Can Reasoning Help Large Language Models Capture Human Annotator Disagreement?

Jingwei Ni[E Z*] Yu Fan[E*] Vilém Zouhar[E] Donya Rooein[B E]
Alexander Hoyle[E] Mrinmaya Sachan[E] Markus Leippold[Z]
Dirk Hovy[B] Elliott Ash[E]
[E]ETH Zürich [Z]University of Zürich [B]Bocconi University
{jingni, yufan, ashe}@ethz.ch

Abstract

Variation in human annotation (i.e., disagreements) is common in NLP, often reflecting important information like task subjectivity and sample ambiguity. Modeling this variation is important for applications that are sensitive to such information. Although RLVR-style reasoning (Reinforcement Learning with Verifiable Rewards) has improved Large Language Model (LLM) performance on many tasks, it remains unclear whether such reasoning enables LLMs to capture informative variation in human annotation. In this work, we evaluate the influence of different reasoning settings on LLM disagreement modeling. We systematically evaluate each reasoning setting across model sizes, distribution expression methods, and steering methods, resulting in 60 experimental setups across 3 tasks. Surprisingly, our results show that RLVR-style reasoning degrades performance in disagreement modeling, while naive Chain-of-Thought (CoT) reasoning improves the performance of RLHF LLMs (RL from human feedback). These findings underscore the potential risk of replacing human annotators with reasoning LLMs, especially when disagreements are important.[1]

1 Introduction

Inter-annotator disagreement is common in NLP annotations (Snow et al., 2008) and often treated as noise to be removed by majority voting (Sabou et al., 2014) or expert aggregation (Hovy et al., 2013). However, these solutions may be misguided, as annotation disagreement can signal a diversity of views and often contains valuable information that enables downstream applications to capture a diversity of human values and interpretations (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 (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).

With the rapid growth of LLMs' capability, evaluating LLMs' ability to capture annotation disagreement is becoming increasingly important. On one hand, more "capable" LLMs achieve better performance in predicting the majority-voted label, and are thus widely adopted to replace human decision-making in applications such as text classification (Pangakis et al., 2023a; Törnberg, 2024; He et al., 2024), chatbot preference annotation (Lee et al., 2024), and LLM-as-a-judge (Calderon et al., 2025; Fan et al., 2025). On the other hand, many of these applications also require understanding the full spectrum of annotator disagreement. However, evaluations typically focus on majority-label prediction, overlooking the modeling of underlying disagreement distributions. As a result, it remains unclear whether the LLMs can reliably automate these applications, by effectively flagging cases with potential annotator disagreement for human oversight.

Prior work evaluates early LLMs and identifies their limitations in modeling annotation disagreement under specific settings (Lee et al., 2023), but have largely overlooked several key factors influencing distribution modeling, such as (1) in-

[*] Equal contributions.

[1] Code and data at https://github.com/EdisonNi-hku/Disagreement_Prediction.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 36–54
March 24-29, 2026 ©2026 Association for Computational Linguistics

context steering methods (e.g., few-shot learning); and (2) distribution expression methods (Meister et al., 2024b). More importantly, the role of reasoning—which significantly enhances LLM performance in various tasks (Wei et al., 2023; DeepSeek-AI, 2025)—is underexplored in prior work (Lee et al., 2023; Chen et al., 2024). Presumably, reasoning can benefit disagreement modeling by enabling LLMs to explore and compare different opinions through CoT. However, reasoning may harm decision making when the problem has hard-to-articulate criteria (Nordgren and Dijksterhuis, 2009; Liu et al., 2024). This may be particularly relevant to RLVR LLMs, which are optimized on tasks with single-deterministic answers—contrasting with the reality that many tasks involve multiple valid perspectives.

To address these gaps, we conduct a comprehensive evaluation of LLMs under different reasoning settings: RLHF LLM with and without CoT, as well as RLVR LLM. Given that the impact of reasoning may be further influenced by other factors such as LLM size, distribution expression, and steering method (Meister et al., 2024b), our evaluation systematically explores the full combinations of (1) 3 reasoning settings; (2) 5 LLM sizes (from 8B to 671B); (3) with or without few-shot steering; and (4) 2 distribution expression methods (Tian et al., 2023; Wei et al., 2024), resulting in 60 prompting settings. We evaluate all settings on 5 datasets of 3 widely studied tasks, following the metrics in prior work: (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.

Surprisingly, we find that RLVR-style reasoning significantly harms disagreement modeling 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, CoT prompting significantly improves disagreement modeling. 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 RLVR-optimized LLMs in disagreement-matter tasks requires extra caution, as these models may overlook critical human disagreements. In summary, our contributions are:

1. We systematically evaluate RLVR and RLHF LLMs in disagreement modeling across 3 tasks, 5 LLM sizes, and 12 prompting settings.
2. We quantitatively reveal the limitations of RLVR-style reasoning in modeling disagreement (§ 6.2), and provide qualitative insights to explain these findings (§ 6.7).
3. Our evaluation further examines the impact offers other relevant factors on disagreement modeling, including distribution expression methods (§ 6.1), the importance of human annotations (§ 6.3), few-shot steering (§ 6.4), and model scale (§ 6.5).

2 Background and Related Work

RLHF and RLVR are two dominant paradigms for LLM alignment. RLHF fine-tunes models using reward models trained on human preference data, optimized via reinforcement learning algorithms like PPO (Ouyang et al., 2022). RLVR instead derives rewards from automatically verifiable properties—such as code execution correctness, passing unit tests, or satisfying mathematical constraints (DeepSeek-AI, 2025). Intuitively, RLHF prioritizes subjective human preference, while RLVR emphasizes objective problem-solving verification.

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 has fruitfully leveraged datasets with annotator metadata (e.g., annotator ID, explanations, and sociodemographic features), enabling annotator modeling and deeper insights into sources of variation (Mostafazadeh Davani et al., 2022; Hu and Collier, 2024; Giorgi et al., 2024; Chen et al., 2024; Chochlakis et al., 2025; Orlikowski et al., 2025). Our evaluation is complementary: we focus on settings where annotator metadata are absent – an increasingly common scenario in large-scale or emergent tasks (e.g., LLMs as judges or annotators, Cui et al., 2024), and assess how well models

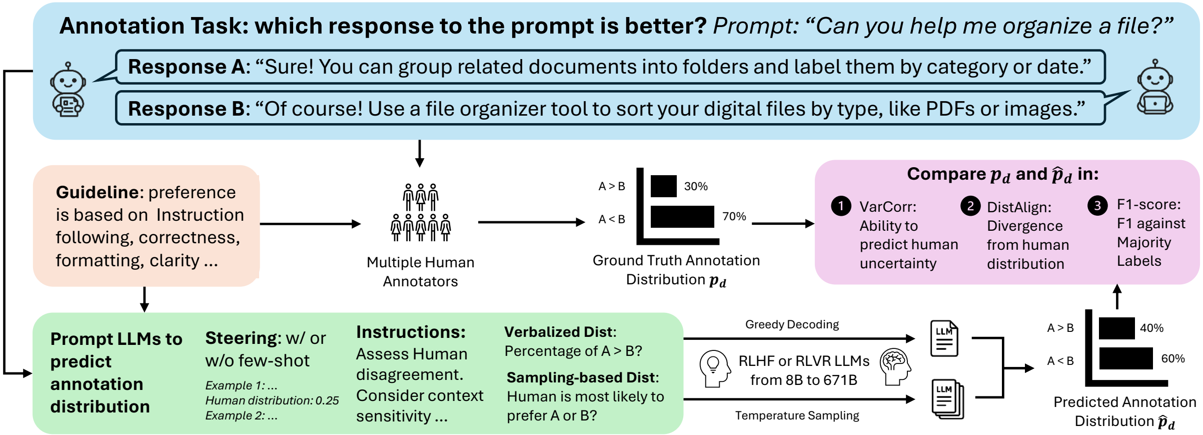
Annotation Task: which response to the prompt is better? Prompt: “Can you help me organize a file?”
Response A: “Sure! You can group related documents into folders and label them by category or date.”
Response B: “Of course! Use a file organizer tool to sort your digital files by type, like PDFs or images.”
Guideline: preference is based on Instruction following, correctness, formatting, clarity ...
Multiple Human Annotators
A > B
A < B
30%
70%
Ground Truth Annotation Distribution p_d
Compare p_d and $\hat{p}_d$ in:
1 VarCorr: Ability to predict human uncertainty
2 DistAlign: Divergence from human distribution
3 F1-score: F1 against Majority Labels
Prompt LLMs to predict annotation distribution
Steering: w/ or w/o few-shot
Example 1: ...
Human distribution: 0.25
Example 2: ...
Instructions: Assess Human disagreement. Consider context sensitivity ...
Verbalized Dist: Percentage of A > B?
Sampling-based Dist: Human is most likely to prefer A or B?
Greedy Decoding
RLHF or RLVR LLMs from 8B to 671B
Temperature Sampling
LLM
LLM
A > B
A < B
40%
60%
Predicted Annotation Distribution $\hat{p}_d$

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.

can capture disagreement in such constrained but realistic contexts.

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; Meister et al., 2024b; Karanjai et al., 2025). The closest prior work to ours is Lee et al. (2023), which reveals LLMs' limited performance on disagreement modeling for Natural Language Inference (NLI). Specifically, they prompt LLMs to predict NLI labels and probe the annotation distribution with the log probabilities of LLM outputs and Monti-Carlo Sampling outcome. However, their evaluation does not fully address several key aspects: (1) Distribution expression methods based on token-level probability and sampling are shown to be ineffective by Meister et al. (2024b) (and our results in § 6.1); (2) Lee et al. (2023) prompt LLMs answer without explicitly instructing LLMs to consider potential disagreement or controversy. Such task-instruction mismatch may hinder LLMs' ability in disagreement modeling; and (3) their study does not investigate the role of reasoning, which can be crucial for LLMs to explore various aspects of disagreements. It also does not consider other factors such as few-shot steering. To address these gaps, we investigate the impact of reasoning with detailed instruction for disagreement modeling, while also examine the influence of distribution expression methods, few-shot steering, and LLM size.

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.[2] 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:

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

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

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

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 $\langle\hat{p}_d\rangle_{d\in D} = \langle 0.1, 0.2\rangle$, 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:

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

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 modeling 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 datasets: 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[3] 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 § 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.

Low Annotation Noise. Annotators' carelessness may lead to divergent labels, instead of systematic disagreements. To reduce such noise, we keep data points with more than 3 annotations for evaluated subsets. For the `HighVar` subsets, there should be at least two annotators disagree with the majority, where the disagreement is less likely due to annotation noise (Sandri et al., 2023). Results in § 6.3 also suggest that our evaluation datasets contain predictable systematic disagreement.

[3] 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.

5 Methodology

We first motivate our evaluation design in § 5.1. Then we describe the implementation and prompt details in § 5.2.

5.1 Evaluation Motivations and Design

Worth Exploring Factors in Distribution Prediction. We start by identifying factors that may affect disagreement modeling, but was not addressed in prior work (Lee et al., 2023; Chen et al., 2024). (1) **Distribution Expression Methods**: we can probe prediction distribution from LLMs by either directly asking for a verbalized probability, or by sampling multiple LLM responses and using the answer frequency as the probability (see math formulas for verbalized and sampling-based distribution in § B). Some previous work find the former more effective (Tian et al., 2023; Meister et al., 2024b) while others have contradictory observations (Wei et al., 2024). (2) **In-Context Steering**: In-context steering methods provide LLMs with specific target group information to enhance distribution prediction. Meister et al. (2024b) find few-shot steering enhances opinion simulation, but its role in disagreement modeling remains underexplored.

Evaluate Combinations of Different Factors. Factors like distribution expression, steering, and LLM size can impact both reasoning and disagreement modeling. To estimate the causal effect of reasoning on disagreement modeling, it is necessary to evaluate all combinations of these factors (i.e., potential confounders) with different reasoning settings. Otherwise, for example, an observed effect of reasoning under a sampling-based distribution method (e.g., Lee et al., 2023) may not generalize to verbalized distribution methods. See § D for detailed causality theories that motivate our design.

5.2 Implementation Details

Prompt-Based Methods. We evaluate three reasoning settings (RLHF LLMs w/ or w/o CoT, or using RLVR LLMs instead) across the combinations of promising settings discussed in the previous section—namely, (1) with or without few-shot steering; (2) verbalized or sampling-based distribution. Hence, there are $3 \times 2 \times 2 = 12$ 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[4].

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 § C and § E respectively.

Fine-tuning Methods. Fine-tuning encoder-only LMs for disagreement modeling 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 § F.

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

	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 :				
	62.5	65.0*	47.5	60.0
Extend Reasoning Twice > Natural Ending :				
	60.0	72.5	50.0	57.5
w/ > w/o Few-Shot:				
	45.3	41.3**	30.7**	37.3*
HS2 w/ > w/o Few-Shot:				
	26.7**	0.0**	6.7**	0.0**
GHC w/ > w/o Few-Shot:				
	80.0**	80.0**	66.7**	53.3
GE-Pos w/ > w/o Few-Shot:				
	53.3	60.0	33.3**	66.7**
GE-Neg w/ > w/o Few-Shot:				
	53.3	53.3	26.7**	53.3
GE-Amb w/ > w/o Few-Shot:				
	13.3**	13.3**	20.0	13.3**
Positive > Negative Scaling:				
	73.3**	70.0**	86.7**	56.7*

Table 1: Win rates (in %) 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 * ($p < 0.05$).

6 Results

This section presents the evaluation results and takeaways. We start from comparing distribution expression methods—verbalized vs. sampling-based distribution. Then, we investigate the role of reasoning settings and other factors. Due to the large number of experiments, we present aggregated results to convey core messages and present the full model-level performance in § G.

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[5]), 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 evaluating LLM disagreement modeling with verbalized distribution, instead of sampling-based approach in prior work (Lee et al., 2023). LLM annotators relying on sampling-based self-consistency to improve majority label prediction may need extra caution, as the sampling-based approach may overlook disagreements (e.g. 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 modeling. We therefore analyze those results separately in § H.

6.2 Reasoning for Disagreement Modeling

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

[5] 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.

		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.

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 (Muennighoff 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[6], 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 modeling.

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

[6] 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.

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.

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 modeling. 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 modeling 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 reasoning settings), which mostly yield consistent trends with the aggregated results. We thus only include the aggregated results in Table 1 and briefly discuss the per-dataset results in § I.

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 modeling. 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 modeling 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 modeling 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 modeling. Negative scaling occurs especially in cases of strong disagreement (HighVar subsets) or on specific datasets (e.g., Neg Random).

6.6 Impact of LLM Size and Steering Method on Reasoning

Will reasoning's effect on disagreement modeling change with different LLM sizes or steering methods? To investigate this, we compare reasoning settings within subsets of conditions where either the steering method or the LLM size is held fixed. Specifically, we evaluate reasoning effects in: (1) all settings with few-shot steering, (2) all settings without few-shot steering, and (3) all settings using specific LLM sizes (e.g., all settings with 8B LLM). Across these subsets, there are no statistically significant observations that contradict those in § 6.2. Thus, the effect of reasoning remains consistent regardless of the steering method or LLM size.

6.7 Qualitative Analysis

To understand why RLVR LLMs perform worse than their RLHF counterparts, we conduct a qualitative analysis on GHC and GoEmotions. Specifically, we sample 20 data points from the HighVar subset, and other 20 from Random with low disagreement, focusing on cases where DeepSeek-R1 and V3 have divergent predictions. We find that **RLVR and RLHF LLMs have different focus of instruction following** although they are prompted exactly the same—In 85% of cases, RLVR LLMs focus on the annotation guideline, assuming humans would objectively follow the guideline in the same way; while RLHF LLMs focus on considering people with diversified background. One potential reason is that RLVR LLMs are optimized on objective math and coding tasks, thus focusing more on the objective / less controversial parts of prompts. More details and examples in § J.

7 Conclusion and Discussion

We evaluate the impact of reasoning on LLM disagreement modeling, with systematic controls of distribution expression, steering, and LLM size. Results show that it requires extra caution to apply RLVR-style reasoning to tasks where annotator disagreements are prevalent and important.

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 modeling, 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).

Interestingly, Yoon et al. (2025) find that RLVR-style reasoning benefits LLMs in calibrating the confidence of their own answers, which seems to contradict our findings at first glance. However, our evaluation suite focuses on predicting human disagreement instead of the models' confidence / uncertainty based on its internal knowledge. The seemingly contradictory results from our work and Yoon et al. (2025) reflect that calibration and disagreement modeling are orthogonal abilities, while both are essential for responsible decision making. For example, there is one data point where 40% of human disagree with the majority label (60%). If a model predicts the majority label with 100% confidence, it achieves zero calibration error. However, if the confidence score is directly interpreted as a disagreement modeling, it fails to capture any critical disagreement.

Moreover, we find that although scaling LLM size and few-shot steering improve disagreement modeling, 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 the impact of LLM reasoning on disagreement modeling and draws observations with statistical significance tests. Through qualitative analyses, we find that RLVR LLMs tend to assume that all annotators would process the annotation guideline in the same objective way, while RLHF LLM tend to consider annotators' diverse background, although they are prompted with both instructions. However, we fail to draw significant qualitative observations to explain other observations in the paper. For example, why does few-shot steering work for some tasks but not others? Why does scaling in LLM size increase some tasks but not others? These questions are critical to providing concrete guidelines for real-world practice of disagreement modeling. Given our focus on reasoning and the complexity of these question, we leave them for future exploration.

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.

Reproducibility. We fully open source our code, prompts, processed datasets, LLM generations, and instructions to reproduce results in https://github.com/EdisonNi-hku/Disagreement_Prediction.

References

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.

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.

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.

Yu Fan, Jingwei Ni, Jakob Merane, Etienne Salimbeni, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Florian Geering, Oliver Dreyer, and 1 others. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.

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.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. 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.

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.

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.

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, Singapore. Association for Computational Linguistics.

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, 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.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press.

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.

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.

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.

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.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. 2025. Reasoning models better express their confidence. *Preprint*, arXiv:2505.14489.

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.

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 Distribution Formulas

We probe prediction distributions from large language models (LLMs) using two approaches.

Sampling-based Distribution. We draw N responses $\{y^{(1)}, \ldots, y^{(N)}\}$ for the same prompt p_s and estimate the predictive distribution as

$$\hat{p}_{\text{sample}}(y \mid p_s) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[y^{(i)} = y\right]. \tag{5}$$

This estimator approximates the model's implicit predictive distribution via empirical answer frequencies. The sampling prompt p_s asks the model to produce the most likely human annotation.

Verbalized Distribution. Alternatively, we directly prompt the LLM to report probabilities over the label space $\mathcal{Y}$, yielding

$$\hat{p}_{\text{verbal}}(y \mid p_v) \in \Delta^{|\mathcal{Y}|-1}, \tag{6}$$

where $\Delta^{|\mathcal{Y}|-1}$ denotes the $(|\mathcal{Y}| - 1)$-dimensional probability simplex. This formulation reflects the model's self-reported uncertainty. The verbalization prompt p_v asks the model to estimate the proportion of humans who would predict label y.

C 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 2̌018Intolerant2̌019 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

D Causal Motivation of Our Evaluation Design

To estimate the causal effect of reasoning (R) on disagreement modeling (Y), it is crucial to account for other experimental factors—such as distribution expression (X_1), steering method (X_2), and LLM size (X_3)—that may influence both R and Y. These act as potential confounders.

Causal Structure. The underlying causal graph can be represented as:

$$X_1, X_2, X_3 \to R \to Y, \quad X_1, X_2, X_3 \to Y$$

where arrows from X_i to R and Y indicate confounding.

Backdoor Adjustment. To identify the causal effect of R on Y, we must block backdoor paths via all X_i. This motivates evaluating all combinations so that comparisons between reasoning settings are not confounded by X_i.

Estimand. The average causal effect (ACE) of reasoning setting R (vs. another reasoning setting R'):

$$\begin{aligned}\text{ACE} = \ \mathbb{E}_{x_1,x_2,x_3}\big[\,&Y(r, x_1, x_2, x_3) \\ &- Y(r', x_1, x_2, x_3)\big]\end{aligned}$$

which requires averaging over all settings of X_1, X_2, X_3.

Conclusion. By systematically evaluating all factor combinations, we obtain unbiased estimates of the causal effect of reasoning, as detailed by standard causal inference theory (Pearl, 2009).

E Inference Details

LLMs. We use the following LLMs—RLHF LLMs: `Llama-3.1-Tulu-3.1-8B`[7]; `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.

F 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.

G 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.

H 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 $62.50^{**}\%$.
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.

I 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

[7] `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

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.

J Qualitative Analysis Details

We provide here qualitative analyses of our major quantitative findings on RLVR vs. RLHF models with CoT reasoning.

We sampled 40 datapoints from the GHC dataset—20 with high human disagreement (HighVar) and 20 with low disagreement (distribution of 0 or 1)—focusing on cases where DeepSeek-R1 and V3 made the most divergent predictions. The RLVR LLM tends to adhere more closely to the annotation guidelines. For example, if there is no explicit use of slurs, profanity, or offensive language, it typically assumes that all annotators would agree on a score of 0. Here is one example data point: *Is it not a glorious sight to see the left freak out over the supreme justice retirement announcement? I ask you Morton Kondracke!*

By contrast, the RLHF LLM extrapolates beyond the guidelines and is more likely to consider perspectives from individuals with diverse backgrounds. For instance, R1 does not consider references to “the left” as targeting a protected group, whereas V3 argues that people identifying with this group may find such language offensive. This is notable, given that both models are explicitly prompted to consider diversity of opinion and background. Here is one example data point: *BREAKING! TECH COMPANIES MEET IN SECRET TO STEAL 2018 ELECTION Left desperate to win midterms.*

When V3 and R1 disagree, it is more likely (87.5% of analyzed cases) that V3’s prediction exhibits higher variance than R1’s. This helps explain why RLHF outperforms on HighVar cases, while RLVR performs better on the Random samples, which contains more items with low annotation variance.

In addition, when posts include exclamation marks or words in all capital letters, RLHF models are more likely to assign higher hate speech scores. RLVR models also assign higher scores than human annotators in these cases, but their scores remain lower than those of RLHF models. For instance, a post with the text “*DEFEND EUROPA*” is assigned hate speech score of 0.3 by an R1 model and 0.6 by the its RLHF counterpart with CoT, where human label is 0. Moreover, hate speech that does not explicitly use vulgar language is less likely to receive high scores from RLHF models (e.g. “*So I get accused of being an Islamophobe by a Muzlime on Twitter, I agree with him that I am an Islamofauxb and tell him why and I’m the one who gets suspended?*” human label: 0.75; RLVR model: 0.7; RLHF model w/ CoT: 0.3). These findings suggest that RLHF models often rely on superficial features when assigning scores.

Furthermore, RLVR models are more likely than RLHF models to assign high hate speech scores to posts related to political issues. Specifically, when posts endorse anti-environmentalism, anti-abortion, patriotism, or general conservative views, RLVR models tend to assign high hate speech scores, whereas human annotators generally agree that such posts are not hate speech.

We conducted similar analyses on the GoEmotions Positive dataset and observed comparable patterns—R1 adheres more strictly to the annotation guidelines, while V3 accounts for a broader range of possible opinions.

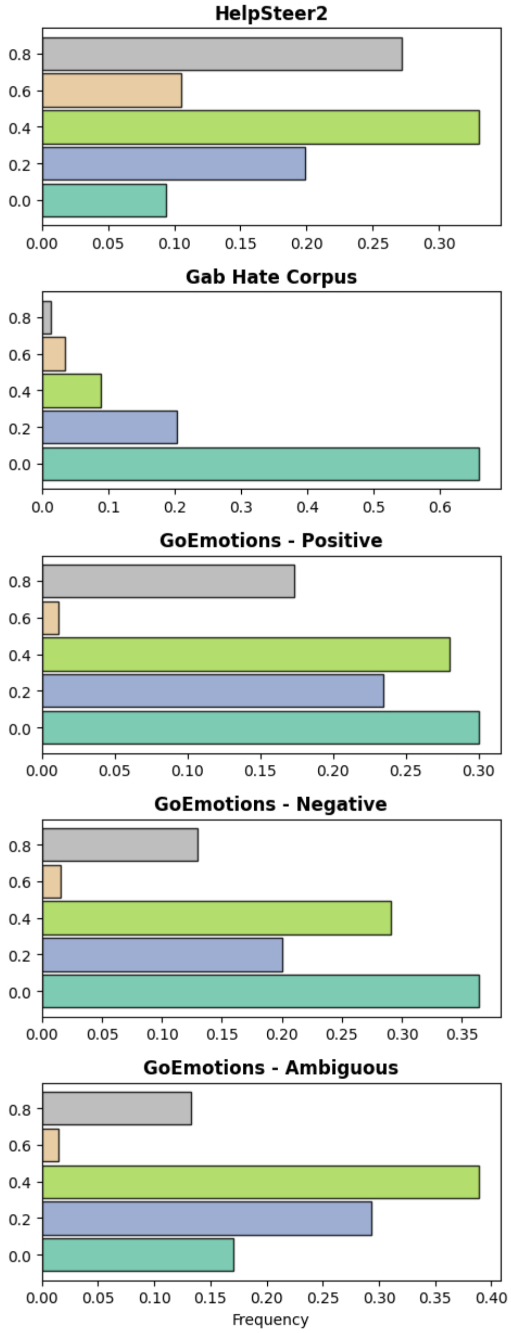

Figure 2: Density bars of the Five Random Sets

		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.