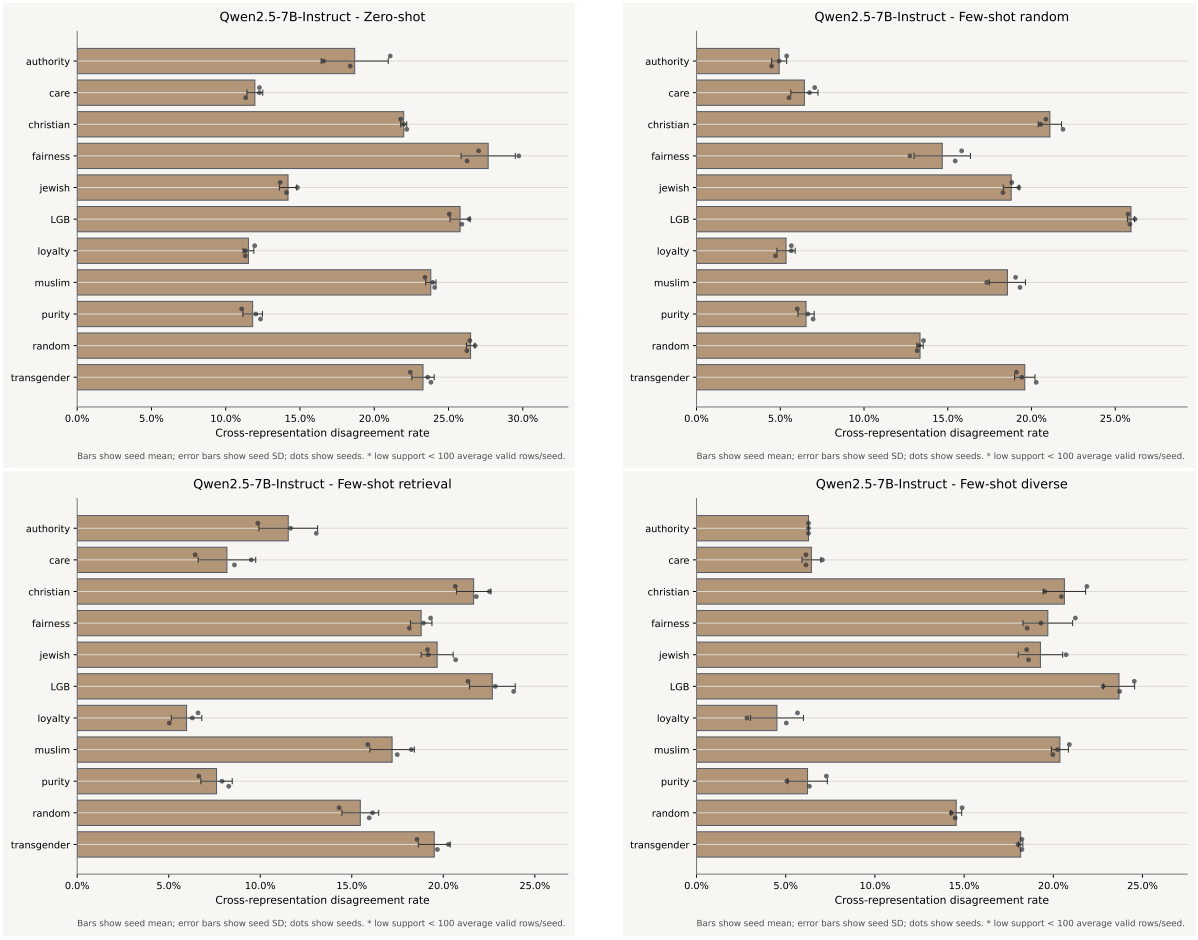

Figure C.3: Sub-category cross-representation decision disagreement for Qwen2.5-7B-Instruct across zero-shot and few-shot settings.

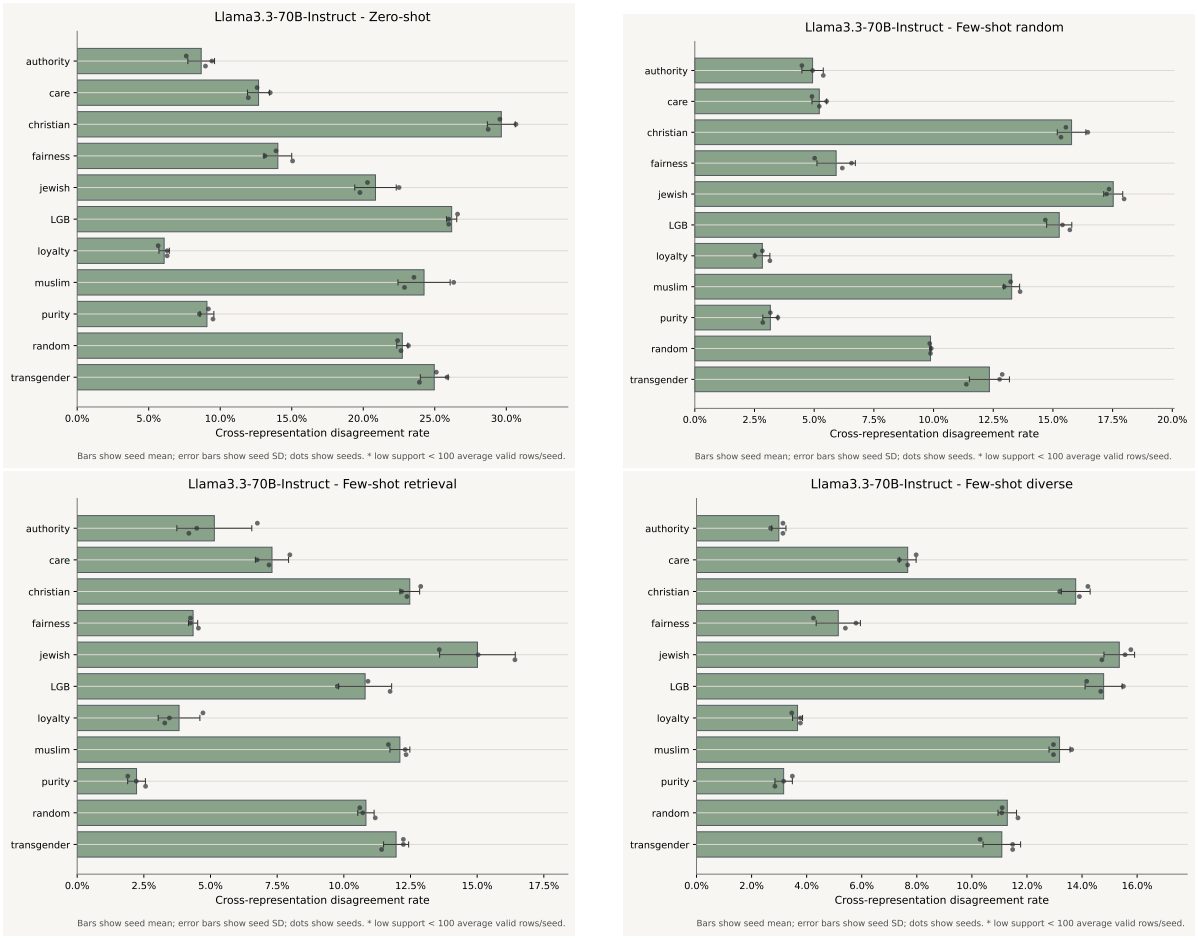

Figure C.4: Sub-category cross-representation decision disagreement for Llama-3.3-70B-Instruct across zero-shot and few-shot settings.

D

Complete performance tables

Table D.1: Complete performance results for Llama-3.3-70B-Instruct. Results are mean ± standard deviation across the three experimental seeds.

Setting	Prompt condition	Macro-F1	Accuracy
Zero-shot	Baseline	0.6100 ± 0.0003	0.6435 ± 0.0004
	Natural Language	0.6226 ± 0.0007	0.6747 ± 0.0012
	Categorical	0.6234 ± 0.0009	0.6728 ± 0.0012
	Numeric	0.6194 ± 0.0019	0.6789 ± 0.0002
	Ordinal	0.6124 ± 0.0016	0.6573 ± 0.0034
	Top-k	0.6059 ± 0.0010	0.6297 ± 0.0018
Few-shot random	Baseline	0.6507 ± 0.0004	0.7082 ± 0.0003
	Natural Language	0.6522 ± 0.0008	0.7162 ± 0.0007
	Categorical	0.6527 ± 0.0011	0.7178 ± 0.0009
	Numeric	0.6482 ± 0.0003	0.7158 ± 0.0003
	Ordinal	0.6512 ± 0.0009	0.7083 ± 0.0008
	Top-k	0.6534 ± 0.0005	0.7084 ± 0.0004
Few-shot retrieval	Baseline	0.6568 ± 0.0009	0.7147 ± 0.0012
	Natural Language	0.6621 ± 0.0028	0.7240 ± 0.0014
	Categorical	0.6654 ± 0.0007	0.7287 ± 0.0012
	Numeric	0.6636 ± 0.0022	0.7267 ± 0.0019
	Ordinal	0.6575 ± 0.0002	0.7163 ± 0.0008
	Top-k	0.6575 ± 0.0004	0.7124 ± 0.0024
Few-shot diversity	Baseline	0.6516 ± 0.0008	0.7019 ± 0.0009
	Natural Language	0.6540 ± 0.0011	0.7120 ± 0.0008
	Categorical	0.6582 ± 0.0010	0.7166 ± 0.0007
	Numeric	0.6533 ± 0.0007	0.7120 ± 0.0009
	Ordinal	0.6527 ± 0.0007	0.7021 ± 0.0006
	Top-k	0.6527 ± 0.0006	0.7003 ± 0.0005

Table D.2: Complete performance results for Qwen2.5-72B-Instruct. Results are mean ± standard deviation across the three experimental seeds.

Setting	Prompt condition	Macro-F1	Accuracy
Zero-shot	Baseline	0.5892 ± 0.0007	0.6024 ± 0.0019
	Natural Language	0.6047 ± 0.0035	0.6218 ± 0.0051
	Categorical	0.5956 ± 0.0039	0.6054 ± 0.0050
	Numeric	0.5844 ± 0.0035	0.5930 ± 0.0047
	Ordinal	0.5791 ± 0.0025	0.5889 ± 0.0038
	Top-*k*	0.5663 ± 0.0021	0.5717 ± 0.0031
Few-shot random	Baseline	0.6608 ± 0.0007	0.6909 ± 0.0012
	Natural Language	0.6618 ± 0.0003	0.7010 ± 0.0002
	Categorical	0.6639 ± 0.0006	0.6975 ± 0.0004
	Numeric	0.6591 ± 0.0004	0.6936 ± 0.0004
	Ordinal	0.6580 ± 0.0003	0.6935 ± 0.0002
	Top-*k*	0.6537 ± 0.0002	0.6832 ± 0.0001
Few-shot retrieval	Baseline	0.6720 ± 0.0011	0.7009 ± 0.0011
	Natural Language	0.6744 ± 0.0008	0.7105 ± 0.0010
	Categorical	0.6759 ± 0.0015	0.7074 ± 0.0014
	Numeric	0.6722 ± 0.0004	0.7045 ± 0.0005
	Ordinal	0.6707 ± 0.0003	0.7039 ± 0.0003
	Top-*k*	0.6697 ± 0.0016	0.6986 ± 0.0017
Few-shot diversity	Baseline	0.6548 ± 0.0018	0.6859 ± 0.0021
	Natural Language	0.6566 ± 0.0007	0.6962 ± 0.0009
	Categorical	0.6601 ± 0.0014	0.6933 ± 0.0014
	Numeric	0.6558 ± 0.0015	0.6889 ± 0.0012
	Ordinal	0.6562 ± 0.0007	0.6913 ± 0.0011
	Top-*k*	0.6496 ± 0.0005	0.6789 ± 0.0004

Table D.3: Complete performance results for Qwen2.5-7B-Instruct. Results are mean ± standard deviation across the three experimental seeds.

Setting	Prompt condition	Macro-F1	Accuracy
Zero-shot	Baseline	0.5750 ± 0.0015	0.5856 ± 0.0021
	Natural Language	0.5775 ± 0.0021	0.5911 ± 0.0035
	Categorical	0.5577 ± 0.0034	0.5617 ± 0.0040
	Numeric	0.5262 ± 0.0031	0.5267 ± 0.0033
	Ordinal	0.5469 ± 0.0034	0.5499 ± 0.0040
	Top-*k*	0.5595 ± 0.0020	0.5659 ± 0.0028
Few-shot random	Baseline	0.6315 ± 0.0002	0.6804 ± 0.0013
	Natural Language	0.6332 ± 0.0009	0.6942 ± 0.0009
	Categorical	0.6405 ± 0.0014	0.6999 ± 0.0003
	Numeric	0.6323 ± 0.0003	0.6775 ± 0.0014
	Ordinal	0.6355 ± 0.0012	0.6934 ± 0.0013
	Top-*k*	0.6383 ± 0.0004	0.6953 ± 0.0016
Few-shot retrieval	Baseline	0.6456 ± 0.0027	0.6878 ± 0.0012
	Natural Language	0.6427 ± 0.0090	0.6879 ± 0.0086
	Categorical	0.6467 ± 0.0083	0.6913 ± 0.0070
	Numeric	0.6323 ± 0.0083	0.6641 ± 0.0093
	Ordinal	0.6419 ± 0.0097	0.6875 ± 0.0070
	Top-*k*	0.6446 ± 0.0073	0.6902 ± 0.0052
Few-shot diversity	Baseline	0.6397 ± 0.0005	0.6852 ± 0.0017
	Natural Language	0.6389 ± 0.0014	0.6949 ± 0.0004
	Categorical	0.6455 ± 0.0004	0.7008 ± 0.0013
	Numeric	0.6385 ± 0.0002	0.6804 ± 0.0020
	Ordinal	0.6411 ± 0.0012	0.6955 ± 0.0028
	Top-*k*	0.6440 ± 0.0001	0.6972 ± 0.0016

E

Exploratory per-annotator performance

Aggregate metrics can conceal heterogeneity across target annotators. The following plots report Macro-F1 separately for each sampled annotator. Within each plot, annotators are ordered by the baseline Macro-F1. Across models and prompting settings, the curves form broad gradients rather than narrow bands around the aggregate mean: some annotator–model combinations are substantially easier to predict than others. The baseline and natural-language curves are often broadly aligned, indicating that relative annotator predictability can be more consequential than the difference between these displayed conditions.

This pattern does not by itself establish that moral-value conditioning is uniformly beneficial. Natural-language prompting does not dominate the baseline for every annotator, setting, or model. Instead, the plots provide descriptive evidence that aggregate scores do not characterize every annotator equally and that representation effects can be heterogeneous. Per-annotator estimates should be interpreted cautiously because they are based on finite and unequal numbers of items and may reflect label consistency, item composition, and class distribution as well as model behaviour.

Table E.1 summarizes the distribution for the Qwen2.5-7B-Instruct zero-shot setting. The baseline ranges from 0.21 to 0.89, with a median of 0.53 and an interquartile range of 0.42–0.62. Thirty-two of the 354 sampled annotators have baseline Macro-F1 of at least 0.70. The analogous natural-language and numeric distributions in Table E.1 are similarly dispersed.

Table E.1: Per-annotator Macro-F1 distribution for Qwen2.5-7B-Instruct in the zero-shot setting. Statistics are computed over 354 sampled annotators.

Condition	**Min.**	**Median**	**Mean**	**IQR**	**Max.**	≥ 0.70
Baseline	0.21	0.53	0.52	0.42–0.62	0.89	32
Natural language	0.19	0.53	0.52	0.43–0.62	0.88	30
Numeric	0.23	0.53	0.52	0.43–0.61	0.88	27

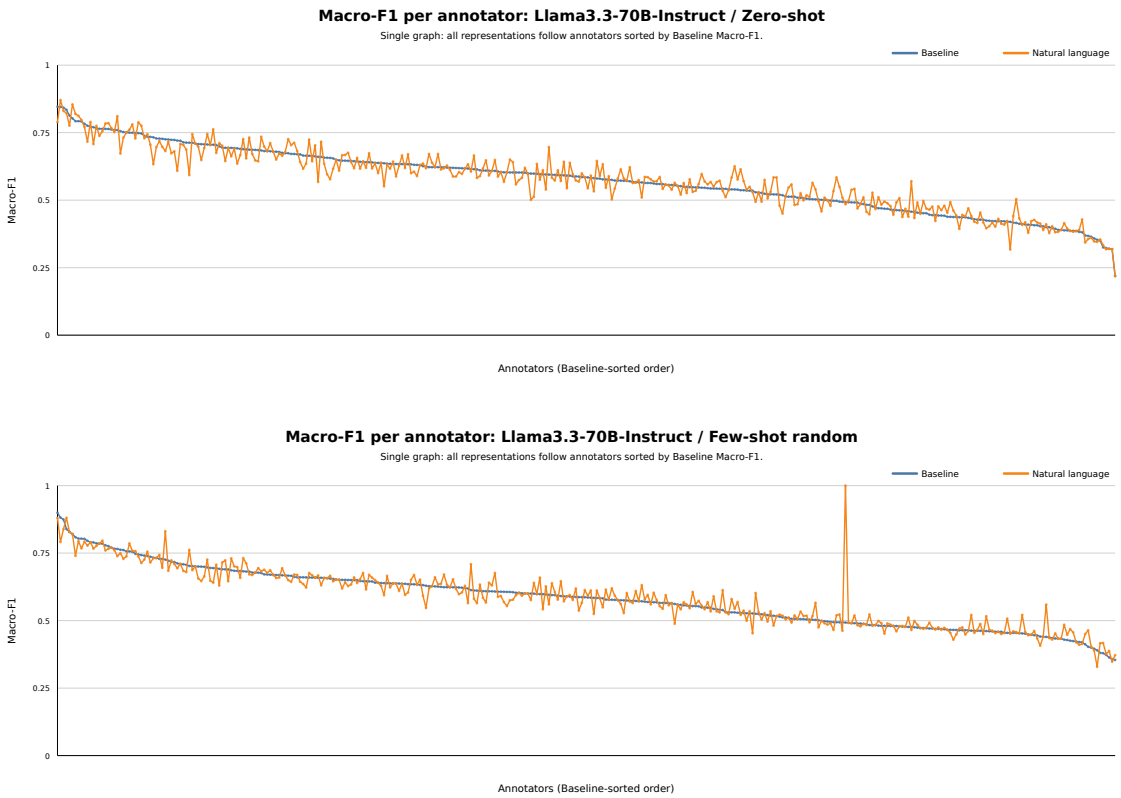

Figure E.1: Per-annotator Macro-F1 for Llama-3.3-70B-Instruct. Top: zero-shot. Bottom: few-shot random. Annotators are sorted by baseline Macro-F1 separately within each panel; horizontal positions are therefore not directly comparable across panels.

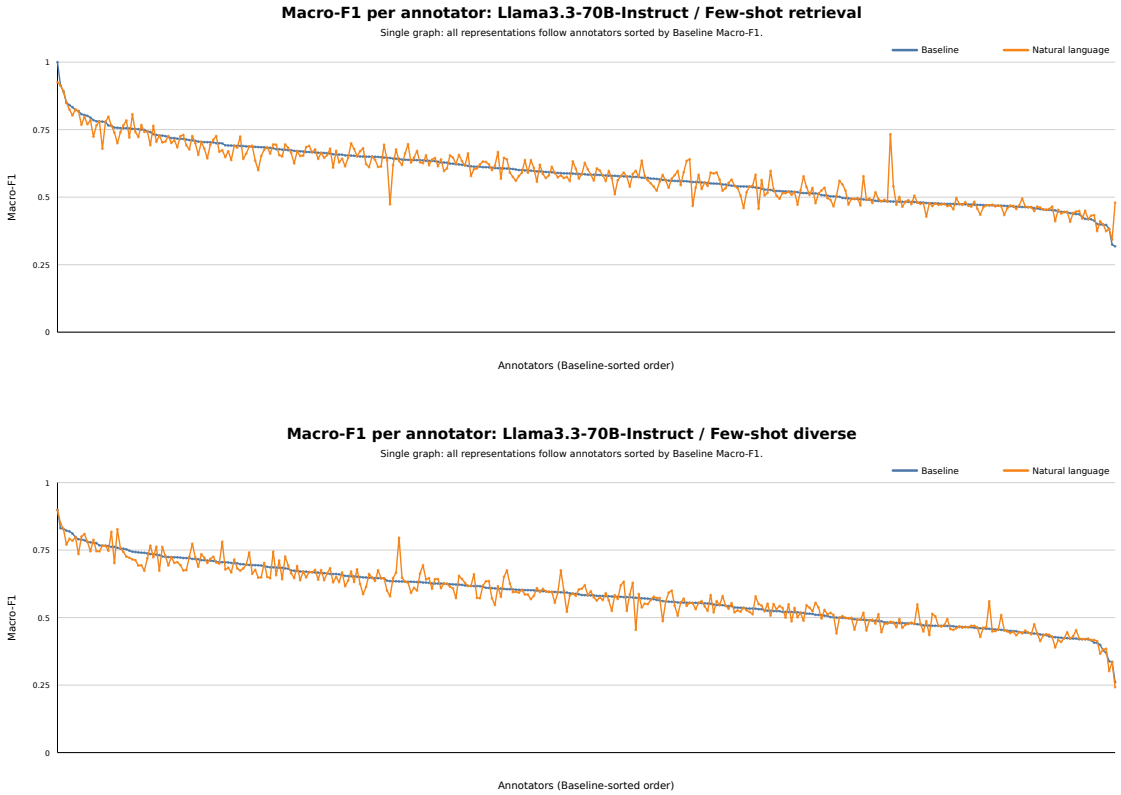

Figure E.2: Per-annotator Macro-F1 for Llama-3.3-70B-Instruct. Top: few-shot retrieval. Bottom: few-shot diversity. Annotators are sorted by baseline Macro-F1 separately within each panel; horizontal positions are therefore not directly comparable across panels.

Figure E.3: Per-annotator Macro-F1 for Qwen2.5-7B-Instruct. Top: zero-shot. Bottom: few-shot random. Annotators are sorted by baseline Macro-F1 separately within each panel; horizontal positions are therefore not directly comparable across panels.

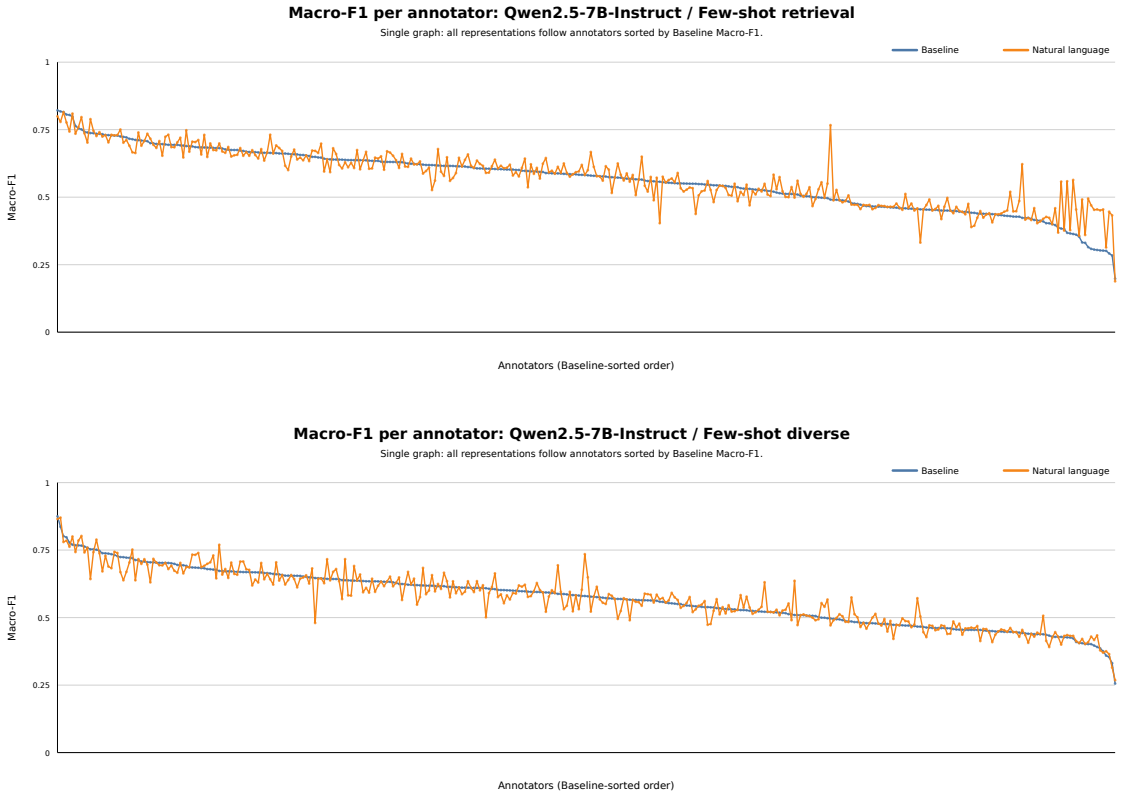

Figure E.4: Per-annotator Macro-F1 for Qwen2.5-7B-Instruct. Top: few-shot retrieval. Bottom: few-shot diversity. Annotators are sorted by baseline Macro-F1 separately within each panel; horizontal positions are therefore not directly comparable across panels.

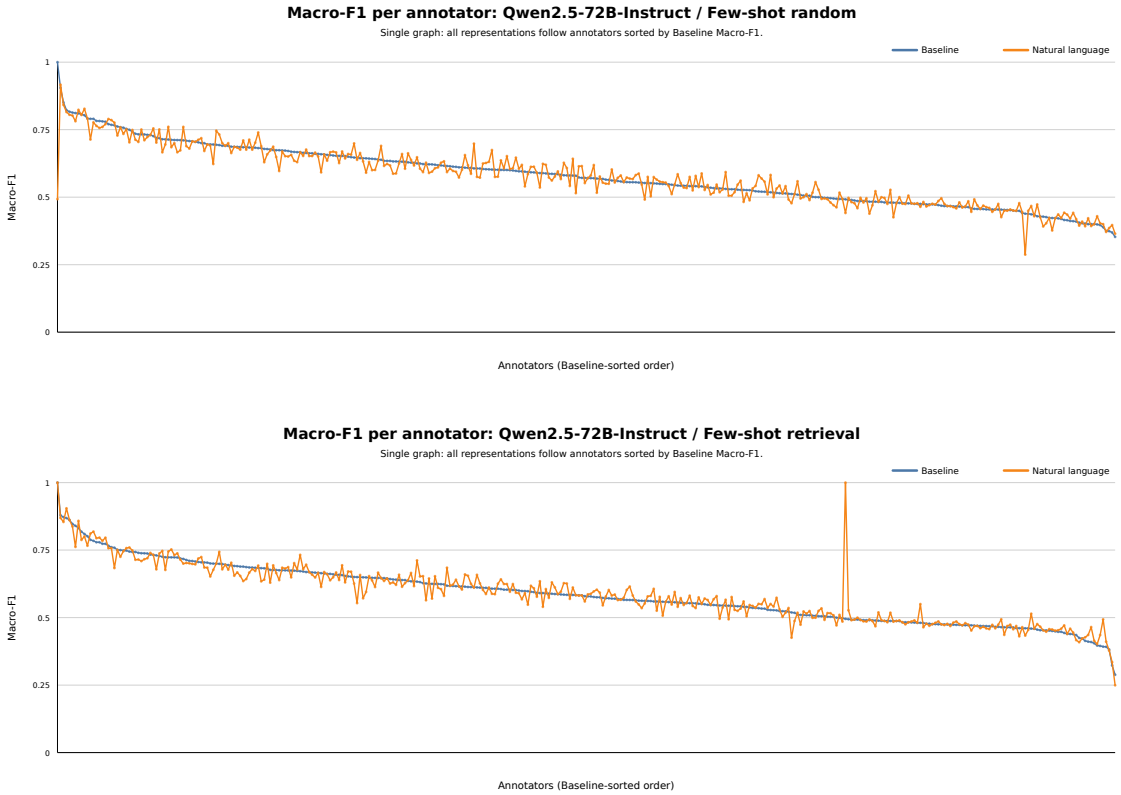

Figure E.5: Per-annotator Macro-F1 for Qwen2.5-72B-Instruct. Top: few-shot random. Bottom: few-shot retrieval. Annotators are sorted by baseline Macro-F1 separately within each panel; horizontal positions are therefore not directly comparable across panels.

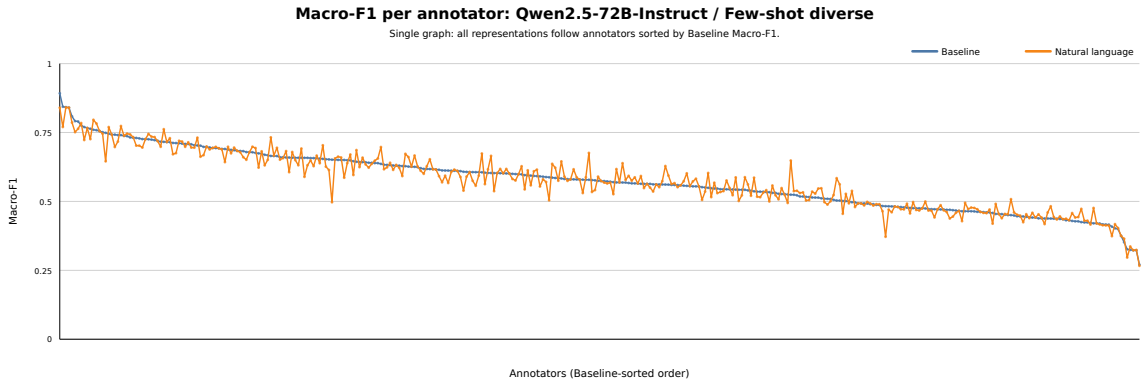

Figure E.6: Per-annotator Macro-F1 for Qwen2.5-72B-Instruct under few-shot diversity selection. Annotators are sorted by baseline Macro-F1 within the panel.

F

Complete McNemar significance matrices

This appendix presents pairwise McNemar significance matrices for Llama-3.3-70B-Instruct, Qwen2.5-72B-Instruct, and Qwen2.5-7B-Instruct across the zero-shot, few-shot random, few-shot retrieval, and few-shot diversity settings. In each matrix, a value of 1 indicates a statistically significant pairwise difference at $\alpha = 0.05$, while a value of 0 indicates no statistically significant difference at this threshold.

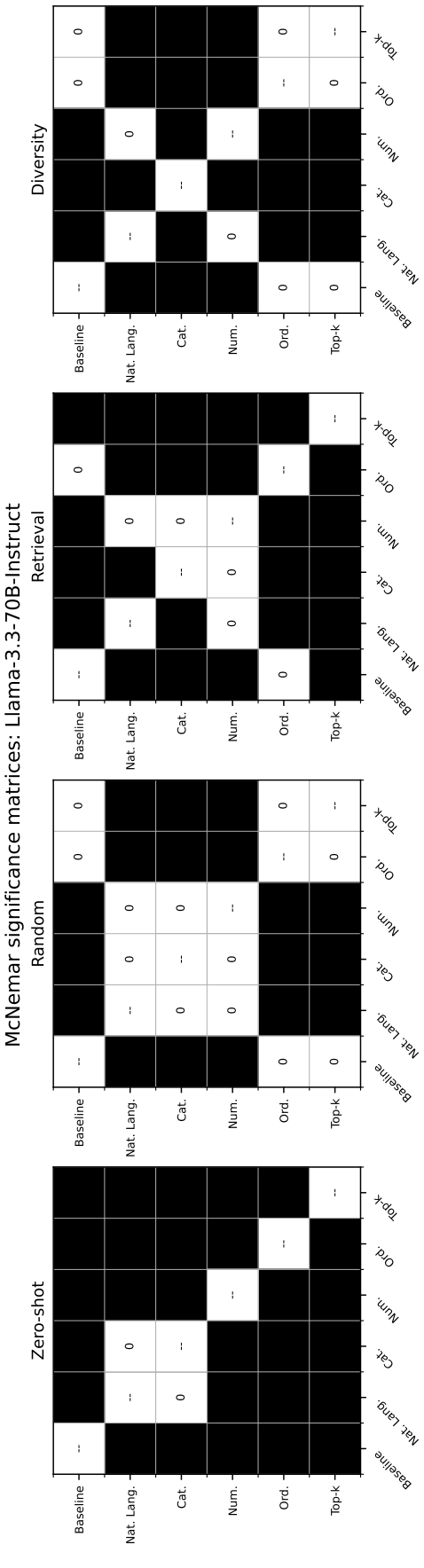

Figure F.1: Pairwise McNemar significance matrices for Llama-3.3-70B-Instruct across the zero-shot, few-shot random, few-shot retrieval, and few-shot diversity settings. Cells marked 1 indicate $p < 0.05$; cells marked 0 indicate $p \geq 0.05$.

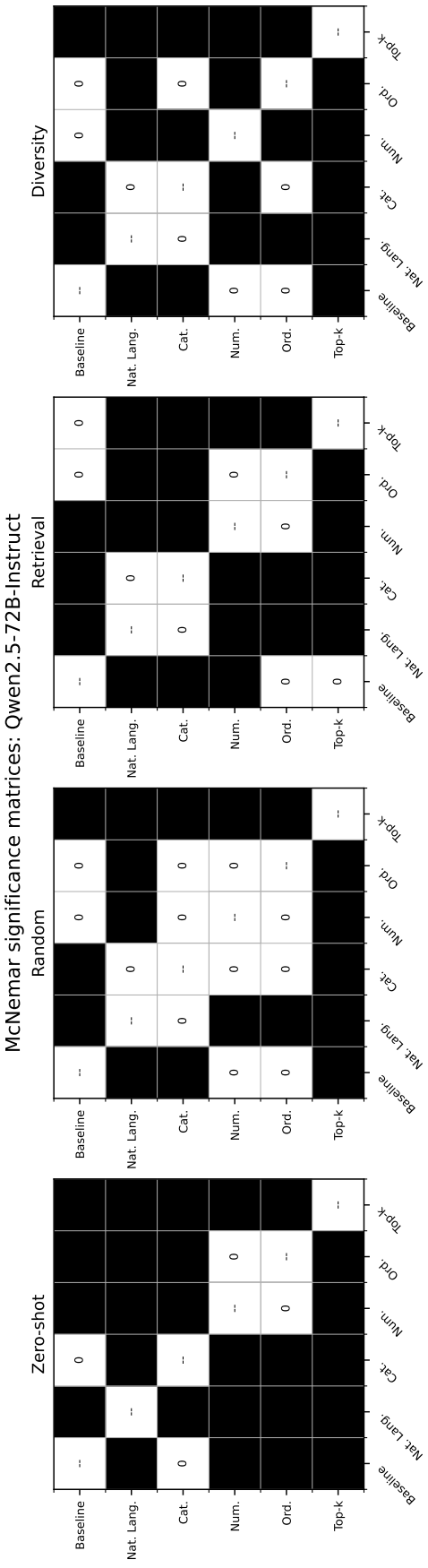

Figure F.2: Pairwise McNemar significance matrices for Qwen2.5-72B-Instruct across the zero-shot, few-shot random, few-shot retrieval, and few-shot diversity settings. Cells marked 1 indicate $p < 0.05$; cells marked 0 indicate $p \geq 0.05$.

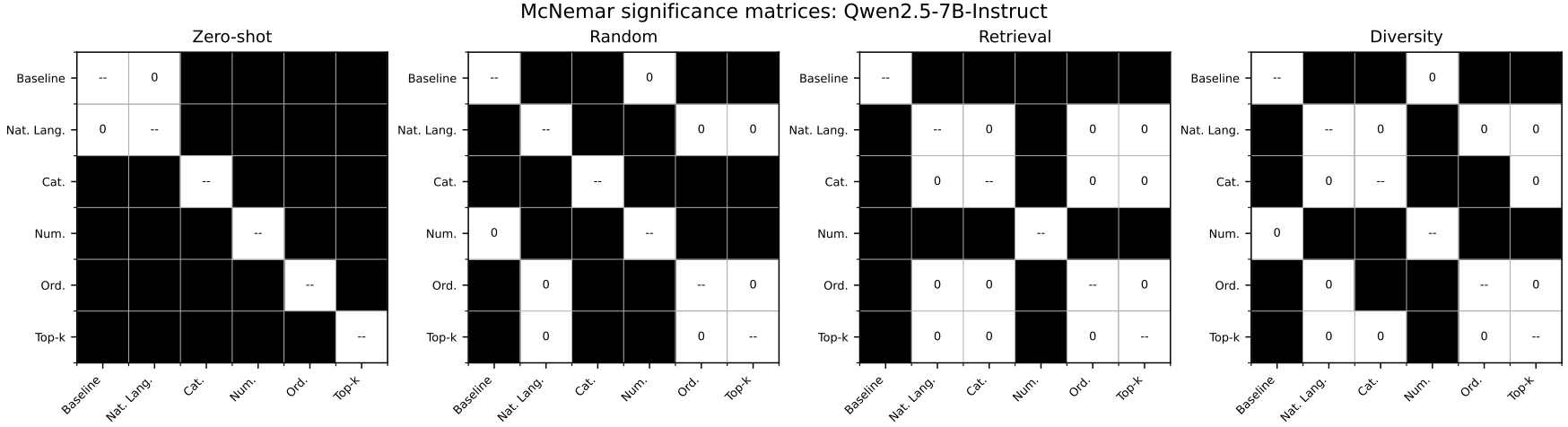

Figure F.3: Pairwise McNemar significance matrices for Qwen2.5-7B-Instruct across the zero-shot, few-shot random, few-shot retrieval, and few-shot diversity settings. Cells marked 1 indicate $p < 0.05$; cells marked 0 indicate $p \geq 0.05$.

Statement on the use of generative AI

OpenAI ChatGPT was used as an assistive tool during the preparation of this thesis. Its use was limited to the following purposes:

- improving grammar, wording, sentence flow, and the organization of text drafted by the author;
- obtaining an initial orientation to some background papers before the author's own close reading, and clarifying terminology or the structure of an argument during that reading;
- assisting with the visual development of the cover image through OpenAI's image-generation functionality.

For background literature, generative-AI output was not treated as academic evidence. The cited papers were subsequently read and checked by the author, and the bibliography was verified against the corresponding publication records. The author decided which sources to cite and is responsible for the interpretation of those sources.

Representative instructions used for text and literature assistance included: "Improve the grammar and sentence flow while preserving the technical meaning and cited claims," and "Explain the contribution, method, and limitations of this paper to support my own reading; distinguish what the paper states from interpretation." These examples describe the type of assistance requested rather than a complete transcript of interactions.

The cover was produced through iterative, author-directed image generation in ChatGPT. A representative visual instruction was: "A contemplative, human-like LLM seated at a round marble table, examining a balance scale with a group of wooden figures on one side and a single figure on the other; realistic, bright photographic style." The author selected and refined the final composition.

The research questions, experimental design, code, data processing, model runs, statistical analyses, interpretation of results, and final claims remain the author's responsibility. Generative AI was not used to create or alter experimental observations or model outputs.

References

[1] Paula Fortuna and Sérgio Nunes. 'A Survey on Automatic Detection of Hate Speech in Text'. In: *ACM Computing Surveys* 51.4 (2018), pp. 1–30. DOI: 10.1145/3232676. URL: https://doi.org/10.1145/3232676.

[2] Thomas Davidson, Dana Warmsley, Michael Macy and Ingmar Weber. 'Automated Hate Speech Detection and the Problem of Offensive Language'. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 2017, pp. 512–515. DOI: 10.1609/icwsm.v11i1.14955. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14955.

[3] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra and Ritesh Kumar. 'SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)'. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 75–86. DOI: 10.18653/v1/S19-2010. URL: https://aclanthology.org/S19-2010/.

[4] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi and Noah A. Smith. 'Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 5884–5906. DOI: 10.18653/v1/2022.naacl-main.431. URL: https://aclanthology.org/2022.naacl-main.431/.

[5] Barbara Plank. 'The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation'. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 10671–10682. DOI: 10.18653/v1/2022.emnlp-main.731. URL: https://aclanthology.org/2022.emnlp-main.731/.

[6] Lora Aroyo and Chris Welty. 'Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation'. In: *AI Magazine* 36.1 (2015), pp. 15–24. DOI: 10.1609/aimag.v36i1.2564. URL: https://doi.org/10.1609/aimag.v36i1.2564.

[7] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco and Davide Bernardi. 'Perspectivist approaches to natural language processing: a survey'. In: *Language Resources and Evaluation* 59 (2025), pp. 1719–1746. DOI: 10.1007/s10579-024-09766-4. URL: https://doi.org/10.1007/s10579-024-09766-4.

[8] Aida Mostafazadeh Davani, Mark Díaz and Vinodkumar Prabhakaran. 'Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations'. In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 92–110. DOI: 10.1162/tacl_a_00449. URL: https://aclanthology.org/2022.tacl-1.6/.

[9] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto and Michael S. Bernstein. 'Jury Learning: Integrating Dissenting Voices into Machine Learning Models'. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2022. DOI: 10.1145/3491102.3502004. URL: https://doi.org/10.1145/3491102.3502004.

[10] Matthias Orlikowski, Paul Röttger, Philipp Cimiano and Dirk Hovy. 'The Ecological Fallacy in Annotation: Modeling Human Label Variation Goes Beyond Sociodemographics'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 1017–1029. DOI: 10.18653/v1/2023.acl-short.88. URL: https://aclanthology.org/2023.acl-short.88/.

[11] Shalom H. Schwartz. 'Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries'. In: *Advances in Experimental Social Psychology*. Ed. by Mark P. Zanna. Vol. 25. Academic Press, 1992, pp. 1–65. DOI: 10.1016/S0065-2601(08)60281-6. URL: https://doi.org/10.1016/S0065-2601(08)60281-6.

[12] Jonathan Haidt and Craig Joseph. 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues'. In: *Daedalus* 133.4 (2004), pp. 55–66. DOI: 10.1162/0011526042365555.

[13] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik and Peter H. Ditto. 'Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism'. In: *Advances in Experimental Social Psychology*. Vol. 47. Academic Press, 2013, pp. 55–130. DOI: 10.1016/B978-0-12-407236-7.00002-4.

[14] Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens and Morteza Dehghani. 'Morality beyond the WEIRD: How the Nomological Network of Morality Varies across Cultures'. In: *Journal of Personality and Social Psychology* 125.5 (2023), pp. 1157–1188. DOI: 10.1037/pspp0000470.

[15] Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker and Vinodkumar Prabhakaran. 'D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation'. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 18511–18526. DOI: 10.18653/v1/2024.emnlp-main.1029. URL: https://aclanthology.org/2024.emnlp-main.1029/.

[16] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein and Sameer Singh. 'Calibrate Before Use: Improving Few-shot Performance of Language Models'. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 12697–12706. URL: https://proceedings.mlr.press/v139/zhao21c.html.

[17] Melanie Sclar, Yejin Choi, Yulia Tsvetkov and Alane Suhr. 'Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting'. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=RIu5lyNXjT.

[18] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang and David Sontag. 'TabLLM: Few-shot Classification of Tabular Data with Large Language Models'. In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. Proceedings of Machine Learning Research. PMLR, 2023, pp. 5549–5581. URL: https://proceedings.mlr.press/v206/hegselmann23a.html.

[19] Paul Röttger, Bertie Vidgen, Dirk Hovy and Janet Pierrehumbert. 'Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks'. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 175–190. DOI: 10.18653/v1/2022.naacl-main.13. URL: https://aclanthology.org/2022.naacl-main.13/.

[20] Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva and Peter H. Ditto. 'Mapping the Moral Domain'. In: *Journal of Personality and Social Psychology* 101.2 (2011), pp. 366–385. DOI: 10.1037/a0021847. URL: https://doi.org/10.1037/a0021847.

[21] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang and Tatsunori Hashimoto. 'Whose Opinions Do Language Models Reflect?' In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 29971–30004. URL: https://proceedings.mlr.press/v202/santurkar23a.html.

[22] Myra Cheng, Tiziano Piccardi and Diyi Yang. 'CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations'. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023, pp. 10853–10875. DOI: 10.18653/v1/2023.emnlp-main.669. URL: https://aclanthology.org/2023.emnlp-main.669/.

[23] S. S. Stevens. 'On the Theory of Scales of Measurement'. In: *Science* 103.2684 (1946), pp. 677–680. DOI: 10.1126/science.103.2684.677. URL: https://www.science.org/doi/10.1126/science.103.2684.677.

[24] James Dougherty, Ron Kohavi and Mehran Sahami. 'Supervised and Unsupervised Discretization of Continuous Features'. In: *Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 194–202. DOI: 10.1016/B978-1-55860-377-6.50032-3. URL: https://doi.org/10.1016/B978-1-55860-377-6.50032-3.

[25] Isabelle Guyon and André Elisseeff. 'An Introduction to Variable and Feature Selection'. In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182. URL: https://www.jmlr.org/papers/v3/guyon03a.html.

[26] Raj Shah, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj and Sashank Varma. 'Numeric Magnitude Comparison Effects in Large Language Models'. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 6147–6161. DOI: 10.18653/v1/2023.findings-acl.383. URL: https://aclanthology.org/2023.findings-acl.383/.

[27] Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A. Bakker, Georgina Evans, Iason Gabriel, Noah Goodman and Verena Rieser. 'Value Profiles for Encoding Human Variation'. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 2047–2095. DOI: 10.18653/v1/2025.emnlp-main.106. URL: https://aclanthology.org/2025.emnlp-main.106/.

[28] Younghun Lee and Dan Goldwasser. 'SOLAR: Towards Characterizing Subjectivity of Individuals through Modeling Value Conflicts and Trade-offs'. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 20836–20851. DOI: 10.18653/v1/2025.emnlp-main.1053. URL: https://aclanthology.org/2025.emnlp-main.1053/.

[29] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. 'Language Models are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877–1901. URL: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[30] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi and Luke Zettlemoyer. 'Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?' In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 11048–11064. DOI: 10.18653/v1/2022.emnlp-main.759. URL: https://aclanthology.org/2022.emnlp-main.759/.

[31] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin and Weizhu Chen. 'What Makes Good In-Context Examples for GPT-3?' In: *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Dublin, Ireland and Online: Association for Computational Linguistics, 2022, pp. 100–114. DOI: 10.18653/v1/2022.deelio-1.10. URL: https://aclanthology.org/2022.deelio-1.10/.

[32] Yiming Zhang, Shi Feng and Chenhao Tan. 'Active Example Selection for In-Context Learning'. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 9134–9148. DOI: 10.18653/v1/2022.emnlp-main.622. URL: https://aclanthology.org/2022.emnlp-main.622/.

[33] Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang and Dacheng Tao. 'Revisiting Demonstration Selection Strategies in In-Context Learning'. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 9090–9101. DOI: 10.18653/v1/2024.acl-long.492. URL: https://aclanthology.org/2024.acl-long.492/.

[34] Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar and Wenming Ye. 'In-Context Learning with Iterative Demonstration Selection'. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 7441–7455. DOI: 10.18653/v1/2024.findings-emnlp.438. URL: https://aclanthology.org/2024.findings-emnlp.438/.

[35] Nils Reimers and Iryna 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

[36] An Yang et al. *Qwen2.5 Technical Report*. 2024. DOI: 10.48550/arXiv.2412.15115. arXiv: 2412.15115 [cs.CL]. URL: https://arxiv.org/abs/2412.15115.

[37] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. DOI: 10.48550/arXiv.2407.21783. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[38] Quinn McNemar. 'Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages'. In: *Psychometrika* 12 (1947), pp. 153–157. DOI: 10.1007/BF02295996.

[39] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff and Yejin Choi. 'Position: A Roadmap to Pluralistic Alignment'. In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 46280–46302. URL: https://proceedings.mlr.press/v235/sorensen24a.html.

[40] Claude E. Shannon. 'A Mathematical Theory of Communication'. In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[41] Edward H. Simpson. 'Measurement of Diversity'. In: *Nature* 163 (1949), p. 688. DOI: 10.1038/163688a0. URL: https://doi.org/10.1038/163688a0.


TU Delft