

Can Large Language Models Be Good Language Teachers?

Liqing Xu^{1*}, Qiwei Li^{2*}, Tianshuo Peng³, Zuchao Li^{2†}, Hai Zhao⁴, Ping Wang^{5†}

¹Shanghai University of International Business and Economics, Shanghai, China,

²School of Artificial Intelligence, Wuhan University, Wuhan, China,

³School of Computer Science, Wuhan University, Wuhan, China,

⁴School of Computer Science, Shanghai Jiao Tong University, Shanghai, China,

⁵School of Information Management, Wuhan University, Wuhan, China

24361024@suibe.edu.cn, qw-line@whu.edu.cn, pengts@whu.edu.cn,
zcli-charlie@whu.edu.cn, zhaohai@cs.sjtu.edu.cn, wangping@whu.edu.cn

Abstract

Large language models (LLMs) have achieved remarkable success across diverse domains. However, their potential as effective language teachers—particularly in complex pedagogical scenarios like teaching Chinese as a second language—remains inadequately assessed. To address this gap, we propose the first pedagogical competence benchmark for LLMs, rigorously evaluating their performance against international standards for Chinese language teachers. Our framework spans three core dimensions: (1) basic knowledge evaluation, covering 32 subtopics across five major categories; (2) international teacher examination, based on data collected from international Chinese teacher certification exams; and (3) teaching practice evaluation, where target LLMs summarize knowledge points and design instructional content for student models, followed by testing the student models to assess the LLM’s ability to distill and teach key concepts. We conduct a comprehensive evaluation of 13 latest multilingual and Chinese LLMs. While most models demonstrate promising pedagogical potential, there remains substantial room for improvement in their teaching capabilities. This study contributes to the development of AI-assisted language education tools capable of rivaling human teaching excellence. The benchmark dataset and evaluation scripts used in this study are publicly available at <https://github.com/Line-Kite/CLTE>.

1 Introduction

In recent years, large language models (LLMs) have witnessed remarkable progress. Models such as GPT-4 (Achiam et al., 2023), Llama

3 (Grattafiori et al., 2024), and Qwen 3 (Yang et al., 2025) have demonstrated extraordinary capabilities in natural language processing, covering a wide range of tasks from text generation to complex question-answering systems. These advancements not only signify a major leap in artificial intelligence technology but also hold great potential for various industries, including education. Benchmark tests play a crucial role in evaluating the performance of these LLMs. They provide a standardized way to measure the capabilities and limitations of different models, which is essential for both researchers to improve the models and users to select the most suitable ones for their specific tasks.

In the field of evaluating LLMs, a diverse array of benchmarks has emerged, catering to different aspects of model performance. For instance, MMLU (Hendrycks et al., 2020) and its extended version MMLU Pro (Wang et al., 2024b) assess models’ knowledge across multiple domains. GSM8K (Cobbe et al., 2021) focuses on mathematical reasoning, while HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) evaluate code generation capabilities. HellaSwag gauges models’ commonsense reasoning skills. In the Chinese context, benchmarks like C-EVAL (Huang et al., 2023) and CMMLU (Li et al., 2023) have been developed to specifically assess the knowledge and reasoning abilities of language models in Chinese language and various disciplines.

However, when it comes to assessing the language teaching capabilities of LLMs, especially in the context of teaching languages like Chinese as a second language, the existing benchmarks fall short. Although benchmarks like CMMLU and C-Eval contain certain language-related content, they have limitations. Firstly, their scopes are too broad, lacking a focused assessment of language teaching-specific skills. Secondly, they mainly test basic knowledge rather than effectively evaluating the practical teaching abilities that are crucial in

*Equal contribution.

†Corresponding author.

This work was supported by the National Natural Science Foundation of China (No. 62306216), the Technology Innovation Program of Hubei Province (No. 2024BAB043) and the National Social Science Fund of China (No. 24&ZD186).

real-world language teaching scenarios, such as the ability to design appropriate teaching plans, explain complex language knowledge in an understandable way, and conduct teaching evaluations.

To fill this gap, we propose the **Chinese Language Teaching Evaluation (CLTE)** benchmark. This benchmark is composed of three core dimensions. The first dimension is basic knowledge, which encompasses 32 sub-topics across five major categories, including linguistics, Chinese culture, and pedagogy. It aims to assess the fundamental knowledge base that a language teacher should possess. The second dimension is international teacher examination. It is based on data collected from international Chinese teacher certification exams, providing a more in-depth and comprehensive evaluation of the LLMs’ knowledge in the field of Chinese language teaching. The third dimension is teaching practice evaluation. In this part, the target LLMs are required to summarize knowledge points and design instructional content for a simulated student model. Then, the student model is tested to evaluate the LLM’s ability to distill key concepts and effectively teach them.

Using the CLTE benchmark, we conduct an extensive evaluation of 13 of the latest multilingual and Chinese LLMs. The results highlight that while these LLMs have demonstrated considerable potential, they still show significant space for improvement when compared to the performance of human specialists on the benchmark. This is especially evident in areas such as linguistic knowledge and teaching practice. This situation can be attributed to several factors. The training data of these models may not comprehensively cover the multifaceted scenarios of language teaching, and the current model architectures may not be optimally designed to address the unique needs of second-language teaching, such as understanding learners’ difficulties and formulating tailored teaching strategies. These insights underscore the importance of further research and development in enhancing LLMs’ language teaching abilities.

Our main contributions are as follows:

- We propose a specialized dataset for evaluating large language models’ capabilities as Chinese language teachers, addressing the unique needs of language teaching assessment.
- We introduce a novel evaluation framework that assesses the teaching abilities of large

models, marking the first attempt to systematically measure their effectiveness in language instruction.

- We analyze existing large language models and reveal significant potential for improvement in Chinese language education, particularly in practical teaching scenarios.

2 Related Work

The rapid advancement of large language models has reshaped natural language processing, with models like GPT series (Achiam et al., 2023; Hurst et al., 2024), DeepSeek series (Guo et al., 2025; Liu et al., 2024), o1 (Jaech et al., 2024), Qwen (Bai et al., 2023; Yang et al., 2024; Team, 2024; Yang et al., 2025), InternLM (Cai et al., 2024), and Llama (Meta AI, 2024; Grattafiori et al., 2024) demonstrating unprecedented capabilities in text generation, reasoning, and cross-domain knowledge integration. General-purpose LLMs such as GPT-4 (Achiam et al., 2023) and Llama 4 (Meta AI, 2024) excel in generating human-like text across diverse topics, while reasoning-oriented models like o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) focus on mathematical reasoning, code generation, and logical inference. These models have demonstrated versatility in various domains, from academic research to professional writing, but their potential in language teaching—particularly in pedagogical design and learner interaction—remains underexplored due to the lack of specialized evaluation frameworks.

Early benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Sarlin et al., 2020) focused on narrow natural language understanding tasks, such as sentiment analysis and textual entailment. However, as LLMs advanced to handle multi-domain knowledge and reasoning, more comprehensive benchmarks emerged. MMLU (Hendrycks et al., 2020) and its professional variant MMLU Pro (Wang et al., 2024b) evaluate models across 57+ subjects using choice questions, with MMLU Pro introducing 10-option questions to challenge advanced reasoning. For mathematical reasoning, GSM8K (Cobbe et al., 2021) provides 8.5K primary-level math problems, while MATH (Hendrycks et al., 2021) and MATH-500 (Lightman et al., 2023) test college-level algebra and calculus. Code generation benchmarks like HumanEval (Huang et al., 2023) and

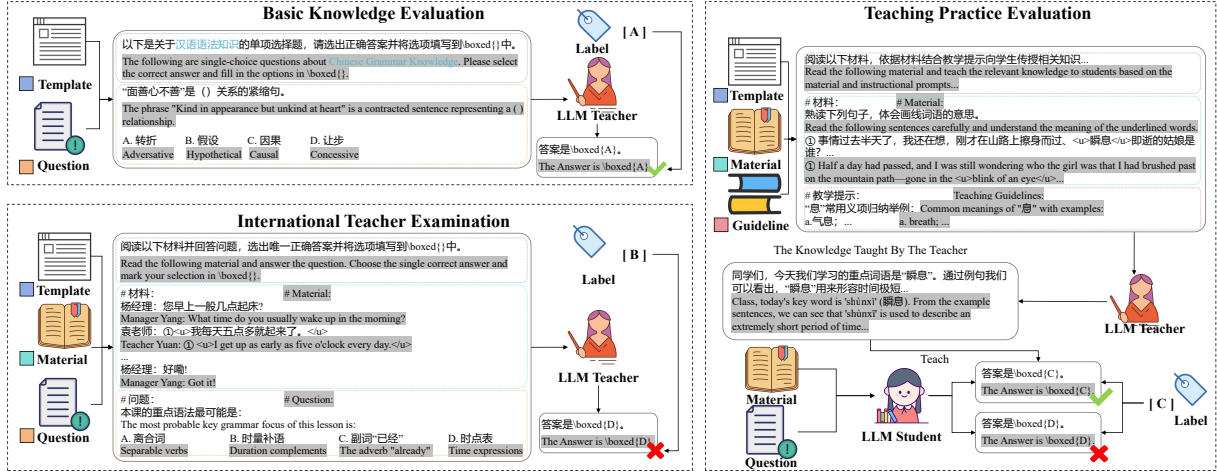


Figure 1: The overall framework of CLTE benchmark. Gray background indicates English translation, and the same applies to subsequent figures.

MBPP (Austin et al., 2021) assess functional correctness in Python programming, while common-sense reasoning is evaluated via HellaSwag (Zellers et al., 2019), ARC (Clark et al., 2018), and DROP (Dua et al., 2019). Specialized benchmarks like TruthfulQA (Lin et al., 2021) focus on factual accuracy to combat model hallucinations, and competitive math benchmarks like AIME 2024/2025 test high-level problem-solving skills. These benchmarks have been instrumental in identifying model strengths in knowledge recall and logical reasoning but are insufficient for evaluating teaching-related competencies.

In the Chinese context, benchmarks like C-Eval (Huang et al., 2023) and CMMLU (Li et al., 2023) have emerged to address language-specific evaluation. C-Eval covers 52 disciplines from Chinese standardized exams, while CMMLU expands to 67 topics, including China-specific domains like teacher certification and cultural knowledge. However, both primarily focus on theoretical knowledge assessment (e.g., linguistics and educational psychology) rather than teaching practice. Other Chinese benchmarks, such as MMCU (Zeng, 2023) (medicine and education), ACLUE (Zhang and Li, 2023) (ancient Chinese understanding), and AGIEval (Zhong et al., 2023) (cross-lingual exams), similarly prioritize knowledge retention over pedagogical application. For example, CMMLU’s “Chinese Pedagogy” subtests assess foundational concepts but do not include teaching practice, such as designing lesson plans or analyzing learner errors. M3KE (Liu et al., 2023), while comprehensive, lacks scenarios that require models to translate knowledge into teachable content or adapt to

diverse learner needs.

A critical limitation across these benchmarks is their focus on static knowledge assessment and logical reasoning, with minimal exploration of teaching practices. Most rely on single-turn question-answering formats, failing to simulate the dynamic interactions inherent in teaching—such as curriculum design, learner-tailored instruction, or formative assessment. For language teaching, which demands skills like content structuring, cultural adaptation, and learner feedback, existing benchmarks provide no framework to evaluate how models transform knowledge into effective instructional materials. The CLTE benchmark addresses this gap by focusing on teaching practice evaluation, where models must design instructional content and demonstrate its effectiveness—dimensions largely absent in current LLM assessment frameworks.

3 CLTE Benchmark

3.1 Overview

As illustrated in Figure 1, our comprehensive evaluation framework assesses large language models’ capabilities in Chinese language teaching through three key dimensions. The Basic Knowledge Evaluation examines foundational knowledge essential for international Chinese education, ensuring linguistic and pedagogical competence. Building upon this, the International Teacher Examination utilizes authentic teaching materials and questions from international teacher certification tests to evaluate fundamental teaching literacy. Most innovatively, the Teaching Practice Evaluation introduces a student-model-based approach to measure instruc-

tional effectiveness: LLMs act as teachers by generating educational content from teaching materials and guidelines, while their performance is quantified by comparing the student model’s pre- and post-instruction test scores, thereby objectively assessing real-world teaching outcomes. This multi-dimensional approach systematically bridges theoretical knowledge, professional standards, and practical teaching efficacy in evaluating LLMs for Chinese language education.

3.2 Benchmark Construction

3.2.1 Data Collection

Our three test tasks involve different types of data sources due to their distinct evaluation purposes. For the Basic Knowledge Evaluation, we primarily collect foundational knowledge questions from publicly available master’s entrance exam papers and mock tests for Teaching Chinese to Speakers of Other Languages (TCSOL). The International Teacher Examination utilizes real-world test questions from the official International Chinese Language Teacher Certification exams. As for the Teaching Practice Evaluation, which assesses practical teaching competence, we constructed the dataset by extracting material-question pairs from Chinese proficiency exam textbooks. To ensure data quality, we hire a TCSOL master’s graduate as an annotator, who manually gather materials, questions, and answers from open sources at a rate of 100 RMB per hour. This meticulous approach guarantees the relevance and accuracy of our evaluation benchmarks.

3.2.2 Annotation Process

We begin by structuring collected professional exam papers and textbook materials. For non-formatted documents like PDFs or images, we leverage the state-of-the-art open-source document parsing framework MinerU (Wang et al., 2024a) to convert them into well-formatted markdown files, ensuring compatibility with special symbols, underlines, and other formatting requirements in educational materials. To address inconsistencies in question-option formatting, we employ regex-based matching for initial organization, followed by manual refinement. To ensure data accuracy, the data is first reviewed and corrected by Chinese International Education specialists. A second reviewer—a computer science master’s graduate—then performs format verification and refinement. This dual-layer validation guarantees both

Task	Number
Basic Knowledge Evaluation	
- Linguistics	307
- Chinese Culture	321
- Pedagogy	163
- World Culture	192
- Cross-cultural Communication	217
International Teacher Examination	
- Materials	232
- Questions	1,044
Teaching Practice Evaluation	
- Materials	120
- Guidelines	120
- Questions	120

Table 1: Data composition of CLTE benchmark.

content precision and structural consistency.

3.2.3 Data Composition

The dataset employed in CLTE benchmark comprises a comprehensive collection of teaching guidelines, instructional materials, and assessment questions designed to evaluate various aspects of international Chinese language education. As illustrated in Table 1, the dataset consists of 120 teaching guidelines spanning fundamental knowledge, international teacher competencies, and teaching practices, along with 352 instructional materials and a total of 2,364 questions. The data is organized into three distinct evaluation tasks, each targeting specific dimensions of pedagogical expertise and model performance.

Basic Knowledge Evaluation focuses on assessing foundational knowledge in Chinese international education, covering five core domains: linguistics, Chinese culture, pedagogy, world culture, and cross-cultural communication. As Figure 2 shown, this task includes 1,200 basic questions, systematically distributed across 32 subdomains.

International Teacher Examination is constructed from authentic assessment materials used in international teacher certification tests. Each data instance consists of an instructional passage accompanied by 2 to 10 single-choice questions. Unlike the Basic Knowledge Evaluation, this task requires models to analyze real-world teaching scenarios and demonstrate integrated linguistic and pedagogical reasoning, thereby better reflecting their practical educational capabilities.

Teaching Practice Evaluation is constructed from



Figure 2: Data classification and distribution of Basic Knowledge Evaluation.

120 teaching materials and guidelines extracted from Chinese proficiency test instructor manuals, along with associated single-choice questions. The questions, materials, and guidelines are interlinked, with each question assessing the knowledge points emphasized in the guidelines. Notably, unlike the previous tasks, the questions in this task are designed for students learning Chinese rather than for teacher evaluation, offering a distinct perspective on the model’s applicability in instructional settings. The data sample analysis of each task can be found in Appendix A.

3.3 Evaluation Criteria

To assess the model’s proficiency in tasks that evaluate knowledge mastery, such as Basic Knowledge Evaluation and International Teacher Examination, we employ a knowledge-based assessment framework. This approach utilizes instruction-answer matching, where the model’s responses are systematically compared against predefined templates to gauge its grasp of foundational and comprehensive knowledge. Additionally, to evaluate the model’s pedagogical capabilities, we introduce an innovative teaching practice assessment methodology. This involves analyzing the performance improvement of a student model before and after interaction with the target model, thereby objectively measuring the large language model’s effectiveness in language instruction. This dual-assessment strategy ensures a rigorous and multi-dimensional evaluation of both knowledge retention and teaching

aptitude.

3.3.1 Knowledge-based Evaluation

To enhance the alignment between predicted answers and single-choice questions, we employed prompt engineering to guide model generation. Specifically, we designed tailored instruction templates for standard single-choice questions and context-based single-choice questions (see Appendix B for details). These templates, combined with the provided materials and questions, were used to prompt the large language model to generate responses in a structured format (denoted as `\boxed{option}`). The model’s output was then matched against the ground truth to evaluate correctness. The final performance was quantified by calculating the average accuracy score across all questions. Instances where the model failed to produce a matching response were automatically classified as incorrect. This approach ensured systematic and reproducible assessment of the model’s knowledge-based reasoning capabilities.

3.3.2 Teaching Practice Evaluation

The Teaching Practice Evaluation task aims to assess the pedagogical effectiveness of large language models by evaluating their ability to enhance a student model’s performance through simulated teaching interactions. To simulate this process, we select an early-stage LLM with relatively weak linguistic and knowledge capabilities as the student model M_s . Specifically, we employ early-stage LLM as M_s and evaluate its baseline performance s_{base} on single-choice questions from a standardized knowledge assessment framework. This initial assessment provides a reference point for measuring the impact of subsequent instructional interventions.

To address the limited instruction-following ability of early-stage models, we construct a specialized fine-tuning dataset derived from 800 non-linguistic discipline-specific questions in the CMMLU dataset. This dataset is used to refine M_s ’s output format stability, ensuring consistent and structured responses during evaluation. The fine-tuning process mitigates formatting inconsistencies that could otherwise obscure the model’s true knowledge retention and comprehension capabilities.

The teaching efficacy of the target instructor model M_t is evaluated by prompting it to generate pedagogical explanations based on given materials and teaching guidelines. M_s then answers the same

set of questions while having access to M_t 's instructional output, yielding an updated score $s_{knowledge}$. The difference between s_{base} and $s_{knowledge}$ serves as a quantitative measure of M_t 's teaching effectiveness, reflecting its ability to convey knowledge and improve the student model's performance. This comparative approach isolates the impact of instructional quality from inherent model capabilities.

4 Experiments

4.1 Experiments Setup

Baselines. We selected the latest versions of classic Chinese models and several high-performance multilingual models, including DeepSeek-V3 (Liu et al., 2024), Qwen3-8B (Yang et al., 2025), Qwen2.5-7B-Instruct (Team, 2024), InternLM3-8B-Instruct (Cai et al., 2024), ChatGLM4-9B-Chat (GLM et al., 2024), Yi-1.5-9B-Chat (Young et al., 2024), GPT-4 (Achiam et al., 2023), GPT-4o-mini (Hurst et al., 2024), GPT-3.5-Turbo (Achiam et al., 2023), Claude-3.5-Haiku (Anthropic, 2022), and Gemini-2.0-Flash (Gemini et al., 2023). Additionally, we evaluated some reasoning-focused models, including DeepSeek-R1 (Guo et al., 2025), o1-mini (Jaech et al., 2024), and Qwen3-8B (Yang et al., 2025).

Model Settings. The local model's max new tokens for inference is set to 4096. All other hyperparameters remain at their default values to ensure stable generation. For local testing, the model is deployed on a single NVIDIA RTX 3090 GPU.

Fine-tuning Settings. To enhance the credibility and robustness of the experiments, we select Qwen-1.8B-Chat (Bai et al., 2023), Qwen-7B-Chat, Qwen-14B-Chat, Yi-6B-Chat (AI et al., 2024), InternLM2-Chat-7B (Cai et al., 2024) as the student models and use LoRA for parameter adjustments. We use a single NVIDIA RTX 3090 GPU to fine-tune the model and batch size is set to 1. For LoRA, we set $r = 16$, $\alpha = 32$, LoRA dropout to 0.05.

4.2 Main Results

The main experimental results are presented in Table 2. As shown, the comprehensive scores of most conversational AI models remain around the passing threshold of 0.6, including both smaller Chinese-specific chat models and larger multilingual models. In comparison, reasoning-oriented models designed for complex problem-solving

demonstrate relatively better performance. However, significant room for improvement remains, as even the top-performing model (DeepSeek-R1) achieves only a 0.778 average score. These findings highlight substantial gaps in current large language models' capabilities for Chinese language instruction, suggesting the need for further advancements in this domain. The results collectively indicate that while some progress has been made, existing systems still fall short of satisfactory performance levels for educational applications.

4.3 Basic Knowledge Evaluation

From the perspective of subtasks, the Basic Knowledge Evaluation task—designed to assess fundamental knowledge mastery—shows relatively better performance across most models, reflecting their strong memorization capabilities. Specifically, DeepSeek's V3 and R1 versions achieved scores of 0.855 and 0.869, respectively. As the latest model, Qwen3-8B also demonstrates competitive results in Chinese language education-related knowledge retention. This trend highlights the robust knowledge retention abilities of current large language models.

In Table 3, we present the performance of various models across different domains in the fundamental knowledge test. DeepSeek-R1 consistently achieves the best results in all domains, followed by DeepSeek-V3 and Gemini-2.0-Flash. Overall, most large language models demonstrate strong performance in Chinese Culture, Pedagogy, World Culture, and Cross-cultural Communication, while showing relatively weaker results in Linguistics, which provides valuable guidance for future enhancements in language teacher models. Notably, the thinking version of Qwen3-8B underperforms its standard conversational counterpart. Upon inspection, we found that the thoughtful Qwen3-8B frequently repeats its reasoning process, leading to excessively long outputs that get truncated. We have statistically analyzed the responses that correctly used the `\boxed{}` format. Since it fails to generate the expected `\boxed{}` format, the matching accuracy of the thinking version of Qwen3-8B (0.801) is significantly lower than that of the chat version (0.998). Results for more specific field can be found in Appendix C.

4.4 International Teacher Examination

In the more challenging and comprehensive International Teacher Examination, most large language

Model Type	Model	BKE	ITE	TPE	AVG
Chat	Yi-1.5-9B-Chat	0.073	0.038	0.669	0.260
	GPT-3.5-Turbo	0.354	0.215	0.669	0.413
	Qwen2.5-7B-Instruct	0.490	0.404	0.665	0.520
	InternLM3-8B-Instruct	0.418	0.523	0.687	0.542
	ChatGLM4-9B-Chat	0.539	0.432	0.676	0.549
	GPT-4	0.666	0.549	0.681	0.632
	Gemini-2.0-Flash	0.756	0.674	0.674	0.701
	GPT-4o-mini	0.638	0.576	0.677	0.630
	Claude-3.5-Haiku	0.679	0.598	0.685	0.654
	Qwen3-8B	0.717	0.605	0.691	0.671
	DeepSeek-V3	<u>0.855</u>	<u>0.765</u>	0.694	<u>0.771</u>
Think	o1-mini	0.661	0.655	0.701	0.672
	Qwen3-8B	0.657	0.578	0.678	0.637
	DeepSeek-R1	0.869	0.767	<u>0.699</u>	0.778

Table 2: Main results. The results of BKE and ITE are obtained by taking the average of three experiments. The results of TPE are obtained by taking the average of five experiments. BKE represents Basic Knowledge Evaluation. ITE represents International Teacher Examination. TPE represents Teaching Practice Evaluation. AVG represents the average result. The best results are highlighted in bold, and the second highest are indicated by underlining. The same applies to subsequent tables.

Model Type	Model	Linguistics	Chinese Culture	Pedagogy	World Culture	Cross-cultural Communication	AVG
Chat	Yi-1.5-9B-Chat	0.094	0.081	0.074	0.057	0.046	0.073
	GPT-3.5-Turbo	0.256	0.355	0.458	0.385	0.387	0.354
	InternLM3-8B-Instruct	0.492	0.333	0.497	0.365	0.424	0.418
	Qwen2.5-7B-Instruct	0.301	0.562	0.603	0.589	0.481	0.490
	ChatGLM4-9B-Chat	0.302	0.559	0.779	0.552	0.656	0.539
	GPT-4o-mini	0.471	0.609	0.814	0.705	0.724	0.638
	GPT-4	0.530	0.629	0.808	0.729	0.750	0.666
	Gemini-2.0-Flash	0.638	0.786	0.853	0.780	0.785	0.756
	Claude-3.5-Haiku	0.510	0.671	0.843	0.750	0.743	0.679
	Qwen3-8B	0.565	0.787	0.822	0.722	0.743	0.717
	DeepSeek-V3	<u>0.818</u>	<u>0.870</u>	<u>0.892</u>	<u>0.880</u>	<u>0.834</u>	<u>0.855</u>
Think	o1-mini	0.544	0.607	0.828	0.703	0.742	0.661
	Qwen3-8B	0.450	0.720	0.828	0.688	0.700	0.657
	DeepSeek-R1	0.843	0.875	0.920	0.886	0.844	0.869

Table 3: Different fields results in Basic Knowledge Evaluation.

models exhibited performance declines. However, DeepSeek’s R1 (0.767) and V3 (0.765) maintain their leading positions, ranking first and second. Notably, InternLM3-8B-Instruct and o1-mini perform better in this comprehensive teacher assessment than in the basic knowledge test. We think

this may reflect their relatively stronger capacity for synthesizing and applying knowledge across contexts.

4.5 Teaching Practice Evaluation

As shown in Table 4, in the teaching practice evaluation, the average baseline performance among the

Model	Qwen-1.8B	Qwen-7B	Qwen-14B	Yi-6B	InternLM2-7B	AVG	AVG Gain
Base	0.557	0.607	0.725	0.787	0.617	0.658	0.000
Yi1.5-9B-Chat	0.598	0.628	0.700	0.762	0.658	0.669	0.011
GPT-3.5-Turbo	0.522	0.645	0.710	0.810	0.658	0.669	0.011
Qwen2.5-7B-Instruct	0.552	0.617	0.707	0.802	0.650	0.665	0.007
InternLM3-8B-Instruct	0.560	0.633	0.725	0.817	0.700	0.687	0.029
ChatGLM4-9B-Chat	0.548	0.622	0.720	0.798	0.692	0.676	0.018
GPT-4	0.588	0.623	0.727	0.793	0.675	0.681	0.023
Gemini-2-Flash	0.567	0.653	0.757	0.777	0.617	0.674	0.016
GPT-4o-mini	0.550	0.622	0.712	0.808	0.692	0.677	0.018
Claude-3.5-Haiku	0.550	0.647	0.715	0.812	0.700	0.685	0.026
Qwen3-8B(Chat)	0.562	0.643	0.725	0.827	0.700	0.691	0.033
DeepSeek-V3	0.585	<u>0.660</u>	0.740	0.810	0.675	0.694	0.036
o1-mini	<u>0.590</u>	0.655	<u>0.743</u>	0.807	<u>0.708</u>	0.701	0.042
Qwen3-8B(Think)	0.563	0.623	0.700	0.802	0.700	0.678	0.019
DeepSeek-R1	<u>0.590</u>	0.665	0.700	<u>0.823</u>	0.717	<u>0.699</u>	<u>0.041</u>

Table 4: Teaching Practice Evaluation on Different Student Models. Base indicates that there is no knowledge information from teachers. AVG Gain represents the average increase in score.

Model	PQ	LOG	FLU	KNW	AVG
Yi-1.5-9B-Chat	74.1	71.5	70.9	74.6	72.8
GPT-3.5-Turbo	78.4	76.5	76.4	79.1	77.6
Qwen2.5-7B-Instruct	75.6	74.0	73.3	76.2	74.8
InternLM3-8B-Instruct	75.6	74.0	73.3	76.2	74.8
ChatGLM4-9B-Chat	81.5	79.6	79.4	82.3	80.7
GPT-4	79.9	78.1	78.5	80.9	79.4
Gemini-2.0-Flash	77.4	74.9	76.0	78.3	76.6
GPT-4o-mini	88.6	85.0	86.4	88.7	87.2
Claude-3-5-Haiku	79.1	78.1	77.1	79.4	78.4
Qwen3-8B(Chat)	89.4	87.4	86.7	89.6	88.3
DeepSeek-V3	<u>91.0</u>	<u>89.0</u>	<u>88.1</u>	<u>91.5</u>	<u>89.9</u>
o1-mini	89.9	87.1	86.7	90.0	88.4
Qwen3-8B(Think)	87.0	86.3	84.5	88.1	86.5
DeepSeek-R1	95.0	92.8	91.0	94.9	93.4

Table 5: GPT-4’s evaluation scores on generated knowledge information of different models. PQ represents Pedagogical Quality. LOG represents Logicity. FLU represents Fluency. KNW represents Knowledgeability.

student models is measured at 0.658. After incorporating instructional knowledge from teacher models, the student model’s scores improved across the board. Interestingly, unlike knowledge mastery outcomes, teaching practice performance did not show a direct correlation with model version or scale. We further analyze the average length of knowledge content generated by each LLM as a "teacher", with results visualized in Figure 3. Notably, o1-mini achieved the best performance while also producing the second longest knowledge segments. In contrast, DeepSeek-V3 delivered competitive results with significantly shorter prompts. A case study (as shown in Appendix D) revealed that o1-mini tended to explain textbook concepts through natural

language descriptions, whereas DeepSeek-V3 condensed knowledge into structured, dictionary-like formats. Despite these stylistic differences, both models effectively identified and presented core educational content. This highlights a promising direction for LLMs in language teaching: adaptable knowledge delivery, whether through elaboration or compression, can enhance pedagogical outcomes.

4.6 Further Study

To further evaluate the characteristics of the knowledge generated by each model, we employ GPT-4 as an independent evaluator to rate the quality of teacher explanations across four key dimensions: Pedagogical Quality (PQ), Logicity (LOG), Fluency (FLU), and Knowledgeability (KNW). As shown in Table 5, the GPT-4’s scores align closely with our method’s results across multiple models, suggesting that score gains partially reflect genuine pedagogical improvement.

4.7 Human Experiments

We also conduct comparisons with human performance. We invited five non-specialists (not majoring in international Chinese education) and five experts (holding a master’s degree or higher in international Chinese education) to participate in the pilot testing of this questionnaire on a voluntary basis. Due to time and cost constraints, we randomly select 10% of the questions from the Basic Knowledge Evaluation and International Teacher Examination to form a question survey. To evaluate the Teaching Practice Evaluation, randomly

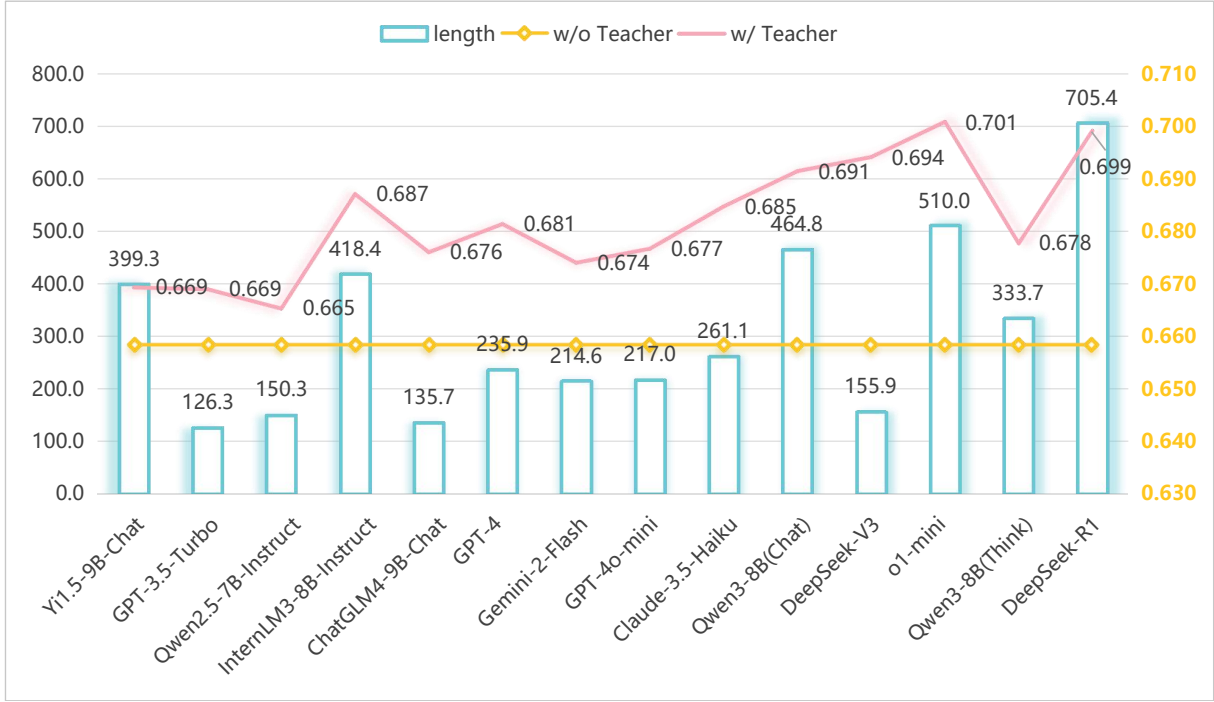


Figure 3: The average character length of the knowledge taught by the teacher.

Field	DeepSeek-R1	Laypeople	Expert
Linguistics	0.843	0.583	0.963
Chinese Culture	0.875	0.592	0.979
BKE Pedagogy	0.920	0.672	0.975
World Culture	0.886	0.536	0.986
Cross-cultural Communication	0.844	0.680	0.976
AVG	0.869	0.613	0.976
ITE	0.767	0.641	0.965
TPE	0.699	0.691	0.750

Table 6: Comparison of performance between DeepSeek-R1 and human.

select 30% of the materials and teaching guideline pairs for writing knowledge content. Their outputs are then tested using the student models, and the results are presented in Table 6.

The experimental results indicate that our best-performing model, DeepSeek-R1, outperforms non-specialists in both knowledge and comprehensive competence in Chinese language education but still lags behind experts. From a knowledge perspective, current large language models already surpass most non-specialists in Chinese language education. Leveraging their extensive knowledge bases, these models can effectively summarize key teach-

ing points, which contributes to the enhancement of instructional quality. While LLMs demonstrate great potential in Chinese language education, a noticeable gap remains compared to true professional educators.

5 Conclusion

This paper introduces the Chinese Language Teaching Evaluation (CLTE) benchmark, a framework designed to assess large language models as Chinese language teachers. It evaluates LLMs across three dimensions: basic knowledge (covering 32 sub-topics), international teacher exams, and teaching practice—where models summarize content, design lessons, and demonstrate effectiveness via simulated student performance. Evaluations of 13 leading multilingual and Chinese LLMs reveal that while they demonstrate significant potential, their performance in language teaching ability still requires improvement, with certain gaps remaining in areas such as linguistics and teaching practice. CLTE contributes the first dedicated teaching benchmark, a practice-oriented evaluation method, and identifies key areas for improvement. While models like DeepSeek-R1 and Qwen3 show potential, they still fall short of human teachers in expertise and adaptability. We will open-source the benchmark data and code under the CC BY 4.0 license.

Limitations

While our benchmark establishes foundational evaluation criteria for AI-driven language instruction, three strategic directions merit future exploration. First, the standardized testing paradigm could be enriched with conversational teaching simulations to better capture dynamic pedagogical interactions. Second, expanding the student model ecosystem across multiple capability tiers (from novice to advanced learners) would enable more nuanced assessment of instructional adaptability. Third, future work should extend the evaluation to incorporate a wider range of pedagogical exercise types—such as error correction, dialogic scaffolding, and adaptive feedback mechanisms—which remain untested in the current benchmark. These enhancements, particularly in light of our preliminary findings showing teaching effectiveness variations across knowledge complexity levels, would further bridge the gap between technical evaluation and authentic educational contexts.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, and 13 others. 2024. [Yi: Open foundation models by 01.ai](#). *Preprint*, arXiv:2403.04652.
- <https://www.anthropic.com/index/introducing-claude> Anthropic. 2022. [Claude](#).
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. [Internlm2 technical report](#). *Preprint*, arXiv:2403.17297.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.
- Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. [Gemini: a family of highly capable multimodal models](#). *ArXiv preprint*, abs/2312.11805.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu,

- Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- 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*.
- Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, and 1 others. 2023. M3ke: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2305.10263*.
- Meta AI. 2024. [Introducing llama 4: Advancing multi-modal intelligence](#). Accessed: 2025-05-20.
- 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.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liquan Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. [Mineru: An open-source solution for precise document content extraction](#). *Preprint*, arXiv:2409.18839.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.
- Hui Zeng. 2023. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*.
- Yixuan Zhang and Haonan Li. 2023. Can large language model comprehend ancient chinese? a preliminary test on aclue. *arXiv preprint arXiv:2310.09550*.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A Sample Analysis

Figure 4 illustrates the data sample for Basic Knowledge Evaluation, which is designed to assess the direct recall and application of knowledge. Each data point is a standalone unit consisting of a single multiple-choice question and its corresponding correct answer. This task evaluates foundational knowledge without the need for external context or interpretive guidance.

Question

“他在家吗？”中“在”的词性是（ ）。

In the sentence “他在家吗？”, the part of speech of “在” is ().

A. 连词 A. Conjunction
B. 介词 B. Preposition
C. 动词 C. Verb
D. 助词 D. Particle

Answer

C

Figure 4: A sample of Basic Knowledge Evaluation in CLTE.

Figure 5 shows the data sample for International Teacher Examination, which is crafted to evaluate a teacher’s experience in practical teaching scenarios. In this task, a single piece of source material serves as the common context for multiple related multiple-choice questions (a one-to-many relationship). This structure tests the ability to comprehend a complex scenario and derive various insights and assessments from it.

Material

玛丽：中村，你看，那件白毛衣怎么样？
Mary: Nakamura, look, how about that white sweater?
中村：挺好看的。不过，白的容易脏。这件蓝的怎么样？
Nakamura: It's quite nice. But white gets dirty easily. How about this blue one?
玛丽：这件毛衣的颜色有点儿深，我喜欢浅颜色的。
Mary: The color of this sweater is a bit too dark. I prefer lighter colors.
中村：那件黄的呢？
Nakamura: What about that yellow one?
玛丽：不错，挺漂亮！就买它吧。
Mary: Not bad, very pretty! Let's get that one.

Question

讲解“毛衣”一词最合适的方法是（ ）。

The most appropriate teaching method to explain the term “毛衣” is ().

A. 语素释义 A. Morpheme explanation
B. 直接对译 B. Direct translation
C. 近义词释义 C. Synonym explanation
D. 实物/图片展示 D. Real object/picture demonstration

Answer

D

Figure 5: A sample of International Teacher Examination in CLTE.

Figure 6 presents the data sample for Teaching Practice Evaluation. This task is built upon a direct and sequential correspondence between its four components: a source material, a teaching guidance prompt, a multiple-choice question (which is directly based on the preceding guidance), and its answer. This design creates a coherent instructional flow. The core objective is to test a teacher’s ability to utilize the provided pedagogical guidance and

source material to effectively extract and present knowledge.

Material

儒：由“儒”构成的词语，《大纲》中只有“儒家”一词，为六级词。
Ru: Among the words formed with ‘儒’, only ‘儒家’ (Confucianism) is included in the ‘Outline’, and it is classified as a Level 6 word.

Guideline

“儒”常用义项归纳举例：
A summary of common meanings of “儒” with examples:
a. 旧时指读书人：儒生、儒医、儒者、名儒、儒将
Historically refers to scholars: 儒生 (Scholar), 儒医 (Confucian Physician), 儒者 (Confucian Scholar), 名儒 (Eminent Confucian Scholar), 儒将 (Scholar-General).
b. 春秋时以孔子为代表的学派：儒家、儒教
The school of thought represented by Confucius during the Spring and Autumn Period: 儒家 (Confucianism), 儒教 (Confucian Doctrine).

Question

选词填空。Banked Cloze
儒雅 儒将 儒家 儒医
Elegant, Scholar-General, Confucian, Confucian Physician
①孔子是“（ ）”学派的创始人。
① Confucius is the founder of the “（ ）” school.
②他是一个以满腔热情投入事业的创造者，一个靠智慧与胆魄书写个人历史的硬汉，一个刚柔并济，处变不惊的（ ）。
② He is a creator who dedicates himself to his cause with full passion, a tough man who writes his personal history with wisdom and courage, and a（ ） who embodies softness within firmness and remains composed in the face of change.
③世医、（ ）、兽医……虽然都有一个医字，却万不可误会他们干的是同一行当。望文生义在这里只能引起误会。
③ Hereditary physicians,（ ），and veterinarians... Although they all share the character “医” (medical), it is a mistake to assume they practice the same profession. Taking the term literally can only lead to misunderstandings here.
④当年看过电影《佐罗》的很多女性朋友特别喜欢这部影片。除了佐罗英俊潇洒的形象外，配音演员童自荣华丽而充满（ ）、贵族气质的声音起了关键作用。
④ Many female friends who watched the movie “Zorro” back in the day particularly loved the film. Besides Zorro's handsome and dashing image, the voice of voice actor Tong Zirong, which was magnificent and brimming with（ ） and an aristocratic temperament, played a key role.
A. 儒家 儒将 儒医 儒雅 A. Confucian, Scholar-General, Confucian Physician, Elegant
B. 儒医 儒将 儒家 儒雅 B. Confucian Physician, Scholar-General, Confucian, Elegant
C. 儒雅 儒医 儒家 儒将 C. Elegant, Confucian Physician, Confucian, Scholar-General

Answer

A

Figure 6: A sample of Teaching Practice Evaluation in CLTE.

B Instruction Template

Figure 7 displays the instruction template for the Basic Knowledge Evaluation task. Questions are inserted into the <Question> placeholder, and contextual cues guide the model to place the answer within the \boxed{ }.

User Template

以下是关于{}的单项选择题，请选出正确答案并将选项填写到\boxed{ }中。
Here is a multiple-choice question about {}. Please select the correct answer and write the option in \boxed{ }.

<Question>

Figure 7: Template for Basic Knowledge Evaluation. The content with angle brackets in the template represents the dataset content to be filled in, and the same applies to subsequent figures.

Figure 8 presents the instruction template for the International Teacher Examination task. It guides the model to answer questions based on provided materials through contextual prompts, thereby eval-

uating the model’s experiential knowledge in teaching scenarios.

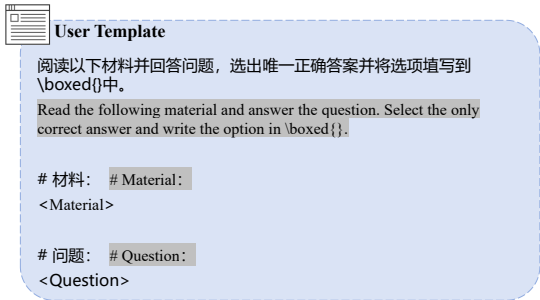


Figure 8: Template for International Teacher Examination.

The Teaching Practice Evaluation task consists of a three-round generation process. As shown in Figure 9, the first round requires the teacher model to generate relevant knowledge content by incorporating both the provided materials and teaching prompts. This generated knowledge will later be used in the question-answering process for the student model in the third round. The second round, illustrated in Figure 10, involves the student model answering questions based solely on the materials and the questions themselves, without any additional knowledge guidance. This stage serves to establish the baseline performance of the student model. In the third round, depicted in Figure 11, the student model answers questions using both the materials and the knowledge generated by the teacher model, thereby allowing an assessment of the performance improvement attributable to the teacher-generated knowledge.

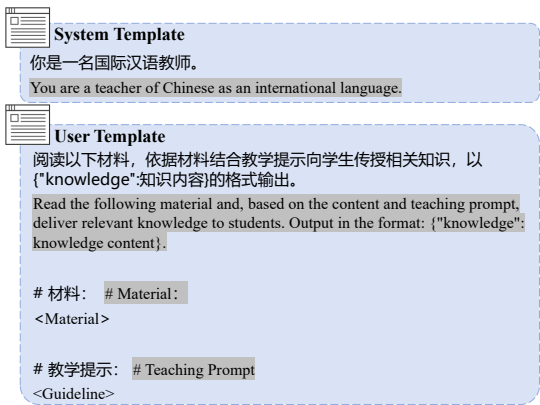


Figure 9: Teacher template for Teaching Practice Evaluation.

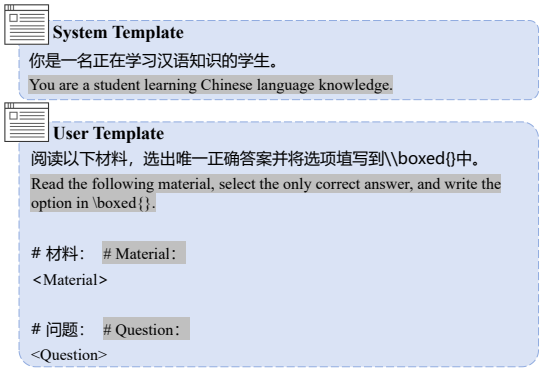


Figure 10: Student template without knowledge for Teaching Practice Evaluation.

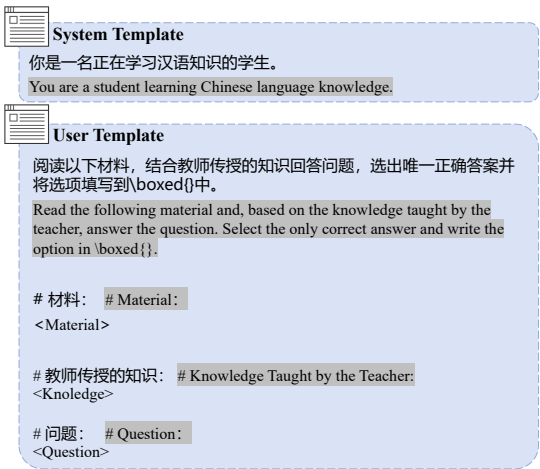


Figure 11: Student template with knowledge for Teaching Practice Evaluation.

C Results in Subfields

The results of different fields in CLTE are shown in Figure 12. It can be observed that most existing LLMs already perform well in areas such as Teacher Competencies, Modern Educational Technology, Educational Psychology, and Western Culture. However, there remains significant room for improvement in their mastery of knowledge related to the Chinese language, including Chinese Grammar, Chinese Rhetoric, Classical Chinese, Chinese Characters, Chinese Lexicon, and Chinese Phonetics. This provides valuable insights for the future development of LLMs in Chinese language education.

D Case Study on Knowledge

The different knowledge generated by o1-mini and DeepSeek-V3 is shown in Figure 13. o1-mini tends to produce descriptive introductions, while DeepSeek-V3 leans toward generating structured introductions.

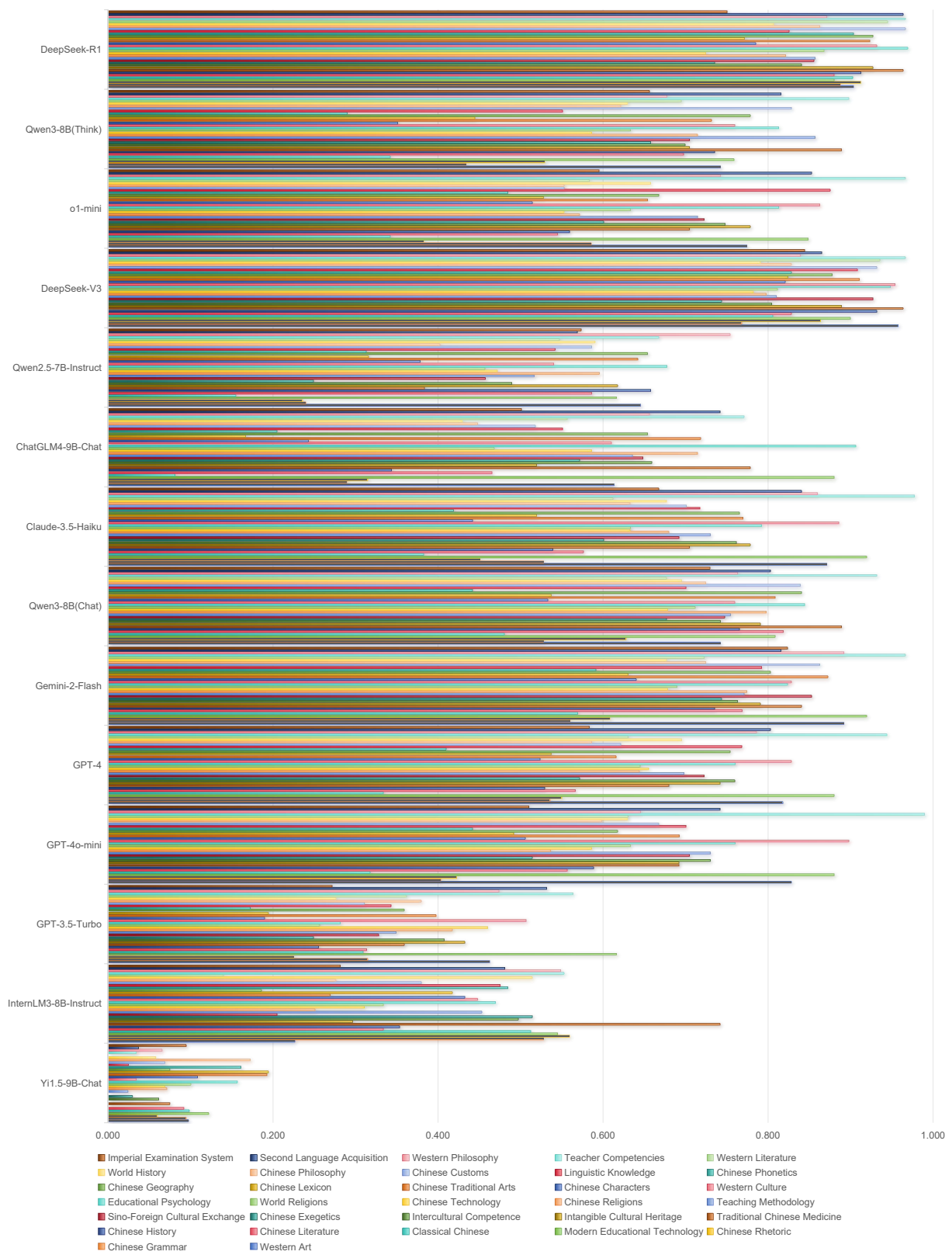


Figure 12: Samples in CLTE.



Material

儒：由“儒”构成的词语，《大纲》中只有“儒家”一词，为六级词。

Ru: Among the words formed with '儒', only '儒家' (Confucianism) is included in the 'Outline', and it is classified as a Level 6 word.



Guideline

“儒”常用义项归纳举例：A summary of common meanings of "儒" with examples:

a. 旧时指读书人：儒生、儒医、儒者、名儒、儒将

a. Historically refers to scholars: 儒生 (Scholar), 儒医 (Confucian Physician), 儒者 (Confucian Scholar), 名儒 (Eminent Confucian Scholar), 儒将 (Scholar-General).

b. 春秋时以孔子为代表的学派：儒家、儒教

b. The school of thought represented by Confucius during the Spring and Autumn Period: 儒家 (Confucianism), 儒教 (Confucian Doctrine).



Knowledge generated by o1-mini

```
{
  "character": "儒",
  "definitions": [{
    "meaning": "旧时指读书人", Historically refers to scholars or educated individuals
    "examples": ["儒生", "儒医", "儒者", "名儒", "儒将"],
      [Scholar, Confucian Physician, Confucian Scholar, Eminent Confucian Scholar, Scholar-General]
    "explanation": "这个义项中的'儒'都用来指代古代的知识分子或学者，特别是那些有学问、有修养的人。比如'儒将'就是指有学者风度的将帅。"
  In this sense, '儒' denotes ancient intellectuals or scholars, particularly those with knowledge and cultivation. For example, '儒将' refers to a military commander who possesses the refined qualities of a scholar.
  ...
}
```



Knowledge generated by DeepSeek-V3

‘儒’在汉语中主要有两种常用义项：The character "儒" in Chinese primarily has two common meanings:

1. 旧时指读书人。例如：儒生、儒医、儒者、名儒、儒将。这些词语通常用来称呼有学问的文人或特定职业背景的读书人。

1. Historically, it referred to scholars or educated individuals. For example: 儒生 (Scholar), 儒医 (Confucian Physician), 儒者 (Confucian Scholar), 名儒 (Eminent Confucian Scholar), 儒将 (Scholar-General). These terms typically denote learned literati or individuals with specific scholarly backgrounds.

2. 春秋时期以孔子为代表的学派。例如：儒家、儒教。儒家思想强调伦理道德、社会规范和教育的重要性，对中国古代社会和文化产生了深远影响。

2. It refers to the school of thought represented by Confucius during the Spring and Autumn Period. For example: 儒家 (Confucianism), 儒教 (Confucian Doctrine). Confucian philosophy emphasizes ethics, social norms, and the importance of education, exerting a profound influence on ancient Chinese society and culture.

需要注意的是，在《大纲》中‘儒家’一词被列为六级词汇，学习和掌握其含义及用法对于理解相关文化和历史知识具有重要意义。

It is worth noting that the term "儒家" (Confucianism) is classified as a Level 6 vocabulary item in the "Outline," and mastering its meaning and usage is essential for understanding related cultural and historical knowledge.

Figure 13: Case study of knowledge generated by o1-mini and DeepSeek-V3. Due to space constraints, the generated results from DeepSeek-V3 have been partially omitted.