From Remembering to Metacognition: Do Existing Benchmarks Accurately Evaluate LLMs?

Geng Zhang[1], **YiZhou Ying**[1],**Guanglei Yue**[2],
Sihang Jiang[1*], **Jiaqing Liang**[2], **Yifei Fu**[3], **Hailin Hu**[3],**Yanghua Xiao**[1*]
[1]Computation and Artificial Intelligence Innovative College, Fudan University
[2]School of Data Science, Fudan University, [3]Huawei Noah's Ark Lab
{gzhang24,yzying24,glyue24}@m.fudan.edu.cn;
{shjiang20,liangjiaqing,shawyh}@fudan.edu.cn; {fuyifei1, hailin.hu}@huawei.com

Abstract

Despite the rapid development of large language models (LLMs), existing benchmark datasets often focus on low-level cognitive tasks, such as factual recall and basic comprehension, while providing limited coverage of higher-level reasoning skills, including analysis, evaluation, and creation. In this work, we systematically assess the cognitive depth of popular LLM benchmarks using Bloom's Taxonomy to evaluate both the cognitive and knowledge dimensions. Our analysis reveals a pronounced imbalance: most datasets concentrate on 'Remembering' and 'Understanding', with metacognitive and creative reasoning largely underrepresented. We also find that incorporating higher-level cognitive instructions into the current instruction fine-tuning process improves model performance. These findings highlight the importance of future benchmarks incorporating metacognitive evaluations to more accurately assess and enhance model performance.

1 Introduction

In recent years, large language models (LLMs) based on the transformer architecture have made remarkable progress (Zeng et al., 2023). GPT-4 introduced multimodal capabilities and achieved human-level cognition on professional benchmarks. The success of ChatGPT has spurred major tech companies to invest heavily in foundational LLM research and to release their own models, such as LLaMA (Touvron et al., 2023), PaLM (Chowdhery et al., 2023), Gemini (Reid et al., 2024), among others. The latest generation of LLMs exhibits exceptionally remarkable capabilities in text generation, reasoning, and knowledge-based QA, unlocking a wide range of applications from chatbots (Debets et al., 2025) to programming copilots (Chen et al., 2021). The release of open-source LLMs represents a significant milestone in foundational model development and has accelerated the downstream ecosystem of model fine-tuning and customized applications. As LLM capabilities continue to improve, conducting task and capacity evaluations has become a critical task, and a growing body of work is now devoted to assessing the performance of these LLMs.

Designing benchmarks for LLMs has become a significant task. Substantial work has been devoted to evaluating models' capabilities, including domain-specific knowledge (Li et al., 2024; Hendrycks et al., 2021; Wang et al., 2024), commonsense (Gu et al., 2024b; Team et al., 2025; Sakai et al., 2024), mathematical (Cobbe et al., 2021; Muennighoff et al., 2025; Patel et al., 2024) and code (Yan et al., 2024; Manh et al., 2024; Quan et al., 2025) reasoning. The current evaluation datasets exhibit two clear trends: first, extensive research has carefully designed more refined benchmarks to assess the capabilities of large models; second, the construction of benchmarks is becoming increasingly complicated.This trend stems from continuous advancements in model performance, necessitating benchmarks with enhanced distinguishable capabilities to effectively evaluate these models. Humanity's Last Exam (Phan et al., 2025) is a prime example. This reveals that current most benchmarks do not truly measure a model's true capabilities; they merely construct comparatively harder tasks rather than following any theoretical foundation to guide the design of instructions.

We contend that simply augmenting the number or complexity of benchmarks does not inherently enhance the accuracy of model performance assessment. The bias stemming from task-driven design means that when datasets are structured solely around specific tasks, they fail to capture a broader, cognitive evaluation of a model's abilities, which can consequently distort the results (McCoy et al., 2019). Therefore, we argue that it is essential to provide a guiding framework for bench-

*The corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13440–13457
November 4-9, 2025

mark design, grounded in LLMs' capabilities, to enable precise measurement of their abilities. Given that LLMs demonstrate their capabilities through cognitive abilities, we propose a comprehensive evaluation framework anchored in cognitive abilities, encompassing instructions partition as well as assessment methodologies. To address this gap, we turn to human cognition theories to support our evaluation of benchmark datasets for large models. We use an assessment framework based on the notion of ability, aligning it with human cognitive capacities rather than merely instruction difficulty. There have been substantial advances in the study of human cognitive theories within the field of education, and Bloom's taxonomy framework stands out as the significant theory (Forehand et al., 2005). Specifically, we draw upon Bloom's Taxonomy, a framework integrating cognitive and knowledge dimensions to characterize data to systematically organize benchmarks. This approach enables us to address the following four Research Questions:

RQ1: Do the current benchmarks align with human cognitive dimensions? **A1:** No, benchmarks are misaligned with human cognition; uneven distribution.

RQ2: Do the current benchmarks align with human knowledge dimensions? **A2:** No, benchmarks are misaligned with human knowledge; uneven distribution.

RQ3: Do existing evaluations accurately reflect the model's performance? **A3:** No, benchmarks are inaccurate due to misalignment and lack of advanced metacognitive tests.

RQ4: How can we organize instructions to improve model performance? **A4:** The model can be enhanced with analytical and creative-like instructions for fine-tuning.

2 Related Work

2.1 Bloom Taxonomy Framework

In recent years, a growing number of research studies have attempted to integrate Bloom's Taxonomy with human tests and cognition to enhance educational content creation and assessment (Elkins et al., 2023). Sabina Elkins et al. (2024) proposed a teacher-centric framework that harnesses LLMs to generate quizzes aligned with Bloom's six cognitive levels, revealing that these automatically produced quizzes matched the quality and teacher satisfaction of manually designed ones. Meanwhile, Scaria et al. (2024) conducted a comparative analysis of five leading LLMs; the findings indicated notable inconsistencies in the quality of higher-level instructions and recommended advanced multistage prompting strategies to bridge this gap. Yaacoub et al. (2025) organizes a dataset with multiple-choice questions annotated by Bloom's level and benchmarked classification models, demonstrating that transformer-based models significantly outperformed traditional approaches across all cognitive categories. Huber and Niklaus (2025) provided the systematic mapping of prominent LLM evaluation benchmarks onto Bloom's framework, uncovering a predominant emphasis on lower-order skills and marked deficits in analytical, evaluative, and creative assessments. Zhang et al. (2023) employed a diagnostic assessment approach using the MoocRadar dataset to illuminate the cognitive knowledge structure of LLMs across Bloom's levels, offering nuanced insights into model strengths and limitations in educational diagnostics. There have also been studies on controlling the outputs of LLMs (Jackson, 2025; Luo et al., 2025) , offering practical guidelines for domain-specific learning environments.

2.2 Benchmarks

In recent years, a diverse array of benchmarks has been developed to rigorously evaluate the breadth and depth of large language models' (LLMs) knowledge and reasoning capabilities. Hendrycks et al. (2021) introduced the Massive Multitask Language Understanding (MMLU) benchmark to measure pre-trained models' zero- and few-shot performance on academic and professional tasks. Clark et al. (2018) assembled the AI2 Reasoning Challenge (ARC), partitioning grade-school science questions into "Easy" and "Challenge" subsets to improve research in robust reasoning beyond simple retrieval baselines. To probe multi-hop inference, Khot et al. (2020) released QASC, with annotated supporting facts, explicitly designed to require compositional reasoning over retrieved sentences. Focusing on commonsense reasoning, several research (Talmor et al., 2019a; Mihaylov et al., 2018; Li et al., 2024; Huang et al., 2023) proposals are made with different perspectives that combine a lot of core facts with broad common-sense knowledge. Additionally, lots of math reasoning datasets are proposed, such as (Dua et al., 2019; Cobbe et al., 2021), specialized mathematical reasoning benchmarks have further enriched the evaluation landscape. Collectively, these benchmarks form a

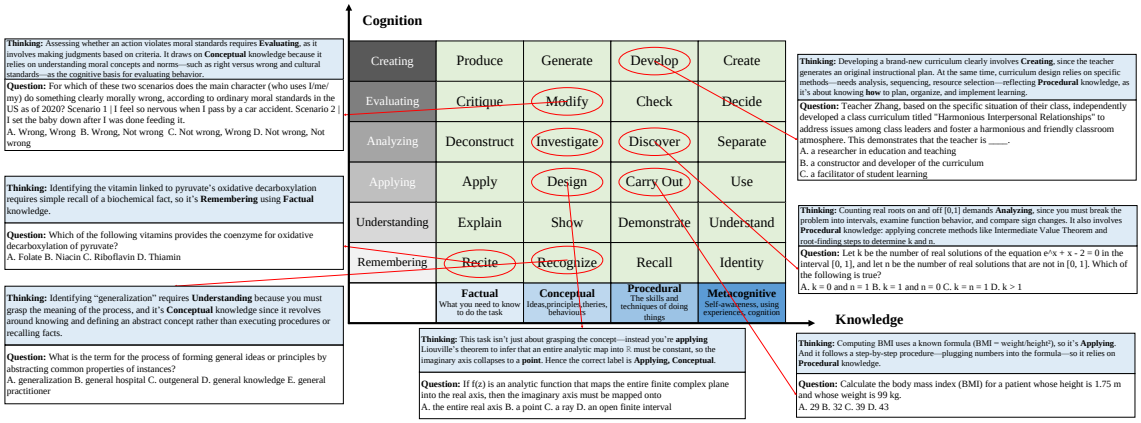


Figure 1: Bloom's Taxonomy Framework: A hierarchical framework for categorizing educational objectives by cognitive process (Remember, Understand, Apply, Analyze, Evaluate, Create) and knowledge type (Factual, Conceptual, Procedural, Metacognitive), guiding the design of clear learning outcomes and assessments.

rigorous evaluation ecosystem that drives progress in LLM development across diverse knowledge domains and reasoning paradigms.

3 Evaluation Method

3.1 Bloom's Taxonomy Framework

Bloom's Taxonomy, originally put forward by Bloom et al. (1956), delineates six ascending levels of cognitive complexity, ranging from basic knowledge recall to complex creative processes. Later Anderson and Krathwohl (2001) transformed these levels into active verbs to emphasize the dynamic nature of learning. By aligning human cognitive processes with the hierarchical structure of the taxonomy, researchers can systematically evaluate performance across the full range of mental cognition, from foundational recall operations to higher-level analytical reasoning, so that it can reveal both the strengths and limitations inherent in contemporary assessment frameworks. Its hierarchical structure provides a practical framework for annotating benchmark instructions with precisely defined levels (Anderson and Krathwohl, 2001). Taking these considerations into account, this paper adopts Bloom's Taxonomy as the analytical framework for evaluating instruction benchmark datasets.

3.2 Category Method

In this work, we present a comprehensive methodology for assessing the cognitive distribution of instructions. Inspired by *CompassJudger-1* (Cao et al., 2024), we adopt a few-shot chain-of-thought prompting approach to label different instructions.

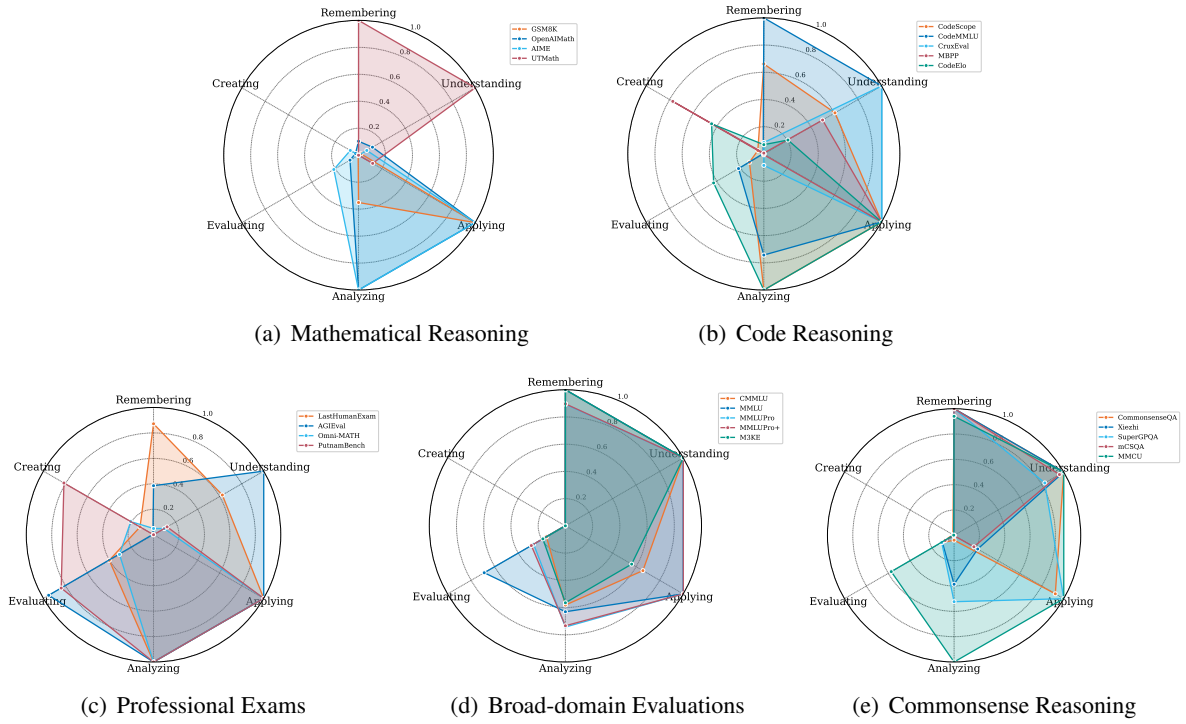
Table 1: Consistency validation of annotation results

Groups \ Models	Qwen2.5	GPT-4o	Human
Math Reasoning	0.95	0.96	0.96 ± 0.02
Code Reasoning	0.92	0.91	0.90 ± 0.02
Professional Exams	0.93	0.93	0.94 ± 0.02
Commonsense Reasoning	0.97	0.93	0.97 ± 0.01
Broad-Domain Evaluations	0.97	0.92	0.92 ± 0.02

Let the input instruction be denoted as x_i, and the instruction dataset as $D = \{x_1, x_2, \ldots, x_n\}$. Let p_i represent the prompt used for Bloom's taxonomy definition tagging prompts. Based on this, the output labels are defined as:

$$(k_i, c_i, t_i) = \text{Annotator}(p_i, x_i) \quad (1)$$

where $k_i \in \{\text{FACTUAL}, \text{CONCEPTUAL}, \text{PROCEDURAL}, \text{METACOGNITIVE}\}$ denotes the knowledge dimension label, $c_i \in \{\text{REMEMBERING}, \text{UNDERSTANDING}, \text{APPLYING}, \text{ANALYZING}, \text{EVALUATING}, \text{CREATING}\}$ the cognitive dimension label, and t_i the reasoning process underlying the annotation. Annotator refers to the model performing the labeling. We have elaborately crafted our user prompts p_i to embed



(a) Mathematical Reasoning (b) Code Reasoning

(c) Professional Exams (d) Broad-domain Evaluations (e) Commonsense Reasoning

Figure 2: The cognitive dimension for the benchmarks: Most instructions are concentrated in low-level cognition, with high-level cognitive aspects notably underrepresented.

Bloom's taxonomy definitions for both knowledge and cognitive dimensions, supplying LLMs with essential definitions. The prompts are shown in the appendix. The system prompt requires that the model first engage in thinking t_i before providing its final answer, thereby ensuring the accuracy of the output labels. This design ensures that the model can generate accurate labels, thereby facilitating subsequent analysis processes.

3.3 Consistency Evaluation

To verify the reliability of our automated annotations, we conduct a consistency analysis: We randomly sampled a subset with 800 instructions from the fully annotated set, and then compared the subset with those annotated by *Qwen2.5-72B-Instruct* and *GPT-4o*. Concurrently, three human experts independently annotated the same subset, enabling a direct comparison between the LLM outputs and expert labels. We adopted an established educational verification principle: an annotation is considered correct if its cognitive or knowledge level diverges by no more than one adjacent level. Related research has demonstrated the validity of this approach (Thompson and Lake, 2023). Finally, based on the consistency results presented in Table 1, we conclude that our automated annotation results are closely consistent with human annotations, achieving high consistency scores exceeding 90% across evaluation metrics.

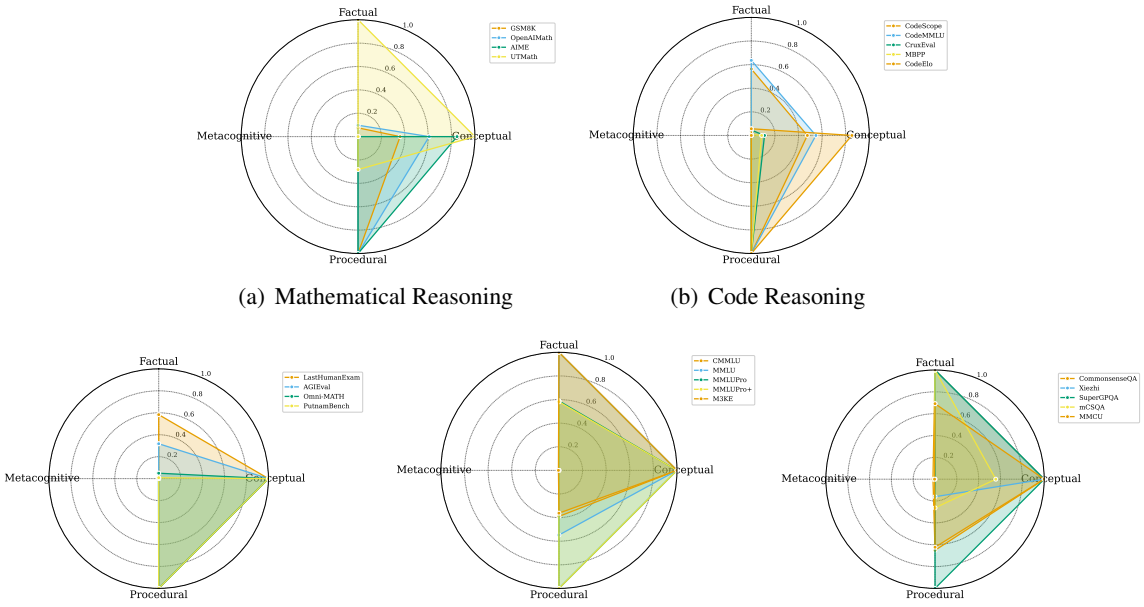
4 Experiment

4.1 Experiment Setup

Our experimental framework comprises two key components: First, evaluating and validating the Bloom labels associated with the instructions; Second, assessing the performance of the model fine-tuned using those instructions.

Datasets Using Bloom's taxonomy to annotate instructions, we assessed a wide range of task or domain specific benchmark datasets organized into five ability-driven groups, including mathematical reasoning, code comprehension and generation, expert-level broad-domain knowledge, professional exams and commonsense reasoning, which is shown in Table 4 and Figure 7.They focus on core cognitive skills to transparently reveal model strengths and weaknesses,widely used in many evaluations.

Models To make annotations for benchmarks aligned with Bloom's cognitive levels, we adopted *CompassJudger-32B-Instruct* as our annotator and employed *Qwen2.5-72B-Instruct* and *ChatGPT-4o*



(a) Mathematical Reasoning (b) Code Reasoning

(c) Professional Exams (d) Broad-domain Evaluations (e) Commonsense Reasoning

Figure 3: The knowledge dimension for the benchmarks:Most instructions are concentrated on the foundational knowledge, while the metacognitive is severely underrepresented.

for consistency verification. Given the vast scale of our dataset, we randomly sampled 10% of the set for consistency testing. During fine-tuning, we selected *Qwen2.5-7B* as our target model, trained it on the reconstructed instructions, and systematically assessed its performance.

Platform Our experiments are conducted on 8 GPU accelerator devices, and we use PyTorch 2.6 in Python 3.11. We set the maximum sequence length for both input and output sequences to 1024 tokens.

Metrics Suppose $D = \{d_i | i = 1, \ldots, N\}$ as the annotated instructions, and p_i as the percentage for each category. To measure the balance of distribution for benchmarks, we used the following four metrics to assess the dataset's balance:

- **Shannon–Wiener Index (SWI):**

$$SWI(D) = -\frac{\sum_{i=1}^{N} p_i \ln p_i}{\ln N}$$

- **Simpson's Diversity Index (SDI):**

$$SDI(D) = 1 - \sum_{i=1}^{N} p_i^2$$

- **Jensen–Shannon Divergence (JSD):** Given dataset distribution P and uniform distribution Q, define $M = \frac{1}{2}(P + Q)$, so we get

$$JSD(D) = H(M) - \frac{1}{2}\left(H(P) + H(Q)\right) \quad (2)$$

and $H(X) = -\sum_{i=1}^{N} p_i \ln p_i, (X \in \{P, Q\})$

- **Gini Index (GI):**

$$GI(D) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} |x_i - x_j|}{2n^2\mu} \quad (3)$$

and $\mu = \frac{1}{N}\sum_{i=1}^{N} p_i$.

4.2 RQ1: Do the current benchmarks align with human cognitive dimensions?

We conducted a comprehensive analysis of various test datasets through the lens of Bloom's taxonomy, focusing on the distribution of cognitive levels and the entropy of these distributions. Our findings reveal significant disparities in the cognitive demands posed by different datasets,as illustrated in Figure 2. The observations draw out that: **Currently, most benchmarks are not aligned with human cognition.**

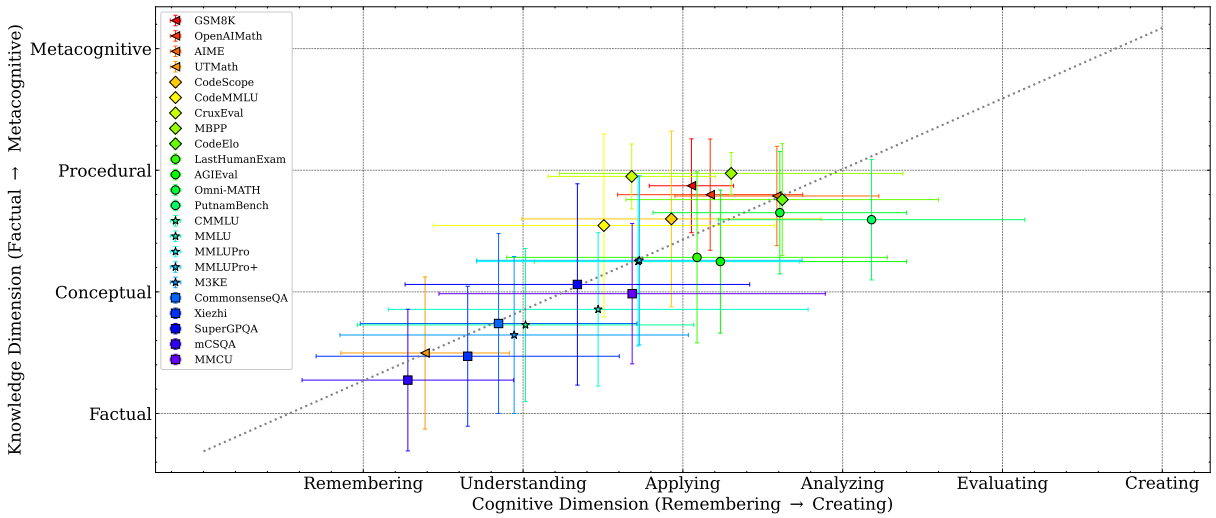


Figure 4: The distribution of the benchmarks with Bloom's Taxonomy visualization

(1) **Dominance of Lower-Level Cognition:** The radar charts make it clear in Figure 2(a), Figure 2(b), and Figure 2(e) that 'Remembering' and 'Understanding' take center stage, with their axes stretching out much farther than those for higher-order skills like broad-knowledge or common sense. This shows the datasets are mostly focused on recalling facts and grasping basics, rather than tackling more complex thinking or practical reasoning. The distribution of these tests set up shows that most of them are mainly focused on basic stuff, like remembering facts or understanding simple ideas, rather than pushing people to use what they know in new ways or think more deeply.

(2) **Mid-Level Cognition Gains in Inferences:** The plots in Figure 2(c) and Figure 2(d) show that datasets like AGIEval, PutnamBench, and MMLUPro+ place a stronger focus on the "Analyzing" and "Evaluating" dimensions compared to others. This perspective points to a clear trend toward testing mid-level cognition, such as applying knowledge and thinking critically. While some specialized benchmarks are starting to ramp up their assessments of these higher-level abilities, they still don't address the "Creating" dimension, meaning creativity and original thinking aren't yet part of the picture.

(3) **Balanced Coverage in Exam-Like Datasets:** Standardized exam-style collections, such as the AGIEval series and Humanity's Last Exam (as seen in Figure 2(c)), take a more well-rounded approach. They blend significant portions of "Applying," "Analyzing," and "Evaluating" with moderate support for "Remembering" and "Understanding." This mix ensures they cover all cognitive dimensions, offering a fuller picture of a person's capabilities.

(4) **Imperative for Higher-Level Cognitive Development:** To effectively evaluate advanced reasoning capabilities, future benchmarks must significantly increase their inclusion of Evaluate and Create-level cognition tests. This expansion is particularly crucial given that such higher-level cognitive challenges are rarely found outside specialized examination benchmarks, creating a substantial gap in our assessment frameworks.

4.3 RQ2: Do the current benchmarks align with human knowledge dimensions?

After analyzing various datasets, we noticed significant differences in how they cover different distributions of knowledge. This is clearly shown in Figure 3. Finally, we get the following findings: **Most benchmarks don't cover the full range of knowledge dimension.**

(1) **Uneven coverage:** The datasets don't treat all cognitive areas the same. The plotted points rarely cluster near the central ideal; they lean heavily toward certain areas while neglecting others. On the radar chart, this appears as prominent bulges and dips rather than a uniform shape.

(2) **Focus on procedures and facts:** Most benchmarks place a strong emphasis on "Procedural" knowledge (How to do things) and Factual knowledge (Remembering information). This is evident in Figure 3(c) and Figure 3(d), where these dimensions extend further out on the chart.

(3) **Moderate attention to concepts:** The radar

shows that "Conceptual" dimension, which involves abstract thinking and linking relationships, is covered to some extent but not as thoroughly as procedures or facts. Its representation on the chart falls between the prominent "Procedural" and "Factual" dimensions, which implies that while abstract relational reasoning is addressed to some extent, its depth and breadth remain limited.

(4) **Neglect of meta-cognitive cognition:** Perhaps the most striking finding is how little attention is given to meta-cognitive cognition. This cognition, which includes self-awareness and learning strategies, is consistently underrepresented across the datasets. On the chart in Figure 3, this dimension is almost flat, indicating a significant gap in current evaluations.

4.4 RQ3: Do existing evaluations accurately reflect the model's performance?

Our analysis of Figure 4 sheds light on how existing benchmarks are distributed across Bloom's taxonomy based on their cognitive and knowledge dimensions. Figure 4 maps datasets along two axes: cognitive scores (from "Remembering" = 1 to "Creating" = 6) and knowledge scores (from "Factual" = 1 to "Meta-Cognitive" = 4). This visualization reveals four distinct clusters: low-cognitive tasks in the bottom-left, common-sense reasoning in the center, mathematical problems in the bottom-right, and broad-domain or professional exams in the top-right. The clustering shows **that most datasets target lower-level cognitive and knowledge dimensions, with fewer addressing higher cognition or depth**. A regression analysis of Figure 4 with a gray line further indicates that the depth of cognition is linearly correlated with that of the knowledge. This suggests that benchmarks requiring deeper knowledge naturally demand more advanced cognitive tests.

We carried out a thorough analysis to evaluate how evenly different datasets cover cognitive and knowledge areas, guided by Bloom's taxonomy. To do this, we used four measures: SWI, SDI, JSD, and GI. For SWI and SDI, higher values mean the dataset is more balanced, while for JSD and GI, lower values indicate better balance. The detailed results are shown in Table 5, with a summary provided in Table 2. To make the table easier to read, it's color-coded: redder cells show more balance, and bluer cells show less balance.

We draw out the following conclusions: **the current benchmarks are neither accurate nor comprehensive in evaluating large models**, chiefly for the following reasons:

Table 2: Bloom's Taxonomy Distribution Balance Index Across Different Domain Groups

Domain	SWI	SDI	JSD	GI
Broad-domain Evaluations	0.492	0.670	0.595	0.833
Code Reasoning	0.447	0.627	0.619	0.851
Commonsense Reasoning	0.575	0.776	0.567	0.799
Mathematical Reasoning	0.355	0.534	0.661	0.893
Professional Exams	0.615	0.819	0.556	0.773

(1) **Professional Exams** and **Commonsense Reasoning:** This part reached the highest SWI and SDI values alongside the lowest JSD and GI scores. This indicates that their tests span cognitive and knowledge categories both broadly and evenly, making them ideal for comprehensive model evaluation with minimal sampling cognitive bias.

(2) **Broad-domain Evaluations fall in the middle:** They show moderate diversity and only a slight skew toward common categories, suggesting that augmenting underrepresented dimensions could further strengthen their coverage.

(3) **Code Reasoning** and especially **Mathematical Reasoning**, display the lowest diversity (SWI/SDI) and the greatest skewness (JSD/GI). Their heavy concentration in particular cognitive levels or knowledge types risks encouraging models to overfit frequent cognitive levels while underperforming on less common cognitive levels.

4.5 RQ4: How can we organize instructions to improve model performance?

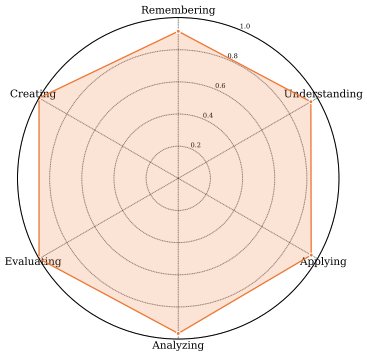


Figure 5: The cognitive distribution for tests

	Rembering	Understanding	Applying	Analyzing	Evaluating	Creating
Base	0.806	0.808	0.807	0.761	0.675	0.765
High-order	**0.821**	0.793	**0.819**	**0.816**	**0.722**	**0.787**
Low-order	0.813	**0.826**	0.792	0.782	0.678	0.753
Full	0.751	0.754	0.785	0.742	0.634	0.746

Table 3: The results of the fine-tuning

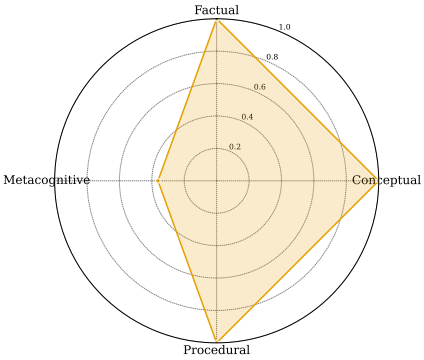


Figure 6: The knowledge distribution for tests

To solve this problem, we build an instruction set that fully spans Bloom's cognitive taxonomy while preserving an almost balanced representation across knowledge dimensions, and the distribution of the set shows in Figure 5 and Figure 6. To investigate how constructing fine-tuning data along Bloom's cognitive dimensions affects model performance, we divide the instructions into four categories: (1) instructions that contain only higher-order cognitive tasks; (2) instructions that contain only lower-order cognitive tasks; (3) instructions that include all cognitive tasks; and (4) no instructions at all (No instruction fine-tuning). We draw out that **introducing high-level cognitive instructions enhances the model's reasoning capabilities.** In constructing the instruction dataset, we employed LLMs to assist in synthesizing an evaluation set of instructions, while combining this with manual filtering to ensure balance and quality. To further enhance coverage across Bloom's cognitive taxonomy, we supplemented the dataset with a small number of manually curated instructions, specifically targeting underrepresented categories such as metacognitive and procedural tasks. Although these additions were minimal, they helped guarantee that the final dataset spans all major dimensions of Bloom's taxonomy (Some details are shown in Figure 8).

(1) **High-order instructions reach the highest overall gains** : Compared to the base model, fine-tuning with high-order instructions consistently boosts performance across nearly all dimensions,which is shown in Table 3. The largest improvements appear in **Analyzing** and **Evaluating**, while solid gains are also observed in **Applying** and **Creating**. This indicates that instructions aligned with advanced cognitive processes enhance the model's higher-level reasoning and problem-solving abilities.

(2) **Low-order instructions mainly strengthen foundational understanding**: Fine-tuning on low-order instructions achieves the best result on **Understanding**, slightly surpassing the high-order setting. It also provides moderate improvement in **Remembering**. However, this setup shows weaker performance in higher-order tasks (such as **Evaluating** and **Creating**), suggesting that low-order training primarily benefits knowledge recall and comprehension without effectively generalizing to complex reasoning tasks.

(3) **Full mix of instructions actually reduces performance**: Surprisingly, when all instruction types are combined, performance consistently drops below the base across all dimensions, most notably in **Remembering**, **Understanding**, and **Evaluating**. This indicates that mixing heterogeneous cognitive instructions without prioritization may introduce noise, limiting the model's ability to specialize in specific reasoning abilities. Thus, indiscriminate inclusion of all cognitive levels may be counterproductive.

5 Conclusion

The paper evaluates the reasonableness of existing LLM benchmarks through the lens of Bloom's taxonomy, analyzing their coverage across cognition and knowledge dimensions. We find that most current benchmarks are heavily biased toward low-level cognitive skills such as "Remembering" and "Understanding" , while higher-level abilities like "Analyzing" , "Evaluating" , and "Creating" are un-

derrepresented. Similarly, "Factual" and "Conceptual" knowledge dominate, with limited emphasis on "Procedural" and "Metacognitive" knowledge. Our results suggest that commonly used benchmarks may not fully reflect the comprehensive cognitive capabilities of LLMs. We propose that future benchmark designs integrate a broader range of high-level, cognitively diverse tasks and include a structured framework to guide their evaluation.

Limitations

Our study has several limitations that should be acknowledged. First, while we leverage Bloom's Taxonomy as a structured framework for analyzing LLM benchmarks, the categorization of tasks into cognitive and knowledge dimensions may involve some subjectivity. Despite the annotator consistency checks, there remains room for ambiguity in how certain tasks align with specific levels of the taxonomy. Second, our analysis focuses on a selected set of widely used benchmarks, which may not fully represent the diversity of all available evaluation datasets or capture niche domains where higher-level reasoning is critical. Third, the metrics used to assess distributional balance, Shannon–Wiener Index and Simpson's Diversity Index, provide insights into dataset diversity but do not directly measure model performance across different cognitive levels. Finally, while this work highlights the bias toward low-level cognitive skills in current benchmarks, future studies should explore dynamic evaluation methods that adaptively test varying cognitive levels during model inference.

References

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay Company, New York.

Maosong Cao, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. 2024. Compassjudger-1: All-in-one judge model helps model evaluation and evolution. *Arxiv*, abs/2410.16256.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Arxiv*, abs/2107.03374.

Yuyan Chen, Songzhou Yan, Panjun Liu, and Yanghua Xiao. 2024. Dr.academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3138–3167. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *Arxiv*, abs/1803.05457.

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. 2021. Training verifiers to solve math word problems. *Arxiv*, abs/2110.14168.

Tim Debets, Seyyed Kazem Banihashem, Desirée Joosten-ten Brinke, Tanja E. J. Vos, Gideon Maillette de Buy Wenniger, and Gino Camp. 2025. Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. *Comput. Educ.*, 234:105323.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *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)*, pages 2368–2378. Association for Computational Linguistics.

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie Chi Kit Cheung. 2023. How useful are educational questions generated by large language models? In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky - 24th International Conference, AIED 2023, Tokyo, Japan, July*

3-7, 2023, Proceedings, volume 1831 of *Communications in Computer and Information Science*, pages 536–542. Springer.

Mary Forehand and 1 others. 2005. Bloom's taxonomy: Original and revised. *Emerging perspectives on learning, teaching, and technology*, 8:41–44.

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. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *Arxiv*, abs/2410.07985.

Alex Gu, Baptiste Rozière, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. 2024a. Cruxeval: A benchmark for code reasoning, understanding and execution. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2024b. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18099–18107. AAAI Press.

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

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *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*.

Thomas Huber and Christina Niklaus. 2025. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246, Abu Dhabi, UAE. Association for Computational Linguistics.

Jonathan Jackson. 2025. Higher order prompting: Applying bloom's revised taxonomy to the use of large language models in higher education. *Studies in Technology Enhanced Learning*, 4(1).

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11260–11285. 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.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *Arxiv*, abs/2305.10263.

Yiming Luo, Ting Liu, Patrick Cheong-Iao Pang, Dana McKay, Ziqi Chen, George Buchanan, and Shanton Chang. 2025. Enhanced bloom's educational taxonomy for fostering information literacy in the era of large language models. *Arxiv*, abs/2503.19434.

Dung Nguyen Manh, Thang Phan Chau, Nam Le Hai, Thong T. Doan, Nam V. Nguyen, Quang Pham, and Nghi D. Q. Bui. 2024. Codemmlu: A multi-task benchmark for assessing code understanding capabilities of codellms. *Arxiv*, abs/2410.01999.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct elec-

tricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

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

Bhrij Patel, Souradip Chakraborty, Wesley A. Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. 2024. AIME: AI system optimization via multiple LLM evaluators. *Arxiv*, abs/2410.03131.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, and 81 others. 2025. Humanity's last exam. *Arxiv*, abs/2501.14249.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. 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*.

Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, Zekun Wang, Jian Yang, Zeyu Cui, Yang Fan, Yichang Zhang, Binyuan Hui, and Junyang Lin. 2025. Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings. *Arxiv*, abs/2501.01257.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, and 34 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Arxiv*, abs/2403.05530.

Sabina Elkins, Ekaterina Kochmar, Jackie C. K. Cheung, and Iulian Serban. 2024. How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 23084–23091. AAAI Press.

Pritish Sahu, Michael Cogswell, Ajay Divakaran, and Sara Rutherford-Quach. 2021. Comprehension based question answering using bloom's taxonomy. In *Proceedings of the 6th Workshop on Representation Learning for NLP, RepL4NLP@ACL-IJCNLP 2021, Online, August 6, 2021*, pages 20–28. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14182–14214. Association for Computational Linguistics.

Nicy Scaria, Suma Dharani Chenna, and Deepak N. Subramani. 2024. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *Artificial Intelligence in Education - 25th International Conference, AIED 2024, Recife, Brazil, July 8-12, 2024, Proceedings, Part II*, volume 14830 of *Lecture Notes in Computer Science*, pages 165–179. Springer.

Saeid Asgari Taghanaki, Ali Asghar Khani, and Amir Khasahmadi. 2024. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms. *Arxiv*, abs/2409.02257.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019a. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *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)*, pages 4149–4158. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019b. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

M.-A-P. Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, and 77 others. 2025. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. *Arxiv*, abs/2502.14739.

Andrew R Thompson and Logan PO Lake. 2023. 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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix,

Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Arxiv*, abs/2302.13971.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. 2024. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. 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*.

Sabine Ullrich and Michaela Geierhos. 2021. Using bloom's taxonomy to classify question complexity. In *4th International Conference on Natural Language and Speech Processing, Trento, Italy, November 12-13, 2021*, pages 273–277. Association for Computational Linguistics.

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*.

Antoun Yaacoub, Jérôme Da-Rugna, and Zainab Assaghir. 2025. Assessing ai-generated questions' alignment with cognitive frameworks in educational assessment. *arXiv preprint arXiv:2504.14232*.

Weixiang Yan, Haitian Liu, Yunkun Wang, Yunzhe Li, Qian Chen, Wen Wang, Tingyu Lin, Weishan Zhao, Li Zhu, Hari Sundaram, and Shuiguang Deng. 2024. Codescope: An execution-based multilingual multitask multidimensional benchmark for evaluating llms on code understanding and generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5511–5558. Association for Computational Linguistics.

Bo Yang, Qingping Yang, and Runtao Liu. 2024. Utmath: Math evaluation with unit test via reasoning-to-coding thoughts. *Arxiv*, abs/2411.07240.

Zhaojian Yu, Yilun Zhao, Arman Cohan, and Xiao-Ping Zhang. 2024. Humaneval pro and MBPP pro: Evaluating large language models on self-invoking code generation. *Arxiv*, abs/2412.21199.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zheyuan Zhang, Jifan Yu, Juanzi Li, and Lei Hou. 2023. Exploring the cognitive knowledge structure of large language models: An educational diagnostic assessment approach. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1643–1650. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

A Testing Benchmarks

The selection of the five categories of evaluation datasets was informed by extensive prior work(Zhong et al., 2024; Polo et al., 2024; Liang et al., 2023), as these datasets have been widely adopted in the literature to represent diverse task types across multiple domains. The benchmarks used in our study are selected from widely adopted evaluation suites for large language models (LLMs), including datasets from MMLU, AGIEval, CMMLU, BBH, and others. These benchmarks are recognized as standard and authoritative in evaluating LLM performance, covering a wide range of task types such as multi-choice QA, reasoning, math, code, and domain-specific knowledge, hence providing strong representativeness and task diversity.

Category	Dataset Name	Size
Mathematical Reasoning	GSM8K(Cobbe et al., 2021)	8,792
	UTMath(Yang et al., 2024)	1,053
	OpenAIMath(Muennighoff et al., 2025)	12,500
	AIME(Patel et al., 2024)	90
Code Reasoning	CodeScope(Yan et al., 2024)	1,949
	CodeMMLU(Manh et al., 2024)	19,913
	CruxEval(Gu et al., 2024a)	800
	MBPP(Yu et al., 2024)	964
	CodeElo(Quan et al., 2025)	408
Professional Exams	AGIEval (various)(Zhong et al., 2024)	7,272
	LastHumanExam(Phan et al., 2025)	2,500
	Omni-MATH(Gao et al., 2024)	4,428
	PutnamBench(Tsoukalas et al., 2024)	522
Broad-domain Evaluations	CMMLU(Li et al., 2024)	11,917
	MMLU(Hendrycks et al., 2021)	14,327
	MMLU-Pro(Wang et al., 2024)	12,102
	MMLU-Pro+(Taghanaki et al., 2024)	12,102
	M3KE(Liu et al., 2023)	20,477
Commonsense reasoning	CommonsenseQA(Talmor et al., 2019b)	10,962
	Xiezhi(Gu et al., 2024b)	54,530
	SuperGPQA(Team et al., 2025)	26,529
	mCSQA(Sakai et al., 2024)	13,636
	MMCU(Li et al., 2024)	9,380

Table 4: Categorized evaluation datasets

B The Diversity and Distributional Imbalance Results of Benchmarks

The table reports detailed results for several datasets under Bloom's Taxonomy conditions, where SWI, SDI, JSD, and GI denote the Shannon–Wiener Index, Simpson's Diversity Index, Jensen–Shannon Divergence, and Gini Coefficient, respectively. For SWI and SDI, higher values (shown in red) indicate more balanced cognitive distributions, whereas lower values (shown in blue) signal greater imbalance. In contrast, for JSD and GI, larger values (blue) correspond to more uniform distributions, and smaller values (red) reflect increased concentration. Overall, our experiments show that professionally designed exam items and commonsense reasoning benchmarks exhibit relatively even coverage across cognitive levels, while professional-task datasets display markedly more skewed distributions.

Table 5: Bloom's taxonomy distribution balance index in various domain groups for details

	Value / Dataset	SWI	SDI	JSD	GI
Mathematical Reasoning	**GSM8K**	0.2061	0.2940	0.7047	0.9329
	OpenAIMath	0.3879	0.5381	0.6339	0.8823
	AIME	0.4506	0.7095	0.6453	0.8632
	UTMath	0.3745	0.5960	0.6603	0.8918
Code Reasoning	**CodeScope**	0.5357	0.6970	0.5696	0.8174
	CruxEval	0.2865	0.4877	0.6898	0.9163
	MBPP	0.2470	0.3868	0.7007	0.9266
	CodeMMLU	0.6207	0.8134	0.5472	0.7761
	CodeElo	0.5466	0.7520	0.5856	0.8196
Professional Exams	**PutnamBench**	0.6125	0.8231	0.5661	0.7734
	Omni-MATH	0.5481	0.7691	0.5856	0.8198
	LastHumanExam	0.6657	0.8468	0.5226	0.7385
	AGIEval	0.6323	0.8384	0.5497	0.7607
Broad-domain Evaluations	**MMCU**	0.6487	0.8248	0.5258	0.7517
	CommonsenseQA	0.5324	0.7366	0.5818	0.8235
	SuperGPQA	0.5585	0.7722	0.5784	0.8137
	Xiezhi	0.4326	0.6267	0.6209	0.8661
	mCSQA	0.2862	0.3905	0.6673	0.9102
Commonsense Reasoning	**MMLU**	0.6084	0.8156	0.5590	0.7767
	CMMLU	0.5426	0.7408	0.5755	0.8163
	MMLUPro	0.5907	0.7963	0.5646	0.7937
	MMLUPro+	0.5893	0.7938	0.5652	0.7950
	M3KE	0.5428	0.7312	0.5700	0.8145

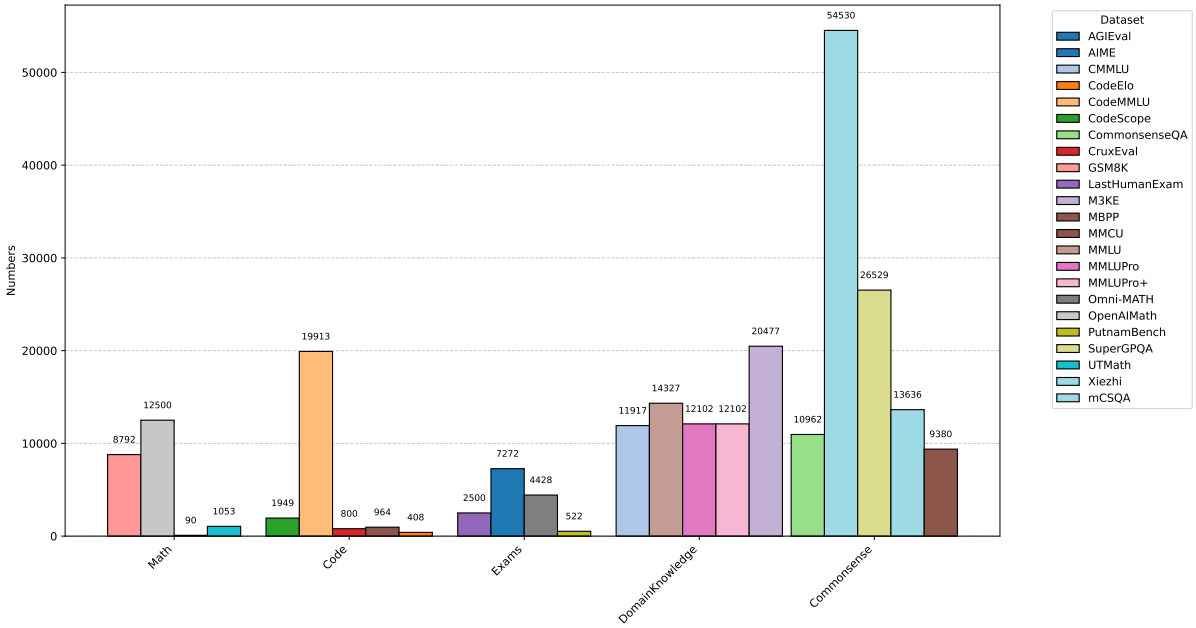


Figure 7: Statistics for each group

C Prompt Templates

We address the potential subjectivity in Bloom's Taxonomy-based categorization by using a standardized prompting protocol, which ensures consistent and deterministic labeling across annotators(Chen et al., 2024). Indeed, Bloom's Taxonomy-based categorization can involve a degree of subjectivity(Ullrich and Geierhos, 2021; Huber and Niklaus, 2025; Chen et al., 2024; Sahu et al., 2021). To mitigate this, we adopted a standardized prompting protocol for annotation: all evaluators were instructed to label each instruction according to a fixed set of Bloom-aligned definitions and criteria, ensuring consistent interpretation across tasks. This approach leads to deterministic labeling once the prompt and instruction are fixed, reducing variability in judgment. This methodology is consistent with recent work on aligning evaluation frameworks with cognitive taxonomies, where standardized prompts effectively ground task classification across annotators.

C.1 Labeling Prompts

System Template: 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>

User Template: Assume you are a data expert. Please classify the following questions according to Bloom's taxonomy.
The revised version of Bloom's taxonomy is divided into two dimensions: cognitive and knowledge.
The cognitive dimensions include:

- **Remembering**: Retrieving, recognizing, and recalling relevant knowledge from long-term memory.
- **Understanding**: Constructing meaning from information through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.
- **Applying**: Carrying out or using a procedure for executing or implementing in a given situation.

- **Analyzing**: Breaking material into its constituent parts and determining how those parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.
- **Evaluating**: Making judgments based on criteria and standards through checking and critiquing.
- **Creating**: Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.

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 questions with bloom's taxonomy**,**DONOT SLOVE THE PROBLEM**, provide your thought process,and then give the answer.
Here is an example:

==========

Input:

```
{
  "passage": "The capital of France is Paris.",
  "question": "What is the capital of France?"
}
```

Output:

```
<think>
The question is about the capital of France, which is a factual question. The answer is Paris.
</think><answer>(Remebering,Factual)</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 IRRELATIVE THINGS, and multiple tags cannot be generated.
Now begin your inputs:

C.2 Fine-Tuning Prompts

System Template: 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>

D Generated samples

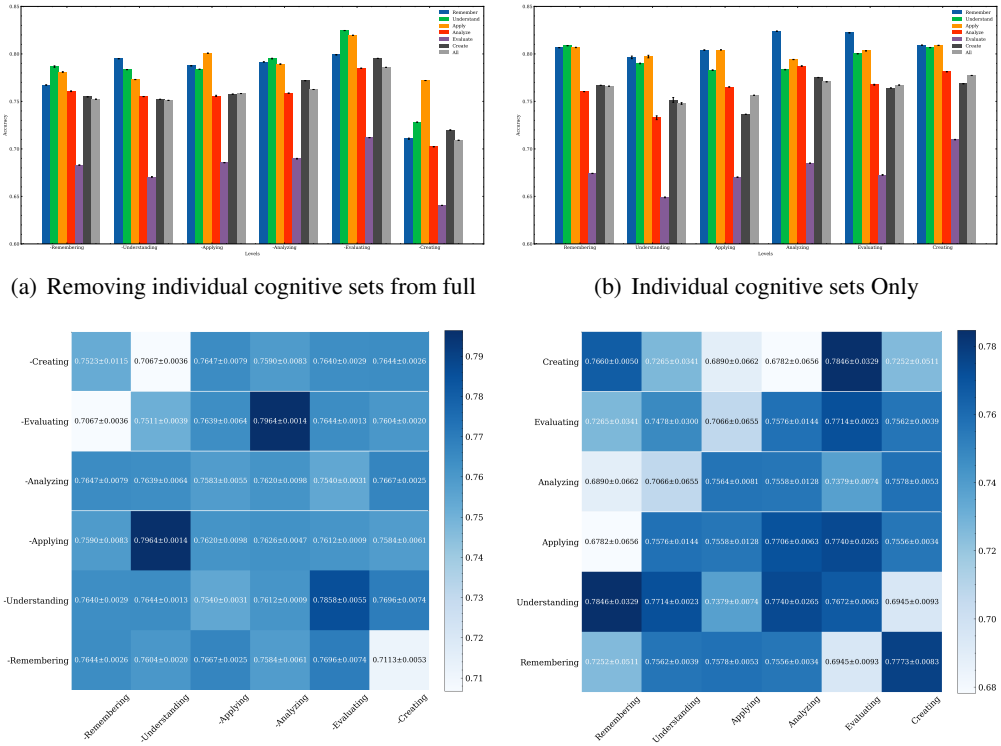
Table 6: Examples of evaluation and training

Knowledge\Cognitive	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual	What is the definition of entropy in thermodynamics?	How does electric current relate to voltage? Explain their relationship.	How can you use the formula of kinetic energy to calculate the motion of a car?	Can you distinguish between mass and weight in physics?	How would you assess the accuracy of Newton's Laws in modern mechanics?	Can you construct an example where force results in no motion?
	A: A measure of the total energy in a system.	A: Voltage causes current to flow through a conductor.	A: By calculating the work done by friction.	A: Mass is the amount of matter; weight is the force due to gravity on that matter.	A: They are completely outdated and no longer used.	A: A book resting on a table with gravity pulling down and table pushing up.
	B: A measure of disorder or randomness in a system.	B: Current causes voltage to appear in a circuit.	B: By using the formula KE = 1/2 mv^2, where m is mass and v is velocity.	B: Mass and weight are exactly the same.	B: They are accurate for everyday objects at low speeds but less accurate at very high speeds or small scales.	B: A car accelerating on a highway.
	C: The force exerted by a system per unit area.	C: Voltage and current are unrelated.	C: By measuring the car's temperature change.	C: Weight measures volume; mass measures density.	C: They perfectly describe all physical phenomena.	C: A ball thrown in the air.
	D: The velocity of particles in a system.	D: Voltage decreases as current increases.	D: By estimating the potential energy at a height.	D: Weight is constant regardless of location; mass changes.	D: They are only applicable in outer space.	D: A person running.
	Answer: B	Answer: A	Answer: B	Answer: A	Answer: B	Answer: A
Conceptual	What is the primary purpose of probability distributions in statistics?	What role do logical quantifiers play in symbolic reasoning?	How would you apply Bayes' theorem to update beliefs?	How can you differentiate between deductive and inductive reasoning?	How would you judge the validity of a syllogistic argument?	Can you formulate a scenario that demonstrates causal inference?
	A: To list all possible outcomes without assigning likelihoods	A: They represent numerical calculations	A: By calculating conditional probabilities based on new evidence	A: Deductive reasoning derives conclusions guaranteed by premises; inductive reasoning generalizes from specific examples	A: By checking if the conclusion logically follows from the premises	A: Studying how smoking causes lung cancer through observational data
	B: To describe how probabilities are assigned to different possible outcomes	B: They specify the quantity of elements that satisfy a property	B: By ignoring prior information	B: Deductive reasoning generalizes from data; inductive reasoning derives conclusions from axioms	B: By measuring the emotional appeal of the argument	B: Listing all symptoms of a disease
	C: To collect raw data from experiments	C: They identify statistical distributions	C: By multiplying probabilities without conditions	C: Both are forms of guesswork	C: By counting the number of words in the argument	C: Describing the parts of a cell
	D: To solve algebraic equations	D: They measure probability	D: By listing all possible hypotheses	D: Inductive reasoning uses logic symbols; deductive reasoning uses statistics	D: By comparing it to scientific data	D: Calculating averages in a dataset
	Answer: B	Answer: B	Answer: A	Answer: A	Answer: A	Answer: A
Procedural	What is the first step in solving a linear equation?	Why is it necessary to understand algorithmic complexity when designing solutions?	Can you apply the Euclidean algorithm to find the GCD of 56 and 98?	Which of the following steps in the quicksort algorithm typically causes the greatest efficiency bottleneck?	How would you evaluate the correctness of a recursive function?	Can you design a decision tree to solve a classification problem with multiple features?
	A: Isolate the variable	A: To make the program easier to debug	A: 14	A: Partitioning the array	A: By checking if it runs without errors	A: Yes, by selecting features, creating branches based on feature values, and defining leaf nodes with classes
	B: Simplify both sides	B: To determine how well the solution performs at scale	B: 7	B: Choosing the pivot	B: By using test cases to verify base and recursive cases	B: Yes, by sorting all input data
	C: Multiply both sides by the same number	C: To reduce the amount of memory usage	C: 28	C: Recursively sorting subarrays	C: By comparing it to an iterative version	C: Yes, by drawing a graph and labeling all nodes randomly

See the next page

Knowledge\Cognitive	Remember	Understand	Apply	Analyze	Evaluate	Create
	D: Graph the equation Answer: B	D: To choose the correct programming language Answer: B	D: 2 Answer: A	D: Copying elements to a temporary array Answer: B	D: By measuring its runtime Answer: B	D: Yes, by performing linear regression first Answer: A
Metacognitive	Which of the following best illustrates a time when your strategy for solving a problem failed? A: You followed a proven method and succeeded immediately. B: You used a familiar method, but it did not work and you had to revise it. C: You guessed randomly and happened to get it right. D: You let someone else solve the problem for you. Answer: B	Why is it important to reflect on your cognitive biases during reasoning? A: Because reflection helps identify and correct flawed thinking patterns. B: Because cognitive biases are always useful in decision-making. C: Because reflecting slows down the problem-solving process. D: Because it helps you avoid doing any reasoning at all. Answer: A	How can you implement a self-monitoring routine while solving logic puzzles? A: By writing down the answer immediately. B: By solving as many puzzles as quickly as possible. C: By regularly checking your thought process and correcting mistakes as you go. D: By skipping over hard problems to save time. Answer: C	Which step would help you break down your own approach to proof-writing to identify flaws? A: Ignoring the logic used in each step. B: Focusing only on the final answer. C: Reviewing each assumption and step in your proof critically. D: Skipping the analysis phase to save time. Answer: C	How would you critique your problem-solving plan after getting an incorrect answer? A: Blame the problem for being too difficult. B: Identify which parts of your plan worked and which led to errors. C: Ignore the mistake and move on. D: Repeat the same plan without changes. Answer: B	What is the best way to devise a personal checklist to improve learning from complex reasoning tasks? A: Write down unrelated tasks to feel productive. B: Include steps for planning, monitoring, and reviewing your reasoning process. C: Avoid listing your mistakes to maintain confidence. D: Rely on memory instead of writing anything down. Answer: B

E The Instruction Fine-Tuning Results



(a) Removing individual cognitive sets from full

(b) Individual cognitive sets Only

(c) Removing two cognitive sets from full

(d) Two cognitive sets Only

Figure 8: Data gain effect diagram

The following results provide a detailed account of the fine-tuned models on individual cognitive dimensions or cognitive groups, as shown in Figure 8.