We thank the Area Chair and Reviewers of the previous round for their valuable suggestions! **In this revision, we address all concerns raised by the previous reviewers.** Changes are summarized as follows:

Concern 1 (Reviewer uD8C, F4PY and Meta Review): *a crucial related work from EMNLP 2023 is missing in the submission (Lee et al. 2023. "Can Large Language Models Capture Dissenting Human Voices?")*

We thank the reviewers for raising this important work that we unfortunately do not encounter when writing the first version. Our work has a different evaluation purpose from Lee et al., 2023. We made the following changes to differentiate our contribution to this prior work:

- **Change 1 in Title:**
  From *"Can Large Language Models Capture Human Annotator Disagreements"* to *"Can Reasoning Help Large Language Models Capture Human Annotator Disagreement?".* This emphasizes that our contribution is focused on identifying the effect of reasoning on disagreement modeling **(lines 131 to 143)**

- **Change 2 in Section 1 Introduction:**
  We emphasize the research gap from **lines 73 to 95**4– evaluation of reasoning and its combination with other factors is underexplored in prior work.

- **Change 3 in Section 2 Background and Related Work:**
  Detailed comparison to Lee et al., 2023 from **lines 178 to 201.** Specifically, Lee et al., 2023 does not explicitly instruct LLMs for disagreement modeling (instruction-task discrepancy). Reasoning and other important factors like distribution expression methods, few-shot steering are also not explored.

Concern 2 (Reviewer uD8C, F4PY and Meta Review): *Lacking in-depth qualitative analysis of the results.*

Change by adding Section 6.7 (lines 611 to 628):
We conduct a comprehensive qualitative analysis, providing insights on why RLVR LLMs are worse when there is high human disagreement. Specifically, 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.

Due to space limitation, we put detailed discussions and examples in **Appendix I (line 1188)**

Concern 3 (Reviewer yt1W and Meta Review): The authors distinguish their study from Lee et al. 2023 addressing their innovation of using CoT reasoning. However, this is more of a technical advancement between 2023 and 2025 and the paper lacks significant methodological innovation.

Change in Section 5.1 (lines 348 to 359):
The methodological differences between our work and Lee et al., 2023 are (1) evaluation focusing on the reasoning of RLVR and RLHF LLMs; and (2) controlled study on full

combinations of other factors. From lines 349 to 360, we present why considering other factors is important, and why the method of Lee et al., 2023 is not sufficient. The inspiration from causal inference theory is detailed in **Appendix C (line 1065)**

Furthermore, we want to highlight that this paper is an evaluation paper instead of a methodological paper (see **lines 131 to 143**). Our evaluation is valuable in 2025. An analogy is that LLM calibration has been evaluated in various 2023 papers (e.g., https://arxiv.org/abs/2305.14975). However, it is still a worth-evaluating problem in 2025, as the reasoning driven by RLVR is an important paradigm shift (e.g., https://arxiv.org/abs/2506.18183, https://arxiv.org/abs/2505.14489, https://arxiv.org/abs/2504.06564).

Concern 4 (Reviewer yt1W and Meta Review): potential annotation noise

Change in Section 4 by adding lines 315 to 324:
We explicitly present evidence and justifications for why our evaluation sets are of limited annotator noise and can imply the true annotation distribution.

Concern 5 (Meta Review): writing clarity in section 4.

Change in Section 4 by following the Meta Review's suggestions on dataset number description (line 287) and emphasize number of tasks in line 133

The Meta Reviewer also has a few other suggestions. However, we found them extremely challenging to address:

Concern 6 (Meta Review): Compare LLM reasoning with Human Reasoning by leveraging HateXplain (with human explanation of annotations); and annotate human reasoning using multiple annotators.

Datasets in our evaluation, although widely studied in literature, do not have human explanations for their annotations. Therefore, it is impossible to compare human reasoning with LLM reasoning. The Meta Review further suggested two solutions:

(1) Use HateXplain: we found HateXplain only has explanations in a form of which tokens are offensive (i.e., why annotate positive). It does not have explanations on why a post does not contain hate speech (i.e., why annotate negative). Therefore, it is impossible to use it to assess human reasoning.
(2) Use multiple annotators to annotate human reasoning: We cannot hire a group of new annotators to **guess** the reasoning behind disagreement of the original annotators. This is not scientifically possible.

Furthermore, our goal is to investigate the effect of LLM reasoning on disagreement modeling, instead of studying the gap between human and LLM reasoning. Therefore, this is an interesting-to-have experiment, but not necessary for the soundness of our conclusion.