Decoupling the Effect of Chain-of-Thought Reasoning: A Human Label Variation Perspective

Beiduo Chen **Tiancheng Hu** **Caiqi Zhang** **Robert Litschko**
Anna Korhonen **Barbara Plank**

MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
Language Technology Lab, University of Cambridge, United Kingdom
{beiduo.chen, robert.litschko, b.plank}@lmu.de
{th656, cz391, alk23}@cam.ac.uk

Abstract

Reasoning-tuned LLMs utilizing long Chain-of-Thought (CoT) excel at single-answer tasks, yet their ability to model Human Label Variation—which requires capturing probabilistic ambiguity rather than resolving it—remains underexplored. We investigate this through systematic disentanglement experiments on distribution-based tasks, employing Cross-CoT experiments to isolate the effect of reasoning text from intrinsic model priors. We observe a distinct "decoupled mechanism": while CoT improves distributional alignment, final accuracy is dictated by CoT content (99% variance contribution), whereas distributional ranking is governed by model priors (over 80%). Step-wise analysis further shows that while CoT's influence on accuracy grows monotonically during the reasoning process, distributional structure is largely determined by LLM's intrinsic priors. These findings suggest that long CoT serves as a decisive LLM decision-maker for the top option but fails to function as a granular distribution calibrator for ambiguous tasks.

1 Introduction

Reasoning-tuned large language models (LLMs) with long CoT reasoning achieve strong performance on many benchmarks (Touvron et al., 2023; Dubey et al., 2024; OpenAI, 2023; Wei et al., 2022; Wang et al., 2023; DeepSeek-AI et al., 2025; Team, 2025c; Hurst et al., 2024), usually measured by accuracy under the assumption of a single correct answer (Hendrycks et al., 2021a; Rein et al., 2023; Wang et al., 2024; Sun et al., 2025; Hendrycks et al., 2021b). However, many real-world tasks are inherently ambiguous or subjective, with human annotators often disagreeing due to genuine semantic uncertainty (Pavlick and Kwiatkowski, 2019; Aroyo and Welty, 2015). Such Human Label Variation (HLV) requires models to predict distributions over plausible answers, making argmax-based evaluation insufficient (Uma et al., 2021; Plank, 2022; Cabitza et al., 2023; Hu et al., 2025). Intuitively, reasoning through intermediate steps might better reflect such variations compared to direct answering (Chen et al., 2025a), motivating us to ask *RQ1: whether long CoT helps models better approximate human label distributions*, and *RQ2: whether any gains come from CoT reasoning or the model's latent parametric knowledge.*

To investigate *RQ1*, we utilize ChaosNLI (Nie et al., 2020), a benchmark capturing collective human opinions. We analyze the latent answer distributions behind CoT using complementary metrics: accuracy for correctness, and Jensen–Shannon Divergence (JSD, Endres and Schindelin 2003) and Spearman's ρ (Spearman, 1961) for distributional and ranking alignment. To further disentangle CoT's role from model-intrinsic priors (*RQ2*), we conduct: i) Cross-CoT experiments, injecting one model's CoT into another to test reasoning transfer; and ii) Step-wise analysis, truncating CoT to track how influence evolves over reasoning steps.

Our analysis uncovers a notable "split influence". While LLMs generally improve distributional alignment (lower JSD) after reasoning, this gain is not uniform across metrics. Using ANOVA to calculate the variance contribution percentage in our Cross-CoT experiments, we find that final accuracy is overwhelmingly determined by the CoT content (≈99%), confirming the strong role of reasoning chain in steering the top-1 answer decision. In stark contrast, the distributional structure—ranking and probability allocation among non-argmax options—is largely immune to CoT, remaining governed by model priors (>80%).

Step-wise analysis further clarifies this dynamic. While all metrics evolve throughout reasoning, changes in accuracy are predominantly driven by CoT and grow monotonically with later steps. By comparison, changes in distributional similarity (JSD and Spearman's ρ) are mostly determined by the LLM's intrinsic behavior. This reveals a di-

Findings of the Association for Computational Linguistics: ACL 2026, pages 21931–21951
July 2-7, 2026 ©2026 Association for Computational Linguistics

chotomy: current long CoT paradigms act as strong LLM decision makers but weak distribution calibrators. CoT tends to progressively concentrate probability mass to lock in the most likely answer latently, but fails to govern the reshaping of the probability landscape for alternative options. This work highlights the structural limitations of current reasoning processes in capturing fine-grained answer uncertainty and motivates the need for distribution-aware reasoning mechanisms.

2 Background

HLV in Natural Language Inference. Unlike the single-label assumption in most benchmarks, NLI is often inherently ambiguous: a premise and hypothesis can elicit a spectrum of plausible interpretations, a phenomenon known as HLV (Plank, 2022). Benchmarks such as ChaosNLI capture this by representing labels as probability distributions rather than single gold labels (Nie et al., 2020; Weber-Genzel et al., 2024; Jiang et al., 2023; Hong et al., 2025a,b). Evaluating models under HLV requires moving beyond standard accuracy to distributional metrics that measure alignment with collective human judgments (Kurniawan et al., 2025; Lee et al., 2023; Leonardelli et al., 2023; Chen et al., 2024, 2025b,a; Ni et al., 2025).

Reasoning under Distributional Uncertainty. Recent LLM advancements emphasize reasoning-intensive paradigms. Long CoT enables models to decompose problems into intermediate steps (Wang et al., 2023; DeepSeek-AI et al., 2025; Team, 2025c; Hurst et al., 2024), effectively reducing uncertainty and producing high-confidence conclusions in deterministic tasks. However, its role in probabilistic HLV settings is less clear. Generating explicit reasoning can inadvertently suppress valid alternative interpretations, potentially biasing the model toward the top-1 choice. While prior work has explored confidence-based calibration (Zhao et al., 2025; Yoon et al., 2025; Mao et al., 2025b), it remains unclear whether CoT actively shapes the full output distribution or mainly rationalizes the final decision, leaving non-argmax probabilities governed by the model's intrinsic priors.

3 Experiments

3.1 Setup

Task We experiment on 3 ChaosNLI subsets: MNLI, SNLI, and αNLI (Bowman et al., 2015; Williams et al., 2018; Bhagavatula et al., 2020). Each instance is annotated by 100 crowdworkers, enabling reliable human judgment distributions (HJD). MNLI and SNLI are three-way classification tasks (entailment, neutral, contradiction), yielding 3-d label distributions. αNLI is a binary-choice task, where annotators select the better hypothesis for a given observation pair, producing 2-d distributions.[1] Dataset details are in Appendix A.

Reasoning LLMs	Abbr.
Qwen/Qwen3-30B-A3B-Thinking-2507 (Team, 2025b)	Qwen
deepseek-ai/DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI et al., 2025)	R1-Llama
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025)	R1-Qwen
allenai/Olmo-3-32B-Think (Olmo et al., 2025)	Olmo
zai-org/GLM-Z1-32B-0414 (GLM et al., 2024)	GLM
ByteDance-Seed/Seed-OSS-36B-Instruct (Team, 2025a)	Seed
openai/gpt-oss-20b (OpenAI, 2025)	GPT

Table 1: Reasoning LLMs and their abbreviation.

Models To comprehensively evaluate the HLV performance of reasoning-tuned LLMs, we select a range of state-of-the-art open-source reasoning models (details in Table 1). All follow a reason-then-answer paradigm: generating a long CoT reasoning process before outputting a final answer.

Evaluation All NLI instances are reformulated as multiple-choice questions. Model predictions are extracted using the first-token probability method (Santurkar et al., 2023; Durmus et al., 2023; Liang et al., 2023), where logits are aggregated and normalized to obtain an output probability distribution over answer options.[2] We measure the HLV alignment between the model-generated distribution and the corresponding HJD using JSD. We also report accuracy. Additionally, we include Spearman's ρ in Section 4.1, which is invariant to monotonic transformations. See details in Appendix B.

3.2 Does CoT Improve HLV Performance?

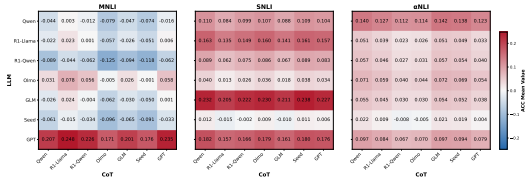
We examine the impact of reasoning by comparing model performance before and after CoT. See Table 2, the effect of CoT on accuracy is mixed. While most models improve on SNLI and αNLI, performance on MNLI is highly unstable. In contrast, JSD consistently decreases across nearly all models and datasets. Importantly, this improved distributional alignment is often decoupled from

[1] ChaosNLI is ideal for HLV evaluation as a rare non–social science benchmark with collective HJDs (Hu et al., 2025).

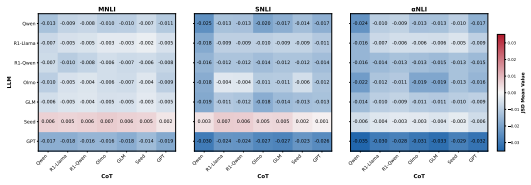
[2] While this approximation may not fully capture downstream decoding dynamics, our analysis focuses on relative comparisons across controlled conditions, where this proxy is consistently applied.

Task	MNLI				SNLI				αNLI			
LLMs/Metrics	$ACC_{start}\uparrow$	$ACC_{last}\uparrow$	$JSD_{start}\downarrow$	$JSD_{last}\downarrow$	$ACC_{start}\uparrow$	$ACC_{last}\uparrow$	$JSD_{start}\downarrow$	$JSD_{last}\downarrow$	$ACC_{start}\uparrow$	$ACC_{last}\uparrow$	$JSD_{start}\downarrow$	$JSD_{last}\downarrow$
Qwen	0,688	0,644	0,093	0,080	0,668	0,778	0,144	0,119	0,749	0,890	0,108	0,084
R1-Llama	0,666	0,689	0,082	0,077	0,615	0,750	0,133	0,123	0,839	0,878	0,098	0,091
R1-Qwen	0,734	0,672	0,080	0,072	0,689	0,764	0,127	0,115	0,832	0,860	0,094	0,081
Olmo	0,614	0,609	0,088	0,082	0,738	0,775	0,133	0,122	0,819	0,863	0,107	0,087
GLM	0,670	0,640	0,082	0,077	0,545	0,756	0,134	0,120	0,834	0,888	0,099	0,088
Seed	0,705	0,614	0,077	0,083	0,766	0,777	0,124	0,127	0,868	0,887	0,098	0,095
GPT	0,437	0,672	0,095	0,077	0,596	0,772	0,145	0,119	0,793	0,872	0,112	0,080

Table 2: Results before and after reasoning. *start* and *last* denote before reasoning and after completion. Red indicates an increase, blue a decrease. Arrows next to metric names show whether higher or lower is better.

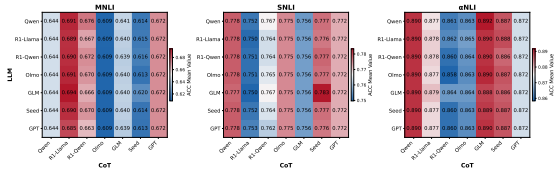


(a) Delta Accuracy ↑

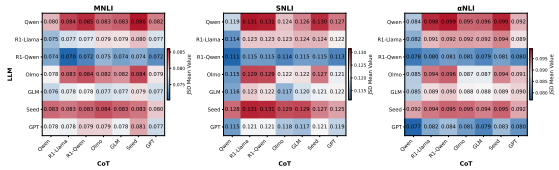


(b) Delta JSD ↓

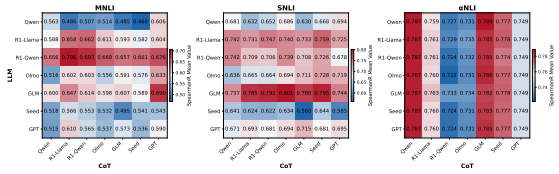
Figure 1: Changes after reasoning for Cross-CoT.



(a) Last Accuracy ↑



(b) Last JSD ↓



(c) Last Spearman's ρ ↑

Figure 2: Last results after reasoning for Cross-CoT.

accuracy: even when the accuracy decreases (e.g., Qwen on MNLI), the output distribution aligns more closely with human judgments on average.

CoT generally reduces JSD, indicating useful signals for HLV, but the benefit varies across models. Models with similar post-CoT accuracy can exhibit substantially different JSD values, and vice versa, raising the key attribution question: *are the gains driven by the semantic content of the reasoning itself, or by model-specific inductive biases when interpreting the reasoning text?*

3.3 Cross-CoT Evaluation

To disentangle the source of JSD improvements, we conduct *Cross-CoT* experiments, injecting reasoning paths from different source models into various inference models. We show the performance changes in Figure 1. On MNLI, accuracy shows mixed patterns. In contrast, consistent with single-model results, JSD improves across nearly all Cross-CoT pairings.[3] Regardless of the reasoning source model, injecting a CoT almost universally reduces divergence from human distributions.

[3]The box plot (Figure 4) in Appendix C shows improvements are widespread across instances, not driven by outliers.

This confirms: **CoT text acts as a portable carrier of HLV-relevant information—reasoning generated by one model can facilitate better distributional alignment in another.** However, the divergent patterns between accuracy and JSD motivate us to examine why these two metrics respond differently to CoT, and how CoT influences them.

4 Analyses

4.1 What Does CoT Determine?

If CoT were the dominant driver, models conditioned on the same CoT should converge to similar outcomes. To test this, we analyze final-step accuracy and JSD. The results (Figure 2(a), 2(b)) reveal a clear dissociation. Accuracy shows a *column-dominant pattern*: for a fixed CoT source, accuracy is nearly identical across inference models, indicating that CoT largely dictates the argmax decision.

Task	MNLI									SNLI									αNLI								
Metric	ACC			JSD			Spearman's ρ			ACC			JSD			Spearman's ρ			ACC			JSD			Spearman's ρ		
Step/Factor	LLM	CoT	Residual	LLM	CoT	Residual	LLM	CoT	Residual	LLM	CoT	Residual	LLM	CoT	Residual	LLM	CoT	Residual	LLM	CoT	Residual	LLM	CoT	Residual	LLM	CoT	Residual
Step 0	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Step 1	97.3%	0.2%	2.5%	96.9%	0.4%	2.7%	92.7%	0.4%	6.9%	96.5%	1.4%	2.2%	94.4%	0.9%	4.7%	83.2%	2.4%	14.4%	93.9%	1.4%	4.7%	95.4%	1.3%	3.3%	94.0%	1.3%	4.8%
Step 2	95.7%	0.5%	3.7%	95.3%	0.7%	4.0%	92.3%	1.8%	5.9%	85.1%	8.6%	6.4%	88.2%	4.4%	7.4%	76.1%	8.0%	15.8%	75.7%	10.5%	13.8%	89.3%	3.9%	6.8%	76.3%	10.4%	13.2%
Step 3	91.8%	1.6%	6.6%	93.4%	1.6%	5.0%	91.6%	2.6%	5.8%	75.1%	16.7%	8.1%	82.2%	8.3%	9.5%	73.2%	12.2%	14.6%	64.1%	22.2%	13.7%	81.4%	8.8%	9.8%	64.3%	22.4%	13.3%
Step 4	89.1%	2.2%	8.7%	91.5%	2.9%	5.6%	90.5%	3.6%	5.9%	62.5%	29.1%	8.5%	75.4%	14.0%	10.6%	69.9%	16.4%	13.7%	52.1%	31.2%	16.8%	80.8%	9.5%	9.6%	51.9%	32.6%	15.5%
Step 5	85.2%	4.3%	10.6%	89.3%	4.2%	6.4%	88.0%	4.9%	7.1%	57.4%	32.4%	10.2%	72.8%	16.6%	10.6%	70.8%	13.0%	16.2%	45.2%	41.7%	13.1%	81.6%	10.4%	8.0%	47.2%	41.3%	11.5%
Step 6	79.4%	8.5%	12.0%	87.9%	5.7%	6.5%	83.6%	7.6%	8.8%	58.9%	29.7%	11.3%	72.5%	17.4%	10.1%	71.9%	10.4%	17.7%	35.4%	53.8%	10.8%	84.9%	9.4%	5.8%	36.4%	53.3%	10.3%
Step 7	66.5%	18.4%	15.1%	87.5%	6.1%	6.4%	77.5%	10.1%	12.4%	53.4%	30.6%	16.0%	73.1%	17.2%	9.7%	73.9%	6.6%	19.5%	22.5%	67.3%	10.2%	87.0%	8.2%	4.9%	22.1%	67.5%	10.4%
Step 8	44.1%	38.3%	17.5%	86.4%	7.0%	6.7%	73.3%	14.5%	12.2%	43.0%	33.2%	23.8%	72.1%	19.1%	8.8%	71.8%	7.9%	20.3%	9.6%	76.2%	14.2%	85.2%	10.1%	4.7%	10.1%	76.9%	13.0%
Step 9	22.4%	65.2%	12.4%	84.3%	8.5%	7.2%	73.0%	16.8%	10.3%	34.8%	43.9%	21.2%	71.3%	19.8%	8.9%	77.8%	6.3%	15.9%	5.3%	83.5%	11.2%	81.4%	13.2%	5.4%	5.4%	84.1%	10.5%
Step 10	0.1%	99.5%	0.4%	83.3%	8.7%	8.1%	71.1%	13.1%	15.7%	0.2%	98.6%	1.2%	67.7%	22.2%	10.1%	81.7%	3.3%	15.0%	0.1%	99.4%	0.5%	76.9%	15.2%	7.9%	0.1%	99.4%	0.5%

Table 3: Step-wise ANOVA results. Each CoT is split into 10 segments by sentence, yielding 11 intermediate answers from no-thinking (step 0) to full-thinking (step 10). ANOVA is computed for each Cross-CoT heatmap. Red numbers indicate the factor dominating the metric at that step. All step-wise heatmaps are in Appendix G.

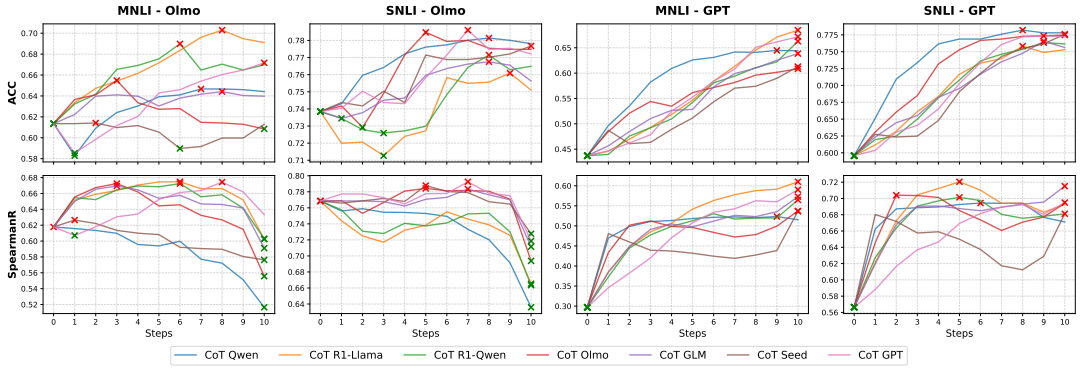


Figure 3: Curve cases for step-wise evaluation. Max and min points are marked. All results are in Appendix H.

In contrast, JSD exhibits a *row-dominant pattern*: final divergence is primarily determined by the inference model, with little sensitivity to the CoT. This suggests that CoT provides a directional signal, while the final distributional shape remains constrained by model-specific priors.

We further analyze Spearman's ρ as a relaxed non-argmax metric (only rankings). Its heatmap (Figure 2(c)) mirrors the row-dominant structure of JSD rather than accuracy.[4] We adopt an additive Analysis of Variance (ANOVA) model to estimate marginal contributions of CoT and the model's priors. Interaction effects, while not explicitly modeled, are captured in the residual term.[5] ANOVA on MNLI confirms this split: CoT explains 99% of the variance in accuracy, but only a small fraction in JSD and Spearman's ρ (8.7% and 13.1%), where model identity dominates (83.3% and 71.1%). This asymmetry exposes a fundamental limitation of current CoT paradigms. CoT is highly effective at explicit decision-making, capable of overriding a model's prior to determine the argmax. However, in the non-argmax space—namely, the ranking and probability allocation over alternative options—its influence sharply diminishes. **Models appear to follow CoT for the final choice, but revert to their latent parametric preferences when distributing uncertainty.**

4.2 When Does CoT Take Control?

Although CoT explains nearly all variance in accuracy, its impact on distributional metrics remains limited. To understand how this asymmetry develops, we apply early stopping to the CoT, truncating it at fixed increments and evaluating intermediate performance.[6] Step-wise ANOVA (Table 3) reveals a sharp divergence. CoT influence on accuracy remains modest during reasoning, then spikes abruptly at the final step, forming a clear inflection. In contrast, its influence on JSD and Spearman's ρ stays uniformly low, with no point at which CoT overrides the model's ranking behavior. To ensure our findings are not artifacts of the chosen probability mapping, we additionally compute JSD using a standard softmax transformation. As detailed in Appendix F, this robustness check yields consistent results, confirming the decoupled mechanism regardless of the normalization scheme.

Notably, the residual variance in our ANOVA decomposition remains consistently small (often under 10%). This indicates that the main effects

[4]Note that, in binary-choice αNLI, ρ is effectively equivalent to accuracy, as only a single non-argmax position exists.

[5]Details of ANOVA are in Appendix D.

[6]Implementation details are elaborated in the Appendix E. All code, logits, and CoT outputs are publicly available at https://github.com/mainlp/CoT-HLV.

overwhelmingly dominate the predictive variance. Because the additive ANOVA model absorbs any unmodeled interactions into the residual term, the minimal magnitude of these residuals assures that the primary conclusions regarding the decoupled mechanism remain robust.

This pattern is also illustrated by representative models (Figure 3). Accuracy often shifts or converges only at the conclusion, while Spearman's ρ fluctuates without a consistent trend. Thus, CoT determines the LLM's final choice but not the structure of uncertainty. Our anecdotal evidence in Appendix I supports that this can be attributed to the CoT format: standard CoT often ends with an explicit conclusion, providing a strong argmax signal, while distributional cues remain implicit. Consequently, **models leverage CoT for decision-making but revert to intrinsic priors for probability allocation, exposing a structural inability of raw CoT to shape answer distributions.**

5 Discussion and Future Work

Our results indicate that scaling inference-time compute via longer CoT is insufficient for resolving complex semantic ambiguity. Rather than preserving or refining uncertainty, standard CoT primarily acts as a mechanism for collapsing a distribution over hypotheses into a single high-confidence argmax prediction. This suggests a broader limitation of current reasoning paradigms: they are primarily optimized for decisiveness, rather than for faithfully representing uncertainty.

Beyond motivating the need for improved HLV evaluation, our findings point to several concrete directions for advancing both modeling and analysis of reasoning systems.

First, to better capture HLV, future work should move beyond implicit reasoning traces and develop explicitly distribution-aware reasoning frameworks. Instead of solely optimizing for the final answer, models should be trained to maintain, calibrate, and communicate relative uncertainties over competing hypotheses throughout intermediate reasoning steps. This may involve new training objectives, decoding strategies, or supervision signals that explicitly reward distributional fidelity.

Scope of Evaluation. Our study focuses on NLI-style multiple-choice question answering (MCQA), a deliberately controlled setting that enables reliable extraction of probability distributions via first-token probabilities. This design avoids the additional noise introduced by open-ended generation, such as answer normalization and semantic equivalence issues. While MCQA is a standard benchmark for evaluating reasoning in LLMs, extending our distributional analysis to settings with larger label spaces and open-ended outputs is a natural next step. Such extensions would require more robust methods for mapping free-form generations to structured belief distributions.

Calibration vs. Distributional Similarity. Our evaluation framework emphasizes relative distributional alignment, measuring how closely model-predicted belief distributions match human annotations using metrics such as JSD and Spearman's ρ. However, this setup does not assess absolute probability calibration with respect to empirical correctness frequencies. A promising direction for future work is to jointly study distributional similarity and calibration, for example by incorporating metrics such as Expected Calibration Error (ECE), to better understand whether reasoning models are not only structurally aligned with human beliefs but also probabilistically well-grounded.

CoT Format Ablations. Finally, while the observed decoupling effect remains consistent across different source and reasoning models in our Cross-CoT framework, we do not isolate the causal contribution of specific reasoning formats. In particular, it remains unclear whether elements such as explicit final-answer cues or structural regularities in CoT directly govern decoding behavior. Future work could investigate this question through controlled format-level ablations, synthetic CoT interventions, or human-authored reasoning traces, enabling a more fine-grained understanding of how reasoning structure influences model predictions.

6 Conclusion

From an HLV perspective, we identify a fundamental "split influence" in CoT reasoning: while reasoning content predominantly determines the LLM's final argmax choice latently, the probability landscape of alternative options remains anchored to the model's intrinsic priors. This exposes a structural limitation where standard CoT effectively collapses ambiguity for decision-making but fails to calibrate fine-grained uncertainty for alternative, plausible answers. Consequently, advancing HLV modeling requires moving beyond implicit reasoning traces toward distribution-aware paradigms.

Limitations

Our work has two main limitations. First, our evaluation relies solely on final human label distributions, as ChaosNLI lacks annotated intermediate reasoning steps. Consequently, our step-wise analysis compares intermediate model outputs against the final human consensus rather than step-specific ground truth. Addressing this limitation would require future improvements in human annotation, where reasoning steps and intermediate answers are collected for direct comparison. An alternative approach could be the use of relative references: for example, treating the intermediate answers generated by a CoT-provider LLM as the gold standard to evaluate the faithfulness of other LLMs. Another possibility is to employ an entailment model to determine whether each reasoning step in a CoT entails the previous step, thereby inferring intermediate answers recursively. However, both of these approaches are highly dependent on the accuracy of the model itself; inaccuracies could introduce evaluation biases.

Second, our study does not include a direct human evaluation of the textual content of the CoTs. Instead, we focus on assessing CoTs in terms of their impact on LLM behavior, especially answer distributions. While this approach emphasizes the effect of reasoning on model outputs, it overlooks the quality of the specific text content. Conducting human evaluation of long CoTs is particularly resource-intensive, given their length and the effort required to annotate individual sentences and their interrelations. Nevertheless, considering the growing importance of CoT reasoning in NLP, carefully designed human evaluation to verify whether extreme values in reasoning metrics correspond to reasonable positions in the text represents a promising direction for future work. Such evaluation could help us better understand the effects of CoT on LLM reasoning.

Ethical Considerations

This work primarily involves the analysis of NLI datasets and open-sourced LLMs. All data used are publicly available and do not contain personally identifiable information. No sensitive or potentially harmful content is generated or utilized in this study. Therefore, we do not anticipate any ethical concerns arising from our work.

Use of AI Assistants The authors acknowledge the use of ChatGPT solely for correcting grammatical errors, enhancing the coherence of the final manuscript.

Acknowledgements

We thank the members of the MaiNLP lab for their insightful feedback on earlier drafts of this paper. We specifically appreciate the suggestions of Siyao Peng, Jana Grimm, Florian Eichin and Benedetta Muscato. We are also grateful to the anonymous reviewers for their constructive feedback. BC acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program. BP and RL are supported by ERC Consolidator Grant DIALECT 101043235. AK is supported by the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 EQUATE (the UK government's funding guarantee for ERC Advanced Grants).

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36(1):15–24.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6860–6868. AAAI Press.

Beiduo Chen, Yang Janet Liu, Anna Korhonen, and Barbara Plank. 2025a. Threading the needle: Reweaving chain-of-thought reasoning to explain human label variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*,

pages 33099–33123, Suzhou, China. Association for Computational Linguistics.

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025b. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10777–10802, Vienna, Austria. Association for Computational Linguistics.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulic, Andreea Bobu, and Nigel Collier. 2026. Value of information: A framework for human-agent communication. *CoRR*, abs/2601.06407.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin andAbhimanyu 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. 2023. Towards measuring the representation of subjective global opinions in language models. *CoRR*, abs/2306.16388.

Dominik Maria Endres and Johannes E. Schindelin. 2003. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2025a. LiTEx: A linguistic taxonomy of explanations for understanding within-label variation in natural language inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34053–34073, Suzhou, China. Association for Computational Linguistics.

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, Benjamin Roth, and Barbara Plank. 2025b. Agree, disagree, explain: Decomposing human label variation in NLI through the lens of explanations. *CoRR*, abs/2510.16458.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *CoRR*, abs/2510.17516.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2025. Training and evaluating with human label variation: An empirical study. *CoRR*, abs/2502.01891.

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.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023.

Yihong Liu, Raoyuan Zhao, Hinrich Schütze, and Michael A. Hedderich. 2026. Large reasoning models are (not yet) multilingual latent reasoners. *CoRR*, abs/2601.02996.

Minjia Mao, Bowen Yin, Yu Zhu, and Xiao Fang. 2025a. Early stopping chain-of-thoughts in large language models. *CoRR*, abs/2509.14004.

Zhenjiang Mao, Artem Bisliouk, Rohith Reddy Nama, and Ivan Ruchkin. 2025b. Temporalizing confidence: Evaluation of chain-of-thought reasoning with signal temporal logic. *CoRR*, abs/2506.08243.

Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2025. Can reasoning help large language models capture human annotator disagreement? *Preprint*, arXiv:2506.19467.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Charles Spearman. 1961. The proof and measurement of association between two things.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, Lei Fang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *CoRR*, abs/2503.21380.

ByteDance Seed Team. 2025a. Seed-oss open-source models. `https://github.com/ByteDance-Seed/seed-oss`.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qwen Team. 2025c. Qwq-32b: Embracing the power of reinforcement learning.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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. *CoRR*, abs/2505.14489.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. 2025. Is chain-of-thought reasoning of llms a mirage? A data distribution lens. *CoRR*, abs/2508.01191.

Raoyuan Zhao, Yihong Liu, Hinrich Schuetze, and Michael A. Hedderich. 2026. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5223–5247, Rabat, Morocco. Association for Computational Linguistics.

A Datasets

To evaluate the model's ability to capture collective human uncertainty and label disagreement, we utilize the **ChaosNLI** dataset (Nie et al., 2020). Unlike standard NLI benchmarks that typically rely on a single "gold" label derived from a majority vote (often among 3–5 annotators), ChaosNLI provides a dense distribution of human annotations.

- **Data Source:** The dataset consists of purely English examples, selected from of SNLI (1514 items, Bowman et al. 2015), MNLI (1599 items, Williams et al. 2018) and αNLI (1532 items, Bhagavatula et al. 2020).
- **Selection Criteria:** The examples were specifically chosen to target ambiguous instances. The authors filtered for examples where the original annotators disagreed (e.g., a 3 vs. 2 vote split) or where the model predictions significantly deviated from the majority label.
- **Annotation Process:** Each example in ChaosNLI is annotated by a crowd of $N = 100$ independent workers. This high volume of annotators allows for the estimation of a true label distribution y_{human} over the three classes (Entailment, Neutral, Contradiction), rather than a deterministic class label. As for αNLI, annotators are asked to select the better hypothesis from a sentence pair for a given observation pair.
- **Objective:** The dataset serves as a testbed for measuring how well a model's predicted probability distribution p_{model} aligns with the distribution of human judgment y_{human}, often measured via Jensen-Shannon Divergence (JSD).

B Evaluation Details

This section elaborates on the details of the experimental setup. We first describe the experiment details, and then introduce how the NLI task is transformed into a multiple-choice question answering (MCQA) format. We finally introduce the procedure for extracting and converting first-token probabilities, followed by a formal definition of the evaluation metrics used in this paper.

B.1 Experiment Details

All LLMs are evaluated using the initial or recommended parameter settings provided by their respective developers, ensuring that each model generates Chain-of-Thought (CoT) outputs consistent with its intended behavior and style. Since our analysis focuses on the model logits rather than the sampled textual outputs, variations in sampling-related parameters (e.g., temperature, top-k, or top-p) do not affect the logits-based evaluations. This design ensures that our comparisons reflect the models' intrinsic preference distributions rather than stochastic differences introduced by the decoding process.

All experiments were conducted on two NVIDIA A100-SXM4-80GB GPUs. On average, complet-

ing a single experiment—which involves generating step-wise intermediate logits for one LLM on a single dataset using a given Chain-of-Thought (CoT)—takes approximately 20 hours. This reflects the computational demands of step-wise evaluation across multiple inference steps and highlights the resource-intensive nature of detailed logit-level analyses for large language models.

B.2 MCQA Format

The conversion to the MCQA format is illustrated in Table 4. Since MNLI and SNLI belong to the same category of NLI datasets, they are transformed using an identical three-way multiple-choice formulation. In contrast, αNLI is converted using a separate binary-choice MCQA format, reflecting its distinct label structure.

B.3 First-Token-Probability and Metrics

B.3.1 First-token Probability

Take MNLI as an example. Conditioned on the prompts described above, we further map LLM outputs from discrete options in `[A, B, C]` to probability distributions, which we treat as model judgment distributions (MJDs). Specifically, we define a one-to-one mapping $f\colon O \to L$ from the option set O to the label space L, where $O = \{A, B, C\}$ and $L = \{\text{ENTAILMENT}, \text{NEUTRAL}, \text{CONTRADICTION}\}$. Both O and L are subject to permutation to mitigate positional and label-order biases.

Let the textual output of an LLM be represented as a sequence of tokens $\boldsymbol{w} = [w_1, w_2, \ldots, w_k]$, where $w_i \in V$, k denotes the output length, and V is the model vocabulary. Instead of using the decoded output, we extract the pre-decoding logits corresponding to the first generated token w_1:

$$\boldsymbol{s}_{w_1} = [s_1, s_2, \ldots, s_n], \quad n = |V|,$$

where s_j denotes the logit associated with the j-th vocabulary token.

We restrict our attention to the subset of logits corresponding to the option tokens in O,

$$\boldsymbol{s}^{O}_{w_1} = [s_A, s_B, s_C],$$

which encode the model's relative preference over the candidate options. Since the normalization transformation preserves the entropy of the original logits, whereas the softmax transformation (especially when applied with a temperature parameter) can alter entropy, we adopt the normalization transformation for our evaluations, because entropy plays a critical role in the computation of JSD, and using a transformation that artificially modifies it could bias the assessment. Therefore, to more accurately measure the LLMs' intrinsic probabilistic preferences and their native reasoning behavior, we rely on the norm transformation rather than softmax in our analysis.

To convert these scores into a probability distribution $\boldsymbol{p}^{O}$, we then apply a normalization step.

$$p^{O}_{\text{norm}}(j) = \frac{s_j}{\sum_{j=1}^{|O|} s_j}, \tag{1}$$

This procedure yields a well-formed probability distribution over labels, enabling fine-grained comparison with human-annotated label distributions.

B.3.2 Rank Correlation Metric

To quantify the agreement between ranked preferences from different sources (e.g., human annotations versus model predictions), we employ rank correlation metrics. Let $\{(x_i, y_i)\}_{i=1}^{n}$ denote paired ranks from two sources.

Spearman's ρ (Spearman, 1961) Spearman's rank correlation coefficient measures the Pearson correlation between ranked variables and is defined as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{2}$$

where $d_i = x_i - y_i$ denotes the rank difference for the i-th item. Spearman's ρ captures monotonic relationships and is robust to nonlinear transformations of the underlying scores.

B.3.3 Distribution-Based Metric

For settings where both human annotations and model outputs are represented as probability distributions, we adopt distributional similarity metrics.

Jensen–Shannon Distance (JSD) (Endres and Schindelin, 2003) Given two discrete probability distributions P and Q, the Jensen–Shannon Distance is defined as:

$$D_{\text{JSD}}(P\|Q) = \sqrt{\frac{1}{2}\left(D_{\text{KL}}(P\|M) + D_{\text{KL}}(Q\|M)\right)}, \tag{3}$$

where $M = \frac{1}{2}(P+Q)$ and $D_{\text{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. JSD is symmetric, bounded, and well-defined even when P and Q contain zero-probability entries, making it suitable for comparing soft label distributions.

Datasets	MCQA Transformation
MNLI & SNLI	Please determine whether the following statement is true (entailment), undetermined (neutral), or false (contradiction) given the context below and select ONE of the listed options and start your answer with a single letter. Context: {premise} Statement: {hypothesis} A. Entailment B. Neutral C. Contradiction Answer:
αNLI	Please determine which of the two hypotheses (A or B) is more likely to explain the transition from the beginning observation to the ending observation and select ONE of the listed options and start your answer with a single letter. Beginning: {begining-observation} Ending: {ending-observation} A. {hypothesis1} B. {hypothesis2} Answer:

Table 4: The MCQA transformation for NLI tasks.

C Box-plot for Cross-CoT Experiments

This section presents the distribution of the Delta JSD metric over all instances in the dataset, aiming to show that the observed reduction in JSD reflects a global trend rather than being driven by a small number of extreme cases that artificially lower the mean.

As illustrated in Figure 4, we visualize per-instance Delta JSD values using box plots, where the central line denotes the median, the boxes correspond to the interquartile range (IQR), and the whiskers extend to $1.5\times$IQR. Across the majority of experimental settings, the distributions are centered below zero, indicating a consistent overall decrease in JSD. These results suggest that the improvement captured by the average JSD is broadly shared across data points, rather than being dominated by a few outliers.

D ANOVA Details

To quantify the relative influence of different factors on the observed scores, we employ a two-way Analysis of Variance (ANOVA) with an additive (no-interaction) design. This statistical framework allows us to decompose the total variance of the dependent variable into contributions from multiple categorical factors.

D.1 Problem Setup

Let y_{ij} denote the observed score associated with the i-th model configuration and the j-th CoT setting. In our implementation, we consider:

- A **model factor** with $I = 7$ levels (indexed by i),
- A **CoT factor** with $J = 7$ levels (indexed by j),
- One observation for each (i, j) combination.

The data are organized into a long-form table with three columns: `score`, `model`, and `param`, where both `model` and `param` are treated as categorical variables.

D.2 Additive Two-Way ANOVA Model

We adopt an additive two-way ANOVA model without interaction terms, formulated as:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \tag{4}$$

where:

- μ is the overall mean score,
- α_i represents the main effect of the i-th model,
- β_j represents the main effect of the j-th CoT,
- ε_{ij} is the residual error term.

This formulation assumes that the effects of the two factors are independent and additive, and that no interaction between model and CoT is modeled. This choice aligns with our goal of isolating the marginal contribution of each factor.

D.3 Estimation via Ordinary Least Squares

The model is estimated using Ordinary Least Squares (OLS), implemented as:

$$\texttt{score} \sim C(\texttt{model}) + C(\texttt{param}), \quad (5)$$

where $C(\cdot)$ denotes categorical encoding. The fitted model is then passed to a Type-II ANOVA procedure, which computes sums of squares for each main effect after accounting for the other factor.

D.4 Variance Decomposition

ANOVA decomposes the total sum of squares (SS) as:

$$SS_{\text{total}} = SS_{\text{model}} + SS_{\text{param}} + SS_{\text{residual}}, \quad (6)$$

where:

- SS_{model} captures variance explained by the model factor,
- SS_{param} captures variance explained by the CoT factor,
- SS_{residual} captures unexplained variance.

Each sum of squares is associated with an F-statistic and corresponding p-value, allowing statistical significance testing of factor effects.

D.5 Contribution Percentage

To improve interpretability, we further compute the **variance contribution percentage** of each factor:

$$\text{Contribution}_k = \frac{SS_k}{SS_{\text{total}}} \times 100\%, \quad (7)$$

where $k \in \{\text{model}, \text{param}, \text{residual}\}$.

This metric reflects the proportion of total variance attributable to each source. In our implementation, these percentages are rounded to one decimal place and reported as the final output of the analysis.

D.6 Interpretation

The resulting contribution percentages provide a clear quantitative comparison of how much variability in the scores is explained by:

- differences between models,
- differences between CoT settings,
- unexplained residual noise.

This two-way additive ANOVA thus serves as an effective tool for disentangling and comparing the marginal effects of multiple experimental factors in our evaluation framework.

Note that our analysis employs an additive two-way ANOVA without explicit interaction terms. Since LLM inference in our framework is fully deterministic—yielding identical pre-decoding logits for a fixed prompt and CoT sequence—multiple runs per experimental cell are unnecessary. Consequently, the additive model is specifically designed to isolate and quantify the marginal effects of the CoT content versus the intrinsic model prior. Any potential interaction effects, along with unexplained variance, are cleanly absorbed into the residual error term (ϵ_{ij}).

E Implementation for Early Stopping

E.1 Accumulative 10% Segmenting of Chain-of-Thoughts

To facilitate analysis of reasoning progression in CoT outputs, we segment each text into ten accumulative portions, corresponding approximately to every 10% of the text length (Mao et al., 2025a; Zhao et al., 2026; Liu et al., 2026). Formally, given a text T of length L, we aim to identify cut points $p_1, p_2, \ldots, p_9$ such that p_i roughly corresponds to $i \cdot L/10$, with the final point $p_{10} = L$.

Our procedure combines sentence-aware and heuristic splitting strategies. First, we parse T into sentences using a syntactic parser and extract all sentence-ending positions. If at least nine sentences are available, we select each cut point p_i as the nearest sentence end to $i \cdot L/10$, ensuring monotonicity of cut points. If fewer than ten sentences exist, we iteratively split the longest existing segment, prioritizing natural boundaries such as punctuation (e.g., semicolons, commas) and spaces, and resorting to midpoint splits when no suitable boundary is found.

Finally, we enforce strict monotonicity of cut points and construct the accumulative segments

$S_1, S_2, \ldots, S_{10}$ where $S_i = T[: p_i]$. This ensures that the last segment always reproduces the full original text. This method preserves original spacing and punctuation, providing natural and interpretable checkpoints for CoT analysis at decile intervals.

E.2 Early-Stopping and Answer Token Extraction

To reliably extract intermediate reasoning outputs, we adopt an *early-stopping* strategy. Specifically, at each accumulative CoT segment cut point, we append a special token sequence signaling the model to terminate reasoning and produce an answer, e.g., "`\n</think>\n\nBased on the reasoning so far, the Answer is:`". This encourages the model to emit the `Answer` token at that intermediate stage, preventing incomplete or excessively long continuations.

After obtaining the logits for the first token following this prompt, we convert them into a probability distribution using the *first-token probability* method. This approach allows us to quantify the model's intermediate answer distribution at each 10% reasoning checkpoint, providing a fine-grained view of decision-making evolution along the CoT.

F Robustness Check for the Softmax Transformation

Using logits (or first-token probabilities) as confidence proxies is standard in prior work on LLM uncertainty and MCQA evaluation (Chen et al., 2024; Dong et al., 2026). Because first-token scores are deterministic given the input and unaffected by sampling hyperparameters, they provide a stable signal of model preference, which our work follows.

Our evaluation focuses on the relative allocation of probability mass across options (Human Label Variation). *We note that softmax introduces an exponential rescaling that may change the entropy of the original score distribution, while entropy is an important factor in uncertainty evaluation*. To better preserve the relative allocation structure across options and to maintain a more pronounced difference, we therefore adopt linear normalization over the candidate option tokens (A/B/C).

Importantly, in our MCQA setting we empirically observe that the logits at the A/B/C positions are consistently positive and well separated across all evaluated reasoning models and prompts—typically above 10. This avoids potential sign issues in linear normalization. We believe this property is partly due to our strongly constrained output format. Under these conditions, linear normalization yields valid non-negative vectors that sum to 1.

We also consider that softmax is a reasonable alternative and include it as an additional robustness check here. Robust check results (MNLI Step-ANOVA with softmax (t=10) JSD, Table 5) below lead to the same conclusions as the linear-normalization JSD results reported in Table 3.

Step	LLM (%)	CoT (%)	Residual (%)
0	100.0	0.0	0.0
1	97.9	0.4	1.7
2	96.7	0.6	2.6
3	95.6	1.3	3.2
4	94.3	1.9	3.8
5	92.8	2.3	4.9
6	91.4	3.1	5.4
7	90.4	3.2	6.4
8	88.8	3.8	7.4
9	86.3	5.6	8.1
10	77.1	13.3	9.6

Table 5: Robust check for step-wise ANOVA on MNLI Softmax (t=10) JSD.

G All Heatmaps for Step-wise Evaluation

This section presents the full set of heatmaps obtained from our step-wise evaluation, which are subsequently used to compute the ANOVA effect sizes reported in Table 3. Specifically, we provide heatmaps for **accuracy** (Figure 5), **JSD** (Figure 6), **Spearman's** ρ (Figure 7).

H All Curves for Step-wise Evaluation

This section presents all curves obtained from the step-wise evaluation, providing a complementary perspective to the heatmaps. Specifically, we show the progression of **accuracy** (Figure 8), **JSD** 9, and **Spearman's** ρ (Figure 10) across inference steps. These curves allow us to track how each metric evolves throughout the reasoning process, revealing dynamic trends in model performance, distributional alignment, and rank correlation over time.

I Analysis of CoT Formats

To further investigate the "split influence" observed in our quantitative results—where CoT determines the final accuracy but leaves the distributional structure (JSD and Spearman's ρ) largely anchored to model priors—we conducted an examination of the generated reasoning traces.

As discussed in Section 4.2, step-wise analysis reveals that accuracy often shifts or converges only at the conclusion, while Spearman's ρ fluctuates without a consistent trend. Thus, CoT determines the LLM's final choice but not the structure of uncertainty. We attribute this to the structural format of CoT: reasoning traces typically culminate in explicit, decisive conclusion statements (e.g., "Therefore, the answer is..."), which strongly steer the accuracy in the final steps.

Table 6 presents anecdotal evidence from the αNLI dataset. We observe a consistent pattern across models:

- **Explicit Conclusion at the End:** The reasoning process consistently ends with a strong, definitive statement identifying the correct option (e.g., "So, A must be the correct choice"). This explicit signal aligns with the sharp rise in accuracy observed in the final steps of our step-wise analysis.
- **Implicit Distributional Weighing:** While the models argue *for* the best option, they rarely explicitly articulate the relative probability of the alternative options in a way that would restructure the output distribution. Consequently, while the final decision is explicitly dictated by the CoT's conclusion, the distributional structure over non-argmax options remains latent and implicit, governed largely by the model's intrinsic priors.

Model & Case	Reasoning Excerpt (Conclusion Phase)	Observation regarding Final Choice vs. Structure
Qwen (Instance 2: Sandy)	"...Therefore, A is better. I think B is a distractor. [...] The instruction says 'select ONE of the listed options'... **So, my response should be just 'A'.**"	**Decisive Locking:** The reasoning concludes by explicitly discarding the alternative and locking onto the single target token 'A', driving the final accuracy without refining the relative probability space.
GLM (Instance 3: Bananas)	"...Therefore, B cannot explain why they're talking about eating a banana. **So, A must be the correct choice.** [...] **Therefore, A is correct.**"	**Convergence to Argmax:** The trace culminates in strong assertions ("must be", "correct"), which serve to fix the model's final decision, explaining why accuracy converges at the end while latent distributions remain implicit.
GPT (Instance 1: Ron)	"...That suggests his actions. So A is more appropriate. **Thus choose A.** [...] We must start answer with a single letter... **So final answer: 'A'.**"	**Explicit Selection:** The reasoning shifts from semantic evaluation to an operational selection command ("Thus choose A"), confirming that the CoT acts as a decision-maker for the top option.

Table 6: Examples of CoT reasoning traces from the αNLI dataset. The excerpts illustrate how CoT reasoning typically ends with explicit conclusion statements. This structural characteristic supports our finding that CoT content determines the final choice (Accuracy) through explicit reasoning, while the underlying structure of uncertainty (JSD/Ranking) remains latent and less affected by these definitive concluding remarks.

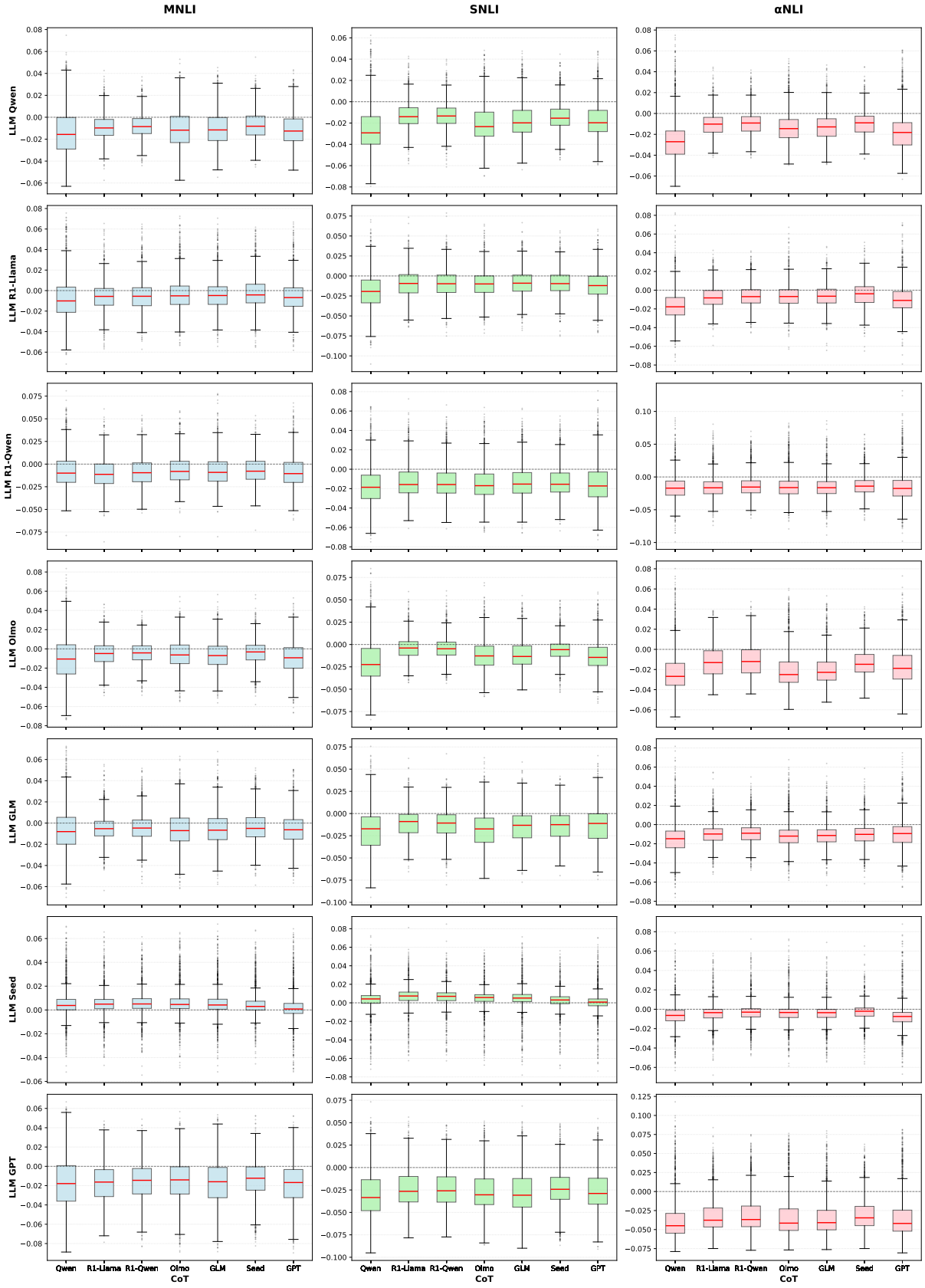


Figure 4: Delta JSD box plot.

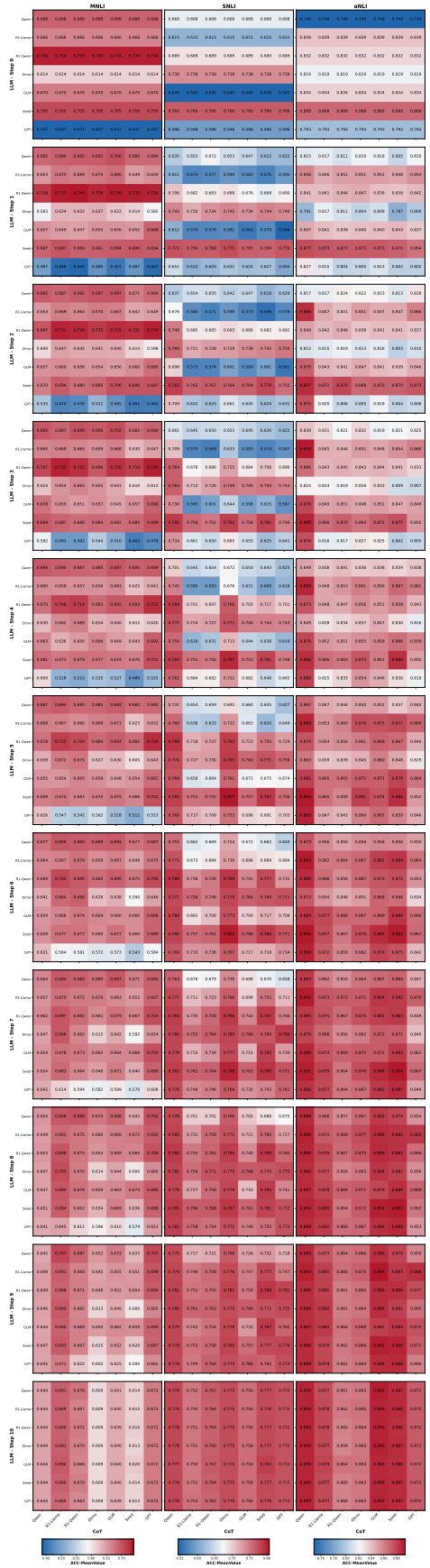


Figure 5: Steps ACC.

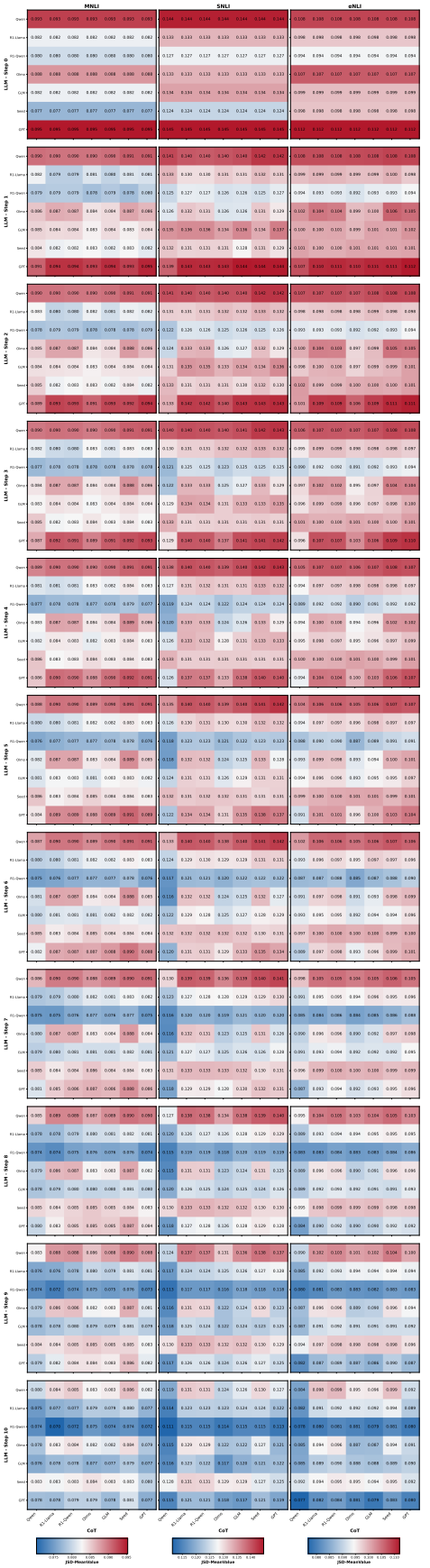


Figure 6: Steps JSD.

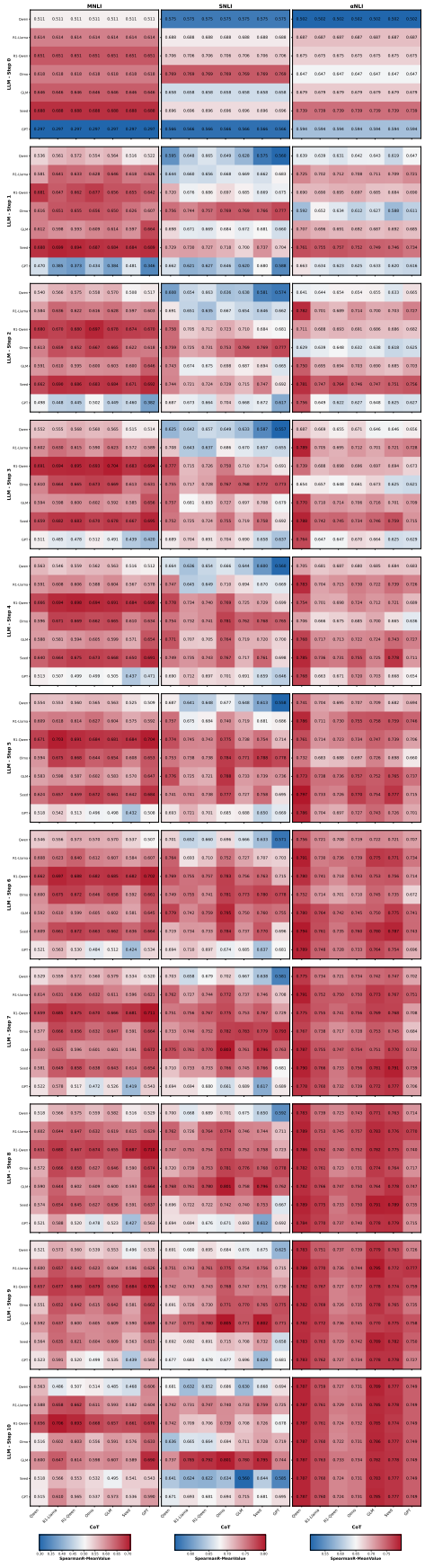


Figure 7: Steps Spearman's ρ.

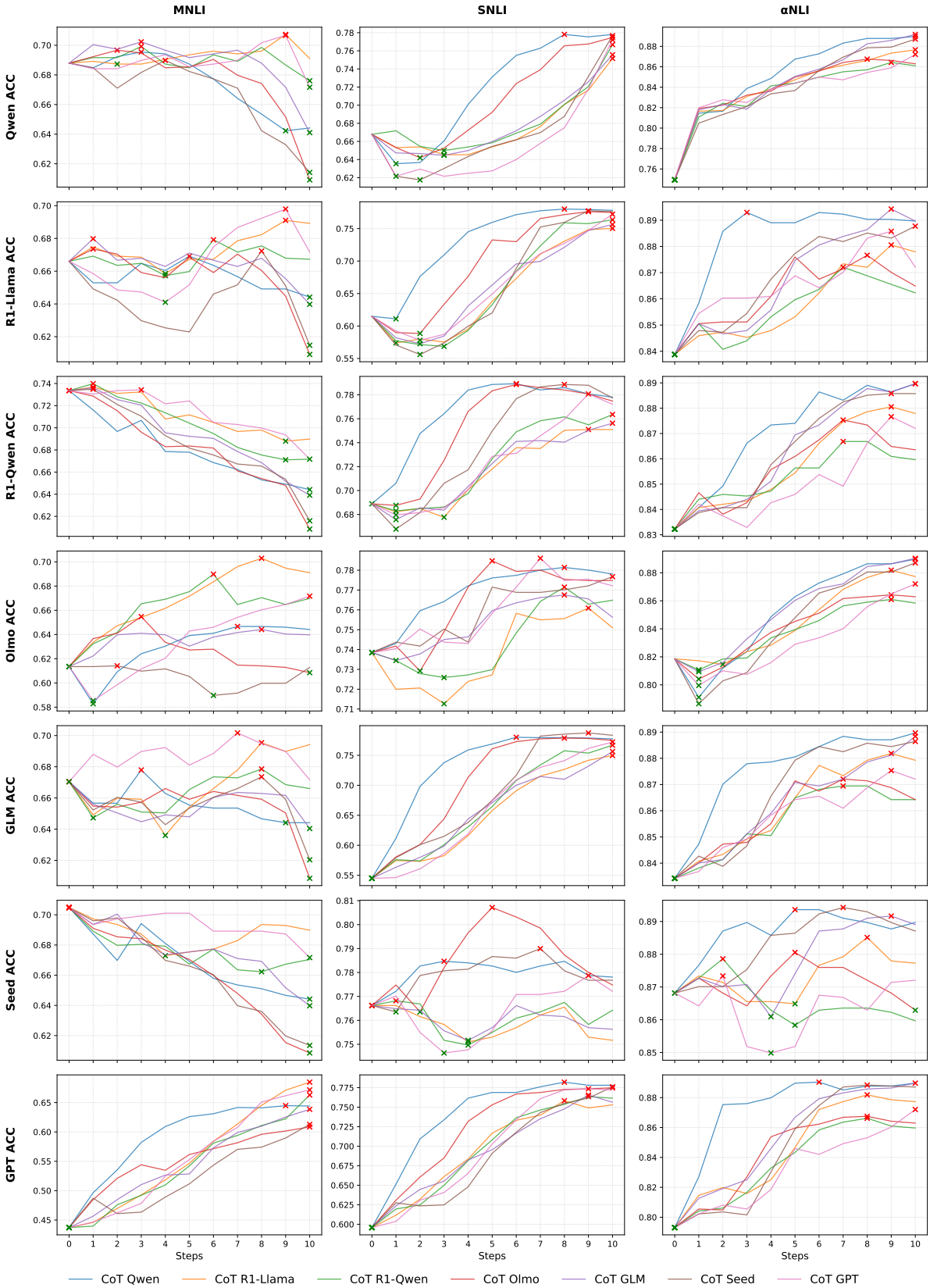


Figure 8: Curves ACC.

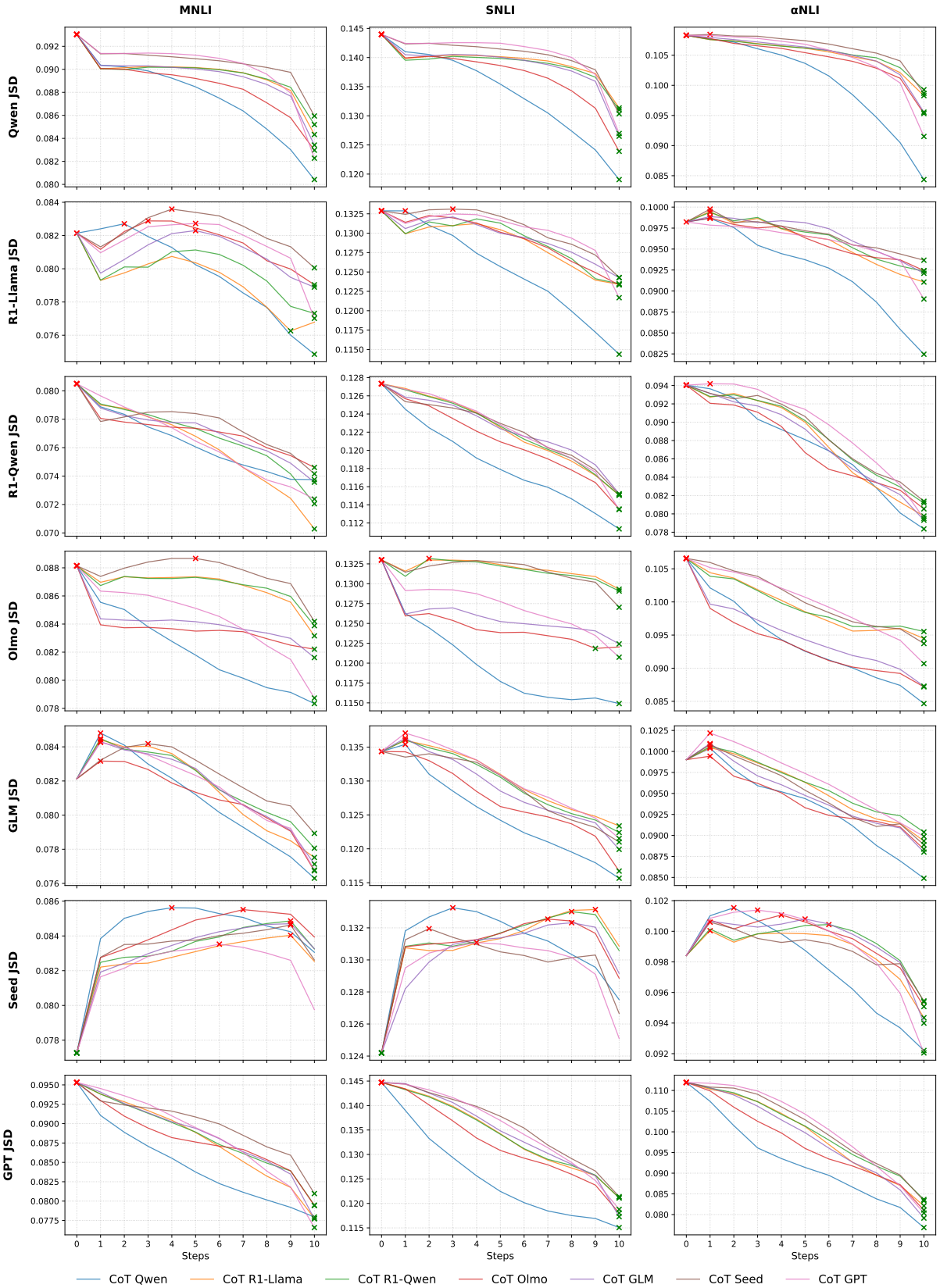


Figure 9: Curves JSD.

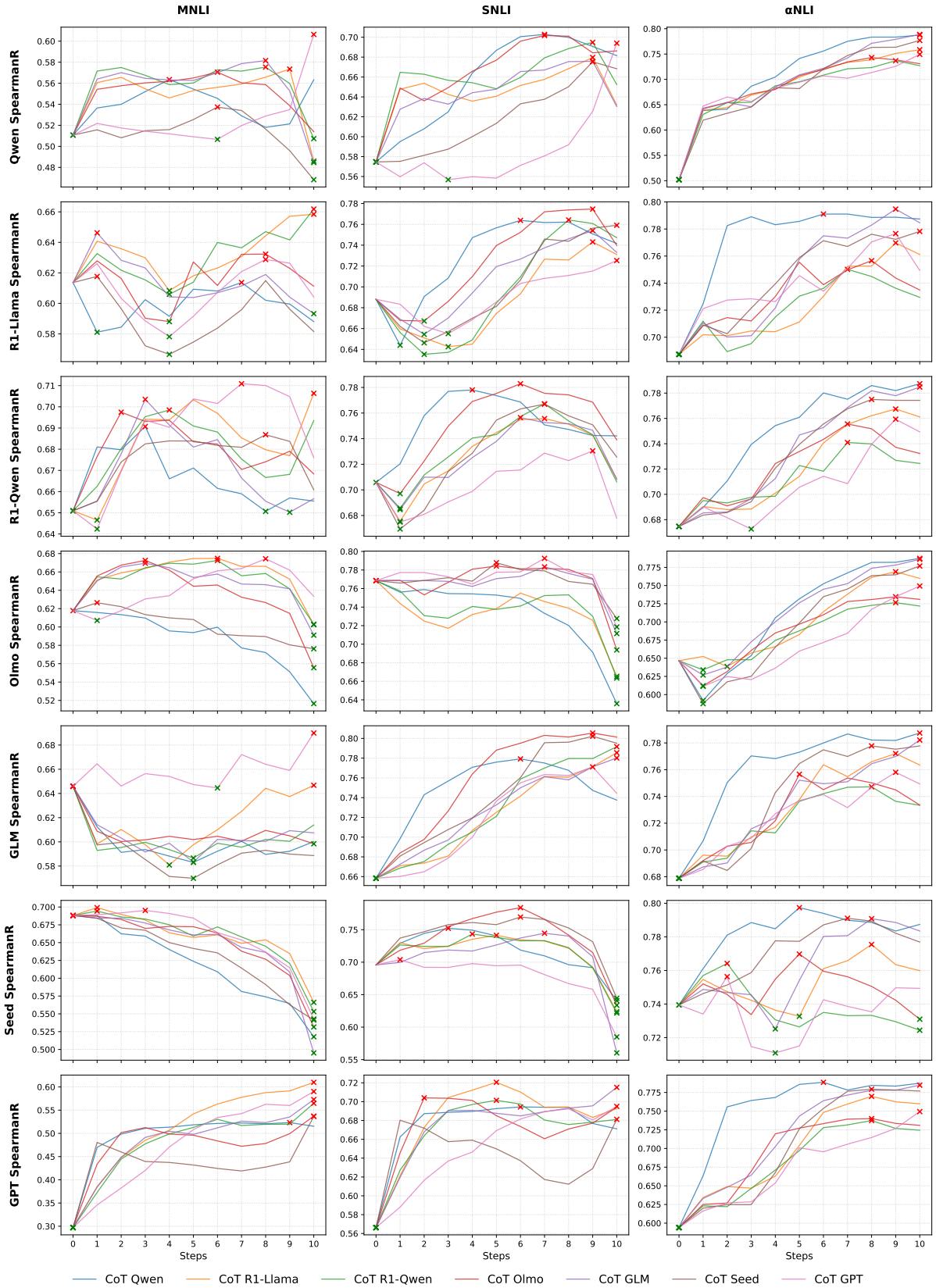


Figure 10: Curves Spearman's ρ.