Can Large Language Models Capture Human Annotator Disagreements?

Anonymous ACL submission

Abstract

Human annotation variation (i.e., annotation disagreements) is common in NLP and often reflects important information such as task subjectivity and sample ambiguity. While Large Language Models (LLMs) are increasingly used for automatic annotation to reduce human effort, their evaluation often focuses on predicting the majority-voted "ground truth" labels. It is still unclear, however, whether these models also capture informative human annotation variation. Our work addresses this gap by extensively evaluating LLMs' ability to predict annotation disagreements without access to repeated human labels. Our results show that LLMs struggle with modeling disagreements, which can be overlooked by majority label-based evaluations. Notably, while RLVR-style[1] reasoning generally boosts LLM performance, it degrades performance in disagreement prediction. Our findings highlight the critical need for evaluating and improving LLM annotators in disagreement modeling.[2]

1 Introduction

The field of NLP rests on annotations where inter-annotator disagreement is common (Snow et al., 2008). Such disagreement is often treated as inconvenient noise due to human error, "solved" by majority voting (Sabou et al., 2014) or expert aggregation (Hovy et al., 2013).

These ad-hoc solutions may be misguided, as annotation disagreement can signal a diversity of views and is often valuable information (Plank, 2022). Human annotators have access to different information sets and are guided by different value systems (Fornaciari et al., 2021; Fuchs et al., 2021). It is therefore not surprising that different annotators give different answers, in particular for subjective tasks such as hate speech detection (e.g. Kennedy et al., 2018) where disagreement often arises from varying sociodemographic and cultural backgrounds (Fleisig et al., 2023). Even seemingly "objective" labeling tasks, such as part-of-speech (POS) tagging, show disagreement due to ambiguous language[3] (Plank et al., 2014; Jiang and de Marneffe, 2022). Generally speaking, disagreement is natural, contains valuable information, and should not be ignored or erased, but actively modeled (Uma et al., 2021; Leonardelli et al., 2023). To model annotator disagreement, previous work has trained models on datasets with multiple annotations per data point, or used behavioral / sociodemographic information for annotator modeling (Mostafazadeh Davani et al., 2022; Fleisig et al., 2023; Hu and Collier, 2024; Giorgi et al., 2024; Chochlakis et al., 2024, 2025; Orlikowski et al., 2025).

All of the above require the existence of multiply-annotated data. But what about datasets and emergent tasks[4] that lack repeated human labels? Collecting repeated human labels can be expensive. LLMs might prove a reasonable substitute for human annotation, especially given their general effectiveness in text classification (Pangakis et al., 2023a; Törnberg, 2024; He et al., 2024b), judging chatbot preferences (Lee et al., 2024), and simulating human opinion (Meister et al., 2024b; Anthis et al., 2025). However, the performance of these LLM annotators is evaluated against a majority label or agreement with humans (He et al., 2024b; Ni et al., 2024). In that setup, pointwise estimates are more important than label distributions, so whether they can capture human annotation disagreement remains an open question. Therefore, we identify the following **practice-evaluation gap**:

[1]Reinforcement learning with verifiable rewards (Lambert et al., 2025; DeepSeek-AI, 2025)

[2]We will fully open-source our code, data, and LLM generations.

[3]E.g., there might be disagreement in the POS tagging of "I saw her *duck*." as duck can either be a noun or verb.

[4]For example, LLM generation evaluation (Zheng et al., 2023) in emergent applications.

While LLM annotators are widely studied and deployed, there is no evaluation of whether they can capture informative human disagreements.

Such evaluation can be particularly important for LLMs optimized on tasks with single-deterministic answers (e.g., RL with verifiable rewards), which contrasts with the reality that many annotation tasks involve multiple valid perspectives. Presumably, training and evaluation with LLM-annotated data that ignore human disagreement may run counter to efforts toward calibrated and pluralistically aligned AI (Sorensen et al., 2024). In other words: rather than measuring whether LLMs can reproduce the majority opinion, we want to know whether they can reproduce the distribution over human answers.

To address this gap, we evaluate LLMs' ability to predict human disagreement in different NLP annotation tasks, following the recommendations of Meister et al. (2024b) to predict human opinion distributions with LLMs. Specifically, we evaluate various training paradigms: LLMs trained with RLVR or RLHF[5], along with other factors: (1) distribution expression (Tian et al., 2023; Wei et al., 2024); (2) few-shot learning; and (3) scaling effects of LLM size. We evaluate all settings on two dimensions: (1) *variance correlation* (VarCorr, Mostafazadeh Davani et al., 2022), measuring how well the LLM-predicted variance correlates to human annotation variance; and (2) *distributional alignment* (DistAlign, Meister et al., 2024a), directly comparing the distributional divergence of LLM and human labels.

Our comprehensive evaluation spans 12 prompting settings, 10 LLMs (ranging from 8B to 671B), and 5 widely studied datasets. We find that RLVR-style reasoning significantly harms disagreement prediction when human annotation variance is high. Moreover, forcing additional reasoning effort (Muennighoff et al., 2025) does not improve the performance of RLVR LLMs. In contrast, for RLHF LLMs, Chain-of-Thought (CoT, Wei et al., 2023) reasoning significantly improves disagreement prediction. Furthermore, RLVR LLMs are better with a *deterministic* goal (e.g., predicting the majority annotation) than with a *probabilistic* goal (e.g., predicting the proportion of human disagreements). Our findings suggest that using LLM annotators—especially with RLVR LLMs and subjective tasks—requires extra caution, as these models may overlook critical human disagreements. In summary, our contributions are:

1. We extensively evaluate using LLMs to predict annotation disagreement.
2. We reveal limitations of reasoning (RLVR) LLMs in disagreement prediction (§ 6.2).
3. Our evaluation offers insights into distribution expression methods (§ 6.1), reasoning (§ 6.2), the importance of human annotations (§ 6.3), few-shot steering (§ 6.4), and model scale (§ 6.5).

2 Related Work

Annotation Disagreement in NLP. Annotation disagreement has been an important area of study with long history (Wiebe et al., 2004; Ovesdotter Alm, 2011; Basile et al., 2021; Uma et al., 2021; Leonardelli et al., 2023). Various qualitative and quantitative analyses show that the majority of disagreement is caused by other systematic reasons (e.g., ambiguity, context sensitivity etc.) rather than random annotation noise (e.g., carelessness) (Plank et al., 2014; Popović, 2021; Jiang and de Marneffe, 2022; Santy et al., 2023; Zhang et al., 2024).

Prior work in modeling disagreement mainly focuses on datasets with repeated annotations and annotator information (e.g., annotator ID and sociodemographic features), which can be used for annotator modeling (Mostafazadeh Davani et al., 2022; Hu and Collier, 2024; Giorgi et al., 2024; Chochlakis et al., 2024, 2025; Orlikowski et al., 2025). However, emergent tasks (e.g., chatbot preference) often lack human annotations (e.g., UltraFeedback, Cui et al., 2024) due to the cost of human data collection and the need for scalability, making it even harder to obtain disagreements with multiple human annotators. Even when multiple annotations are available (e.g., HelpSteer2, Wang et al., 2025b), annotator information might be missing, making it challenging to model individual annotators' behavior or persona. Therefore, it is important to evaluate LLM annotators' ability to capture disagreement without modeling extensive repeated human labels.

Distribution Prediction with LLM. The extensive training corpus of LLMs may enable them to simulate different opinions and predict distribution in real-world (Grossmann et al., 2023; Ziems et al., 2024), and numerous previous studies use LLMs to predict the distribution of political opinions (Argyle et al., 2023; Durmus et al., 2024; Karanjai et al.,

[5] RLHF refers to LLMs with RL from human feedback (Ouyang et al., 2022) but without test-time scaling on RLVR.

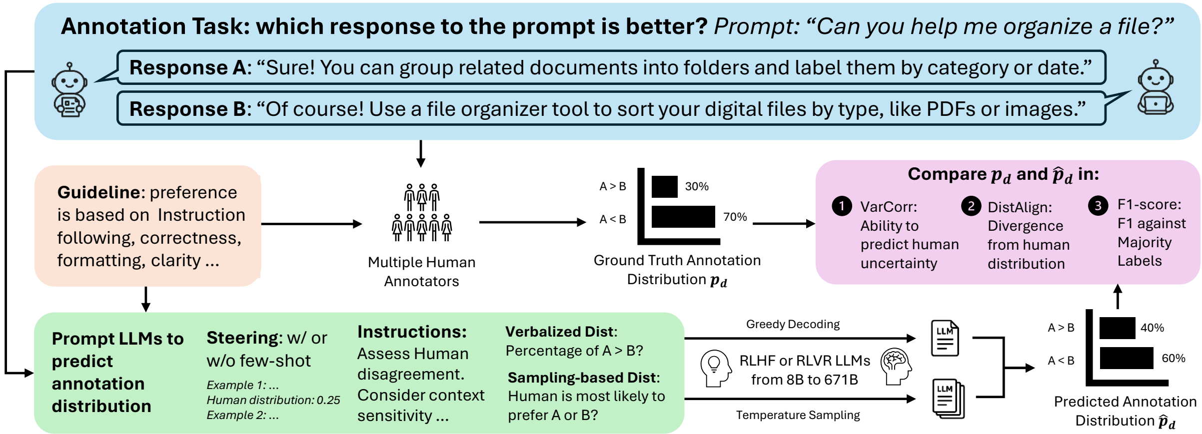

Figure 1: An illustration of our evaluation: We start with a task with guidelines for both human and LLM annotators. The LLM predictions of the annotation distributions are then compared with true human label distribution.

2025; Jiang et al., 2025). Meister et al. (2024b) highlight that the performance of distribution prediction is highly dependent on the target task (e.g., political vs. non-political). Hence, we extend the evaluation of distribution prediction to disagreement in NLP annotation, an interesting yet under-explored area in existing work. We also evaluate the under-studied role of LLM scale and test-time reasoning in distribution prediction.

Automatic Annotation. Despite the prevalence of LLM-automated annotation (Tan et al., 2024), its evaluation ignores disagreement modeling. LLM annotators are evaluated by accuracy (He et al., 2024b; Törnberg, 2023), downstream fine-tuning performance (Lee et al., 2024; Ni et al., 2024, 2025), and agreement with human annotators (He et al., 2024a; Ni et al., 2024). An LLM annotator is validated as reliable if it achieves higher average agreement with human than inter-human agreement (Ni et al., 2024; Calderon et al., 2025). However, this justification ignores the rich information in disagreement between humans. To the best of our knowledge, no prior work has evaluated the LLMs' ability in simulating a group of annotators and predicting the annotation distribution.

3 Problem Formalization

In this section, we formalize the problem of predicting human annotation disagreement and visualize it in Fig. 1. Let $d \in D$ be a datapoint from a dataset D, for which we have a set of n annotations $\mathbf{A_d} = \{a_{d,i} | a_{d,i} \in \{0, 1\}, i \in \{1, 2, ..., n\}\}$ from different human annotators, indicating if d is a positive (1) or negative (0) sample.[6] We assume that the n annotators are representative of the annotator population, so human annotation on d follows a Bernoulli distribution H_d parameterized by:

$$p_d = \frac{|\{a_{d,i} = 1 | a_{d,i} \in \mathbf{A_d}\}|}{n} \quad (1)$$

where p_d denotes the probability that a human annotator labels d positive. The variance of human annotation is $\sigma_d^2 = p_d(1 - p_d)$.

Given human disagreement as the gold label, a machine learning algorithm is tasked with simulating and predicting it. Specifically, through techniques such as fine-tuning, prompting, or sampling, a model can predict a Bernoulli distribution $\hat{H}_d$ regarding how likely a human will annotate d positive, parameterized by $\hat{p}_d$. Then, the variance of the machine-predicted annotation is $\hat{\sigma}_d^2 = \hat{p}_d(1 - \hat{p}_d)$.

To evaluate the model's annotation distribution against humans', we employ two dimensions of evaluation from prior work:

Variance Correlation. In automatic annotation, it is crucial for LLMs to identify samples that are likely to elicit disagreements between human annotators. To evaluate this ability, we adopt the variance correlation metric from Mostafazadeh Davani et al. (2022), which quantifies to what extent higher model uncertainty indicates higher human uncertainty. The formula is:

$$\text{VarCorr} = \text{Corr}\left(\langle \sigma_d^2 \rangle_{d \in D}, \langle \hat{\sigma}_d^2 \rangle_{d \in D}\right) \quad (2)$$

where Corr denotes the Pearson's Correlation (Pearson, 1895).

Distributional Alignment. Although VarCorr captures the alignment of uncertainty, it fails to capture the exact gap between the annotation distributions. For example, if $\langle p_d \rangle_{d \in D} = \langle 0.4, 0.5 \rangle$ and

[6] For simplicity, we study the binary classification problem. Multi-label classification problem with m labels is equivalent to m binary classification problems.

$\langle \hat{p}_d \rangle_{d \in D} = \langle 0.1, 0.2 \rangle$, the model achieves perfect VarCorr but underestimates the human disagreement. Similarly, $\langle p_d, \hat{p}_d \rangle = \langle 0.2, 0.8 \rangle$ shares the same variance, but has contradictory distribution. Therefore, we adopt Distributional Alignment from Meister et al. (2024b), formalized by:

$$\text{DistAlign} = \frac{1}{|D|} \sum_{d \in D} \|p_d - \hat{p}_d\|_1 \qquad (3)$$

which measures the exact difference between two distributions. Importantly, DistAlign cannot fully substitute VarCorr in evaluating uncertainty. For example, given the gold labels of samples $\langle p_1, p_2 \rangle = \langle 0.33, 0.4 \rangle$, model prediction (A) $\langle \hat{p}_1, \hat{p}_2 \rangle = \langle 0.4, 0.33 \rangle$ is better than (B) $\langle \hat{p}_1, \hat{p}_2 \rangle = \langle 0.15, 0.4 \rangle$ in DistAlign. However, (B) has better VarCorr than (A) and correlates better with human uncertainty.

Therefore, both VarCorr and DistAlign are important dimensions to evaluate the prediction of disagreement.

F1 on Majority Label. LLMs (especially with RLVR) are optimized to predict the majority labels. Therefore, we adopt F1-score to study the difference between disagreement prediction and majority label prediction. Specifically, we compute $\text{F1}(\langle \mathbb{1}\{p_d > 0.5\} \rangle_{d \in D}, \langle \mathbb{1}\{\hat{p}_d > 0.5\} \rangle_{d \in D})$ where $\mathbb{1}$ is the indicator function. We drop data points with p_d or $\hat{p}_d$ equal to 0.5 to avoid biased tie-break.

4 Datasets

Hate speech detection (Warner and Hirschberg, 2012; Waseem, 2016) and emotion classification (Hirschberg et al., 2003; Mihalcea and Liu, 2006) are two broadly studied tasks in annotation disagreement. We follow Mostafazadeh Davani et al. (2022) and include Gab Hate Corpus (hereafter GHC; Kennedy et al., 2018) and GoEmotions (Demszky et al., 2020) for our evaluation. GoEmotion is a multi-label classification dataset. We divide it into three binary classification problems—annotating whether a post contains (1) positive / negative / ambiguous emotions, or not (0). GoEmotion Subtasks hereafter referred to as Pos, Neg, and Amb. Furthermore, we include HelpSteer2 (hereafter HS2; Wang et al., 2025b), which consists of multiple annotators' preferences for the helpfulness of chatbot responses. Therefore, our evaluation includes five tasks: hate speech detection, chatbot preference classification, and classifications of positive, negative, and ambiguous emotions.

We further derive two subsets of interest from the dataset of each task: (1) `Random` subset: a randomly sampled subset with 1k data points; and (2) `HighVar` subset: a subset of 200[7] data points where at least two annotators disagree with the majority label, and where the overall proportion of the minority label $(1 - p_d)$ falls between $\frac{1}{3}$ and $\frac{1}{2}$ to ensure high annotation variance. `Random` keeps the original data distribution, containing a lot of samples where human achieves agreement and certain samples where human disagrees. It is useful for evaluating VarCorr—how a model is helpful in predicting human annotation variance. `HighVar` contains samples with potential systematic disagreement (e.g., two annotators disagree with the other three). Therefore, it is useful in evaluating DistAlign—when there exist separate opinions, can a model detect that and predict an aligned distribution? Dataset preparation details can be found in App. A.

Notably, we do not evaluate F1 and VarCorr on `HighVar`, as predicting majority labels or annotation variance is ill-defined when human annotators already exhibit high annotation variance.

5 Methodology

To effectively evaluate LLMs' ability in disagreement prediction, it is important to prompt them correctly. Therefore, we first survey previous work to identify promising distribution prediction methods worth exploring in our evaluation (§ 5.1). Then we describe the implementation details of these methods and relevant baselines (§ 5.2).

5.1 Existing Methods for LLM Distribution Prediction

Distribution Expression Method. Literature in LLM calibration suggests two approaches for LLM to express a distribution: (1) asking for a verbalized probability (Tian et al., 2023); and (2) sampling multiple LLM responses and using the answer frequency as the probability. Tian et al. (2023) show that a verbalized distribution is better, while Wei et al. (2024) draw an opposite conclusion. In distribution prediction, Meister et al. (2024b) finds that verbalized distributions achieve good performance, but sampling-based distributions remain underexplored, especially when combined with reasoning. Therefore, we explore both verbalized and

[7]Size of `HighVar` is determined by the limited number of data points with at least two disagreements. The size of `Random` is determined for budget control.

sampling-based distribution expression methods.

The Effects of Reasoning. Test-time reasoning significantly enhances LLM performance in deterministic reasoning tasks like math and code generation (Wei et al., 2023; DeepSeek-AI, 2025). However, no previous work explores the role of reasoning in probabilistic annotation disagreement. On one hand, reasoning can benefit the prediction of disagreements by giving LLMs the chance to explore and compare different opinions; on the other hand, reasoning may harm decision making, especially when the problem is subjective or has hard-to-articulate criteria (Nordgren and Dijksterhuis, 2009; Liu et al., 2024). In this work, we compare three settings: RLHF LLMs with and without CoT, and RLVR-style reasoning.

In-Context Steering Methods. In-context steering refers to providing LLMs with information about the target group being simulated to help distribution prediction. We investigate the impact of few-shot prompting on predicting annotation disagreement, a method shown effective by previous work (Meister et al., 2024b). Other common steering methods include persona steering (Santurkar et al., 2023) and annotator modeling (Chochlakis et al., 2024, 2025). However, we do not include these methods because (1) for many tasks (e.g., chatbot preference), demographic information might have limited relevance to disagreements, and annotator information might often be unavailable; and (2) piror work has highlighted notable limitations in both prompt-based annotator modeling (Chochlakis et al., 2024, 2025) and persona steering (Meister et al., 2024b; Hu and Collier, 2024).

5.2 Implementation Details

Prompt-Based Methods. We evaluate the combinations of promising settings discussed in the previous section—namely, the combinations of (1) with or without few-shot steering; (2) verbalized or sampling-based distribution; and (3) RLHF LLMs with or without CoT, or using RLVR LLMs instead. Hence, there are $2 \times 2 \times 3 = 12$ settings to be evaluated in total.

To make RLHF and RLVR LLMs comparable, we use DeepSeek-R1 series LLMs (DeepSeek-AI, 2025) (e.g., DeepSeek-R1-Distill-Llama-70B) and corresponding RLHF LLMs sharing the same base LLM (e.g., Llama-3.3-70B-Instruct). To investigate the effect of scaling in LLM size, we experiment LLMs of 8B, 14B, 32B, 70B, and 671B parameters[8].

The prompt structure is illustrated in Fig. 1. For few-shot illustration, We carefully balance the 5 examples—2 of human-agreed positives and negatives correspondingly, and 1 human-disagreed—to avoid introducing spurious bias (Turpin et al., 2023) to distribution prediction. For verbalized probability, we follow Meister et al. (2024b) to directly ask for the proportion of human annotators that may annotate the sample positive. For sampling-based distributions, we ask for the most likely human label and sampling 10 times with a temperature of 0.7 for conventional LLMs, and 0.6 for reasoning LLMs, following the official recommendation.

Furthermore, all prompts present LLMs with the same annotation guidelines as in the original dataset papers, which are likely the guidelines presented to human annotators. This may increase LLMs' chance to capture human disagreement caused by the context or natural ambiguity of annotation guidelines. We also explicitly prompt LLMs to assess potential disagreement and consider context sensitivity (e.g., cultural, social, linguistic ambiguity) that may influence the interpretation. Full prompts and inference hyperparameter / budget are detailed in App. B and App. C respectively.

Fine-tuning Methods. Fine-tuning encoder-only LMs for disagreement prediction is a straightforward way to use human labels (Mostafazadeh Davani et al., 2022; Fleisig et al., 2023). Therefore, we fine-tune ModernBERT-large (Warner et al., 2024) and DeBERTa-V3-large (He et al., 2023) to regress onto the positive annotation probability of human p_d. The loss function is:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|D_{\text{train}}|} \sum_{d \in D_{\text{train}}} (\hat{p}_d - p_d)^2 \qquad (4)$$

where $\hat{p}_d = \text{LM}(d)$ is the prediction of the encoder-only LM; and D_{train} denotes a randomly sampled training set. Fine-tuning baselines require thousands of data points and repeated human labels to capture the target distribution. This is not applicable for most automatic annotation tasks with limited human labels without majority voting aggregation. Fine-tuning details are in App. D.

6 Results

This section presents the evaluation results and takeaways. We start from comparing distribution

[8] We exclude 7B LLMs because their base LLM, Qwen2.5-7B-Math, is specialized for mathematical tasks and therefore unsuitable for the current task.

Setting	Random VarCorr	Random DistAlign	Random F1	HighVar DistAlign
Verbalized > Sampling:				
	95.0%**	92.5%**	28.3%**	98.3%**
RLVR > RLHF:				
	40.0%	62.0%*	36.0%**	18.0%**
RLHF CoT > RLHF w/o CoT :				
	64.0%**	72.0%**	66.0%**	70.0%**
Extend Reasoning Once > Natural Ending :				
	62.50%	65.00%*	47.50%	60.00%
Extend Reasoning Twice > Natural Ending :				
	60.00%	72.50%	50.00%	57.50%
w/ > w/o Few-Shot:				
	45.3%	41.3%**	30.7%**	37.3%*
HS2 w/ > w/o Few-Shot:				
	26.67%**	0.00%**	6.67%**	0.00%**
GHC w/ > w/o Few-Shot:				
	80.00%**	80.00%**	66.67%**	53.33%
GE-Pos w/ > w/o Few-Shot:				
	53.33%	60.00%	33.33%**	66.67%**
GE-Neg w/ > w/o Few-Shot:				
	53.33%	53.33%	26.67%**	53.33%
GE-Amb w/ > w/o Few-Shot:				
	13.33%**	13.33%**	20.00%	13.33%**
Positive > Negative Scaling:				
	73.33%**	70.00%**	86.67%**	56.67%*

Table 1: Win rates of the left settings with Wilcoxon signed-rank tests. We evaluate on the `Random` and `HighVar` subsets. The intensity of green and red indicates how strongly the left setting wins over or loses to the right one. Statistically significant wins or losses are marked with ** ($p < 0.01$) and * ($p < 0.05$).

expression methods—verbalized vs. sampling-based distribution. Then, we investigate the role of steering method and different reasoning paradigms. Due to the large number of experiments, we present aggregated results to convey core messages and present the full model-level performance in App. E.

6.1 Verbalizing or Sampling?

We compare verbalized and sampling-based distributions across 120 controlled experimental settings, varying only the distribution expression method. These settings span 4 LLM sizes (8B, 14B, 32B, and 70B[9]), 3 reasoning paradigms (RLVR, RLHF with and without CoT), 5 datasets, and 2 steering strategies (few-shot or no steering).

The winning rates of the verbalized distribution in different metrics are shown in the first row of Table 1, combined with the results of the Wilcoxon test (Wilcoxon, 1992) to show statistical significance. We observe that the verbalized method significantly outperforms in predicting annotation distribution (VarCorr and DistAlign). However, the sampling-based method is better in predicting the majority label (F1). This indicates that predicting the majority label and disagreement are different tasks that require separate evaluations.

Takeaway: we recommend using verbalized distribution in disagreement prediction, and evaluating LLM annotators on both majority label and disagreement prediction—especially those rely on sampling-based self-consistency to improve majority label prediction (Pangakis et al., 2023b; Ni et al., 2024; Zhou et al., 2025; Wang et al., 2025a).

Given the significantly better performance of verbalized distribution, we focus the analyses in the following sections on results obtained with this method. Sampling-based methods yield better majority label prediction, which lies outside the scope of disagreement prediction. We therefore analyze those results separately in App. F.

6.2 Reasoning in Disagreement Prediction

We compare reasoning methods—(1) RLHF LLMs without reasoning; (2) RLHF LLMs with CoT reasoning; and (3) lengthy reasoning with RLVR LLMs—across 50 controlled settings, varying only the reasoning methods. Controlled settings span 5 LLM sizes (8B, 14B, 32B, 70B, 671B), 5 datasets, and 2 steering strategies (few-shot or no steering).

Results on `Random` and `HighVar` are presented in Table 2 and Table 3 respectively. We aggregate the results of 5 LLM sizes by the average and best scores to enable straightforward comparisons between reasoning methods. Rows 2 and 3 of Table 1 present the comparisons of (1) RLVR vs. RLHF (w/ or w/o CoT); and (2) RLHF w/ vs. w/o CoT across 50 controlled settings.

When comparing RLVR LLMs with their RLHF counterparts, we observe that (1) on `HighVar` where humans strongly disagree with each other, RLVR LLMs achieve significantly worse performance in both aggregated scores in Table 3 and setting-level comparisons summarized in Table 1. (2) On `Random`, results are more mixed but RLVR model does not significantly outperform their RLHF counterparts, as Table 1 row 2 shows. However, the Table 1 row 3 shows that CoT reasoning in RLHF LLMs improves the performance on both `Random` and `HighVar`, compared to without CoT.

To better understand the effect of long reasoning with RLVR LLMs, we force these models to think longer by replacing the end of thinking token “</think>” with “Wait”, which effectively boosts performance for math reasoning (Muen-

[9]We exclude the 671B model due to the high cost of sampling-based prediction.

		HelpSteer2			Gab Hate Corpus			GE-Positive			GE-Negative			GE-Ambiguous		
		VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑
Fine-Tuning-Based Methods																
ModernBERT		0.003	0.269	0.559	0.426	0.141	0.368	0.277	0.187	0.681	0.487	0.180	0.584	0.249	0.198	0.528
DeBERTa-V3		0.020	0.272	0.578	0.554	0.115	0.495	0.336	0.178	0.745	0.530	0.168	0.670	0.289	0.186	0.631
*Verbalized Distribution & **w/o** Few-shot Steering*																
Avg	No-CoT	0.143	0.254	0.718	0.362	0.229	0.294	0.183	0.249	0.607	0.337	0.265	0.561	0.096	0.273	0.440
	CoT	0.177	0.250	0.677	0.363	0.203	0.373	0.192	0.226	0.638	0.329	0.246	0.570	0.116	0.252	0.431
	R1	0.136	0.247	0.705	0.374	0.177	0.394	0.236	0.215	0.633	0.331	0.242	0.556	0.121	0.257	0.395
Best	No-CoT	0.183	0.236	0.741	0.461	0.158	0.376	0.241	0.220	0.721	0.444	0.265	0.583	0.126	0.256	0.547
	CoT	0.230	0.231	0.715	0.399	0.164	0.434	0.233	0.209	0.675	0.389	0.246	0.581	0.183	0.230	0.534
	R1	0.188	0.230	0.722	0.426	0.148	0.463	0.274	0.201	0.674	0.419	0.241	0.596	0.147	0.233	0.463
Verbalized Distribution + Few-shot Steering																
Avg	No-CoT	0.098	0.291	0.683	0.355	0.205	0.372	0.197	0.240	0.573	0.241	0.275	0.526	0.055	0.306	0.450
	CoT	0.139	0.279	0.686	0.380	0.182	0.405	0.200	0.226	0.619	0.321	0.250	0.566	0.098	0.276	0.450
	R1	0.100	0.281	0.608	0.416	0.159	0.393	0.236	0.212	0.589	0.359	0.233	0.538	0.107	0.279	0.333
Best	No-CoT	0.163	0.258	0.710	0.459	0.142	0.553	0.249	0.210	0.658	0.411	0.226	0.576	0.088	0.268	0.534
	CoT	0.182	0.266	0.692	0.436	0.147	0.467	0.243	0.211	0.680	0.409	0.219	0.580	0.135	0.248	0.512
	R1	0.128	0.255	0.678	0.449	0.135	0.447	0.252	0.205	0.675	0.402	0.214	0.593	0.118	0.267	0.437

Table 2: Performance on `Random` (randomly sampled) subsets of all datasets, aggregating 8B–671B results by Average or Best. Color intensity reflects relative performance within each column. RLVR LLMs shows no significant advantage over RLHF LLMs. Fine-tuning outperforms prompting on all datasets except HS2.

		HS2↓	GHC↓	Pos↓	Neg↓	Amb↓
Fine-Tuning-Based Methods						
ModernBERT		0.094	0.246	0.148	0.153	0.138
DeBERTa-V3		0.109	0.256	0.166	0.191	0.153
*Verbalized Distribution & **w/o** Few-shot Steering*						
Avg	No-CoT	0.272	0.233	0.294	0.279	0.223
	CoT	0.202	0.207	0.237	0.217	0.193
	R1	0.240	0.222	0.260	0.261	0.246
Best	No-CoT	0.240	0.182	0.249	0.222	0.165
	CoT	0.180	0.170	0.205	0.173	0.156
	R1	0.206	0.204	0.217	0.239	0.195
Verbalized Distribution + Few-shot Steering						
Avg	No-CoT	0.284	0.236	0.233	0.227	0.233
	CoT	0.279	0.211	0.237	0.234	0.231
	R1	0.286	0.232	0.260	0.260	0.283
Best	No-CoT	0.216	0.188	0.178	0.159	0.204
	CoT	0.254	0.193	0.202	0.193	0.159
	R1	0.251	0.204	0.218	0.228	0.231

Table 3: DistAlign Performance on `HighVar` (high annotation variance) subset of all datasets. RLVR LLMs constantly underperforms RLHF LLMs on both Avg and Best. Fine-tuning outperforms prompting on all datasets except GHC.

nighoff et al., 2025). We force longer reasoning twice, and compare to the results to natural ending. The controlled comparisons span 40 settings—4 LLM sizes[10], 2 steering methods, and 5 datasets. The row 4 and 5 of Table 1 show the results, where forcing longer reasoning rarely leads to statistically significant improvements.

Moreover, RLVR underperforms RLHF on majority label prediction (F1) with verbalized distribution as shown by Table 1. However, when applying sampling-based method, RLVR significantly outperforms RLHF on F1 (win rate 62.5%**). This may be because, in sampling, LLMs are prompted to predict the most likely human label (i.e., majority label), while considering disagreement. This *deterministic* goal is more suitable for RLVR LLMs than the *probabilistic* goal of predicting the proportion of disagreement. However, the sampling-based method still leads to worse distributional prediction as discussed in § 6.1.

Takeaway: CoT reasoning with RLHF LLMs may benefit the prediction of disagreement. However, people should be more cautious about lengthy reasoning with RLVR LLMs, which can significantly harm the performance in probabilistic disagreement prediction.

6.3 Human Labels are Important

To study whether it is necessary to gather repeated human labels for disagreement modeling, we compare small LMs – ModernBERT and DeBERTa-V3 – fine-tuned on large-scale human annotations, to the best LLM results. From Table 2 and Table 3, we observe that fine-tuned small encoder-only LMs outperforms LLMs on GHC `Random`, HS2 `HighVar`, and all GoEmotions subsets, indicating the value of real human annotations in predicting disagreement. However, LLM-based methods are also promising, achieving better performance on HS2 `Random` and GHC `HighVar` without human annotations.

Takeaway: incorporating human labels is highly beneficial for accurate disagreement modeling, while LLM-based methods also demonstrate strong potential due to their cost efficiency and solid performance on certain tasks.

[10]We exclude the 671B DeepSeek-R1 since this model is accessed through API, which does not allow forcing longer reasoning

	HS2 Random			HighVar	GHC Random			HighVar	Pos Random			HighVar	Neg Random			HighVar	Amb Random			HighVar
	VarCorr	DistAlign	F1	DistAlgin	VarCorr	DistAlign	F1	DistAlgin	VarCorr	DistAlign	F1	DistAlgin	VarCorr	DistAlign	F1	DistAlgin	VarCorr	DistAlign	F1	DistAlgin
*Verbalized Distribution but **w/o** Few-shot Steering*																				
No-CoT	0.702	0.703	0.945	-0.037	-0.345	-0.049	0.277	0.722	0.568	0.586	0.825	0.690	-0.402	-0.197	0.539	0.196	0.818	0.224	0.428	-0.046
CoT	0.913	0.738	0.447	-0.097	0.441	0.485	0.799	0.261	0.786	0.593	0.582	0.260	-0.303	-0.280	0.686	-0.096	0.899	0.854	0.329	0.138
R1	0.852	0.790	0.726	-0.668	0.083	-0.400	0.628	0.862	-0.059	0.598	0.470	0.853	-0.700	-0.333	0.306	0.873	0.518	0.934	0.657	0.667
Verbalized Distribution + Few-shot Steering																				
No-CoT	0.906	0.804	0.507	0.399	0.275	0.298	0.240	0.175	0.578	0.593	0.778	-0.289	-0.167	-0.235	0.030	-0.819	0.014	0.023	0.584	0.172
CoT	0.692	0.252	-0.209	-0.230	0.457	0.463	0.587	-0.379	0.503	0.428	0.777	-0.047	-0.170	-0.455	0.299	-0.604	0.504	0.327	0.457	-0.105
R1	0.653	-0.104	-0.811	-0.488	0.151	0.056	0.539	0.671	0.639	0.700	-0.299	0.789	-0.714	-0.570	-0.152	0.792	0.449	0.204	0.862	0.504

Table 4: Correlation of performance and log-number of LLM parameters ($log(8)$ to $log(671)$). Green and red intensity reflects the degree of positive / negative scaling.

6.4 Few-Shot Steering

Meister et al. (2024b) show that LLMs exhibit strong few-shot steerability in distribution prediction. Therefore, we investigate whether few-shot illustrations can steer LLMs for better disagreement prediction. Few-shot is compared to zero-shot prompting across 75 controlled settings—spanning 5 LLM sizes (8B to 671B), 3 reasoning settings, and 5 datasets. Comparisons are summarized in the sixth row of Table 1. Few-shot steering decreases the performance on 4 metrics, with statistically significant drop in 3 of them.

Observing Table 2 and Table 3, we notice that few-shot steering seems to help certain tasks (e.g., GHC Random) but harm others (e.g., HS2). Therefore, we separately evaluate the effect of few-shot steering on each dataset (see the lower half of Table 1 before the last row). The results show that few-shot steering significantly harms disagreement prediction on HS2 and GE-Pos, but improves performance on GHC Random and GE-Neg HighVar.

Takeaway: few-shot steering can be helpful, but its effectiveness varies across tasks and datasets.

We also perform similar per-dataset analyses in earlier sections (e.g., comparing CoT vs. no-CoT), which mostly yield consistent trends with the aggregated results or lacks statistical significance. We thus only include the aggregated results in Table 1 and briefly discuss the per-dataset results in App. G.

6.5 Scaling Effect of LLM Size

Our coverage of LLMs from 8B to 671B allows exploring the scaling effect of LLM size in disagreement prediction. Specifically, we compute the correlation between performance improvement and the increase of log-number of parameters. Table 4 reports the Pearson's coefficients spanning 30 settings—5 datasets, 2 steering methods, and 3 reasoning settings. The comparison across 30 settings are summarized in the last row of Table 1. Scaling LLM size can improve disagreement prediction with statistical significance. However, the improvement is less significant on HighVar while more significant for majority label prediction (F1). Table 4 also shows that different datasets seem to have different scaling effect. Conducting Wilcoxon Test for each dataset, we find that there is statistical significant negative scaling on the disagreement prediction of Neg Random. Other trends are consistent with the results observed across all datasets.

Takeaway: Scaling LLM size may more effectively boost majority label prediction than disagreement prediction. Negative scaling occurs especially in cases of strong disagreement (HighVar subsets) or on specific datasets (e.g., Neg Random).

7 Discussion and Conclusion

LLM annotators are widely used, but their ability to capture informative human disagreement remains under-explored. Addressing this gap, we comprehensively evaluate LLMs in disagreement prediction, covering widely studied tasks, and common settings of LLM usage.

RLHF LLMs exhibit greater potential than RLVR LLMs in predicting disagreements (§ 6.2). This may be because RLVR optimization on verifiable and deterministic answers harms the ability to capture multiple debatable answers. In contrast, reasoning (CoT) with RLHF LLMs improves disagreement prediction, suggesting that the reduced performance of RLVR is not necessarily due to reasoning itself. This may also be related to recent observations that RLVR models can hallucinate more than RLHF models in some tasks (Metz and Weise, 2025).

Moreover, we find that although scaling LLM size and few-shot steering improve disagreement prediction, these methods are not more effective than a data-centric approach—fine-tuning small LLMs with thousands of human data (§ 6.3). Given the scarcity of repeated human labels, future work may explore how to leverage human data more efficiently.

Limitations

This work evaluates LLMs in disagreement prediction and draws observations with statistical significance tests. However, it does not analyze the causes of the observations. For example, what are the exact causes of RLVR worse than RLHF LLMs? Why does few-shot steering work for some datasets but not others? These questions are critical for providing concrete guidelines for real-world practice. As the first work studying disagreement modeling in LLM annotation, we prioritize evaluation **breadth** to include broad potential settings in reasoning, distribution expression, in-context steering, and LLM size. This gives us advantages in (1) addressing promising settings in prior work (§ 5.1); and (2) conducting a statistical significance check thanks to the large number of experiments. However, it also limits us in analysis **depth** and we leave the critical causal analyses of the observations to future work.

Ethics Statement

Data Privacy or Bias. We use publically available datasets (GHC, GoEmotions, and HelpSteer2) which have no data privacy issues or bias against certain demographics. All artifacts we use are under licenses allowing research usage. We also notice no ethical risks associated with this work.

References

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *Preprint*, arXiv:2504.02234.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

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.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *Preprint*, arXiv:2501.10970.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2024. The strong pull of prior knowledge in large language models and its impact on emotion recognition. *Preprint*, arXiv:2403.17125.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2025. Aggregation artifacts in subjective tasks collapse large language models' posteriors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5513–5528, Albuquerque, New Mexico. Association for Computational Linguistics.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

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, Singapore. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Lukas M Fuchs, Yu Fan, and Christian von Scheve. 2021. Value differences between refugees and german citizens: insights from a representative survey. *International Migration*, 59(5):59–81.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188, Miami, Florida, USA. Association for Computational Linguistics.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024a. Annollm: Making large language models to be better crowdsourced annotators. *Preprint*, arXiv:2303.16854.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024b. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Julia Hirschberg, Jackson Liscombe, and Jennifer Venditti. 2003. Experiments in emotional speech. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 1–7.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Shapeng Jiang, Lijia Wei, and Chen Zhang. 2025. Donald trumps in the virtual polls: Simulating and predicting public opinions in surveys using large language models. *Preprint*, arXiv:2411.01582.

Rabimba Karanjai, Boris Shor, Amanda Austin, Ryan Kennedy, Yang Lu, Lei Xu, and Weidong Shi. 2025. Synthesizing public opinions with llms: Role creation, impacts, and the future to edemorcacy. *Preprint*, arXiv:2504.00241.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, and 1 others. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

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, Toronto, Canada. Association for Computational Linguistics.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *Preprint*, arXiv:2410.21333.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024a. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16485–16498, Miami, Florida, USA. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024b. Benchmarking distributional alignment of large language models. *Preprint*, arXiv:2411.05403.

Cade Metz and Karen Weise. 2025. A.i. is getting more powerful, but its hallucinations are getting worse. *The New York Times*. Accessed: 2025-05-10.

Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144.

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.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

Jingwei Ni, Tobias Schimanski, Meihong Lin, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2025. DIRAS: Efficient LLM annotation of document relevance for retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5238–5258, Albuquerque, New Mexico. Association for Computational Linguistics.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1890–1912, Bangkok, Thailand. Association for Computational Linguistics.

Loran F. Nordgren and Ap Dijksterhuis. 2009. The devil is in the deliberation: Thinking too much reduces preference consistency. *Journal of Consumer Research*, 36(1):39–46.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *Preprint*, arXiv:2502.20897.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023a. Automated annotation with generative ai requires validation. *ArXiv*, abs/2306.00176.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023b. Automated annotation with generative ai requires validation. *Preprint*, arXiv:2306.00176.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *ArXiv*, abs/2402.05129.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Preprint*, arXiv:2305.04388.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *Preprint*, arXiv:2304.06588.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. 2025a. Cream: Consistency regularized self-rewarding language models. *Preprint*, arXiv:2410.12735.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2025b. Helpsteer2-preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Frank Wilcoxon. 1992. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY.

Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2024. Diverging preferences: When do annotators disagree and do models know? *Preprint*, arXiv:2410.14632.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Xin Zhou, Yiwen Guo, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Self-consistency of the internal reward models improves self-rewarding language models. *Preprint*, arXiv:2502.08922.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Dataset Preparation

For all datasets, we only use the data points with at least 4 annotators for both training and evaluation to ensure annotation quality. Data points with 3 annotations may have one annotator disagree with the others, and the disagreement might be caused by random annotation error (e.g., a wrong click).

As shown by (Sandri et al., 2023), 2 annotators making random mistake might be 100 times less likely than 1 annotator doing that.

After this filtering, we randomly select 2,000 data points from the 3,330 Gab Hate Corpus samples, 2,000 data points from the 20,014 GoEmotions samples, and 1,250 data points from the 2,467 HelpSteer2 samples as training data; and 1K datapoints for `Random` subsets for testing. The size of training set is strategically picked so that there are enough annotations with high human annotation variance to form the `HighVar` subsets. HelpSteer2 has a smaller training set because it has less datapoints with at least 4 annotations. Therefore, we shrink its training sets' size to ensure the size of evaluation sets.

The distributions of human annotation p_d of each dataset are presented in Fig. 2.

B Prompts

Here we use Gab Hate Corpus prompts to showcase our prompt design. Tasks for other datasets can be found in our Github repository.

One example prompt for Gab Hate Corpus is illustrated bellow, with verbalized distribution and w/o few-shot steering. It is used for RLVR LLMs and for RLHF w/o reasoning, since no reasoning is required in output format. In the instruction part, the model is guided to consider potential disagreement.

GHC Prompt Example

You are a helpful research assistant with expertise in detecting hate speech from social media posts.

Objective:
You are tasked with evaluating social media posts to assess the distribution of labels assigned by annotators regarding whether a post constitutes hate speech. The goal is to predict the percentage of annotators who classify the post as hate speech and those who classify it as not hate speech, based on the specific aspects of hate speech it may exhibit.

A post contains hate speech if it contains any of the following aspects:
- Assaults on Human Dignity: Does the post demean or degrade individuals or groups based on race, ethnicity, gender, religion, sexual orientation, or other protected characteristics?
- Calls for Violence: Does the post incite or encourage physical harm or violence against individuals or groups?
- Vulgarity and/or Offensive Language: Does the post contain profanity, slurs, or other offensive language that may or may not be directed at individuals or groups?

Instruction:
- Review the Post: Carefully read the social media post and consider its content in relation to the three aspects of hate speech listed above.
- Assess Agreement: Human annotators may have high disagreement if the post contains ambiguous language or context, or if it is likely to be interpreted differently based on subjective views about the severity or nature of hate speech. They may have low disagreement if the post clearly falls into one or more of the hate speech categories or is obviously not hate speech.
- Context Sensitivity: Consider the cultural, social, and linguistic context that may influence the interpretation of the post.

Output Format:
[Percentage]: a floating-point number indicating the percentages of annotators who classify the post as hate speech, in a squared bracket, e.g., [0.3]. Omit explanations and only give the percentage.
Here is the post: post

For sampling-based distribution, the objective and output format changes to follows, where the LLM is asked to predict the "most likely" annotation from human.

Sampling-Based Distribution

...
Objective:
You are tasked with evaluating social media posts to determine whether they contain hate speech or not. The goal is to predict the label most likely assigned by human annotators, classifying each post as either containing hate speech or not, based on the specific aspects of hate speech it expresses.

...task_description...

Output Format:
[Prediction]: your prediction of how the majority of human annotators will label the post (0 or 1 in squared brackets). [1] if it contains hate speech; [0] if it contains no hate speech. Omit explanations and only give the percentage.
...

When using RLHF LLMs with CoT, we change the output format requirements to:

RLHF CoT

...
Output Format:
[Reasoning]: your careful reasoning about the content of the post, focusing on the clarity or ambiguity of the post's content concerning aspects of hate speech.
[Prediction]: your prediction of how the majority of human annotators will label the post (0 or 1 in squared brackets). [1] if it contains hate speech; [0] if it contains no hate speech.
...

For few-shot steering, we add the following in-context examples. The few-shot illustrations are carefully picked to avoid biasing the output distribution (Turpin et al., 2023).

GHC Few-Shot Steering

...
Here are some sample posts:

Post: Lots of references to an oligarchy in the shadows who wants to depopulate the planet and sees common folk as insects to be crushed; but who are they? https://www.zerohedge.com/news/2018-08-26/retired-... #Geopolitics #Ebola #Africa #China | Percetage of postive human labels: 0.25

Post: The minority that will outbreed you and make you the minority | Percetage of postive human labels: 0.50

Post: GOTTA REMEMBER...THE LEFT IS INSANELY DESPERATE CAUSE THEY KNOW THEY'RE ON THE PRECIPICE OF GOING TO GITMO FOR TREASON...IT'S ALREADY IN MOTION...SO LET EM SCRAMBLE AND TALK SHIT...THEIR TIME IS LIMITED...LIKE THE DEVIL'S... | Percetage of postive human labels: 1.00

Post: Dozens at Facebook Unite to Challenge Its 2018Intolerant2019 Liberal Culture https://nyti.ms/2NsJ2rJ It would be nice if it was hundreds, but it's a start. | Percetage of postive human labels: 0.00

Post: I am officially done trying to have peaceful conversation with liberals. They are animals. They deserve to be removed from our republic. Either your believe in America or you are a fucking traitor. Get out of America. You are not welcomed by those of us who love our country. | Percetage of postive human labels: 0.75

C Inference Details

LLMs. We use the following LLMs—RLHF LLMs: `Llama-3.1-Tulu-3.1-8B`[11]; `Qwen2.5-14B-Instruct`; `Qwen2.5-32B-Instruct`; `Llama-3.3-70B-Instruct`, and `DeepSeek-V3`. RLVR LLMs: `DeepSeek-R1-Distill-Llama-8B`; `DeepSeek-R1-Distill-Qwen-14B`; `DeepSeek-R1-Distill-Qwen-32B`; `DeepSeek-R1-Distill-Llama-70B`; and `DeepSeek-R1`.

Framework and Hyperparameters. For 8B to 70B LLMs, we rely on a cluster with 4 GH200 GPUs for local inference. We use vLLM for fast inference. For R1-series RLVR LLMs, we use all official recommended settings, including a temperature of 0.6, and always add <think> at the beginning of assistant message. For RLHF LLMs, we use temperature 0 for verbalized distribution and 0.7 for sampling-based distribution. All other hyperparameters are set to default without restriction on generation length. For the 671B LLMs, we use DeepSeek API with recommended settings.

Computational Cost. The majority of inference cost goes to RLVR LLMs. For the RLVR LLMs of 70B, 32B, 14B, and 8B, the inference costs 100, 40, 20, and 10 GPU hours correspondingly, where the majority is spent on sampling-based distribution which requires sampling 10 times. For RLHF LLMs, especially without CoT, the cost is much less. The RLHF LLMs of 70B, 32B, 14B, and 8B cost 40, 20, 10, 10 GPU hours correspondingly with the cost of CoT and no-CoT settings combined. Note that model loading times are not counted into GPU cost. The API cost of DeepSeek-R1 and DeepSeek-V3 costs roughly 40 USD in total.

Packages for Evaluation. Scipy is used to calculate Pearson's Correlations and Wilcoxon Tests.

D Fine-Tuning Details

We use Huggingface to fine-tune and evaluate fine-tuned ModernBERT-large and DeBERTa-V3-large. We use a learning rate of 5e-5, a weight decay of 0.01, a batch size of 128, and a epoch number of 5. All other hyperparameters are set to default.

E Results w/o Aggregation

Here we present the performance of all LLMs with different settings regarding distribution expression, steering, and reasoning, which can be used to calculate all the aggregated results in § 6. Results on `Random` and `HighVar` subsets are presented in Table 5 and Table 6, respectively.

F Majority Label Prediction

In § 6.1, we observe that sampling-based method achieves better majority label prediction (F1) than verbalized distribution. The prediction of majority labels lies outside the scope of this project, so we analyze those observations in this appendix section to fully reveal the potential of sampling-based methods. We draw the following observations with statistical significance.

1. RLVR LLMs outperform RLHF LLMs, with a win rate 62.50**%.

2. RLHF w/ CoT outperforms w/o CoT, with a win rate 62.50**%.

[11] `Llama-3.1-8B-Instruct` from Meta refuse classify hate speeches, so we use Tulu-3.1 which is also based on Llama-3.1-8B

3. Few-shot steering improves the F1 of GHC with a rate of 66.67**%, but decrease the HS2, Pos, and Neg where the win rates are 6.67**%, 33.33**%, and 26.67**% correspondingly.

All other trends on F1 do not have statistical significance.

G Per-Dataset Results

When comparing RLVR with RLHF LLMs on each dataset, the trends are mostly consistent with Table 1 row 2 on `Random` F1 and `HighVar` DistAlign. For `Random` VarCorr and DistAlgin, we further find that following observations with statistical significance: (1) RLVR underperforms RLHF on HS2 `Random`; and (2) RLVR outperforms RLHF on Pos `Random`. The trends in Table 1 summarizes this observation, as RLVR vs. RLHF has more mixed results on distribution prediction of `Random` subsets, compared to `HighVar` subsets.

For CoT vs. w/o CoT on RLHF LLMs, per-dataset comparison shows that on all datasets, CoT either significantly outperforms w/o CoT, or CoT slightly underperforms w/o CoT but without statistical significance.

Furthermore, extending reasoning with RLVR LLMs does not lead to significant change to the performance on all datasets; while verbalized distribution constantly performs significantly better than sampling-based distribution on all datasets.

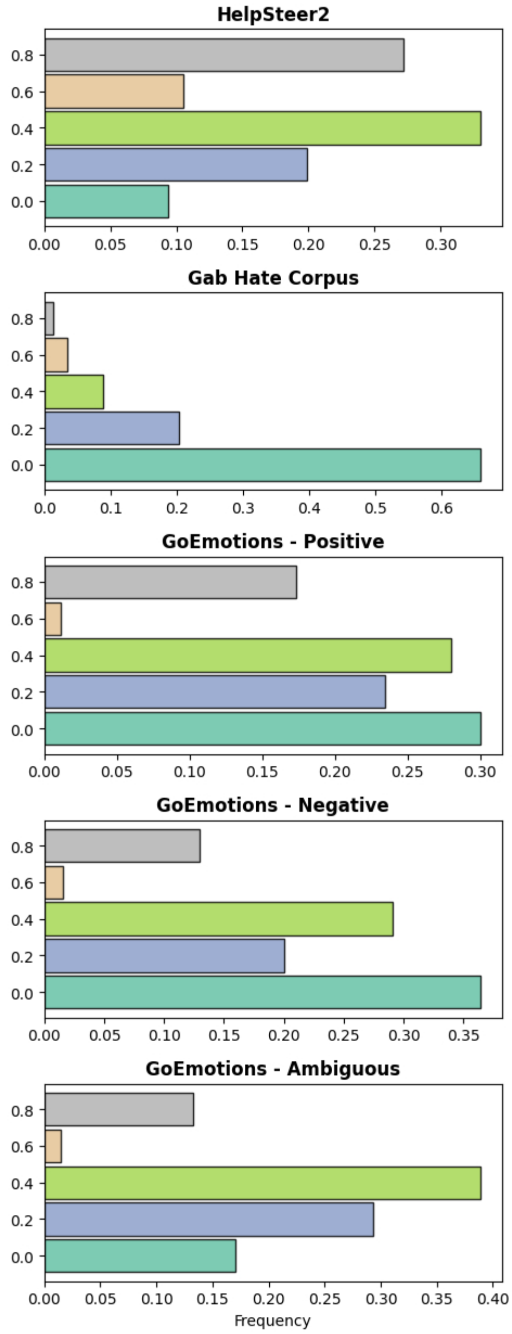

Figure 2: Density bars of the Five Random Sets

		HelpSteer2			Gab Hate Corpus			GE-Positive			GE-Negative			GE-Ambiguous		
		VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑	VarCorr↑	DistAlign↓	F1↑
*Verbalized Distribution & **w/o** Few-shot Steering*																
Llama-8B	No-CoT	0.043	0.277	0.699	0.283	0.290	0.225	0.109	0.357	0.504	0.282	0.294	0.517	0.045	0.309	0.499
	CoT	0.127	0.273	0.699	0.262	0.265	0.270	0.121	0.269	0.631	0.256	0.269	0.566	0.089	0.273	0.514
	R1	0.053	0.281	0.695	0.298	0.194	0.230	0.186	0.240	0.547	0.301	0.273	0.456	0.136	0.268	0.408
Qwen-14B	No-CoT	0.147	0.251	0.713	0.442	0.206	0.294	0.175	0.228	0.637	0.344	0.280	0.558	0.083	0.265	0.392
	CoT	0.132	0.256	0.566	0.399	0.194	0.372	0.194	0.222	0.647	0.374	0.239	0.573	0.068	0.266	0.392
	R1	0.109	0.252	0.675	0.426	0.153	0.400	0.256	0.214	0.670	0.419	0.215	0.596	0.076	0.268	0.339
Qwen-32B	No-CoT	0.172	0.245	0.721	0.461	0.158	0.376	0.195	0.220	0.552	0.444	0.198	0.583	0.102	0.256	0.273
	CoT	0.193	0.234	0.706	0.398	0.164	0.400	0.210	0.214	0.594	0.389	0.216	0.562	0.084	0.257	0.270
	R1	0.151	0.243	0.713	0.425	0.148	0.463	0.262	0.209	0.625	0.398	0.212	0.581	0.123	0.269	0.330
Llama-70B	No-CoT	0.171	0.263	0.717	0.337	0.238	0.274	0.241	0.221	0.620	0.409	0.245	0.579	0.126	0.258	0.487
	CoT	0.205	0.257	0.697	0.376	0.208	0.389	0.202	0.209	0.644	0.379	0.234	0.567	0.155	0.230	0.448
	R1	0.180	0.230	0.722	0.351	0.193	0.428	0.274	0.201	0.674	0.332	0.234	0.595	0.125	0.247	0.436
Deepseek	V3-no-CoT	0.183	0.236	0.741	0.288	0.254	0.302	0.194	0.220	0.721	0.208	0.307	0.568	0.123	0.280	0.547
	V3-CoT	0.230	0.231	0.715	0.381	0.186	0.434	0.233	0.216	0.675	0.246	0.273	0.581	0.183	0.234	0.534
	R1	0.188	0.231	0.721	0.370	0.196	0.447	0.204	0.209	0.649	0.206	0.274	0.552	0.147	0.233	0.463
Verbalized Distribution + Few-shot Steering																
Llama-8B	No-CoT	0.049	0.293	0.658	0.111	0.365	0.147	0.070	0.325	0.409	0.052	0.340	0.450	0.005	0.347	0.489
	CoT	0.067	0.297	0.692	0.215	0.282	0.230	0.142	0.255	0.526	0.197	0.276	0.540	0.123	0.267	0.494
	R1	0.065	0.297	0.676	0.353	0.186	0.258	0.234	0.224	0.546	0.352	0.245	0.456	0.086	0.279	0.290
Qwen-14B	No-CoT	0.086	0.317	0.710	0.459	0.142	0.553	0.207	0.224	0.584	0.371	0.226	0.557	0.079	0.289	0.375
	CoT	0.139	0.267	0.685	0.428	0.147	0.467	0.205	0.226	0.639	0.387	0.224	0.580	0.029	0.296	0.386
	R1	0.114	0.255	0.674	0.442	0.135	0.444	0.216	0.214	0.608	0.402	0.214	0.593	0.105	0.267	0.234
Qwen-32B	No-CoT	0.108	0.290	0.655	0.434	0.145	0.387	0.249	0.210	0.582	0.288	0.241	0.555	0.088	0.268	0.383
	CoT	0.144	0.266	0.680	0.436	0.154	0.397	0.205	0.213	0.591	0.394	0.230	0.567	0.072	0.302	0.368
	R1	0.066	0.298	0.558	0.449	0.149	0.386	0.247	0.205	0.610	0.365	0.223	0.570	0.118	0.306	0.291
Llama-70B	No-CoT	0.083	0.299	0.684	0.431	0.166	0.378	0.229	0.227	0.633	0.411	0.236	0.576	0.083	0.310	0.471
	CoT	0.182	0.297	0.687	0.413	0.164	0.467	0.243	0.211	0.656	0.409	0.219	0.576	0.132	0.248	0.490
	R1	0.127	0.261	0.678	0.433	0.161	0.447	0.231	0.211	0.675	0.352	0.229	0.592	0.118	0.274	0.411
Deepseek	V3-no-CoT	0.163	0.258	0.710	0.343	0.208	0.396	0.229	0.212	0.658	0.085	0.331	0.490	0.028	0.317	0.534
	V3-CoT	0.164	0.271	0.686	0.406	0.164	0.462	0.206	0.226	0.680	0.220	0.300	0.566	0.135	0.268	0.512
	R1	0.128	0.291	0.455	0.403	0.162	0.429	0.252	0.206	0.509	0.322	0.257	0.479	0.107	0.270	0.437
*Sampling-Based Distribution & **w/o** Few-shot Steering*																
Llama-8B	No-CoT	0.021	0.423	0.695	0.357	0.158	0.398	0.002	0.286	0.631	0.097	0.273	0.564	0.027	0.358	0.521
	CoT	0.063	0.440	0.699	0.215	0.207	0.355	0.061	0.289	0.631	0.143	0.308	0.566	0.004	0.374	0.496
	R1	0.121	0.447	0.697	0.149	0.233	0.330	0.169	0.232	0.690	0.089	0.312	0.586	0.099	0.292	0.494
Qwen-14B	No-CoT	0.090	0.361	0.669	0.135	0.203	0.354	0.080	0.271	0.629	0.047	0.332	0.567	0.031	0.382	0.426
	CoT	0.070	0.318	0.688	0.202	0.210	0.350	0.098	0.267	0.649	0.083	0.324	0.593	0.043	0.361	0.495
	R1	0.124	0.282	0.705	0.287	0.165	0.406	0.145	0.250	0.686	0.234	0.281	0.595	0.050	0.306	0.469
Qwen-32B	No-CoT	0.091	0.348	0.702	0.142	0.187	0.376	0.092	0.264	0.623	0.124	0.297	0.590	0.042	0.366	0.402
	CoT	0.118	0.287	0.702	0.280	0.165	0.430	0.157	0.251	0.627	0.208	0.290	0.589	0.025	0.349	0.458
	R1	0.073	0.294	0.759	0.244	0.169	0.414	0.184	0.233	0.685	0.192	0.285	0.607	0.071	0.301	0.442
Llama-70B	No-CoT	0.024	0.412	0.673	0.074	0.263	0.298	0.006	0.291	0.644	0.043	0.367	0.565	0.014	0.393	0.513
	CoT	0.124	0.357	0.693	0.146	0.216	0.337	0.046	0.289	0.649	0.053	0.361	0.560	0.030	0.355	0.516
	R1	0.091	0.278	0.751	0.175	0.208	0.344	0.158	0.240	0.699	0.112	0.313	0.591	0.063	0.315	0.484
Sampling-Based Distribution + Few-shot Steering																
Llama-8B	No-CoT	0.003	0.414	0.698	0.004	0.313	0.257	0.064	0.373	0.563	0.097	0.386	0.522	0.067	0.476	0.504
	CoT	0.006	0.440	0.697	0.150	0.237	0.332	0.070	0.275	0.646	0.098	0.326	0.565	0.088	0.299	0.313
	R1	0.022	0.445	0.699	0.114	0.236	0.339	0.182	0.227	0.689	0.181	0.275	0.607	0.060	0.290	0.483
Qwen-14B	No-CoT	0.084	0.357	0.685	0.151	0.208	0.348	0.087	0.298	0.634	0.087	0.320	0.570	0.084	0.417	0.504
	CoT	0.062	0.316	0.697	0.266	0.175	0.394	0.121	0.282	0.646	0.139	0.324	0.579	0.037	0.333	0.222
	R1	0.121	0.290	0.692	0.322	0.158	0.389	0.137	0.257	0.673	0.209	0.281	0.601	0.068	0.310	0.488
Qwen-32B	No-CoT	0.101	0.381	0.687	0.142	0.183	0.375	0.111	0.263	0.646	0.111	0.301	0.585	0.034	0.372	0.493
	CoT	0.130	0.281	0.709	0.272	0.166	0.416	0.120	0.253	0.661	0.111	0.320	0.564	0.051	0.330	0.358
	R1	0.019	0.308	0.743	0.246	0.164	0.419	0.174	0.237	0.701	0.161	0.290	0.604	0.084	0.299	0.473
Llama-70B	No-CoT	0.025	0.433	0.703	0.018	0.231	0.335	0.090	0.300	0.646	0.120	0.326	0.593	0.023	0.438	0.505
	CoT	0.077	0.322	0.715	0.158	0.192	0.391	0.022	0.303	0.644	0.098	0.323	0.590	0.100	0.329	0.389
	R1	0.063	0.288	0.749	0.234	0.184	0.388	0.148	0.247	0.687	0.197	0.299	0.592	0.069	0.320	0.475

Table 5: Performance on `Random` (randomly sampled) subsets of all datasets.

		HS2↓	GHC↓	Pos↓	Neg↓	Amb↓
*Verbalized Distribution & **w/o** Few-shot Steering*						
Llama-8B	No-CoT	0.182	0.317	0.284	0.296	0.165
	CoT	0.178	0.222	0.205	0.229	0.156
	R1	0.204	0.280	0.263	0.291	0.232
Qwen-14B	No-CoT	0.236	0.293	0.328	0.318	0.258
	CoT	0.230	0.200	0.295	0.239	0.235
	R1	0.216	0.235	0.284	0.262	0.283
Qwen-32B	No-CoT	0.253	0.240	0.303	0.222	0.261
	CoT	0.242	0.199	0.252	0.173	0.226
	R1	0.227	0.242	0.281	0.257	0.284
Llama-70B	No-CoT	0.294	0.262	0.307	0.277	0.225
	CoT	0.170	0.180	0.210	0.207	0.165
	R1	0.235	0.236	0.257	0.255	0.235
Deepseek	V3-no-CoT	0.199	0.248	0.249	0.282	0.210
	V3-CoT	0.217	0.207	0.223	0.237	0.184
	R1	0.227	0.206	0.217	0.239	0.195
Verbalized Distribution + Few-shot Steering						
Llama-8B	No-CoT	0.225	0.274	0.178	0.188	0.204
	CoT	0.254	0.226	0.222	0.232	0.159
	R1	0.255	0.234	0.263	0.276	0.276
Qwen-14B	No-CoT	0.357	0.188	0.231	0.213	0.245
	CoT	0.289	0.193	0.271	0.240	0.278
	R1	0.251	0.236	0.270	0.255	0.286
Qwen-32B	No-CoT	0.317	0.232	0.240	0.159	0.259
	CoT	0.307	0.203	0.239	0.193	0.305
	R1	0.341	0.239	0.278	0.270	0.360
Llama-70B	No-CoT	0.306	0.266	0.296	0.269	0.246
	CoT	0.256	0.209	0.202	0.196	0.173
	R1	0.273	0.249	0.272	0.271	0.262
Deepseek	V3-no-CoT	0.216	0.218	0.219	0.305	0.210
	V3-CoT	0.288	0.226	0.251	0.309	0.241
	R1	0.308	0.204	0.218	0.228	0.231
*Sampling-Based Distribution & **w/o** Few-shot Steering*						
Llama-8B	No-CoT	0.408	0.333	0.274	0.339	0.240
	CoT	0.440	0.365	0.341	0.381	0.315
	R1	0.461	0.386	0.334	0.405	0.274
Qwen-14B	No-CoT	0.433	0.476	0.451	0.492	0.447
	CoT	0.298	0.402	0.397	0.437	0.354
	R1	0.293	0.389	0.381	0.415	0.338
Qwen-32B	No-CoT	0.429	0.469	0.449	0.474	0.442
	CoT	0.327	0.417	0.400	0.427	0.372
	R1	0.349	0.398	0.375	0.422	0.336
Llama-70B	No-CoT	0.467	0.478	0.446	0.495	0.451
	CoT	0.338	0.430	0.400	0.469	0.379
	R1	0.316	0.434	0.379	0.443	0.353
Sampling-Based Distribution + Few-shot Steering						
Llama-8B	No-CoT	0.380	0.393	0.353	0.389	0.384
	CoT	0.435	0.383	0.342	0.392	0.259
	R1	0.448	0.391	0.349	0.381	0.286
Qwen-14B	No-CoT	0.415	0.456	0.447	0.483	0.453
	CoT	0.297	0.403	0.403	0.436	0.398
	R1	0.321	0.381	0.384	0.415	0.327
Qwen-32B	No-CoT	0.430	0.465	0.443	0.469	0.451
	CoT	0.330	0.419	0.389	0.420	0.379
	R1	0.356	0.400	0.370	0.421	0.332
Llama-70B	No-CoT	0.457	0.481	0.461	0.482	0.481
	CoT	0.333	0.434	0.427	0.449	0.385
	R1	0.323	0.425	0.385	0.422	0.363

Table 6: DistAlign Performance on `HighVar` (high annotation variance) subset of all datasets.