EduBench: A Comprehensive Benchmarking Dataset for Evaluating Large Language Models in Diverse Educational Scenarios

Bin Xu*, **Yu Bai***, **Huashan Sun***, **Yiguan Lin***
Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao† **, Heyan Huang**
School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China
{binxu, yubai, gyang}@bit.edu.cn

Abstract

As large language models continue to advance, their application in educational contexts remains underexplored and under-optimized. In this paper, we address this gap by introducing the first diverse benchmark tailored for educational scenarios, incorporating synthetic data containing 9 major scenarios and over 4,000 distinct educational contexts. To enable comprehensive assessment, we propose a set of multi-dimensional evaluation metrics that cover 12 critical aspects relevant to both teachers and students. We further apply human annotation to ensure the effectiveness of the model-generated evaluation responses. Additionally, we succeed to train a relatively small-scale model on our constructed dataset and demonstrate that it can achieve performance comparable to state-of-the-art large models (e.g., Deepseek V3, Qwen Max) on the test set. Overall, this work provides a practical foundation for the development and evaluation of education-oriented language models. Code and data are released at `https://github.com/ybai-nlp/EduBench`.

1 Introduction

Large Language Models (LLMs) have recently shown remarkable potential in educational contexts, offering capabilities such as problem-solving, interactive dialogue, and decision-making (Jia et al., 2021; Rouzegar and Makrehchi, 2024). These abilities make LLMs promising tools for tasks ranging from personalized tutoring to educational content generation. However, despite growing interest, research on their practical deployment in education remains limited.

A key limitation of previous work is their narrow focus on knowledge-intensive tasks (Rein et al., 2023; Team et al., 2025; Cobbe et al., 2021), which fails to reflect the diverse educational scenarios encountered in real-world settings. These efforts often overlook the complexity introduced by varying roles (e.g., professors, students, psychological counselors), whose needs, goals, and interaction styles differ significantly. More importantly, existing benchmarks (Huang et al., 2024b; Koutcheme et al., 2024b; Yang et al., 2024b) rarely align **task interactions with learners' cognitive levels and scenario-specific objectives**, leading to evaluation dimensions that lack pedagogical relevance. In contrast, our work introduces **EduBench**, a benchmark designed not only to support educational applications, but to promote the development of robust and goal-aligned evaluation mechanisms that reflect the diversity of modern educational needs. By going beyond purely knowledge-based tasks (Koutcheme et al., 2024a; Ng and Fung, 2024; Huang et al., 2024a), the EduBench re-centers educational AI research on the core values of education—holistic development, personalized support, and context-aware learning—while systematically constructing data that captures the rich landscape of roles, domains, and learning scenarios.

For constructing the data, we create a diverse dataset considering 9 different education scenarios, including assignment judging, proposing a study plan given the profile of specific students, suggesting for psychological health, and etc. We design several educational contexts in each scenario to further facilitate the diversity of the data, such as question difficulties, student grades (e.g., elementary school students, high school students, graduate students, etc.), and different subjects. All the above categories formulate different education contexts with a total number of over 4,000. For these different contexts, we create different querying data, resulting in a dataset containing 18,821 data points.

In designing our evaluation metrics, we focus on three principal aspects. First, **Scenario Adaptation** determines if the model's response appropriately addresses the conditions and constraints defined

*Equal contribution.
†Corresponding author.

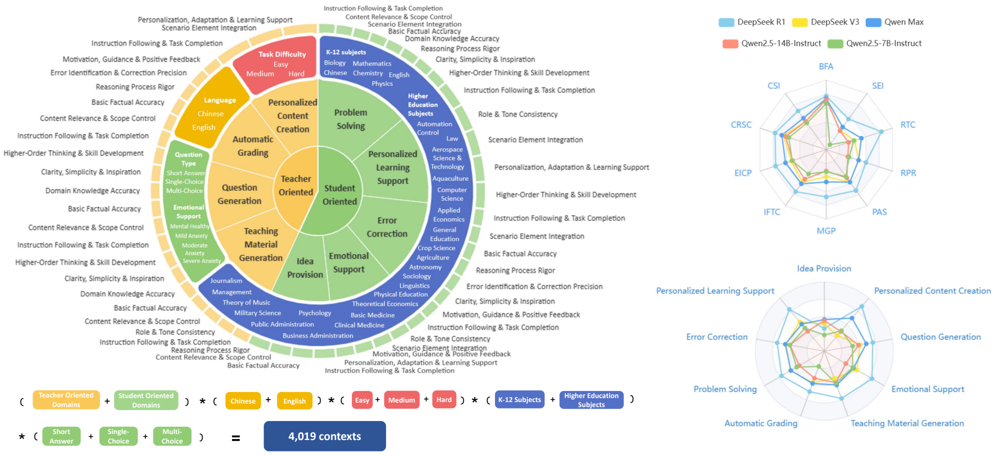

Figure 1: The left section presents our 9 educational scenarios, along with their multi-dimensional educational contexts and corresponding metrics. The right section illustrates the results from human evaluation on EduBench.

by specific educational scenarios, task instructions, and role-playing requirements. Second, **Factual & Reasoning Accuracy** scrutinizes both the factual correctness of the information presented and the logical soundness of the reasoning applied. Third, **Pedagogical Application** examines the response's adherence to established educational principles and its potential to positively impact students' learning experiences. Each major aspect contains 4 distinct sub-metrics to evaluate the responses in a finer-grained way. These metrics are systematically allocated across nine educational scenarios to ensure comprehensive coverage of diverse task demands.

When evaluating the responses of different models, we notice a trade-off between accuracy and cost when comparing human and LLM evaluators. Hence, we chose to investigate the evaluation capabilities of different LLMs by comparing their evaluative responses with those of human judges in a small test set consisting of 198 diverse data points. Our experiments on several LLMs reveal that DeepSeek V3 achieves the best alignment with human annotator. Results of using it to benchmark five different LLMs show that: 1. The models' understanding of the scoring guidelines remains imperfect. 2. Not all models as evaluator align closely with human ratings; DeepSeek V3 stands out for its consistency, whereas GPT-4o performs relatively weaker. 3. Smaller models in general would perform worse than large models.

As smaller language models typically underperform significantly larger models, we further highlight the utility of our data in boosting the performance of smaller models. This is achieved through knowledge distillation from powerful larger models, aiming to bridge the performance gap. We implement a multi-source distillation strategy, extracting expertise for each scenario from the LLMs with the highest performance on that scenario. Our results demonstrate that a 7B model can attain performance comparable to the state-of-the-art 671B Deepseek V3 model using a dataset size of around 17,000 training samples.

We summarize our contribution as follows:

1. We present the first LLM-powered educational benchmark with the largest scalable scenario and context collection (4,000+ contexts), along with a 12-dimension evaluation system.

2. We establish a series of findings and will release all the model-generated and human-annotated data that could benefit the LLM research community regarding both educational applications and LLM-based evaluations.

3. We show that our data could benefit smaller models to achieve comparable performance with powerful state-of-the-art LLMs.

2 Related Work

2.1 Evaluation Benchmarks of LLMs

The evolution of evaluation frameworks has been pivotal in benchmarking general LLM capabilities, yet they remain limited in address-

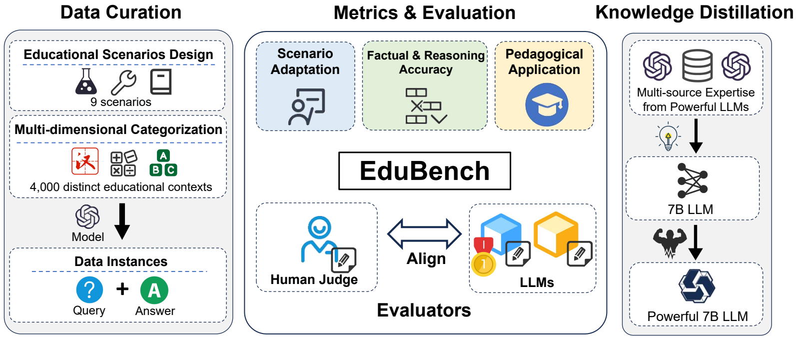

Figure 2: An overview of our EduBench. The left part illustrates our data curation procedure. In the middle part, we showcase demonstrations of our three main evaluation principles and our investigation into the alignment of LLMs with the human judge. The right part shows how our data can boost the performance of smaller models.

ing domain-specific challenges, especially in education. Early benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Sarlin et al., 2020) established standardized linguistic tasks, while MMLU (Hendrycks et al., 2021) extended evaluations to cross-disciplinary knowledge. Building upon these, recent initiatives such as AlpacaEval (Dubois et al., 2024) quantify instruction-following fidelity via pairwise comparisons.

Despite these advancements, mainstream benchmarks still emphasize general capabilities over educational applicability. This gap has spurred specialized benchmarks that target educational scenarios, although significant limitations persist. For instance, EduNLP (Huang et al., 2024b) offers a modular framework spanning 8 disciplines and 5 downstream tasks, but its pedagogical relevance is limited by narrow scenario diversity and reliance on generic metrics. Code repair has emerged as a key educational application of LLMs, particularly in programming pedagogy. Koutcheme et al. (2024a) benchmarks LLMs' abilities in both code repair and explaining errors in natural language across varied contexts and error types. Building on this, Koutcheme et al. (2024b) introduces two realistic code repair scenarios—classroom tasks and open-source contributions—and builds multi-institutional datasets reflecting distinct coding norms and error patterns. In academic Q&A assessment, persistent limitations also appear. GPQA (Rein et al., 2023) evaluates graduate-level reasoning across three disciplines, while SuperGPQA (Team et al., 2025) broadens this to 285 fields but maintains a rigid format that hampers interactive pedagogy. Language learning adaptations (Huang et al., 2024a) further examine LLMs' potential to simplify educational texts, using textbook-based content to enhance readability.

Despite progress in general and domain-specific benchmarks, their application to education remains limited by narrow scenarios—typically single-subject and fact-based Q&A—overlooking authentic contexts like adaptive feedback and collaborative learning.

2.2 LLMs for Education Applications

With strong problem solving, interactive, and sound judgment capabilities, LLMs hold significant promise for education, prompting extensive research into their use across diverse scenarios and contexts.

For example, LLMs have been extensively used to generate educational content (Jia et al., 2021; Ghanem et al., 2022) in subjects across mathematics (Liyanage and Ranathunga, 2019; Rouzegar and Makrehchi, 2024), computer science (Logacheva et al., 2024; Sarsa et al., 2022; Frankford et al., 2024a), medicine (Berman et al., 2024) and so on. These models enable efficient large-scale content creation while preserving scenario specificity and linguistic coherence, making them valuable tools for instructional design.

In parallel, researchers have increasingly employed LLMs as interactive tutors. Such systems guide students through programming concepts (Vadaparty et al., 2024; Frankford et al., 2024b), assist with code debugging and explanation (MacNeil et al., 2023; Wang et al., 2024c; Yang et al., 2024a), support the acquisition of mathematical skills (Pardos and Bhandari, 2023), help simulate classroom learning processes (Zhang et al., 2025), and even train social communication strategies (Yang et al., 2024b). Moreover, in the context of teacher development, some efforts have used

LLMs to simulate student behavior or provide instructional feedback (Markel et al., 2023), thus improving teacher pedagogical strategies through AI-driven interaction.

Another active line of work involves the use of LLMs as automated assessment tools. These models have been applied to evaluate essays (Cao et al., 2020; Yang et al., 2020; Kim and Kim, 2024) or mathematical answers (Jiang et al., 2024; Urrutia and Araya, 2023), offering scoring and diagnostic feedback. For example, previous studies have used LLMs to classify algebra errors (Heickal and Lan, 2024) or identify logical errors in student code (McNichols et al., 2023), providing timely and detailed responses that can support self-directed learning.

Despite growing interest in LLMs for education, current implementations often rely on surface-level metrics like ROUGE or accuracy, overlooking essential pedagogical aspects such as scaffolding, engagement, and long-term learning outcomes. Moreover, most studies address isolated tasks—e.g., quiz generation or one-off tutoring—without considering the full instructional cycle.

3 Dataset Collection

In this section, we describe the dataset construction methodology used in this paper. Our objective is to create a comprehensive benchmark that reflects realistic and diverse scenarios in the educational domain. The dataset covers both students' and teachers' needs across multiple functional capacities, incorporating varied levels of difficulty and content modalities.

3.1 Scenario Design

We categorize and organize a set of key education-related tasks (Gan et al., 2023; Wang et al., 2024b), each reflecting a distinct dimension of education. Based on these tasks, we define a "scenario" as a representative educational scenario that involves typical user roles, cognitive skill demands, input-output formats, and evaluation criteria. Building on the distinction in typical user roles, we categorize these tasks into Student-Oriented Scenarios (Ng and Fung, 2024; Team et al., 2025; Koutcheme et al., 2024a) and Teacher-Oriented Scenarios (Zhang et al., 2025; Ghanem et al., 2022; Huang et al., 2024a). This conceptualization enables us to structure a wide range of tasks grounded in real-world educational practice.

To support both learner assistance and instructional tasks, we categorize domains into two primary types.

Student-Oriented Scenarios: Problem Solving (Q&A), Error Correction (EC), Idea Provision (IP), Personalized Learning Support (PLS), Emotional Support (ES).

Teacher-Oriented Scenarios: Question Generation (QG), Automatic Grading (AG), Teaching Material Generation (TMG), Personalized Content Creation (PCC).

These scenarios simulate key educational roles and scenarios, enabling realistic evaluation of LLM capabilities in classroom, tutoring, and curriculum design settings. A detailed breakdown of each scenario is provided in Appendix A.

3.2 Educational Context Design

To ensure scenario realism, learner alignment, and meaningful evaluation, we construct diverse educational domain contexts that reflect the conditions under which tasks naturally occur. Each context is designed to capture variation across four primary dimensions: **subject taxonomy**, **task difficulty**, **language setting**, and **question type**. This diversity enables EduBench to cover a wide range of real-world educational scenarios, from K–12 to postgraduate levels, and from basic recall to complex reasoning.

Following disciplinary category design of EduNLP (Huang et al., 2024b) and SuperGPQA (Team et al., 2025), we categorize tasks based on educational stages (e.g., K–12 vs. higher education) and align them with different cognitive and pedagogical goals. Each task is assigned a difficulty level (easy, medium, hard) to reflect expected learner proficiency. EduBench currently supports both **Chinese and English** tasks to encourage multilingual model evaluation. Question types are specified based on domain functionality: most scenarios adopt common types such as *Single Choice*, *Multiple Choice*, and *Short Answer*, while specialized scenarios follow scenario-specific definitions. Details of all the educational context design can be found in Appendix B.

3.3 Question Generation for Scenario Tasks

Building on the structured scenario design, we generate benchmark data using a systematic and scalable pipeline tailored to educational scenarios. First, we organize our coverage across nine major scenarios grounded in an educational competency framework (see Appendix A). Within each scenario, we apply multi-dimensional categorization, including subject taxonomy, task difficulty, and question type, to define fine-grained task scenarios.

For each scenario, we design prompt templates that reflect realistic user intents, then use GPT-4o to generate consistent data instances. For example, in the mathematics domain, a scenario like "Middle School – Short Answer Question" guides the model in producing relevant QA pairs. Further prompt design details and generated samples are available in Appendix E.

4 Evaluation Metric Design

Evaluating complex domains such as education is challenging due to multiple participants and multidimensionality. Previous studies have employed large language models (LLMs) for evaluation purposes (Koutcheme et al., 2024b; Team et al., 2025; Yang et al., 2024b; Wang et al., 2024a; Ng and Fung, 2024). Building on this foundation, we proposed a newly designed, structured set of evaluation metrics that include 12 dimensions and reflect both pedagogical goals and scenario-specific expectations to enhance the accuracy and interpretability of model evaluation. Our design comprises three core dimensions: Scenario Adaptation, Factual & Reasoning Accuracy, Pedagogical Application.

4.1 Overall Design

To enhance the accuracy and interpretability of model evaluation, we design a series of evaluation metrics. By incorporating them into the evaluation prompts, we guide the model to score according to the given metrics and provide detailed justifications during the evaluation process. The approach of introducing metrics or principles in evaluation has been widely validated (Liu et al., 2025; Sharma et al., 2025; Bai et al., 2022), with some being manually designed and others generated by the model before or during the evaluation. The formal expression of this process is as follows:

$$\text{Score} = \frac{1}{n}\sum_{i=1}^{n} \text{Evaluator}_i(x, y, \mathcal{M}) \quad (1)$$
$$\text{where } \mathcal{M} \in \{\text{Handcrafted}, \text{Model generated}\}$$

4.2 Metric Design

We design distinct evaluation metrics based on 3 scenarios, with 4 metrics for each scenario to cover its key aspects, resulting in 12 different metrics.

Scenario Adaptation The Scenario Adaptation metric evaluates whether model responses are contextually appropriate and aligned with educational expectations. It is assessed across four dimensions: 1) *Instruction Following & Task Completion*, 2) *Role & Tone Consistency*, 3) *Content Relevance & Scope Control*, 4) *Scenario Element Integration*.

Factual & Reasoning Accuracy The Factual & Reasoning Accuracy metric assesses the correctness of information and the rigor of reasoning in model responses. It includes four sub-metrics: 1) *Basic Factual Accuracy*, 2) *Domain Knowledge Accuracy*, 3) *Reasoning Process Rigor*, 4) *Error Identification & Correction Precision*.

Pedagogical Application This metric evaluates whether responses embody sound educational principles and effectively support student learning. It consists of the following sub-metrics: 1) *Clarity, Simplicity & Inspiration*, 2) *Motivation, Guidance & Positive Feedback*, 3) *Personalization, Adaptation & Learning Support*, 4) *Higher-Order Thinking & Skill Development*.

The detailed explanations of these metrics are provided in Appendix C.

4.3 Dynamic Metric Allocation

Given the diversity of educational tasks covered in EduBench, a one-size-fits-all evaluation approach is insufficient. Not all metrics are equally relevant or applicable across the nine distinct scenarios. For example, Emotional Support tasks emphasize contextual empathy and scenario alignment more than factual precision, while scenarios like Problem Solving and Question Generation rely heavily on rigorous reasoning and factual correctness.

To address this, we design a flexible evaluation framework that dynamically allocates appropriate metrics based on the specific requirements of each scenario. Each scenario is associated with a tailored subset of evaluation dimensions that best reflect its instructional goals, content characteristics, and target outcomes. This ensures both fairness and relevance in the evaluation process. Detailed metric-scenario mappings and allocation rules are provided in Appendix C.4.

4.4 Human-guided LLM-based Evaluation

Evaluating open-ended educational tasks is challenging due to the subjectivity and complexity of responses. While human annotation offers high-quality judgments, it is costly and difficult to scale.

On the other hand, relying solely on LLM-based evaluation raises concerns about reliability and consistency across diverse scenarios.

We adopt a human-guided evaluation framework by constructing a high-quality test set of 198 diverse examples (11 per scenario in both English and Chinese), annotated by an expert judge across various task types. This dataset serves as a benchmark to assess the alignment of different LLMs with human evaluation standards. This approach supports scalable and efficient evaluation aligned with human standards.

5 Experiments

5.1 Experiment Settings

Response generation We selected 5 representative models: DeepSeek R1 (Guo et al., 2025), DeepSeek V3 (Liu et al., 2024), Qwen Max, Qwen2.5-14B-Instruct, and Qwen2.5-7B-Instruct (Qwen, 2024). This selection provides a broad view of how models of varying sizes and types, such as standard and reasoning models, handle educational tasks.

Response evaluation We selected the following LLMs: QwQ Plus (Qwen, 2025), GPT-4o (OpenAI et al., 2024), DeepSeek R1 (Guo et al., 2025), and DeepSeek V3 (Liu et al., 2024), as evaluators due to their strong scenario understanding, broad knowledge, and accurate intent recognition. These models assess responses using our defined metrics, guided by dedicated prompts that include language-specific descriptions to avoid multilingual bias. Prompt details are provided in Appendix D.

Test set We use 198 test samples (99 Chinese, 99 English), comprising questions from Section 3.3 and responses from five models. For data selection of the test set, we sample all the data points from different educational sub-context to ensure the diversity and comprehensiveness of our evaluation.

We present the evaluation results from the best-performing evaluator model: Deepseek V3, while the complete results of model generation and evaluation are detailed in Appendix J.

5.2 Evaluation Details

As mentioned in Section 4.4, to ensure the rationality and verifiability of the evaluation, we employ both model-based and human-based point-wise evaluations to assess the responses from different models. Specifically, each QA pair is evaluated separately by each evaluation model and by one human annotator. All evaluations are based on the 12 metrics that we have defined in Section 4.

During model evaluation, metric information is embedded into the prompt, requiring models to output individual metric scores in a single response. In human evaluation, the annotator studies the metrics in advance and adheres strictly to the criteria during annotation. We adopt a point-wise evaluation strategy, as preliminary experiments reveal significant positional bias in pair-wise settings (Appendix I.3). The scoring guidelines are detailed in Appendix F.

5.3 Experiment Results

Model-evaluation results DeepSeek R1 demonstrates the best overall performance across different metrics, while Qwen2.5-7B-Instruct performs the worst in Table 1. Moreover, DeepSeek R1 performs the best on Higher-Order Thinking & Skill Development, and Qwen2.5-7B-Instruct is the least satisfactory in Error Identification & Correction Precision, with both models showing a clear gap compared to others. In specific scenarios, DeepSeek R1 remains the best, while Qwen2.5-7B-Instruct outperforms Qwen2.5-14B-Instruct in scenarios like Emotional Support and Personalized Content Creation in Table 3. This shows the gap between models with smaller sizes is not very large and it drives us to choose the 7B model as the student model during our distillation experiments in Section 6.

Human-evaluation results DeepSeek R1 and Qwen2.5-7B-Instruct still demonstrate the best and worst performance respectively in Table 1, consistent with the results from model-based evaluation. Unlike model evaluation, human annotator shows noticeably lower satisfaction with the performance of all five models on the Reasoning Process Rigor metric. Qwen2.5-7B-Instruct performs particularly poorly on this metric, scoring only 5.90. In contrast, DeepSeek R1 shows consistently strong performance on the Motivation, Guidance & Positive Feedback metric, even when other models fall short. At the scenario level, DeepSeek R1 remains far ahead in Table 3, while the performance gap between 7B and 14B of Qwen models is relatively small, making the 7B model a cost-effective choice in resource-constrained settings.

Evaluator	Model	BFA	CSI	CRSC	DKA	EICP	HOTS	IFTC	MGP	PAS	RPR	RTC	SEI	Average
DeepSeek V3	DeepSeek R1	9.51	**8.75**	**9.44**	**9.45**	**7.61**	**8.53**	**9.47**	7.76	**9.64**	**8.85**	**9.14**	**9.06**	**8.93**
	DeepSeek V3	**9.57**	8.61	9.25	9.27	7.23	7.98	9.21	7.56	8.94	8.76	9.00	8.59	8.66
	Qwen Max	9.38	8.53	9.12	9.23	7.43	7.99	9.16	**7.85**	9.05	8.57	9.00	8.61	8.66
	Qwen2.5-14B-Instruct	9.28	8.50	9.03	9.14	7.14	7.81	8.94	7.55	8.71	8.35	8.82	8.25	8.46
	Qwen2.5-7B-Instruct	9.27	8.55	9.08	9.12	6.77	7.86	8.96	7.05	8.95	8.42	8.82	8.53	8.44
Human	DeepSeek R1	**8.97**	**8.60**	**8.98**	**8.94**	**8.86**	**8.56**	**8.77**	**8.20**	**9.26**	**7.95**	**8.91**	**8.92**	**8.74**
	DeepSeek V3	8.77	7.77	8.40	7.89	8.11	7.25	8.10	7.70	7.42	7.03	7.80	7.47	7.89
	Qwen Max	8.81	8.01	8.52	8.27	8.23	7.59	8.10	7.70	7.89	7.31	8.09	7.74	8.02
	Qwen2.5-14B-Instruct	8.74	7.76	8.26	7.79	7.86	6.88	7.77	6.97	7.02	7.01	7.59	7.03	7.56
	Qwen2.5-7B-Instruct	8.49	7.63	8.04	7.82	7.45	6.93	7.65	7.05	7.38	5.90	7.82	7.35	7.46

Table 1: Metric-Level average scores evaluated by DeepSeek V3 and human evaluators under various metrics. For simplicity, we use abbreviations for the metrics. Full names of each metric can be found in Table 2.

Abbreviation	Full Name
IFTC	Instruction Following & Task Completion
RTC	Role & Tone Consistency
CRSC	Content Relevance & Scope Control
SEI	Scenario Element Integration
BFA	Basic Factual Accuracy
DKA	Domain Knowledge Accuracy
RPR	Reasoning Process Rigor
EICP	Error Identification & Correction Precision
CSI	Clarity, Simplicity & Inspiration
MGP	Motivation, Guidance & Positive Feedback
PAS	Personalization, Adaptation & Learning Support
HOTS	Higher-Order Thinking & Skill Development

Table 2: The abbreviations for all of our sub-metrics

5.4 Analysis

5.4.1 Consistency Results

Consistency between evaluation models The models exhibit a high degree of consistency in Table 5, with Kendall's W values for nearly all models above 0.5, most around 0.6, indicating strong consistency. DeepSeek V3 shows the highest consistency with other models, and its ranking of average scores for the response models aligns closely with those of other models.

Consistency between human and models The overall rankings of the generation models from both human evaluation and model evaluation show similar trends. We also evaluated the Kendall's Coefficient of Concordance between different evaluators in Table 5. The evaluation scores from models do not align precisely with human judgments which may be attributed to their limited evaluation metric understanding. Overall, DeepSeek V3 exhibits the highest correlation with human evaluations, while GPT-4o shows the lowest. This pattern may be attributed to the relatively larger model sizes and broad training data distribution of DeepSeek V3.

5.4.2 Model Behavior Analysis

Model evaluations tend to assign higher scores than human annotator. Our results demonstrate that models assign scores approximately one point higher than human's at both metric and scenario levels. In the Q&A scenario, they assign scores above 9 versus human scores of 6–7, creating an almost two-point discrepancy that highlights a significant divergence in evaluation standards. We attribute this gap to two factors: 1) Models may misinterpret scoring criteria, which could be improved through post-training, as current evaluators are not reward models. 2) RLHF training makes models reluctant to give negative feedback. However, we believe post-training could mitigate this issue, and we'll explore this direction in future work.

Larger models typically outperform smaller ones across scenarios. Top models like DeepSeek R1 excel overall, while smaller ones (e.g., Qwen2.5-7B-Instruct) only succeed in limited tasks and lag in complex metrics (Domain Knowledge Accuracy, etc). Notably, for smaller models, size is not definitive – the 7B Qwen2.5 occasionally surpasses its 14B version.

6 Multi-source Distillation

Data selection based on EduBench To fully leverage the strengths of different response generation models across various scenario, we adopt a multi-source distillation pipeline. For each task, we select the best-performing model on the test set as the response generator, using it to answer educational domain questions and construct the training dataset for the distillation model, details are shown in Appendix K. Through the distillation pipeline, we obtain a training set of 17,000 samples covering various subtasks across all 9 educational scenarios.

Results As shown in Table 4, after distillation, the performance of the 7B model significantly improved across 10 of the 12 metrics, achieving performance comparable to that of state-of-the-art models. Notably, it outperforms all other models, including DeepSeek R1 and Qwen Max, in terms of the Reasoning Process Rigor metric.

Evaluator	Model	Q&A	PLS	EC	IP	AG	TMG	ES	QG	PCC	Average
DeepSeek V3	DeepSeek R1	9.49	**9.65**	**9.27**	**8.75**	**7.27**	**9.45**	**9.38**	**9.33**	**9.71**	**9.14**
	DeepSeek V3	**9.68**	9.04	9.14	8.53	7.05	9.34	9.00	9.06	8.92	8.86
	Qwen Max	9.18	8.88	9.06	8.52	7.23	9.24	9.04	9.05	9.29	8.83
	Qwen2.5-14B-Instruct	9.07	8.72	8.97	8.30	6.77	9.21	8.74	9.02	8.80	8.62
	Qwen2.5-7B-Instruct	9.15	9.07	9.01	8.47	6.44	9.21	8.85	8.69	9.00	8.65
Human	DeepSeek R1	7.17	**9.11**	**8.71**	**8.80**	**8.42**	**8.86**	**9.15**	**8.79**	**9.35**	**8.71**
	DeepSeek V3	7.45	8.12	8.16	8.17	7.84	7.56	8.08	8.01	7.03	7.82
	Qwen Max	**7.72**	7.94	8.21	8.15	7.89	7.99	7.85	8.39	8.42	8.06
	Qwen2.5-14B-Instruct	7.66	7.38	7.92	7.56	7.55	7.84	7.31	7.91	7.36	7.61
	Qwen2.5-7B-Instruct	6.78	7.63	7.93	7.74	6.79	7.86	7.79	7.55	7.42	7.50

Table 3: Scenario-Level average scores evaluated by DeepSeek V3 and human evaluator. Max values in each column per evaluator are bolded. Full names of each scenarios can be found in Section 3.1.

Model	BFA	CSI	CRSC	DKA	EICP	HOTS	IFTC	MGP	PAS	RPR	RTC	SEI	Average
DeepSeek R1	<u>9.51</u>	**8.75**	**9.44**	**9.45**	**7.61**	**8.53**	**9.47**	<u>7.76</u>	**9.64**	<u>8.85</u>	**9.14**	**9.06**	**8.93**
DeepSeek V3	**9.57**	<u>8.61</u>	9.25	<u>9.27</u>	7.23	7.98	9.21	7.56	8.94	8.76	9.00	8.59	8.66
Qwen Max	9.38	8.53	9.12	9.23	<u>7.43</u>	7.99	9.16	**7.85**	9.05	8.57	9.00	8.61	8.66
Qwen2.5-14B-Instruct	9.28	8.50	9.03	9.14	7.14	7.81	8.94	7.55	8.71	8.35	8.82	8.25	8.46
Qwen2.5-7B-Instruct	9.27	8.55	9.08	9.12	6.77	7.86	8.96	7.05	8.95	8.42	8.82	8.53	8.44
Distillation Qwen2.5-7B	9.26	8.56	<u>9.27</u>	8.95	6.89	<u>8.43</u>	<u>9.41</u>	7.32	<u>9.56</u>	**9.26**	<u>9.09</u>	<u>8.95</u>	<u>8.75</u>

Table 4: Performance comparison of our distillation model with other models based on metric-level evaluations by DeepSeek V3. Best results are bold while second best results are underlined. For our distillation model, we use Qwen2.5-7B-Instruct as the base model. For simplicity, we use abbreviations for the metrics. Full names of each metric can be found in Table 2.

Model	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human
DeepSeek R1	-	0.55	0.61	**0.65**	0.63
GPT-4o	0.55	-	0.57	**0.58**	0.56
QwQ-Plus	0.61	0.57	-	**0.62**	0.63
DeepSeek V3	0.65	0.58	0.62	-	0.63
Human	0.63	0.56	0.63	**0.63**	-

Table 5: **Kendall's** W between different evaluation models and human evaluation.

7 Discussion

In this section, we discuss the implications for future research and application development in the intersection of education and artificial intelligence carried by our work.

We believe that our work lays a foundational step for LLM-based educational research, offering a comprehensive benchmark and evaluation framework that captures the **diverse roles, scenarios, and needs** present in real-world education. By systematically incorporating scenarios like psychological counseling, assignment grading, and personalized study planning, we bring previously overlooked scenarios into the research landscape, encouraging deeper exploration across subject areas, learner profiles, and task types. Our benchmark can serve as a springboard for future research in designing better **better benchmarks that fulfill more diverse needs**, **robust evaluation models**, **scenario-adapted educational LLMs**, and even **LLM agents** that can perform multi-role, interactive support in classrooms or digital learning environments. Moreover, this work has immediate **practical value for educators and institutions**, offering structured tools that can help enhance efficiency, personalize learning, and reduce workload. The synthetic data construction methods we employ also open up new possibilities for **scalable, low-cost training and evaluation**, though future work could further improve context richness, realism, and dynamic data generation. Ultimately, we hope this work inspires the community to build stronger foundation models for trustworthy and effective educational AI systems.

8 Conclusion

In this work, we present the first comprehensive benchmark for evaluating LLMs in diverse educational scenarios. By incorporating data across 9 major domains and over 4,000 distinct educational contexts, it integrates model-generated query data to reflect real-world needs. We further introduce a set of multi-dimensional evaluation metrics spanning 12 critical aspects, addressing the perspectives of both educators and learners. Human annotations are employed to validate the quality and relevance of model-generated outputs, enhancing the benchmark's reliability. Extensive experiments show that smaller models trained on our dataset can rival state-of-the-art LLMs, underscoring the potential for efficient education-oriented LLMs. We believe this benchmark could serve as a valuable resource for the community and inspire further research in optimizing LLMs for educational applications.

Limitations

This work has several limitations that point to promising directions for future research. First, due to cost constraints, we relied on only one human annotator for evaluation, which may limit the reliability and generalizability of our findings; expanding the annotator pool could improve robustness. Second, the set of LLMs we evaluated is relatively limited, and including a wider variety of models would offer a more comprehensive understanding of system performance. Third, all of our query data was generated by models, which may not fully reflect realistic or diverse user intent. Future work could benefit from incorporating more human-written queries. Additionally, while our work explores the correlation between human and model evaluations, there is still room to improve alignment. We also employed only basic prompt engineering techniques; more sophisticated prompting strategies or the use of LLM agents may lead to better results. Moreover, most of our evaluation metrics and task scenarios were manually designed, and automating this process could enhance scalability and consistency. Finally, our methods have not yet been tested in real-world educational environments with practitioners, which will be important for validating practical applicability and impact.

Author Contributions

- **Bin Xu**: Conceived and designed the analysis; collected the data; performed the analysis; wrote the paper.
- **Yu Bai**: Developed the idea; conceived and designed the analysis; manage the project; performed the analysis; wrote the paper.
- **Huashan Sun**: Conceived and designed the analysis; collected the data; performed the analysis; wrote the paper.
- **Yiguan Lin**: Conceived and designed the analysis; collected the data; performed the analysis; wrote the paper.
- **Siming Liu**: Conceived and designed the analysis; performed the analysis; wrote the paper; provided training to the annotation company.
- **Xinyue Liang**: Performed the analysis; wrote the paper.
- **Yaolin Li**: Wrote the paper.
- **Yang Gao**: Conceptualization; lead and manage the project; provided computing resources and funding.
- **Heyan Huang**: Provided computing resources.

Ethical Consideration

In this study, all experiments involving human annotator comply fully with ethical standards set by a professional annotation company. The annotator participates voluntarily and is fully informed about the experimental procedures and task requirements before starting. Clear guidelines are provided, and the annotator receive sufficient training to ensure consistency and fairness in the annotation process. All data remain anonymized, and the privacy of annotator is strictly protected. This study ensures that no tasks involve potential harm or ethical concerns, and all experiments follow relevant ethical guidelines.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jonathan Berman, Lise McCoy, and Troy Camarata. 2024. LLM-Generated Multiple Choice Practice Quizzes for Pre-Clinical Medical Students; Use and Validity. *Physiology*, 39(S1):376.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-Adaptive Neural Automated Essay Scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1011–1020, Virtual Event China. ACM.

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

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

Eduard Frankford, Ingo Höhn, Clemens Sauerwein, and Ruth Breu. 2024a. A Survey Study on the State of the Art of Programming Exercise Generation Using Large Language Models. In *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*, pages 1–5, Würzburg, Germany. IEEE.

Eduard Frankford, Clemens Sauerwein, Patrick Bassner, Stephan Krusche, and Ruth Breu. 2024b. AI-Tutoring in Software Engineering Education.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large Language Models in Education: Vision and Opportunities. *arXiv preprint*. Issue: arXiv:2311.13160 arXiv:2311.13160 [cs].

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer Von Der Ohe, and Alona Fyshe. 2022. Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146, Dublin, Ireland. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hasnain Heickal and Andrew Lan. 2024. Generating Feedback-Ladders for Logical Errors in Programming using Large Language Models. *arXiv preprint*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Chieh-Yang Huang, Jing Wei, and Ting-Hao Kenneth Huang. 2024a. Generating Educational Materials with Different Levels of Readability using LLMs. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 16–22, Honolulu HI USA. ACM.

Zhenya Huang, Yuting Ning, Longhu Qin, Shiwei Tong, Shangzi Xue, Tong Xiao, Xin Lin, Jiayu Liu, Qi Liu, Enhong Chen, and Shijing Wang. 2024b. EduNLP: Towards a Unified and Modularized Library for Educational Resources. *arXiv preprint*. Issue: arXiv:2406.01276 arXiv:2406.01276 [cs].

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2021. EQG-RACE: Examination-Type Question Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13143–13151.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. LLMs can Find Mathematical Reasoning Mistakes by Pedagogical Chain-of-Thought. *arXiv preprint*.

Seungyoon Kim and Seungone Kim. 2024. Can Language Models Evaluate Human Written Text? Case Study on Korean Student Writing for Education. *arXiv preprint*.

Charles Koutcheme, Nicola Dainese, and Arto Hellas. 2024a. Using Program Repair as a Proxy for Language Models' Feedback Ability in Programming Education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 165–181, Mexico City, Mexico. Association for Computational Linguistics.

Charles Koutcheme, Nicola Dainese, Sami Sarsa, Juho Leinonen, Arto Hellas, and Paul Denny. 2024b. Benchmarking Educational Program Repair. *arXiv preprint*. Issue: arXiv:2405.05347 arXiv:2405.05347 [cs].

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

Vijini Liyanage and Surangika Ranathunga. 2019. A Multi-language Platform for Generating Algebraic Mathematical Word Problems. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 332–337, Kandy, Sri Lanka. IEEE.

Evanfiya Logacheva, Arto Hellas, James Prather, Sami Sarsa, and Juho Leinonen. 2024. Evaluating Contextually Personalized Programming Exercises Created with Generative AI. *arXiv preprint*.

Stephen MacNeil, Andrew Tran, Arto Hellas, Joanne Kim, Sami Sarsa, Paul Denny, Seth Bernstein, and Juho Leinonen. 2023. Experiences from Using Code Explanations Generated by Large Language Models in a Web Software Development E-Book. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 931–937, Toronto ON Canada. ACM.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT-based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, pages 226–236, Copenhagen Denmark. ACM.

Hunter McNichols, Mengxue Zhang, and Andrew Lan. 2023. Algebra Error Classification with Large Language Models. *arXiv preprint*.

Chee Ng and Yuen Fung. 2024. Educational Personalized Learning Path Planning with Large Language Models. *arXiv preprint*. Issue: arXiv:2407.11773 arXiv:2407.11773 [cs].

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *arXiv preprint arXiv: 2410.21276*.

Zachary A. Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint*.

Team Qwen. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Team Qwen. 2025. Qwq-32b: Embracing the power of reinforcement learning.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint*. Issue: arXiv:2311.12022 arXiv:2311.12022 [cs].

Hamdireza Rouzegar and Masoud Makrehchi. 2024. Generative AI for Enhancing Active Learning in Education: A Comparative Study of GPT-3.5 and GPT-4 in Crafting Customized Test Questions.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic Generation of Programming Exercises and Code Explanations using Large Language Models.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, and 1 others. 2025. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*.

M.-A.-P. Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, and 78 others. 2025. SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines. *arXiv preprint*. Issue: arXiv:2502.14739 arXiv:2502.14739 [cs].

Felipe Urrutia and Roberto Araya. 2023. Automatically Detecting Incoherent Written Math Answers of Fourth-Graders. *Systems*, 11(7):353.

Annapurna Vadaparty, Daniel Zingaro, David H. Smith, Mounika Padala, Christine Alvarado, Jamie Gorson Benario, and Leo Porter. 2024. CS1-LLM: Integrating LLMs into CS1 Instruction. *arXiv preprint*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and 1 others. 2024a. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024b. Large Language Models for Education: A Survey and Outlook. *arXiv preprint*. Issue: arXiv:2403.18105 arXiv:2403.18105 [cs] version: 1.

Tianyu Wang, Nianjun Zhou, and Zhixiong Chen. 2024c. Enhancing Computer Programming Education with LLMs: A Study on Effective Prompt Engineering for Python Code Generation. *arXiv preprint*.

Boyang Yang, Haoye Tian, Weiguo Pian, Haoran Yu, Haitao Wang, Jacques Klein, Tegawendé F. Bissyandé, and Shunfu Jin. 2024a. CREF: An LLM-based Conversational Software Repair Framework for Programming Tutors. *arXiv preprint*.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S. Bernstein, and John Mitchell. 2024b. Social Skill Training with Large Language Models. *arXiv preprint*. Issue: arXiv:2404.04204 arXiv:2404.04204 [cs].

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. Simulating Classroom Education with LLM-Empowered Agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.

A Detailed Scenario Descriptions

This appendix provides comprehensive definitions and explanations of all task scenarios included in EduBench. Each scenario captures a specific educational use case, featuring unique roles, goals, and evaluation considerations. The descriptions are organized into two categories—student-oriented and teacher-oriented—based on the primary user group served by the tasks.

A.1 Student-Oriented Scenarios

This section outlines scenarios in which AI systems directly assist students in their learning journey. The tasks are designed to reflect authentic student needs, including solving academic problems, receiving feedback on errors, obtaining personalized learning recommendations, and even receiving emotional support. These scenarios emphasize interaction quality, content accuracy, and adaptability to individual learning situations.

- **Problem Solving:** The ability of an AI system to accurately solve questions posed by students across various subjects and difficulty levels.
- **Error Correction:** The capacity to identify and correct student errors in assignments, exams, or daily exercises. Errors can range from obvious mistakes to subtle issues such as variable misuse in code or logical flaws in mathematical reasoning. Evaluation focuses on the accuracy of error detection and the quality of correction.
- **Idea Provision:** This includes answering student queries about knowledge points, homework guidance, or exam preparation. It is subdivided into basic factual explanations, step-by-step solution analysis, and general academic advice. Responses are evaluated for accuracy, clarity, and informativeness.
- **Personalized Learning Support:** Based on student profiles (e.g., skill level, learning goals), the system recommends learning paths, exercises, or reading materials tailored to individual needs. Effectiveness is judged by the relevance, difficulty alignment, and usefulness of the recommendations.
- **Emotional Support:** This involves detecting a student's emotional state (e.g., anxiety before exams) from text and offering appropriate supportive feedback or suggestions. Scenarios include pre-exam stress, post-exam frustration, or social isolation. Evaluation metrics include emotion classification accuracy, specificity of emotional cues, and quality of suggestions.

A.2 Teacher-Oriented Scenarios

This section focuses on scenarios where AI systems are used to support educators in instructional design, assessment, and personalized teaching. These tasks capture the typical responsibilities of teachers—such as generating exam questions, grading, and preparing learning materials—and evaluate how effectively AI can augment or automate these functions to improve teaching efficiency and quality.

- **Question Generation:** Generating questions based on specified topics, difficulty levels, and knowledge scopes. This includes both single-topic and multi-topic (comprehensive) question generation. Advanced requirements involve generating explanations and formatting full exams. Evaluation focuses on question quality, relevance, and structural coherence.
- **Automatic Grading:** Supporting grading of objective questions (e.g., multiple-choice, fill-in-the-blank) and subjective tasks (e.g., project reports) based on scoring rubrics. Feedback generation is also supported. Metrics include scoring accuracy, reasonableness, and feedback informativeness.
- **Teaching Material Generation:** Automatically generating educational content such as slides, teaching plans, and lecture notes. This includes content structuring and supplementing with relevant external materials like images or references.

- **Personalized Content Creation:** Generating differentiated content for students based on their learning levels or personal profiles. This includes both individualized assignments and tiered content design (e.g., differentiated learning objectives, teaching strategies, and assessments for varying student levels). Evaluation focuses on the internal validity of each item and cross-tier consistency.

B Educational Context Details

Subject Taxonomy: We follow a two-tier classification system to reflect academic breadth:

- **K–12 Subjects:**
  - Chinese, Mathematics, English, Physics, Chemistry, Biology, History, Geography.
- **Higher Education Subjects:**
  - **Sciences:** Mathematics, Physics, Chemistry, Biology, Astronomy
  - **Engineering:** Computer Science, Automation Control, Aerospace Science and Technology
  - **Agriculture:** Aquaculture, Crop Science
  - **Economics:** Applied Economics, Theoretical Economics
  - **Education:** General Education, Physical Education
  - **Management:** Business Administration, Public Administration
  - **Medicine:** Basic Medicine, Clinical Medicine
  - **Social Sciences and Humanities:** Sociology, Psychology, History, Law, Management
  - **Literature and Arts:** Linguistics, Journalism, Theory of Music
  - **Military Science**

Task Difficulty Design: Tasks are divided into three levels:

- **Easy** – Basic knowledge and low cognitive load.
- **Medium** – Intermediate tasks requiring moderate understanding.
- **Hard** – Complex problems demanding critical reasoning and expertise.

Language:
- EduBench currently supports tasks in **Chinese and English**.

Question Types: The question type dimension captures the format of interaction or evaluation expected from the model:

- **Standard Scenarios (e.g., Problem Solving, Idea Provision, Grading):**
  - Single Choice
  - Multiple Choice
  - Short Answer
- **Emotional Support Scenarios:**
  - Mental Healthy
  - Mild Anxiety
  - Moderate Anxiety
  - Severe Anxiety
- **Personalized Learning Support & Personalized Content Creation:**
  - No explicit question type; task outputs are tailored recommendations or generated content based on learner profiles.

C Evaluation Metric Design Details

C.1 Scenario Adaptation Criteria

The Scenario Adaptation metric evaluates whether the model's output aligns with the scenario-specific expectations and pedagogical goals. Below are detailed descriptions of its four sub-components:

- **Instruction Following & Task Completion:** This sub-metric measures the model's ability to accurately interpret and complete assigned tasks, such as solving problems, correcting errors, or generating questions, while adhering to the required output format and constraints.
- **Role & Tone Consistency:** This dimension evaluates whether the language style, tone, and level of expertise in the response are appropriate for the designated role (e.g., teacher, teaching assistant, peer) and the target learner group (e.g., primary school students, university students).
- **Content Relevance & Scope Control:** The response is assessed for its focus on the specified topic or knowledge area, as well as its ability to stay within the intended difficulty level, subject boundaries, and content scope.
- **Scenario Element Integration:** This sub-metric measures the degree to which the model effectively incorporates scenario-specific information, such as prior student responses, individual learning preferences, or stated pedagogical objectives. This is especially important in personalized learning and interactive tutoring contexts.

C.2 Factual & Reasoning Accuracy Criteria

This metric evaluates whether a model's response is grounded in factual correctness and logical rigor, particularly in scenario-intensive or multi-step tasks. It includes the following sub-components:

- **Basic Factual Accuracy:** This sub-metric examines the accuracy of objective information, including definitions, formulas, factual statements, code syntax, and terminology.
- **Domain Knowledge Accuracy:** It assesses the appropriateness and depth of subject-specific knowledge presented in the response, ensuring alignment with disciplinary standards across domains such as mathematics, law, and computer science.
- **Reasoning Process Rigor:** This criterion focuses on the completeness and logical validity of the model's reasoning in tasks that require multi-step derivations, explanations, or justifications.
- **Error Identification & Correction Precision:** In contexts involving diagnostics or feedback, this sub-metric evaluates the model's ability to accurately detect, localize, and correct errors without introducing false positives or negatives.

C.3 Pedagogical Application Criteria

This metric evaluates whether a model's response demonstrates pedagogical effectiveness and contributes meaningfully to learning outcomes. It includes the following sub-components:

- **Clarity, Simplicity & Inspiration:** This sub-metric assesses whether the explanation is articulated clearly and accessibly, using appropriate language to promote understanding and stimulate student interest or engagement.
- **Motivation, Guidance & Positive Feedback:** It evaluates the model's ability to encourage learners through constructive feedback and supportive guidance, promoting confidence and independent thinking rather than relying on direct answers alone.
- **Personalization, Adaptation & Learning Support:** This criterion measures the response's ability to adapt based on the learner's background, proficiency level, and individual needs, including tailored suggestions, scaffolded prompts, and relevant resource recommendations.
- **Higher-Order Thinking & Skill Development:** This sub-metric examines whether the response promotes advanced cognitive skills, such as critical thinking, problem-solving, creative reasoning, and the ability to transfer knowledge to new contexts.

C.4 Metric allocation for each scenario

To ensure that evaluation is both fair and context-sensitive, we dynamically allocate evaluation metrics based on the instructional characteristics and goals of each scenario. Not all metrics are applicable across all scenarios: for example, reasoning rigor is essential in problem-solving, while emotional and adaptive support is critical in student guidance tasks. The following table summarizes the allocation of the designed metrics (Section 4.3) across the nine scenarios in EduBench:

Evaluation Metric	Error Correction	Idea Provision	Grading	Answering Questions	Material Generation	Question Generation	Mental Health	Personalized Content Creation	Learning Support
Instruction Following & Task Completion	✓	✓	✓	✓	✓	✓	✓	✓	✓
Role & Tone Consistency					✓		✓		
Content Relevance & Scope Control		✓	✓	✓	✓	✓			✓
Scenario Element Integration	✓	✓					✓	✓	✓
Basic Factual Accuracy	✓	✓	✓	✓	✓	✓			
Domain Knowledge Accuracy		✓			✓	✓			
Reasoning Process Rigor	✓	✓	✓	✓					
Error Identification & Correction Precision	✓		✓						
Clarity, Simplicity & Inspiration	✓	✓			✓	✓			
Motivation, Guidance & Positive Feedback	✓		✓				✓		
Personalization, Adaptation & Learning Support							✓	✓	✓
Higher-Order Thinking & Skill Development		✓			✓	✓			✓

Table 6: Allocation of evaluation metrics across nine educational scenarios in EduBench.

This allocation ensures that each scenario is evaluated according to the dimensions most critical to its pedagogical purpose. For instance, cognitive rigor is emphasized in analytical tasks (e.g., problem-solving and grading), while adaptive support and contextual integration are prioritized in student-facing personalization tasks (e.g., learning path design or mental health feedback). This scenario-aware evaluation design enhances the interpretability, accuracy, and instructional relevance of the benchmark results.

D Evaluation Prompt Design

To ensure consistent and fair human evaluation across multiple educational tasks and languages, we carefully design a suite of evaluation prompts aligned with our 12 fine-grained metrics (see Section 4). Each prompt is tailored to elicit targeted human judgment on a specific aspect of model behavior, such as factual accuracy, instruction following, reasoning rigor, or pedagogical impact.

Model Evaluation Prompt Design

I will provide you with an educational question and its corresponding answer. Please evaluate the given answer based on the provided assessment criteria and scoring principles, and output the score along with the reasons in JSON format.
Scoring principles: {principle}
Question: {question}
Answer: {response}
The JSON format is defined as follows:

```
{
  "detailed_scores": [
    {
      "principle": "principle 1",
      "score": 0,
      "reason": ""
    },
    ...
  ]
}
```

E Prompt Design Examples

This appendix provides illustrative examples of the prompt templates used during the question data generation process described in Section 3.3. Each example corresponds to a specific task scenario, domain, and target competency, and is designed to ensure the quality, diversity, and controllability of the generated data. The question design and repsonse generaion prompts are as follows: Problem Solving E.1, Error Correction E.2, Idea Provision E.3, Personalized Learning Support E.4, Emotional Support E.5, Question Generation E.6, Automatic Grading E.7, Teaching Material Generation E.8, Personalized Content Creation E.9.

E.1 Problem Solving

Problem Solving Prompt Design

Please freely generate an appropriate question based on the following subject and difficulty level, and provide a standard answer. The question type is {question_type}.
If the type is a short-answer question, for certain subjects, provide code and mathematical calculations if necessary. Do not return extra content.

Subject: {subject}
Difficulty Level: {level}

Return in JSON format:

```
"question":
"answer":
```

Description: This prompt corresponds to the following capability requirement:
- **Capability Requirement:** Problem-Solving Ability
- **Basic Description:** The ability to solve problems raised by students.
- **Scenario Design:** Answering questions across different subjects.

Prompt Design for Obtaining Model Responses in This Scenario

{question}. Please generate the corresponding answer based on the question.
"Answer":
Return in JSON format.

E.2 Error Correction

Error Correction Prompt Design

You are an expert teacher in all subjects, helping students correct errors. Please freely generate a question and an incorrect student answer based on the following subject and difficulty level, and provide corrections. The question type is {question_type}.
If the type is a short-answer question, for certain subjects, provide code and mathematical calculations if necessary. Do not return extra content.

Subject: {subject}
Difficulty Level: {level}

Return in JSON format:

```
"question":
"original_answer":
"corrected_answer":
"correction_explanation":
```

Description: This prompt corresponds to the following capability requirement:

- **Capability Requirement:** Error Correction Ability
- **Basic Description:** The ability to identify and correct errors in students' homework, exams, or daily learning, providing targeted improvement suggestions.
- **Scenario Design:** Basic error correction: Input incorrect solutions, identify errors, and correct them (e.g., code, math).

Prompt Design for Obtaining Model Responses in This Scenario

{question}{original_answer} You are providing error correction services for students' answers. Please provide a "Corrected Answer" and "Error Explanation" based on this question and the original answer.
"Corrected Answer":
"Error Explanation":
Return in JSON format.

E.3 Idea Provision

Idea Provision Prompt Design

You are an expert teacher in all subjects, helping students with problem-solving guidance instead of providing standard answers. Please freely generate a question based on the following subject and difficulty level, and provide guidance without giving the answer. The question type is {question_type}.
If the type is a short-answer question, for certain subjects, provide code and mathematical calculations if necessary. Do not return extra content.

Subject: {subject}
Difficulty Level: {level}

Return in JSON format:

```
"question":
"provided_guidance":
```

Description: This prompt corresponds to the following capability requirement:
- **:** Q&A Ability
- **Basic Description:** The ability to answer students' learning questions in real-time, covering knowledge point explanations, homework assistance, exam preparation guidance, etc.
- **Scenario Design:** - Basic knowledge questions: Explanation of knowledge points (accuracy, simplicity, inspiration). - Question-solving analysis: Explanation of problem-solving steps, decomposition of complex problems, analysis of involved knowledge points, and summarization of experience. - General: Including study techniques, exam strategies, etc.

Prompt Design for Obtaining Model Responses in This Scenario

{question}Please provide an approach based on this question.
"Provided Approach":
Return in JSON format.

E.4 Personalized Learning Support

Personalized Learning Support Prompt Design

You are a personalized service customization expert, providing tailored services to improve learning efficiency. Please freely generate a specific and appropriate student profile based on the following subject and difficulty level, and provide learning path planning suggestions and personalized recommendations. The question type is {question_type}.
If the type is a short-answer question, for certain subjects, provide code and mathematical calculations if necessary. Do not return extra content.

Subject: {subject}
Difficulty Level: {level}

Return in JSON format:

```
"student_profile":
"learning_path_planning_suggestions":
"personalized_recommendations":
```

Description: This prompt corresponds to the following capability requirement:
- **Capability Requirement:** Personalized Service
- **Basic Description:** Based on student profiles, provide customized services to enhance learning efficiency.
- **Scenario Design:** - Learning path planning: Arrange future courses based on current ability levels and learning goals. - Personalized recommendations: Generate or recommend practice exercises and reading materials based on weak knowledge areas and learning habits.

Prompt Design for Obtaining Model Responses in This Scenario

{student_profile} Based on the student profile, provide "Learning Path Planning" and "Personalized Recommendations",
"Learning Path Planning":
"Personalized Recommendations":
Returned in JSON format.

E.5 Emotional Support

Emotional Support Prompt Design

You are an intelligent assistant capable of identifying students' emotional states, analyzing the causes of emotional issues, and providing relevant comfort and advice. Please freely generate a multi-turn conversation with a student studying {subject}, identify their emotional state, analyze the causes of emotional issues, and provide relevant comfort and advice. The anxiety level is {anxiety_level}.
Do not return extra content.

Academic Level: {level}

Return in JSON format:

```
"conversation_with_student":
"emotional_state_analysis":
"comfort_and_advice":
```

Description: This prompt corresponds to the following capability requirement:
- **Capability Requirement:** Psychological Support
- **Basic Description:** The ability to identify students' emotional states, analyze the causes of emotional issues, and provide relevant comfort and advice.
- **Scenario Design:** - Emotional recognition: Identify emotional states (e.g., normal, mild anxiety, severe anxiety) based on conversations with students. - Comfort and advice: Provide targeted comfort and advice for different emotional issues (e.g., pre-exam anxiety, post-exam frustration, social isolation). - Task scenario: Input: Student's text description or questions; Output: Emotional state recognition, targeted advice, and possible follow-up action suggestions. - Evaluation metrics: Recognition accuracy, emotional granularity, cause identification, response relevance, and practicality of advice.

Prompt Design for Obtaining Model Responses in This Scenario

{conversation_with_student}{anxiety_level} Please provide "Emotional State Analysis" and "Comfort & Suggestions" based on the student's emotional state and conversation.
"Emotional State Analysis":
"Comfort & Suggestions":
Return in JSON format.

E.6 Question Generation

Question Generation Prompt Design

You are an expert teacher in all subjects, generating appropriate questions based on knowledge scope and question types. Please freely generate a knowledge point and corresponding question based on the following subject and difficulty level, provide guidance, and give an answer. The question type is {question_type}.
If the type is a short-answer question, for certain subjects, provide code and mathematical calculations if necessary. Do not return extra content.

Subject: {subject}
Difficulty Level: {level}

Return in JSON format:

```
"knowledge_point":
"question":
"provided_guidance":
"answer":
```

Description: This prompt corresponds to the following capability requirement:
- **Capability Requirement:** Question Generation
- **Basic Description:** Generate questions based on knowledge scope and question types, considering difficulty levels.
- **Scenario Design:** - Basic question generation: Generate questions for different difficulty levels, question types, and knowledge scopes. - Comprehensive question generation: Cross-reference multiple knowledge points to generate questions. - Additional requirements: - Provide solutions and step-by-step scoring references for questions. - Compile quizzes or exams based on syllabi to form structured classroom assessments.

Prompt Design for Obtaining Model Responses in This Scenario

{knowledge_point}{subject}{question_type}{level}
Please generate a question based on the subject, academic level, knowledge point, and question type.
"Question":
Return in JSON format.

E.7 Automatic Grading

Automatic Grading Prompt Design

You need to implement: 1. Objective grading: Grade multiple-choice, true/false, and fill-in-the-blank questions; provide step-by-step scoring for open-ended questions. 2. Subjective grading: Evaluate large assignments and lab reports comprehensively (e.g., workload, completeness, knowledge application). 3. Personalized feedback: Generate constructive feedback, including potential knowledge gaps and learning suggestions.
Please freely generate a question and a student's answer based on the following subject and difficulty level, and grade the answer. The question type is {question_type}.
If the type is a short-answer question, for certain subjects, provide code and mathematical calculations if necessary. Do not return extra content.

Subject: {subject}
Difficulty Level: {level}

Return in JSON format:

```
"question":
"student_answer":
"grading":
"grading_details":
"personalized_feedback":
```

Description: This prompt corresponds to the following capability requirement:
- **Capability Requirement:** Automated Homework Grading
- **Basic Description:** Automatically grade students' homework and analyze results to provide suggestions.
- **Scenario Design:** - Objective grading: Grade multiple-choice, true/false, and fill-in-the-blank questions; provide step-by-step scoring for open-ended questions. - Subjective grading: Evaluate large assignments and lab reports comprehensively (e.g., workload, completeness, knowledge application).
- Personalized feedback: Generate constructive feedback, including potential knowledge gaps and learning suggestions.

Prompt Design for Obtaining Model Responses in This Scenario

{question}{student_answer} Please provide "Score", "Scoring Details", and "Personalized Feedback" based on the question and student's answer.
"Score":
"Scoring Details":
"Personalized Feedback":
Return in JSON format.

E.8 Teaching Material Generation

Teaching Material Generation Prompt Design

You are responsible for helping teachers generate high-quality teaching materials, including lesson plans, presentations, and lecture notes. Based on textbook chapters or knowledge points, automatically generate structured lesson plans, including learning objectives, key points, difficult points, and classroom activity designs. The question type is {question_type}.
If the type is a short-answer question, for certain subjects, provide code and mathematical calculations if necessary. Do not return extra content.

Subject: {subject}
Difficulty Level: {level}

Return in JSON format:

```
"knowledge_point":
"teaching_materials":
```

Description: This prompt corresponds to the following capability requirement:
- **Capability Requirement:** Teaching Material Generation
- **Basic Description:** Generate high-quality teaching materials, including lesson plans, presentations, and lecture notes.
- **Scenario Design:** - Course PPT or lecture note generation: Automatically generate slides or detailed notes based on textbook chapters or knowledge points. - Lesson plan generation: Automatically generate structured lesson plans, including learning objectives, key points, difficult points, and classroom activity designs. - Course-related material generation or retrieval: Generate or search for relevant materials, such as images, teaching cases, links, and references.

Prompt Design for Obtaining Model Responses in This Scenario

{knowledge_point} Please provide "Teaching Material" based on this knowledge point. The teaching material should include teaching objectives, key points and difficulties, classroom activity design, etc.
"Teaching Material":
Return in JSON format.

E.9 Personalized Content Creation

Personalized Content Creation Prompt Design

You are an intelligent assistant capable of generating personalized learning content or tasks based on individual student differences. Please freely generate a student profile for a student studying {subject}, and consider the following three aspects: 1. One-on-one: Customize practice questions or reading materials based on specific student profiles. 2. Tiered teaching: For the same course content, generate different teaching objectives, methods, assessment methods, and homework assignments for students at different levels. 3. Other: Combine other capability requirements to design differentiated ability evaluation data for individual students, study groups, and classes.
Do not return extra content.
Academic Level: {level}
Strictly return in JSON format:

```
"student_profile":
"personalized_learning_content_or_tasks":
```

Description: This prompt corresponds to the following capability requirement:
- **Capability Requirement:** Personalized Content Generation
- **Basic Description:** Generate personalized learning content or tasks based on individual student differences.
- **Scenario Design:** - One-on-one customization: Based on specific student profiles, generate tailored practice questions or reading materials. - Tiered teaching: For the same course content, generate different teaching objectives, methods, assessment methods, and homework assignments for students at different levels. - Other considerations: Combine other capability requirements to design differentiated ability evaluation data for individual students, study groups, and classes.

Prompt Design for Obtaining Model Responses in This Scenario

{student_profile} Based on the student profile, generate a 'personalized learning content or task' for each student.
"personalized learning content or task":
Return in JSON format.

F Comprehensive Evaluation Metric Scoring Details

To systematically evaluate the quality of AI-generated responses in educational settings, we designed a comprehensive human evaluation rubric comprising three main dimensions, each containing several fine-grained criteria. Each criterion is rated on a 10-point scale with clearly defined level anchors to guide consistent judgment. Details are listed as follows:

F.1 Instructional Quality

F.1.1 Instruction Following & Task Completion (IFTC)

Description: Did it fully understand and execute the user's instruction? Was the core task (e.g., solving problems, error correction, question generation) completed? Is the output formatting correct?

- **9-10:** Fully understood and precisely executed all instructions; achieved core task with perfect accuracy; output format is fully compliant.
- **7-8:** Accurately understood main instructions and correctly completed the task; core goals are well achieved; format is mostly correct with only minor omissions or deviations.
- **5-6:** Understood the general intent but may miss some details; task largely completed but with some inaccuracies or omissions; formatting attempts present but with notable flaws.
- **3-4:** Misunderstood part of the instruction; low task completion or major errors; formatting mostly incorrect.
- **1-2:** Completely misunderstood or ignored instructions; task not completed or totally incorrect; formatting is chaotic or irrelevant.

F.1.2 Role & Tone Consistency (RTC)

Description: Does the language style, tone, and level of professionalism match the assigned role (e.g., teacher, teaching assistant, peer) and the target learner group (e.g., elementary, college)?

- **9-10:** Excellent role-playing (e.g., teacher/TA); language style, professionalism, and tone (e.g., encouraging/serious) are perfectly aligned with the assumed role and audience.
- **7-8:** Role and tone are mostly consistent and appropriate for the scenario, with minor deviation in individual expressions.

- **5-6:** Attempts to match the role and tone can be seen, but overall consistency is weak; some expressions are disconnected from the role/scenario.
- **3-4:** Significant mismatch in role and tone; comes across as unnatural or inconsistent.
- **1-2:** No reflection of assigned role/tone; expression entirely inconsistent with the scenario.

F.1.3 Content Relevance & Scope Control (CRSC)

Description: Is the content tightly aligned with the specified topic, theme, or question? Is it kept within the specified difficulty level, scenario, or scope?

- **9-10:** Content is highly relevant to the specified topic/theme/question; strictly within required difficulty/scope/discipline without redundant or irrelevant information.
- **7-8:** Overall relevance is high; scope control is good with possibly a small amount of slightly off-topic or mildly overreaching information.
- **5-6:** Mostly relevant, but includes some off-topic or out-of-scope content; scope control needs improvement.
- **3-4:** Poor relevance; includes a significant amount of irrelevant information or is largely outside scope.
- **1-2:** Content is largely irrelevant or completely outside the specified scope.

F.1.4 Scenario Element Integration (SEI)

Description: Did it effectively use scenario-specific information (e.g., previous student answers, learning preferences, specific teaching goals)? Especially important in personalized, Q&A, or error-correction contexts.

- **9-10:** Fully integrated all key scenario elements (e.g., student history, learning preferences); output is highly personalized and well-matched to the teaching context.
- **7-8:** Used major scenario elements effectively; response is targeted, possibly overlooks minor details but does not affect overall results.
- **5-6:** Some use of scenario information, but integration is shallow; personalization or contextual fit is average.
- **3-4:** Only surface-level reference to scenario information; did not integrate core elements effectively; weak contextual connection.
- **1-2:** Completely ignored scenario-specific information; output is generic, templated, and irrelevant to the scenario.

F.2 Content Accuracy

F.2.1 Basic Factual Accuracy (BFA)

Description: Are objective facts such as concept definitions, formulas, dates, terminology, code syntax, legal clauses correctly presented?

- **9-10:** All stated factual elements (definitions, formulas, dates, terms, syntax, etc.) are completely accurate.
- **7-8:** Vast majority of facts are correct; possibly contains very minor, non-critical typos or omissions.
- **5-6:** Most facts are correct, but there are some notable factual errors that require review.
- **3-4:** Contains several or key factual inaccuracies; information is not trustworthy.
- **1-2:** Riddled with factual errors; information is completely incorrect or misleading.

F.2.2 Domain Knowledge Accuracy (DKA)

Description: Is the use of subject matter knowledge (math, programming, law, finance, etc.) not only correct but also appropriately specialized and aligned with domain standards?

- **9-10:** Subject matter application is not only accurate but also shows appropriate depth and rigor; adheres to industry or academic standards.
- **7-8:** Proper use of professional knowledge reflecting a good degree of proficiency; minor shortcomings in depth or detail not affecting validity.
- **5-6:** Basic accuracy in subject knowledge, but somewhat surface-level or lacking rigor; some confusion or omissions of non-core concepts.
- **3-4:** Significant errors or major omission in subject-specific content; lacks professionalism.
- **1-2:** Serious domain errors; completely incorrect or misleading; does not meet any professional standards.

F.2.3 Reasoning Process Rigor (RPR)

Description: For content requiring reasoning (e.g., math steps, code logic, legal arguments, case analysis), is the logical flow complete and sound?

- **9-10:** Reasoning is complete, clear, and rigorous; all steps are correct; arguments are strong and free of logical fallacies.
- **7-8:** Reasoning is largely correct and logically coherent with minor issues in individual steps or details that do not affect the conclusion.
- **5-6:** Reasoning is visible but contains unclear logic, missing steps, or insufficient argumentation, affecting the overall outcome.
- **3-4:** Reasoning has major logical flaws, confusion in steps, or critical omissions; reliability is low.
- **1-2:** Virtually no valid reasoning; logic is chaotic; steps are incorrect or irrelevant.

F.2.4 Error Identification & Correction Precision (EICP)

Description: In error correction scenarios, are errors precisely identified (no missed or false positives)? Are the corrections correct and optimal?

- **9-10:** Precisely identified all errors (no omission or false positives); provided completely correct, clear, and optimal correction suggestions.
- **7-8:** Correctly located most major errors; suggestions are generally accurate and effective with only minor omissions or less-than-perfect advice.
- **5-6:** Identified some errors but with clear omissions or false positives; suggestions are partially correct but may lack clarity, completeness or optimality.
- **3-4:** Inaccurate error detection with critical omissions or many false positives; suggestions contain errors or are hard to comprehend.
- **1-2:** Completely failed to detect errors; provided entirely incorrect or misleading correction advice.

F.3 Pedagogical Effectiveness

F.3.1 Clarity, Simplicity & Inspiration (CSI)

Description: Are explanations, descriptions, and feedback clear, concise, and easy for the target learners to understand? Is the delivery inspiring and thought-provoking?

- **9-10:** Extremely clear and concise explanations; fully accessible for target learners; vibrant and engaging delivery that inspires deep thought and interest.
- **7-8:** Clear and easy to understand; appropriate for learner level; somewhat thought-provoking and can trigger reflection.
- **5-6:** Generally understandable but may be wordy, complex, or dull; limited inspirational impact.
- **3-4:** Lacks clarity; uses excessive jargon or complex structures; difficult to comprehend; uninspiring.
- **1-2:** Confusing and hard to follow; disregards learner needs; offers no inspiration and may cause confusion.

F.3.2 Motivation, Guidance & Positive Feedback (MGP)

Description: Does the interaction provide encouragement and support? Is constructive and positive language used? In answering or tutoring, does it guide thinking or just give away answers?

- **9-10:** Strongly supportive and encouraging; consistently uses constructive and positive language; offers highly effective heuristic guidance instead of simply giving answers.
- **7-8:** Generally supportive tone and positive language; provides useful guidance though occasionally too direct.
- **5-6:** A mix of encouragement and neutral/critical language; guidance is inconsistent—sometimes helpful, sometimes overly direct or lacking.
- **3-4:** Lacks encouragement and support; language is neutral or mildly negative; rarely guides, often just answers or remains unhelpful.
- **1-2:** Negative or discouraging tone; no motivation or support; fails to guide or gives misleading suggestions.

F.3.3 Personalization, Adaptation & Learning Support (PAS)

Description: Can it provide differentiated content, advice, or feedback based on a student's level, traits, or needs? Does it recommend effective learning paths or resources?

- **9-10:** Highly personalized content/advice/feedback based on student level/traits/needs; resource and learning path suggestions are accurate, practical, and valuable.
- **7-8:** Demonstrates some adaptation to student situation; provides relevant learning advice or resources with good utility.
- **5-6:** Attempts personalization but with limited effectiveness; recommendations are generic and of limited value.
- **3-4:** Little to no personalization; output is the same for everyone; learning support is insufficient or unrelated.
- **1-2:** No personalization; output may conflict with student needs; offers no or incorrect learning support.

F.3.4 Higher-Order Thinking & Skill Development (HOTS)

Description: Does the interaction or content help foster students' critical thinking, creativity, problem-solving, or knowledge transfer skills?

- **9-10:** Skillfully designed to promote critical/creative thinking, problem-solving, or transfer of knowledge (e.g., through open-ended questions, comparative analysis, case study, project-based tasks).
- **7-8:** Includes guiding questions or moderately challenging tasks that positively support the development of higher-order thinking (e.g., analysis, evaluation, application).
- **5-6:** Some attempt to encourage higher-order thinking (e.g., simple reflective questions), but limited in depth and scope; mainly focused on rote understanding or basic application.
- **3-4:** Interaction/content mostly revolves around memory and comprehension; rarely addresses higher-order thinking tasks.
- **1-2:** Completely ignores higher-order skill development; encourages rote memorization and repetition; may inhibit thinking flexibility.

G Human Annotator Cost

The cost for each QA pair is $2.22. We provided 198 questions (99 in English, 99 in Chinese) with 5 responses per question, totaling 990 QA pairs. The final cost is approximately $2,194.

H Detailed Correlation Analysis Between Model Evaluation and Human Assessment across Dimensions

Model	Instruction Following & Task Completion					Role & Tone Consistency				
	DeepSeek R1	**GPT-4o**	**QwQ-Plus**	**DeepSeek V3**	**Human**	**DeepSeek R1**	**GPT-4o**	**QwQ-Plus**	**DeepSeek V3**	**Human**
DeepSeek R1	-	0.56	0.66	0.69	0.63	-	0.48	0.55	0.65	0.71
GPT-4o	0.56	-	0.55	0.52	0.53	0.48	-	0.55	0.56	0.53
QwQ-Plus	0.66	0.55	-	0.63	0.58	0.55	0.55	-	0.58	0.59
DeepSeek V3	0.69	0.52	0.63	-	0.65	0.65	0.56	0.58	-	0.7
Human	0.63	0.53	0.58	**0.65**	-	**0.71**	0.53	0.59	0.7	-
	Content Relevance & Scope Control					**Scenario Element Integration**				
	DeepSeek R1	**GPT-4o**	**QwQ-Plus**	**DeepSeek V3**	**Human**	**DeepSeek R1**	**GPT-4o**	**QwQ-Plus**	**DeepSeek V3**	**Human**
DeepSeek R1	-	0.57	0.61	0.65	0.61	-	0.55	0.59	0.7	0.66
GPT-4o	0.57	-	0.6	0.6	0.55	0.55	-	0.53	0.56	0.54
QwQ-Plus	0.61	0.6	-	0.62	0.64	0.59	0.53	-	0.61	0.63
DeepSeek V3	0.65	0.6	0.62	-	0.61	0.7	0.56	0.61	-	0.7
Human	0.61	0.55	**0.64**	0.61	-	0.66	0.54	0.63	**0.7**	-

Table 7: **Kendall's** W between different evaluation models and human evaluation in Instructional Quality.

I Pre-Experiments

I.1 Existence of self-preference

Given the potential influence of both the response generation model and the evaluation model, it is crucial to verify whether a model tends to favor its own responses. To examine this, we use each of the three response-generation models as evaluators to assess the responses they themselves generated. Specifically, for each question, we construct pairwise comparisons among the responses generated by the three models. The evaluator model is then asked to select the better response from each pair. By counting the number of times each model's output is preferred and comparing the distribution of win rates across different evaluators, we assess whether models exhibit self-preference biases.

As shown in the Table 10, the overall trends are consistent across the three different evaluators, with no substantial differences observed in the exact win rates. Notably, none of the evaluators exhibit a strong

Model	Basic Factual Accuracy					Domain Knowledge Accurac				
	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human
DeepSeek R1	-	0.51	0.66	0.68	0.59	-	0.6	0.59	0.58	0.57
GPT-4o	0.51	-	0.57	0.56	0.59	0.6	-	0.59	0.62	0.56
QwQ-Plus	0.66	0.57	-	0.62	0.63	0.59	0.59	-	0.64	0.64
DeepSeek V3	0.68	0.56	0.62	-	0.54	0.58	0.62	0.64	-	0.54
Human	0.59	0.59	**0.63**	0.54	-	0.57	0.56	**0.64**	0.54	-
	Reasoning Process Rigor					Error Identification & Correction Precision				
	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human
DeepSeek R1	-	0.55	0.59	0.69	0.62	-	0.53	0.56	0.54	0.52
GPT-4o	0.55	-	0.57	0.6	0.55	0.53	-	0.55	0.58	0.59
QwQ-Plus	0.59	0.57	-	0.57	0.64	0.56	0.55	-	0.68	0.67
DeepSeek V3	0.69	0.6	0.57	-	0.65	0.54	0.58	0.68	-	0.66
Human	0.62	0.55	0.64	**0.65**	-	0.52	0.59	**0.67**	0.66	-

Table 8: **Kendall's** W between different evaluation models and human evaluation in Content Accuracy.

Model	Clarity, Simplicity & Inspiration					Motivation, Guidance & Positive Feedback				
	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human
DeepSeek R1	-	0.56	0.56	0.54	0.57	-	0.54	0.58	0.59	0.67
GPT-4o	0.56	-	0.57	0.52	0.51	0.54	-	0.58	0.57	0.58
QwQ-Plus	0.56	0.57	-	0.56	0.56	0.58	0.58	-	0.53	0.61
DeepSeek V3	0.54	0.52	0.56	-	0.56	0.59	0.57	0.53	-	0.54
Human	**0.57**	0.51	0.56	0.56	-	**0.67**	0.58	0.61	0.54	-
	Personalization, Adaptation & Learning Support					Higher-Order Thinking & Skill Development				
	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human	DeepSeek R1	GPT-4o	QwQ-Plus	DeepSeek V3	Human
DeepSeek R1	-	0.61	0.71	0.74	0.74	-	0.58	0.69	0.69	0.68
GPT-4o	0.61	-	0.56	0.62	0.59	0.58	-	0.62	0.6	0.6
QwQ-Plus	0.71	0.56	-	0.71	0.71	0.69	0.62	-	0.68	0.66
DeepSeek V3	0.74	0.62	0.71	-	0.72	0.69	0.6	0.68	-	0.68
Human	**0.74**	0.59	0.71	0.72	-	**0.68**	0.6	0.66	**0.68**	-

Table 9: **Kendall's** W between different evaluation models and human evaluation in Pedagogical Effectiveness.

preference for the responses they themselves generated. These results suggest that model self-preference is not a significant concern, thereby granting us greater flexibility in the choice of evaluation models.

I.2 Discrimination between normal model and reasoning model

After ruling out self-preference biases, we further investigate whether models of different types—such as reasoning models and normal models—exhibit biases toward each other. To this end, we design an experiment in which two external models, QwQ (a reasoning model) and GPT-4o (a normal model), are used to evaluate the responses generated by DeepSeek V3 and DeepSeek R1. The evaluation follows the same pairwise comparison protocol as in the self-preference setting: for each question, the evaluator selects the better response from a pair. By comparing the evaluation outcomes across the two evaluators, we aim to assess potential inter-model biases.

As shown in Table 11, the reasoning model exhibits a clear preference for responses generated by other reasoning models, whereas the general-purpose model demonstrates a more balanced evaluation. Notably, the reasoning model's bias is substantial, with a win ratio as skewed as 9 to 1 in favor of reasoning models. These results highlight the importance of incorporating both reasoning and normal evaluators in the assessment process to mitigate evaluation bias. Relying on only one type of model may lead to unfair or distorted conclusions.

I.3 Discrimination between different positions

When using pairwise evaluation, a natural concern arises as to whether evaluators might exhibit positional bias—i.e., favoring responses based on their order of presentation. To control for this factor, we randomly sample 200 instances from the evaluation dataset and present them to the evaluator in both the original and reversed order. For each comparison, we record the position of the selected response (e.g., incrementing the Former count if the first response is chosen, and Latter if the second is chosen). This allows us to

Evaluator	Win times
Qwen Max	Qwen Max(68) < DeepSeek V3(101) < DeepSeek R1(193)
DeepSeek V3	Qwen Max(82) < DeepSeek V3(86) < DeepSeek R1(196)
DeepSeek R1	Qwen Max(68) < DeepSeek V3(110) < DeepSeek R1(186)

Table 10: To investigate whether models exhibit a preference for their own generations, we employ each data-generating models as evaluators to assess the datasets they produced. Specifically, we construct pairwise comparisons by selecting answers generated by two different models at a time. The evaluator is then asked to choose the better response from each pair. By aggregating the number of times each model's outputs are preferred, we examine potential self-preference biases.

Evaluator	Gen_Models	Win_times
Normal	Normal	442
	Reasoning	399
Reasoning	Reasoning	914
	Normal	100

Table 11: Comparison of evaluation results between different model types.

assess whether response position systematically influences evaluation outcomes.

Evaluator	Order	Former_win	Latter_win
GPT-4o	Normal	**873**	117
	Reverse	**742**	248
QwQ	Normal	605	385
	Reverse	387	603

Table 12: The table presents the evaluation results after reversing the order of the responses, allowing us to examine whether the evaluator exhibits any positional bias.

The experimental results reveal that GPT-4o exhibits a notable positional bias, with the former response being selected significantly more often than the latter, even after the order is reversed. As shown in Table 12, when the evaluated responses are identical aside from their order, the former response is chosen at a disproportionately high rate—by a factor of 3 to 7—compared to the latter. In contrast, QwQ demonstrates a more balanced evaluation, with selection counts remaining consistent before and after the reversal, indicating minimal positional bias.

J Extra Results

We present the additional experimental results in this section, including the evaluation results of the five response models by the three additional evaluators: GPT-4o, DeepSeek R1, and QwQ-Plus. The evaluation is conducted across two dimensions: metric-level and task-level assessments. The task-level score is the average score of the metrics under each task. The detailed results can be found in Table 15 and Table 14.

K Distillation Training Setting

After obtaining human and model evaluation results on sample data, we can optimize the selection of model-generated data based on evaluation scores to maximize data quality and efficiency.

We propose two data selection strategies. The first strategy involves selecting the best generation model within each scenario. The specific process includes calculating the average scores of all evaluation metrics

Scenario	Category Dimensions	All Data			For Training		
		Chinese	English	Total	Chinese	English	Total
Problem Solving	Duration*Difficulty*Subject	1,306	1,328	2,634	1,272	1,284	2,556
Error Correction	Duration*Difficulty*Subject	620	1,350	1,970	603	1,334	1,937
Idea Provision	Duration*Difficulty*Subject	1,342	1,350	2,692	1,300	1,322	2,622
Personalized Learning Support	Duration*Subject	348	561	909	67	435	502
Emotional Support	Duration*Anxiety Level	1,344	1,074	2,418	1,331	1,059	2,390
Question Generation	Duration*Difficulty*Subject	1,358	1,338	2,696	1,331	1,322	2,653
Automatic Grading	Duration*Difficulty*Subject	931	1,073	2,004	912	1,058	1,970
Teaching Material Generation	Duration*Difficulty*Subject	1,324	1,347	2,671	1,306	1,255	2,561
Personalized Content Creation	Duration*Subject	568	259	827	557	235	792
Total		9,141	9,680	18,821	8,679	9,304	17,983

Table 13: The number of scenarios in the dataset and the specific quantities for each scenario.

Evaluator	Model	Q&A	PLS	EC	IP	AG	TMG	ES	QG	PCC	Average
	DeepSeek R1	**9.81**	**9.83**	**9.05**	**9.11**	7.74	**9.46**	**9.71**	**9.22**	**9.73**	**9.29**
	DeepSeek V3	9.67	9.12	8.97	8.82	**8.32**	9.31	9.34	8.65	9.23	9.05
DeepSeek R1	Qwen Max	9.07	9.11	8.86	8.84	7.99	9.15	9.40	8.89	9.29	8.96
	Qwen2.5-14B-Instruct	8.94	8.79	8.68	8.23	7.83	9.06	8.52	8.35	8.80	8.58
	Qwen2.5-7B-Instruct	8.34	9.01	8.64	8.16	6.64	9.33	8.75	8.23	9.06	8.46
	DeepSeek R1	9.49	**9.65**	**9.27**	**8.75**	**7.27**	**9.45**	**9.38**	**9.33**	**9.71**	**9.14**
	DeepSeek V3	**9.68**	9.04	9.14	8.53	7.05	9.34	9.00	9.06	8.92	8.86
DeepSeek V3	Qwen Max	9.18	8.88	9.06	8.52	7.23	9.24	9.04	9.05	9.29	8.83
	Qwen2.5-14B-Instruct	9.07	8.72	8.97	8.30	6.77	9.21	8.74	9.02	8.80	8.62
	Qwen2.5-7B-Instruct	9.15	9.07	9.01	8.47	6.44	9.21	8.85	8.69	9.00	8.65
	DeepSeek R1	9.32	**9.38**	9.05	8.78	8.51	**9.25**	**9.15**	8.98	**9.08**	**9.06**
	DeepSeek V3	9.22	9.15	**9.14**	8.77	8.54	9.12	9.05	**9.00**	8.95	8.99
GPT-4o	Qwen Max	**9.50**	9.17	9.01	8.69	**8.70**	8.99	8.96	8.92	9.05	8.99
	Qwen2.5-14B-Instruct	9.34	9.25	8.92	8.51	8.11	8.99	9.11	8.77	8.82	8.87
	Qwen2.5-7B-Instruct	9.22	9.17	8.92	**8.84**	8.04	9.05	9.00	8.62	8.94	8.87
	DeepSeek R1	**9.85**	**9.87**	**9.24**	**9.05**	**8.78**	**9.75**	**9.85**	**9.09**	**9.88**	**9.49**
	DeepSeek V3	9.59	9.43	9.06	8.66	8.18	9.29	9.66	8.47	9.24	9.06
QwQ-Plus	Qwen Max	9.90	9.25	9.03	8.78	8.11	9.54	9.56	8.79	9.70	9.18
	Qwen2.5-14B-Instruct	9.83	9.21	9.05	8.23	7.88	9.22	9.45	8.48	9.02	8.94
	Qwen2.5-7B-Instruct	9.02	9.28	8.79	8.82	7.16	9.33	9.31	7.98	9.35	8.78
	DeepSeek R1	7.17	**9.11**	**8.71**	**8.80**	**8.42**	**8.86**	**9.15**	**8.79**	**9.35**	**8.71**
	DeepSeek V3	7.45	8.12	8.16	8.17	7.84	7.56	8.08	8.01	7.03	7.82
Human	Qwen Max	**7.72**	7.94	8.21	8.15	7.89	7.99	7.85	8.39	8.42	8.06
	Qwen2.5-14B-Instruct	7.66	7.38	7.92	7.56	7.55	7.84	7.31	7.91	7.36	7.61
	Qwen2.5-7B-Instruct	6.78	7.63	7.93	7.74	6.79	7.86	7.79	7.55	7.42	7.50

Table 14: Scenario-Level Average Scores Evaluated by Different Evaluators. Max values in each column per evaluator are bolded. Full names of each scenarios can be found in Section 3.1.

from both human and model evaluators on the sample data, ranking the generation models in each scenario according to these average values, and then selecting samples generated by the best-performing model in each scenario from all distilled data.

The second strategy focuses on selecting optimal models for each evaluation metric. We calculate the average score of each generation model on individual metrics, rank them accordingly to identify the best model for each metric. During the final distilled data screening phase, if a piece of data was generated by a model that has been recognized as optimal in any evaluation metric, it will be included in the final finetuning dataset.

We use the following settings for model training: The learning rate is set to 1.0×10^{-5}, with a batch size of 1 per GPU device. Gradient accumulation is applied over 8 steps, resulting in an effective batch size of 8. For parameter updating, we employ full fine-tuning, where all model parameters are updated during training. All experiments are conducted on 4 NVIDIA A100 GPUs, each with 40GB of memory.

Evaluator	Model	BFA	CSI	CRSC	DKA	EICP	HOTS	IFTC	MGP	PAS	RPR	RTC	SEI	Average
DeepSeek R1	DeepSeek R1	9.55	**8.67**	**9.64**	**9.53**	8.66	**8.39**	**9.61**	7.30	**9.80**	**9.17**	**9.64**	**9.45**	**9.12**
	DeepSeek V3	**9.58**	8.47	9.48	9.30	**9.32**	7.53	9.39	**7.48**	8.92	9.05	9.32	9.10	8.91
	Qwen Max	9.42	8.49	9.46	9.24	9.09	7.67	9.25	7.44	8.97	8.62	9.34	9.05	8.84
	Qwen2.5-14B-Instruct	9.08	8.28	9.20	8.82	8.98	7.16	8.87	6.86	8.20	8.57	9.02	8.51	8.46
	Qwen2.5-7B-Instruct	8.73	8.22	9.00	9.00	8.30	7.27	8.72	6.61	8.68	8.05	9.23	8.55	8.36
DeepSeek V3	DeepSeek R1	9.51	**8.75**	**9.44**	**9.45**	**7.61**	**8.53**	**9.47**	7.76	**9.64**	**8.85**	**9.14**	**9.06**	**8.93**
	DeepSeek V3	**9.57**	8.61	9.25	9.27	7.23	7.98	9.21	7.56	8.94	8.76	9.00	8.59	8.66
	Qwen Max	9.38	8.53	9.12	9.23	7.43	7.99	9.16	**7.85**	9.05	8.57	9.00	8.61	8.66
	Qwen2.5-14B-Instruct	9.28	8.50	9.03	9.14	7.14	7.81	8.94	7.55	8.71	8.35	8.82	8.25	8.46
	Qwen2.5-7B-Instruct	9.27	8.55	9.08	9.12	6.77	7.86	8.96	7.05	8.95	8.42	8.82	8.53	8.44
GPT-4o	DeepSeek R1	9.48	**8.73**	**9.59**	**9.17**	9.05	**8.35**	9.13	8.45	**9.18**	**8.89**	**9.11**	**8.65**	**8.98**
	DeepSeek V3	9.54	8.72	9.51	9.05	**9.14**	8.05	**9.16**	**8.59**	8.95	8.75	9.02	8.63	8.93
	Qwen Max	**9.58**	8.65	9.43	8.83	9.07	8.08	9.14	8.56	8.97	8.89	8.95	8.64	8.90
	Qwen2.5-14B-Instruct	9.45	8.51	9.44	8.88	8.93	7.83	9.02	8.20	8.88	8.60	9.07	8.43	8.77
	Qwen2.5-7B-Instruct	9.45	8.57	9.38	8.85	8.59	8.00	9.01	8.20	8.85	8.65	9.02	**8.65**	8.77
QwQ-Plus	DeepSeek R1	**9.78**	**8.47**	**9.78**	**9.82**	**9.70**	**8.19**	**9.65**	**8.35**	**9.86**	**9.61**	**9.70**	**9.58**	**9.37**
	DeepSeek V3	9.42	8.25	9.57	9.09	9.52	7.22	9.36	7.62	9.23	9.23	9.39	9.32	8.93
	Qwen Max	9.64	8.39	9.59	9.47	9.30	7.48	9.45	7.68	9.39	9.10	9.48	9.36	9.03
	Qwen2.5-14B-Instruct	9.49	8.20	9.48	8.98	9.20	7.10	9.15	7.64	8.77	8.83	9.41	9.06	8.78
	Qwen2.5-7B-Instruct	9.08	8.10	9.31	8.98	8.91	7.02	9.03	7.18	9.09	8.61	9.30	9.33	8.66
Human	DeepSeek R1	**8.97**	**8.60**	**8.98**	**8.94**	**8.86**	**8.56**	**8.77**	**8.20**	**9.26**	**7.95**	**8.91**	**8.92**	**8.74**
	DeepSeek V3	8.77	7.77	8.40	7.89	8.11	7.25	8.10	7.70	7.42	7.03	7.80	7.47	7.89
	Qwen Max	8.81	8.01	8.52	8.27	8.23	7.59	8.10	7.70	7.89	7.31	8.09	7.74	8.02
	Qwen2.5-14B-Instruct	8.74	7.76	8.26	7.79	7.86	6.88	7.77	6.97	7.02	7.01	7.59	7.03	7.56
	Qwen2.5-7B-Instruct	8.49	7.63	8.04	7.82	7.45	6.93	7.65	7.05	7.38	5.90	7.82	7.35	7.46

Table 15: Metric-level average scores evaluated by different evaluators under various metrics. For simplicity, we use abbreviations for the metrics. Full names of each metric can be found in Table 2.

Model	BFA	CSI	CRSC	DKA	EICP	HOTS	IFTC	MGP	PAS	RPR	RTC	SEI	Average
DeepSeek R1	**9.21**	**8.80**	**9.26**	**9.41**	**9.02**	**8.77**	**9.10**	**8.07**	**9.41**	**8.13**	9.25	**8.94**	**8.95**
DeepSeek V3	9.01	7.05	8.73	8.16	8.18	6.87	8.38	7.43	7.53	7.05	9.21	7.66	7.94
Qwen Max	9.05	7.41	8.85	8.55	8.21	7.54	8.52	7.53	8.24	7.41	9.25	8.05	8.22
Qwen2.5-14B-Instruct	8.96	7.13	8.56	7.92	8.00	6.54	8.01	6.76	6.98	7.13	9.10	7.20	7.69
Qwen2.5-7B-Instruct	8.81	6.14	8.37	8.01	7.33	6.77	7.99	6.94	7.55	8.09	**9.26**	7.60	7.74

Table 16: Metric-Level Human Evaluation Results over Three Rounds

Model	Q&A	PLS	EC	IP	AG	TMG	ES	QG	PCC	Average
DeepSeek R1	7.65	**9.56**	**8.83**	**9.05**	**8.40**	**9.09**	**9.30**	**8.91**	**9.42**	**8.91**
DeepSeek V3	7.58	7.37	8.32	8.34	7.79	8.01	8.35	8.17	7.82	7.97
Qwen Max	**7.95**	8.79	8.27	8.43	7.78	8.34	8.37	8.69	8.21	8.31
Qwen2.5-14B-Instruct	7.93	7.44	7.89	7.79	7.58	7.92	7.60	8.16	7.21	7.72
Qwen2.5-7B-Instruct	7.04	7.64	8.07	8.16	6.77	8.21	8.12	7.73	7.53	7.70

Table 17: Task-Level Human Evaluation Results over Three Rounds