Can Large Language Models Be Good Language Teachers?

Anonymous ACL submission

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, covering 32 subtopics across five major categories (linguistics, Chinese culture, pedagogy, etc.); (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 a student model, followed by testing the student model to assess the LLM's ability to distill and teach key concepts. We conduct a comprehensive evaluation of 13 latest multilingual and Chinese LLMs. The results reveal that most existing models struggle to achieve a 60% overall score, highlighting significant room for improvement. This study contributes to the development of AI-assisted language education tools capable of rivaling human teaching excellence.

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., 2024) 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 real - world language teaching scenarios, such as the ability to design appropriate teaching plans, explain complex language points 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 conducted an extensive evaluation of 13 of the latest multilingual and Chinese LLMs. The results highlight that while these models have made significant strides in general language processing, their performance in language teaching tasks reveals substantial room for improvement. Most models did not surpass an overall score of 60%, indicating that there are still considerable challenges to overcome in developing LLMs with proficient language teaching capabilities. 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 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 (LLMs) 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 (Massive Multitask Language Understanding) (Hendrycks et al., 2020) and its professional variant MMLU Pro (Wang et al., 2024) 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 MBPP (Austin et al., 2021) assess functional correctness in Python programming, while commonsense 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. As CMMLU and C-Eval highlight, even advanced models struggle with tasks requiring applied knowledge and pedagogical reasoning, underscoring the need for benchmarks that bridge theoretical knowledge and practical teaching capabilities. 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' (LLMs) 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 instructional 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 multidimensional 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 collected 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 hired a TCSOL master's graduate as an annotator, who manually gathered 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.

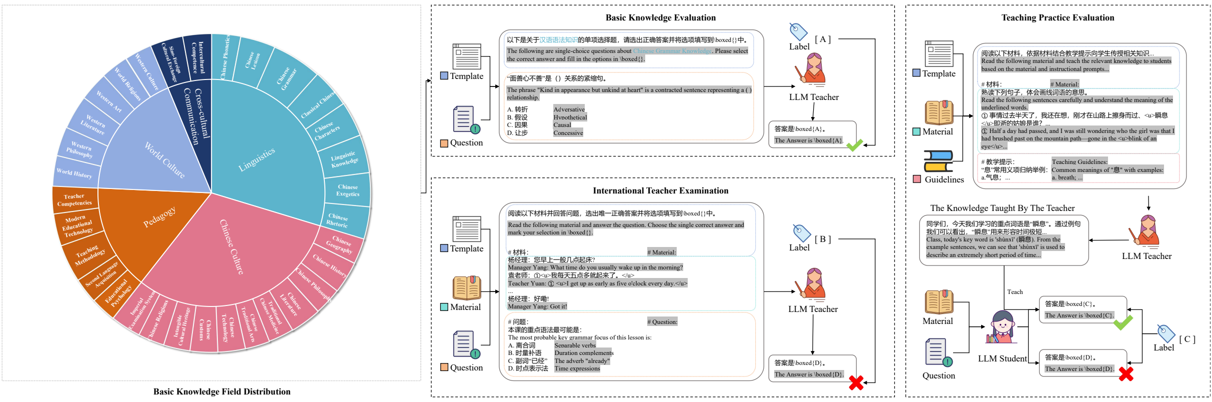

Figure 1: The overall framework of CLTE benchmark.

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 to convert them into well-formatted Markdown, 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, annotations are first performed and reviewed by Chinese International Education specialists, after which a second reviewer—a computer science master's graduate—conducts a final format verification. This dual-layer validation guarantees both content precision and structural consistency.

3.2.3 Data Composition

Task	Number
Basic Knowledge Evaluation	
- Linguistics	309
- Chinese Culture	322
- Pedagogy	157
- World Culture	188
- Cross-cultural Communication	65
International Teacher Examination	
- Materials	164
- Questions	742
Teaching Practice Evaluation	
- Materials	77
- Guidelines	77
- Questions	77

Table 1: Data composition of CLTE benchmark.

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 77 teaching guidelines spanning fundamental knowledge, international teacher competencies, and teaching practices, along with 241 instructional materials and a total of 1,860 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 1 shown, this task includes 1,041 basic questions, systematically distributed across 32 subdomains. Each subdomain contains a balanced number of questions, ranging from 26 to 53, ensuring a thorough and nuanced evaluation of the model's grasp of fine-grained knowledge.

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.

This task is constructed from 77 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 \box{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 (LLMs) 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 Qwen-7B-Chat (Bai et al., 2023) 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, 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), and Yi-1.5-9B-Chat (Young et al., 2024). We also tested several high-performance multilingual models, such as 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 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. We select Qwen-7B-Chat (Bai et al., 2023) as the student model 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 fail to reach 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 0.78 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 0.825 and 0.865, respectively. As a next-generation 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 Qwen3-8B. Overall, most large language models demonstrate strong performance in Chinese Culture, Pedagogy, and World Culture, while showing relatively weaker results in Linguistics and Cross-cultural Communication, 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. Since it fails to generate the expected \box{} format, its matching accuracy (0.783) is significantly lower than that of the standard version (0.997). Results for more specific field can be found in Appendix 5.

4.4 International Teacher Examination

In the more challenging and comprehensive International Teacher Examination, most large language models exhibited performance declines. However, DeepSeek's R1 (0.815) and V3 (0.767) maintained their leading positions, ranking first and second. Notably, InternLM3-8B-Instruct and o1-mini performed 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. From a linguistic perspective, Chinese-oriented LLMs generally outperformed multilingual models in evaluating Chinese language teaching expertise, highlighting their advantage in domain-specific knowledge mastery.

4.5 Teaching Practice Evaluation

In the teaching practice evaluation, the baseline performance of the student model was measured at 0.556. After incorporating instructional prompts from large language models (LLMs), 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 analyzed the average length of knowledge content generated by each LLM as a "teacher," with results visualized in Figure 2. Notably, o1-mini achieved the best performance while also producing the longest knowledge segments. In

Model Type	Language	Model	BKE	ITE	TPE	AVG
Chat	Chinese	Yi-1.5-9B-Chat	0.078	0.039	0.621	0.246
	Multilingual	GPT-3.5-Turbo	0.347	0.253	0.603	0.401
	Chinese	Qwen2.5-7B-Instruct	0.474	0.419	0.623	0.505
	Chinese	InternLM3-8B-Instruct	0.397	0.531	0.587	0.505
	Chinese	ChatGLM4-9B-Chat	0.492	0.472	0.623	0.529
	Multilingual	GPT-4	0.604	0.437	0.647	0.563
	Multilingual	Gemini-2.0-Flash	0.628	0.577	0.587	0.597
	Multilingual	GPT-4o-mini	0.592	0.561	0.597	0.583
	Multilingual	Claude-3-5-Haiku	0.632	0.588	0.590	0.603
	Chinese	Qwen3-8B	0.691	0.632	0.649	0.657
	Chinese	DeepSeek-V3	0.825	0.767	0.647	0.746
Think	Multilingual	o1-mini	0.588	0.629	**0.673**	0.630
	Chinese	Qwen3-8B	0.616	0.582	0.639	0.612
	Chinese	DeepSeek-R1	**0.865**	**0.815**	0.660	**0.780**

Table 2: Main results. The result was 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.

Model Type	Language	Model	Linguistics	Chinese Culture	Pedagogy	World Culture	Cross-cultural Communication	AVG
Chat	Chinese	Yi-1.5-9B-Chat	0.094	0.096	0.057	0.059	0.015	0.078
	Multilingual	GPT-3.5-Turbo	0.259	0.329	0.446	0.436	0.354	0.347
	Chinese	InternLM3-8B-Instruct	0.466	0.342	0.459	0.372	0.262	0.397
	Chinese	Qwen2.5-7B-Instruct	0.311	0.550	0.548	0.574	0.400	0.474
	Chinese	ChatGLM4-9B-Chat	0.288	0.556	0.732	0.537	0.431	0.492
	Multilingual	GPT-4o-mini	0.456	0.606	0.752	0.665	0.569	0.592
	Multilingual	GPT-4	0.476	0.640	0.688	0.686	0.600	0.604
	Multilingual	Gemini-2.0-Flash	0.537	0.668	0.711	0.712	0.500	0.628
	Multilingual	Claude-3-5-Haiku	0.505	0.637	0.790	0.707	0.615	0.632
	Chinese	Qwen3-8B	0.573	0.761	0.771	0.723	0.615	0.691
	Chinese	DeepSeek-V3	0.777	0.863	0.854	0.840	0.754	0.825
Think	Multilingual	o1-mini	0.502	0.553	0.752	0.670	0.538	0.588
	Chinese	Qwen3-8B	0.447	0.686	0.739	0.681	0.585	0.616
	Chinese	DeepSeek-R1	**0.864**	**0.876**	**0.866**	**0.856**	**0.831**	**0.865**

Table 3: Different fields results in Basic Knowledge Evaluation. AVG represents the average result. The best results are highlighted in bold, and the second highest are indicated by underlining.

contrast, DeepSeek-V3 delivered competitive results with significantly shorter prompts. A case study 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 Human Experiments

We also conduct comparisons with human performance. 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 178-question survey. This survey is distributed to 25 non-specialists (non-majors in international Chinese education) and 25 experts (master's degree holders or above in international Chinese education). To evaluate the Teaching Practice Evaluation, we recruit both ordinary participants and experts to write teaching materials based on 77 datasets. Their outputs are

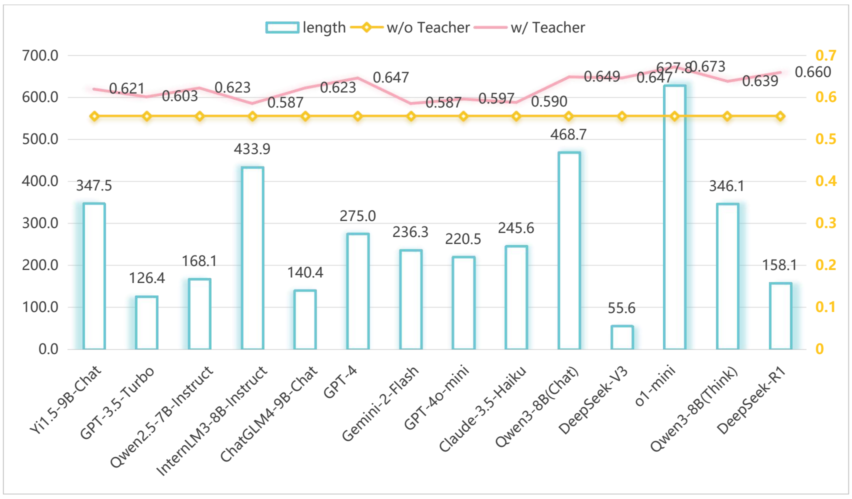

Figure 2: The average length of the knowledge taught by the teacher.

Field		DeepSeek-R1	Laypeople	Expert
BKE	Linguistics	0.864	0.600	0.987
	Chinese Culture	0.876	0.613	0.936
	Pedagogy	0.866	0.737	0.947
	World Culture	0.856	0.654	0.962
	Cross-cultural Communication	0.831	0.667	0.954
	AVG	0.865	0.6542	0.9572
ITE		0.815	0.559	0.949
TPE		0.662	0.610	0.779

Table 4: Comparison of performance between DeepSeek-R1 and human.

then tested using a student model, and the results are presented in Table 4.

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 (LLMs) already surpass most non-specialists in Chinese language education. Thanks to their vast knowledge base, they effectively summarize key teaching points, thereby improving instructional quality. While LLMs demonstrate great potential in Chinese language education, a noticeable gap remains compared to true professional educators.

5 Conclusion

This paper proposes the Chinese Language Teaching Evaluation (CLTE) benchmark, a specialized framework designed to assess large language models' (LLMs) capabilities as Chinese language teachers, addressing critical gaps in existing evaluation methods. The CLTE benchmark systematically evaluates LLMs across three core dimensions: basic knowledge (covering 32 sub-topics in linguistics, Chinese culture, and pedagogy), international teacher examination (leveraging certification exam data for in-depth knowledge assessment), and teaching practice evaluation. For the latter, LLMs must summarize knowledge points, design instructional content for a simulated student model, and demonstrate teaching effectiveness through student performance, establishing a novel paradigm for evaluating practical teaching skills. Through comprehensive evaluations of 13 state-of-the-art multilingual and Chinese LLMs, our results reveal that while these models show promising general language processing abilities, their performance in language teaching remains inadequate (mostly below 60% overall), highlighting substantial limitations in pedagogical adaptation, curriculum design, and learner-centered instruction. Our work makes three key contributions: introducing the first dedicated benchmark for language teaching assessment, developing a practice-oriented evaluation methodology with simulated teaching scenarios, and identifying critical improvement areas for LLMs in educational applications. Although current models like DeepSeek-R1 and Qwen3 exhibit remarkable potential as language teachers, they still fall far short of real human language teachers in pedagogical expertise, adaptive instruction, and contextual understanding, underscoring the need for fundamental advances before achieving authentic teaching competence.

Limitations

While our benchmark establishes foundational evaluation criteria for AI-driven language instruction, two 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 – a crucial next step given our preliminary findings showing teaching effectiveness variations across knowledge complexity levels. These enhancements 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*.

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, 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*.

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 multimodal 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*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. 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*.

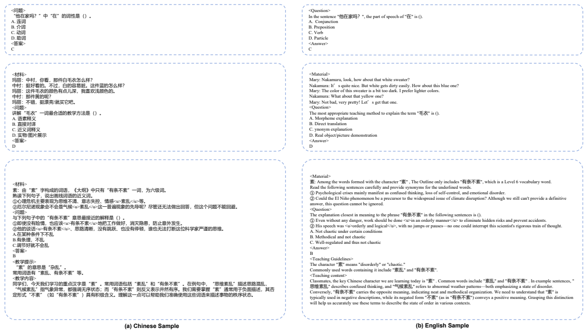

Figure 3: Samples in CLTE.

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

The samples of three task in CLTE are shown in Figure 3.

B Instruction Template

The templates of three task in CLTE are shown in Figure 4.

C Results in Various Fields

The rsesults of differnt fields in CLTE are shown in Figure 3.

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。

<question>

Task 1 Prompt Template

Here is a multiple-choice question about {}. Please select the correct answer and write the option in \boxed{}.

<question>

Task 1 Prompt Template

阅读以下材料并回答问题，选出唯一正确答案并将选项填写到\\boxed{}中。

材料:
<text>

问题:
<question>

Task 2 Prompt Template

Read the following material and answer the question. Select the only correct answer and write the option in \boxed{}.

Material：
<text>

Question：
<question>

Task 2 Prompt Template

system_prompt: 你是一名国际汉语教师。

阅读以下材料，依据材料结合教学提示向学生传授相关知识，以{"knowledge":知识内容}的格式输出。

材料:
<text>

教学提示
<edu prompt>

Task 3 Teacher's Prompt Template

System prompt: You are a teacher of Chinese as an international language.

Read the following material and, based on the content and teaching prompt, deliver relevant knowledge to students. Output in the format: {"knowledge": <knowledge content>}.

Material：
<text>

Teaching Prompt：
<edu prompt>

Task 3 Teacher's Prompt Template

system_prompt: 你是一名正在学习汉语知识的学生。

阅读以下材料，选出唯一正确答案并将选项填写到\\boxed{}中。

材料:
<text>

问题:
<question>

Task 3 Student's Prompt Template

System prompt: You are a student learning Chinese language knowledge.

Read the following material, select the only correct answer, and write the option in \boxed{}.

Material：
<text>

Question：
<question>

Task 3 Student's Prompt Template

system_prompt: 你是一名正在学习汉语知识的学生。

阅读以下材料，结合教师传授的知识回答问题，选出唯一正确答案并将选项填写到\\boxed{}中。

材料:
<text>

教师传授的知识:
<knoledge>

问题:
<question>

Task 3 Student's Prompt Template

System prompt: You are a student learning Chinese language knowledge.

Read the following material and, based on the knowledge taught by the teacher, answer the question. Select the only correct answer and write the option in \boxed{}.

Material：
<text>

Knowledge taught by the teacher:
<knoledge>

Question:
<question>

Task 3 Student's Prompt Template

Figure 4: Templates in CLTE.

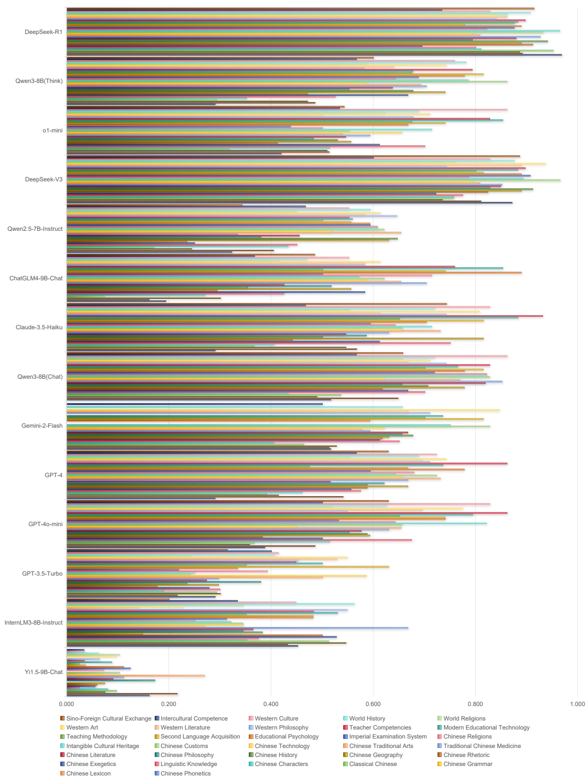

Figure 5: Samples in CLTE.