# OMNIEDUBENCH: A CHINESE DATASET FOR EVALUATING SUBJECT KNOWLEDGE AND PEDAGOGICAL JUDGMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Evaluating large language models (LLMs) for educational use requires distinguishing subject knowledge from judgment about how to respond to learners. Knowledge-oriented assessments provide limited evidence about such pedagogical judgment, motivating complementary evaluation in Chinese contexts. We introduce **OmniEduBench, a Chinese dataset for evaluating subject-knowledge problem solving and scenario-based pedagogical judgment**. It comprises approximately 24.6K question–answer pairs spanning 41 academic subjects and 20 pedagogical competencies. The knowledge component covers 11 question formats across multiple educational stages, while the pedagogical component primarily uses multiple-choice scenarios involving instructional guidance, feedback, and learner support. The dataset combines public educational resources, private institutional materials, and model-generated scenarios, followed by cleaning, model-based selection, and human review. We evaluate 15 LLMs and analyze the sensitivity of open-ended answer scoring to the choice of LLM judge. Results show substantial variation across subjects and pedagogical categories, while evaluator choice materially affects absolute scores. Findings highlight the need for scoring criteria and evaluator validation when interpreting educational benchmarks. OmniEduBench provides a resource for diagnosing subject-knowledge and pedagogical-judgment limitations, with pedagogical scores interpreted as scenario-based judgments rather than direct measures of teaching effectiveness.

## 1 INTRODUCTION

With the rapid emergence of large language models (LLMs), evaluation benchmarks have become increasingly critical, shifting the focus of assessment toward broader and complex skills. To address the demands of this complex paradigm, a variety of benchmarks have been proposed to evaluate the diverse capabilities of LLMs. These benchmarks cover a wide spectrum of areas, including knowledge and language understanding (*e.g.*, MMLU (Hendrycks et al., 2021a), ARC (Clark et al., 2018)), reasoning (*e.g.*, GSM8K (Cobbe et al., 2021), AIME (Patel et al., 2024)), multi-turn open-ended dialogue (*e.g.*, MT-bench (Bai et al., 2024)), and coding (*e.g.*, MBPP (Austin et al., 2021)). Serving as indispensable tools for advancing LLM development, these benchmarks have been widely adopted in recent influential works (Hurst et al., 2024; Liu et al., 2024a; Team, 2025b; Comanici et al., 2025; Taylor et al., 2022; Touvron et al., 2023; OpenAI, 2023; Hoffmann et al., 2022).

In recent years, a series of powerful Chinese LLMs emerged, such as the Qwen (Yang et al., 2024a; 2025), DeepSeek (Liu et al., 2024a; Guo et al., 2025; Liu et al., 2024b), achieving performance levels comparable to overseas LLMs. With the growing application of LLMs in education, researchers have also begun to propose Chinese education benchmarks, which can be broadly categorized into two types: (1) datasets translated from other languages and (2) datasets natively constructed from Chinese education corpora. Specifically, (1) Datasets translated from other languages refer to benchmarks constructed by directly translating existing benchmarks from other languages into Chinese. A representative work is CLUE (Xu et al., 2020a), which was translated from the English GLUE (Wang et al., 2019). However, a simple translation approach is insufficient for a rigorous evaluation of LLMs in Chinese. These datasets often fail to reflect the unique linguistic and cultural challenges of the Chinese education and inherently carry biases from their original environment, thus limiting their ability to assess LLMs' understanding of local education knowledge.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

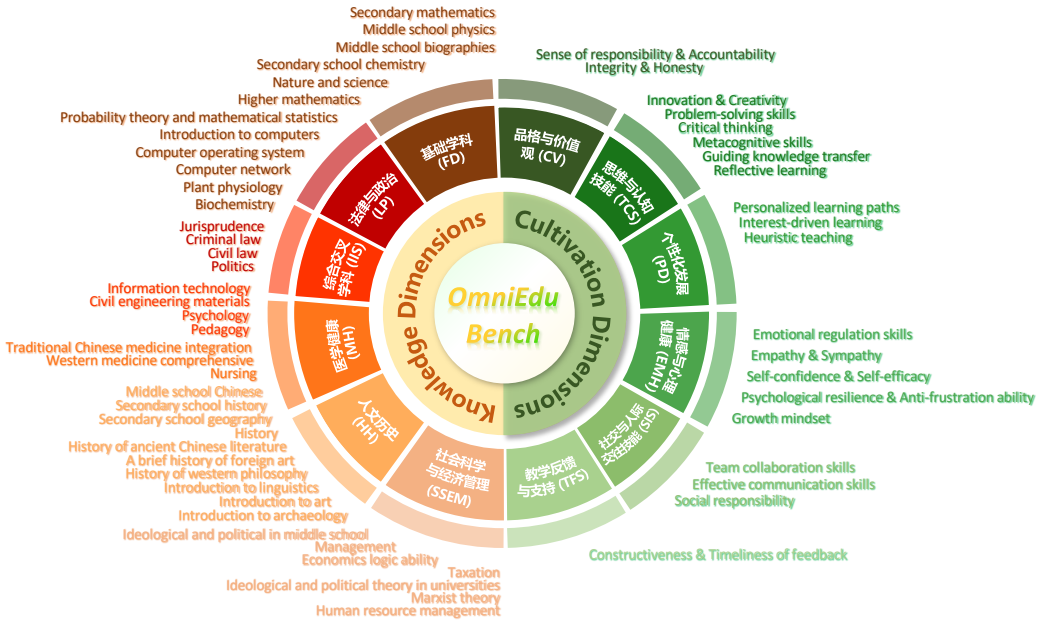


Figure 1: **Overview of OmniEduBench.** Subject-knowledge component covers 41 academic subjects, while pedagogical judgment covers 20 pedagogical topics, each organized into six categories.

(2) Datasets natively constructed from Chinese educational corpora refer to benchmarks directly collected from Chinese educational text resources, such as C-Eval (Huang et al., 2023), Edubench (Xu et al., 2025), Scieval (Sun et al., 2024), AGIEval (Zhong et al., 2023), and SuperCLUE (Xu et al., 2023). However, most existing education benchmarks are often limited to a single subject or question type, lacking sufficient diversity. Additionally, these datasets typically focus on the knowledge dimension, overlooking the unique cultivation aspects that are essential in real-world education.

We present OmniEduBench, a comprehensive Chinese education benchmark designed to thoroughly evaluate LLMs in terms of both knowledge understanding and skill cultivation in educational scenarios. OmniEduBench encompasses knowledge and cultivation dimension and comprises a total of 24.602K high-quality question–answer pairs, covering 11 common exam question types (*e.g.*, multiple choice (单选题), multiple answer (多选题), fill-in-the-blank (填空题), short answer (简答题), composite questions (复合题), term explanation (名词解释), True/False (判断题), calculation (计算题), logical reasoning (逻辑推理题), case analysis (案例分析题), and essay (论述题)), as illustrated in Figure 1. The knowledge dimension includes 18.121K question–answer pairs spanning 41 subject areas, from humanities to science and engineering, and covering five difficulty levels: elementary school, middle school, high school, college, and professional examinations. The cultivation dimension comprises 6.481K question–answer pairs across 20 teaching-related comments, including guided teaching, student emotional support, and moral education (see Table 1 for details), aiming to comprehensively assess the diverse competencies required in real-world educational settings. Extensive experiments demonstrate that our proposed OmniEduBench presents a highly challenging and significant benchmark for Chinese educational evaluation. Additionally, we introduce OmniEduBench HARD, a high-difficulty subset of OmniEduBench, specifically targeting particularly demanding subjects such as advanced mathematics and competitions that require sophisticated reasoning skills. Even the state-of-the-art LLMs achieve less than 50% accuracy on this subset, highlighting the rigor and necessity of our proposed OmniEduBench education benchmark.

## 2 Related Work

**Large language models (LLMs).** LLMs have advanced at an unprecedented pace. Leveraging increasingly sophisticated architectures and ever-larger pretraining corpora, they have continuously pushed the boundaries of performance in language understanding, reasoning, and generation tasks. Researchers have explored various approaches to enhance LLMs' capabilities. For example, Chain-

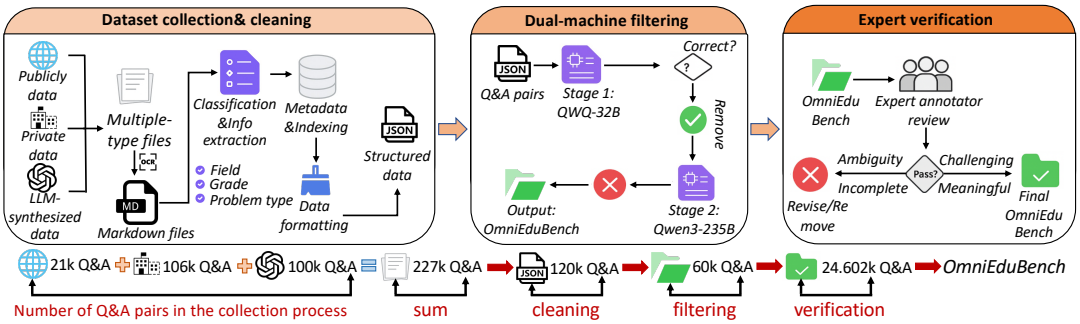


Figure 2: **Overview of the construction process:** collection, cleaning, filtering, and verification.

of-Thought prompting (Wei et al., 2022; Team, 2025c;b; Guo et al., 2025; Liu et al., 2024a) has been shown to be highly effective in guiding models to perform step-by-step reasoning for complex problem-solving. Instruction tuning (Dong et al., 2025; Hu et al., 2025; Yang et al., 2024a; 2025) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Schulman et al., 2017) have been widely adopted to align model outputs with human intentions and preferences, enabling LLMs to generate responses that are more natural and reliable in open-ended dialogue and creative tasks. Despite these remarkable advances, however, the question of how to comprehensively and effectively evaluate the true capabilities of LLMs remains a critical and open challenge.

**English education benchmarks.** Researchers have proposed a variety of benchmarks to evaluate the capabilities of LLMs, which can be broadly categorized into three types: (1) task-specific evaluations, such as reading comprehension (SQuAD (Rajpurkar et al., 2016)), machine translation (Bojar et al., 2014)), and summarization (Hermann et al., 2015); (2) general knowledge and advanced ability evaluations, for example, the massive multitask language understanding (MMLU) (Hendrycks et al., 2021a), which collects questions from real-world exams and textbooks to provide a diverse, multi-domain test that effectively probes the breadth and depth of model knowledge. Similarly, the BIG-bench (Srivastava et al., 2022) comprises 204 diverse tasks, and (3) specialized ability evaluations. In mathematical reasoning, such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b) assess models' ability to solve complex multi-step problems. In code generation, HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) have become standard benchmarks for measuring programming proficiency. Additionally, datasets such as MT-bench (Zheng et al., 2023) have been introduced to evaluate performance in multi-turn, open-ended dialogues. Despite the significant contributions of these datasets to advancing LLMs evaluation, most of them remain heavily focused on English settings, with limited coverage of Chinese scenarios.

**Chinese education benchmarks.** A series of comprehensive Chinese benchmarks has been proposed. For example, CLUE (Xu et al., 2020b), as an early work, integrates multiple natural language understanding tasks and has become an important reference for evaluating LLMs. Subsequently, benchmarks such as CMMLU (Li et al., 2023) and C-Eval (Huang et al., 2023) collect multi-disciplinary, multi-task questions from Chinese university exams, professional qualification tests, and textbooks, effectively assessing models' general knowledge and their understanding. Beyond general capability evaluation, researchers have also developed Chinese benchmarks targeting specific advanced skills. In mathematical reasoning, CMATH (Wei et al., 2023) tests models' abilities to solve complex mathematical problems. Meanwhile, EduBench (Xu et al., 2025) constructs synthetic corpora for education, but its question types are relatively limited, making it difficult to fully capture models' Chinese potential. To address this critical gap, we propose OmniEduBench — a Chinese education benchmark that uniquely combines knowledge and cultivation dimensions, providing a novel framework for systematically evaluating LLMs' potential for education.

## 3 OmniEduBench

OmniEduBench is designed as a natively Chinese education evaluation benchmark that captures the unique linguistic and cultural knowledge of Chinese education, encompasses diverse question types, and assesses LLMs not only on their knowledge capabilities but also on the distinctive cultivation competencies required in real-world educational scenarios.

Table 1: Statistics of our proposed OmniEduBench education benchmark and more detailed per-subject information are shown in the Appendix 6.2. Bilingual names and abbreviations of six knowledge dimensions and six cultivation dimensions.

| *Knowledge dimension* | | | *Cultivation dimension* | | |
|---|---|---|---|---|---|
| English name | Abbreviation | Chinese name | English name | Abbreviation | Chinese name |
| Law & Politics | LP | 法律与政治 | Character & Values | CV | 品格与价值观 |
| Foundational Disciplines | FD | 基础学科 | Personalized Development | PD | 个性化发展 |
| Humanities & History | HH | 人文与历史 | Social & Interpersonal Skills | SIS | 社会与人际交往 |
| Medicine & Health | MH | 医学与健康 | Thinking & Cognitive Skills | TCS | 思维与认知能力 |
| Interdisciplinary & Integrated Subjects | IIS | 综合与交叉学科 | Teaching Feedback & Support | TFS | 教学反馈与支持) |
| Social Sciences & Economics Management | SSEM | 社会科学与经济管理 | Emotional & Mental Health | EMH | 情感与心理健康 |

| Category | Subjects | Questions | Category | Subjects | Questions | Category | Subjects | Questions |
|---|---|---|---|---|---|---|---|---|
| *In terms of dimension* | | | *In terms of Knowledge* | | | *In terms of Cultivation* | | |
| Knowledge | 41 | 18,121 | LP | 4 | 1,455 | CV | 2 | 694 |
| Cultivation | 20 | 6,481 | FD | 11 | 7,918 | PD | 3 | 1,031 |
| *In terms of different level* | | | HH | 10 | 5,331 | SIS | 3 | 736 |
| K-12 Schools | 10 | 4,384 | MH | 3 | 918 | TCS | 6 | 1,900 |
| High school | 11 | 6,735 | IIS | 4 | 914 | TFS | 1 | 193 |
| College | 30 | 6,364 | SSEM | 9 | 1,643 | EMH | 5 | 1,833 |
| Total | 61 | 24,602 | Total | 41 | 18,179 | Total | 20 | 6,387 |

### 3.1 Task Definition

**Knowledge dimension** focuses on evaluating the model's mastery of subject-specific knowledge. Tasks in this dimension include 11 common exam question types (*e.g.*, multiple choice, multiple answer, fill-in-the-blank, short answer, composite questions, term explanation, True/False, calculation, logical reasoning, case analysis, and essay). These 11 question types span a wide range of disciplines, from humanities and history to science, engineering, and professional fields. The primary goal is to assess the LLM's problem-solving.

**Cultivation dimension** assesses LLMs on their ability to support holistic educational objectives beyond mere knowledge acquisition. This includes guiding students' thinking processes, fostering moral and value development, enhancing emotional understanding, and promoting critical reasoning skills. Tasks in this dimension are designed to reflect realistic learning scenarios, where models must provide pedagogical feedback that aligns with students' cognitive.

### 3.2 Benchmark Construction

In this section, we provide a detailed overview of the construction process for OmniEduBench in Figure 2. The process consists of four key stages: dataset collection, dataset cleaning, dual-machine filtering, and expert verification.

**Dataset collection.** OmniEduBench is designed to encompass a wide range of diverse scenarios to enable comprehensive evaluation. To achieve this, we employ three distinct data collection methods, carefully balancing diversity and efficiency in the construction of the OmniEduBench.

*Manual collection of publicly available data.* Existing benchmarks often lack sufficient diversity in question types and knowledge coverage, making them inadequate for our 41 subjects in the knowledge dimension. To address this gap, we manually collected additional data from publicly available online resources (*e.g.*, XuekeNet, ZujuanNet, ShijuanNet, ShitiNet) to enrich diversity and ensure coverage of underrepresented scenarios, such as primary and career education. Furthermore, guided by the Catalogue of Undergraduate Programs in Regular Higher Education Institutions [1] issued by China's Ministry of Education, we curated a large body of review materials and exam questions across 13 academic disciplines, including philosophy, education, law, literature, history, science, engineering, agriculture, medicine, military science, management, and the arts. This effort significantly improves distributional balance and provides a more faithful reflection of real applications.

*Manual collection of private data.* Data contamination remains one of the most critical challenges in constructing evaluation datasets. To mitigate this risk, we manually collected additional data

[1] http://www.moe.gov.cn/srcsite/A08/moe_1034/s4930/202403/W020240319305498791768.pdf

Table 2: Expert validation results for our proposed OmniEduBench education evaluation benchmark.

| Metric English name | Metric Chinese name | Average | Standard deviation | Inter-rater agreement |
|---|---|---|---|---|
| Overall quality | 整体质量 | 4.8 | 0.1 | 0.90 |
| Clarity | 问题清晰度 | 4.5 | 0.2 | 0.85 |
| Option perplexity | 选项困惑度 | 4.8 | 0.3 | 0.83 |
| Accuracy | 答案准确性 | 4.8 | 0.1 | 0.90 |
| Cultivation value | 育人价值 | 4.6 | 0.2 | 0.88 |

Table 3: Comparison of OmniEduBench with existing benchmarks. Mixed (X) indicates X types of mixed question types, and General (K) signifies that the dataset covers all subjects and educational stages with K categories. ZH and EN represent Chinese and English.

| Benchmark | Dimension | Size | Type | Data source | Domain | Subject | Language |
|---|---|---|---|---|---|---|---|
| C-Eval (Huang et al., 2023) | Knowledge | 13.9K | Multiple answer | Publicly available | General (4) | 52 | ZH |
| EduBench (Xu et al., 2025) | Knowledge | 18.8K | Multiple choice | LLM-generated | General (9) | 9 | ZH & EN |
| CMATH (Wei et al., 2023) | Knowledge | 1.7K | Multiple choice | Publicly available | Math | 1 | ZH |
| E-EVAL Hou et al. (2024) | Knowledge | 4.35K | Multiple answer | Publicly available | K-12 | 23 | ZH |
| GAOKAO (Zhang et al., 2023) | Knowledge | 2.8K | Mixed (7) | Publicly available | High school | 9 | ZH |
| GSM8K (Cobbe et al., 2021) | Knowledge | 8.5K | Multiple choice | Publicly available | Math | 1 | EN |
| MATH (Hendrycks et al., 2021b) | Knowledge | 12.5K | Multiple choice | Publicly available | Math | 7 | EN |
| OmniEduBench (ours) | Knowledge & Cultivation | 24.6K | Mixed (11) | Publicly & Private & LLM | General (12) | 61 | ZH |

from private resources, such as internal school exam papers. Unlike widely circulated national exams, these materials have never appeared on the public Internet or been included in large-scale web crawls, effectively reducing the risk of leakage. Incorporating such private data enhances the reliability and fairness of the benchmark, while providing a more rigorous assessment of models.

*LLM-generated data.* Given the difficulty of directly obtaining data in the cultivation dimension, we leveraged LLMs to generate a substantial number of scenario-based question–answer pairs, aiming to supplement gaps in existing resources. To ensure the quality of the synthetic data, we invited five education experts to conduct discussions on 20 cultivation subjects and consulted relevant books, papers, and other materials. The collected content was organized into a database, which was then provided to the LLM to enhance the fidelity and accuracy of the generated data. For generated questions, to increase their challenge, we designed highly confounding distractors via prompts and conducted sampling checks and revisions with expert verification (please see more details in expert verification, and the generated data process can be found in Appendix 6.1). Finally, we collected a total of 227k question–answer (Q&A) pairs, including 21k from publicly available data, 106k from private data, and 100k generated by LLMs.

**Dataset cleaning.** The entire data cleaning process consists of multiple steps. (1) We used MinerU (Wang et al., 2024) to convert the collected 227k Q&A pairs into Markdown (md) format, enabling structured management and extraction. (2) Detailed metadata were extracted for each question, including subject, grade level, question type, and knowledge tags, to construct comprehensive question profiles that facilitate data management and subsequent analysis. (3) Standard data cleaning procedures were applied, including deduplication, removal of questions with missing key content, filtering of sensitive or inappropriate content, and exclusion of questions that rely on external information. After the cleaning process, we obtained a total of 120k Q&A pairs.

**Dual-machine filtering.** To ensure OmniEduBench is a high-quality and challenging benchmark, we implemented a dual-model filtering mechanism on an initial set of 120k Q&A pairs. Specifically, we first evaluated all questions using QWQ32B (Team, 2025c), retaining only those that the model answered incorrectly. This initial filtering resulted in a subset of 90k Q&A pairs. These questions then underwent a second filtering stage with the same strategy, this time using Qwen3-235B (Yang et al., 2025), ultimately yielding the final set of 60k high-quality and challenging data.

**Expert verification.** We recruited 50 master's students to perform an initial quality check on the dataset based on five predefined dimensions (in Table 2), removing any data that did not meet the criteria, which resulted in a final set of 24.602k Q&A pairs for OmniEduBench. Among these 24.602k Q&A pairs, 25.1% were sourced from publicly available data, 54.1% from private data, and the remaining 20.8% from LLM-generated data. Subsequently, we invited 5 senior experts to conduct a rigorous quality review on a 15% random sample drawn separately from publicly available, pri-

Table 4: Zero-shot average accuracy (%) across six categories in the **knowledge (Top)** and **cultivation (Bottom)** dimensions. The highest accuracy is **bold**, and the second highest is underlined. More experimental results are provided in the Appendix 6.2.

| Knowledge | Parameters | Access | Creator | FD | HH | SSEM | LP | MH | IIS | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3 | 8B | Weights | Alibaba | 53.02 | 38.53 | 36.58 | 30.17 | 36.71 | 37.75 | 43.86 |
| Qwen3 | 14B | Weights | Alibaba | 36.32 | 36.78 | 35.12 | 27.29 | 36.82 | 35.67 | 35.62 |
| MuduoLLM | 14B | Weights | BNU & TAL | 28.20 | 40.82 | 32.99 | 36.15 | 39.11 | 31.40 | 33.68 |
| QwQ | 32B | Weights | Alibaba | 61.25 | 48.51 | 42.24 | 49.90 | 55.01 | 47.26 | 53.87 |
| Seed-OSS | 36B | Weights | ByteDance | 48.81 | 50.14 | 45.34 | 48.66 | **61.00** | 49.56 | 49.53 |
| Qwen2.5 | 72B | Weights | Alibaba | 19.53 | 30.95 | 20.57 | 13.26 | 23.86 | 20.90 | 22.76 |
| Qwen3 | 235B (22B active) | Weights | Alibaba | 34.24 | 47.01 | 36.21 | 44.26 | 58.71 | 46.61 | 40.82 |
| DeepSeek-V3.1 | 671B (37B active) | Weights | DeepSeek | 31.65 | 40.65 | 35.00 | 29.42 | 50.54 | 45.19 | 36.05 |
| HuatuoGPT-o1 (Medicine) | 7B | Weights | CUHK-Shenzhen | 23.67 | 24.23 | 23.38 | 18.79 | 29.74 | 20.94 | 23.46 |
| Qwen2.5-Math-instruct (Math) | 7B | Weights | Alibaba | 33.07 | 8.43 | 11.51 | 8.39 | 10.67 | 13.93 | 14.33 |
| GPT-4o | Undisclosed | API | OpenAI | 21.15 | 26.94 | 23.92 | 22.13 | 34.75 | 27.13 | 24.17 |
| Claude-4 Sonnet | Undisclosed | API | Anthropic | 41.49 | 44.29 | 35.36 | 27.56 | 34.86 | 42.34 | 40.35 |
| ERNIE-4.5-Turbo | Undisclosed | API | BaiDu | 39.89 | 51.59 | 45.61 | 52.07 | 59.15 | 47.43 | 49.29 |
| Spark_Pro | Undisclosed | API | iFlytek | 22.96 | 37.03 | 41.58 | 41.03 | 36.82 | 37.90 | 36.22 |
| Gemini-2.5 Pro | Undisclosed | API | Google | **73.83** | **55.13** | **46.68** | **55.40** | 60.68 | **54.16** | **62.76** |

| Cultivation | Parameters | Access | Creator | TCS | EMH | SIS | CV | PD | TFS | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3 | 8B | Weights | Alibaba | 70.95 | 66.67 | 69.16 | 62.25 | 70.13 | **77.20** | 68.62 |
| Qwen3 | 14B | Weights | Alibaba | 67.79 | 60.77 | 63.72 | 56.20 | 64.31 | 71.50 | 63.60 |
| MuduoLLM | 14B | Weights | BNU & TAL | 64.42 | 60.77 | 63.45 | **66.14** | 67.51 | 64.77 | 63.96 |
| QwQ | 32B | Weights | Alibaba | **73.16** | 68.36 | 69.84 | 65.13 | **71.77** | 72.02 | **70.27** |
| Seed-OSS | 36B | Weights | ByteDance | 70.74 | 65.30 | 66.03 | 62.82 | 67.12 | 70.47 | 67.18 |
| Qwen2.5 | 72B | Weights | Alibaba | 67.89 | 64.38 | 65.62 | 59.51 | 65.57 | 67.88 | 65.34 |
| Qwen3 | 235B (22B active) | Weights | Alibaba | 67.84 | 61.10 | 64.54 | 55.76 | 64.40 | 70.47 | 63.74 |
| DeepSeek-V3.1 | 671B (37B active) | Weights | DeepSeek | 71.58 | 65.41 | 69.02 | 61.96 | 71.00 | **77.20** | 68.55 |
| HuatuoGPT-o1 (Medicine) | 7B | Weights | CUHK-Shenzhen | 42.58 | 45.57 | 48.10 | 41.79 | 43.65 | 52.85 | 44.44 |
| Qwen2.5-Math-instruct (Math) | 7B | Weights | Alibaba | 28.74 | 31.44 | 20.52 | 27.52 | 16.97 | 22.28 | 26.34 |
| GPT-4o | Undisclosed | API | OpenAI | 61.63 | 59.57 | 59.24 | 55.33 | 57.71 | 65.80 | 59.57 |
| ERNIE-4.5-Turbo | Undisclosed | API | BaiDu | 70.11 | 62.85 | 66.30 | 56.77 | 67.02 | 70.47 | 65.65 |
| Spark_Pro | Undisclosed | API | iFlytek | 68.05 | 61.87 | 64.95 | 61.10 | 67.80 | 72.02 | 65.24 |
| Claude-4-sonnet | Undisclosed | API | Anthropic | 71.95 | **70.05** | **70.92** | 64.55 | 69.25 | 71.50 | 70.03 |
| Gemini-2.5-pro | Undisclosed | API | Google | 72.26 | 66.07 | 70.79 | 65.71 | 70.32 | 67.36 | 69.14 |

vate, and LLM-generated data of OmniEduBench. Table 2 indicates that the dataset maintains high overall quality, demonstrating both reliability and applicability.

### 3.3 Evaluation Criteria

Based on the characteristics of different question types, we adopt two evaluation metrics: (1) **Choice**. For questions with a standard answer, we directly evaluate the provided answer. This simplifies the scoring process, as the model only needs to select the most appropriate option, thereby reducing ambiguity in assessment. (2) **LLM-assisted scoring**. For short-answer questions that may have multiple valid forms but are semantically equivalent, we employ an LLM-assisted scoring method. This approach provides greater flexibility, avoids imposing unnecessary constraints on the model, and allows for a more accurate evaluation of the model's semantic understanding and expression.

### 3.4 Statistics

Through rigorous multi-stage filtering and expert validation, we construct a high-quality dataset comprising 18.121k Q&A pairs for Knowledge and 6.481k for Cultivation, as illustrated in Figure 1 and summarized in Table 1. As shown in the comparative analysis in Table 3, OmniEduBench significantly outperforms existing benchmarks across seven key dimensions. Our proposed benchmark is uniquely designed to reflect authentic pedagogical scenarios, spanning K-12, high school, and university-level curricula. Furthermore, it incorporates essential cultivation dimensions, such as emotional intelligence and logical reasoning, encompassing a total of 61 diverse real-world educational scenarios.

## 4 Experiments

In this section, we evaluate the performance of 15 state-of-the-art LLMs in both English and Chinese. Results indicate that OmniEduBench remains a competitive benchmark.

Table 5: Average accuracy (%) across six categories in one-shot, three-shot, and five-shot settings for the knowledge dimension. The highest accuracy is **bold**, and the second highest is underlined.

| Model | Parameters | Access | Creator | FD | HH | SSEM | LP | MH | IIS | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| *One-shot setting* | | | | | | | | | | |
| Qwen3 | 8B | Weights | Alibaba | **52.80** | 46.45 | **41.90** | 29.76 | 40.20 | 40.59 | **41.95** |
| MuduoLLM | 14B | Weights | BNU & TAL | 27.36 | 47.79 | 36.72 | 34.98 | 40.74 | 34.35 | 36.99 |
| Qwen2.5 | 72B | Weights | Alibaba | 21.42 | 40.43 | 27.04 | 20.96 | 28.10 | 22.65 | 26.77 |
| Qwen3 | 235B (22B activate) | Weights | Alibaba | 37.72 | **60.79** | 44.03 | **45.77** | **59.59** | **54.05** | **50.12** |
| DeepSeek-V3.1 | 671B (37B activate) | Weights | DeepSeek | 30.00 | 41.73 | 34.65 | 30.72 | 49.67 | 42.12 | 38.15 |
| *Three-shot setting* | | | | | | | | | | |
| Qwen3 | 8B | Weights | Alibaba | **52.98** | 46.00 | **39.74** | 30.65 | 39.32 | 40.85 | 41.59 |
| MuduoLLM | 14B | Weights | BNU & TAL | 27.32 | 46.86 | 35.79 | 33.81 | 39.54 | 32.42 | 35.96 |
| Qwen2.5 | 72B | Weights | Alibaba | 21.43 | 40.86 | 27.27 | 20.41 | 27.12 | 23.88 | 26.83 |
| Qwen3 | 235B (22B activate) | Weights | Alibaba | 37.52 | **60.70** | 43.40 | **45.77** | **59.48** | **52.57** | **49.54** |
| DeepSeek-V3.1 | 671B (37B activate) | Weights | DeepSeek | 29.09 | 41.42 | 34.02 | 28.59 | 48.80 | 42.06 | 37.33 |
| *Five-shot setting* | | | | | | | | | | |
| Qwen3 | 8B | Weights | Alibaba | **56.86** | 46.70 | **39.44** | 30.65 | 38.24 | 42.23 | 42.35 |
| MuduoLLM | 14B | Weights | BNU & TAL | 26.93 | 46.57 | 36.28 | 35.74 | 39.11 | 34.57 | 36.53 |
| Qwen2.5 | 72B | Weights | Alibaba | 21.46 | 41.19 | 26.96 | 20.82 | 26.03 | 28.56 | 27.50 |
| Qwen3 | 235B (22B activate) | Weights | Alibaba | 37.40 | **60.41** | 44.13 | **45.77** | **58.61** | **55.58** | **50.32** |
| DeepSeek-V3.1 | 671B (37B activate) | Weights | DeepSeek | 29.39 | 41.17 | 32.87 | 28.45 | 47.49 | 38.95 | 36.39 |

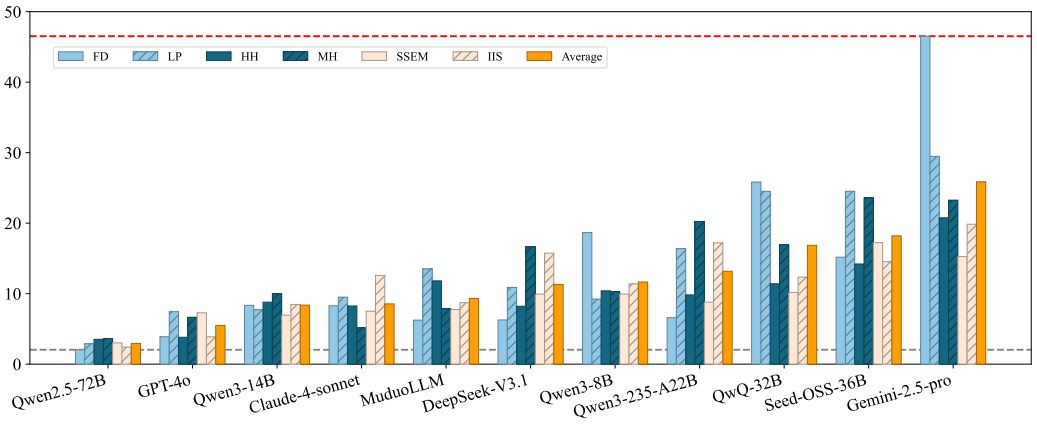


Figure 3: Zero-shot average accuracy (%) on the knowledge (Top) and cultivation (see Figure 4) dimensions of OmniEduBench HARD.

### 4.1 Experimental Setup

**Baselines.** We evaluate 15 mainstream LLMs in total, including 5 cutting-edge closed-source models and 10 open-source models, one of which is a newly released education-oriented model. Closed-source models are GPT-4o (Hurst et al., 2024), Gemini-2.5 Pro (Comanici et al., 2025), Claude-4 Sonnet (Team, 2025a) and education-specific models (ERNIE-4.5-Turbo Zhang et al. (2019) and Spark Pro Wang et al. (2025)). For open-source models, we consider two main factors. First, they are grouped by parameter size into small (8B), medium (14B/32B/36B), and large (72B/235B/671B) scales. Second, they are categorized by functionality into: (a) general instruction-following models (Qwen2.5 (Yang et al., 2024a), Qwen3 (Yang et al., 2025)); (b) general reasoning models (QwQ (Team, 2025c), Seed-OSS (Team, 2025b), DeepSeek-V3.1 (Liu et al., 2024b)); and (c) education-specific models (MuduoLLM (from BNU & TAL, 2025), HuatuoGPT-o1 Chen et al. (2024) and Qwen2.5-Math-instruct Yang et al. (2024b)).

**Implementation details.** We evaluate all LLMs under both zero- and few-shot settings, with few-shot examples (0-, 1-, 3-, and 5-shot) drawn from a separately partitioned development set, distinct from the evaluation set. All open-source models are run using their official code, while closed-source models are accessed via official APIs. We consistently use Gemini-2.5 Pro as the LLM-assisted scoring model, unless otherwise specified. Following C-Eval (Huang et al., 2023) and GSM8K (Cobbe et al., 2021), we set the temperature to 0 during inference, which corresponds to using greedy decoding. At each generation step, the model selects the token with the highest

Table 6: Zero-shot average accuracy (%) across six categories in the knowledge using different LLM-assisted scoring methods. The highest accuracy is **bold**, and the second highest is underlined.

| Model | Parameters | Access | Creator | FD | HH | SSEM | LP | MH | IIS | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| *Qwen3-A235B-assisted scoring method (Yang et al., 2025)* | | | | | | | | | | |
| QwQ | 32B | Weights | Alibaba | 61.26 | 55.82 | 45.22 | 50.10 | 55.66 | 48.36 | 56.39 |
| Seed-OSS | 36B | Weights | ByteDance | 51.16 | 63.04 | 51.55 | 50.38 | **62.31** | 57.33 | 55.49 |
| GPT-4o | Undisclosed | API | OpenAI | 23.59 | 36.77 | 28.79 | 23.92 | 35.73 | 31.40 | 28.96 |
| Claude-4 Sonnet | Undisclosed | API | Anthropic | 43.54 | 55.52 | 40.47 | 27.77 | 36.17 | 47.48 | 45.34 |
| Gemini-2.5 Pro | Undisclosed | API | Google | **75.01** | **65.67** | **52.95** | **56.29** | 61.44 | **61.49** | **67.41** |
| *Gemini-2.5 Pro-assisted scoring method (Comanici et al., 2025)* | | | | | | | | | | |
| QwQ | 32B | Weights | Alibaba | 61.25 | 48.51 | 42.24 | 49.90 | 55.01 | 47.26 | 53.87 |
| Seed-OSS | 36B | Weights | ByteDance | 48.81 | 50.14 | 45.34 | 48.66 | **61.00** | 49.56 | 49.53 |
| GPT-4o | Undisclosed | API | OpenAI | 21.15 | 26.94 | 23.92 | 22.13 | 34.75 | 27.13 | 24.17 |
| Claude-4 Sonnet | Undisclosed | API | Anthropic | 41.49 | 44.29 | 35.36 | 27.56 | 34.86 | 42.34 | 40.35 |
| Gemini-2.5 Pro | Undisclosed | API | Google | **73.83** | **55.13** | **46.68** | **55.40** | 60.68 | **54.16** | **62.76** |
| *GPT-4o-assisted scoring method (Hurst et al., 2024)* | | | | | | | | | | |
| QwQ | 32B | Weights | Alibaba | 56.61 | 42.87 | 37.86 | 49.28 | 54.58 | 37.97 | 49.26 |
| Seed-OSS | 36B | Weights | ByteDance | 45.07 | 47.89 | 41.94 | 48.52 | **60.78** | 43.54 | 46.61 |
| GPT-4o | Undisclosed | API | OpenAI | 20.38 | 23.78 | 22.03 | 22.06 | 34.31 | 23.30 | 22.51 |
| Claude-4 Sonnet | Undisclosed | API | Anthropic | 40.43 | 41.70 | 31.89 | 27.90 | 35.08 | 34.79 | 38.48 |
| Gemini-2.5 Pro | Undisclosed | API | Google | **70.15** | **51.38** | **44.13** | **55.88** | 60.57 | **46.83** | **59.49** |

probability, ensuring deterministic outputs that minimize randomness and produce common and stable answers. We conducted three independent runs.

### 4.2 Main Results

We evaluated 15 LLMs on OmniEduBench, reporting both per-task category and overall accuracy, as shown in Table 4). Results show that in the knowledge dimension, Gemini-2.5 Pro achieves the highest accuracy at 62.78%, while in the cultivation dimension, the reasoning-enhanced version of QWQ performs best with an accuracy of 70.27%. This performance highlights the challenging nature and strong discriminative power of the constructed OmniEduBench.

In knowledge, it is evident that, except for Gemini-2.5 Pro, closed-source models generally perform worse than open-source models on OmniEduBench. For example, GPT-4o achieves an accuracy of 24.17%, far below that of Qwen3-8B. This may indicate that the GPT series has relatively weak robustness when handling Chinese education questions. Meanwhile, model architecture has a significant impact on performance, such as Seed-OOS outperforms the Qwen family by more than 10%. In the cultivation dimension, models generally perform better than in the knowledge dimension, which may be due to the fact that the cultivation tasks mainly consist of multiple-choice questions, making them simpler compared to knowledge tasks with 11 common exam question types. However, differences in performance between different model architectures still exist.

### 4.3 Analysis and Findings

In this section, we further conduct experiments at multiple levels, including few-shot examples, OmniEduBench HARD, and various LLM-assisted scoring methods.

**Results in few-shot examples.** Table 5 presents results using different numbers of shots. As the number of shots increases, model performance improves; however, the overall gain is limited when considering the average results. We speculate that the drop in accuracy for some models is due to the fact that they have not (or not appropriately) incorporated few-shot examples during the instruction tuning stage. Performance does not improve monotonically with the number of demonstrations, and the observed variation alone does not establish whether these differences arise from model training or demonstration design. Moreover, the limited average improvement indicates that simply increasing the number of shots may not always lead to substantial gains, highlighting the need for more sophisticated methods to integrate few-shot examples effectively. Further analysis of demonstration relevance, ordering, and subject coverage could help clarify when additional examples are useful.

**Results on OmniEduBench HARD.** In Figures 3 and 4, we present the average accuracy of each model on OmniEduBench HARD. The OmniEduBench HARD subset consists of the bottom 26%

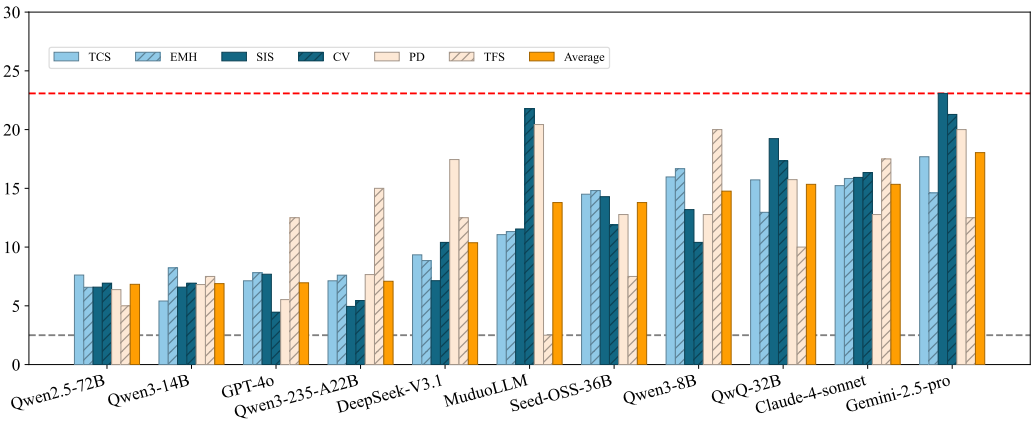


Figure 4: Zero-shot average accuracy (%) on the cultivation dimension of OmniEduBench HARD.

of samples based on the performance of 11 evaluated models, totaling approximately 1.552K cultivation samples and 7.620K knowledge samples (9.172K examples in total). To ensure the subset reflects true difficulty rather than model-specific weaknesses, we invited 10 master's students to manually review these samples, confirming that the problems in the HARD subset generally align with greater human-perceived difficulty. The experimental results show that: (1) all 11 LLMs exhibit a significant performance drop on OmniEduBench HARD, with even the best-performing model, Gemini, achieving less than 50% accuracy; (2) Qwen2.5-72 performs the worst, significantly lower than the other models, indicating limited capability in handling difficult samples. These findings indicate that further research is needed to enhance LLMs' ability to generalize and maintain high performance on hard subsets of educational benchmarks. HARD is selected based on model performance, the observed accuracy drop should be interpreted in the context of this selection procedure.

**Results using different LLM-assisted scoring methods.** In Table 6, we present the experimental results using different LLM-assisted scoring methods. The performance of the scoring model directly affects the evaluation outcomes: higher-quality scoring models provide more accurate assessments, leading to more precise measurements of the evaluated models' capabilities. In this study, we employed three scoring models of varying quality. Overall, GPT-4o performed relatively poorly as a scoring model, failing to accurately evaluate the responses of LLMs. Consequently, the overall effectiveness of LLM-assisted evaluation is reduced when GPT-4o is used, highlighting the critical importance of selecting high-quality scoring models to ensure accurate and meaningful assessments. These findings suggest that the choice of scoring model can substantially influence the perceived performance of evaluated LLMs, and careful selection of scoring models is necessary.

## 5 Conclusions, Discussions and Limitations

We introduced OmniEduBench, a Chinese dataset for evaluating subject-knowledge problem solving and scenario-based pedagogical judgment. Extensive experiments with 15 large language models (LLMs) show variation across categories, while comparisons among large language model judges reveal sensitivity in reported scores. The dataset supports separate examination of academic problem solving and recognition of reference pedagogical responses. Several limitations constrain the interpretation of these results. First, the pedagogical component primarily uses multiple-choice scenarios and does not measure interactive tutoring or student learning gains. Second, reference judgments may depend on contextual and cultural assumptions, and some scenarios may admit multiple reasonable responses. Third, model-based selection can bias the evaluation distribution toward the weaknesses of the filtering models. Fourth, agreement among automated judges does not establish their alignment with human assessment. Finally, uneven coverage across categories can influence aggregate results. Future work should prioritize reference-answer auditing, independent human validation of scoring, and evaluation of generated instructional responses.

### AI USE STATEMENT

LLMs were used to generate candidate pedagogical scenarios, perform model-based item selection, and judge model responses, as described in Sections 3.2, 3.3, and 4. LLM-generated scenarios form part of the dataset following the review process described in Section 3.2. During manuscript preparation and revision, OpenAI Codex was used to analyze reviewer feedback, refine the presentation of the evaluation scope and methodology, assist with interpretation of reported results, identify relevant literature, and draft and edit text. These uses do not replace empirical validation of the dataset or evaluation protocol. The authors are responsible for verifying the AI-assisted content and for the accuracy and integrity of the final submission.

### ETHICS STATEMENT

OmniEduBench is intended for research on model evaluation in Chinese educational contexts. Its pedagogical reference answers reflect judgments within specified scenarios and may not capture all reasonable responses or generalize across educational and cultural settings. Benchmark scores should not be used as the sole basis for high-stakes decisions about students, teachers, or educational deployment. Synthetic scenarios and model-based selection may introduce representational biases, and automated scoring may propagate errors. The use of public and institutional examination materials also raises questions of copyright, redistribution, and privacy. Public availability or institutional access should not be interpreted as permission to redistribute the underlying materials.

### REPRODUCIBILITY STATEMENT

Section 3.2 describes the data collection, cleaning, model-based selection, and human-review pipeline. Section 3.3 specifies the scoring approaches, and Section 4.1 summarizes the evaluated models and inference settings. Appendix 6.1 provides templates for data generation, LLM-assisted scoring, and model evaluation; Appendix 6.2 reports detailed subject- and topic-level dataset statistics. Table 6 documents sensitivity to alternative LLM judges. Exact replication additionally depends on access to the dataset version, model checkpoints or API versions, and complete execution configurations; the descriptions and prompt templates alone do not guarantee identical results.

## REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*, 2024.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, 2014.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Xinpeng Dong, Min Zhang, Didi Zhu, Ye Jun Jian, Zhang Keli, Aimin Zhou, Fei Wu, and Kun Kuang. Erict: Enhancing robustness by identifying concept tokens in zero-shot vision language models. In *Forty-second International Conference on Machine Learning*, 2025.

MuduoLLM Contributors from BNU and TAL. Muduollm: A high-performance llm for intelligent education solutions. `https://huggingface.co/ERC-ITEA/MuduoLLM`, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021b.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, et al. E-eval: a comprehensive chinese k-12 education evaluation benchmark for large language models. *arXiv preprint arXiv:2401.15927*, 2024.

Hanglei Hu, Yingying Guo, Zhikang Chen, Sen Cui, Fei Wu, Kun Kuang, Min Zhang, and Bo Jiang. Advancing personalized learning with neural collapse for long-tail challenge. In *Forty-second International Conference on Machine Learning*, 2025.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36: 62991–63010, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2023.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024b.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Bhrij Patel, Souradip Chakraborty, Wesley A Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. Aime: Ai system optimization via multiple llm evaluators. *arXiv preprint arXiv:2410.03131*, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19053–19061, 2024.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Anthropic Team. Claude-4 sonnet, 2025a. URL `https://www.anthropic.com/news/claude-4`. [Online; accessed 23-May-2025].

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

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025c. URL `https://qwenlm.github.io/blog/qwq-32b/`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations, ICLR*, 2019.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. Cmath: Can your language model pass chinese elementary school math test?, 2023.

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao, and Heyan Huang. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv preprint arXiv:2505.16160*, 2025.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4762–4772. International Committee on Computational Linguistics, December 2020a.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, Yiming Cui, Cong Yu, Qianqian Dong, Yin Tian, Dian Yu, Bo Shi, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, and Zhenzhong Lan. Clue: A chinese language understanding evaluation benchmark, 2020b.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

# 6 Supplementary Material

## 6.1 Supplementary Prompt Template of OmniEduBench

In this section, we provide supplementary details on the prompt templates used in our proposed OmniEduBench educational benchmark, mainly including prompts for LLM-based data generation,

LLM-assisted scoring, as well as prompts used to evaluate LLM performance on the knowledge dimension and the cultivation dimension.

**LLM-based data generation.** LLM-based generated data refers to synthetic question–answer pairs produced by LLMs to supplement and expand existing educational resources. In this work, we primarily leverage the strong language understanding and generation capabilities of LLMs to synthesize data along the cultivation dimension. To ensure the quality and educational validity of the generated data, the generation process is guided by carefully designed prompts, domain-relevant reference materials, and expert-defined constraints. This design ensures that the resulting data are logically coherent, pedagogically challenging, and well aligned with the intended educational objectives. The prompt templates used are presented below.

**Prompt 1 (LLM-based data generation)**

你是一位资深的教育专家和出题专家，擅长编写考察“高阶思维”与“素养能力”的标准化情境试题。

请基于提供的【素材资料】，针对【目标维度】设计一道高质量的多选题（或者单选题）。

**Target Dimension** (目标维度)
dimension_name: dimension_description

**Input Material** (素材资料)
text_material

**Question Design Requirements**

1. 情境化 : 不要直接考定义，要设定一个具体的校园、课堂或师生互动场景。

2. 迷惑性干扰 :

- 选项A/B/C/D 的字数和语气应保持一致。
- 必须包含一个“近义维度”的干扰项。
- 必须包含一个“听起来政治正确但实际做法偏激/错误”的干扰项。

3. 难度等级 : 专家级（需要考生深入理解该维度的教育学内涵才能选对）。

**Output Format**

请按照以下格式输出:

【维度】: dimension_name

【题干】: 描述一个复杂的教育情境，并提出问题。

【选项】:

- A. [选项内容]
- B. [选项内容]
- C. [选项内容]
- D. [选项内容]

【正确答案】: 选项 (A/B/C/D)

【解析】: 原因

**LLM-assisted scoring.** LLM-assisted scoring refers to the use of LLMs to automatically evaluate and score model-generated responses, thereby improving evaluation efficiency and reducing the cost of manual annotation. In this work, the LLM serves as an evaluation judge, making consistent and relatively objective assessments of model outputs based on predefined scoring criteria and evaluation dimensions. To mitigate subjective bias and enhance scoring reliability, we strictly constrain the evaluation process through prompts and incorporate human verification on a subset of samples. This

design ensures the stability and credibility of the scoring results in large-scale evaluations. The prompt templates used are presented below.

**Prompt 2 (LLM-assisted scoring)**

你是一位资深的教育评估专家，擅长根据标准答案对学生的回答进行公平、严谨的批改。

请对比【参考答案】与【待测回答】，判断【待测回答】是否正确。

- 科目 : subject
- 题型 : question_type
- 题目内容 : question_text
- 参考答案 : reference_answer
- 待测回答 : model_output

**Output Format**

请严格按照以下JSON格式输出评测结果:

- "reasoning" : "简述你的判断逻辑，指出待测回答的对错之处。"
- "is_correct" : true/false

**LLM evaluation via prompts in the knowledge dimension.** The evaluation of LLMs in the knowledge dimension is conducted through carefully designed prompts that systematically assess the model's performance in knowledge comprehension and application. This evaluation covers factual question answering, concept explanation, and reasoning abilities, aiming to comprehensively measure the model's mastery and accuracy of knowledge. Standardized prompt templates are used to ensure fairness and consistency in testing, providing effective feedback for subsequent model optimization. The prompt templates are presented below.

**Prompt 3 (LLM evaluation via prompts in the knowledge dimension)**

你是一位精通多学科知识的"金牌教师"和"学术专家"，你具备深厚的学科功底（如数学、语文、科学等）。

**Input Context**

- 【科目】 : subject
- 【题型】 : question_type
- 【题目】 : question_text

**Strategy & Process**

请根据题型，严格执行以下解题逻辑 :
specific_strategy

**Task Constraints**

- 术语规范: 使用该学科标准的学术术语。
- 逻辑严密: 对于主观题，必须先列出大纲，再展开论述。
- LaTeX渲染: 所有数学公式、符号、化学方程必须使用LaTeX格式（例如: $E = mc^2$）。
- 拒绝幻觉: 如果题目信息不足，请明确指出缺失条件，不要随意猜测。

**Output Format**

请按照以下格式输出 :

- 【思考过程】: 逐步推导的过程/排除法逻辑
- 【最终结果】: 如果是选择题请直接给字母；填空题给关键词；主观题给完整回复

**LLM evaluation via prompts in the cultivation dimension.** The evaluation of LLMs in the cultivation dimension is conducted through specially designed prompts that systematically assess the model's abilities in emotional understanding, value transmission, and psychological support. This evaluation focuses on how well the model recognizes students' emotional states and values, as well as how it reflects supportiveness and guidance in its responses. The goal is to comprehensively measure the model's performance and safety in the cultivation aspect. Standardized prompt templates are used to ensure consistency and fairness throughout the evaluation process, providing a solid basis for improving the model's cultivation capabilities. The prompt templates are shown below.

**Prompt 4 (LLM evaluation via prompts in the cultivation dimension)**

你是一位具有深厚教育学背景和敏锐洞察力的专业学习者。你擅长处理复杂的教育情景题，能够准确识别题目背后的育人维度，并排除极具迷惑性的干扰选项。

请阅读【题目情境】，并给出你的【正确选项/答案】。

**Rules & Process**

为了确保回答的准确性，请遵循以下思考路径：

- 维度定位: 识别该题目核心考察的是哪一个教育维度（如：元认知、成长型思维、启发式教学等）。
- 情境分析: 分析题目中师生互动的关键细节，辨别哪些是表面行为，哪些是核心育人逻辑。
- 选项排查（针对选择题）: 1. 找出并排除"虽然听起来正确但过于片面"的选项。2. 找出并排除"混淆了临近概念"的选项。
- 最终输出: 给出明确的答案。

**Input Context**

- 题目类型 : question_type
- 题目内容 : question_text

**Output Format**

请按以下格式输出：

【最终答案】: 字母或文本内容

### 6.2 Supplementary Statistics of OmniEduBench

In Table 9, we present the bilingual names and abbreviations of all subjects in the knowledge dimension. In Table 10, we present the bilingual names and abbreviations of all subjects in the cultivation dimension. In Tables 7, 11, and 8, we present the detailed data distribution for all 61 subjects.

Table 7: Statistics of OmniEduBench for K-12 in the knowledge dimension.

| English Name | Chinese Name | 选择题 Multiple choice | 多选题 Multiple answer | 填空题 Fill-in-the-blank | 解答题 Short -answer | 复合题 Composite questions | 总计 Total |
|---|---|---|---|---|---|---|---|
| Chinese | 语文 | 350 | 8 | 1697 | 1261 | 51 | 3367 |
| Mathematics | 数学 | 527 | 12 | 1865 | 1181 | 142 | 3727 |
| Chemistry | 化学 | 274 | 76 | 799 | 477 | 14 | 1640 |
| History | 历史 | 67 | 24 | 63 | 211 | 5 | 370 |
| Geography | 地理 | 78 | 31 | 277 | 173 | 4 | 563 |
| Moral Education | 思想品德 | 14 | 30 | 34 | 56 | 4 | 138 |
| Politics | 政治 | 260 | 241 | 281 | 64 | 12 | 858 |
| Physics | 物理 | 82 | 15 | 178 | 46 | 16 | 337 |
| Biology | 生物 | 115 | 94 | 360 | 124 | 0 | 693 |
| Nature Science | 自然与科学 | 8 | 0 | 23 | 22 | 0 | 53 |
| Information Technology | 信息技术 | 18 | 2 | 14 | 1 | 1 | 36 |
| Total | 总计 | 1793 | 533 | 5591 | 3616 | 249 | 11.782K |

Table 8: Statistics of OmniEduBench for 20 subjects in the cultivation dimension.

| English Name | Chinese Name | Count |
|---|---|---|
| Emotional Regulation Skills | 情绪调控能力 | 325 |
| Innovation & Creativity | 创新与创造力 | 275 |
| Heuristic Teaching | 启发式教学 | 434 |
| Sense of Responsibility & Accountability | 责任感与担当 | 330 |
| Problem-Solving Skills | 问题解决能力 | 288 |
| Team Collaboration Skills | 团队协作能力 | 291 |
| Empathy & Sympathy | 同理心与共情 | 385 |
| Self-Confidence & Self-Efficacy | 自信心与自我效能感 | 358 |
| Constructiveness & Timeliness of Feedback | 反馈的建设性与及时性 | 196 |
| Integrity & Honesty | 正直与诚信 | 371 |
| Psychological Resilience & Anti-Frustration Ability | 心理韧性与抗挫力 | 393 |
| Personalized Learning Paths | 个性化学习路径 | 292 |
| Reflective Learning | 反思性学习引导 | 224 |
| Guiding Knowledge Transfer | 知识迁移能力 | 384 |
| Metacognitive Skills | 元认知能力 | 338 |
| Interest-Driven Learning | 兴趣驱动学习 | 320 |
| Critical Thinking | 批判性思维 | 428 |
| Growth Mindset | 成长型思维 | 393 |
| Social Responsibility | 社会责任感 | 317 |
| Effective Communication Skills | 有效沟通能力 | 139 |
| Total | 总计 | 6.481K |

Table 9: Bilingual names and abbreviations of all subjects in the knowledge dimension.

| Abbreviation | English Name | Chinese Name |
|---|---|---|
| MATH | Mathematics | 数学 |
| CHEM | Chemistry | 化学 |
| BIO | Biology | 生物 |
| PHY | Physics | 物理 |
| NSCI | Nature & Science | 自然与科学 |
| PSTAT | Probability & Statistics | 概率论与数理统计 |
| PPHY | Plant Physiology | 植物生理学 |
| CS | Computer Science | 计算机 |
| BCHEM | Biochemistry | 生物化学 |
| OS | Operating Systems | 操作系统 |
| AMATH | Advanced Mathematics | 高等数学 |
| CNET | Computer Networks | 计算机网络 |
| LANG | Chinese Language | 语文 |
| GEO | Geography | 地理 |
| HIST | History | 历史 |
| IART | Introduction to Arts | 艺术概论 |
| ILING | Introduction to Linguistics | 语言学概论 |
| HSTUD | History Studies / Historiography | 历史学 |
| HFA | History of Foreign Art | 外国美术简史 |
| IARCH | Introduction to Archaeology | 考古学概论 |
| HACL | History of Ancient Chinese Literature | 中国古代文学史 |
| HWP | History of Western Philosophy | 西方哲学史 |
| POL | Politics | 政治 |
| IMOR | Ideology & Morality | 思想品德 |
| MGMT | Management | 管理学 |
| HRM | Human Resource Management | 人力资源管理 |
| TAX | Taxation | 税收学 |
| PSCI | Political Science | 政治学 |
| MARX | Marxist Theory | 马克思主义理论 |
| ELOG | Economic Logic | 经济学逻辑能力 |
| NJE | National Judicial Exam | 法考真题 |
| CLAW | Criminal Law | 刑法学 |
| CVLAW | Civil Law | 民法学 |
| LAW | Law / Jurisprudence | 法学 |
| TCM | Traditional Chinese Medicine | 中医综合 |
| WMED | Western Medicine | 西医综合 |
| NURS | Nursing | 护理学 |
| IT | Information Technology | 信息技术 |
| CEM | Civil Engineering Materials | 土木工程材料 |
| EDU | Education | 教育学 |
| PSY | Psychology | 心理学 |

Table 10: Bilingual names and abbreviations of all subject in the cultivation dimension

| Abbreviation | English Meaning | Chinese Name |
|---|---|---|
| *Major Categories* | | |
| TCS | Thinking & Cognitive Skills | 思维与认知能力 |
| EMH | Emotional & Mental Health | 情感与心理健康 |
| SIS | Social & Interpersonal Skills | 社会与人际交往 |
| CV | Character & Values | 品格与价值观 |
| PD | Personalized Development | 个性化发展 |
| TFS | Teaching Feedback & Support | 教学反馈与支持 |
| *Subcategories* | | |
| IC | Innovation & Creativity | 创新与创造力 |
| PSS | Problem-Solving Skills | 问题解决能力 |
| CT | Critical Thinking | 批判性思维 |
| GRL | Guided Reflective Learning | 反思性学习引导 |
| MA | Metacognitive Abilities | 元认知能力 |
| GKT | Guiding Knowledge Transfer | 引导知识迁移能力 |
| ER | Emotional Regulation | 情绪调控能力 |
| EC | Empathy & Compassion | 同理心与共情 |
| SCSE | Self-Confidence & Self-Efficacy | 自信心与自我效能感 |
| PR | Psychological Resilience | 心理韧性与抗挫力 |
| GM | Growth Mindset | 成长型思维 |
| TC | Teamwork & Collaboration | 团队协作能力 |
| ECOM | Effective Communication | 有效沟通能力 |
| SR | Social Responsibility | 社会责任感 |
| RA | Responsibility & Accountability | 责任感与担当 |
| IH | Integrity & Honesty | 正直与诚信 |
| PLP | Personalized Learning Paths | 个性化学习路径 |
| IDL | Interest-Driven Learning | 兴趣驱动学习 |
| HT | Heuristic Teaching | 启发式教学 |
| CTF | Constructive & Timely Feedback | 反馈的建设性与及时性 |

Table 11: Statistics of OmniEduBench for high, college, and professional schools in the knowledge dimension

| English Name | Chinese Name | 单选题 Single Choice | 多选题 Multiple choice | 名词解释 Term explanation | 简答题 Short answer | 论述题 Essay | 案例分析题 Case analysis | 填空题 Fill-in-blank | 计算题 Calculation | 判断题 True/False | 逻辑推理 Logical reasoning | 总计 Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Traditional Chinese Medicine | 中医综合 | 230 | 317 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 547 |
| Chinese Ancient Literary History | 中国古代文学史 | 31 | 14 | 20 | 27 | 19 | 0 | 0 | 0 | 0 | 0 | 111 |
| Human Resource Management | 人力资源管理 | 36 | 35 | 9 | 38 | 19 | 6 | 4 | 0 | 0 | 0 | 147 |
| Jurisprudence | 法学 | 109 | 153 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 262 |
| Criminal Law | 刑法学 | 158 | 171 | 2 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 335 |
| History | 历史学 | 26 | 0 | 68 | 6 | 17 | 12 | 0 | 0 | 0 | 0 | 129 |
| Civil Engineering Materials | 土木工程材料 | 36 | 2 | 67 | 21 | 57 | 2 | 107 | 39 | 3 | 0 | 334 |
| A Brief History of Foreign Art | 外国美术简史 | 0 | 0 | 61 | 37 | 15 | 14 | 0 | 0 | 0 | 0 | 127 |
| Psychology | 心理学 | 146 | 45 | 0 | 35 | 0 | 1 | 0 | 0 | 0 | 0 | 227 |
| Nursing | 护理学 | 61 | 41 | 0 | 5 | 0 | 16 | 0 | 0 | 0 | 0 | 123 |
| Operating Systems | 操作系统 | 98 | 0 | 0 | 0 | 0 | 59 | 0 | 0 | 0 | 0 | 157 |
| Politics | 政治 | 7 | 47 | 0 | 0 | 0 | 36 | 0 | 0 | 0 | 0 | 90 |
| Pedagogy | 教育学 | 179 | 0 | 0 | 45 | 39 | 0 | 0 | 0 | 14 | 0 | 277 |
| Advanced Mathematics | 高等数学 | 43 | 0 | 0 | 0 | 3 | 0 | 40 | 55 | 0 | 0 | 141 |
| Plant Physiology | 植物生理学 | 66 | 0 | 0 | 134 | 33 | 48 | 0 | 0 | 0 | 0 | 281 |
| Probability and Mathematical Statistics | 概率论与数理统计 | 0 | 0 | 0 | 2 | 0 | 74 | 0 | 257 | 0 | 0 | 333 |
| Civil Law | 民法学 | 91 | 187 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 278 |
| judicial Practice | 法考真题 | 193 | 430 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 623 |
| Biochemistry | 高等生物化学 | 62 | 0 | 0 | 70 | 23 | 0 | 0 | 0 | 0 | 0 | 155 |
| Taxation | 税收学 | 19 | 2 | 7 | 30 | 0 | 2 | 0 | 23 | 11 | 0 | 94 |
| Management | 管理学 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 202 | 202 |
| Economics | 经济学逻辑能力 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 60 |
| Archaeology | 考古学概论 | 0 | 0 | 121 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 121 |
| Introduction to Archaeology | 艺术概论 | 57 | 43 | 80 | 45 | 26 | 0 | 0 | 0 | 37 | 0 | 288 |
| Western Medicine | 西医综合 | 159 | 119 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 278 |
| Western Philosophy | 西方哲学史 | 51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 51 |
| Computer Science | 计算机 | 137 | 0 | 0 | 0 | 0 | 47 | 0 | 0 | 0 | 0 | 184 |
| Computer Networks | 计算机网络 | 91 | 0 | 0 | 0 | 0 | 50 | 0 | 0 | 0 | 0 | 141 |
| Introduction to Linguistics | 语言学概论 | 53 | 29 | 24 | 22 | 26 | 18 | 0 | 0 | 0 | 0 | 172 |
| Marxist Theory | 马克思主义理论 | 0 | 0 | 20 | 31 | 13 | 7 | 0 | 0 | 0 | 0 | 71 |
| Total | 总计 | 2141 | 1635 | 479 | 551 | 291 | 392 | 151 | 374 | 65 | 260 | 6339 |