

Article

CPG-EVAL: Evaluating the Readiness of Large Language Models as Assistants and Teammates in Language Teaching

Dong Wang 

Faculty of Education and Integrated Arts and Sciences, Waseda University, Shinjuku-ku 169-8050, Tokyo, Japan; wangdong@aoni.waseda.jp

Abstract

Large language models (LLMs) have begun to function as assistants or teammates in language learning, teaching, and research. However, what prerequisites are required for LLMs to reliably play these roles, and how such prerequisites should be measured, remains under-discussed. This study focuses on measuring Pedagogical Grammar Pattern Recognition (P-GPR) and establishes the Chinese Pedagogical Grammar Evaluation (CPG-EVAL), a multi-tiered benchmark designed to evaluate P-GPR within International Chinese Language Education. CPG-EVAL operationalizes grammar–instance correspondence through five task types that progressively increase contextual load and interference. We evaluate multiple proprietary and open-source LLMs as well as human participants. Results show a monotonic ordering across groups (humans > larger-scale models > semi-larger-scale models > smaller-scale models). In comparison with human participants, LLM performance is more sensitive to task-format complexity. In addition, we identify a set of completely failed items that consistently mislead all evaluated LLMs, exposing shared and systematic weaknesses in current models’ pedagogical grammar recognition. Overall, this study provides an operational framework for diagnosing the capabilities and risks of LLMs when they are deployed as assistants or teammates in grammar-related language-education tasks and offers empirical reference for safer and more syllabus-aligned use of LLMs in educational settings.

Keywords: large language model; teacher language awareness; benchmark; pedagogical grammar; generative AI; teaching Chinese as a second language



Academic Editors: Antony Bryant, Horacio Saggion, Paolo Bellavista, Kenji Suzuki, Roberto Montemanni, Andreas Holzinger and Min Chen

Received: 10 December 2025

Revised: 25 January 2026

Accepted: 27 January 2026

Published: 6 February 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In recent years, rapid advancements in generative models have significantly expanded the practical applications of LLMs in foreign language education. Models like ChatGPT have begun to fulfill diverse roles within language learning, including supporting autonomous learning, generating adaptive content, and aiding teachers in instructional design and research processes [1,2]. This trend reflects high expectations surrounding their capabilities for natural language generation and reasoning. However, it also raises a fundamental question: what role should AI play in future language education?

Strong domain knowledge alone is not a sufficient condition for effective teaching ability [3]. In the field of language education, even native speakers with excellent linguistic intuition and language proficiency are not necessarily qualified to serve as foreign language teachers [4]. Similarly, although LLMs may produce highly natural text or achieve excellent performance on language tests, such outcomes only indicate strong language performance or NLP competence, rather than pedagogical competence. Moreover, a growing

body of evidence suggests that current LLMs are unable to replace human experts across multiple dimensions, including capability, responsibility, and educational ethics [5]. A more appropriate positioning for LLMs, at the present stage, is therefore as assistants or teammates that operate under human expert supervision and strict alignment with instructional content, rather than as autonomous substitutes for teachers or researchers [6].

In foreign language education, pedagogical grammar is an essential and largely unavoidable component of instruction [7–9]. This implies that if an LLM cannot reliably align with pedagogical grammar, its use in educational tasks—such as textbook analysis, test-item generation, instructional support, or learner error diagnosis—lacks the prerequisite knowledge and capabilities. Nevertheless, to date, no systematic evaluation methodology has been proposed for assessing LLMs specifically with respect to pedagogical grammar.

Against this background, the present study focuses on how to quantitatively evaluate LLMs' ability to align with foreign language teaching syllabi through grammatical pattern recognition. Specifically, we examine whether LLMs can correctly recognize and map pedagogical grammar descriptions to corresponding linguistic instances, a foundational capability required for their use as controlled instructional agents in language education. We propose a low-cost and efficient method for rapidly constructing a multi-tiered, specialized benchmark to assess LLMs' knowledge of pedagogical grammar. Building on this benchmark, we further conduct evaluations involving both LLMs and human participants.

2. Related Work

This section situates the concept of pedagogical grammatical pattern recognition within theoretical frameworks of foreign language education and then reviews related studies on the evaluation of LLM capabilities.

Pedagogical grammar refers to a meta-linguistic system designed to facilitate target language acquisition by second language learners [10]. Rather than describing language as an abstract system, pedagogical grammar organizes linguistic knowledge into teachable and learnable units aligned with instructional practice. For example, expressions such as “nice and friendly”, “big and comfortable”, or “Turkish and English” can be abstracted into a pedagogical grammar rule such as FORM: combining two adjectives with “and” [7]. This abstraction illustrates how pedagogical grammar systematically generalizes instructional rules from concrete language instances. Ref. [11] conceptualizes the mapping between pedagogical grammar items and language instances as Pedagogical Grammar Pattern Recognition (P-GPR). P-GPR refers to an operational process through which a system determines whether a given language instance instantiates a specific pedagogical grammar item and identifies the appropriate grammar item based on a concrete linguistic example. As such, P-GPR represents a minimal yet indispensable operational prerequisite for grammar-related instructional decision-making and thus a core component of LLMs' readiness for pedagogically aligned language-teaching tasks.

From the perspective of second language acquisition research, P-GPR can be regarded as a core operational prerequisite of teacher language awareness (TLA) (extensive discussions of TLA can be found in [12–14]). Ref. [15] defines TLA as “teachers' cognitions (knowledge and beliefs) about language in general and the language they teach.”

Importantly, TLA is not merely an abstract construct but is manifested through concrete instructional actions. Ref. [16] notes that, in classroom contexts, TLA strongly shapes teachers' ability to mediate linguistic input, highlight salient grammatical features, provide appropriate exemplification and clarification, monitor both learners' and their own output, guide learners toward useful generalizations, and reduce potential sources of confusion, while continuously reflecting on the instructional impact of these decisions.

A common implicit prerequisite underlying all these instructional tasks is the ability to establish reliable, bidirectional mappings between abstract pedagogical grammar descriptions and concrete language instances. In other words, any agent engaged in language teaching—whether a human teacher or an artificial intelligence system—should be able to perform P-GPR in order to support grammar-related instructional activities. Despite its centrality to language teaching practice, we are not aware of any prior work that has systematically operationalized or measured P-GPR as a target capability in LLMs.

This gap motivates the present study. As shown in Table 1, existing benchmarks provide comprehensive evaluation frameworks for general language understanding, reasoning, and multilingual knowledge. However, no dedicated benchmark has been proposed to assess LLMs' pedagogical grammar competence in foreign language education. At the macro level, this absence may lead to systematic overestimation or misinterpretation of LLM capabilities in educational policy and curriculum-related decision-making. At the instructional level, it increases the risk that educators and learners rely on LLMs that are not adequately aligned with pedagogical needs. From a development perspective, the lack of domain-specific evaluation signals also limits researchers' ability to diagnose and improve LLMs for language teaching contexts.

To address this gap, we introduce a benchmark specifically designed to evaluate the pedagogical grammar knowledge of LLMs, enabling a more precise and education-oriented assessment of their suitability for foreign language instruction.

Table 1. Comparison of representative benchmarks and their coverage of pedagogical grammar.

Benchmark	Domain Focus	Language	Pedagogical Grammar
GLUE [17]	NLU	English	No
SuperGLUE [18]	NLU + Reasoning	English	No
CLUE [19]	NLU	Chinese	No
SuperCLUE [20]	NLU + Reasoning	Chinese	No
MMLU [21]	General Knowledge	Multilingual	No
BIG-bench [22]	Reasoning	Multilingual	No
PubMedQA [23]	Medicine	English	No
FINQA [24]	Finance	English	No
LegalBench [25]	Law	English	No
E-EVAL [26]	K–12 Education	Chinese	No
C-EVAL [27]	Multi-discipline knowledge	Chinese	Minimal

3. The Construction of CPG-EVAL

This section describes in detail the conceptual framework and construction of the CPG-EVAL. Figure 1 provides an overview.

3.1. Problem Definition & Task Design

Taken together, prior work suggests that the ability to establish reliable, bidirectional mappings between pedagogical grammar items and concrete language instances—referred to as Pedagogical Grammar Pattern Recognition—constitutes a minimal yet foundational competence for any LLM intended to function as a pedagogically aligned instructional agent. However, this fundamental prerequisite of foreign language teaching has received little explicit attention in existing assessment frameworks. Neither language proficiency tests for learners (e.g., TOEFL, IELTS, HSK, JLPT), nor examinations targeting teachers' pedagogical knowledge (e.g., CELTA, CTC SOL), nor current benchmarks for LLMs are designed to directly evaluate grammar–instance correspondence ability. This section details the construction of CPG-EVAL, which is designed to fill the above gap. Before detailing the specific construction procedures, it is necessary to briefly revisit the pedagogical relationship between grammar items and language instances. In instructional practice, pedagogical

cal grammar items function as abstract, curriculum-defined descriptions, while language instances serve as their concrete realizations in actual linguistic input. P-GPR emerges precisely from this abstraction–instantiation relationship, which forms the conceptual basis for the design of the CPG-EVAL benchmark. Ref. [11] conceptualizes this evaluative process as P-GPR and formalizes it as follows:

$$f(G_i, L_i) \rightarrow \{0, 1\} \quad (1)$$

Here, G_i represents a grammar item, and L_i denotes a {language instance} (curly braces {} are used to indicate placeholder variables). The task requires determining whether G_i and L_i correspond/match ($\rightarrow 1$) or do not correspond/mismatch ($\rightarrow 0$). Starting from this abstract formulation, the construction of CPG-EVAL can be understood as a process of concretizing each component of P-GPR for empirical evaluation. Specifically, we describe the selection of pedagogical grammar items (see Section 3.2), the construction of corresponding language instances (see Section 3.3), and the prompt design used to operationalize P-GPR in evaluation tasks (see Section 3.4).

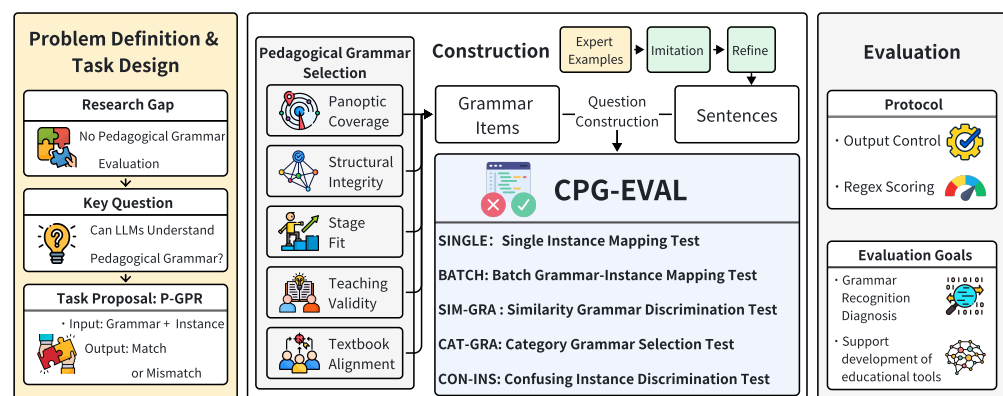


Figure 1. Overall architecture of the CPG-EVAL benchmark. **Note.** For “Problem Definition & Task Design”, see Section 3.1; for “Grammar Items and Sentences,” see Sections 3.2 and 3.3; for “Question Construction,” see Section 3.4. For details on the evaluation procedure, refer to Sections 4 and 5.

3.2. Grammar Item

To transform basic task modeling into an operational evaluation benchmark, it is necessary to address issues related to the selection of the benchmark knowledge framework (Pedagogical Grammar Selection) and the construction of a corresponding example sentence database for grammar items.

The choice of pedagogical grammar is particularly critical, as it reflects the perspectives of a specific expert community and embodies the core values of the evaluation instrument. In the field of English as a Second or Foreign Language (ESL), the English Grammar Profile [7] compiles 1222 statements of grammatical competence, each outlining what learners are expected to achieve at various CEFR levels. In the context of teaching Chinese as a second language, the Chinese Grammar Learning Manual [8] (hereafter abbreviated as CGLM) was developed by Chinese Testing International following the official publication of the Chinese Proficiency Grading Standards for International Chinese Language Education [28], commonly referred to as the “Standards”. The CGLM is a systematically designed grammar syllabus textbook intended to support the implementation of the Standards and cater to the needs of global Chinese language learners.

Beyond the CGLM, numerous pedagogical grammar frameworks are available in various textbooks and other instructional resources. However, for our initial attempt at establishing a benchmark for teaching Chinese grammar knowledge specifically targeted at

LLMs, we selected the CGLM as the knowledge core for the CPG-EVAL due to the following advantages:

- **Comprehensive Coverage:** The CGLM offers a grammar outline highly aligned with the Standards, covering all proficiency levels from beginner to advanced, and aligned with international examination systems such as HSK and YCT.
- **Architectural Integrity:** Grammar items in the CGLM are not listed in isolation; they are instead systematically categorized and integrated according to dimensions such as phonetics, vocabulary, semantics, and pragmatics, establishing explicit interconnections and a pedagogically sequenced progression of grammatical functions.
- **Stage Appropriateness:** The CGLM adjusts the cognitive load for learners at different proficiency levels, ensuring that grammar items at each level are both challenging and achievable.
- **Teaching Effectiveness:** Developed from over two decades of grammar teaching and assessment practices, the CGLM incorporates extensive empirical teaching data and classroom feedback, verifying the strong correlation between grammar points and teaching activities and demonstrating its effectiveness.
- **Educational Compatibility:** The CGLM aligns closely with the grammar distribution patterns in mainstream international Chinese textbooks, such as “Developing Chinese”, “New Practical Chinese Reader”, and “HSK Standard Course”, ensuring high compatibility between the grammar benchmark and textbook content.

In this study, we constructed a structured database based on the grammar items in the GLM. After excluding four items serving purely indexing purposes and without practical teaching content, we retained 739 grammar items of substantial pedagogical significance to form the basis of the CPG-EVAL grammar items.

3.3. Language Instance

In the CGLM, each grammar item is accompanied by nine example sentences designed to illustrate the linguistic phenomena described by that grammar item. These example sentences have been thoroughly reviewed by the editorial committee of the CGLM and revised by Chinese linguists. However, these expert-curated data are not suitable to serve as a {language instance} for testing purposes, as they may already be included in the training datasets of some large language models (LLMs), which could introduce significant bias. To balance professionalism and open accessibility, this study employs synthetic data “equivalent to expert data in terms of its relationship to grammar items”.

The synthesis process involves the following steps: (1) Using the official example sentences from the CGLM as templates, rewritten versions were generated by DeepSeek-v3; (2) Three native Chinese speakers, including the author, reviewed these sentences to ensure their compliance with Chinese linguistic conventions and to confirm that the synthetic data accurately preserved the relationship between the original expert examples and their respective grammar items; (3) The synthesized sentences were categorized in a database according to their corresponding grammar items. Through this method, a total of 739 grammar items $\times$ nine sentences = 6651 synthetic sentences were obtained as language instances for constructing benchmark questions.

Additionally, we constructed highly confusing instances designed to be particularly challenging, aiming to further evaluate the LLMs’ resistance to interference (see Section 3.4).

3.4. Question Construction

This section details the specific types of questions included in CPG-EVAL. Building upon [11]’s Single Instance Mapping Test, the present study expands the assessment methods to evaluate the multidimensional capabilities of LLMs. All actual prompts were written in Chinese; however, English is used in the following examples for ease of presentation:

1. **SINGLE: Single Instance Mapping Test** specifically assesses whether LLMs can accurately identify the correspondence between a single language instance and the indicated grammar item, following the basic format of P-GPR defined by [11]. SINGLE includes two subtasks: SINGLE-T (answer is T) and SINGLE-F (answer is F).

Prompt template: 请判断: 在句子 {Language Instance} 中, 是否包含语法点 {Grammar Item}, 如果包含, 请仅回答 T, 如果不包含, 请仅回答 F (Please evaluate: Does {Language Instance} contain the grammar point {Grammar Item}, If yes, output: T; if no, output: F.)?

Example: 请判断: 在句子 “我不会弹钢琴。” 中, 是否包含语法点 [能愿动词: 会]? 如果包含, 请仅回答 T, 如果不包含, 请仅回答 F. (Please evaluate: Does the sentence “wǒ bù huì tán gāng qín.” (I cannot play the piano) contain the grammar point [qíng tài dòng cí: huì] (Modal verb: huì—can, will)? If yes, output: T; if no, output: F.

Answer: T

2. **BATCH: Batch Grammar-Instance Mapping Test** assesses the ability of LLMs to determine the correspondence between multiple language instances and the indicated grammar item. It therefore probes whether an LLM can (i) maintain stable grammar-instance judgments under higher contextual load and (ii) reliably follow item-wise evaluation instructions by assigning a separate T/F label to each sentence, rather than collapsing the set into a single global impression.

Prompt template: 请对 1. 号到 9. 号每个句子依次进行判断: 如果包含语法点 {Grammar Item}, 回答 T. 如果不包含, 回答 F. (Please evaluate sentences 1 through 9 one by one: If a sentence contains the grammar point {Grammar Item}?, respond with T. If not, respond with F. Use only T or F for each sentence. No explanation is needed.) {Language Instance} × 9

Example: 请对 1. 号到 9. 号每个句子依次进行判断: 如果包含语法点 [能愿动词: 能], 回答 T. 如果不包含, 回答 F. (Please evaluate sentences 1 through 9 one by one: If a sentence contains the grammar point [néng yuàn dòng cí: néng] (Modal verb: néng - be able to)?, respond with T. If not, respond with F. Use only T or F for each sentence. No explanation is needed.)

1. 我不会弹钢琴。(I can’t play the piano.)
2. 我妹妹会画画。(My younger sister can draw.)
3. 他会参加下周的会议吗? (Will he attend next week’s meeting?)
4. 你会游泳吗? (Can you swim?)
5. 她会跳舞, 不会唱歌。(She can dance but can’t sing.)
6. 这孩子会说话了。(This child can speak now.)
7. 爷爷不会用电脑。(Grandpa can’t use a computer.)
8. 你会不会做饭? (Can you cook or not?)
9. 明天她不会来学校。(She won’t come to school tomorrow.)

Answer: FFFFFFFF

3. **SIM-GRA: Similarity Grammar Discrimination Test** evaluates whether LLMs can distinguish between grammatically similar grammar items. The four distractor grammar items are derived by calculating semantic vector similarity using embedding models. (Specifically, we use OpenAI’s text-embedding-3-large model to generate vector representations for each of the 739 grammar entries. Then, we compute the

similarity scores between each grammar entry and the remaining 738 entries. The four most similar grammar entries are selected as distractor options based on similarity scores). This design explicitly tests whether LLMs can map observed linguistic forms back to the correct pedagogical abstraction, rather than merely verifying a proposed rule, thereby challenging their ability to perform fine-grained grammatical discrimination under conditions of high conceptual similarity.

Prompt template: 句子 {Language Instance} 最适合作为以下哪个语法点的例句? 请只使用选项字母回答。(Which grammar point does the sentence {Language Instance} best exemplify? Please answer using only the option letter.)

{Grammar Item} × 5

Example: 句子“他没有哥哥。”最适合作为以下哪个语法点的例句? 请只使用选项字母回答。(Which grammar point does the sentence “tā méi yǒu gē gē.” (He doesn’t have an older brother.) best exemplify? Please answer using only the option letter.)

- A. “有”字句: 表示附着: 主语 + 动词 + 有 + 宾语 (“yǒu”-construction: indicating attachment - Subject + Verb + yǒu + Object)
- B. “有”字句: 表示领有 (“yǒu”-construction: indicating possession)
- C. “有”字句: 表示存在、具有: 主语 + 有 + 着 + 宾语 (“You”-construction: indicating existence/possession with subject + yǒu + zhe + object)
- D. “有”字句: 表示比较 (“yǒu”-construction: indicating comparison)
- E. “有”字句: 表示存在 (“yǒu”-construction: indicating existence)

Answer: B

4. **CAT-GRA: Category Grammar Selection Test** specifically evaluates the ability of LLMs to differentiate among grammar items within the same grammatical category. In CAT-GRA, all grammar items belonging to the same grammatical category as the correct answer are provided as options. Rather than merely increasing the number of options, this task is intentionally designed to challenge LLMs’ language processing abilities within a highly specialized and systematically organized pedagogical grammar knowledge context, where distinctions are subtle, formally constrained, and instructionally meaningful.

Prompt template: 句子 {Language Instance} 最适合作为以下哪个语法点的例句?(Which grammar point does the sentence {language instance} best exemplify?)

Example: 句子“苹果是红的。”最适合作为以下哪个语法点的例句? (Which grammar point does the sentence “píng guǒ shì hóng de.” (The apple is red.) best exemplify?)

- 1063.“是”字句: 表示等同或类属 (“Shì”-construction: expressing equivalence or class membership)
- 1064.“是”字句: 表示说明或特征 (“Shì”-construction: indicating description or characteristic)
- 1065.“是”字句: 表示存在 (“Shì”-construction: expressing existence)
- 3113. 用“是”强调 (Using “Shì” for emphasis)

Answer: 1064

5. **CON-INS: Confusing Instance Discrimination Test** is designed to examine whether LLMs can resist misleading linguistic cues and correctly reject language instances that appear similar to a target grammar item but do not actually instantiate it (CON-INS includes two subtasks: CON-INS-F10 (all instances do not contain the targeted grammar point) and CON-INS-T5F5 (half of the instances contain the targeted grammar point, half do not). CON-INS can be viewed as a variant of BATCH, increasing the difficulty at the language instance level by requiring LLMs to resist interference from linguistic similarities. Initially, 314 grammar items were selected following these two criteria, and DeepSeek-v3 was employed to generate sentences featuring similar lin-

guistic forms but not aligned with the targeted grammar items. Two native-speaking Chinese language teachers specializing in teaching Chinese as a second language then screened and revised the sentences).

Confusing instances primarily involve three types:

- (a) Grammatical markers embedded in lexical words. For example, (1) 会 as a modal verb (“can, will”) vs. 会议 (“meeting”), where 会 is part of a noun; (2) 可 as a degree adverb (“can”) vs. 可持续 (“sustainable”), where 可 functions as part of an adjective.
- (b) Identical surface forms expressing different grammatical meanings. For example, (1) 躺着, where 着 marks the continuation of a state, vs. 跑着, where 着 marks the continuation of an action; 当 in 当你去中国时 (“when you go to China”), where 当 functions as a preposition introducing a temporal clause, vs. 当 in 他当经理 (“he serves as a manager”), where 当 functions as a verb meaning “to assume a role.”
- (c) Different forms expressing similar grammatical meanings. For example: (1) 发挥着作用, where 着 marks an ongoing action, vs. 正在发挥作用, where 正在 marks an ongoing action; (2) 这 (this, colloquial/modern) vs. 此 (this, formal/archaic).

Prompt template: 请对 1. 号到 10. 号每个句子依次进行判断: 如果包含语法点 {Grammar Item}, 回答 T. 如果不包含, 回答 F. (Please evaluate sentences 1 through 10 one by one: If a sentence contains the grammar point {Grammar Item}, respond with T. If not, respond with F.) {Language Instance} × 10

Example: 请对 1. 号到 10. 号每个句子依次进行判断: 如果包含语法点 [经历态: 用动态助词 “过” 表示], 回答 T. 如果不包含, 回答 F. (Please evaluate sentences 1 through 10 one by one. If a sentence contains the grammar point [Experiential aspect: expressed by the dynamic particle “guo”] (in someone’s opinion, from someone’s perspective), respond with T. If not, respond with F.)

1. 我看过那本书。(I have read that book.)
2. 我没吃过火锅。(I have never eaten hot pot.)
3. 她过敏体质, 不能吃海鲜。(She has an allergic constitution and cannot eat seafood.)
4. 他没学过法语。(He has never studied French.)
5. 她当过学生会主席。(She once served as the student union president.)
6. 这本书我已经过目了, 内容很不错。(I have already skimmed through this book; the content is very good.)
7. 她过分谦虚, 反而让人觉得不真实。(She is overly modest, which makes her seem insincere.)
8. 她过生日那天, 收到了很多礼物。(On her birthday, she received many gifts.)
9. 他过马路时, 差点被车撞到。(He almost got hit by a car while crossing the street.)
10. 她学过法语。(She has studied French.)

Answer: TTFTTFFFFT

For the total number of questions in each question bank, see the “Number of Questions” column in Table 2.

Table 2. Distribution of question items by type in the CPG-EVAL benchmark.

Question Bank	Number of Questions
SINGLE-T	6651
SINGLE-F	11,968
BATCH-T	739
BATCH-F	11,968
SIM-GRA	6651
CAT-GRA	5283
CON-INS-F10	314
CON-INS-T5F5	314

Note. SINGLE and BATCH test grammar recognition in single or multiple sentences, with subtypes T/F indicating presence or absence of the grammar item. SIM-GRA and CAT-GRA assess fine-grained grammar selection among similar or same-category items. CON-INS tasks evaluate robustness against confusing instances. See Section 3.4 for detailed definitions and examples.

3.5. Sampling Procedure for CPG-EVAL: Core

The full CPG-EVAL question bank contains as many as 43,888 items, resulting in significant consumption of computational resources and making it unsuitable for experiments involving human participants. Therefore, we prepared a streamlined version of the question bank, referred to as CPG-EVAL: core. To construct the CPG-EVAL: core subset, a stratified random sampling method was employed based on the full question pool of CPG-EVAL. The sampling process aimed to ensure representativeness across different task types and proficiency levels (Lv1–Lv7), while maintaining statistical validity under commonly accepted survey sampling assumptions. The required sample size for each major subset (e.g., SINGLE-T, BATCH-F) was first determined using the standard sample size formula for finite populations:

$$n = \frac{N \cdot Z^2 \cdot p(1-p)}{E^2 \cdot (N-1) + Z^2 \cdot p(1-p)} \quad (2)$$

where

- N is the total number of relevant items in the population (e.g., 2844 elementary-level SINGLE-T questions);
- $Z = 1.96$ corresponds to the 95% confidence level;
- $p = 0.5$ is the assumed proportion (used to yield the maximum required sample size);
- $E = 0.05$ is the margin of error.

Applying this formula yields a required sample size of $n = 339$ for the elementary-level SINGLE-T subset. To compute the sample size specific to each level (e.g., Lv1), proportional allocation was used. For example, if Lv1 accounts for 729 out of the 2844 elementary SINGLE-T items, the Lv1-specific sample size is calculated as follows:

$$n_{\text{SINGLE-T-Lv1}} = \frac{729}{2844} \cdot 339 \approx 87 \quad (3)$$

This procedure was applied across all five task types (SINGLE, BATCH, SIM-GRA, CAT-GRA, CON-INS) and all seven proficiency levels (Lv1–Lv7), ensuring that the resulting CPG-EVAL core set preserves both coverage and balance in pedagogical grammar difficulty.

Subsequently, the 5935 sampled questions were divided into four parts, and four participants were invited to complete the test, with each spending an average of 386 min on the questions. (All four participants were native Chinese speakers with master's degrees in linguistics, and none had received targeted training based on CGLM. During the ex-

periment, participants were instructed to answer independently, relying solely on their existing knowledge and understanding, without consulting external materials. To assess participants' performance under optimal conditions, the testing was conducted over multiple sessions, with each session lasting no more than 60 min continuously. Compensation was provided based on the total time spent answering questions, thereby minimizing potential influences arising from conflicts of interest between the experimental procedure and participants' incentives. Additionally, regarding the specific Chinese teaching experience of the four participants, Human-1 had 8 years of experience, Human-2 had 5 years of experience, Human-3 had short-term teaching internship experience, and Human-4 had no practical teaching experience). During manual verification, we identified 20 questions (approximately 0.3% of the sample) with incorrect pre-specified gold labels. (The following example illustrates how incorrect pre-specified gold labels can arise during the construction process. In the sentence synthesis stage, an LLM was instructed to generate a sentence containing the modal particle “而已” (just). The model generated the sentence: “我不过是随便翻翻而已, 没有认真读。” (I was just casually flipping through it, nothing more than that; I didn't read it carefully). This sentence was subsequently incorporated into a single-choice grammar identification item (SIM-GRA). However, one of the answer options corresponded to the fixed pattern “不过……而已” (just / only …and nothing more). As a result, the sentence simultaneously instantiated two distinct pedagogical grammar items, making more than one option defensible as a correct answer and thereby violating the single-correct-answer assumption of the task. More generally, when generating synthetic examples, it is difficult to guarantee with complete reliability that a sentence (i) maintains sufficient linguistic diversity, (ii) instantiates a designated target grammar item, and (iii) excludes other closely related grammar items that are formally or functionally adjacent. The example above represents only one of several possible error-inducing scenarios; in practice, the causes of incorrect gold labels are more complex and cannot be reliably resolved through simple rule-based or pattern-matching approaches. Addressing these issues in a systematic and scalable manner remains an important direction for future work). After correction and filtering, the final version consists of a rigorously human-verified set of 5915 questions, which constitutes CPG-EVAL: core. CPG-EVAL: core is published separately in the repository. When evaluators wish to avoid the approximately 0.3% error caused by contentious items, or aim to save computational resources, they can use CPG-EVAL: core for a faster and more economical evaluation.

4. Evaluation

4.1. Setup and Models

We employed CPG-EVAL to evaluate multiple open-source and proprietary large language models under a zero-shot setting, in which no demonstrations or supplementary information beyond the task instructions and answer choices described in Section 3.4 were provided. All prompts were issued in Chinese, using the exact prompt templates specified in Section 3.4. The decoding temperature was fixed at 0, with a maximum output length of 1000 tokens. To assess result stability, each model was evaluated three times under this configuration; the variation in accuracy across runs did not exceed 0.002 and therefore did not materially affect the results reported in Section 4.2. The list of evaluated models (In this study, models accessed via Weights were downloaded directly from the official Hugging Face repositories provided by the model developers, which are publicly available at https://huggingface.co/{publisher}/{model_name} (e.g., <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>, accessed on 1 June 2025). models accessed via APIs were used through interfaces officially provided by their respective developers) is provided in Table 3.

Table 3. Model specifications: parameters and access method. Note. For more information about models, see the following references: Qwen [29], GLM [30], internlm2_5 [31], Llama [32], DeepSeek [33], GPT [34,35], and Doubao [36].

Model	#Parameters	Access
Qwen2.5-7B-Instruct	7B	Weights
internlm2_5-7b-chat	7B	Weights
Llama-3.1-8B-Instruct	8B	Weights
GLM-4-9B-Chat	9B	Weights
Qwen2.5-70B-Instruct	70B	Weights
DeepSeek-V3-250324	660B MoE	API
GPT-4o-2024-08-06	Undisclosed	API
GPT-4o-mini-2024-07-18	Undisclosed	API
Doubao-1-5-pro-32k-250115	Undisclosed	API
Doubao-1-5-lite-32k-250115	Undisclosed	API
Qwen2.5-MAX-250409	Undisclosed	API

To calculate scores objectively, we utilized regular expressions to extract the answers from the LLM outputs. To reduce noise that could affect statistical accuracy, explicit instructions such as “Please respond using only the letter corresponding to the correct answer” were embedded in the prompts to enforce output control, thereby ensuring consistency and clarity in the LLM-generated responses. Finally, we extracted answers from the model responses using regex scoring and calculated each model’s scores across the five evaluation categories.

4.2. Results

The evaluation results (in CPG-EVAL, each BATCH question is comprised of nine sub-questions, and each CON-INS question is comprised of ten sub-questions. Accuracy is calculated independently at the sub-question level. For example, if a CON-INS question has the correct answer “TTTTFTTFFF” and the model responds with “TTTTFTFTTT”, the model correctly answers six out of ten sub-questions, thus scoring 0.6 points (with a full score of 1 point per question)) for the five types of tests are shown in Table 4. In Figure 2, we use box plots to present the accuracy results of four groups across five question types: Human (Human-1, 2, 3, 4), larger-scale models (Doubao-1-5-pro, GPT-4o, DeepSeek-v3, Qwen2.5-Max), semi-larger-scale models (Doubao-1-5-lite, Qwen2.5-72B, GPT-4o-mini), and smaller-scale models (Qwen2.5-7B, glm-4-9b, Llama-3.1-8B, internlm2_5-7b).

To further distinguish the gap in ability between humans and models of different scales, we fitted a mixed-effects model separately for each task at the item $\times$ subject level, with random intercepts for items and subjects and a fixed effect of group order (For tasks with binary item scores (SINGLE, SIM-GRA, CAT-GRA), we used a binomial GLMM; for tasks scored as proportions (BATCH, CON-INS), we used a linear mixed model (LMM). One-sided adjacent contrasts tested the ordered hypothesis). For each task, we tested the ordered hypothesis with one-sided adjacent contrasts (humans > larger-scale models > semi-larger-scale models > smaller-scale models). Across all tasks, adjacent contrasts were significant in the expected direction (one-sided $p < 0.05$), indicating a statistically robust monotonic ordering of group performance. The accuracy distributions across tasks and subject groups are shown in Figure 2.

Table 4. The results are based on the CPG-EVAL: core subset evaluation results across five test types (SINGLE, BATCH, SIM-GRA, CAT-GRA, CON-INS) and average scores. The bold formatting highlights the highest value in each individual item for ease of comparison.

Model	SINGLE		BATCH		SIM-GRA	CAT-GRA	CON-INS		Average
	T	F	T*9	F*9			F*10	T*5&F*5	
RANDOM	0.500	0.500	0.500	0.500	0.200	0.100	0.500	0.500	0.413
Doubao-1-5-pro	0.985	0.944	0.987	0.942	0.949	0.939	0.842	0.960	0.950
GPT-4o	0.977	0.939	0.926	0.934	0.938	0.906	0.823	0.935	0.933
DeepSeek-v3	0.952	0.961	0.940	0.897	0.910	0.870	0.792	0.922	0.916
Qwen2.5-Max	0.989	0.905	0.849	0.918	0.903	0.879	0.796	0.945	0.910
Doubao-1-5-lite	0.952	0.938	0.921	0.889	0.908	0.859	0.716	0.902	0.903
Qwen2.5-72B	0.949	0.954	0.939	0.897	0.885	0.851	0.689	0.873	0.900
GPT-4o-mini	0.921	0.928	0.950	0.712	0.868	0.809	0.554	0.843	0.843
Qwen2.5-7B	0.906	0.930	0.717	0.779	0.838	0.608	0.600	0.732	0.797
glm-4-9b	0.953	0.794	0.930	0.522	0.765	0.727	0.428	0.767	0.750
Llama-3.1-8B	0.979	0.781	0.803	0.549	0.795	0.626	0.416	0.726	0.734
internlm2_5-7b	0.969	0.796	0.737	0.304	0.928	0.546	0.371	0.652	0.692
Ave.LLM	0.958	0.897	0.882	0.758	0.881	0.784	0.639	0.842	0.848
Human-2	0.968	0.996	0.987	0.999	0.968	0.929	0.920	0.976	0.972
Human-1	0.988	0.985	0.960	0.991	0.939	0.961	0.961	0.973	0.972
Human-3	0.995	0.970	0.951	0.979	0.950	0.938	0.958	0.934	0.964
Human-4	0.957	0.974	0.976	0.967	0.913	0.899	0.897	0.941	0.943
Ave.Human	0.977	0.981	0.969	0.984	0.943	0.932	0.934	0.956	0.963

Note. SINGLE = Single Instance Mapping Test; BATCH = Batch Grammar-Instance Mapping Test; SIM-GRA = Similarity Grammar Discrimination Test; CAT-GRA = Category Grammar Selection Test; CON-INS = Confusing Instance Discrimination Test. For detailed task descriptions, see Section 3.4. The results for human participants are based on the CPG-EVAL: core Subset; for further details, please refer to Section 3.5. For model specifications, refer to Table 3.

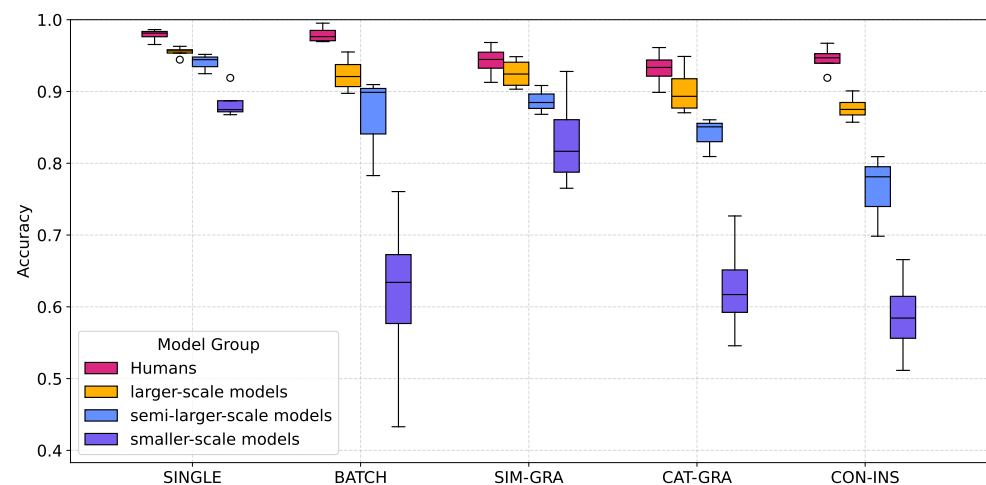


Figure 2. Accuracy Distributions Across Tasks and Subject Groups. Note. Distributions are based on CPG-EVAL: core. Humans (Human-1, 2, 3, 4), larger-scale models (Doubao-1-5-pro, GPT-4o, DeepSeek-v3, Qwen2.5-Max), semi-larger-scale models (Doubao-1-5-lite, Qwen2.5-72B, GPT-4o-mini), and smaller-scale models (Qwen2.5-7B, glm-4-9b, Llama-3.1-8B, internlm2_5-7b).

5. Analysis and Discussion

5.1. Results of SINGLE and BATCH

In the CPG-EVAL evaluation framework, the SINGLE task represents the simplest scenario by merely requiring LLMs to judge whether instructional grammatical descriptions match particular language instances. Therefore, it offers an intuitive reflection of an LLM's capability in recognizing the linguistic phenomena defined by pedagogical grammar rules.

Overall, analysis of SINGLE and BATCH tasks reveals four critical observations:

- (1) The performance of LLMs on SINGLE-T and BATCH-T tasks is significantly better than on SINGLE-F and BATCH-F tasks. This indicates that errors primarily arise from

- false positives, i.e., mistakenly identifying a language instance that does not contain a specific grammar item as one that does. It is particularly noteworthy that, except for Qwen2.5-7B, all other smaller models exhibit a serious false positive problem.
- (2) Some smaller-scale LLMs demonstrate recognition abilities on the SINGLE-T task (such as Llama-3.1-8B) and the BATCH-T task (such as glm-4-9b) that are comparable to, or even exceed, those of certain larger models.
 - (3) LLMs employ different strategies when determining the presence or absence of a grammar item within a language instance. Some models tend to be more assertive—more likely to provide affirmative responses—while others are relatively conservative and more inclined to give negative responses.
 - (4) In the more complex batch task BATCH-F, larger models exhibit a clear advantage when handling negative instances (i.e., instances that do not conform to grammatical rules). Under the difficulty posed by negative instances and the demand to follow batch judgment instructions, the accuracy of some smaller-scale models (such as internlm2_5-7b) even falls below that of the random baseline.

These results offer the following implications: (1) Current LLMs applied to real-world pedagogical contexts are likely to misclassify sentences as containing specific grammatical phenomena that are not present. (2) For general users, such as teachers and learners, larger-scale models are more reliable when used as tools for language education and learning. For developers, it is possible to enhance the practicality of certain smaller-scale models by decomposing tasks that involve multiple language instances into several SINGLE-type tasks.

This reinforces the importance of a systematically multi-tiered evaluation framework for diagnosing and better understanding gaps and limitations in LLMs' capabilities.

5.2. From the Results of SIM-GRA and CAT-GRA

If LLMs are to be used in foreign language teaching, it is essential for them to have fine-grained grammatical discrimination abilities. The two types of tasks explored in this study, SIM-GRA and CAT-GRA, share considerable similarity: both require LLMs to identify the grammatical rule that best describes a given linguistic instance. The primary distinction between these two task types lies in how distractors are generated, either semantically or grammatically.

The results from the SIM-GRA and CAT-GRA tasks yield several key observations:

- (1) Models with larger scales and more comprehensive training demonstrate greater robustness and generalization when confronted with fine-grained grammatical distinctions and specialized forms of interference.
- (2) Smaller-scale models, such as LLaMA-3.1-8B and InternLM2.5-7B, achieved relatively high scores in the SIM-GRA task (81.2% and 92.3%, respectively), but their performance dropped significantly in the CAT-GRA task (61.3% and 54.3%, respectively).

Those findings highlight that literal and categorical semantic similarities present fundamentally different challenges to LLMs. In other words, the more specialized and systematically organized the language teaching task, the more demanding it becomes for these models.

In conclusion, the current generation of LLMs commonly encounters difficulties in handling detailed, specialized interference within the structured knowledge systems of pedagogical grammar. This limitation is especially pronounced for smaller-scale models, indicating that for fine-grained grammar discrimination tasks within foreign language teaching scenarios, selecting larger language models is currently advisable due to their consistently superior performance.

5.3. From the Results of CON-INS

When responding to CON-INS tasks, LLMs particularly need to focus on overcoming interference arising from linguistic form and semantics when making judgments at the grammatical-functional level. The results from the CON-INS tasks yield several key observations:

- (1) As shown in Figure 2, among the five task types, CON-INS appears to exhibit the highest discriminative power, as the boxplots show the most pronounced separation between human participants, larger-scale, semi-larger-scale, and smaller-scale models.
- (2) As indicated by the comparison between F10 and T5F5 in Table 4, the greater the interference from confusing instances, the more likely models are to make errors.
- (3) Large-scale models perform relatively well, but there remains a significant accuracy gap between them and humans.
- (4) For the majority of smaller-scale models, accuracy on F10 falls below the random baseline.

These results indicate the following: (1) Confusing instances pose a significant challenge to the capabilities of current LLMs. This highlights persistent obstacles to the effective use of these models as language-teaching assistants. The observed performance gap underscores the importance of enhancing grammatical discrimination abilities and resistance to interference in small-scale, cost-effective language models designed for educational settings. (2) CON-INS is more sensitive to performance differences across model tiers and may serve as a more informative subtask for evaluating pedagogical grammar competence.

5.4. From Completely Failed Items

This subsection examines a subset of benchmark items for which all evaluated LLMs consistently produced incorrect responses, which we refer to as completely failed items. Unlike partially failed cases, these items reveal systematic weaknesses shared across models and therefore provide particularly informative diagnostic signals about current limitations in pedagogical grammar recognition. Based on an error analysis of these items, the observed failures can be broadly categorized into three recurring types, distinguished by the interaction between surface form, semantic interpretation, and token-level processing. The following examples illustrate each type.

Type I: Errors induced by formal similarity overriding contextual semantic interpretation. The item [temporal adverb: zǎowǎn] means “sooner or later”, indicating that a result will eventually occur. However, in the sentence “爷爷 [早晚] 都会去公园遛弯儿” (Grandpa will go for a walk in the park [zǎowǎn] (morning and evening)) (CON-INS-F10), the phrase [早晚] (zǎowǎn) is difficult to interpret as a temporal adverb (sooner or later); rather, it should be understood as “morning and evening”. When LLMs were asked whether the sentence contains the [temporal adverb: zǎowǎn], all tested models gave the wrong answer.

Type II: Errors induced by semantic similarity overriding formal distinctions. If the previous example can be explained by LLMs’ possible inadequate understanding of the linguistic term “temporal adverb”, then the next “completely failed item” may seem rather baffling from a human perspective. Please evaluate: in the sentence “这次会议的参与者 [仅] 有十个人” (There are [jǐn] (only) ten participants in this meeting) (SINGLE-F), does it contain the grammar point of [范围、协同副词: 只] (scope/coordinating adverb: zhǐ (only))? From a human perspective, “仅” (only) and “只” (only) are obviously two different words, so the clear answer is “does not contain”. However, all tested LLMs judged that it does. A possible reason is that, in this context, “jǐn” can be replaced by “zhǐ”, and after substitution,

the truth-conditional meaning of the sentence remains unchanged. A similar case can be observed where models incorrectly judge that the sentence “这道菜 [既] 美味, [又] 营养。” (“This dish is [both] delicious and [also] nutritious.”) contains the grammar point “又 …, 又 …”.

Type III: Errors induced by token-level overlap and segmentation effects. In this category, misjudgments arise from token-level segmentation, where a target grammatical form appears as a substring of a larger lexical unit, leading LLMs to incorrectly infer the presence of the grammar item. For example, in the SINGLE task, consider the sentence “[我们] 去公园, [你们] 去哪儿?” (“We are going to the park; where are you going?”) and the grammar item [personal pronouns: [我](I/me), [你](you), [您](you), [他](he/him), [她](she/her)], which refers to singular personal pronouns. Although the sentence only contains plural forms, all evaluated models incorrectly judged that the grammar item is present. In Chinese, this error can be attributed to the fact that the characters used for singular personal pronouns (e.g., “我”) also appear within their corresponding plural forms (e.g., “我们”), which appears to mislead the models. A similar phenomenon can be observed in cases where LLMs incorrectly judge that the sentence “我们坚决不让类似的错误 [再次] 出现。” (“We firmly will not allow similar mistakes to occur [again].”) contains the grammar item [frequency/repetition adverb: 再]. Although the character “再” appears within the word “再次”, the two forms differ substantially in usage (for example, one can say “再睡一小时” [sleep for another hour], but not “* 再次睡一小时”). Despite this distinction, LLMs consistently fail to differentiate between the standalone grammatical use of “再” and its occurrence as part of a compound word.

Type IV: Hallucinations with unknown causes. In this category, models incorrectly identify the presence of a grammar item despite the absence of any corresponding formal or semantic cues. For example, the evaluated models incorrectly judged that the sentence “这家咖啡馆很安静, 但 WiFi 信号不太好。” (“This café is very quiet, but the WiFi signal is not very good.”) contains the grammar point [conjunction (connective word or phrase): 与 2], where “与 2” is a shorthand used in International Chinese Education to denote the second usage of 与. In this sentence, the character “与” does not appear, nor are there any words or phrases expressing a similar meaning (e.g., and). Nevertheless, the models consistently judged that the grammar point 与 2 is present.

These “total failure items” expose the current limitations in the capabilities of LLMs. These findings provide important clues for further developing more targeted adversarial samples for future research.

6. Conclusions

This study argues that reliable bidirectional mapping between pedagogical grammar items and concrete language instances—formalized as Pedagogical Grammar Pattern Recognition (P-GPR)—constitutes a minimal yet indispensable operational prerequisite specifically for grammar-related instructional decision-making when LLMs are used in foreign language education. To operationalize and measure this capability, we constructed CPG-EVAL, a multi-tiered benchmark designed to evaluate P-GPR within International Chinese Language Education, and evaluated both human participants and a range of open-source and proprietary LLMs.

In Section 5, this paper presents several critical observations and implications. Among them, we consider the following three points to be particularly significant. (1) Compared with human participants, the complexity of task formats has a more pronounced impact on the performance of LLMs. (2) When handling certain types of simple tasks (SINGLE-T), some LLMs have already matched or even surpassed human participants in performance. However, LLMs generally exhibit a notable tendency toward false positives. (3) Influenced

by semantic similarity, LLMs may sometimes make judgments that contradict formal evidence, with some errors appearing baffling from a human perspective.

At the same time, the scope of this study should be recognized. CPG-EVAL targets P-GPR as a foundational mapping competence rather than “teaching ability” in a broad sense; it does not directly evaluate higher-level pedagogical decision-making, instructional ethics, or long-horizon classroom interaction. Despite these boundaries, if an LLM can perform P-GPR reliably, then grammar-related educational applications can be substantially “de-risked” by constraining model outputs to syllabus-aligned grammar inventories, for example, through a retrieval-augmented generation (RAG) framework [37]. Beyond single-model usage, a lightweight LLM with strong P-GPR competence can serve as a dedicated specialist module within a multi-agent system, collaborating with other agents (e.g., content retrievers, feedback planners, rubric checkers) to support tasks such as fine-grained textbook analysis, test-item review, and high-quality feedback on learner production under teacher supervision.

This study represents an initial step toward capability-oriented evaluation for positioning large language models as assistants and teammates in foreign language teaching and research. Future research will move beyond this initial abstraction by examining LLM evaluation from more concrete teaching perspectives, including diverse instructional tasks, learner profiles, and usage scenarios. Together, these directions aim to promote safer and more reliable integration of LLMs into language education by aligning model behavior with pedagogical grammar systems and real teaching practice.

Funding: This research was supported by Waseda University Special Research Projects (Grant No. 2025E-008). The APC was funded by Waseda University.

Data Availability Statement: The CPG-EVAL dataset and analysis code are openly available at <https://github.com/wd-github-2017/CPG-EVAL> (accessed on 26 January 2026).

Acknowledgments: Declaration of generative AI and AI-assisted technologies in the writing process. Statement: During the preparation of this work, the author used ChatGPT-5 in order to improve the readability and language of the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Conflicts of Interest: No potential conflict of interest was reported by the author.

References

1. Karataş, F.; Abedi, F.Y.; Ozek Gunyel, F.; Karadeniz, D.; Kuzgun, Y. Incorporating AI in Foreign Language Education: An Investigation into ChatGPT’s Effect on Foreign Language Learners. *Educ. Inf. Technol.* **2024**, *29*, 19343–19366. <https://doi.org/10.1007/s10639-024-12574-6>.
2. Li, B.; Lowell, V.L.; Wang, C.; Li, X. A Systematic Review of the First Year of Publications on ChatGPT and Language Education: Examining Research on ChatGPT’s Use in Language Learning and Teaching. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100266. <https://doi.org/10.1016/j.caeai.2024.100266>.
3. Monk, D.H. Subject Area Preparation of Secondary Mathematics and Science Teachers and Student Achievement. *Econ. Educ. Rev.* **1994**, *13*, 125–145. [https://doi.org/10.1016/0272-7757\(94\)90003-5](https://doi.org/10.1016/0272-7757(94)90003-5).
4. Medgyes, P. *The Non-Native Teacher*; Macmillan: London, UK, 1994.
5. Shi, Y.; Yu, K.; Dong, Y.; Chen, F. Large Language Models in Education: A Systematic Review of Empirical Applications, Benefits, and Challenges. *Comput. Educ. Artif. Intell.* **2026**, *10*, 100529. <https://doi.org/10.1016/j.caeai.2025.100529>.
6. Giannakos, M.; Azevedo, R.; Brusilovsky, P.; Cukurova, M.; Dimitriadis, Y.; Hernandez-Leon, D.; Järvelä, S.; Mavrikis, M.; Rienties, B. The Promise and Challenges of Generative AI in Education. *Behav. Inf. Technol.* **2025**, *44*, 2518–2544. <https://doi.org/10.1080/0144929X.2024.2394886>.
7. O’Keeffe, A.; Mark, G. The English Grammar Profile of Learner Competence: Methodology and Key Findings. *Int. J. Corpus Linguist.* **2022**, *22*, 457–489. <https://doi.org/10.1075/ijcl.14086.oke>.

8. Ying, C.; Wang, H.; Jin, H.; Li, Y.; Liu, Y. (Eds.) *Chinese Proficiency Grading Standards for International Chinese Language Education Grammar Learning Manual (Elementary, Intermediate, Advanced)*; Beijing Language and Culture University Press: Beijing, China, 2022.
9. Sunakawa, Y. *A Handbook of Japanese Grammar Patterns for Teachers and Learners*; Kurosio Publishers: Tokyo, Japan, 2015.
10. Odlin, T. (Ed.) *Perspectives on Pedagogical Grammar*; Cambridge Applied Linguistics, Cambridge University Press: Cambridge, UK, 1994. <https://doi.org/10.1017/CBO9781139524605>.
11. Wang, D. Evaluation of Large Language Models' Foreign Language Teaching Ability: An Experimental Study Focusing on Pedagogical Grammar. In Proceedings of the The 103rd Language and Speech Understanding and Dialogue Processing Study Group, Tokyo, Japan, 20–22 March 2025; Volume 103, pp. 80–85. https://doi.org/10.11517/jsaislud.103.0_80.
12. Andrews, S. Why do L2 teachers need to 'know about language'? Teacher metalinguistic awareness and input for learning. *Lang. Educ.* **1999**, *13*, 161–177. <https://doi.org/10.1080/09500789908666766>.
13. Andrews, S. *Teacher Language Awareness*, 1 ed.; Cambridge University Press: Cambridge, UK, 2007. <https://doi.org/10.1017/CBO9780511497643>.
14. Wang, W.; Yan, Y. A review of teacher language awareness (2015–2024): Current trends and future directions. *J. Lang. Teach.* **2024**, *4*, 10–20. <https://doi.org/10.54475/jlt.2024.023>.
15. Andrews, S.; Svalberg, A.M.L. Teacher Language Awareness. In *Language Awareness and Multilingualism*; Cenoz, J., Gorter, D., May, S., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 219–231. https://doi.org/10.1007/978-3-319-02240-6_17.
16. Andrews, S. The language awareness of the L2 teacher: Its impact upon pedagogical practice. *Lang. Aware.* **2001**, *10*, 75–90. <https://doi.org/10.1080/09658410108667027>.
17. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018*; Linzen, T.; Chrupała, G.; Alishahi, A., Eds.; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 353–355. <https://doi.org/10.18653/v1/W18-5446>.
18. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019*; Curran Associates, Inc.: New York, USA 2019, Volume 32, pp. 3266–3280. <https://doi.org/10.48550/arXiv.1905.00537>.
19. Xu, L.; Hu, H.; Zhang, X.; Li, L.; Cao, C.; Li, Y.; Xu, Y.; Sun, K.; Yu, D.; Yu, C.; et al. CLUE: A Chinese Language Understanding Evaluation Benchmark. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 13–18 September 2020; pp. 4762–4772. <https://doi.org/10.18653/v1/2020.coling-main.419>.
20. Xu, L.; Li, A.; Zhu, L.; Xue, H.; Zhu, C.; Zhao, K.; He, H.; Zhang, X.; Kang, Q.; Lan, Z. SuperCLUE: A Comprehensive Chinese Large Language Model Benchmark. *arXiv* **2023**, arXiv:2307.15020.
21. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring Massive Multitask Language Understanding. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 26 April–1 May 2020. <https://doi.org/10.48550/arXiv.2009.03300>.
22. Srivastava, A.; Rastogi, A.; Rao, A.; Shole, A.A.M.; Abid, A.; Fisch, A.; Brown, A.R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Trans. Mach. Learn. Res.* **2022**, *2023*, 1–95. <https://doi.org/10.48550/arXiv.2206.04615>.
23. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 2567–2577. <https://doi.org/10.18653/v1/D19-1259>.
24. Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.H.; Routledge, B.; et al. FINQA: A Dataset of Numerical Reasoning over Financial Data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3697–3711.
25. Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; Zambrano, D.; et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 44123–44279.
26. Hou, J.; Ao, C.; Wu, H.; Kong, X.; Zheng, Z.; Tang, D.; Li, C.; Hu, X.; Xu, R.; Ni, S.; et al. E-EVAL: A Comprehensive Chinese k-12 Education Evaluation Benchmark for Large Language Models. *arXiv* **2024**, arXiv:2401.15927.
27. Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; et al. C-EVAL: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 62991–63010.

28. National Language Commission. *Chinese Proficiency Grading Standards for International Chinese Language Education*; Beijing Language and Culture University Press: Beijing, China, 2021.
29. Team Qwen. Qwen2.5 Technical Report. *arXiv* **2024**, arXiv:2412.15115. <https://doi.org/10.48550/arXiv.2412.15115>.
30. Team GLM.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv* **2024**, arXiv:2406.12793. <https://doi.org/10.48550/ARXIV.2406.12793>.
31. Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; et al. InternLM2 Technical Report. *arXiv* **2024**, arXiv:2403.17297. <https://doi.org/10.48550/ARXIV.2403.17297>.
32. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. <https://doi.org/10.48550/ARXIV.2407.21783>.
33. DeepSeek-AI. DeepSeek-V3 Technical Report. *arXiv* **2025**, arXiv:2412.19437. <https://doi.org/10.48550/arXiv.2412.19437>.
34. OpenAI. GPT-4o-2024-08-06. 2024. Available online: <https://platform.openai.com/docs/models/gpt-4o> (accessed on 1 June 2025).
35. OpenAI. GPT-4o-mini-2024-07-18. 2024. Available online: <https://platform.openai.com/docs/models/gpt-4o-mini> (accessed on 1 June 2025).
36. Doubao Team. Doubao 1.5 Pro. 2025. Available online: https://team.doubao.com/en/special/doubao_1_5_pro (accessed on 1 June 2025).
37. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 9459–9474.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.