K-12EduBench: A Benchmark for Evaluating Large Language Models' Knowledge, Problem-Solving, and Educational Goal Cognition in K-12 Education

Yuqing Ye[1*], Xuan Zhou[1*], Zhifu Chen[1*], Dandan Li[1], Hengnian Gu[1], Jin Peng Zhou[2], Dongdai Zhou[†]

[1]Northeast Normal University, Changchun, Jilin, China
[2]Cornell University, Ithaca, New York, United States
{yeyuqing,zhouxuan711,zhifuchen,lidandan,guhn546,ddzhou}@nenu.edu.cn, jpzhou01@gmail.com

Abstract

Large language models hold great promise for transforming K-12 education, but there is an urgent need for systematic evaluation of their core educational capabilities. Existing benchmarks often overlook educational goal cognition and overemphasize answer accuracy, thereby failing to capture deeper subject-level knowledge ability and problem-solving ability. To address this gap, we introduce K-12EduBench: a benchmark for evaluating LLMs' subject-level knowledge ability, subject-specific problem-solving ability, and educational goal cognition ability in K-12 education. K-12EduBench comprises four components: (1) a dataset of 2,640 objective and 619 subjective questions across nine subjects, annotated with answers, problem-solving processes, and cognitive-level labels; (2) nine Item Response Theory (IRT) models for estimating subject-level knowledge ability; (3) evaluation methods and metrics for assessing multi-step problem-solving ability; and (4) prompts and scoring rubrics for measuring alignment with target cognitive levels. Experiments on advanced LLMs show that education-optimized models consistently outperform general-purpose ones across all three abilities, while under-scaled models lag substantially. We observe a strong positive correlation between subject-level knowledge ability and subject-specific problem-solving ability. Despite gains in educational goal cognition ability, current models—even those tailored for education—still fall short of real-world instructional needs.

Code — https://github.com/shida-edu4ai/K-12EduBench

Introduction

Large language models (LLMs) have advanced rapidly and are being applied in K-12 education (Hu et al. 2024; Tai and Chen 2024). Both general-purpose models (e.g., ChatGPT and Claude) and education-optimized ones (e.g., EduChat and iFlytek Spark) are deployed, but their true subject-level abilities remain opaque, motivating the need for a comprehensive K-12 benchmark.

Existing benchmarks for educational LLMs fall into two rough categories. The first emphasizes foundational performance via exam-style accuracy across K-12 subjects, measuring subject-level knowledge ability and general reasoning (e.g., C-Eval and E-Eval (Huang et al. 2023a; Hou et al. 2024)). The second targets task-specific effectiveness, such as the Pedagogy Benchmark (interdisciplinary and special education knowledge (Lelièvre et al. 2025)), Dr. Academy (question-asking framed by Bloom's taxonomy (Chen et al. 2024)), and SciBench (science problem solving via accuracy (Wang et al. 2023)).

Teaching and learning are structured problem-solving activities: they begin with domain challenges that trigger cognition and require the integration of knowledge, reasoning, and intentional alignment to instructional objectives (Popper 1999; Dewey 1933). To evaluate LLMs in K-12 education—whether acting as tutors, content generators, or learners—we advocate for evaluating three subject-level abilities: subject-level knowledge ability, subject-specific problem-solving ability, and educational goal cognition ability. These correspond respectively to mastery of curricular content in a given subject, the capacity to carry out multi-step reasoning and strategy selection to solve domain problems, and the understanding of the intended educational goal behind a task so that model behavior aligns with pedagogical intent.

Each of these abilities exposes a distinct gap in common evaluation practices. Raw answer accuracy conflates question difficulty and examinee ability, obscuring subject-level knowledge ability (Embretson and Reise 2013). Problem-solving ability cannot be captured by final correctness alone; it demands process-aware assessment of trajectories, intermediate reasoning, and error recovery (Wiggins 1998). Educational goal cognition such as the Bloom's taxonomy (Anderson and Krathwohl 2001) is required because effective feedback and instruction depend on whether the model comprehends and aligns with the target cognitive objective. Existing benchmarks largely collapse these dimensions—overrelying on accuracy, merging knowledge and reasoning, and typically omitting educational goal cognition—thus failing to provide a holistic characterization of LLMs' core subject-level educational abilities. This motivates a unified benchmark that jointly measures all three.

To address these gaps, we introduce **K-12EduBench**: the first comprehensive K-12 benchmark that jointly evaluates LLMs' **subject-level knowledge ability**, **subject-specific**

[*]These authors contributed equally.
[†]Corresponding author.

problem-solving ability, and **educational goal cognition ability**. K-12EduBench comprises four components:

- High-quality multidisciplinary dataset: 3,259 exam questions (2,640 objective, 619 subjective) curated from over 20 years of teaching resources across nine subjects—Chinese, Mathematics, English, Physics, Chemistry, Biology, Politics, History, and Geography. Each question is annotated with answers, problem-solving processes, and cognitive-level labels.
- IRT-based subject-level knowledge ability evaluation: we estimate latent subject knowledge ability via Item Response Theory, decoupling item properties (e.g., difficulty and discrimination) from model proficiency to produce more reliable mastery estimates.
- Process-based evaluation of subject-specific problem-solving ability: drawing on cognitivist problem-solving principles, we analyze entire solution trajectories—assessing effectiveness, factual consistency, pragmatic relevance, and coherence—to capture reasoning, planning, strategy selection, and organization.
- Educational goal cognition evaluation: using the revised Bloom's taxonomy as the reference, we assess models' ability to recognize and differentiate the intended cognitive level of instructional content.

We conducted large-scale experiments and in-depth analyses to measure LLMs across these three subject-level abilities and their interactions. The main findings are:

- LLMs exhibit substantial heterogeneity in educational performance: models such as Doubao-Pro-32K, Doubao-Lite-32K, and GeneralV3.5 achieve strong results across all three abilities, whereas EduChat-SFT-13B and LLaMA-3.1-8B lag behind.
- Education-optimized models—particularly those with sufficient scale—tend to outperform general-purpose counterparts; smaller-scale variants (even if education-optimized) show degraded performance.
- The results (see Table 2) show a significant positive correlation between subject-level knowledge ability (via IRT) and subject-specific problem-solving ability. Subject-level knowledge ability correlates more strongly with accuracy (ACC) than with problem-solving ability, suggesting that problem-solving captures reasoning components not fully explained by knowledge mastery alone.
- Educational goal cognition ability remains limited: both education-optimized and general-purpose models struggle to accurately classify the intended instructional cognitive level. They often overestimate cognitive level when surface textual cues imply higher-order thinking, yet under-recognize deeper pedagogical intent, leading to systematic misalignment.

Related Work

Foundational Educational Benchmarks

Foundational benchmarks assess subject knowledge and general reasoning, typically via answer accuracy across disciplines. Representative examples include MMLU (Hendrycks et al. 2021a) and its Chinese variant CMMLU (Li et al. 2024), MMCU (Zeng 2023), C-Eval (Huang et al. 2023b), and Xiezhi (Gu et al. 2024) for breadth. Some span educational stages: M3KE (Liu et al. 2023) covers elementary to university levels, and E-Eval (Hou et al. 2024) focuses on K-12. CK12 (You et al. 2024) decomposes fine-grained knowledge points and incorporates mathematical reasoning. Real-world test suites such as AGIEval (Zhong et al. 2024) (standardized exams with error analysis) and Gaokao (Zhang et al. 2024) (China's national college entrance exam) further probe knowledge and reasoning in deployed formats.

Specialized and Pedagogical Benchmarks

Other benchmarks target domain-specific problem solving or pedagogical competencies. SciBench (Wang et al. 2023) evaluates scientific reasoning in chemistry, physics, and math, analyzing both accuracy and reasoning errors. MATH (Hendrycks et al. 2021b) uses competition-level math problems to stress advanced reasoning, and ChemBench (Mirza et al. 2024) focuses on chemistry problem solving. Pedagogical-oriented evaluations include MathTutorBench (Macina et al. 2025) (tutoring dialogue skills such as student modeling, error detection, and scaffolding), Dr. Academy (Chen et al. 2024) (question generation informed by Bloom's taxonomy), the Pedagogy Benchmark (Lelièvre et al. 2025) (cross-disciplinary and special education knowledge), and CoMTA (Miller and DiCerbo 2024) (judging student responses in complex math tutoring).

Despite their contributions, existing benchmarks share key limitations: they largely depend on final answer accuracy, blur the distinction between knowledge and reasoning, and rarely evaluate subject-specific problem-solving trajectories or educational goal cognition as defined by Bloom's taxonomy. They also lack a unified educational science grounding for jointly characterizing these core abilities.

K-12EduBench

Overview

To overcome existing benchmarks' limitations, we introduce *K-12EduBench*, a comprehensive benchmark for evaluating LLMs' subject-level knowledge, subject-specific problem-solving, and educational goal cognition abilities in K-12 education.The benchmark contains 3,259 high-quality questions (2,640 objective and 619 subjective) across nine subjects and 701 knowledge points. Each question is accompanied by verified answers, reasoning processes, and cognitive-level annotations, reviewed by subject experts for consistency and validity (see Appendix A), and sourced from real educational contexts. Grounded in educational measurement theory and instructional design principles, the benchmark comprises four core components: (1) the multidisciplinary question set, annotated with correct answers, detailed reasoning processes, and cognitive-level objectives; (2) subject-specific Item Response Theory (IRT) models to estimate latent knowledge ability, allowing robust proficiency estimation independent of question difficulty;

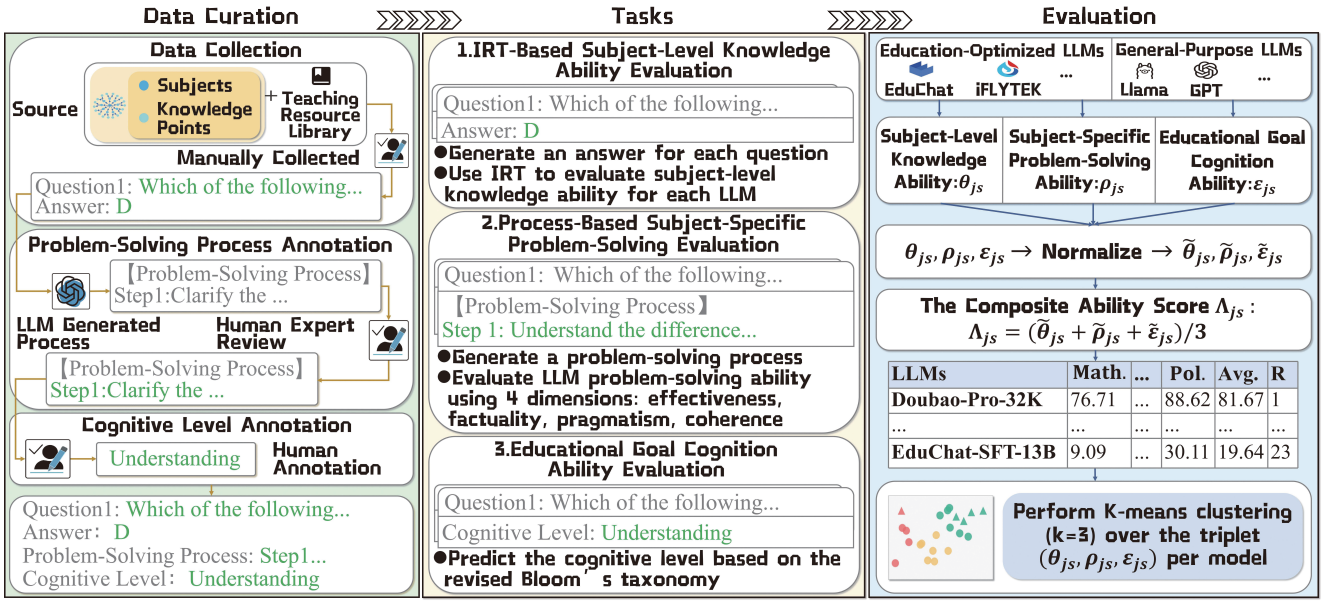


Figure 1: Overview of the K-12EduBench pipeline. (Left) Data curation: a high-quality dataset is assembled across nine K-12 subjects, with each question annotated with reference answers, standardized problem-solving processes, and cognitive-level labels. (Middle) Tasks: three subject-level evaluations—knowledge ability, problem-solving ability, and educational goal cognition. (Right) Evaluation: tailored metrics are applied to comprehensively assess LLMs' performance across these abilities.

(3) evaluation methods and metrics for assessing problem-solving ability via process-based analysis of reasoning and solution paths; and (4) evaluation methods and metrics to assess educational goal cognition, ensuring alignment with intended instructional objectives based on Bloom's taxonomy. The overall evaluation framework is illustrated in Figure 1.

Data Curation

Data Annotation Our source data is drawn from a proprietary database containing over 20 years of continuously collected teaching resources. To ensure high-quality annotations, subject-matter experts first manually reviewed all source data, verifying the reference answers and knowledge labels to establish a reliable annotation foundation. Annotations then primarily targeted two areas: annotating standardized problem-solving processes, and labeling each question with its cognitive level.

Problem-Solving Process Annotation: We developed a hybrid annotation pipeline combining automatic generation (by a well-performing general-purpose LLM) with subsequent expert revision. Specifically, an LLM first generated initial step-by-step problem-solving processes given the question, answer, and provided explanations (see Appendix A1, Figure 3). Educational experts then conducted rigorous evaluations and standardization of these generated solutions according to four clearly-defined evaluation dimensions: *effectiveness*, *factuality*, *pragmatism*, and *coherence*. During revision, experts ensured completeness and correctness of each step (effectiveness and factuality), logical organization and clarity of expression (coherence), and alignment of step-level scoring with importance in solving the task (pragmatism). To account for differences between humanities and science subjects, separate revision guidelines were designed (see Appendix A1, Tables 5 and 6).

Cognitive Level Annotation: To ground the benchmark firmly in educational science, we designed a dedicated annotation protocol based on the revised Bloom's taxonomy (*Annotation Protocol for Educational Goal Cognition*, Appendix A2). The annotation approach operationalizes the "teaching-assessment integration" philosophy from Bloom's taxonomy into three key principles: *Cognitive Path Integrity*, *Essential Task Judgment*, and *Core Level Identification*. These principles emphasize capturing the underlying instructional goals and learning objectives embedded within each question.

The annotation workflow comprised three sequential steps: (1) identifying the main cognitive task requirements embedded in the question, (2) analyzing solution steps to infer corresponding cognitive levels, and (3) synthesizing a single cognitive-level label for the entire question. Clear and detailed annotation criteria guided each cognitive level judgment (provided in Appendix A2). Each question received independent annotations from two experts, with disagreements resolved by a third expert's review to ensure reliable and consistent labeling.

Tasks

Using the curated multidisciplinary dataset, we design three evaluation tasks to comprehensively measure LLMs' core subject-level educational abilities: (1) *subject-level knowledge ability*, evaluated using IRT; (2) *subject-specific problem-solving ability*, assessed via process-based anal-

ysis; and (3) *educational goal cognition ability*, determined through alignment with Bloom's taxonomy. Formally, we denote each evaluated LLM as LLM_j, and represent its subject-level knowledge, problem-solving, and educational goal cognition abilities on subject s as θ_{js}, ρ_{js}, and ε_{js}, respectively. Specifically, we define the subject set $S = \{\text{Math., Chin.,Engl., Phys.,Chem., Biol.,Hist., Geogr., Pol.}\}$, representing Mathematics, Chinese, English, Physics, Chemistry, Biology, History, Geography, and Politics. For each subject $s \in S$, we have a corresponding question set $Q_s = \{q_{s1}, q_{s2}, \ldots, q_{sN}\}$, where each question q_{si} is accompanied by a reference answer a_{si}, a standardized problem-solving process r_{si}, and an expert-annotated educational goal cognition label t_{si}. Each evaluation task utilizes dedicated prompts and scoring criteria to ensure methodological consistency across subjects.

Subject-Level Knowledge Ability via IRT We model each evaluated LLM, LLM_j as an examinee and estimate its subject-level knowledge ability θ_{js} using Item Response Theory (Embretson and Reise 2013) (IRT), which disentangles a model's latent proficiency from item characteristics. The intuition is that raw accuracy conflates question difficulty, discrimination, and chance success, whereas IRT provides a principled probabilistic model of how likely a model with ability θ_{js} answers each question correctly. For each subject s, we fit a three-parameter logistic (3PL) model over its item set. Let q_{si} be the i-th question in subject s with item parameters α_{si} (discrimination), β_{si} (difficulty), and γ_{si} (guessing). The probability that LLM_j answers q_{si} correctly is

$$P_{jsi} = \gamma_{si} + (1 - \gamma_{si}) \cdot \frac{1}{1 + \exp\left(-\alpha_{si}(\theta_{js} - \beta_{si})\right)}. \quad (1)$$

Each response is binarized (correct vs. incorrect) according to the reference answer a_{si} using the scoring rubric (see Appendix B1), and the subject-specific IRT model jointly estimates the item parameters and abilities θ_{js} via the EM algorithm. The resulting θ_{js} serves as a calibrated measure of LLM_j's knowledge mastery in subject s, robust to differences in item design. The prompt templates and training process of the IRT model are provided in Appendix B1.

Process-Based Subject-Specific Problem-Solving Evaluation Grounded in cognitive models of problem solving (Newell, Shaw, and Simon 1958), we assess each model's subject-specific problem-solving ability ρ_{js} by evaluating the generated solution trajectories. For each question q_{si}, LLM_j produces a step-by-step solution process $\hat{r}_{si}^{(j)}$ in a standardized format(see Appendix B2, Figure 5). We adopt an *LLM-as-judge* (Zheng et al. 2023) framework: a dedicated (and calibrated) evaluator—implemented via structured prompting of a high-quality LLM and aligned with expert judgments—scores each generated process along four complementary dimensions that map to the cognitive stages of problem solving:

- **Effectiveness:** captures problem representation and operator application, i.e., whether the process addresses the core subproblems and applies correct solution steps toward a_{si}, as compared to the reference process r_{si}.
- **Factuality:** evaluates operator selection and current-state evaluation by checking that each step is factually accurate, grounded in the given question q_{si}, and free of contradictions or misuse of domain knowledge.
- **Pragmatism:** measures efficiency and relevance of the chosen steps—whether the model avoids unnecessary detours and uses appropriate strategies to reach the correct answer a_{si}.
- **Coherence:** assesses the logical flow across steps, ensuring smooth transitions, no unjustified leaps, and an internally consistent reasoning chain.

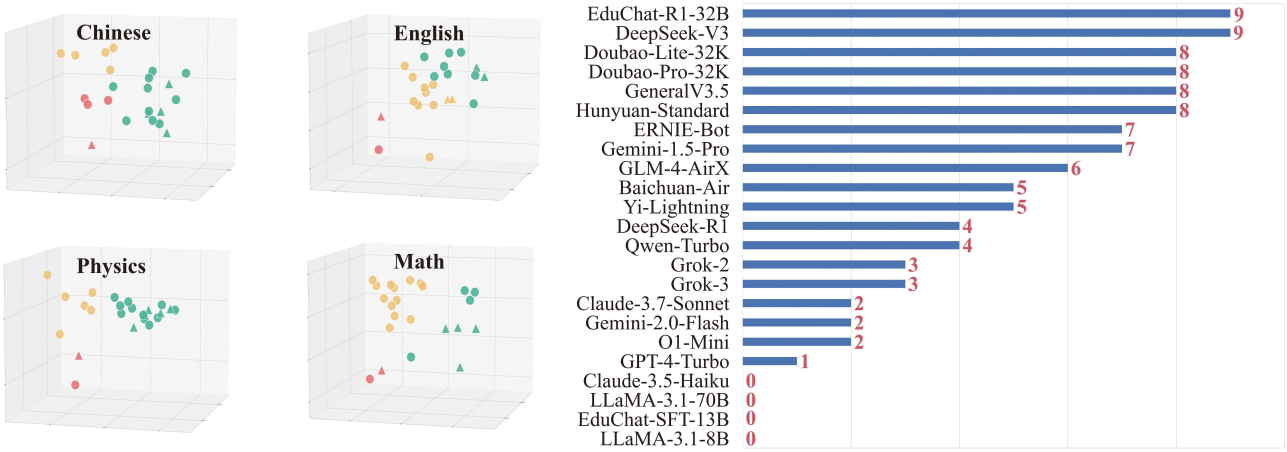
Each dimension yields an question-level score $s_{jsi}^{(d)}$ (for $d \in \{\text{eff}, \text{fact}, \text{prag}, \text{coh}\}$), which are aggregated into a composite problem-solving score for question q_{si}; these question-level scores are then pooled over all questions in subject s to produce ρ_{js} (see Appendix B2). This process-based scoring framework captures nuanced reasoning quality beyond final correctness, enabling a fine-grained estimate of LLM_j's problem-solving ability in each subject. To assess how well this *LLM-as-judge* framework aligns with expert judgments, we randomly sampled 90 questions and asked graduate students in educational technology to annotate process-based problem-solving using the same rubric as the LLM judge. The LLM evaluator achieved a Pearson correlation of r=0.84 ($p < 0.01$) with expert scores, indicating strong consistency and reliability for large-scale evaluation(see Appendix B2).

Educational Goal Cognition Ability Evaluation Educational goal cognition ability ε_{js} captures whether LLM_j recognizes and aligns with the intended instructional objective of a question in subject s. Operationally, this is framed as a multi-class classification task: for each question q_{si}, the model predicts a cognitive level $\hat{t}_{si}^{(j)}$ from the six mutually exclusive categories of the revised Bloom's taxonomy (remembering, understanding, applying, analyzing, evaluating, creating), where the ground-truth label t_{si} is provided by expert annotation. This evaluation is motivated by the fact that effective instruction and feedback depend not only on correctness, but on understanding *what kind* of thinking the question is targeting—e.g., whether it is testing recall versus higher-order analysis—so that model behavior can be pedagogically appropriate.

In practice, each model is prompted with a concise description of the taxonomy and asked to assign one of the six levels to each question (prompt templates in Appendix B3, Figure 7). Because the cognitive levels are ordinal (e.g., remembering is closer to understanding than to creating), we evaluate performance using both classification accuracy and mean absolute error (MAE) over the integer encoding of levels which penalizes larger deviations more heavily and respects the graded similarity between adjacent categories. We also analyze the predicted-level distribution and confusion patterns to uncover systematic biases such as overestimation toward higher-order levels or under-recognition of deeper pedagogical intent. These signals are combined to produce ε_{js}, a calibrated measure of educational goal cognition ability for subject s.

LLMs	Math.	Chin.	Engl.	Phys.	Chem.	Biol.	Hist.	Geogr.	Pol.	Avg.	R
Doubao-Pro-32K	76.71	71.44	85.57	84.74	90.91	88.86	63.62	84.55	88.62	81.67	1
DeepSeek-V3	86.35	76.12	83.26	83.69	66.56	86.68	82.45	71.17	80.76	79.67	2
Doubao-Lite-32K	70.78	63.65	67.08	70.50	85.01	84.74	72.82	76.25	77.17	74.22	3
GeneralV3.5	65.45	69.21	70.76	69.69	77.30	84.17	73.89	73.63	71.35	72.83	4
ERNIE-Bot	58.83	73.70	83.38	70.63	69.75	78.93	80.22	57.33	72.63	71.71	5
EduChat-R1-32B	72.35	67.11	75.16	74.95	70.88	78.20	67.50	66.58	69.99	71.41	6
Baichuan4-Air	72.66	77.98	79.73	70.46	49.53	71.96	77.54	78.80	62.90	71.28	7
GLM-4-AirX	59.96	72.18	60.98	59.15	74.86	78.78	75.79	76.59	74.21	70.28	8
Yi-Lightning	73.32	73.87	68.83	66.45	61.27	83.40	75.83	68.37	57.47	69.87	9
DeepSeek-R1	89.23	71.12	78.06	77.85	56.56	71.93	59.80	55.00	62.66	69.13	10
Hunyuan-Standard	55.19	67.48	50.14	70.87	71.81	79.55	81.59	65.05	72.88	68.28	11
Gemini-1.5-Pro	65.02	53.71	73.64	61.00	73.33	75.68	73.58	70.89	64.11	67.88	12
Qwen-Turbo	63.22	62.43	62.12	69.31	66.55	73.53	58.31	75.67	56.61	65.31	13
Grok-2	62.19	73.54	79.31	74.54	56.18	71.35	65.23	46.47	48.12	64.10	14
Gemini-2.0-Flash	70.52	62.21	67.69	73.72	61.43	66.37	80.04	44.49	49.97	64.05	15
Grok-3	61.11	64.91	68.77	72.55	59.67	68.18	68.01	54.91	57.05	63.91	16
Claude-3.7-Sonnet	73.79	62.08	77.89	63.16	52.20	64.06	65.55	49.17	42.95	61.20	17
GPT-4-Turbo	58.24	56.98	72.74	56.58	49.07	58.31	46.10	57.49	47.94	55.94	18
O1-Mini	82.22	41.52	56.88	76.48	48.43	62.69	43.70	43.58	34.63	54.46	19
LLaMA-3.1-70B	45.37	43.89	59.01	55.53	45.18	52.16	46.82	50.23	48.70	49.65	20
Claude-3.5-Haiku	60.24	46.10	63.26	37.32	43.35	48.92	45.46	30.74	29.46	44.98	21
LLaMA-3.1-8B	20.22	30.84	18.06	19.35	21.82	20.23	24.55	22.85	19.29	21.91	22
EduChat-SFT-13B	9.09	11.51	15.87	14.38	31.30	22.76	18.02	23.75	30.11	19.64	23

Table 1. Composite educational ability performance of evaluated LLMs on nine K-12 subjects. Each cell reports the composite score Λ_{js} (higher is better) for model LLM_j on subject s (math, Chinese, English, physics, chemistry, biology, history, geography and politics). "Avg." is the per-model average across subjects, and "R" is the overall rank by aggregated composite ability. Models in **boldface** have been optimized for educational scenarios. The table is sorted by rank, with Doubao-Pro-32K achieving the highest overall composite performance and EduChat-SFT-13B the lowest.



(a) Model Clustering by Performance in Core Subjects.

(b) Frequency of Top-Tier Rankings Across Subjects.

Figure 2: Analysis of multidimensional subject-level ability and consistency. (a) K-means clustering ($k = 3$) of 23 LLMs in the 3D ability space shown for four representative subjects. Color encodes cluster membership (green=high-performing, yellow=medium, red=low), and marker shape differentiates model type (circle=general-purpose, triangle=education-optimized). (b) Number of subjects (out of nine) in which each model attains a top-tier composite ranking, illustrating that education-optimized models tend to achieve more consistent high performance across subjects.

Experiments

Models We evaluate a diverse collection of 23 LLMs spanning general-purpose and education-optimized paradigms (e.g., GPT-4 Turbo, Claude-3.5-Haiku, EduChat-R1-32B, Doubao-Pro-32K), using either official APIs or publicly released weights. All models are exercised in a zero-shot setting with standardized prompts (details, model list, and prompt templates in Appendix C1) to assess their

raw subject-level capabilities without fine-tuning.

Metrics We quantify each model's subject-level performance via the three abilities: knowledge ability θ_{js} (from IRT), problem-solving ability ρ_{js} (from the O3-based LLM-as-judge process scoring), and educational goal cognition error ε_{js} (mean absolute error in Bloom's taxonomy classification, lower is better). To make these signals comparable and ensure higher-is-better semantics, we define

$$x_{js}^{(1)} = \theta_{js}, \quad x_{js}^{(2)} = \rho_{js}, \quad x_{js}^{(3)} = -\varepsilon_{js},$$

and normalize each dimension across all models within subject s. First, we standardize:

$$z_{js}^{(i)} = \frac{x_{js}^{(i)} - \mu_i^s}{\sigma_i^s}, \tag{2}$$

where μ_i^s and σ_i^s are the mean and standard deviation of $\{x_{1s}^{(i)}, \ldots, x_{Js}^{(i)}\}$, and $i \in \{1, 2, 3\}$. Then we apply min-max scaling to map to $[5, 95]$:

$$\tilde{x}_{js}^{(i)} = 5 + 90 \cdot \frac{z_{js}^{(i)} - \min_j z_{js}^{(i)}}{\max_j z_{js}^{(i)} - \min_j z_{js}^{(i)}}. \tag{3}$$

We denote $\tilde{\theta}_{js} = \tilde{x}_{js}^{(1)}$, $\tilde{\rho}_{js} = \tilde{x}_{js}^{(2)}$, and $\tilde{\varepsilon}_{js} = \tilde{x}_{js}^{(3)}$. The composite ability score is the arithmetic mean, i.e., $\Lambda_{js} = (\tilde{\theta}_{js} + \tilde{\rho}_{js} + \tilde{\varepsilon}_{js})/3$, and models are ranked per subject by Λ_{js}. Further implementation details, including the o3 prompt design and calibration, are provided in Appendix B2.

Main Results

We evaluate all models on the three subject-level abilities and aggregate them into the composite score Λ_{js}; the per-subject results and overall ranking are reported in Table 1 (raw component scores in Appendix C2). Results reveal substantial heterogeneity: models such as Doubao-Pro-32K, DeepSeek-V3, and Doubao-Lite-32K consistently rank near the top, demonstrating strong subject-level knowledge, problem-solving, and educational goal cognition. In contrast, models like EduChat-SFT-13B and LLaMA-3.1-8B fall to bottom, indicating clear room for improvement. After controlling for scale (i.e., excluding small-capacity variants), education-optimized models tend to achieve higher composite ranks than general-purpose counterparts, suggesting that targeted optimization for educational use confers measurable advantages in these core abilities.

To examine the joint distribution of abilities, we perform K-means clustering ($k = 3$) over the triplet $(\theta_{js}, \rho_{js}, \varepsilon_{js})$ per model; the resulting clusters are visualized in Figure 2(a). The top cluster ("high-performing") includes EduChat-R1-32B, DeepSeek-V3, Doubao-Lite-32K, Doubao-Pro-32K, GeneralV3.5, Hunyuan-Standard, ERNIE-Bot, and Gemini-1.5-Pro. A middle cluster captures models with mixed strength (e.g., GLM-4-AirX, Baichuan-Air, Qwen-Turbo), and a lower-performing cluster contains weaker or under-scaled models (e.g., Claude-3.7-Sonnet, GPT-4-Turbo, EduChat-SFT-13B). Figure 2(b) complements this by showing each model's frequency of achieving top-tier subject rankings, highlighting that the highest-ranked models not only score well on average but also deliver consistent performance across subjects.

Taken together, these patterns imply that education-oriented optimization improves both peak and stable performance in the multidimensional ability space, while scale and model design remain limiting factors for some models. The clustering and consistency analysis also reveal that strong performance in one dimension does not guarantee uniform strength across all three—validating the necessity of a joint, subject-level evaluation framework like K-12EduBench.

Discussion

Subject-Level Knowledge and Its Relation to Problem Solving and Accuracy We compute Pearson correlations across all evaluated models (per subject) between the subject-level knowledge ability θ_{js} and both the problem-solving ability ρ_{js} and raw answer accuracy (ACC). Results are reported in Table 2. All correlations are positive and statistically significant ($p < 0.001$), confirming that higher knowledge proficiency is associated with better problem-solving performance and greater likelihood of correct answers. Notably, the correlation between θ_{js} and ACC is consistently stronger than that between θ_{js} and ρ_{js} in most subjects (with English being a close exception), indicating that while IRT-derived knowledge ability closely tracks correctness, problem-solving ability captures additional reasoning quality beyond mere knowledge mastery. At the same time, the substantial θ_{js}-ρ_{js} correlations (e.g., ranging roughly from 0.65 to 0.88) reinforce that subject knowledge is a foundational component of effective problem solving, but not sufficient to fully explain differences in multi-step reasoning performance. Furthermore, Tables 12-15 reveal that 17 LLMs diverge from ACC-based rankings: models with higher θ_{js} tend to perform better on more difficult and discriminative questions (see Appendix C2 and D1 for details). These findings collectively support the validity of the mixed modeling approach, showing that IRT enables finer ability stratification beyond what accuracy alone can reveal.

Integrated Ranking versus Single-Dimension Rankings To understand how balanced versus imbalanced strengths across abilities affect overall evaluation, we compare each model's composite ranking R (from Λ_{js}) with its individual rankings on knowledge ability (R_θ), problem-solving ability (R_ρ), and educational goal cognition (R_ε). Rankings for all models are reported in Appendix D1 (Tables 15–18 for single dimensions and Table 19 for full comparison). Overall, 12 models exhibit relatively stable positions between single-dimension and composite rankings, while 11 show substantial shifts, reflecting imbalances among the three abilities.

Table 3 highlights representative cases. DeepSeek-V3 maintains a high composite rank (2) despite a lower rank on educational goal cognition ($R_\varepsilon = 9$), benefiting from strong knowledge and problem-solving abilities. DeepSeek-R1, by contrast, ranks 2nd on problem solving but drops to 10th overall due to weaker performance in the other dimensions, especially educational goal cognition. Claude-3.7-Sonnet ranks first on educational goal cognition yet falls

Correlation Pair	Math.	Chin.	Engl.	Phys.	Chem.	Biol.	Hist.	Geogr.	Pol.
θ vs. ρ	0.702^{***}	0.744^{***}	0.808^{***}	0.653^{***}	0.879^{***}	0.824^{***}	0.707^{***}	0.672^{***}	0.761^{***}
θ vs. ACC	0.992^{***}	0.987^{***}	0.801^{***}	0.948^{***}	0.989^{***}	0.971^{***}	0.981^{***}	0.939^{***}	0.951^{***}

Table 2. Pearson correlations between subject-level knowledge ability θ_{js} and problem-solving ability ρ_{js}, and between θ_{js} and raw answer accuracy (ACC), computed across all models for each subject. All reported correlations are significant at $p < 0.001$ (***). The consistently higher θ-ACC correlations (relative to θ-ρ) suggest that while knowledge ability aligns closely with correctness, problem-solving ability captures additional multi-step reasoning quality beyond raw knowledge mastery.

LLMs	$\mathbf{R}_\theta$	$\mathbf{R}_\rho$	$\mathbf{R}_\varepsilon$	$\mathbf{R}$
DeepSeek-V3	2	3	9	2
DeepSeek-R1	4	2	21	10
Claude-3.7-Sonnet	18	6	1	17

Table 3. Comparison of single-dimension rankings (knowledge ability R_θ, problem-solving ability R_ρ, educational goal cognition R_ε) versus the integrated composite ranking R for selected models. Large discrepancies between individual and fused rankings reveal imbalanced ability profiles, while consistency indicates more balanced performance.

LLMs	ε (MAE)	$\mathbf{R}_\varepsilon$
Claude-3.7-Sonnet	0.81	1
EduChat-R1-32B	0.83	2
Gemini-2.0-Flash	0.87	3
Grok-2	0.88	4
Grok-3	0.91	5
DeepSeek-R1	1.15	21
LLaMA-3.1-8B	1.23	22
EduChat-SFT-13B	1.34	23
Random Guess	1.65	–
Human	0.77	–

Table 4. Top-5 and bottom-3 LLMs by educational goal cognition error ε_{js}. Lower ε_{js} indicates better alignment with intended instructional cognitive level. All models exceed random guessing but remain below human performance.

to 17th in the integrated ranking, indicating that excellence in a single dimension is insufficient for holistic educational capability. Case analyses (e.g., Figure 8 in Appendix D3) reveal that models with balanced multi-dimensional performance produce not only correct answers but coherent reasoning and better-aligned cognitive objective predictions. These patterns demonstrate that the composite, fused ranking more accurately reflects the multidimensional educational strengths of LLMs than any single metric alone.

Educational Goal Cognition Performance and Error Patterns Table 4 summarizes the best and worst performers on educational goal cognition (full rankings in Appendix D1, Table 17). All LLMs achieve substantially lower error than the random baseline, but still higher error than human evaluation(based on annotations from three freshmen on 90 questions, details in Appendix D2). This indicates that LLMs possess a distinguishable understanding of instructional objectives, but their performance does not reach the human level and absolute errors remain nontrivial—top models (e.g., $\varepsilon_{js} \approx 0.8$) typically err by less than one cognitive level on average, while weaker ones deviate more, reflecting inconsistent alignment. Among the top performers, only EduChat-R1-32B is education-optimized, likely benefiting from its base architecture (e.g., large-scale pretraining) and explicit mechanisms such as an "Educational Chain-of-Thinking" that support structured cognition. Other education-optimized models do not uniformly dominate, suggesting that targeted architectural or prompting enhancements are required to consistently improve this ability.

Error analysis reveals two systematic failure modes. First, **overestimation** occurs when models rely on superficial lexical cues: for instance, interpreting phrasing like "which of the following statements is incorrect" as a signal for "Evaluating" (higher-order) when the true task only requires "Applying," thus violating the principle of task essence. Second, **underestimation** arises when models execute known procedures without fully modeling upstream analytic decisions—for example, applying a combinatorial formula directly (labeling it as "Applying") while missing the deeper "Analyzing" step of correctly framing the problem. These errors reflect breakdowns in cognitive path completeness and show that models sometimes fail to discern the intended depth of thinking even if they can produce plausible answers. Representative illustrative cases are in Appendix D3.

In summary, while LLMs are better than random at recognizing instructional goals, their educational goal cognition remains limited: they frequently over- or under-estimate cognitive levels, leading to systematic misalignment. Improving sensitivity to task essence and enforcing cognitive path integrity are crucial next steps to close this gap for both general-purpose and education-optimized models.

Conclusion

We present K-12EduBench, a comprehensive benchmark for evaluating LLMs' subject-level knowledge, problem-solving, and educational goal cognition abilities in K-12 education. Grounded in IRT, problem-solving theory, and the revised Bloom's Taxonomy, it addresses limitations of existing benchmarks by disentangling mastery, reasoning, and instructional intent. Empirically, education-optimized models (at sufficient scale) tend to outperform general-purpose ones, yet imbalances persist and educational goal cognition remains weak. K-12EduBench delivers unified diagnostics for selecting and improving education-aligned LLMs, promoting research on education-aligned model development.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant No.62477005, the Humanities and Social Science Fund of Ministry of Education of China under grant No.24YJA880104 and the Jilin Province Key R&D Plan Project under grant No.20230201063GX.

References

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

Chen, Y.; Wu, C.; Yan, S.; Liu, P.; and Xiao, Y. 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)*, 3138–3167.

Dewey, J. 1933. *How we think: A restatement of reflective thinking to the educative process*. Heath (DC).

Embretson, S. E.; and Reise, S. P. 2013. *Item response theory for psychologists*. Psychology Press.

Gu, Z.; Zhu, X.; Ye, H.; Zhang, L.; Wang, J.; Zhu, Y.; Jiang, S.; Xiong, Z.; Li, Z.; Wu, W.; et al. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18099–18107.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. arXiv:2103.03874.

Hou, J.; Ao, C.; Wu, H.; Kong, X.; Zheng, Z.; Tang, D.; Li, C.; Hu, X.; Xu, R.; Ni, S.; et al. 2024. E-eval: a comprehensive Chinese k-12 education evaluation benchmark for large language models. *arXiv preprint arXiv:2401.15927*.

Hu, B.; Zheng, L.; Zhu, J.; Ding, L.; Wang, Y.; and Gu, X. 2024. Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design. *IEEE Transactions on Learning Technologies*, 17: 1445–1459.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; et al. 2023a. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36: 62991–63010.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36: 62991–63010.

Lelièvre, M.; Waldock, A.; Liu, M.; Aspillaga, N. V.; Mackintosh, A.; Portela, M. J. O.; Lee, J.; Atherton, P.; Ince, R. A. A.; and Garrod, O. G. B. 2025. Benchmarking the Pedagogical Knowledge of Large Language Models. arXiv:2506.18710.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, 11260–11285.

Liu, C.; Jin, R.; Ren, Y.; Yu, L.; Dong, T.; Peng, X.; Zhang, S.; Peng, J.; Zhang, P.; Lyu, Q.; Su, X.; Liu, Q.; and Xiong, D. 2023. M3KE: A Massive Multi-Level Multi-Subject Knowledge Evaluation Benchmark for Chinese Large Language Models. arXiv:2305.10263.

Macina, J.; Daheim, N.; Hakimi, I.; Kapur, M.; Gurevych, I.; and Sachan, M. 2025. MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors. arXiv:2502.18940.

Miller, P.; and DiCerbo, K. 2024. Llm based math tutoring: Challenges and dataset.

Mirza, A.; Alampara, N.; Kunchapu, S.; Ríos-García, M.; Emoekabu, B.; Krishnan, A.; Gupta, T.; Schilling-Wilhelmi, M.; Okereke, M.; Aneesh, A.; et al. 2024. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

Newell, A.; Shaw, J. C.; and Simon, H. A. 1958. Elements of a theory of human problem solving. *Psychological review*, 65(3): 151.

Popper, K. R. 1999. *All life is problem solving*. Psychology Press.

Tai, T.-Y.; and Chen, H. H.-J. 2024. Improving elementary EFL speaking skills with generative AI chatbots: Exploring individual and paired interactions. *Computers & Education*, 220: 105112.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Wiggins, G. 1998. *Educative Assessment. Designing Assessments To Inform and Improve Student Performance.* ERIC.

You, W.; Wang, P.; Li, C.; Ji, Z.; and Bai, J. 2024. CK12: A Rounded K12 Knowledge Graph Based Benchmark for Chinese Holistic Cognition Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19431–19439.

Zeng, H. 2023. Measuring Massive Multitask Chinese Understanding. arXiv:2304.12986.

Zhang, X.; Li, C.; Zong, Y.; Ying, Z.; He, L.; and Qiu, X. 2024. Evaluating the Performance of Large Language Models on GAOKAO Benchmark. arXiv:2305.12474.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2299–2314.