

CogBench: Benchmarking Cognitive Alignment of Large Language Models in Educational Question Answering

Tong Lu¹, Zhichun Wang^{*1,2,3}, Yuanhao Sun¹, Yaoyu Zhou¹
Mingrui Li¹, Yiming Guan¹ and Zhiyong Bai⁴

¹ School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China

² Beijing Key Laboratory of Artificial Intelligence for Education, Beijing 100875, China

³ Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing 100875, China

⁴ Faculty of Education, Beijing Normal University, Beijing 100875, China
ethanlu@mail.bnu.edu.cn, zcwang@bnu.edu.cn

Abstract

Large language models (LLMs) possess strong capabilities in language understanding and generation, as well as remarkable problem-solving abilities. In the educational domain, a representative application is to employ LLMs as learning assistants that answer students' questions and support their learning processes. In such scenarios, it is crucial for the model to perceive a student's cognitive level and provide explanations that are appropriate to that level. However, whether LLMs can effectively accomplish this task has not yet been thoroughly investigated. To address this gap, we introduce CogBench, an evaluation benchmark designed to assess the cognitive alignment capabilities of LLMs in educational QA. CogBench comprises 2.1K mathematics questions, each associated with multiple valid solutions that rely on knowledge and reasoning at different cognitive levels. Building on this structure, we formulate three cognition-aware evaluation tasks and propose three complementary metrics to quantify cognitive alignment from multiple perspectives. Extensive experiments on 11 representative LLMs reveal that, while models can often produce correct answers, they still struggle to consistently generate explanations that are aligned with the intended cognitive level. These results highlight substantial room for improvement and establish CogBench as a diagnostic benchmark for advancing cognitively aligned educational AI systems. The code and dataset are released at <https://github.com/kg-bnu/CogBench>.

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities in natural language understanding and generation, leading to their rapid adoption in educational applications such as intelligent tutoring systems (Scarlatos et al., 2025), automated question answering (Luo et al., 2024), and learning

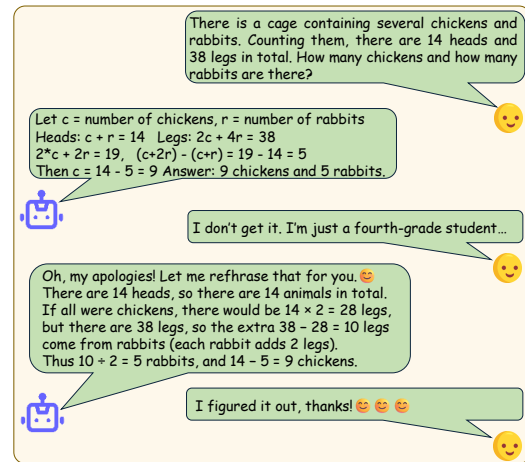


Figure 1: Illustration of cognitive alignment when using LLMs in educational QA.

assistance tools (Aisyah et al., 2025). In these scenarios, LLMs are increasingly expected not only to provide correct answers, but also to facilitate learning by providing explanations that are pedagogically effective and adaptive to individual learners. This shift positions LLMs as potential personalized tutors rather than mere information retrieval systems, raising new challenges.

A fundamental requirement of effective instruction is **cognitive alignment**: explanations should be tailored to the learner's current knowledge state and cognitive level. Given the same question, learners at different cognitive levels require substantially different forms of explanation. For example (as illustrated in Figure 1), for a classic *heads-and-legs* word problem, in which the numbers of animals are inferred from the total counts of heads and legs, an explanation for lower-grade elementary students typically relies on concrete reasoning or step-by-step trial and error, whereas middle school students are often taught to formulate and solve the same problem using algebraic equations. However, despite their strong general reasoning abilities, existing LLMs tend to produce explanations with a

*Corresponding author.

fixed level of abstraction, resulting in responses that may be overly complex for younger learners or insufficiently rigorous for more advanced students. This cognitive mismatch limits the effectiveness of LLMs as educational assistants.

Prior research on LLMs in education has primarily focused on answer correctness, reasoning accuracy, or overall helpfulness. Existing evaluation benchmarks such as C-EVAL (Huang et al., 2023), CMMath (Li et al., 2025b), EduBench (Xu et al., 2025), E-EVAL (Hou et al., 2024) and MDK12-Bench (Zhou et al., 2025) assess the problem-solving abilities of LLMs across diverse educational stages, various teaching scenarios, and multimodal educational questions. However, these benchmarks typically include different questions for different grades or levels, making it difficult to disentangle problem difficulty from explanation adaptability. As a result, they do not systematically assess whether LLMs can generate explanations that are cognitively aligned with learners when faced with the same question. Moreover, existing evaluation protocols rarely examine explanation appropriateness in terms of knowledge usage and reasoning depth, leaving a critical gap in the understanding of LLM behavior in pedagogical contexts.

In this work, we investigate the cognitive alignment of LLMs in educational question answering. Specifically, we examine whether, given an identical question and an explicit specification of the learner’s cognitive level, an LLM can generate responses that are not only factually correct but also appropriate in terms of knowledge complexity and reasoning depth. To support this study, we introduce **CogBench**, a benchmark designed to systematically evaluate the cognitive alignment capabilities of LLMs in educational scenarios. CogBench is constructed by transforming existing K–12 mathematical questions into a one-question–multiple-solutions form, where each problem is associated with multiple valid solution traces that rely on different reasoning strategies and knowledge abstractions. This formulation enables controlled comparison of model behavior across cognitive levels without confounding factors introduced by varying question difficulty.

CogBench assesses the adaptive capacity of LLMs through three cognition-aware QA tasks: (i) direct question answering without cognitive constraints, (ii) grade-constrained solution generation requiring alignment with a specified student grade, and (iii) knowledge-constrained solution

generation that restricts reasoning to designated curriculum knowledge. Together, these tasks assess not only whether models can solve problems, but whether they can adapt how they solve them in response to cognitive constraints. To quantitatively measure cognitive alignment, we further propose three complementary evaluation metrics: *Cognitive Accuracy*, which measures answer correctness under cognitive constraints; *Knowledge Consistency*, which assesses the extent and appropriateness of knowledge utilization; and *Knowledge Divergence*, which quantifies variations in knowledge usage across different cognitive levels.

Our experimental results show that mainstream LLMs prioritize factual correctness, often at the expense of cognitive alignment. Cognition-aware fine-tuning substantially improves alignment with learners’ cognitive levels, but introduces a slight reduction in standard accuracy, revealing an inherent trade-off between correctness and pedagogical appropriateness. These findings underscore both the limitations of current LLMs as educational tutors and the value of targeted training for improving their pedagogical reliability.

In summary, our contributions are threefold:

(1) We formalize cognitive alignment as a key requirement for educational QA, highlighting the need for LLM-generated explanations to match learners’ cognitive levels in reasoning depth and curriculum knowledge, beyond answer correctness.

(2) We introduce CogBench, a benchmark that transforms K–12 mathematics problems into a one-question–multiple-solutions format, enabling systematic evaluation of LLMs’ ability to adapt explanations under unconstrained, grade-constrained, and knowledge-constrained settings.

(3) We propose three metrics for evaluating the LLMs’ capability of cognitive alignment, including Cognitive Accuracy, Knowledge Consistency, and Knowledge Divergence. We conduct extensive experiments on 11 LLMs and provide fine-grained analyses of their cognitive alignment behavior through both automatic metrics and human evaluation.

2 CogBench Benchmark

CogBench is designed to evaluate the cognitive alignment of LLMs in educational question answering by reformulating K–12 mathematics problems into a one-question–multiple-solutions benchmark, where each question is associated with several

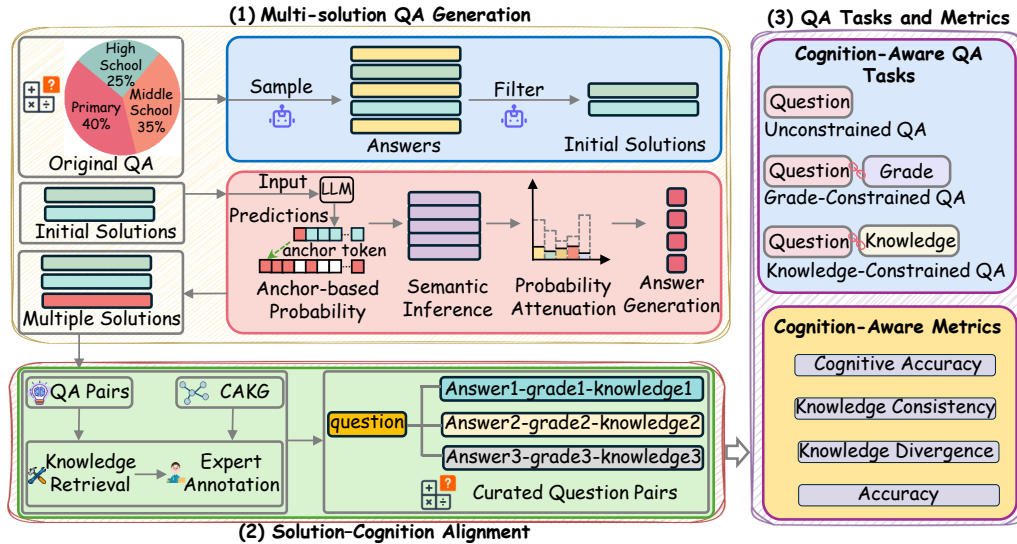


Figure 2: The overview of our framework.

correct solution traces reflecting different reasoning strategies and levels of knowledge abstraction. Each solution is further aligned with curriculum-defined knowledge components and grade levels using a curriculum-grounded knowledge graph, enabling fine-grained assessment of whether model-generated explanations appropriately match specified cognitive constraints. As illustrated in Figure 2, CogBench consists of three core components: multi-solution QA generation, solution-cognition alignment, and cognition-aware evaluation tasks and metrics.

2.1 Multi-solution QA Generation

To obtain multiple valid solutions for each question, we adopt a two-step generation process that encourages solution diversity by manipulating decoding hyperparameters and applying probability attenuation to the model’s output distribution.

2.1.1 Multi-turn Solution Sampling via Controlled Decoding

We first generate diverse solution traces using a multi-turn sampling strategy inspired by (Wang et al., 2023). Specifically, each question is input into an LLM multiple times, with decoding parameters including *top-k sampling*, *temperature*, and *nucleus sampling* varied to encourage diverse reasoning paths. The generated responses are evaluated against the corresponding gold-standard answers, and only those yielding correct final answers are retained. This step produces an initial set of valid solutions that differ in reasoning trajectories and knowledge usage.

2.1.2 Generating Additional Solutions via Probability Attenuation

While LLMs encode substantial knowledge in their parameters, their outputs are often biased toward high-probability tokens and are sensitive to prompt variations (Li et al., 2025a). To further explore the internal knowledge of language models and encourage the generation of distinctive solutions, we introduce a probability attenuation mechanism that reduces the likelihood of reproducing previously generated or semantically similar answers. Specifically, we identify key concepts in existing solutions as anchor tokens (Chang et al., 2023) and construct an anchor-based probability distribution. Using the Moore–Penrose pseudoinverse of the model’s embedding matrix, we infer a semantic representation of previously generated answers and attenuate their corresponding token probabilities during decoding. The model then samples from the adjusted probability distribution, encouraging the discovery of novel yet valid solutions that may rely on different knowledge and reasoning patterns.

2.2 Curriculum-Grounded Cognition Level Alignment

To enable principled mapping between solution strategies and cognitive levels, we construct a curriculum-based, cognition-aware knowledge graph and use it to align multiple solutions of the same question with different levels of expected mathematical understanding. Automatic alignment based on semantic similarity is further validated and refined by education experts, resulting in a

high-quality, cognition-aware evaluation dataset.

2.2.1 Cognition-Aware Knowledge Graph

We extract cognition-aware mathematical knowledge spanning the full K–12 curriculum from official educational standards. Since students’ cognitive abilities develop progressively from elementary to advanced stages, we organize the extracted knowledge into a *cumulative* knowledge graph, where subgraphs corresponding to higher grade levels subsume those of lower grades and encode increasingly sophisticated reasoning requirements.

Unlike prior educational knowledge graphs that primarily focus on abstract concepts or topic hierarchies (Jhajj et al., 2024; Sun et al., 2024; Lu et al., 2026), the proposed Cognition-Aware Knowledge Graph (CAKG) emphasizes fine-grained problem-solving knowledge by explicitly modeling solution strategies and procedural reasoning patterns. By associating each triple with specific grade levels and curricular requirements, the CAKG provides an explicit bridge between solution content and expected cognitive proficiency.

2.2.2 Solution–Cognition Alignment

Given multiple valid solutions for the same question, our objective is to assign each solution to an appropriate cognitive level according to the knowledge components and reasoning strategies it employs, rather than its final correctness alone. To this end, we align solution traces with the curriculum-grounded CAKG via semantic matching in a shared representation space. Formally, we encode both CAKG triples and solution texts using the Qwen3 Embedding model (Zhang et al., 2025). For each knowledge triple $t_i \in \mathcal{T}$ and solution s_j , we obtain vector representations:

$$\mathbf{V}_{t_i} = \text{Enc}(t_i), \mathbf{V}_{s_j} = \text{Enc}(s_j) \quad (1)$$

where $\text{Enc}(\cdot)$ denotes the embedding encoder. Each solution embedding \mathbf{V}_{s_j} is treated as a semantic query, and its relevance to curriculum knowledge is measured by cosine similarity:

$$\text{Sim}(s_j, t_i) = \frac{\mathbf{V}_{s_j}^\top \mathbf{V}_{t_i}}{\|\mathbf{V}_{s_j}\|_2 \|\mathbf{V}_{t_i}\|_2}. \quad (2)$$

We then retrieve the top- k knowledge triples most relevant to s_j , which is denoted as $\mathcal{R}_k(s_j)$. Since each knowledge triple in the CAKG is associated with one or more grade levels reflecting curriculum expectations, the retrieved set $\mathcal{R}_k(s_j)$ induces a candidate cognitive level for the solution.

To ensure annotation reliability, we adopt a human-in-the-loop curation procedure in which education experts review the retrieved knowledge triples and the induced grade-level assignments. Experts verify semantic relevance, correct potential mismatches, and refine the final solution–knowledge–grade alignment. This expert validation step mitigates embedding noise and curriculum ambiguity, resulting in a high-quality benchmark that supports fine-grained evaluation of cognitive alignment.

2.3 Cognition-Aware QA Tasks and Metrics

In CogBench, we design three complementary QA tasks with increasing levels of cognitive constraint. These tasks probe different aspects of a model’s ability to control and adapt its reasoning depth and knowledge usage.

Task 1: Unconstrained QA. In this setting, the model is provided only with the problem statement and is asked to produce a solution without any explicit cognitive guidance. This task evaluates the model’s default problem-solving behavior and serves as a strong baseline for comparison.

Task 2: Grade-Constrained QA. In this task, the model is given the problem statement together with a specified grade level, and is instructed to generate a solution that can be understood by a student at that grade.

Task 3: Knowledge-Constrained QA. In this task, the model is provided with the problem statement and an explicit set of required knowledge, and is instructed to solve the problem using only the specified knowledge.

To evaluate the cognitive alignment of large language models beyond standard answer accuracy (ACC), we introduce three complementary metrics: Cognitive Accuracy (CA), Knowledge Consistency (KC), and Knowledge Divergence (KD). Together, these metrics assess not only whether a model produces correct answers, but also whether its reasoning and knowledge usage are consistent with curriculum-defined cognitive expectations.

Cognitive Accuracy (CA). CA measures the proportion of responses that are both *answer-correct* and *cognitively aligned*. A response is considered incorrect if the model either produces an incorrect final answer or fails to generate a solution that conforms to the required cognitive level. Formally,

$$\text{CA} = \frac{\#(\text{cognition-aligned correct answers})}{\#(\text{total answers})}. \quad (3)$$

Knowledge Consistency (KC). KC quantifies how well a model’s solution adheres to the curriculum-prescribed knowledge at a given grade level. Given a question Q and a target grade g , let K_g denote the set of knowledge components associated with grade g in the CAKG, and let K_m denote the set of knowledge components adopted by the model. KC is defined as

$$\text{KC} = \frac{|K_m \cap K_g|}{|K_g|}. \quad (4)$$

Knowledge Divergence (KD). KD measures the extent to which a model differentiates its knowledge usage across cognitive levels. Intuitively, a cognitively aligned model should employ increasingly distinct sets of knowledge components as the target grade level changes. We compute KD using the average pairwise Jaccard distance between the knowledge sets used at different cognitive levels:

$$\text{KD} = \frac{2}{g(g-1)} \sum_{1 \leq i < j \leq g} \left(1 - \frac{|K_i \cap K_j|}{|K_i \cup K_j|} \right), \quad (5)$$

where g denotes the total number of cognitive levels (grades), and K_i represents the set of knowledge components used by the model for grade level i . The normalization factor ensures that KD lies in the range $[0, 1]$, with higher values indicating greater differentiation in knowledge usage across cognitive levels.

When evaluating LLMs, the generated answers and their associated knowledge are automatically compared with the answers and annotated knowledge at different cognitive levels for the same questions using these described evaluation metrics.

2.4 Benchmark Statistics

We curate and annotate 2.1K math problems, including 1.2K open-source Olympiad problems¹ and 0.9K problems from CMMath (Li et al., 2025b), spanning primary, middle, and high school levels across K–12. Specifically, we use Qwen3-30B-A3B as the base model to generate multiple solutions using the method illustrated in Figure 2(1). On average, each question has 2.16 solutions in different cognitive levels and 3.2 knowledge components. As shown in Figure 3 and Figure 4, each question includes **at least two solutions** at different cognitive levels, and the dataset spans primary (grades 1–6), middle (grades 7–9), and high school (grades 10–12), accounting for 40%, 35%, and 25% of the problems, respectively.

¹<https://www.aoshuku.com/>

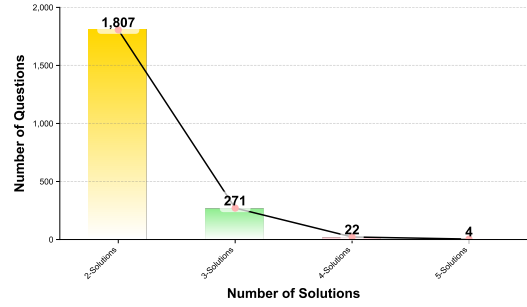


Figure 3: Question distribution by number of solutions.

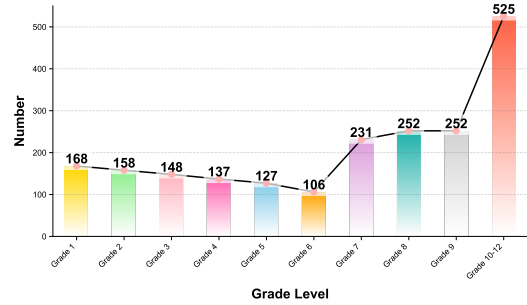


Figure 4: Question distribution by cognitive levels.

3 Experiments

We conduct comprehensive evaluations of the cognitive alignment of LLMs, with the goal of addressing the following research questions:

RQ1: To what extent can mainstream LLMs adapt their responses to align with learners at different cognitive levels?

RQ2: To what extent does model fine-tuning improve LLMs’ ability to generate cognitively aligned responses?

RQ3: To what extent are different performance aspects of LLMs correlated, and what do these correlations reveal about the coupling or decoupling of these abilities in educational question answering?

RQ4: To what extent do the proposed automatic metrics agree with human judgments from educational experts in assessing cognitive alignment?

3.1 Evaluated LLMs

We evaluated 11 LLMs including 6 open-source LLMs and 5 proprietary advanced LLMs. Open-source LLMs are DeepSeek-Chat (DeepSeek-AI et al., 2025), GPT-OoS-20B (OpenAI et al., 2025), Glm4 (GLM et al., 2024), Glm4.5 (Sun et al., 2024), Qwen3-30B-A3B (Yang et al., 2025) and Qwen3-235B-A22B (Yang et al., 2025). Proprietary LLMs are GPT-5 (OpenAI, 2025), GPT-5-Min (OpenAI, 2025), Claude-4-Opus (Anthropic, 2025), Claude-4-Sonnet (Anthropic, 2025),

Model	Unconstrained			Grade-Constrained				Knowledge-Constrained		
	ACC	CA	KC	ACC	CA	KC	KD	ACC	CA	KC
Claude-4-Opus	0.804	0.433	0.546	0.773	0.536	0.753	0.778	0.691	0.430	0.557
Claude-4-Sonnet	0.773	0.423	0.557	0.742	0.505	0.722	0.790	0.687	0.361	0.517
DeepSeek-V3	0.845	0.476	0.567	0.845	0.520	0.614	0.755	0.850	0.469	0.542
Gemini-3-Pro	0.889	0.534	0.604	0.857	0.560	0.642	0.596	0.877	0.473	0.539
Glm-4	0.824	0.460	0.550	0.837	0.512	0.615	0.738	0.852	0.467	0.524
Glm-4.5	0.853	0.481	0.574	0.841	0.519	0.611	0.756	0.858	0.462	0.548
GPT-5	0.942	0.517	0.552	0.918	0.542	0.591	0.687	0.928	0.466	0.497
GPT-5-Mini	0.861	0.505	0.587	0.852	0.536	0.631	0.579	0.848	0.452	0.523
GPT-OoS-20B	0.886	0.525	0.592	0.854	0.529	0.627	0.561	0.859	0.460	0.529
Qwen3-235B-A22B	0.881	0.517	0.593	0.865	0.538	0.620	0.731	0.900	0.442	0.487
Qwen3-30B-A3B	0.873	0.511	0.587	0.860	0.521	0.610	0.724	0.891	0.438	0.487

Table 1: Model Performance Comparison. ACC = Accuracy, CA = Cognitive Accuracy, KC = Knowledge Consistency, and KD = Knowledge Divergence. Knowledge divergence (KD) quantifies differences in the model’s knowledge utilization across prompts with varying cognitive levels and is therefore computed only under grade-constrained settings.

Gemini-2.5-Flash and Gemini-3-Pro. We evaluated LLMs through API endpoints in September and December 2025.

3.2 Main results (RQ1)

Table 1 summarizes the overall performance of 11 LLMs. Under the unconstrained QA setting, where no cognitive-level cues are provided and models must implicitly infer the appropriate level from the question alone, both cognitive accuracy and knowledge consistency are substantially lower than standard answer accuracy. Specifically, the highest cognitive accuracy and knowledge consistency reach only 0.534 and 0.604, respectively, compared to a peak accuracy of 0.942. These results indicate that, although LLMs can occasionally produce responses that align with the cognitive level implied by the original exam questions, such alignment is incidental rather than systematic. The large gap between accuracy and cognition-aware metrics suggests that models predominantly rely on advanced parametric knowledge acquired during pretraining, even when simpler or more curriculum-aligned reasoning would be more appropriate. Overall, these findings highlight that, *in the absence of explicit cognitive constraints, mainstream LLMs still struggle to achieve fine-grained and reliable alignment with the intended cognitive levels of educational questions.*

When explicit cognitive-level specifications are provided, LLMs exhibit moderate but consistent improvements in cognitive alignment, whereas con-

straining models with curriculum-aligned knowledge does not yield the same benefit. As shown in Table 1, prompting LLMs with a specified cognitive level increases the best cognitive accuracy and knowledge consistency to 0.560 and 0.753, respectively, compared to the unconstrained QA setting. In addition, the highest knowledge divergence rises to 0.790, indicating that models are more sensitive to cognitive-level distinctions and can adjust their reasoning patterns across different learner specifications. In contrast, explicitly providing cognitively aligned knowledge leads to a degradation in alignment metrics. This counterintuitive effect can be attributed to the fact that knowledge injection tends to activate higher-level reasoning patterns and more specialized linguistic expressions associated with that knowledge during pretraining (Murthy et al., 2025). As a result, models are biased toward advanced parametric knowledge, making it difficult to flexibly adapt their explanations to the intended cognitive level, even when the injected knowledge itself is curriculum-aligned.

3.3 Results of Finetuned LLMs (RQ2)

To examine whether the cognitive alignment capability of LLMs can be enhanced through training, we fine-tune the Qwen3-30B-A3B model using the benchmark dataset, which is split into training, validation, and test sets with a 4:3:3 ratio. As shown in Table 2, fine-tuning leads to clear improvements in both cognitive accuracy and knowledge consistency, while standard accuracy and

knowledge divergence exhibit slight decreases. It indicates that *training explicitly oriented toward cognitive alignment encourages the model to prioritize curriculum-appropriate reasoning and knowledge usage, even at the cost of reduced reliance on advanced parametric knowledge that may yield higher standard accuracy*. The reduction in knowledge divergence further suggests that fine-tuning narrows the model’s solution space, making its reasoning strategies more consistent across cognitive levels rather than maximizing differentiation. While this may limit diversity in knowledge usage, it reflects a stronger inductive bias toward generating explanations that conform to specified cognitive constraints. Overall, these results demonstrate that fine-tuning introduces an inherent trade-off between factual optimization and cognitive alignment, with the latter benefiting from targeted training.

We further analyze the individual contributions of supervised fine-tuning (SFT) and direct preference optimization (DPO). The comparison between Qwen-FT w/o DPO and Qwen-FT w/o SFT reveals that DPO plays a dominant role in improving cognitive alignment, whereas SFT alone yields only marginal gains. This suggests that, although SFT enables the model to encode multi-level cognitive knowledge, the absence of explicit generative policy constraints guided by cognitive preferences limits its ability to reliably produce cognitively aligned responses. In contrast, DPO effectively enforces alignment at the generation level by optimizing the model toward cognitively preferred outputs. Additional details of the fine-tuning procedure are provided in Appendix A.

Model	ACC	CA	KC	KD
Qwen (base model)	0.88	0.47	0.54	0.72
Qwen-FT	0.83	0.63	0.68	0.61
w/o DPO	0.86	0.46	0.54	0.69
w/o SFT	0.85	0.57	0.62	0.65

Table 2: Comparison of cognitive alignment performance before and after fine-tuning. Qwen = Qwen3-30B-A3B, Qwen-FT = fine-tuned Qwen3-30B-A3B, w/o DPO = Qwen-FT_{w/o DPO}, w/o SFT = Qwen-FT_{w/o SFT}.

3.4 Inter-correlation Among Accuracy, Cognitive Alignment, and Knowledge Usage (RQ3)

As illustrated in Figure 5, a significant positive linear correlation is observed between cognitive

accuracy and standard accuracy, with the latter consistently exceeding the former. It indicates that, although models capable of producing correct answers tend to achieve higher cognitive accuracy, factual correctness alone does not guarantee cognitive alignment. Instead, *LLMs exhibit a systematic bias toward prioritizing answer correctness over generating responses that are aligned with learners’ cognitive levels*.

In contrast, cognitive accuracy shows a strong positive correlation with knowledge consistency, suggesting that the appropriate and curriculum-aligned use of knowledge is a critical prerequisite for producing cognitively aligned responses. This finding highlights that *cognitive alignment is closely coupled with how knowledge is utilized, rather than merely whether an answer is correct*.

Interestingly, knowledge divergence exhibits little to no correlation with the other metrics. This observation reflects the fact that knowledge divergence captures a distinct aspect of model behavior, which is largely independent of both answer correctness and alignment quality.

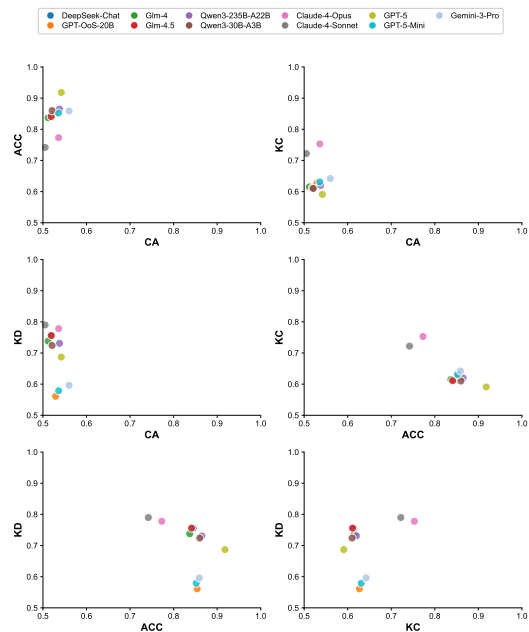


Figure 5: The relationship among diverse metrics.

3.5 Results of Human Evaluation (RQ4)

To validate the effectiveness of the proposed automatic metrics, we conduct a human evaluation involving five experts in education. We randomly sample 500 questions and ask the evaluators to assess the cognitive alignment of responses generated by all models, focusing on the appropriateness and

Metric	Criteria	Score
Consistency	Evaluates whether the knowledge level employed by LLMs in answering pedagogical questions aligns with the cognitive levels of students.	0 – 5
Diversity	Evaluates whether the knowledge used by LLMs when answering pedagogical questions varies according to the cognitive level of the students.	0 – 5

Table 3: Human evaluation metrics.

differentiation of reasoning across cognitive levels. Table 3 shows the human evaluation metrics and criteria.

Figure 6 shows that results of human evaluation. Comparing to the results in Table 1, the trends observed in human judgments of Consistency and Diversity closely mirror those measured by the automatic metrics cognitive accuracy and knowledge divergence, respectively. This qualitative agreement suggests that *the proposed metrics capture key aspects of cognitive alignment that are also recognized by human evaluators.*

To further quantify the agreement between human and automatic evaluations, we compute the Pearson correlation coefficient (normalized to 100) for each model, as illustrated in Figure 7. The results show strong and consistent correlations between cognitive accuracy and human-rated Consistency, as well as between knowledge divergence and human-rated Diversity, across different LLMs. These findings indicate that *the proposed automatic metrics reliably reflect expert judgments of cognitive alignment and demonstrate substantial potential as scalable alternatives to human evaluation in assessing the cognitively aligned capabilities of LLMs.*

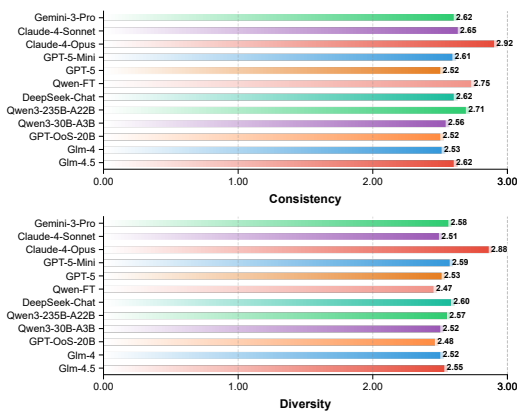


Figure 6: Results of human evaluation.

4 Related Work

Educational Benchmarks. Existing benchmarks primarily assess LLM via accuracy on educational

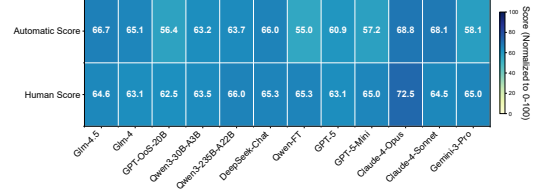


Figure 7: Correlation between automatic and human evaluations.

questions across disciplines, including EXAMS-V (Das et al., 2024), E-EVAL (Hou et al., 2024), CMATH (Wei et al., 2023), M3KE (Liu et al., 2023), CMMaTH (Li et al., 2025b), MathEval (Liu et al., 2025a), Mathador-LM (Kurtic et al., 2024), MDK12-Bench (Zhou et al., 2025), and DeepMath-103K (He et al., 2025). Some benchmarks further annotate knowledge structures or domains (e.g., CK12 (You et al., 2024), MathBench (Liu et al., 2024), MATHSCAPE (Liang et al., 2025)) or evaluate educational problem-solving by treating LLMs as question creators (e.g., Dr.Academy (Chen et al., 2024), EduBench (Xu et al., 2025)). However, none explicitly assess LLMs’ cognitive alignment across learners with diverse cognitive levels.

Knowledge Boundary. Inspired by knowledge boundary probing (Li et al., 2025a; Deng et al., 2025; Du et al., 2025; Yin et al., 2024), we study whether LLMs can generate solutions aligned with diverse human cognitive levels. Existing methods expand knowledge boundaries to improve reasoning via internal knowledge learning and probing (Liang et al., 2024; Azaria and Mitchell, 2023; Feng et al., 2024; Chen et al., 2025; Liu et al., 2025b; Ren et al., 2025; Tan et al., 2023; Shi et al., 2024), but tend to bias LLMs toward high-level knowledge, causing cognitive misalignment.

5 Conclusion

In this work, we introduce a benchmark for evaluating the adaptive alignment capabilities of LLMs in answering questions across different cognitive levels. Our results reveal substantial space for improvement, and future work will focus on enhancing LLMs’ cognitive alignment in terms of both

disciplinary breadth and cognitive depth.

Limitations

This work investigates the extent to which LLMs can adaptively generate responses aligned with students' cognitive levels in K–12 education. However, there are still several limitations. First, our evaluation is currently confined to the mathematics domain and a fixed set of cognitive levels; extending the benchmark to a broader range of subjects and more diverse cognitive progressions is necessary to assess the generality of our findings. Second, the cognitive knowledge graph used in this study, while curriculum-grounded, remains limited in coverage and granularity. Developing a more comprehensive and fine-grained cognitive knowledge base would further strengthen both the evaluation and the training of cognition-aware LLMs. Third, fine-tuning for cognitive alignment slightly degrades standard accuracy, suggesting a trade-off between pedagogical appropriateness and maximal problem-solving performance. Further investigation is required to balance these objectives in real-world educational systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276026) and the Fundamental Research Funds for the Central Universities (No. 2253500001).

References

- Nurul Aisyah, Muhammad Dehan Al Kautsar, Arif Hidayat, Raqib Chowdhury, and Fajri Koto. 2025. [Evaluating vision-language and large language models for automated student assessment in Indonesian classrooms](#). *Preprint*, arXiv:2506.04822.
- Anthropic. 2025. Claude opus 4 & sonnet 4 system card. <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf>.
- Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it's lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.
- Haw-Shiuan Chang, Ruei-Yao Sun, Kathryn Ricci, and Andrew McCallum. 2023. [Multi-CLS BERT: An efficient alternative to traditional ensembling](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–854, Toronto, Canada. Association for Computational Linguistics.
- Xiang Chen, Kun Yue, Wenjie Liu, Zhenyu Zhang, and Liang Duan. 2025. [Dual-kernel graph community contrastive learning](#). *Preprint*, arXiv:2511.08287.
- Yuyan Chen, Chenwei Wu, 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)*, pages 3138–3167, Bangkok, Thailand. Association for Computational Linguistics.
- Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. [EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Yang Deng, Moxin Li, Liang Pang, Wenxuan Zhang, and Wai Lam. 2025. [Unveiling knowledge boundary of large language models for trustworthy information access](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25*, page 4086–4089, New York, NY, USA. Association for Computing Machinery.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *Preprint*, arXiv:2305.14314.
- Lin Du, Lu Bai, Jincheng Li, Lixin Cui, Hangyuan Du, Lichi Zhang, Yuting Chen, and Zhao Li. 2025. [Lgan: An efficient high-order graph neural network via the line graph aggregation](#). *Preprint*, arXiv:2512.10735.
- Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. [Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.
- Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others.

2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. [Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning](#). *Preprint*, arXiv:2504.11456.
- Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. 2024. [E-EVAL: A comprehensive Chinese k-12 education evaluation benchmark for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7753–7774, Bangkok, Thailand. Association for Computational Linguistics.
- 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*, volume 36, pages 62991–63010. Curran Associates, Inc.
- Gaganpreet Jhajj, Xiaokun Zhang, Jerry Ryan Gustafson, Fuhua Lin, and Michael Pin-Chuan Lin. 2024. [Educational knowledge graph creation and augmentation via llms](#). In *Generative Intelligence and Intelligent Tutoring Systems: 20th International Conference, ITS 2024, Thessaloniki, Greece, June 10–13, 2024, Proceedings, Part II*, page 292–304, Berlin, Heidelberg. Springer-Verlag.
- Eldar Kurtic, Amir Moeini, and Dan Alistarh. 2024. [Mathador-LM: A dynamic benchmark for mathematical reasoning on large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17020–17027, Miami, Florida, USA. Association for Computational Linguistics.
- Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025a. [Knowledge boundary of large language models: A survey](#). *Preprint*, arXiv:2412.12472.
- Zhongzhi Li, Ming-Liang Zhang, Pei-Jie Wang, Jian Xu, Rui-Song Zhang, Yin Fei, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Jiaxin Zhang, and Cheng-Lin Liu. 2025b. [CMMaTH: A Chinese multi-modal math skill evaluation benchmark for foundation models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2690–2726, Abu Dhabi, UAE. Association for Computational Linguistics.
- Hao Liang, Linzhuang Sun, Minxuan Zhou, Zirong Chen, Meiyi Qiang, Mingan Lin, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. [Mathscape: Benchmarking multimodal large language models in real-world mathematical contexts](#). *Preprint*, arXiv:2408.07543.
- Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. [Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation](#). In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 44–58, Bangkok, Thailand. Association for Computational Linguistics.
- 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](#). *Preprint*, arXiv:2305.10263.
- Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. [MathBench: Evaluating the theory and application proficiency of LLMs with a hierarchical mathematics benchmark](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6884–6915, Bangkok, Thailand. Association for Computational Linguistics.
- Tianqiao Liu, Zui Chen, Zhensheng Fang, Weiqi Luo, Mi Tian, and Zitao Liu. 2025a. [Matheval: A comprehensive benchmark for evaluating large language models on mathematical reasoning capabilities](#). *Frontiers of Digital Education*, 2(2):16.
- Yuhan Liu, Cong Xu, Lu Liu, Yihua Wang, Feiyu Chen, Qi Jia, Yaqian Zhao, Zhichun Wang, and Xiang Li. 2025b. [DeMAC: Enhancing multi-agent coordination with dynamic DAG and manager-player feedback](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14072–14098, Suzhou, China. Association for Computational Linguistics.
- Tong Lu, Zhichun Wang, Yaoyu Zhou, Yiming Guan, Zhiyong Bai, and Junsheng Du. 2026. [Scimkg: A multimodal knowledge graph for science education with text, image, video and audio](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15466–15474.
- Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. [Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation](#). In *Proceedings of the*

- 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.
- Rudra Murthy, Praveen Venkateswaran, Prince Kumar, and Danish Contractor. 2025. **Kcif: Knowledge-conditioned instruction following**. *Preprint*, arXiv:2410.12972.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. **gpt-oss-120b and gpt-oss-20b model card**. *Preprint*, arXiv:2508.10925.
- OpenAI. 2025. Gpt-5 system card. <https://cdn.openai.com/gpt-5-system-card.pdf>.
- Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. **Investigating the factual knowledge boundary of large language models with retrieval augmentation**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3697–3715, Abu Dhabi, UAE. Association for Computational Linguistics.
- Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. **Training llm-based tutors to improve student learning outcomes in dialogues**. In *Artificial Intelligence in Education*, pages 251–266, Cham. Springer Nature Switzerland.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. **REPLUG: Retrieval-augmented black-box language models**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.
- Jianing Sun, Zhichao Zhang, and Xueli He. 2024. **Llm4edukg: Llm for automatic construction of educational knowledge graph**. In *2024 International Conference on Networking and Network Applications (NaNA)*, pages 269–275.
- Chenmien Tan, Ge Zhang, and Jie Fu. 2023. **Massive editing for large language models via meta learning**. *ArXiv*, abs/2311.04661.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. **Llama 2: Open foundation and fine-tuned chat models**. *Preprint*, arXiv:2307.09288.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. **Self-consistency improves chain of thought reasoning in language models**. In *The Eleventh International Conference on Learning Representations*.
- Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. **Cmath: Can your language model pass chinese elementary school math test?** *Preprint*, arXiv:2306.16636.
- Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao, and Heyan Huang. 2025. **Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios**. *Preprint*, arXiv:2505.16160.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. **Qwen3 technical report**. *Preprint*, arXiv:2505.09388.
- Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. **Benchmarking knowledge boundary for large language models: A different perspective on model evaluation**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, Bangkok, Thailand. Association for Computational Linguistics.
- Weihao You, Pengcheng Wang, Changlong Li, Zhilong Ji, and Jinfeng Bai. 2024. **Ck12: A rounded k12 knowledge graph based benchmark for chinese holistic cognition evaluation**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:19431–19439.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. **Qwen3 embedding: Advancing text embedding and reranking through foundation models**. *Preprint*, arXiv:2506.05176.
- Pengfei Zhou, Fanrui Zhang, Xiaopeng Peng, Zhaopan Xu, Jiaxin Ai, Yansheng Qiu, Chuanhao Li, Zhen Li, Ming Li, Yukang Feng, Jianwen Sun, Haoquan Zhang, Zizhen Li, Xiaofeng Mao, Wangbo Zhao, Kai Wang, Xiaojun Chang, Wenqi Shao, Yang You, and Kaipeng Zhang. 2025. **Mdk12-bench: A multi-discipline benchmark for evaluating reasoning in multimodal large language models**. *Preprint*, arXiv:2504.05782.

APPENDIX

A	Details of LLM Fine-tuning	12
B	Details of Experiment Settings	12
C	The effectiveness of probability attenuation	12
D	Prompts	13
E	Case Study	13
F	Ethics Statement	14

A Details of LLM Fine-tuning

To explore whether the cognitive alignment ability of LLMs can be enhanced, we use a fine-tuning strategy to the existing Qwen3-30B-A3B model. Following the evaluation, we adopt a two-stage training framework based on QLoRA (Detmers et al., 2023) to investigate whether language models could improve adaptive cognitive alignment capability through fine-tuning, enabling the training of large models under limited GPU memory constraints. We leverage the proposed CAKG together with the curated QA pairs to construct the instruction–response dataset $\mathcal{D}_{\text{SFT}} = \{(z_i, s_i)_{i=1}^N\}$, where z_i denotes either a knowledge point or an instruction derived from CAKG or from a given question, and s_i provides the corresponding explanation or solution aligned with a specific cognitive level. Each question q_j has multiple solutions $\{s_i^1, s_i^2, s_i^3, \dots, s_i^L\}_{l=1}^L$ aligned with the varying cognitive levels of the students. Our objective in the supervised fine-tuning stage is to maximize the likelihood of the appropriate response conditioned on its instruction:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T_i} \log p_{\theta}(s_{i,t} | z_i, s_{i,<t}) \quad (6)$$

where T_i is the length of the target sequence y_i and θ are the model parameters.

To further align the model with diverse cognitive levels, we utilize direct preference optimization with a curated preference dataset $\mathcal{D}_{\text{DPO}} = \{(q_j, s_j^+, s_j^-)\}_{j=1}^M$, where s_j^+ is the preferred answer corresponded to a specific cognitive level, s_j^-

is the less preferred alternative for question q_j . The loss is defined as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{M} \sum_{j=1}^M \log \sigma \left(\beta \left[\log \frac{p_{\theta}(s_j^+ | q_j)}{p_{\text{ref}}(s_j^+ | q_j)} - \log \frac{p_{\theta}(s_j^- | q_j)}{p_{\text{ref}}(s_j^- | q_j)} \right] \right) \quad (7)$$

where $\sigma(\cdot)$ denotes the sigmoid function, p_{ref} is the reference model distribution, and β is a temperature scaling parameter.

Finally, we minimize the combined objective:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{SFT}}(\theta) + \lambda \mathcal{L}_{\text{DPO}}(\theta) \quad (8)$$

where λ controls the trade-off between supervised fidelity and preference alignment.

B Details of Experiment Settings

All experiments were conducted in a Python 3 environment with PyTorch 2. We evaluated LLMs through API endpoints in September and December 2025. Model fine-tuning was carried out on a single NVIDIA A100 GPU. All hyperparameters of the LLMs are listed in Table 4. β is a temperature scaling parameter and λ controls the trade-off between supervised fidelity and preference alignment, both used in model training.

Parameter	Value
temperature	1.0
max_length	1000
top_p	0.9
β	0.23
λ	0.27

Table 4: Hyperparameters.

C The effectiveness of probability attenuation

To explore the effectiveness of probability attenuation, we investigate whether LLMs possess an inherent cognitive-aware capability to generate more cognitively aligned answers under attenuated probability distributions. As shown in Table 5, under the grade-constrained QA setting, most models demonstrate consistent improvements in cognitive accuracy and knowledge consistency after using

Model	Pa	ACC	CA	KC	KD
LLaMa2-7B (Touvron et al., 2023)	✗	0.710	0.403	0.215	0.581
	✓	0.692	0.427	0.290	0.563
LLaMa3-8B (Grattafiori et al., 2024)	✗	0.706	0.462	0.303	0.479
	✓	0.659	0.513	0.384	0.424
Qwen3-8B (Yang et al., 2025)	✗	0.794	0.480	0.254	0.636
	✓	0.712	0.556	0.398	0.672
LLaMa2-13B (Touvron et al., 2023)	✗	0.767	0.432	0.265	0.612
	✓	0.734	0.483	0.347	0.571
Qwen3-30B-A3B (Yang et al., 2025)	✗	0.860	0.521	0.610	0.724
	✓	0.814	0.558	0.634	0.665

Table 5: The impact of probability attenuation on cognitive alignment in LLMs. Pa = Probability attenuation.

probability attenuation, *indicating that LLMs intrinsically encode knowledge across multiple cognitive levels and that attenuation can bias decoding toward grade-aligned knowledge.* However, by suppressing knowledge representations that exceed the specific grade, probability attenuation also disrupts the models’ original optimal reasoning paths, leading to a general decline in overall answer accuracy. This constrained decoding narrows the effective candidate knowledge components and reduces knowledge divergence for most models. Although Qwen3-8B exhibits an opposite trend, suggesting that the effect of attenuation depends on the model’s internal knowledge organization. *These results reveal a trade-off between maximizing answer accuracy through broader knowledge access and constraining cognitive alignment for education.*

D Prompts

Here are the key prompts used in this work:

Prompts of generating answer based question titles
<p>[System Input]: You are an expert in K–12 education. Please respond in a rigorous academic tone.</p> <p>[User Input]: Please generate an answer for the following math question, including a brief mathematical analysis: Question: {title} Please provide: 1. A brief problem-solving approach 2. Detailed calculation steps 3. The final answer Please reply in brief academic Chinese.</p> <p>[Answer]:</p>

Prompts of generating answer based question titles and grades
<p>[System Input]: You are an expert in K–12 education. Please respond in a rigorous academic tone.</p> <p>[User Input]: Please use the following math problem and grade information to generate an answer with a brief math analysis: Question: {title} Grade: {grade} You must based on the knowledge level of students in this grade to answer question, please provide: 1. A suitable problem-solving approach for this grade 2. Detailed calculation steps 3. The final answer Please reply in brief academic Chinese.</p> <p>[Answer]:</p>

Prompts of generating answer based question titles and their aligned knowledge
<p>[System Input]: You are an expert in K–12 education. Please respond in a rigorous academic tone.</p> <p>[User Input]: please generate an answer with a brief mathematical analysis based on the following math problem and related knowledge points: Question: {title} Related Knowledge Points: {knowledge} You must based on the provided Related Knowledge Points to answer question: 1. A problem-solving approach based on the relevant knowledge points 2. Brief calculation steps 3. Final answer Please reply in brief academic Chinese.</p> <p>[Answer]:</p>

E Case Study

As illustrated in Figure. 8, standard answers were generated using the proposed evaluation framework. The results indicate that the fine-tuned Qwen3-30B-A3B effectively generates aligned answers for learners across diverse cognitive levels. This case study shows that *the fine-tuned model adapts its reasoning strategies to different cognitive levels, rather than only generating the correct results.* Answers at lower cognitive levels rely on intuitive and step-by-step reasoning, while higher-grade responses employ more abstract mathematical concepts and formal expressions. This demonstrates that the fine-tuned model can capture not only answer correctness, but also alignment across different cognitive levels.

F Ethics Statement

The dataset used in this study was sourced from an open-access dataset released with an academic paper as well as Olympiad mathematics problems collected from a public website. In the data processing, we strictly avoided the use of any personally identifiable information, and the dataset includes no inappropriate content. In addition, the data annotation was conducted through a collaborative workflow involving five expert annotators in education and an automated annotation system. All annotators were fully informed in advance about the intended use and presentation of the data, ensuring that the entire annotation process complied with relevant ethical standards. Specifically, we first generated initial solutions using a pre-written Python program, after which annotators performed verification and quality control. To assess annotation reliability, we computed Krippendorff's alpha, which exceeded 0.87. To reasonably manage workload, each annotator worked no more than four hours per day and was compensated in accordance with the local statutory hourly minimum wage.

Question: 计算 $1+3+5+7+9=?$
Question: Calculate $1+3+5+7+9=?$

Standard Answers

grade1: 我们可以在可能的情况下把数字进行分组, 使它们凑成 10, 这样可以简化加法运算。下面是一步一步的方法: **###** 第一步: 寻找能凑成 10 的数对 $-1+9=10$ $-3+7=10$ -5 单独留下 **###** 第二步: 把数对和剩余的数相加 -10 (来自 $1+9$) -10 (来自 $3+7$) -5 (剩余的) **###** 第三步: 把结果相加 $-10+10+5=25$ **###** 最终答案: $1+3+5+7+9=25$ (使用“凑十法”)
Grade 1: We can group numbers to make tens where possible, which helps simplify addition. Here's how you can do it step by step: **###** Step 1: Look for pairs that make 10 $-1+9=10$ $-3+7=10$ -5 is left alone. **###** Step 2: Add the pairs and the remaining number -10 (from $1+9$) -10 (from $3+7$) -5 (leftover) **###** Step 3: Add the results $-10+10+5=25$ **###** Final Answer: $1+3+5+7+9=25$ (using the Make-ten strategy)

grade6: 我们可以利用前 n 个奇数之和的性质: $> 前 n 个奇数之和等于 n^2$ 。 $>$ 也就是说: $> 1+3+5+\dots+(2n-1)=n^2$ **###** 第一步: 数一数有多少项数列是: $1, 3, 5, 7, 9$ 这是前 5 个奇数 **###** 第二步: 应用公式因为这是前 5 个奇数, 所以可以使用公式: $1+3+5+7+9=5^2=25$ **###** 最终答案: 25
Grade 6: We can use the property of the sum of the first n odd numbers: $> The sum of the first n odd numbers is equal to n^2$. $>$ That is, $> 1+3+5+\dots+(2n-1)=n^2$ **###** Step 1: Count the number of terms The sequence is: $1, 3, 5, 7, 9$ This is the first 5 odd numbers **###** Step 2: Apply the formula Since it's the first 5 odd numbers, we can use the formula: $1+3+5+7+9=5^2=25$ **###** Final Answer: 25

grade10: 要求和为: $1+3+5+7+9$ 我们可以使用等差数列前 n 项和的公式: $S_n = \frac{n}{2}(a_1+a_n)$ 其中: $-S_n$ 表示前 n 项的和, $-n$ 是项数, $-a_1$ 是首项, $-a_n$ 是末项 **###** 第一步: 确定各项 $-a_1=1$ $-a_n=9$ $-n=5$ 数列: $1, 3, 5, 7, 9$ $-n=5$ **###** 第二步: 代入公式 $S_5 = \frac{5}{2}(1+9) = \frac{5}{2} \times 10 = 5 \times 5 = 25$ **###** 最终答案: 25 这是等差数列 $1+3+5+7+9$ 的和
Grade 10: To calculate the sum of the arithmetic sequence: $1+3+5+7+9$ We can use the formula for the sum of the first n terms of an arithmetic sequence: $S_n = \frac{n}{2}(a_1+a_n)$ Where: $-S_n$ is the sum of the first n terms, $-n$ is the number of terms, $-a_1$ is the first term, $-a_n$ is the last term. **###** Step 1: Identify the terms - First term $a_1=1$ - Last term $a_n=9$ - The sequence: $1, 3, 5, 7, 9$ - Number of terms $n=5$ **###** Step 2: Plug into the formula $S_5 = \frac{5}{2}(1+9) = \frac{5}{2} \times 10 = 5 \times 5 = 25$ **###** Final Answer: 25 This is the sum of the arithmetic sequence $1+3+5+7+9$.

Qwen3-30B-A3B

grade1: 做得很好, 你已经尝试解决这个问题了! 让我们一步一步把这些数加起来: $1+3=4+5=9+7=16+9=25$ 所以, $1+3+5+7+9=25$ 干得漂亮!
Grade 1: Great job on trying to solve this! Let's add the numbers step by step: $1+3=4$ $4+5=9$ $9+7=16$ $16+9=25$ So, $1+3+5+7+9=25$ Nice work!

grade6: 当然可以! 让我们一步一步把这些数加起来: $1+3=4+5=9+7=16+9=25$ 所以, $1+3+5+7+9=25$ 做得很好! 这些都是奇数, 把它们加起来得到 25.
grade6: Sure! Let's add the numbers step by step: $1+3=4$ $4+5=9$ $9+7=16$ $16+9=25$. So, $1+3+5+7+9=25$ Nice job! These are all odd numbers, and adding them gives you 25.

grade10: 当然可以! 让我们一步一步把这些数加起来: $1+3=4+5=9+7=16+9=25$ 所以, $1+3+5+7+9=25$ 。 **答案:** 25 。 **小提示:** 这些是连续的奇数。前 n 个奇数的和总是等于 n^2 。 这里我们有 5 个奇数 $(1, 3, 5, 7, 9)$, 所以 $5^2=25$ 。
grade10: Sure! Let's add the numbers step by step: $1+3=4$ $4+5=9$ $9+7=16$ $16+9=25$. So, $1+3+5+7+9=25$. **Answer:** 25 . **Bonus tip:** These are consecutive odd numbers. The sum of the first n odd numbers is always n^2 . Here, we have 5 odd numbers $(1, 3, 5, 7, 9)$, so $5^2=25$. That's a neat shortcut!

Fine-tuned Qwen3-30B-A3B

grade1: 当然可以! 让我们一步一步把这些数加起来。我们要加的是: $1+3+5+7+9$ 让我们来一找能凑成 10 的数对: $-1+9=10$ $-3+7=10$ 现在我们有 两个 10: $10+10=20$ 还剩下 5。
现在再相加: $20+5=25$ 最终答案: 25
grade1: Of course! Let's add these numbers step by step. We are adding: $1+3+5+7+9$ Let's look for pairs that add up to 10: $1+9=10$ $3+7=10$ Now we have two tens: $10+10=20$ That leaves 5. Now add them together: $20+5=25$ **Final answer:** 25

grade6: 当然可以! 让我们利用奇数的性质来计算这个和: $1+3+5+7+9$ 我们知道奇数有一个特别的规律。奇数比偶数多 1 的数, 它们依次是: $1, 3, 5, 7, 9, \dots$ 当我们把一串奇数相加时: 前 n 个奇数的和等于 n^2 (n 的平方) ——这是一个非常有趣的规律! **###** 让我们把它用在这里: 我们要加的是: $1+3+5+7+9$ 这些数有多少个呢? 让我们来数一数: -1 第一个奇数 -3 第二个 -5 第三个 -7 第四个 -9 第五个 所以, 一共有 5 个奇数。现在使用这个规律: 前 5 个奇数的和 $=5^2=25$ **###** 所以答案是: 25
grade6: Of course! Let's use the properties of odd numbers to calculate this sum: $1+3+5+7+9$ We know that odd numbers follow a special pattern. An odd number is one more than an even number, and they go like this: $1, 3, 5, 7, 9, \dots$ When we add a sequence of odd numbers: The sum of the first n odd numbers is equal to n^2 (n squared) — a very interesting rule! Let's apply it here: We are adding: $1+3+5+7+9$ How many numbers are there? Let's count them: 1 —first odd number 3 —second 5 —third 7 —fourth 9 —fifth So, there are 5 odd numbers in total. Now apply the rule: The sum of the first 5 odd numbers $=5^2=25$ Therefore, the answer is: 25

grade10: 当然可以! 让我们使用等差数列的方法来计算 $1+3+5+7+9$ 。我们可以观察到, 这些数是: $1, 3, 5, 7, 9$ 它们构成了一个等差数列, 其中: $-$ 首项 $a=1$ $-$ 公差 $d=2$ $-$ 项数 $n=5$ 现在使用等差数列求和公式: $S_n = \frac{n}{2}(2a+(n-1)d)$
把数值代入公式: $S_5 = \frac{5}{2}(2 \times 1 + (5-1) \times 2) = \frac{5}{2}(2+8) = \frac{5}{2} \times 10 = 5 \times 5 = 25$
最终答案: 25
grade10: Of course! Let's use the arithmetic sequence method to calculate $1+3+5+7+9$. We can observe that these numbers are: $1, 3, 5, 7, 9$ They form an arithmetic sequence, where: First term $a=1$ Common difference $d=2$ Number of terms $n=5$ Now we use the sum formula for an arithmetic sequence: $S_n = \frac{n}{2}(2a+(n-1)d)$ Substitute the values into the formula: $S_5 = \frac{5}{2}(2 \times 1 + (5-1) \times 2) = \frac{5}{2}(2+8) = \frac{5}{2} \times 10 = 5 \times 5 = 25$ Final answer: 25

Figure 8: The fine-tuned Qwen3-30B-A3B could generate aligned answers across diverse cognitive levels. English translations are provided beneath the relevant Chinese text in blue.