

HIGHER-ORDER COGNITIVE CHAIN-OF-THOUGHT IS ENOUGH: EVALUATION, ANALYSIS, AND OPTIMIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Chain-of-Thought (CoT) data has become essential for advancing large language models’ reasoning capabilities, yet current quality assessment methods neglect the quality of underlying reasoning processes and thus undermine their effectiveness. To address these challenges, we propose a CoT data quality assessment framework from a cognitive perspective, grounded in Bloom’s Taxonomy as our core theoretical foundation. Through systematic analysis of existing CoT datasets, we reveal that current CoT data exhibits significant distributional biases toward intermediate-order cognitive operations, failing to adequately represent the full spectrum of human-level cognitive capabilities. These findings demonstrate systematic inadequacies in reasoning quality across multiple benchmarks, with models struggling to reproduce sophisticated cognitive processes essential for complex problem-solving. Based on these insights, we propose a simple-yet-effective cognitive-guided CoT data enhancement approach that supplements datasets with minimal higher-order cognitive CoT data. Consequently, we introduce a simple-yet-effective CoT data enhancement method that rapidly enhances model performance using minimal additional high-order cognitive CoT data, experiments demonstrates the effectiveness of cognitive-aware CoT dataset construction and evaluation.

1 INTRODUCTION

Chain-of-Thought (CoT) data plays a pivotal role in advancing large language models (LLMs). Besides CoT data significantly enhances the accuracy of complex problem-solving by providing structured reasoning examples (Wei et al., 2022; Zhang et al., 2023; Wei et al., 2022; Suzgun et al., 2023), it effectively induces models to develop sophisticated reasoning capabilities, especially in reinforcement learning (RL) based models like Deepseek-R1 (DeepSeek-AI et al., 2025). By incorporating CoT data, RL-based reasoning LLMs can learn to break down complex problems into manageable steps (DeepSeek-AI et al., 2025; Wen et al., 2025), which not only boosts performance but also enhances the interpretability of the model’s decision-making process, making CoT data invaluable for improving the overall effectiveness and trustworthiness of these advanced LLMs. Hence, constructing and employing CoT datasets becomes an important task (Zhang et al., 2023; Suzgun et al., 2023; Yin et al., 2025).

Researchers have developed various CoT data construction methodologies, primarily encompassing two technical approaches: direct utilization of pre-existing CoT datasets and acquisition via knowledge distillation (Wang et al., 2023a; Chen et al., 2023). Direct utilization typically involves manual annotation, wherein domain experts meticulously construct high-quality reasoning chains (Li et al., 2023a). Conversely, knowledge distillation leverages large pre-trained language models to automatically generate CoT data. To ensure CoT data quality, these methods introduce additional verification mechanisms to filter out low-quality CoT data, such as teacher-generated rationales with post-hoc filtering (Li et al., 2023b), self-distillation via self-consistency that retains rationales yielding consistent answers (Wang et al., 2023b), and verification-based distillation that employs explicit verifiers to validate intermediate reasoning steps (Ling et al., 2023).

However, despite the proliferation of CoT construction efforts, we argue that current CoT datasets suffer from fundamental quality assessment challenges that undermine their effectiveness. Current research exhibits a pronounced trend toward scalability, with newly constructed CoT datasets not only growing in data volume but also in length (Wang et al., 2023b; Fang et al., 2025; Golovneva et al., 2023; Nguyen et al., 2024). This “bigger and longer” development strategy lacks robust quality assessment foundations, implying three fundamental limitations. **(1) Existing evaluations ignore the quality of reasoning processes.** Existing metrics concentrating exclusively on length and accuracy cannot adequately characterize the semantic and cognitive properties intrinsic to reasoning chains; they induce fundamental gaps in assessing the cognitive essence of the reasoning processes Prasad et al. (2023). **(2) Existing evaluations may introduce systematic biases.** Longer CoT frequently exhibits extraneous content, erroneous logical steps, or task-irrelevant information, whereas cognitively sound reasoning is characteristically concise and demonstrates direct problem-solving efficacy (Ling et al., 2023; Li et al., 2023a; Wang et al., 2023b). **(3) Existing evaluation metrics fail to reflect data quality defects in terms of model performance.** Emergent empirical evidence indicates that neither length nor scale of CoT data correlates with substantive reasoning capability gains Kejriwal et al. (2024); Sylolypavan et al. (2023).

To address these limitations, cognitive theories can be leveraged for CoT quality assessment. First, cognitive theory inherently prioritizes the quality of reasoning processes. Since CoT data fundamentally aligns with human cognitive operations, cognitive theory provides normative frameworks for assessing the quality of intermediate reasoning operations (Wei et al., 2022; Nguyen et al., 2024). Second, cognitive theory emphasizes the depth of processing, establishing normative quality benchmarks that mitigate quality degradation resulting from superficial pursuit of length or scale metrics, thereby ensuring evaluation criteria correspond to authentic cognitive rigor. Finally, cognitive science has established robust causal links between hierarchical cognitive operations and performance efficacy, providing mechanistic foundations for fine-grained diagnostic evaluations of model reasoning capabilities.

Specifically, we employ Bloom’s Taxonomy Bloom et al. (1956) as our core cognitive framework, a theory that categorizes cognitive abilities into six cognitive levels (remembering, understanding, applying, analyzing, evaluating, and creating), which is particularly suited for systematically parsing the types of cognitive operations within CoT. We have developed methods to automatically map CoT steps to Bloom’s Taxonomy levels and constructed an innovative evaluation metric system based on this mapping, focusing on four key research questions:

RQ1: How well do current CoT datasets represent compared to human cognitive processes?

RQ2: Can LLMs reproduce human-level cognitive processes?

RQ3: What are the characteristics of the cognitive processes exhibited by LLMs?

RQ4: How can CoT datasets be optimized to enhance their cognitive expressiveness?

The main contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to systematically investigate CoT data quality assessment from a cognitive perspective, establishing a novel evaluation paradigm grounded in cognitive principles.
- We develop a comprehensive evaluation framework anchored in Bloom’s Taxonomy that enables precise quantification of cognitive complexity within CoT reasoning sequences, providing both accuracy and computational efficiency.
- We conduct extensive experiments to analyze the quality of existing CoT datasets and knowledge-distilled CoT data, revealing systematic biases and cognitive inadequacies that compromise reasoning quality across multiple benchmarks.
- We propose a simple-yet-effective cognitive-guided approach for CoT data enhancement, with experimental validation demonstrating substantial improvements in model reasoning capabilities using little CoT data.

2 RELATED WORK

CoT Datasets Recent efforts in CoT dataset construction have emphasized creating coherent reasoning chains across diverse tasks and domains, thereby enhancing the interpretability and logical capabilities of LLMs. Early large-scale efforts, such as ThoughtSource(Ott et al., 2023), aggregate CoT examples from multiple domains, including science, mathematics, medicine, and general question answering, while standardizing formats and providing tools for evaluation, fine-tuning, and reasoning analysis. Other works such as LogiCoT(Liu et al., 2023) and OmniThought(Cai et al., 2025), introduce large-scale CoT dataset and show the effectiveness of them. Mathematics has emerged as a central domain for CoT research, with datasets(Yu et al., 2024; Ling et al., 2017; Amini et al., 2019; Patel et al., 2024; Mitra et al., 2024) providing complementary resources with varying scale, annotation format, and reasoning granularity, supporting tasks such as symbolic manipulation, algebraic reasoning, and multi-hop deduction., and reverse reasoning to enhance existing benchmarks (Cobbe et al., 2021; Lightman et al., 2024). Generally, CoT datasets have progressed toward broader domain coverage and more sophisticated reasoning processes.

CoT Data Evaluation Recent research has increasingly focused on assessing the quality of CoT data. Numerous studies have attempted to quantify and evaluate the complexity of CoT data through various approaches. For instance, Budagam et al. (2024) analyzed CoT data complexity using statistical indicators such as length, entropy, and logical leaps, while Li et al. (2024) introduced structural entropy and cognitive tree representations. Alternative methods have examined reasoning complexity through semantic stability (Yao et al., 2023) and logical consistency (Xu et al., 2025), employing embedding trajectory geometry and logical template matching as key analytical tools. Additionally, Wang et al. (2024) framed model generation as progressive cognitive unfolding, introducing the Local Disbalance Rate metric to capture the interplay between cognitive complexity and generative behavior. Notably, research has revealed that increasing reasoning chain length or depth does not guarantee improved performance (Lee et al., 2025; Kang et al., 2024). Overall, existing CoT data evaluation remains predominantly focused on analyzing surface-level complexity of reasoning chains, lacking deeper insights into their underlying cognitive capabilities.

3 METHODOLOGY

3.1 BLOOM’S TAXONOMY FRAMEWORK

Bloom’s Taxonomy serves as our theoretical framework for analyzing the quality of CoT datasets. Bloom’s Taxonomy, first introduced by Bloom et al. (1956) and Anderson & Krathwohl (2001), provides a hierarchical classification of cognitive processes encompassing six levels: Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating, alongside four knowledge dimensions: Factual, Conceptual, Procedural, and Metacognitive.

We selected Bloom’s Taxonomy as our theoretical foundation for three reasons. (1) Bloom’s Taxonomy describes cognitive operational processes, which perfectly align with the core requirement of CoT data assessment, to reveal the connection between reasoning processes and quality. (2) Bloom’s Taxonomy provides a structured framework with dual-dimensional structure (cognitive processes and knowledge types), enabling precise characterization of the cognitive-knowledge coupling patterns, thereby offering an interpretable mechanism for model performance. (3) Bloom’s Taxonomy possesses a comprehensive conceptual system and standardized assessment dimensions that ensure both theoretical scientific reliability and operational standards convertible to automated evaluation.

3.2 PROBLEM DEFINITION

To systematically capture the cognitive structure of reasoning in CoT data, we formalize a taxonomy-based annotation scheme. This scheme leverages Bloom’s Taxonomy to evaluate both the inherent cognitive difficulty of the instructions and the step-wise cognitive progression of reasoning. By representing CoT annotation through a formal mathematical expression, we can explicitly define the procedure and goals of the annotation task, thereby further enabling a clear and rigorous characterization of both the task itself and the intended annotation outcomes. The set of cognitive levels and

knowledge levels in Bloom’s Taxonomy are denoted as:

$$\begin{aligned}\mathcal{C} &= \{\text{REMEMBERING, UNDERSTANDING, APPLYING, ANALYZING, EVALUATING, CREATING}\} \\ \mathcal{K} &= \{\text{FACTUAL, CONCEPTUAL, PROCEDURAL, METACOGNITIVE}\}.\end{aligned}$$

Consider giving a CoT instance $x = (Q, CoT, A)$, where Q denotes the input question, $CoT = (s_1, s_2, \dots, s_n)$ is the generated reasoning chain consisting of intermediate steps s_i , A is the final answer. Then, we present two cognitive annotation tasks with different granularities: **instruction and trajectory cognitive annotation**.

Definition 3.1 (Instruction Cognitive Annotation) *The **instruction cognitive annotation** assigns a single cognitive label $c^* \in \mathcal{C}$ and a knowledge label $k^* \in \mathcal{K}$ to a CoT instance, reflecting the highest cognitive level required to correctly respond to the given instruction. Formally, given the CoT, question Q , answer A , and the annotation prompt P_I , the label is defined as*

$$L_I(CoT \mid Q, A, P_I) = (c^*, k^*, t_I^*), \quad (1)$$

where t_I^* represents the reasoning process used in the annotation, capturing the annotator’s thought process in determining the highest cognitive level required to answer the instruction.

For example, a simple factual recall problem may be labeled as REMEMBERING, whereas a multi-step logical deduction problem may require ANALYZING or higher.

Definition 3.2 (Cognitive Trajectory Annotation) *The **cognitive trajectory annotation** captures the cognitive labels of individual reasoning steps within a CoT. Each reasoning step $s_i \in CoT$ is paired with a cognitive label $c_i \in \mathcal{C}$, producing a cognitive trajectory $(c_1 \rightarrow c_2 \rightarrow \dots \rightarrow c_n)$ that aligns one-to-one with the reasoning steps within the annotation prompt P_T :*

$$L_T(CoT \mid Q, A, P_T) = (\tau, t_T^*), \quad (2)$$

$$\tau = (c_1 \rightarrow c_2 \rightarrow \dots \rightarrow c_n), \quad (3)$$

where t_T^* denotes the annotator’s thought process followed during annotation, reflecting how each step s_i was analyzed to determine its cognitive label c_i .

The proposed annotation scheme not only formalizes cognitive labeling, but also enables structural analysis of CoT sequences from two complementary perspectives:

Depth The highest cognitive level reached in the reasoning chain, reflecting overall problem difficulty;

Progression The coherence of cognitive steps across the chain, indicating whether reasoning follows a consistent trajectory or shows regressions/redundancies.

Together, these two perspectives support a fine-grained evaluation of reasoning quality, considering both the sophistication and the structural integrity of the thought process.

3.3 COGNITIVE ANNOTATION WITH LLMs

We adopt *Qwen2.5-72B-Instruct* as the primary annotator for cognitive-level labeling, inspired by the strong evaluation results of *CompassJugder-32B-Instruct* (Cao et al., 2024). Specifically, we use the instruction-tuned *Qwen2.5-72B-Instruct* as a few-shot annotator for both model-generated and human-written CoTs. The annotation prompts are constructed by combining concise operational definitions of Bloom’s cognitive processes and knowledge dimensions with representative examples, thereby grounding the labeling in explicit decision rules.

To assess reliability, we randomly select 1,000 CoT instances, and engage three cognitive science experts for independent annotation. We compute Cohen’s Kappa (κ) against three experts, and the results are shown in Table 1. ***Qwen2.5-72B-Instruct* achieves the highest agreement across both knowledge and cognitive dimensions, with $\kappa > 0.88$ indicating near-human consistency and demonstrating the effectiveness of the model-based annotation pipeline.** The consistency evaluation result also demonstrates the effectiveness of model-based annotation, details are shown in Appendix 7.

Table 1: Cohen’s Kappa (κ) consistency scores computed on 1,000 sampled CoT instances, comparing multiple LLM annotators with three experts.

	Knowledge Dimension				Cognitive Dimension			
	Expert 1	Expert 2	Expert 3	Average	Expert 1	Expert 2	Expert 3	Average
Qwen2.5-32B-Instruct	0.9758	0.9801	0.9700	0.9753	0.8533	0.8718	0.8649	0.8633
Qwen3-32B	0.9646	0.9707	0.9609	0.9654	0.7811	0.7754	0.7811	0.7792
CompassJudge-2-32B-Instruct	0.9818	0.9790	0.9794	0.9801	0.8383	0.8517	0.8424	0.8441
Qwen2.5-72B-Instruct	0.9874	0.9874	0.9868	0.9872	0.8820	0.8832	0.8914	0.8855
ChatGPT4-mini	0.9646	0.9707	0.9609	0.9654	0.7811	0.7754	0.7811	0.7792

3.4 METRICS

For quantifying the CoT data quality of a dataset, we define $\bar{V}^c$ and $\bar{V}^k$ as the mean values for the cognitive and knowledge dimensions, respectively, and denote ϵ^c and ϵ^k as their respective standard deviations. Specifically, following Bloom’s Taxonomy hierarchy from low to high, we annotate each CoT instance with cognitive dimension values from 1 to 6, where 1 denotes the lowest cognitive level (Remembering) and 6 represents the highest level (Creating). Similarly, we annotate the knowledge dimension values from 1 to 4, ranging from lower-order to higher-order knowledge types.

We further introduce a new metric **Bloom’s Taxonomy Weighted Level Index** W_c and W_k in order to quantify the distance between each cognitive and knowledge dimension’s distribution and the uniform distribution. Specifically, we define $W_c = \min(N_C \cdot p_c, 1)$, $W_k = \min(N_K \cdot p_k, 1)$, where $N_C = 4$, $N_K = 6$ are the number of dimensions of cognitive and knowledge, respectively, and p_c, p_k are its relative frequency. The min function ensures the metric to 1 for dimensions with $p_k \leq \frac{1}{N_K}$ and $p_c \leq \frac{1}{N_C}$.

4 EXPERIMENT

To evaluate and improve the quality of existing CoT datasets, we conduct a series of experiments to answer the following research questions:

RQ1: How well do current CoT datasets represent compared to human cognitive processes?

RQ2: Can LLMs reproduce human-level cognitive processes?

RQ3: What are the characteristics of the cognitive processes exhibited by LLMs?

RQ4: How can CoT dataset be optimized to enhance their cognitive expressiveness?

4.1 EXPERIMENT SETUP

Datasets We evaluate the quality of 10 different widely-used CoT datasets (Cobbe et al., 2021; Mahdavi et al., 2025; Albalak et al., 2025; Lu et al., 2024; Gao et al., 2025; Hendrycks et al., 2021; Ling et al., 2017), which can be categorized into 2 groups according to their usage. **(1) Human-authored CoT datasets.** We assembled a diverse collection of human-written CoT corpora. The datasets cover a wide range of mathematical reasoning tasks, from standard exercises to Olympiad-level problems, and constitute the primary source of high-quality human CoTs that capture natural cognitive processes. **(2) LLM-distilled CoT datasets.** We collected distilled CoTs from state-of-the-art models. Distillation was performed on representative benchmarks selected for their domain coverage and ability to reflect realistic model reasoning. Details of the CoT datasets are shown in Appendix (Table 2 and Table 3).

Settings Our experiments are conducted on 4 accelerated GPUs, and we use PyTorch 2.6 in Python 3.11. We set the maximum sequence length for both input and output sequences to 1536 tokens.

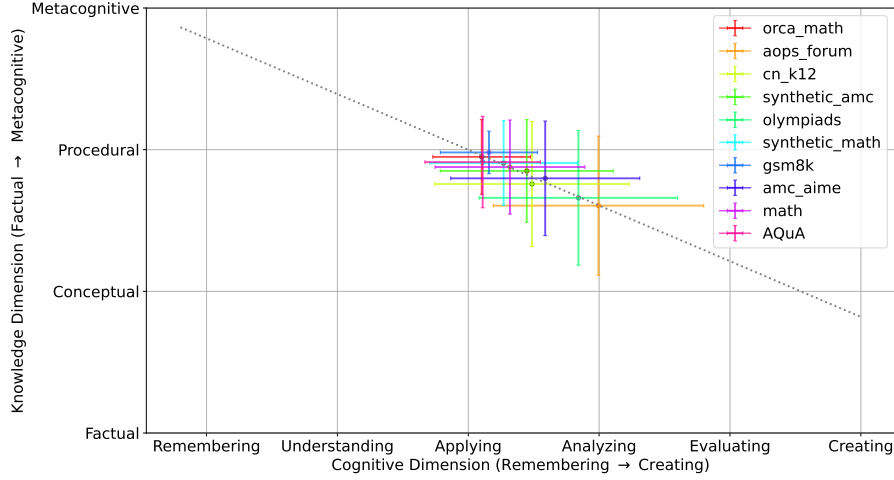


Figure 1: The distribution of current CoT data across Bloom’s two dimensions.

4.2 ANALYSIS

4.2.1 RQ1: HOW WELL DO CURRENT CoT DATASETS REPRESENT COMPARED TO HUMAN COGNITIVE PROCESSES?

Observation 1: Current CoT datasets demonstrate systematic bias toward intermediate-order cognitive processes and knowledge types. Figure 1 illustrates that most CoTs fall between Applying and Analyzing, indicating an emphasis on intermediate-order cognitive reasoning. However, current CoT datasets exhibit marked deficiency in higher-order cognitive operations (Evaluating, Creating). As shown in Figure 2a, a minimal proportion of CoT data operates at higher-order cognitive levels (Evaluating and Creating). Similarly, Figure 1 reveals a pronounced concentration of existing CoT datasets on conceptual and procedural knowledge types. Figure 2b further demonstrates that the vast majority of reasoning chains operate at the conceptual and procedural knowledge level, higher-order knowledge (metacognitive) remains markedly underrepresented across all evaluated datasets.

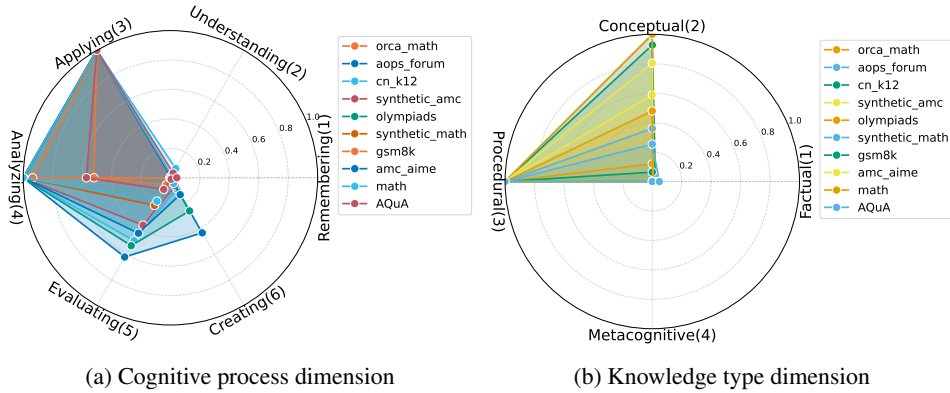


Figure 2: Distribution of cognitive process and knowledge types in CoT datasets.

Observation 2: Cognitive-Knowledge dimensional misalignment. Existing CoT datasets exhibit systematic mismatches between cognitive and knowledge dimensions that contradict human learning patterns. While optimal cognitive development follows a hierarchical progression where cognitive complexity increases with knowledge abstraction (Krathwohl, 2002), Figure 1 shows anomalous pairings, notably Applying-Procedural and Analyzing-Conceptual, which violate this fundamental cognitive-pedagogical principle.

Highlights 1

Current CoT datasets suffer from severe distributional imbalance and systematic inconsistency with human cognitive patterns.

4.2.2 RQ2: CAN LLMs REPRODUCE HUMAN-LEVEL COGNITIVE PROCESSES?

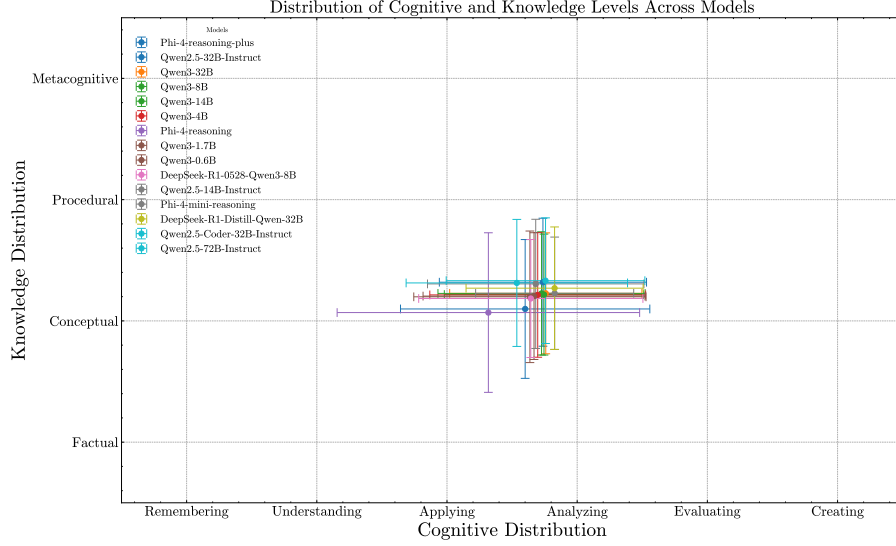


Figure 3: The distribution of distilled CoT data across Bloom’s two dimensions.

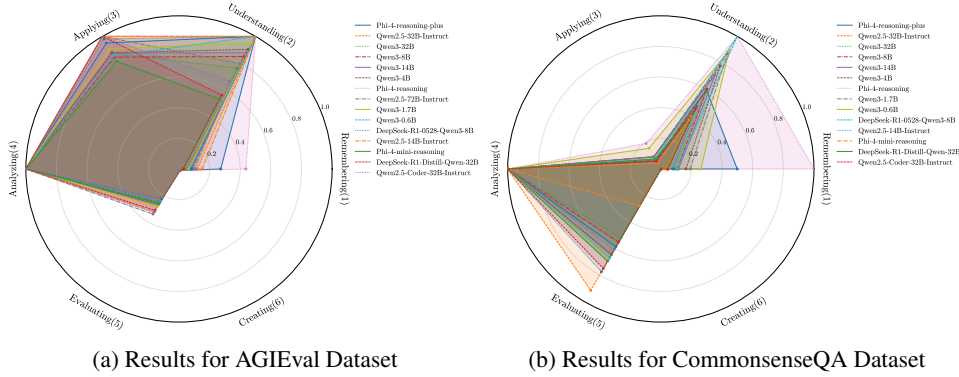


Figure 4: The details for the LLM’s distilled CoT results.

Observation 3: Dimensional Concentration. Consistent with existing CoT datasets, distilled CoT data exhibits systematic concentration in intermediate cognitive processes and knowledge dimensions, replicating the distributional bias observed in current CoT datasets. As shown in Figure 3, distilled CoTs concentrate in Applying and Analyzing operations, with sparse representation in both lower-order (Remembering, Understanding) and higher-order (Evaluating, Creating) operations. In the knowledge dimension, most models rely on Conceptual and Procedural knowledge, rather than factual recall or metacognitive reasoning.

Observation 4: Task-Specific Diversity. Distilled CoT data demonstrates exceptionally high diversity with pronounced distributional inconsistency across different task domains, indicating significant task-dependent variation in cognitive-knowledge patterns. We show the distribution across four different task categories in Figure 4, including professional exams, commonsense reasoning, mathematics, and code reasoning. In AGIEval and CruxEval, reasoning concentrates on Understand-

ing and Applying, with Analyzing as a secondary component. CommonsenseQA displays higher proportions of Analyzing and Evaluating, reflecting increased critical assessment demands, while Omni-MATH emphasizes Applying and Analyzing, highlighting procedural and problem-solving reasoning. Detailed experimental results are shown in Appendix (Figure 7).

Observation 5: Human-Aligned Coupling. Figure 3 shows that distilled CoT data achieves alignment with human cognitive patterns through proper dimensional coupling: lower-order cognitive processes consistently pair with lower-order knowledge types, while higher-order cognitive operations align with higher-order knowledge domains.

Highlights 2

While distilled CoT data exhibits high diversity and consistency with human cognition, its pronounced distributional imbalance critically constrains overall data quality.

4.2.3 RQ3: WHAT ARE THE CHARACTERISTICS OF THE COGNITIVE PROCESSES EXHIBITED BY LLMs?

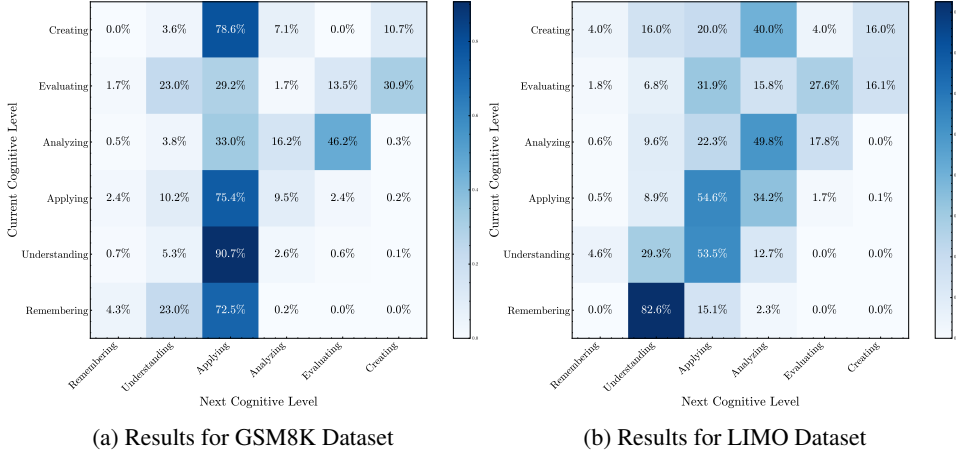


Figure 5: Cognitive trajectory annotation results.

Figure 5 presents cognitive trajectory matrices for 2 different types of datasets, where each cell represents the transition probability from one cognitive level to the next in CoT reasoning, as annotated using Bloom’s Taxonomy. This visualization captures the sequential cognitive dynamics of model-generated reasoning, showing how steps evolve across tasks. More detailed results are shown in Appendix (Figure 9).

Observation 6: Cognitive Hierarchy Compliance. All datasets demonstrate cognitive transitions that adhere to human cognitive hierarchies, with lower-order cognitive processes (Remembering/Understanding) predominantly transitioning to adjacent or higher-level cognitive processes, while reverse transitions (e.g., from Analyzing back to Remembering) occur infrequently, confirming alignment with established cognitive progression patterns (Wei et al., 2022; Huber & Niklaus, 2025).

Observation 7: Task-Specific Transition Characteristics. Different datasets exhibit distinct cognitive transition signatures reflecting their underlying task nature: GSM8K shows pronounced Understanding-Applying transitions (90.7%), indicative of algorithmic problem-solving characteristics; AQuA demonstrates strongest Understanding-Applying patterns (73.5%) with more distributed transitions, reflecting quantitative reasoning complexity; while LIMO exhibits strong Remembering-Understanding transitions (82.6%) coupled with notable Analyzing-Analyzing self-loops (49.8%), characteristic of iterative logical reasoning processes.

Highlights 3

While distilled CoT data exhibits strong task-specificity and human-aligned cognitive transition hierarchies, it inevitably suffers from persistent distributional biases.

4.2.4 RQ4: HOW CAN THE COGNITIVE PROCESSES OF LLMs BE OPTIMIZED TO ENHANCE THEIR COGNITIVE EXPRESSIVENESS?

Figure 6 illustrates the performance trends of instruction fine-tuning on three representative benchmarks, including CMMLU, MMLU-Pro (Broad Domain Knowledge), CommonsenseQA (Commonsense Reasoning), and AGIEval (Professional Exams). We compare the effect of high-order CoT instructions (corresponding to ANALYZING, EVALUATING, CREATING in Bloom’s Taxonomy) with low-order CoT instructions (corresponding to REMEMBERING, UNDERSTANDING, APPLYING). We present the average accuracy of all evaluations in the figure.

The results show distinct trends: for high-order cognitive level CoTs, the accuracy initially increases as the dataset size grows, but gradually saturates after around 400–600 instructions. In contrast, low-order cognitive level CoTs exhibit an increase in accuracy at small scales, but performance begins to decline as more data are added. This indicates that fine-tuning with only low-order cognitive level CoTs cannot sustainably improve LLM reasoning ability. These findings suggest that **a relatively small but high-quality subset of high-order cognitive level CoTs (approximately 300 examples) is sufficient to boost LLM performance across diverse benchmarks**, whereas simply enlarging low-order cognitive level CoT datasets may even harm performance. This highlights that a small but cognitively rich set of instructions can serve as a “golden dataset” for instruction fine-tuning, offering an efficient strategy for enhancing reasoning capabilities in LLMs.

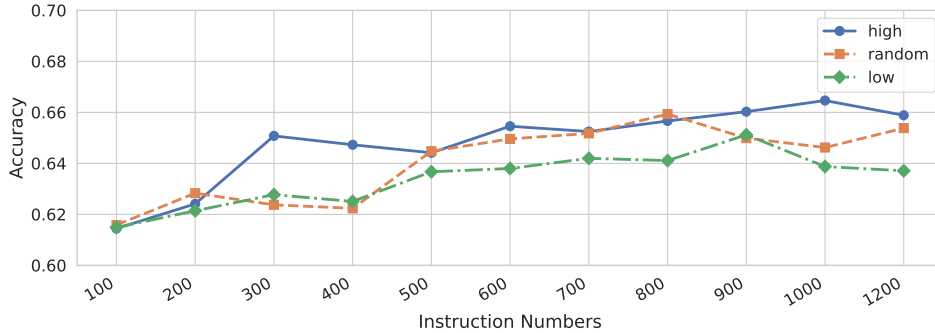


Figure 6: The fine-tuning results on Qwen2.5-7B-Base.

Highlights 4

Supplementing minimal amounts of high-order CoT data enables rapid model performance enhancement and stabilization.

5 CONCLUSION

In this paper, we systematically investigate CoT data quality assessment from a cognitive perspective and propose a comprehensive evaluation framework grounded in Bloom’s Taxonomy. Our analysis reveals that existing CoT datasets typically suffer from distributional biases that fail to adequately represent human-level cognitive capabilities. Consequently, we introduce a simple-yet-effective CoT data enhancement method that rapidly enhances model performance using minimal additional high-order cognitive CoT data. Future efforts must prioritize the creation and integration of higher-order and metacognitive CoT data to unlock the full potential of language models for creative and reflective tasks.

REFERENCES

- Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *CoRR*, abs/2502.17387, 2025. doi: 10.48550/ARXIV.2502.17387. URL <https://doi.org/10.48550/arXiv.2502.17387>.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 2357–2367. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1245. URL <https://doi.org/10.18653/v1/n19-1245>.
- Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.
- Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. Handbook i: cognitive domain. *New York: David McKay*, pp. 483–498, 1956.
- Devichand Budagam, Sankalp KJ, Ashutosh Kumar, Vinija Jain, and Aman Chadha. Hierarchical prompting taxonomy: A universal evaluation framework for large language models. *CoRR*, abs/2406.12644, 2024. doi: 10.48550/ARXIV.2406.12644. URL <https://doi.org/10.48550/arXiv.2406.12644>.
- Wenrui Cai, Chengyu Wang, Junbing Yan, Jun Huang, and Xiangzhong Fang. Reasoning with omnithought: A large cot dataset with verbosity and cognitive difficulty annotations. *CoRR*, abs/2505.10937, 2025. doi: 10.48550/ARXIV.2505.10937. URL <https://doi.org/10.48550/arXiv.2505.10937>.
- Maosong Cao, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. Compassjudge-1: All-in-one judge model helps model evaluation and evolution. *CoRR*, abs/2410.16256, 2024. doi: 10.48550/ARXIV.2410.16256. URL <https://doi.org/10.48550/arXiv.2410.16256>.
- Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. MCC-KD: multi-cot consistent knowledge distillation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6805–6820. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.454. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.454>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL <https://arxiv.org/abs/2110.14168>.
- 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, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,

- R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL <https://doi.org/10.48550/arXiv.2501.12948>.
- Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=fL4qWkSmtM>.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=yaqPf0KA1N>.
- Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=xYlJRpzZtsY>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021*. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>.
- Thomas Huber and Christina Niklaus. Lms meet bloom’s taxonomy: A cognitive view on large language model evaluations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 5211–5246. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.coling-main.350/>.
- Liwei Kang, Zirui Zhao, David Hsu, and Wee Sun Lee. On the empirical complexity of reasoning and planning in llms. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 2897–2936. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.164. URL <https://doi.org/10.18653/v1/2024.findings-emnlp.164>.
- Mayank Kejriwal, Henrique Santos, Ke Shen, Alice M. Mulvehill, and Deborah L. McGuinness. A noise audit of human-labeled benchmarks for machine commonsense reasoning. *Scientific Reports*, 14(1):8609, Apr 14 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-58937-4. URL <https://doi.org/10.1038/s41598-024-58937-4>.
- David R Krathwohl. A revision of bloom’s taxonomy: An overview. *Theory into practice*, 41(4): 212–218, 2002.
- Ayeong Lee, Ethan Che, and Tianyi Peng. How well do llms compress their own chain-of-thought? A token complexity approach. *CoRR*, abs/2503.01141, 2025. doi: 10.48550/ARXIV.2503.01141. URL <https://doi.org/10.48550/arXiv.2503.01141>.
- Changcheng Li, Xiangyu Wang, Qiuju Chen, Xiren Zhou, and Huanhuan Chen. MTMT: consolidating multiple thinking modes to form a thought tree for strengthening LLM. *CoRR*, abs/2412.03987, 2024. doi: 10.48550/ARXIV.2412.03987. URL <https://doi.org/10.48550/arXiv.2412.03987>.

- Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also “think” step-by-step. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 2665–2679. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-LONG.150. URL <https://doi.org/10.18653/v1/2023.acl-long.150>.
- Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also “think” step-by-step. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2665–2679, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.150. URL <https://aclanthology.org/2023.acl-long.150/>.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=v8L0pN6EOi>.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 158–167. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1015. URL <https://doi.org/10.18653/v1/P17-1015>.
- Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/72393bd47a35f5b3bee4c609e7bba733-Abstract-Conference.html.
- Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. Logi-cot: Logical chain-of-thought instruction tuning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 2908–2921. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.191. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.191>.
- Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Juntao Pan, Mingjie Zhan, and Hongsheng Li. MathGenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2732–2747, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.151. URL <https://aclanthology.org/2024.acl-long.151/>.
- Sadegh Mahdavi, Muchen Li, Kaiwen Liu, Christos Thrampoulidis, Leonid Sigal, and Renjie Liao. Leveraging online olympiad-level math problems for llms training and contamination-resistant evaluation. *CoRR*, abs/2501.14275, 2025. doi: 10.48550/ARXIV.2501.14275. URL <https://doi.org/10.48550/arXiv.2501.14275>.
- Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *CoRR*, abs/2402.14830, 2024. doi: 10.48550/ARXIV.2402.14830. URL <https://doi.org/10.48550/arXiv.2402.14830>.

- Thi Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 2862–2883. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.168. URL <https://doi.org/10.18653/v1/2024.findings-acl.168>.
- Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther, and Matthias Samwald. Thoughtsource: A central hub for large language model reasoning data. *CoRR*, abs/2301.11596, 2023. doi: 10.48550/ARXIV.2301.11596. URL <https://doi.org/10.48550/arXiv.2301.11596>.
- Bhrij Patel, Souradip Chakraborty, Wesley A. Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. AIME: AI system optimization via multiple LLM evaluators. *CoRR*, abs/2410.03131, 2024. doi: 10.48550/ARXIV.2410.03131. URL <https://doi.org/10.48550/arXiv.2410.03131>.
- Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 10066–10086. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.622. URL <https://doi.org/10.18653/v1/2023.emnlp-main.622>.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.824. URL <https://doi.org/10.18653/v1/2023.findings-acl.824>.
- Aneeta Sylolypavan, Derek H. Sleeman, Honghan Wu, and Malcolm Sim. The impact of inconsistent human annotations on AI driven clinical decision making. *npj Digit. Medicine*, 6, 2023. doi: 10.1038/S41746-023-00773-3. URL <https://doi.org/10.1038/s41746-023-00773-3>.
- Andrew R Thompson and Logan PO Lake. Relationship between learning approach, bloom’s taxonomy, and student performance in an undergraduate human anatomy course. *Advances in Health Sciences Education*, 28(4):1115–1130, 2023.
- Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. SCOTT: self-consistent chain-of-thought distillation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 5546–5558. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-LONG.304. URL <https://doi.org/10.18653/v1/2023.acl-long.304>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 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, 2023b. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- Yu Wang, Shiwan Zhao, Zhihu Wang, Heyuan Huang, Ming Fan, Yubo Zhang, Zhixing Wang, Haijun Wang, and Ting Liu. Strategic chain-of-thought: Guiding accurate reasoning in llms through strategy elicitation. *CoRR*, abs/2409.03271, 2024. doi: 10.48550/ARXIV.2409.03271. URL <https://doi.org/10.48550/arXiv.2409.03271>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language

- models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *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*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
- Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Tanglifu Tanglifu, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. Light-rl: Curriculum SFT, DPO and RL for long COT from scratch and beyond. In Georg Rehm and Yunyao Li (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 318–327, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.24. URL <https://aclanthology.org/2025.acl-industry.24/>.
- Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *CoRR*, abs/2502.18600, 2025. doi: 10.48550/ARXIV.2502.18600. URL <https://doi.org/10.48550/arXiv.2502.18600>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chuji Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiayi Yang, Jingren Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025a. doi: 10.48550/ARXIV.2505.09388. URL <https://doi.org/10.48550/arXiv.2505.09388>.
- An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report. *CoRR*, abs/2501.15383, 2025b. doi: 10.48550/ARXIV.2501.15383. URL <https://doi.org/10.48550/arXiv.2501.15383>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.
- Huifeng Yin, Yu Zhao, Minghao Wu, Xuanfan Ni, Bo Zeng, Huaiyu. wh Huaiyu. wh, Tianqi Shi, Liangying Shao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1 v2: Towards widening the distillation bottleneck for reasoning models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pp. 23506–23516. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.acl-long.1145/>.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=N8N0hgNDRt>.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=5NTt8GFjUHkr>.

6 DATA COLLECTIONS

To address different research questions, we compiled a variety of datasets from multiple domains. These datasets include human-constructed datasets, model-distilled datasets, and model fine-tuning datasets. Detailed examples of the datasets can be found in Table 2, Table 3, and Table 4.

Table 2: Human-authored Chain-of-Thought datasets.

Dataset	Domain	Scale	Source	Usage
Orac_math	Mathematics	~200K	Human + synthetic	Training / CoT examples
AOPS_forum	Math contest problems	Thousands	AoPS forum	Multi-step reasoning
CN_K12	K-12 math (Chinese)	Thousands	Textbooks / exams	Cross-lingual tasks
Synthetic_AMC	AMC-style math	Thousands–10K	Synthetic generation	Contest-style reasoning
Olympiads	Math Olympiad	Thousands	Exam archives	Hard math reasoning
Synthetic_math	General synthetic math	>100K	Synthetic generation	Broad math reasoning
GSM8K	Grade-school math	8,000	Human-authored	Benchmark / evaluation
AMC_AIME	AMC + AIME contests	Thousands	Contest archives	Hard math reasoning
MATH	Math competition dataset	12,500	MATH benchmark	Step-by-step reasoning
AQuA	Algebra/word problems	Thousands	Human-authored	Logic + math reasoning

Table 3: Model-distilled Chain-of-Thought datasets.

Dataset	Domain	Scale	Usage
AQuA	Math/logic QA	Thousands	Distilled CoT training
GSM8K	Grade-school math	8,000	Distilled CoT training
LIMO	Logic + math reasoning	Thousands	Distilled reasoning analysis
MATH	Extended MATH	Thousands	Complex reasoning

Table 4: Evaluation benchmarks for LLM performance.

Dataset	Domain	Scale	Source	Usage
AGIEval	Expert knowledge + math	Thousands	Exam-based	Cross-domain evaluation
MMLU-Pro	Broad expert knowledge	Thousands	MMLU series	General evaluation
CMMLU	Chinese / multilingual MMLU	Thousands	Exam-based	Multilingual reasoning
CommonSenseQA	Commonsense reasoning	12K	CommonsenseQA benchmark	Commonsense evaluation

7 ANNOTATION CONSISTENCY ANALYSIS

7.1 CONSISTENCY EVALUATION

Given the inherent ambiguity and subjectivity involved in annotating cognitive levels from CoT reasoning, it is essential to assess the consistency and reliability of the automated labeling process. A representative sample of 1000 CoT instances was randomly selected from the annotated dataset. Each instance was independently labeled by three large language models, including *Qwen2.5-72B-Instruct* Yang et al. (2025b), *Qwen3-32B* Yang et al. (2025a), and *ChatGPT-4o-mini*, as well as by three human annotators with relevant expertise. This multi-annotator setup enables a robust comparison between machine-generated and human-assigned cognitive labels. To evaluate alignment between different annotators, we adopt a relaxed consistency criterion inspired by educational assessment frameworks: a cognitive or knowledge label is considered acceptable if it differs from the reference label by at most one adjacent level in Bloom’s taxonomy hierarchy. Prior research Thompson & Lake (2023) supports the validity of this tolerance range in cognitive categorization tasks.

To assess the confidence level of model-based labeling, we apply confidence interval estimation based on a two-step procedure. (1) We treat the majority vote among human annotators as the ground truth. For each CoT instance, we compare the label assigned by model *M* to this reference label.

Table 5: Overview of Models by Series, Name, and Application Domain

Series	Model Name	Application Domain
Qwen3	Qwen3-0.6B	General tasks, light reasoning
	Qwen3-1.7B	General tasks, light reasoning
	Qwen3-4B	General tasks, reasoning and generation
	Qwen3-8B	General tasks, reasoning and generation
	Qwen3-14B	Advanced reasoning, complex tasks
	Qwen3-32B	Advanced reasoning, complex tasks
Phi4	Phi4-reasoning-plus	Advanced reasoning and logic tasks
	Phi4-reasoning	Reasoning and analytical tasks
	Phi4-minireasoning	Lightweight reasoning tasks
DeepSeek	DeepSeek-R1-Distill-Qwen-32B	Model distillation and knowledge retrieval
	DeepSeek-R1-0528-Distill-Qwen3-8B	Model distillation and knowledge retrieval
Qwen2.5	Qwen2.5-32B-Instruct	Instruction-tuned, general tasks
	Qwen2.5-14B-Instruct	Instruction-tuned, general tasks
	Qwen2.5-72B-Instruct	Instruction-tuned, general tasks
	Qwen2.5-Coder-32B-Instruct	Programming instructions, code generation

A prediction is marked as correct if it exactly matches the human-assigned label (or falls within the acceptable tolerance range, as defined above). The model’s accuracy is then calculated as the proportion of correct predictions over the total number n of samples $\hat{p}$. (2) We construct a confidence interval for the observed accuracy using the binomial proportion confidence interval formula:

$$CI = \hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \quad (4)$$

where $Z_{\alpha/2}$ refers to the Z-score corresponding to the chosen confidence level. This interval quantifies the statistical reliability of the model’s annotation accuracy, offering a formal measure of the model’s consistency with human judgment. The results show that the annotated model *Qwen2.5-72B-Instruct* aligns closely with human annotations and reasoning, exhibiting a very high consistency interval in Table 6 and Table 7; similar results are also observed for Cohen’s Kappa (κ).

7.2 ANALYSIS

Table 6: The Cognitive Dimension Consistency for $CI(\alpha = 95\%)$

	Expert 1		Expert 2		Expert 3		Average	
	low	high	low	high	low	high	low	high
Qwen2.5-32B-Instruct	0.8297	0.8768	0.8497	0.8939	0.8417	0.8881	0.8403	0.8863
Qwen3-32B	0.7537	0.8084	0.7479	0.8030	0.7536	0.8086	0.7517	0.8067
CompassJugder-2-32B-Instruct	0.8125	0.8642	0.8268	0.8765	0.8161	0.8687	0.8185	0.8698
Qwen2.5-72B-Instruct	0.8604	0.9036	0.8618	0.9047	0.8701	0.9127	0.8641	0.9070
ChatGPT4-mini	0.7537	0.8084	0.7479	0.8030	0.7536	0.8086	0.7517	0.8067

Table 7: The Knowledge Dimension Consistency for $CI(\alpha = 95\%)$

	Expert 1		Expert 2		Expert 3		Average	
	low	high	low	high	low	high	low	high
Qwen2.5-32B-Instruct	0.9644	0.9873	0.9697	0.9905	0.9570	0.9831	0.9637	0.9869
Qwen3-32B	0.9519	0.9774	0.9591	0.9824	0.9475	0.9743	0.9528	0.9780
CompassJugder-2-32B-Instruct	0.9720	0.9917	0.9685	0.9896	0.9687	0.9902	0.9697	0.9905
Qwen2.5-72B-Instruct	0.9792	0.9956	0.9792	0.9956	0.9783	0.9954	0.9789	0.9955
ChatGPT4-mini	0.9519	0.9774	0.9591	0.9824	0.9475	0.9743	0.9528	0.9780

8 TEST DETAILS

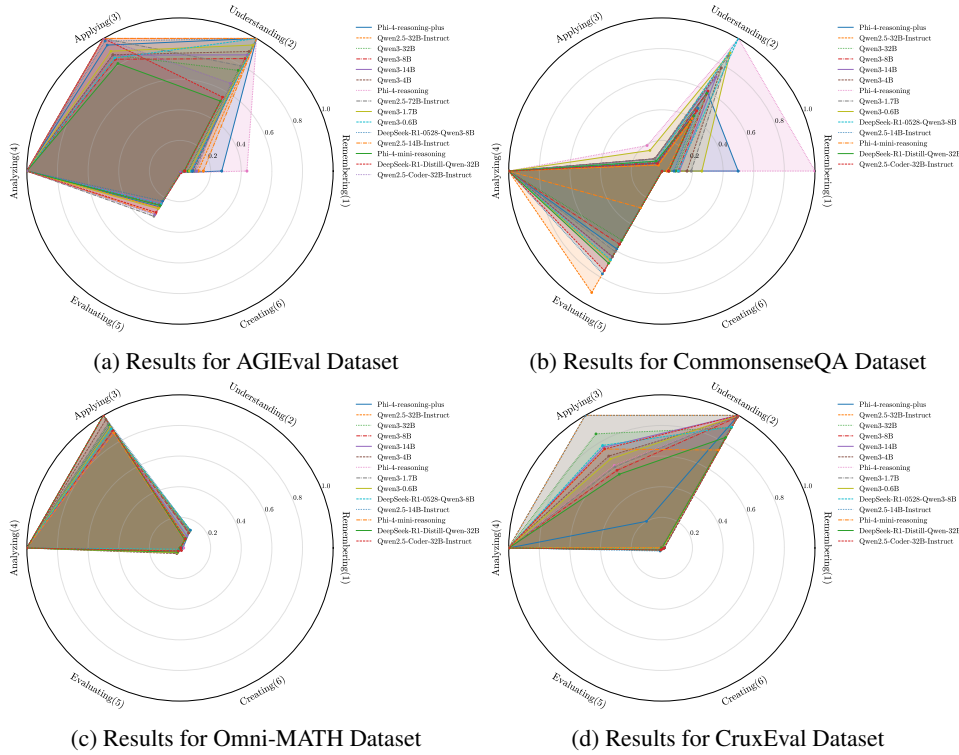


Figure 7: The whole details for the LLM’s distilled CoT results.

Figure 8 illustrates how the average cognitive level of distilled CoT reasoning in the Qwen3 series changes with model scale. Larger models generally produce reasoning with higher cognitive scores, but the improvement varies by benchmark. CommonsenseQA shows the most pronounced increase, Omni-MATH remains consistently high with marginal gains, AGIEval rises gradually, and CruxEval peaks early for smaller models before stabilizing. This indicates that scaling benefits reasoning performance, but its effect is task-dependent.

9 INSTRUCTION PROMPTS

9.1 PROMPTS TEMPLATE

We used the same system prompt as our evaluation target, and then employed different user prompts to specify various annotation tasks. The system prompt is shown as follows:

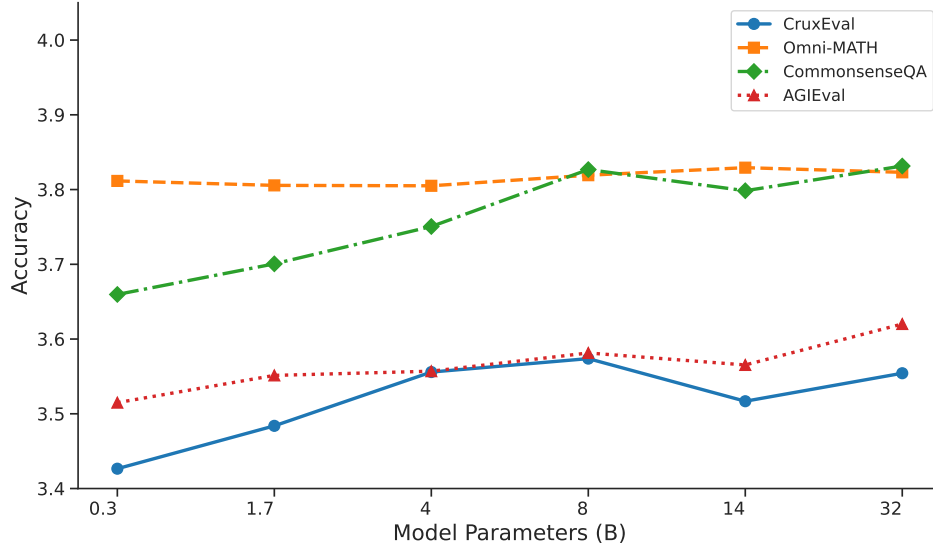


Figure 8: Average cognitive level scores of distilled CoT reasoning from the Qwen3 LLM series across different model scales on four benchmarks.

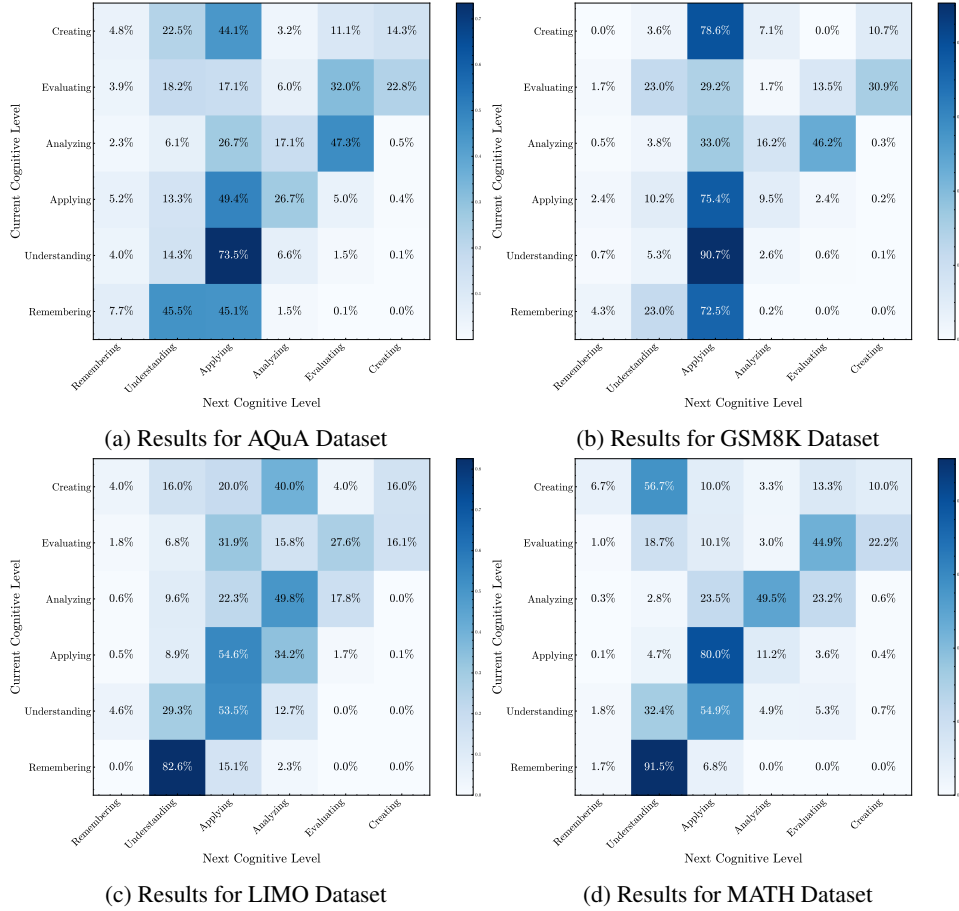


Figure 9: Cognitive trajectory annotation results: CoT reasoning mainly spans lower-order to mid-order cognition, with rare transitions to higher-order processes.

System Prompt

You are a helpful assistant facilitating meaningful dialogue between users and assistants.
The user poses a question, and the assistant provides a solution by first reasoning through the problem before delivering a response.

Please make sure to display the complete thought process in your outputs, including
<think></think> in think sections, <answer></answer> in answer section.

****Example Output:**** <think>thinking process</think><answer>Final answer</answer>

Instruction cognitive annotations Instruction-level cognitive annotation refers to the labeling of the average cognitive processes within a chain of thought, as defined in detail in the main text. In the user prompts, we provide much more detailed definitions of Bloom’s cognitive processes and the corresponding annotation methods. To ensure the annotation process is more accurate, the prompts guide the model to first reason and then provide an answer, resulting in more precise responses. The user prompt is shown as follows:

User Prompt

Assume you are a data annotation expert. Please use the "question" and the "thinking process" to classify the following "thinking process" according to Bloom's Taxonomy cognitive framework. First, explain your reasoning and analytical process for the classification, then provide its corresponding cognitive process dimension and knowledge dimension.

The revised version of the Bloom's taxonomy is divided into two dimensions: cognitive and knowledge. The cognitive dimensions include:

- **Remembering**: Thinking that focuses on retrieving or recognizing previously learned information from memory, such as recalling specific facts, terms, dates, or basic concepts without necessarily interpreting them.

- **Understanding**: Thinking that involves processing and interpreting information to demonstrate comprehension, such as explaining a concept in one's own words, summarizing a text, or classifying items into categories based on their meaning.

- **Applying**: Thinking that centers on using acquired knowledge, skills, or procedures in new or practical situations, such as solving a problem with a learned method, executing a task based on a rule, or adapting a concept to a different context. It bridges theory and practice.

- **Analyzing**: Thinking that entails breaking down complex information into its individual parts to examine relationships, patterns, or underlying structures. This could include comparing ideas, identifying causes and effects, or organizing data to reveal insights about how the pieces fit together.

- **Evaluating**: Thinking that involves assessing or critiquing information, arguments, or methods based on specific standards or criteria. Examples include judging the reliability of a source, weighing the strengths and weaknesses of an approach, or determining the quality of a solution.

- **Creating**: Thinking that focuses on combining or synthesizing elements to produce something new and original, such as designing a project, constructing a novel solution, or generating innovative ideas by integrating prior knowledge in unique ways.

The dimensions of knowledge include:

- **Factual**: Basic elements such as terminology, facts, and discrete pieces of information (the "what").

- **Conceptual**: Relationships among ideas, theories, models, and structures (the "why").

- **Procedural**: How to do something—methods, techniques, and criteria for using skills and algorithms (the "how").

- **Metacognitive**: Awareness and regulation of one's own cognition—strategies for learning and self-assessment (the "knowing about knowing").

Please **categorize the following think process with bloom's taxonomy**, **DONOT** solve THE PROBLEM, provide your thought process, and then give the answer. Here is an example:

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

User Prompt
<pre> ===== Input: ```json { "question": "<Question>", "think_process": "<Thinking>", "answer": "<answer>" } ``` Output: <think> Why fits Bloom's Taxonomy some levels </think><answer> ("cognitive", "knowledge")</answer> ===== Write the answers in the tags with format (cognitive, knowledge), and there is only one tag for cognitive dimension and one tag for knowledge dimension,DO NOT GENERATE OTHER IR- RELATIVE THINGS, and multiple tags cannot be generated. Now begin your inputs: </pre>

Trajectory cognitive annotations The **cognitive trajectory annotation** captures the cognitive labels of individual reasoning steps within a *CoT*. We defined detailed Bloom's cognitive definitions and then divided them into several parts according to the chain-of-thought process, mapping each part to its corresponding cognitive level. The user prompt is shown as follows:

System Prompt

You will be given a long chain-of-thought (CoT) reasoning text. Your task is to segment this text into a series of clear, logically complete reasoning steps, and annotate each step with its corresponding level in Bloom's taxonomy. The cognitive dimensions include:

- **Remembering**: Thinking that focuses on retrieving or recognizing previously learned information from memory, such as recalling specific facts, terms, dates, or basic concepts without necessarily interpreting them.

- **Understanding**: Thinking that involves processing and interpreting information to demonstrate comprehension, such as explaining a concept in one's own words, summarizing a text, or classifying items into categories based on their meaning.

- **Applying**: Thinking that centers on using acquired knowledge, skills, or procedures in new or practical situations, such as solving a problem with a learned method, executing a task based on a rule, or adapting a concept to a different context. It bridges theory and practice.

- **Analyzing**: Thinking that entails breaking down complex information into its individual parts to examine relationships, patterns, or underlying structures. This could include comparing ideas, identifying causes and effects, or organizing data to reveal insights about how the pieces fit together.

- **Evaluating**: Thinking that involves assessing or critiquing information, arguments, or methods based on specific standards or criteria. Examples include judging the reliability of a source, weighing the strengths and weaknesses of an approach, or determining the quality of a solution.

- **Creating**: Thinking that focuses on combining or synthesizing elements to produce something new and original, such as designing a project, constructing a novel solution, or generating innovative ideas by integrating prior knowledge in unique ways.

Do not omit any part of the original content. Each step should represent a distinct unit of thought, such as a single observation, inference, recall, or comparison.

You are only allowed to segment the original CoT content into multiple parts and assign a single Bloom's taxonomy label to each part.

The output must be **a sequence containing only individual Bloom-level labels**. **WARNING: DONOT GENERATE ANY OTHER IRRELEVANT CONTENTS!** **The output can only be a python list and cannot contain any other irrelevant content! And any steps should be one of Bloom's taxonomy tags.**

=====

Here is an Example:

```

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217

```

System Prompt

Input:

```

```json
{
 "question": "<Question>",
 "think_process": "<Thinking>",
 "answer": "<answer>"
}
```

Output: <think> why the annotation tags is set, give the thinking for annotation </think>
<answer>
[
  ["bloom-cognitiveA", "raw_text"],
  ["bloom-cognitiveB", "raw_text"],
  ...,
  ["bloom-cognitiveN", "raw_text"]
]
</answer>
=====

Now begin the normal Input:

```

Algorithm 1: Cognitive Annotation of Chain-of-Thought

```

1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

```

Input: Dataset $D = \{x_1, \dots, x_i, \dots, x_N\}$

Output: Average cognitive annotation D_I , Cognitive trajectory annotation D_T

Initialization:

Set average cognitive annotation prompt p_I and cognitive trajectory annotation prompt p_T ;

Load dataset D and format into $x_I = (q, CoT, a, p_I)$, $x_T = (q, CoT, a, p_T)$;

Initialize annotated sets $D_I = \emptyset$, $D_T = \emptyset$;

Initialize remaining sets $D_E^{(I)} = D_E^{(T)} = D$;

for $j \in \{I, T\}$ **do**

while $D_E^{(j)} \neq \emptyset$ **do**

Sample a batch($i : i + B$) from $D_E^{(j)}$: $x_{i:i+B} = (q_{i:i+B}, CoT_{i:i+B}, a_{i:i+B}, p_{i:i+B}^{(j)})$;

$O_{i:i+B} \leftarrow LLM(x_{i:i+B}^{(j)})$;

for $k \in (i : i + B)$ **do**

if O_k satisfies annotation requirements **then**

$D^{(j)} \leftarrow D^{(j)} \cup O_{x_k^{(j)}}$;

$D_E^{(j)} \leftarrow D_E^{(j)} - O_{x_k^{(j)}}$;

return D_I, D_T

9.2 CASE STUDY

Here are some examples of annotated thought chains:

A case for Instruction-Level Annotation

"question": "Harly's animal shelter has 80 dogs. She adopts out 40% of them but then has to take back 5 because of personality conflicts with other dogs in their adopted homes. How many dogs does she have now?", "think_process": "How many dogs are adopted out? ** First find how many dogs are adopted out: $80 \text{ dogs} * 40\% = \langle\langle 80 * 40 * .01 = 32 \rangle\rangle 32 \text{ dogs}$ How many dogs does Harly have now? ** Then subtract the number of dogs adopted out from the total number of dogs and add the number who were returned to find the final number of dogs: $80 \text{ dogs} - 32 \text{ dogs} + 5 \text{ dogs} = \langle\langle 80 - 32 + 5 = 53 \rangle\rangle 53 \text{ dogs}$ ", "answer": "53", "response": "<think>

Why this fits Bloom's Taxonomy:

- + The person calculates the number of dogs adopted out by applying the percentage to the total number of dogs (this shows they can use mathematical procedures).
- + They then adjust the total number of dogs by subtracting those adopted out and adding those returned (this shows they can apply operations to solve a problem).
- + They are using procedural knowledge — methods, techniques, and criteria for using skills and algorithms to solve a practical problem.

</think><answer>(Applying,Procedural)</answer>"

A case for Instruction-Level Annotation

"question": "Meso can type 15 pages in 5 minutes. Tyler can type the same 15 pages in 3 minutes. How many minutes would it take Meso and Tyler to type 40 pages working together?", "think_process": "Meso types 15 pages / 5 minutes = $\langle\langle 15 / 5 = 3 \rangle\rangle 3 \text{ page per minute}$. Tyler types 15 pages / 3 minutes = $\langle\langle 15 / 3 = 5 \rangle\rangle 5 \text{ pages per minute}$. Together they can type $3 + 5 = \langle\langle 3 + 5 = 8 \rangle\rangle 8 \text{ pages per minute}$. Working together, it would take them $40 \text{ pages} / 8 \text{ pages per minute} = \langle\langle 40 / 8 = 5 \rangle\rangle 5 \text{ minutes}$ ",

"answer": "5", "response": "<think>

Why this fits Bloom's Taxonomy:

- + The person calculates the typing speed of Meso and Tyler by dividing the number of pages by the time it takes (this shows they can apply mathematical operations to solve a problem).
- + They then combine the typing speeds to find the total pages per minute when working together (this shows they can apply the concept of combining rates).
- + Finally, they calculate the time it would take to type 40 pages together by dividing the total pages by the combined rate (this shows they can apply the concept of time, rate, and work relationships).

</think><answer>(Applying,Procedural)</answer>"

A case for Trajectory-Level Annotation

"question": "Michael is traveling on the interstate at an average speed of 50km/hr (taking into account all the necessary stops for gas, relaxation, etc.). If he goes on like this for 3 days, and the distance between Alaska and Texas is 6000 km, what percentage of this distance has he covered?", "answer": "There are 24 hours in each day, so he has been traveling for a total of $3 \times 24 = 72$ hours

At a rate of 50km/hr he has covered $50 \times 72 = 3600$ km

3600km as a percentage of 6000km is $(3600/6000) \times 100\% = 60\%$,

"answer": "60" <think> Step 1: Recall basic time conversion (hours in a day) and multiplication to find total travel time. Step 2: Apply the formula distance = speed $\times$ time to calculate total distance traveled. Step 3: Apply percentage calculation formula to find what fraction of the total distance has been covered. </think> <answer> [["Remembering", "There are 24 hours in each day, so he has been traveling for a total of $3 \times 24 = 72$ hours."], ["Applying", "At a rate of 50km/hr he has covered $50 \times 72 = 3600$ km."], ["Applying", "3600km as a percentage of 6000km is $(3600/6000) \times 100$ "] </answer>

A case for Trajectory-Level Annotation

"question": "To make 1 liter of juice, Sam needs 5 kilograms of oranges. Each kilogram of oranges costs \$3. How much money would Sam have to spend to make 4 liters of juice?",

"think_process": "How many kilograms of oranges does Sam need to make 4 liters of juice? **

To make 4 liters of juice, Sam needs $4 \times 5 = 20$ kilograms of oranges.

How much would 20 kilograms of oranges cost? ** If one kilogram of oranges costs \$3,

then 20 kilograms would cost $20 \times 3 = 60$.", "answer": "60", "response":

"<think> The CoT reasoning involves a series of calculations where Sam determines the total amount of oranges needed and the total cost. Each calculation step applies a learned method (multiplication) to a practical scenario, which aligns with the **Applying** level in Bloom's taxonomy. No steps involve merely recalling facts, interpreting concepts, analyzing structures, evaluating, or creating something new; they are all straightforward applications of arithmetic to solve a problem. Therefore, every step is labeled as **Applying**. </think> <answer> [["Applying", "How many kilograms of oranges does Sam need to make 4 liters of juice?"], ["Applying", "To make 4 liters of juice, Sam needs $4 \times 5 = 20$ kilograms of oranges."], ["Applying", "How much would 20 kilograms of oranges cost?"], ["Applying", "If one kilogram of oranges costs 3, then 20 kilograms would cost $20 \times 3 = 60$."]] </answer>"