Seeing All Sides: Multi-Perspective In-Context Learning for Subjective NLP

Anonymous ACL submission

Abstract

Modern language models excel at factual reasoning but struggle with value diversity: the multiplicity of plausible human perspectives. Tasks such as hate speech or sexism detection expose this limitation, where human disagreement captures the diversity of perspectives that models need to account for, rather than dataset noise. In this paper, we explore whether multi-perspective in-context learning (ICL) can align large language models (LLMs) with this diversity without parameter updates. We evaluate four LLMs on five datasets across three languages (English, Arabic, Italian), considering three label-space representations (aggregated hard, disaggregated hard, and disaggregated soft) and five demonstration selection and ordering strategies. Our multi-perspective approach outperforms standard prompting on aggregated English labels, while disaggregated soft predictions better align with human judgments in Arabic and Italian datasets. These findings highlight the importance of perspective-aware LLMs for reducing bias and polarization, while also revealing the challenges of applying ICL to socially sensitive tasks. We further probe the model faithfulness using XAI, offering insights into how LLMs handle human disagreement.

1 Introduction

Traditional classification aggregates multiple annotator labels into a *single* ground truth, which works under full agreement but fails for subjective NLP tasks, where disagreement can arise from ambiguity, differing viewpoints, or multiple plausible interpretations (Plank et al., 2014). For example, hate speech and offensive language detection often involve contentious annotations, as interpretations vary with personal experiences and cultural backgrounds (Davani et al., 2021; Akhtar et al., 2021), making these tasks highly subjective and complex (Del Arco et al., 2021; Husain and

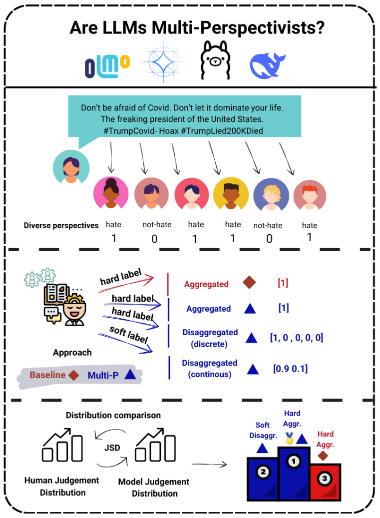


Figure 1: Overall, multi-perspective prompting outperforms the baseline, particularly for aggregated English and disaggregated Arabic & Italian soft labels, highlighting both task- and language-dependent behavior.

Uzuner, 2021). Recent research shows that leveraging disagreements in human-annotated datasets can improve model performance and confidence (Casola et al., 2023; Davani et al., 2022; Muscato et al., 2024). This perspective-driven paradigm, known as Perspectivism (Basile et al., 2021), advocates designing systems that are *perspective-aware*, *responsible*, and *socially intelligent* (Liu et al., 2023; Kovač et al., 2023). Despite these advances, LLMs—including GPT-series models—struggle to capture the diversity of user preferences (Pavlovic and Poesio, 2024; Feng et al., 2024) across sociodemographic groups (Wang et al., 2025) and remain prone to inherent biases in ambiguous, subjective contexts. In this paper, we examine the interplay between ICL and human disagreement in subjective NLP tasks, investigating **whether multi-perspective ICL can serve as an effective and practical alternative to fine-tuning for capturing diverse perspectives**, especially under limited annotated data and computational constraints.

We systematically investigate how different ICL configurations influence model alignment with human disagreement across subjective NLP tasks. Our evaluation setup supports **four open-source LLMs** evaluated on **five datasets** spanning **five subjective tasks**: the first three in **English** (hate speech, offensive, and abusive language detection), the fourth in **Arabic** (sexism detection), and the fifth in **Italian** (irony detection). We assess **three label-space representations** (aggregated hard, disaggregated hard, and soft) and **five demonstration selection and ordering strategies**, including textual similarity, annotator disagreement, two-stage ranking, random, and Curriculum Learning (CL) (Liu et al., 2024), to compare our **multi-perspective ICL approach**—which explicitly instructs models to consider diverse viewpoints—against a **standard prompting baseline**.

The multi-perspective approach outperforms the baseline on aggregated labels in English datasets but struggles with subjective nuances; in Arabic and Italian datasets, disaggregated soft predictions better align with human judgments, demonstrating effectiveness across multilingual and culturally diverse contexts. These findings highlight the need for **perspective-aware LLMs** to mitigate polarization and societal bias (Figure 1), while also illustrating ongoing challenges in applying ICL to **subjective and socially sensitive tasks**.

The most related work to ours is Pavlovic and Poesio (2024), which uses a single closed LLM (GPT-3.5-turbo) via zero-shot ICL with role-playing to compare human and model opinion distributions, producing only disaggregated soft labels on our English and Arabic benchmark datasets. We extend this line of research with the following key contributions, forming a systematic and extensible evaluation framework:

- **Prompting strategies:** Evaluating both zero-shot and few-shot ICL settings.
- **Demonstration design:** Incorporating task-appropriate demonstration selection and ordering strategies for subjective tasks.
- **Label representation:** Expanding the label space to include aggregated hard as well as disaggregated hard and soft labels.
- **Explainability:** Leveraging Integrated Gradients (IG) attribution method to probe model faithfulness in subjective NLP tasks.

Through these contributions, this evaluation setup provides a **unified, systematic, and extensible framework** for assessing LLMs' capabilities and limitations in capturing human disagreement via ICL **across multiple languages**.

2 Background & Related Work

2.1 Socially Intelligent and Responsible LLMs

Socially intelligent AI systems are those capable of understanding and reasoning about intentions, emotions, and mental states (Sap et al., 2022; Qiu et al., 2022). By interacting and adapting to subjective preferences and beliefs (Mittelstädt et al., 2024; Kirk et al., 2025), they can align more effectively with diverse user needs (Mathur et al., 2024). Within Responsible AI, social intelligence supports pluralistic alignment, helping LLMs reduce harm and deliver societal benefit beyond mere performance gains (Sorensen et al., 2024; Tahaei et al., 2023; Kirk et al., 2024).

2.2 Learning from Human Disagreement

Human disagreement, traditionally viewed as noisy crowd-sourcing error, is increasingly recognized as plausible human label variation (HLV) (Plank, 2022), particularly in subjective tasks such as hate speech or sexism detection, where a single ground truth may not exist. Traditional approaches rely on hard (discrete) aggregated labels via majority voting, whereas perspectivist methods (Frenda et al., 2024) preserve disagreement using hard or soft (continuous) disaggregated labels. Models fine-tuned on disaggregated soft labels have been shown to outperform those trained on aggregated hard labels (Van Der Meer et al., 2024). Other strategies include annotator-specific ensembles (Akhtar et al., 2021), multi-task architectures (Davani et al., 2022), and incorporating socio-demographic information (Fleisig et al., 2023).

2.3 In-Context Learning (ICL)

ICL enables LLMs to learn new tasks by analogy without parameter updates (Dong et al., 2024b; Winston, 1980). Zero-shot ICL uses no examples, while few-shot provides a small demonstration set. ICL excels at complex reasoning (Wei et al., 2022) and role-playing (Kong et al., 2024), but it is computationally costly, less efficient than PEFT (Liu et al., 2022a), and highly sensitive to prompt design, including example selection and ordering (Gao et al., 2021; Liu et al., 2024; Peng et al., 2024).

Although cost-effective, ICL remains underexplored in subjective tasks. While LLMs can approximate expert label distributions (Chen et al., 2024), it is unclear whether this extends to non-expert disagreements, with prompt sensitivity further limiting reliability. Performance depends on demonstration selection and ordering: random or human-curated examples (Brown et al., 2020; Kazemi et al., 2023) give mixed results, while kNN (Liu et al., 2022b), latent similarity (Wang et al., 2024), and perplexity (Gonen et al., 2023) favor relevant examples. Ordering strategies like chain-of-thought (Qin et al., 2024) and entropy-based (Lu et al., 2022) reduce prompt sensitivity.

3 Prompting LLMs for Multi-Perspective (*MultiP*)

We assess whether LLMs can capture diverse viewpoints by prompting them with explicit multi-perspective (*MultiP*) instructions[1]. Formally, given a subjective task t with input text x and annotations $A = \{a_1, \ldots, a_n\}$, the model M predicts $\hat{y}$ via zero- or few-shot ICL with the following components (Wang et al., 2022a):

- Task Definition: Defines a subjective task t and explains how input text x is mapped to the appropriate label space for t.
- Label Space: A specification of the expected output label $\hat{y}$, whether as an aggregated or disaggregated hard or soft label l.
- Demonstration Examples: A reference input-output pairs $D = \{x'_j, y'_j\}$, for $j = \{1, ..., m\}$ (with m examples), only for few-shot learning.

To evaluate LLMs' *MultiP* capabilities via ICL, we test variations in three prompt components: task definition t (Section 3.1), label space l (Section 3.2), and demonstration arrangement (Section 3.3). Our *MultiP* prompt template is shown in Box 3.1 ↑ in green, and a few-shot prompt sample, where t contains the instructions for hate speech detection with an aggregated hard label (l) is illustrated in Box 3.1 ↓ in purple.

[1] https://anonymous.4open.science/r/ICL_with_Disagreement-F3AC

Our *MultiP* Prompt Template

TASK DEFINITION (t):
- Hate speech
- Offensive language
- Abusive language

LABEL SPACE (l):
- **Hard**: Aggregated or Disaggregated
- **Soft** : Disaggregated

DEMONSTRATION EXAMPLE(S) (D):
- `(text, hard agg.)`: (e.g., `yes`)
- `(text, hard disagg.)`: (e.g., `[0,0,1,1,0]`)
- `(text, soft)`: (e.g., `[0.7, 0.3]`)

INPUT:
- Tweet (x): `{text}`
- Answer ($\hat{y}$): `[output]`

Example *MultiP* Prompt for Hate Speech

[t] Does the following tweet contain hate speech, particularly xenophobia or islamophobia? **The task is subjective, so please answer considering different perspectives** from Muslim immigrants as well as others from different backgrounds.
[l] There are two options: *yes* and *no*.
[D] Examples: Any future terrorist attack in Europe will be blame on Brexit by the lmsm, yes
Now consider the following example and only output your option without punctuation.
[x] Tweet: What the referendum seem to have mean to alarm number a vote for anyone look foreign to leave immediately
[$\hat{y}$] Answer:

3.1 Task Definition

Typically, LLMs are prompted to directly answer questions (e.g., Classify the following tweet as hate speech (Antypas et al., 2023)) without accounting for task subjectivity. In this study, we explore two approaches: the baseline (*BS*) and the *MultiP* method.

Baseline (*BS*) Priming The *BS* represents a scenario in which M *is prompted to produce a single aggregated label*, ignoring the subjectivity or ambiguity of t. Specifically, the example in Box 3.1 ↓ (highlighted in purple) is adjusted: t excludes the bold statement, while l remains unchanged, yielding $\hat{y}$ as an aggregated hard label.

Multi-Perspective (*MultiP*) Priming To mitigate M's tendency to favor a dominant viewpoint and enhance its contextual understanding of t, the model is *explicitly instructed to consider multiple perspectives* from different viewpoints, as highlighted in the bold statement in Box 3.1 ↓ (purple). Inspired by Pavlovic and Poesio (2024); Lan et al. (2024), we prompt M *to adopt the role of an expert in task* t, applying role-playing (*RL*) for both the *BS* and *MultiP* approaches.

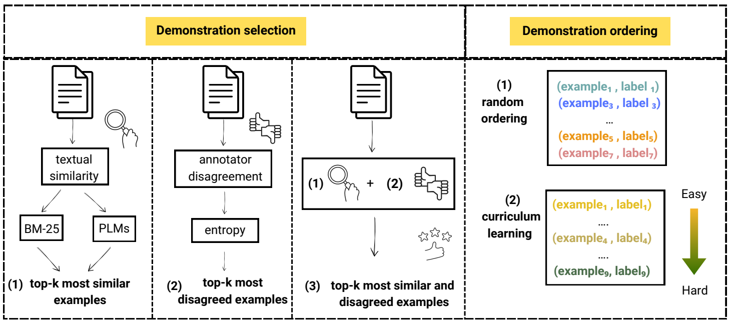


Figure 2: *MultiP* ICL with three demonstration selection strategies—BM25/PLM similarity, entropy-based annotator disagreement, and a two-stage similarity + re-ranking—alongside two ordering strategies: random and CL.

3.2 Label Space

To better model subjectivity, we define three types of label predictions in our *MultiP* framework: **aggregated hard** (discrete), **disaggregated hard** (discrete), and **disaggregated soft** (continuous) labels. When the model M is prompted to generate a prediction $\hat{y}$ within a label space l, all demonstration examples are likewise represented within the same space l.

3.2.1 Hard Labels

Aggregated A single ground-truth label (e.g., hate speech or not) is obtained through vote aggregation, typically majority voting.

Disaggregated Individual labels from multiple annotators capture the diversity of human judgments. With n annotators, M is prompted to generate a label set A', where each $a'_i \in A'$ for $i = 1, \ldots, n$, represents the predicted annotation from each annotator. To obtain disaggregated soft labels from aggregated or disaggregated hard labels[2], we extract the probability scores from M corresponding to the predicted labels (Lee et al., 2023).

3.2.2 Soft Labels

Disaggregated A probability distribution over the possible classes represents the likelihood of each category. For example, in a binary hate speech classification task (t), M is prompted to output probabilities as $[P_{hate}, P_{not\_hate}]$, with values summing to one—a condition that LLMs consistently satisfy.

3.3 Demonstration Examples

Prior work improves demonstration organization (Dong et al., 2024a), mainly for objective tasks, but its effectiveness in subjective tasks remains underexplored. Moreover, annotator disagreement, crucial in subjective tasks, is rarely considered in example selection. To address this, we investigate different methods for selecting (Section 3.3.1) and ordering (Section 3.3.2) demonstration examples in subjective tasks, as illustrated in Figure 2, inspired by Liu et al. (2024).

3.3.1 Demonstration Selection

We investigate few-shot ICL by comparing traditional example selection based on textual similarity (Peng et al., 2024; Zhang et al., 2024) with methods that also incorporate annotator disagreement into the selection process, exploring three strategies for k-shot selection (one per class).

Textual Similarity We compute the similarity between target test examples and the training set, retrieving the top-k most similar texts based on a fixed threshold[3]. Similarity is measured using: (1) BM25 (Robertson et al., 2009), and (2) following Peng et al. (2024), cosine similarity between sentence embeddings from pre-trained language models (PLMs), specifically all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-distilroberta-v1, and all-mpnet-base-v2 from huggingface[4].

Annotator Disagreement We hypothesize that examples with high annotator disagreement are more informative. Thus, we select the top-k samples with highest annotation entropy, where higher entropy denotes greater ambiguity and lower entropy reflects clearer consensus.

Two-Stage Ranking Preliminary experiments reveal that prior methods often yield examples that are either too similar or too diverse, limiting generalization. To mitigate this, we adopt a two-stage approach (Dang et al., 2013) that balances textual similarity and annotator disagreement: the top-k most similar training examples are first retrieved,

[2]This transformation is only required for evaluation; see Section 4.3 for details.

[3]Thresholds from 0.5 to 0.8 are tested, with 0.7 yielding the best performance in terms of cosine similarity.

[4]https://huggingface.co/sentence-transformers

then re-ranked by disagreement to select those that are both relevant and diverse for few-shot ICL.

3.3.2 Demonstration Ordering

We use random ordering as a baseline and explore curriculum learning (CL) (Liu et al., 2024), where examples are ranked by annotation entropy. We hypothesize that the model benefits from learning *simpler*, high-*agreement* examples before more *challenging*, high-*disagreement* ones.

4 Experimental Setup

4.1 Datasets

We use five benchmark datasets in English, Arabic, and Italian, each corresponding to a different subjective task t, from the SemEval 2023 (LeWiDi) *Learning with Disagreements*[5] competition (Leonardelli et al., 2023; Casola et al., 2024). These datasets cover a range of subjective tasks with varying numbers of annotations n (Table 1).

Dataset	Train	Test	Dev	Tot. Class	Ann.	Full Agr. (%)	Subj. Task
HS-Brexit	784	168	168	2	6	69%	Hate speech
MD-Agr	6592	3057	1104	2	5	42%	Offensive lang.
ConvAbuse	2398	840	812	2	3-8	86%	Abusive lang.
ArMIS	657	145	141	2	3	65%	Misoginy and sexism
MultiPICo	600	200	200	2	3	64%	Irony

Table 1: Dataset statistics

Hate Speech on Brexit HS-Brexit (Akhtar et al., 2021) contains English tweets around the Brexit vote, filtered for keywords on immigrants and Brexit, annotated by six annotators to capture opinions on immigrants' role in UK society.

Multi-Domain Agreement MD-Agreement (Leonardelli et al., 2021) covers three popular 2020 topics—Covid-19, the US Presidential elections, and the Black Lives Matter movement. Tweets were collected using topic-specific keywords and each instance was labeled by five annotators.

ConvAbuse ConvAbuse (Cercas Curry et al., 2021) comprises English conversations between users and three conversational AI systems (Alana v2, Eliza, and CarbonBot), each with distinct goals. Each example is labeled by three to eight annotators[6].

ArMIS This dataset (Almanea and Poesio, 2022) contains Arabic tweets with binary labels, annotated by three annotators (conservative male, moderate female, liberal female) to study political leaning effects on sexism judgments.

MultiPICo This dataset (Casola et al., 2024) contains short post-reply conversations from Twitter and Reddit annotated for irony; we focus on the Italian subset of the nine-language corpus.

4.2 LLMs

We prompt four open-source instruction-tuned LLMs—Olmo-7b-Instruct[7], Llama-3-8b-Instruct[8], Gemma-7b-it[9], Deepseek-7b-chat[10]—following the methodology (Section 3). Decoding was performed via greedy search to ensure reproducibility.

4.3 Evaluation metrics

This *MultiP* paradigm uses soft metrics[11], such as Jensen-Shannon Divergence (JSD) and Cross-Entropy (CE) (Uma et al., 2021), rather than hard metrics like precision or F1, which consider only the most probable class and miss nuances in ambiguous instances (Wang et al., 2022b). We use JSD as the primary soft metric to measure distance between probability distributions, and CE to assess the model's confidence in its top prediction[12].

Due to space constraints, we report only the best-performing models; full results are available in our GitHub repository, with macro F1 as the primary hard metric. In Section 5.3, micro F1 is reported for comparison with state-of-the-art results, and macro F1 is reported only for the MultiPICo baseline (Casola et al., 2024). Higher values indicate better performance for hard metrics, while lower values are better for soft metrics.

5 Results

We evaluate *MultiP* ICL on five subjective tasks spanning three languages: English, Arabic, and Italian **comparing the *BS* with *MultiP* in zero-shot ($0S$) and few-shot (FS) settings**. In both $0S$ and FS settings[13], we examine the effect of

[5] https://le-wi-di.github.io

[6] Labels are in binary format (Vitsakis et al., 2023).

[7] https://huggingface.co/allenai/OLMo-7B-Instruct

[8] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[9] https://huggingface.co/google/gemma-7b-it

[10] https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat

[11] Soft metrics assess models' ability to predict both preferred and alternative interpretations (Rizzi et al., 2024).

[12] Aggregated and disaggregated hard labels are converted to soft labels only during evaluation (Section 3.2).

[13] Larger versions of all tables and figures are provided in Appendix A and B

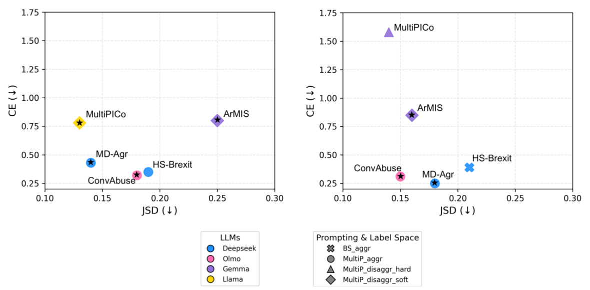


Figure 3: **Best $0S$ (left) and FS models (right) per dataset, including *BS* and *MultiP* with or without *RL*.** Each plot shows JSD vs. CE, with points colored by model, shaped by *BS/MultiP* & *aggr/disaggr*, and annotated by dataset. RL is marked by a black "$\star$".

the label space l (Section 3.2); in FS, we additionally assess the impact of demonstration examples D (one-shot per class) via different selection and ordering strategies (Section 3.3).

5.1 Impact of the Label Space

Zero-shot ($0S$) *The MultiP approach outperforms BS on aggregated English labels, showing lower JSD and CE (left plot, Figure 3). For Arabic and Italian, disaggregated soft predictions better align with human judgments, demonstrating robust performance across multilingual and culturally diverse settings.*

For all three English datasets, $MultiP\_aggr$ (●) performs best, whereas for the Arabic and Italian datasets, $MultiP\_disaggr\_soft$ (◆) achieves the highest performance. Regarding the LLMs, DeepSeek-7b-chat performs best for hate speech (HS-Brexit) and offensive language (MD-Agr.), achieving (0.19 JSD, 0.35 CE) and (0.14 JSD, 0.43 CE with RL), respectively. Olmo-7b-Instruct leads on abusive language (ConvAbuse) (0.18 JSD, 0.32 CE); Gemma-7b-it on Arabic sexism (ArMIS) (0.25 JSD, 0.80 CE); and Llama3-8b-it on Italian irony (MultiPICo) (0.13 JSD, 0.78 CE). Results indicate that, except for HS-Brexit, the best-performing $0S$ configurations include role-playing (RL), highlighting its effectiveness for modeling subjective disagreement.

Few-shot (FS) *The MultiP approach outperforms BS in most cases, except for the HS-Brexit dataset (right plot, Figure 3), achieving lower JSD and CE on aggregated English labels. For Italian, disaggregated hard predictions, and for Arabic, disaggregated soft predictions better align with human judgments.*

For two English datasets, $MultiP\_aggr$ (●) performs best (MD-Agr: JSD 0.18, CE 0.25; ConvAbuse: JSD 0.15, CE 0.31), whereas for HS-Brexit, *BS_aggr* achieves the highest performance (JSD 0.21, CE 0.39). For the non-English datasets, $MultiP\_disaggr\_soft$ (◆) and $MultiP\_disaggr\_hard$ (▲) achieve the top performance for Arabic (ArMIS: JSD 0.16, CE 0.85) and Italian (MultiPICo: JSD 0.14, CE 1.58), respectively. Consistent with the $0S$ results, DeepSeek-7b-chat is the best-performing LLM for HS-Brexit and MD-Agr. For ConvAbuse, Olmo-7b-Instruct achieves the highest performance, while for both non-English datasets, Gemma-7b-it performs best. Unlike in $0S$ settings, Gemma-7b-it is also the top-performing model for the Italian MultiPICo. Similar to the $0S$ settings, for most tasks, the best-performing FS configuration includes role-playing (RL).

The findings are largely consistent across $0S$ and FS settings. *MultiP* outperforms *BS* across all subjective tasks except hate speech in FS. The best-performing LLM for each dataset is the same in both $0S$ and FS settings, except for the Italian MultiPICo. Furthermore, RL enables LLMs to better capture the nuances of human disagreement in these cross-lingual subjective contexts, regardless of whether demonstration examples are provided (FS) or not ($0S$). A two-tailed paired t-test on JSD scores for the best model and configuration of each dataset shows statistical significance ($p < 0.05$).

5.2 Few-shot (FS): Impact of the Demonstration Examples

Demonstration Selection *BM25-based selection (●) achieves the best performance on two English datasets, entropy-based selection (▲) on the third, and two-stage ranking (◆) on the non-English datasets, achieving lower JSD and CE (left plot, Figure 4).*

Specifically, using BM25 (●), MD-Agr. achieves JSD = 0.16 and CE = 0.25, while ConvAbuse reaches JSD = 0.14 and CE = 0.24. In contrast, entropy-based selection (▲) performs best for HS-Brexit, with JSD = 0.18 and CE = 0.39. For the non-English datasets, two-stage ranking (◆) attains the highest performance: ArMIS achieves JSD = 0.13 and CE = 0.73, and MultiPICo achieves JSD = 0.13 and CE = 1.07. The results indicate that the effectiveness of demonstration selection varies between English and non-English datasets, as shown in the left plot of Figure 4 under random ordering.

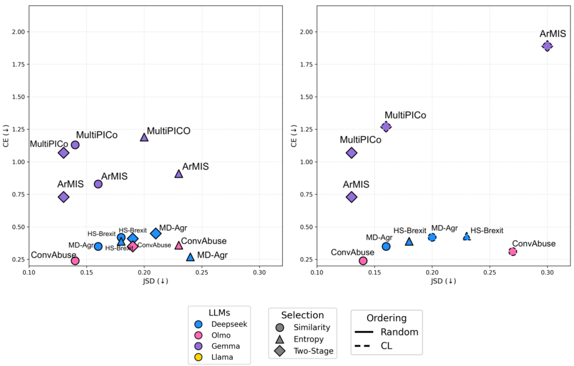


Figure 4: **Best FS models per dataset - left: demonstration selection (BM25, entropy, and two-stage ranking with random ordering), right: demonstration ordering (random & CL)**. Each plot shows JSD vs. CE, with points colored by model, shaped by the selection method (left), line style indicating ordering method (right), and annotated by dataset.

Demonstration Ordering *For all datasets, random ordering (——) outperforms the CL (- - -), achieving lower JSD and CE (right plot, Figure 4).* The effect of demonstration ordering is consistent across tasks and languages, with random ordering (——) outperforming CL (- - -) for all datasets in the right plot (Figure 4), which compares each dataset's best selection strategy.

5.3 Comparison with State-of-the-Art Models

Our best models achieve competitive or superior performance compared with LeWiDi, GPT-3.5, and MultiPICo baselines, attaining low CE and strong micro and macro F1 scores (Table 2), effectively capturing human disagreement across languages and tasks.

Our results are comparable to those from the LeWiDi competition (Leonardelli et al., 2023), which employs fine-tuned BERT-based models, and to Pavlovic and Poesio (2024), the study most closely aligned with ours, which applies GPT-3.5_disaggr_soft_$0S$ to model human label disagreement. Both the baseline and the best LeWiDi_BS_aggr models predict aggregated labels, with the latter employing an ensemble trained on multiple annotators' labels. As shown in Table 2, for a fair comparison, the best $aggr$ models (in $0S$ or FS) are evaluated against the fine-tuned LeWiDi_BS_aggr models, while the best $disaggr\_soft$ models in $0S$ are compared with GPT-3.5_disaggr_soft_$0S$. We report CE alongside micro F1[14].

[14] Micro F1 measures overall performance but may overlook class imbalance.

Dataset	Approach	*micro* F1(↑)	CE(↓)
HS-Brexit	LeWiDi_BS_aggr (Baseline)	84.20	2.72
	LeWiDi_BS_aggr (Best)	**92.90**	**0.24**
	Our_$MultiP$_aggr_$0S$ (Best)	89.29	0.35
	GPT-3.5_disaggr_soft_$0S$	69.60	5.04
	Our_$MultiP$_disaggr_soft_$0S$ (Best)	**70.24**	**0.52**
MD-Agr	LeWiDi_BS_aggr (Baseline)	53.40	7.39
	LeWiDi_BS_aggr (Best)	**84.60**	0.47
	Our_$MultiP$_aggr_FS (Best)	75.23	**0.25**
	GPT-3.5_disaggr_soft_$0S$	52.00	3.83
	Our_$MultiP$_disaggr_soft_$0S$ (Best)	**74.81**	**0.32**
ConvAbuse	LeWiDi_BS_aggr (Baseline)	74.10	3.48
	LeWiDi_BS_aggr (Best)	**94.20**	**0.19**
	Our_BS_aggr_$0S$_RL (Best)	82.02	0.29
	GPT-3.5_disaggr_soft_$0S$	**90.20**	3.75
	Our_$MultiP$_disaggr_soft_FS (Best)	77.21	**0.39**
ArMIS	LeWiDi_BS_aggr (Baseline)	41.70	8.90
	LeWiDi_BS_aggr (Best)	**83.20**	**0.46**
	Our_$MultiP$_aggr_$0S$ (Best)	44.82	9.87
	GPT-3.5_disaggr_soft_$0S$	25.60	6.67
	Our_$MultiP$_disaggr_soft_$0S$_RL (Best)	**58.62**	**0.80**

Table 2: **Best-performing $0S$ and FS models across BS and $MultiP$ configurations, evaluated in both $aggr$ and $disaggr$ label spaces, with/without RL, and compared against state-of-the-art**. The best $aggr$ models (in $0S$ or FS) are compared against the fine-tuned baseline and the best LeWiDi_BS_aggr models, while the best $disaggr\_soft\_0S$ models are compared with GPT-3.5_disaggr_soft_$0S$. The best scores are shown in **bold**, and the second-best are underlined.

We compare the best models, selected by lowest CE, separately for each dataset. In comparison with LeWiDi_BS_aggr models, our best model achieves the lowest CE on MD-Agr. (0.25), ranks second for HS-Brexit and ConvAbuse (CE 0.35 and 0.29), and shows comparable performance to the LeWiDi baseline on ArMIS (9.87). Against GPT-3.5_disaggr_soft_$0S$, our best models achieve the lowest CE across all four datasets, including ArMIS (Arabic), HS-Brexit (0.52), MD-Agr (0.32), ConvAbuse (0.39), and ArMIS (0.80). In terms of micro F1, compared with LeWiDi_BS_aggr, our best models consistently rank second across all four datasets: HS-Brexit (89.29), MD-Agr (75.23), ConvAbuse (82.02), and ArMIS (44.82). In contrast to GPT-3.5_disaggr_soft_$0S$, our best model outperforms GPT on all datasets except ConvAbuse, achieving micro F1 scores of HS-Brexit (70.24), MD-Agr (74.81), and ArMIS (58.62). Notably, our best model strongly outperforms GPT on ArMIS and shows a slight improvement on HS-Brexit.

In addition to the four datasets, for MultiPICo, we compare our best model with the top model reported in the original paper using macro F1 (Casola et al., 2024). The top $MultiP$ model is GPT-3.5_aggr_$0S$ for aggregated labels (53.30), followed by PolyLM (Polyglot LLM)[15] (42.60). Our best model, DeepSeek-7b-chat in $0S$ settings, achieves 45.77 (BS_aggr) and

[15] https://huggingface.co/DAMO-NLP-MT/polylm-chat-13b)

44.38 ($MultiP$_aggr), both ranking second overall. These results are excluded from Table 2, as MultiPICo (Casola et al., 2024) reports macro F1.

6 Feature Attribution for Explaining Model Predictions

Inspired by Zhou et al. (2024), we use the feature attribution method to provide insights into model decision-making process. Particularly, we probe model faithfulness at the instance level by prompting the model to generate explicit class labels in both the $aggr$ (yes/no) and $disaggr$ (probability score) formats. For attribution, we employ Integrated Gradients (IG) (Sundararajan et al., 2017), a post-hoc attribution method that quantifies the contribution of each input feature by integrating gradients from a baseline input[16] (Kokhlikyan et al., 2020).

To ensure explanations are faithful to the model's reasoning, we first require that the predicted label from the black-box LLM matches the label obtained via IG attribution for each instance; when this alignment holds, the highlighted tokens reflect the input features that contributed to the model's decision. We further evaluate faithfulness by examining whether higher-attribution tokens have a proportionally greater impact on the black-box LLM's output. Specifically, we mask the top-3 IG-identified tokens[17] and measure the resulting prediction shifts: for $aggr$, we check for label flips, while for $disaggr$, we compute the JSD between pre- and post-masking probability distributions, applying a threshold[18] to identify significant deviations.

HS-Brexit (English) vs. MultiPICo (Italian) We explore model faithfulness[19] in capturing human disagreement in subjective NLP tasks. For this preliminary analysis, we select the five instances with the lowest JSD scores per approach, prioritizing cases where the predicted distribution is closest to the gold distribution. The selected approaches are BS and $MultiP$ ($aggr$) for HS-Brexit, and BS and $MultiP$ ($disaggr$) for MultiPICo. Our experiments indicate that $aggr$ performs better for English datasets, whereas $disaggr$ is more effective for non-English ones (Section 5.1). The selected instances are available in our GitHub repository. In HS-Brexit, both approaches yield a similar class distribution among the selected examples[20], whereas in MultiPICo, the distribution differs[21]. As the black-box LLM, we use the best models for each dataset: DeepSeek-7b-chat for hate-speech (HS-Brexit) and Llama3-8b-it for Italian irony detection (MultiPICo) in $0S$ settings.

In HS-Brexit, $MultiP$ exhibits higher faithfulness (0.80) than BS (0.40), while in MultiPICo, both approaches perform comparably (0.60). This divergence reflects differing subjectivity profiles of the datasets: HS-Brexit's polarized, lexically explicit cues are readily isolated by attribution-based methods, amplifying masking effects, whereas MultiPICo relies on implicit, context-dependent cues, limiting token-level salience and producing negligible differences between BS and $MultiP$.

7 Conclusion & Future Work

We present an adaptable evaluation framework that yields novel insights into LLM alignment. Our results show that multi-perspective ICL improves alignment with diverse human judgments in socially sensitive tasks, including hate speech and sexism detection. Across five datasets, three languages, and three label-space representations, our approach enhances performance on aggregated English labels and more accurately captures human disagreement in Arabic and Italian through disaggregated soft predictions. Instance-level XAI analysis reveals how LLMs rely on specific tokens when handling human disagreement, highlighting the importance of perspective-aware prompting for building faithful and culturally sensitive models.

This work has several potential implications. While our framework provides insights, it cannot verify whether annotators' perspectives reflect the populations in benchmark datasets, highlighting the need for annotation frameworks that capture population diversity. Although prior work shows that incorporating annotator demographics can improve robustness, we argue that value profiling via perspectives is more important and also offers privacy-preserving benefits. Future studies could examine how attention patterns or latent representations correlate with specific human perspectives, enabling more principled alignment interventions.

[16]Experiments use the Captum library:https://captum.ai.

[17]Empirically selected based on attribution scores.

[18]Set to 0.2, determined from preliminary analysis.

[19]Fraction of cases in which the model changes its prediction after masking the top-attribution tokens.

[20]BS and $MultiP$ (*hateful*:2, *not hateful*:3)

[21]BS (*hateful*:1, *not hateful*:4), $MultiP$ (*hateful*:2, *not hateful*:3)

Limitations

This study has several limitations. Our analysis is constrained by limited resources, as perspectivism remains an emerging paradigm. Consequently, we rely on the LeWiDi competition datasets and the Italian MultiPICo, which currently serve as the main benchmarks in perspectivist NLP. Our focus on subjective tasks—hate speech, offensive language, abusive language, sexism, and irony detection—reflects dataset availability; however, their binary classification setup limits generalization to more complex, multi-class scenarios. This study relies on benchmark datasets and does not assess whether annotators' perspectives capture the diversity of modeled populations. Future annotation frameworks for subjective NLP tasks should better reflect population diversity.

Previous research (Ding et al., 2022) argues that annotators' demographics should be considered and recorded during annotation to support more robust model training. In contrast, we argue that capturing annotators' perspectives (Deng et al., 2023) through value profiling (Peter and Devlin, 2025)—rather than their identities—is crucial for building more inclusive, perspective-aware, and better-performing models. Beyond this, we leave the exploration of permutation invariance in prompt construction with disaggregated hard labels for future work, since it was not feasible within our current computational and time constraints.

Additionally, the current study provides only instance-level insights into modeling human disagreement via model faithfulness. A systematic investigation using diverse XAI approaches is needed to better understand model reasoning and its correlation with specific human perspectives. Future work will extend the presented adaptable framework to a broader range of model architectures and access modalities, including both open-source and proprietary LLMs such as Claude 3.5[22] and Gemini Pro[23].

Ethics Statement

This research examines how LLMs capture diverse human perspectives in subjective tasks, such as hate speech and sexism detection. We treat human disagreement as a meaningful reflection of value pluralism, designing experiments to respect this diversity rather than suppress it. All datasets used are publicly available or anonymized, complying with privacy standards and data protection regulations. Experiments follow ACL's ethical guidelines, with care taken to avoid harm, bias, or unfair representation. The multi-perspective ICL evaluation framework supports pluralistic modeling and reduces polarization, while instance-level insights increase transparency and interpretability. No personal data was collected or released. We encourage community discussion on ethical annotation, demographic diversity, and perspective-aware, responsible NLP model deployment. All research protocols align with ACL standards for privacy, fairness, transparency, and responsible use. We remain committed to ongoing assessment of ethical risks and limitations associated with this work.

Impact Statement

Our research addresses the risk of societal biases being embedded in NLP systems. By focusing on LLMs and employing multi-perspective prompting, we aim to promote inclusivity and diversity in model behavior. This study highlights the challenges of applying LLMs to subjective tasks, demonstrating the need for perspective-aware approaches. We encourage the NLP community to develop models that are more inclusive, socially responsible, and capable of understanding diverse human viewpoints, ultimately fostering more effective and equitable human–machine interaction.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. SuperTweetEval: A challenging, unified and heterogeneous benchmark for social media NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12590–12607, Singapore. Association for Computational Linguistics.

[22] https://www.anthropic.com/news/claude-3-5-sonnet

[23] https://deepmind.google/models/gemini/pro/

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Silvia Casola, Soda Lo, Valerio Basile, Simona Frenda, Alessandra Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "seeing the big through the small": Can LLMs approximate human judgment distributions on NLI from a few explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.

Van Dang, Michael Bendersky, and W Bruce Croft. 2013. Two-stage learning to rank for information retrieval. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*, pages 423–434. Springer.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2021. Hate speech classifiers learn human-like social stereotypes. *Preprint*, arXiv:2110.14839.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Flor Miriam Plaza Del Arco, Arturo Montejo-Ráez, L Alfonso Urena Lopez, and María-Teresa Martín-Valdivia. 2021. Offendes: A new corpus in spanish for offensive language research. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2021)*, pages 1096–1108.

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *arXiv preprint arXiv:2305.14663*.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024a. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024b. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. Lambada: Backward chaining for automated reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568.

Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. 2025. Why human-ai relationships need socioaffective alignment. *arXiv preprint arXiv:2502.02528*.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and 1 others. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Jiaming Zhou, and Haoqin Sun. 2024. Self-prompt tuning: Enable autonomous role-playing in llms. *arXiv preprint arXiv:2407.08995*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 891–903.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022b. 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*, pages 100–114.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. 2024. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in ai agents: Technical challenges and open questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560.

Justin M Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14(1):27449.

Benedetta Muscato, Praveen Bushipaka, Gizem Gezici, Lucia Passaro, Fosca Giannotti, and 1 others. 2024. Multi-perspective stance detection. In *CEUR WORKSHOP PROCEEDINGS*, volume 3825, pages 208–214. CEUR-WS.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. 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)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Oriane Peter and Kate Devlin. 2025. Decentralising llm alignment: A case for context, pluralism, and participation. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES-25)*.

Barbara Plank. 2022. 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*, pages 10671–10682.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7441–7455.

Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 146–158, Edinburgh, UK. Association for Computational Linguistics.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, and 1 others. 2023. Human-centered responsible artificial intelligence: Current & future trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Michiel Van Der Meer, Neele Falk, Pradeep Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective nlp tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. ilab at semeval-2023 task 11 le-wi-di: Modelling disagreement or modelling perspectives? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2024. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022a. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Yimeng Chen, Shimin Tao, Jiaxin Guo, Chang Su, Min Zhang, and Hao Yang. 2022b. Capture human disagreement distributions by calibrated networks for natural language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1524–1535.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Patrick H Winston. 1980. Learning and reasoning by analogy. *Communications of the ACM*, 23(12):689–703.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.

Wei Zhou, Heike Adel, Hendrik Schuff, and Ngoc Thang Vu. 2024. Explaining pre-trained language models with attribution scores: An analysis in low-resource settings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6867–6875, Torino, Italia. ELRA and ICCL.

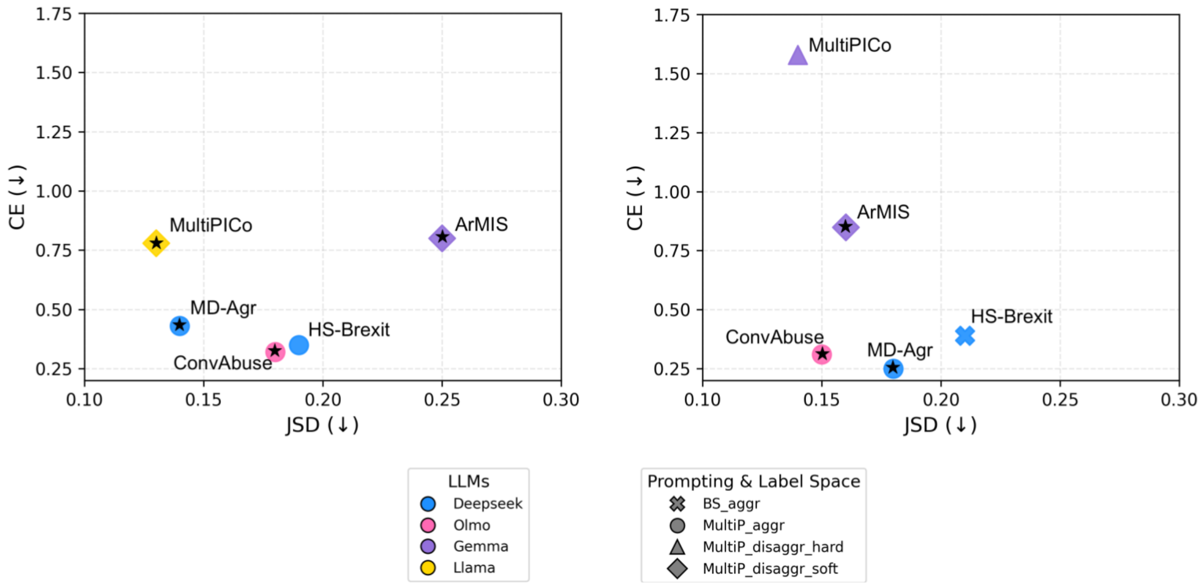
Appendix

A Zero-shot ($0S$) and Few-shot (FS) Results

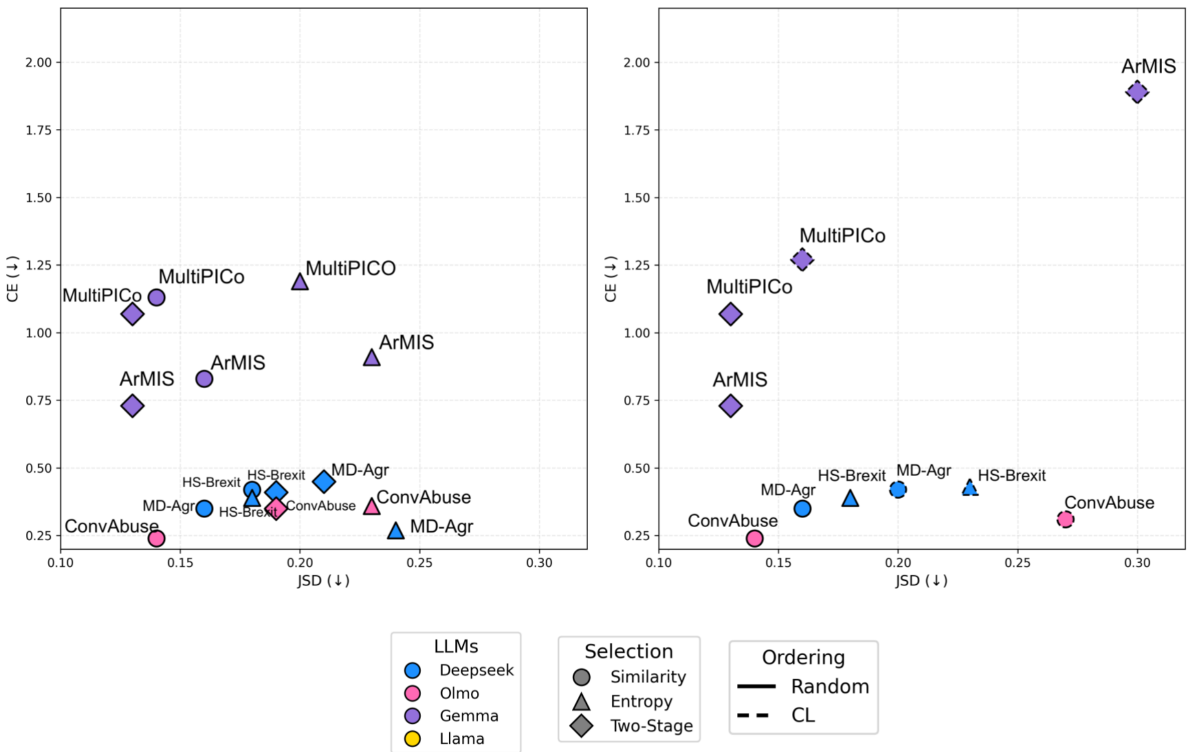
This section provides an extended view of Figures 3 and 4. Figure 5a shows the impact of label space in zero-shot ($0S$) prompting, while Figure 5b illustrates the effect of demonstration selection and ordering across tasks and languages.

B State-of-the-art Results

This section presents Table 3, comparing our best zero-shot ($0S$) and few-shot (FS) models against state-of-the-art approaches from the LeWiDi competition, as summarized in Section 5.3.



(a) Best $0S$ (left) and and FS models (right) per dataset, including *BS* and *MultiP* with or without *RL*, marked by a black "⋆".



(b) Best FS models per dataset- left: demonstration selection (BM25, entropy, and two-stage ranking with random ordering), right: demonstration ordering (random & CL).

Figure 5: Best models per dataset. (a) Zero-shot & Few-shot- Impact of Label Space (b) Few-shot (FS)- Impact of Demonstration Examples.

Dataset	Approach	*micro* F1(↑)	CE(↓)
HS-Brexit	LeWiDi_BS_aggr (Baseline)	84.20	2.72
	LeWiDi_BS_aggr (Best)	**92.90**	**0.24**
	Our_$MultiP$_aggr_$0S$ (Best)	<u>89.29</u>	<u>0.35</u>
	GPT-3.5_disaggr_soft_$0S$	69.60	5.04
	Our_$MultiP$_disaggr_soft_$0S$ (Best)	**70.24**	**0.52**
MD-Agr	LeWiDi_BS_aggr (Baseline)	53.40	7.39
	LeWiDi_BS_aggr (Best)	**84.60**	<u>0.47</u>
	Our_$MultiP$_aggr_FS (Best)	<u>75.23</u>	**0.25**
	GPT-3.5_disaggr_soft_$0S$	<u>52.00</u>	<u>3.83</u>
	Our_$MultiP$_disaggr_soft_$0S$ (Best)	**74.81**	**0.32**
ConvAbuse	LeWiDi_BS_aggr (Baseline)	74.10	3.48
	LeWiDi_BS_aggr (Best)	**94.20**	**0.19**
	Our_BS_aggr_$0S$_RL (Best)	<u>82.02</u>	<u>0.29</u>
	GPT-3.5_disaggr_soft_$0S$	**90.20**	<u>3.75</u>
	Our_$MultiP$_disaggr_soft_FS (Best)	<u>77.21</u>	**0.39**
ArMIS	LeWiDi_BS_aggr (Baseline)	41.70	<u>8.90</u>
	LeWiDi_BS_aggr (Best)	**83.20**	**0.46**
	Our_$MultiP$_aggr_$0S$ (Best)	<u>44.82</u>	9.87
	GPT-3.5_disaggr_soft_$0S$	<u>25.60</u>	<u>6.67</u>
	Our_$MultiP$_disaggr_soft_$0S$_RL (Best)	**58.62**	**0.80**

Table 3: **Best-performing $0S$ and FS models across BS and $MultiP$ configurations, evaluated in both $aggr$ and $disaggr$ label spaces, with/without RL, and compared against state-of-the-art**. The best $aggr$ models (in $0S$ or FS) are compared against the fine-tuned baseline and the best LeWiDi_BS_aggr models, while our best $disaggr\_soft\_0S$ models are compared with GPT-3.5_disaggr_soft_$0S$. The best scores are shown in **bold**, and the second-best are <u>underlined</u>.