科技与中文教学

Journal of Technology and Chinese Language Teaching

Volume 17 Number 1, June 2026
二〇二六年六月　第十七卷第一期

ISSN: 1949-260X
http://www.tclt.us/journal

科技与中文教学

Journal of Technology and Chinese Language Teaching

A peer-reviewed online publication with in-print supplement
ISSN: 1949-260X http://www.tclt.us/journal

Volume 17 Number 1, June 2026

Managing editor for this issue: Jun Da

Articles

Columns

科技与中文教学

Journal of Technology and Chinese Language Teaching

A peer-reviewed online publication with in-print supplement
ISSN: 1949-260X http://www.tclt.us/journal

Sponsor

Department of World Languages, Literatures, and Cultures, Middle Tennessee State University

Contacts

URL: http://www.tclt.us/journal
Email: editor@tclt.us

Journal of Technology and Chinese Language Teaching Volume 17 Number 1, June 2026
http://www.tclt.us/journal/2026v17n1/linzuckerbellassen.pdf pp. 1-29

Prompting Large Language Models for CEFR-EBCL-Aligned Chinese L2 Learning: An Empirical Study of Sinographic Constraint Compliance
(透過提示工程引導大型語言模型進行符合 CEFR–EBCL 標準的漢語第二語言學習：一項關於漢字限制遵循性的實證研究)

Lin-Zucker, Miao
(林季苗)
University of Jean Moulin Lyon 3
miao.lin-zucker@univ-lyon3.fr

Bellassen, Joël
(白樂桑)
Institut National des Langues et Civilisations Orientales
joel.bellassen@inalco.fr

Zucker, Jean-Daniel
(諸葛梁)
IRD/Sorbonne University
jean-daniel.zucker@ird.fr

Abstract: Large Language Models (LLMs) are increasingly used in Chinese as a Second Language (L2) learning, yet their ability to comply with pedagogical constraints specific to the Chinese writing system remains underexplored. This study examines whether system prompts aligned with the CEFR–EBCL framework enable LLMs to generate learner-facing Chinese texts that respect sinographic thresholds at the A1, A1+ and A2 levels. We conducted controlled experiments using two models (GPT-4.1 and GPT-4.1-mini) across ten EBCL-related written tasks. Prompt conditions with and without explicit character lists were compared. Model outputs were automatically analyzed to quantify instruction deviation, defined as the proportion of characters outside the target EBCL set. Results indicate that including explicit character lists significantly reduces out-of-list character production at the A1 and A1+ levels, particularly with GPT-4.1. At the A2 level, this effect becomes marginal. These findings provide empirical evidence on the pedagogical value and limits of prompt-based control of ChatGPT outputs for CEFR–EBCL-aligned Chinese L2 learning.

摘要：大型語言模型（LLMs）在漢語作為第二語言學習中的應用日益普及，但其是否能有效遵循漢字書寫系統所特有的教學限制，仍缺乏實證研究。本研究探討在 CEFR–EBCL 框架下，系統提示是否能引導大型語言模型在 A1、A1+ 與 A2 級別生成符合漢字門檻的漢語學習文本。研究以 GPT-4.1 與 GPT-4.1-mini 兩種模型為對象，圍繞十項 EBCL 書面語言任務進行受控實驗，比較提示中是否提供明確漢字列表的差異，並以「指令偏離度」量化模型輸出中超出目標漢字集合的比例。結果顯示，在 A1 與 A1+ 級別中，加入漢字列表能顯著降低不符合門檻的漢字生成比例，而在 A2 級別中，此效果趨於有限。本研究為基於提示工程控制 ChatGPT 輸出、以支援對應 CEFR–EBCL 標準的漢語二語學習，提供了實證依據。

Keywords: Chinese as a Second Language, Large Language Models, Prompt Engineering, CEFR–EBCL Alignment, Sinographic Constraints

關鍵詞: 漢語第二語言，大型語言模型，提示工程，對應 CEFR–EBCL 標準，漢字限制

1. Introduction

ChatGPT is currently one of the most widely used large language model–based chatbots for communication, information seeking, and learning activities (B. Li et al., 2024). In the field of second language acquisition, and particularly in the teaching and learning of Chinese as a Foreign Language (CFL), the rapid diffusion of large language models (LLMs) has renewed long-standing debates on the role of technology in language pedagogy, learner autonomy, and individualized learning trajectories (Glaser, 2023; Imran & Almusharraf, 2023). Unlike earlier conversational agents, LLM-based systems such as ChatGPT are capable of producing extended, coherent, and context-sensitive discourse while dynamically adjusting lexical and syntactic choices. These properties make them attractive in language learning contexts, where exposure, interaction, and feedback play central roles in the acquisition process (Wang et al., 2025).

When applied to Chinese language learning, however, the integration of LLM-based chatbots raises issues that cannot be reduced to those observed for alphabetic languages. Chinese combines two core linguistic units—the word and the character—and two partially dissociated strata of competence: oral-lexical competence and graphic-sinographic competence. This structural specificity poses a significant challenge to aligning Chinese language teaching with proficiency frameworks originally designed for phonographic languages. Within the European educational context, this challenge has been addressed through the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) and, more specifically, for Chinese, through the European Benchmarks for the Chinese Language (EBCL) project (Guder, 2014, 2015). A central contribution of the EBCL project is the explicit introduction of sinographic thresholds—250 characters for A1-aligned competencies, 320 for A1+, and 630 for A2-aligned competencies.

Building on this framework, the present study addresses the following research question: to what extent can a general-purpose large language model such as ChatGPT be constrained, through carefully designed system prompts, to generate written Chinese aligned with CEFR- and EBCL-defined sinographic thresholds? Rather than evaluating ChatGPT as a global pedagogical tool, our focus is on instruction adherence, operationalized as instruction deviation, that is, the degree to which the model respects explicit constraints on character usage imposed by the prompt.

Our working hypothesis is that prompt engineering—particularly the use of system prompts specifying pedagogical roles and explicit character lists—can partially compensate for the absence of fine-tuning and enable LLMs to function as CEFR–EBCL-aligned, level-aware tutoring systems for Chinese learners (Ekin, 2023; Pryzant et al., 2023). By crossing lexical recurrence with sinographic recurrence, such prompts may help ensure rich interaction while maintaining strict control over written input.

This article makes three main contributions. First, it situates the use of generative AI for Chinese language learning within the CEFR framework, with a specific focus on written competencies. Second, it proposes and tests a set of system prompts designed for competencies aligned with A1, A1+, and A2 descriptors, with and without explicit character lists. Third, it provides a quantitative analysis of instruction deviation based on large-scale experiments conducted with two ChatGPT models (GPT-4.1 and GPT-4.1-mini) across multiple EBCL-aligned reading and writing tasks.

The remainder of the article is structured as follows. Section 2 reviews related work on LLM-based language instruction and situates the study within CEFR- and EBCL-aligned research on Chinese as a Foreign Language. Section 3 introduces the linguistic and pedagogical framework underlying the study, focusing on the sinographic specificity of Chinese. Section 4 presents the methodology and experimental design. Sections 5 and 6 report and discuss the experimental results. Finally, Section 7 concludes the article by outlining limitations and perspectives for future research.

2. Background and Related Work: Generative AI and Chinese L2 within CEFR-EBCL

This section situates the present study within recent research on large language model–based chatbots for language learning, with a specific focus on Chinese as a Foreign Language (CFL) in CEFR- and EBCL-aligned contexts. Rather than providing a historical overview of conversational agents, we focus on developments that directly motivate our prompt-based experimental approach.

2.1 From rule-based chatbots to LLM-based systems in language learning

Conversational agents have been explored for language-related purposes since the early days of artificial intelligence. Early systems such as ELIZA (Weizenbaum, 1966) and later rule-based chatbots such as ALICE (Wallace, 2009) demonstrated that scripted dialogue could support limited forms of interaction. However, their pedagogical impact remained constrained by shallow contextual memory, rigid rule sets, and extensive manual scripting.

A major paradigm shift occurred with the emergence of large language models (LLMs) trained on massive text corpora and built on deep neural architectures. Models such as GPT-3 and its successors introduced the ability to generate extended, coherent, and context-sensitive discourse, marking a clear rupture with earlier rule-based approaches (Kalyan, 2024). As noted by Adamopoulou and Moussiades (2020), most surveys of chatbots published prior to this period do not account for this transformation, underscoring the novelty of the current landscape for language learning research.

In the context of CFL, this shift is particularly significant, as LLMs can dynamically generate written Chinese. This generative capacity raises a central question for pedagogy: whether such output can be constrained in ways that remain compatible with established educational principles and proficiency frameworks.

2.2 LLM-based chatbots and language learning: recent empirical trends

The rapid diffusion of LLM-based chatbots has generated a growing body of empirical research in language education (Cong, 2024). Meta-analyses and systematic reviews suggest that chatbot-assisted language learning generally yields positive effects compared with non-chatbot conditions, particularly in terms of learner engagement, exposure to input, and perceived usefulness (Huang et al., 2022; Labadze et al., 2023; Wang et al., 2025).

Focusing specifically on ChatGPT, B. Li et al. (2024) provide a systematic review of the first year of publications on ChatGPT and language education. They identify recurring themes such as personalization, feedback quality, ethical concerns, and academic integrity, while also noting that relatively little attention has been paid to the relationship between chatbot-generated language and externally defined proficiency descriptors.

Taken together, this literature establishes the relevance of LLMs for language learning, but it also reveals a methodological gap: most studies evaluate learner performance or perceptions, rather than analyzing the extent to which models can be constrained to produce language aligned with explicit pedagogical frameworks.

2.3 ChatGPT in Chinese as a Foreign Language research

Within the broader field of LLM-assisted language learning, a growing number of studies focus specifically on Chinese. Research has shown that ChatGPT can support conversational practice, grammar learning, critical thinking, and differentiated instruction in CFL contexts (Jiang et al., 2024; B. Li et al., 2024; J. Li et al., 2023; Zhao et al., 2024).

While these studies converge in recognizing ChatGPT as a flexible pedagogical tool, prompts are typically treated as task-level instructions, and linguistic output is evaluated indirectly through learner outcomes. The question of whether ChatGPT can reliably respect explicit linguistic constraints—particularly those related to the Chinese writing system—remains largely unexplored.

The present study departs from this learner-centered perspective by shifting the analytical focus to model behavior and by treating prompts as objects of systematic experimental manipulation.

3. Generative AI for Chinese as a Foreign Language: A Sinographic Perspective

This section outlines the linguistic and pedagogical framework that underpins the experimental design, focusing on the specificity of Chinese writing and its articulation with CEFR- and EBCL-aligned proficiency descriptors.

3.1 The pedagogical specificity of Chinese writing

Chinese language teaching is characterized by a structural specificity that distinguishes it from the teaching of alphabetic languages: the coexistence of two minimal units, the word and the character, and two partially dissociated strata of competence, oral-lexical competence and graphic-sinographic competence. This structural specificity has long been recognized in the pedagogy of Chinese (Bellassen, 1989, 2009, 2018, 2024; DeFrancis et al., 1966; Guo, 1985) and poses a major challenge to aligning Chinese teaching with proficiency frameworks originally designed for phonographic languages.

The EBCL project did not emerge in a vacuum. In several European educational contexts, the integration of Chinese into CEFR-aligned language curricula from the early 2000s onward raised fundamental questions regarding the “eurocompatibility” of Chinese as a non-alphabetic language. In response, European scholars and curriculum designers progressively developed approaches that explicitly recognized sinographic competence as a prerequisite for written proficiency, notably through character thresholds. These early initiatives anticipated the dualistic perspective later formalized by the EBCL framework, which dissociates the progression of oral and written competences in Chinese and adapts CEFR descriptors accordingly (Bellassen & Zhang, 2008; Bellassen, 2012; Zhang-Colin & Gianninoto, 2022; Lin-Zucker, 2024).

This dissociation underlies long-standing debates in CFL pedagogy between so-called monistic approaches, which treat characters as purely instrumental representations of spoken language, and dualistic approaches, which regard sinographic competence as a core, autonomous dimension of language learning. Dualistic approaches emphasize character frequency, combinatorial capacity, and lexical recurrence as organizing principles for the progression of written Chinese.

3.2 Chinese within the CEFR and the EBCL framework

The integration of Chinese into CEFR-based language education has brought these theoretical issues into sharper focus. The CEFR, originally designed for phonographic languages, devotes limited attention to orthographic competence and does not provide operational tools for handling logographic writing systems.

In response to this limitation, the European Benchmarks for the Chinese Language (EBCL) project explicitly introduced graphemic competence as a prerequisite for written reception and production (Guder, 2014). By defining sinographic thresholds—250 characters for competencies aligned with A1 descriptors, 320 for A1+, and 630 for those aligned with A2 descriptors—the EBCL framework operationalizes the principle that written competence in Chinese must be explicitly constrained and scaffolded (Guder, 2015).

These thresholds provide a concrete basis for evaluating written input and output in CFL contexts and are central to the experimental design adopted in the present study.

The present study explicitly assumes that the EBCL character inventories constitute a pedagogically desirable reference point. This assumption is grounded in the EBCL tradition of controlled sinographic input (Guder, 2014; Bellassen, 2018), in which limiting and sequencing the character load is held to support readability and to reduce cognitive overload for beginning learners. Readers working outside the EBCL tradition (for instance, within purely communicative or frequency-driven approaches) may not share this assumption. The study, therefore, does not claim that character-list compliance in itself improves learning or readability; it evaluates whether generative models can be made to respect an externally defined pedagogical standard, while the empirical validation of that standard, for example, through measures of learner performance, is left to future research.

3.3 Prompt-based control as a pedagogical interface

Prompt engineering has recently been conceptualized as a form of natural language programming that conditions LLM behavior through explicit instructions specifying tasks, roles, and constraints (Liu et al., 2023). Among the different prompt types, system prompts play a central role, as they define the model's behavior across an interaction session, including pedagogical role and linguistic boundaries.

From a language education perspective, prompt-based control offers a practical alternative to fine-tuning. While fine-tuning requires access to model weights, training data, and computational resources that are generally unavailable in educational contexts, prompt engineering allows teachers and researchers to shape model behavior through transparent, reproducible, and easily adjustable instructions.

In the context of CFL, prompt-based control makes it possible to impose explicit sinographic constraints aligned with CEFR- and EBCL-defined descriptors, thereby coupling lexical recurrence with sinographic recurrence. This approach directly addresses one of the central pedagogical tensions in Chinese language learning: reconciling rich interaction with controlled progression in written input.

4. Methodology

This section presents the experimental design used to evaluate the extent to which large language models (LLMs) can be constrained through system prompts to generate written Chinese aligned with CEFR–EBCL sinographic thresholds. We describe the research questions, task selection, prompt design, model configurations, and the quantitative metric used to assess adherence to instructions. Technical terms related to large language models and prompt engineering are defined in Appendix B.

4.1 Research questions

The study addresses the following research questions:

- RQ1. To what extent does the explicit inclusion of EBCL-aligned character lists in system prompts reduce instruction deviation in LLM-generated Chinese output?

- RQ2. Does the effect of explicit character lists vary across proficiency levels (A1, A1+, A2)?

- RQ3. Do model size and architecture (GPT-4.1 vs. GPT-4.1-mini) influence the degree of instruction deviation under identical prompt conditions?

These questions reflect the central objective of the study: to assess whether prompt-based control alone, without fine-tuning, is sufficient to align LLM output with externally defined sinographic constraints.

4.2 Task selection and EBCL alignment

To ensure pedagogical relevance and CEFR compatibility, experimental tasks were selected from the European Benchmarks for the Chinese Language (EBCL) framework. We focused exclusively on tasks involving written reception and written production, as sinographic constraints primarily affect written competencies.

The selected tasks correspond to EBCL descriptors related to reading comprehension and written expression at levels A1, A1+, and A2, including activities such as overall reading comprehension, reading correspondence, and the production of short written messages and forms. These tasks reflect common learning objectives in early-stage Chinese language instruction and provide a controlled context for evaluating written output. In this work, the sets of characters associated with each proficiency level are defined not only according to the EBCL/CEFR descriptors but also with reference to the actual distribution of character usage frequencies in a standard corpus. The use of character frequency data (for example, Da, 2004) have allowed to quantitatively rationalize the target character lists, ensuring that each threshold reflects distinct levels of text coverage. The complete EBCL-aligned character lists for each proficiency level (A1: 250 characters, A1+: 320 characters, A2: 630 characters) are provided in Appendix D.

Ten EBCL-aligned written tasks were selected for this study, distributed across three categories: reading comprehension (RW1–RW5), written production (PW1–PW2), and written interaction (IW1–IW3). These task types represent core written activities encountered by beginner-level learners in authentic communicative situations. Each task was systematically instantiated across all three proficiency levels (A1, A1+, A2), yielding 30 distinct task-level combinations. The complete list of task names and their integration into the system prompts is provided in Appendix C.

4.3 Prompt design and experimental conditions

The experiment relies exclusively on system prompts, understood as initial instructions that define the chatbot's pedagogical role and constrain its behavior throughout the interaction session. Unlike user prompts, which vary during interaction, system prompts provide a stable experimental condition.

For each proficiency level (A1, A1+, A2), two system prompt conditions were designed:

- Condition L (List): the system prompt explicitly includes a list of target characters corresponding to the EBCL sinographic threshold for the level.
- Condition NL (No List): the system prompt specifies the target level and instructs the model to remain within its character threshold, but without providing an explicit character list.

In both conditions, the system prompt assigns the model the role of a Chinese language tutor and constrains its output to written Chinese aligned with the specified EBCL level. The prompts also include a self-verification instruction requiring the model to rephrase its output if characters outside the allowed set are detected.

The full system prompt used for levels A1, A1+, and A2 is provided in Appendix C.

4.4 Models and experimental setup

Two versions of ChatGPT were evaluated in this study: GPT-4.1 and GPT-4.1-mini. These models were deliberately selected as widely deployed, general-purpose LLMs that are currently accessible to students and teachers through the standard ChatGPT interface, either in free or subscription-based usage. As such, they reflect the models that learners are most likely to encounter in authentic educational settings. The objective of this study is not to benchmark state-of-the-art systems, but to examine whether differences in model capacity, within realistically accessible models, affect compliance with explicit sinographic constraints under controlled prompting conditions.

GPT-4.1 and GPT-4.1-mini provide a meaningful contrast: they share the same general architecture and instruction-following paradigm, while differing in computational capacity, response stability, and cost—factors that are directly relevant for pedagogical deployment. To ensure scalability and reproducibility, all outputs were generated programmatically via the OpenAI API (Python SDK, standard endpoint). These API-based generations were designed to faithfully reproduce the outputs a student would obtain through the standard ChatGPT user interface, given identical system prompts and generation parameters, thereby combining experimental rigor with pedagogical realism.

For each combination of proficiency level (A1, A1+, A2), prompt condition (with or without a character list), task type (RW, PW, IW), and model, ten generations were produced per condition to mitigate the stochastic variability inherent to autoregressive language models. Data were generated in early 2026. All experiments were conducted with identical generation parameters, with temperature set to 0.7 to approximate realistic learner-facing usage rather than maximal constraint satisfaction. Deterministic decoding (temperature = 0) was therefore not included, as the objective was to capture typical instructional behavior; a systematic ablation over decoding temperatures is left for future work. The maximum number of generated tokens was set to 1000, top_p to 1.0, and both frequency_penalty and presence_penalty to 0.0. No fine-tuning, retrieval-augmented generation, or external tools were employed in order to isolate the effect of prompt-based control alone, which is the primary mechanism available to learners and teachers in practice. All system prompts, character lists, task templates, and evaluation scripts are provided in the appendices to ensure full reproducibility of the experimental setup.

Statistical comparisons between models and prompt conditions were conducted using Welch's t-test for independent samples with unequal variances assumed. Significance levels are reported as follows: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and not significant for $p \geq 0.05$.

It should be noted that the ten generations obtained for each condition originate from the same underlying model and prompt, and therefore do not constitute fully independent observations in the strict statistical sense. Treating repeated generations as independent samples may overestimate the effective sample size. The significance levels reported below should accordingly be interpreted as indicative characterizations of within-condition variability rather than as formal population-level inferences; this limitation is further discussed in Section 6.5.

4.5 Evaluation metric: instruction deviation

To quantitatively assess adherence to instruction, we introduce the instruction deviation metric. Instruction deviation is defined as the proportion of characters in the model's output that do not belong to the target EBCL character set for the specified proficiency level. Formally, for a generated output containing N characters, of which k characters fall outside the allowed EBCL set, instruction deviation is computed as k/N. Character-level analysis was performed automatically by comparing each generated character with the reference character lists for levels A1, A1+, and A2. This metric enables fine-grained evaluation of constraint compliance independently of semantic adequacy or pedagogical quality.

Instruction deviation was computed for each generated output and aggregated by condition, proficiency level, and model. This approach enables direct comparison between prompt conditions and model versions, and provides a quantitative basis for evaluating the effectiveness of prompt-based control. For deviation analysis, only CJK Unified Ideographs were counted as characters. Punctuation marks, Arabic numerals, Latin letters, whitespace, and formatting symbols were excluded from all character-level measurements. Deviation rates were computed on the final output presented to the learner after the self-verification step. This choice reflects a pedagogical perspective, where compliance of the delivered content—rather than intermediate drafts—is the relevant criterion.

5. Results

This section reports the quantitative results of the experiments described in Section 4. Results are presented descriptively, without interpretative discussion, which is reserved for Section 6.

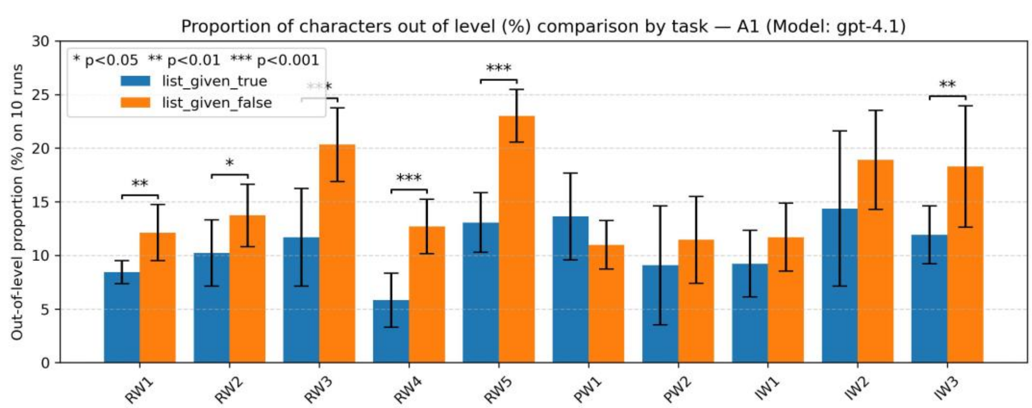
5.1 Overview of experimental data

The experimental design crossed four factors: model (GPT-4.1, GPT-4.1-mini), proficiency level (A1, A1+, A2), task type (ten EBCL-aligned written tasks), and prompt condition (with explicit character list vs. without list). For each combination, ten generations were produced (10 runs per condition), yielding a total of 1,200 model outputs (600 per model). The key metric is the out-of-level ratio: the percentage of Chinese characters in the model's response that fall outside the authorized EBCL character set for the target level.

For each output, instruction deviation was computed as defined in Section 4.5. Mean instruction deviation values and standard deviations were calculated for each condition and aggregated by level and model.

5.2 Effect of explicit character lists on instruction deviation

For the GPT-4.1 model (Figure 1 below), the inclusion of a character list is associated with a marked reduction in instruction deviation at levels A1 and A1+. With the character list, GPT-4.1 achieves a mean out-of-level ratio of 10.74% at A1 level (vs. 15.32% without list, Δ=4.6 pp), 8.85% at A1+ level (vs. 12.01% without list, Δ=3.2 pp), and 3.91% at A2 level (vs. 3.92% without list, Δ≈0 pp). Statistical significance tests show that 6 out of 10 tasks at A1 level and 6 out of 10 tasks at A1+ level exhibit significant improvement ($p<0.05$) when the list is provided. At level A2, instruction deviation remains low in both conditions, and the difference between the two conditions is negligible.



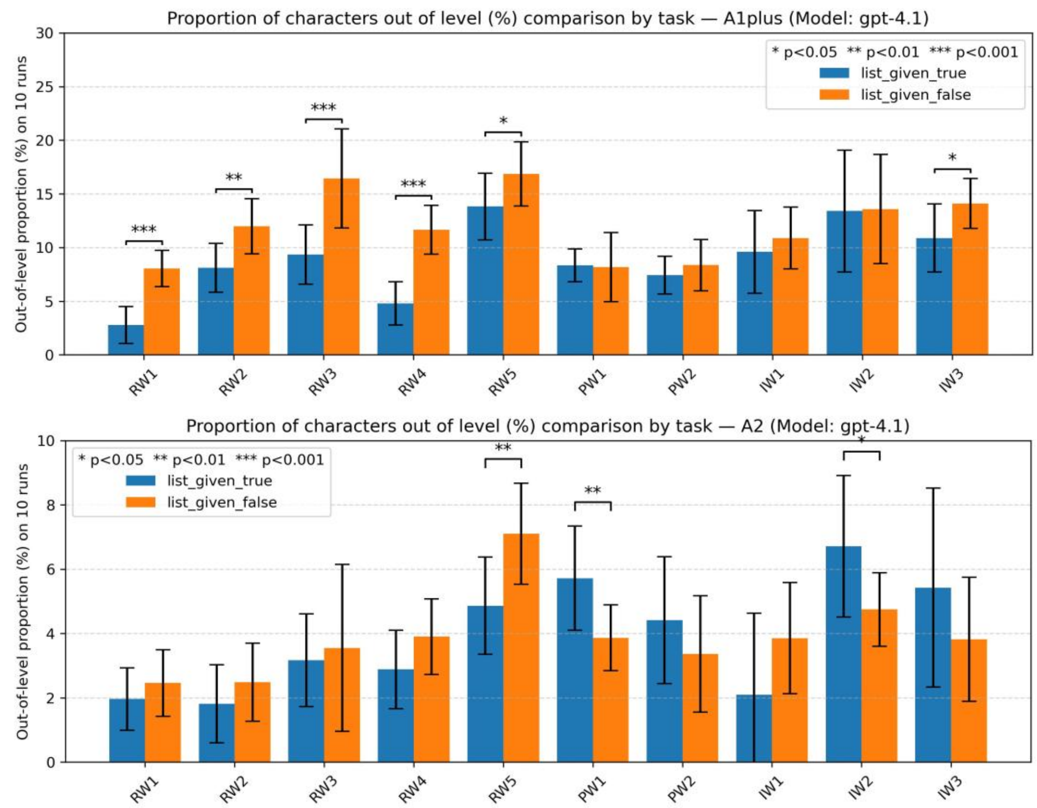


Figure 1 Effect of providing an EBCL-aligned character list on the proportion of out-of-level characters generated by the GPT-4.1 model across tasks and proficiency levels. For each task (RW1–RW5, PW1–PW2, IW1–IW3), results compare conditions with and without an explicit character list in the system prompt. Bars represent the mean percentage of characters not belonging to the target level, averaged over 10 runs; error bars indicate standard deviation. Statistical significance between conditions is indicated as follows: p < 0.05, p < 0.01, p < 0.001. Results are shown separately for A1 (top), A1+ (middle), and A2 (bottom) levels.

For the GPT-4.1-mini model (Figure 2 below), instruction deviation values are generally higher than for GPT-4.1 across all levels. With the character list, GPT-4.1-mini achieves a mean out-of-level ratio of 15.55% at A1 level (vs. 18.59% without list), 12.30% at A1+ level (vs. 14.40% without list), and 5.29% at A2 level (vs. 7.18% without list). The effect of providing the character list is less pronounced than for GPT-4.1, with only 3-4 out of 10 tasks showing statistically significant improvement at each level.

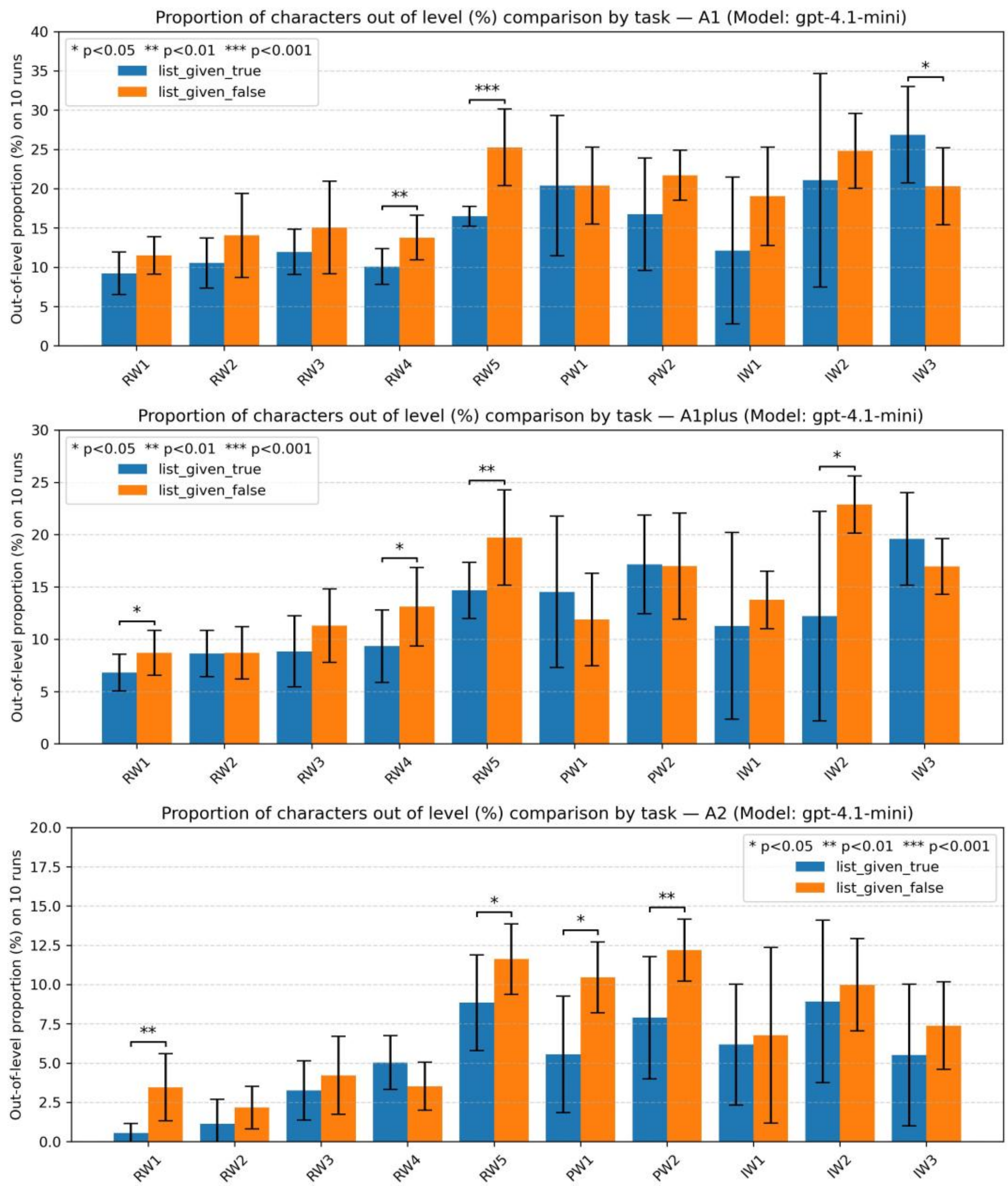


Figure 2 Effect of providing an EBCL-aligned character list on the proportion of out-of-level characters generated by the GPT-4.1-mini model across tasks and proficiency levels. For each task (RW1–RW5, PW1–PW2, IW1–IW3), results compare conditions with and without an explicit character list in the system prompt. Bars represent the mean percentage of characters not belonging to the target level, averaged over 10 runs; error bars indicate standard deviation. Statistical significance between conditions is indicated as follows: p < 0.05, p < 0.01, p < 0.001. Results are shown separately for A1 (top), A1+ (middle), and A2 (bottom) levels.

5.3 Comparison between GPT-4.1 and GPT-4.1-mini

Across all proficiency levels and prompt conditions, GPT-4.1 consistently exhibits lower instruction deviation than GPT-4.1-mini, indicating a stronger ability to comply with explicit sinographic constraints. When character lists are provided, GPT-4.1 achieves a global mean out-of-level ratio of 7.8%, compared to 11.0% for GPT-4.1-mini, corresponding to a 3.2 percentage point advantage. This gap is especially marked at the A1 and A1+ levels, where GPT-4.1 systematically produces outputs containing fewer out-of-list characters across most task types.

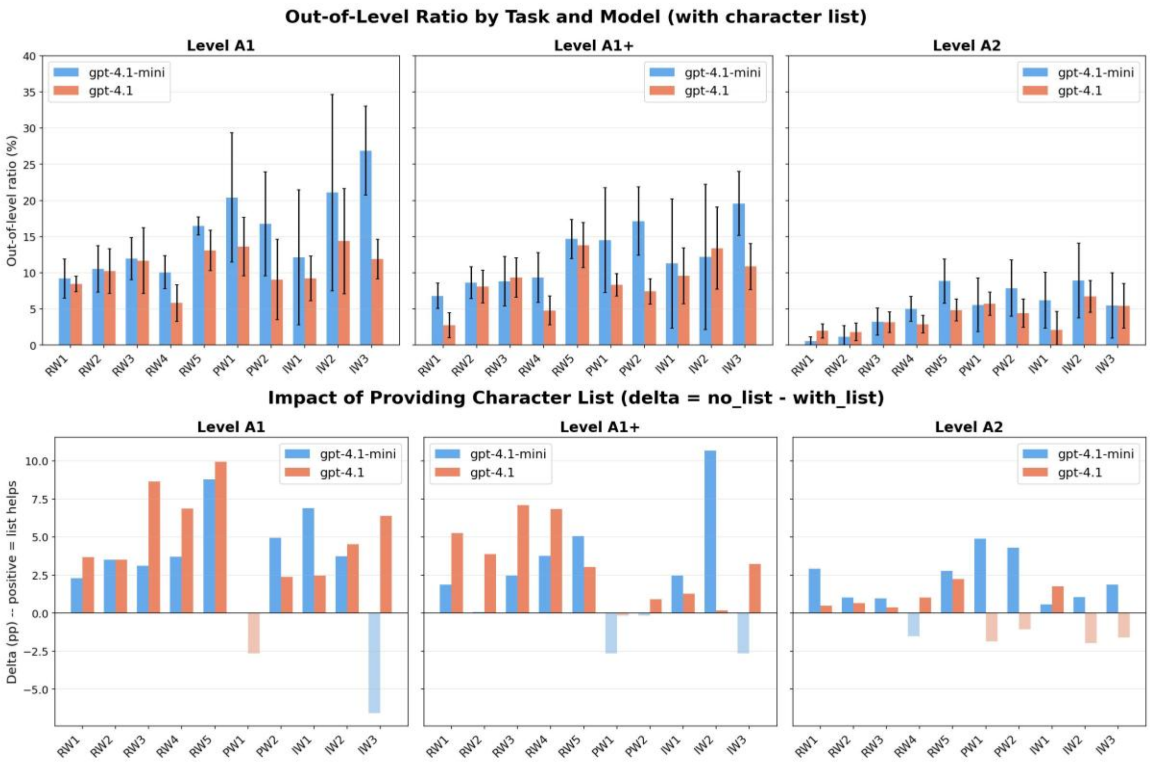


Figure 3 Comparison between GPT-4.1 and GPT-4.1-mini across tasks and proficiency levels. The top panel shows the mean proportion of out-of-level characters produced with an EBCL-aligned character list for levels A1, A1+, and A2 (means over 10 runs; error bars = standard deviation). The bottom panel shows the effect of the character list, expressed as the difference between no-list and with-list conditions (delta = no-list − with-list), by task, level, and model.

Beyond average performance, the two models differ in reliability and stability. GPT-4.1 shows perfect reliability, with a 100% success rate across all 600 runs, whereas GPT-4.1-mini presents a lower success rate (96.2%), with 23 failed generations producing no Chinese characters at all. These failures are not uniformly distributed but are concentrated in interactive writing tasks (IW2, IW3) at the A1 level, suggesting greater fragility of the smaller model when simultaneously handling interactional structure and strict character-level constraints.

At the A2 level, instruction deviation remains relatively low for both models, regardless of prompt condition (GPT-4.1: 3.91%; GPT-4.1-mini: 5.29% with character list). However, GPT-4.1 continues to demonstrate greater stability across repeated generations, as reflected by a substantially lower variance (mean standard deviation of 3.9% compared to 6.0% for GPT-4.1-mini with character lists). This indicates that, even when mean performance converges, GPT-4.1 yields more predictable and consistent outputs, a property of particular importance in pedagogical contexts.

Figure 3 (top) summarizes the comparison between GPT-4.1 and GPT-4.1-mini in terms of mean out-of-level ratios across tasks and CEFR-EBCL levels when a character list is provided. Figure 3 (bottom) complements this analysis by explicitly visualizing the impact of the character list for each model, expressed as the difference between the no-list and with-list conditions. Together, these two panels highlight both GPT-4.1's superior overall performance and its greater responsiveness to prompt-based constraint reinforcement.

5.4 Summary of quantitative findings

The quantitative analyses yield several consistent patterns. At the A1 and A1+ levels, the inclusion of explicit EBCL-aligned character lists in the system prompt leads to a substantial reduction in instruction deviation for GPT-4.1, with mean improvements of 4.6 percentage points at A1 and 3.2 percentage points at A1+. By contrast, at the A2 level, instruction deviation remains low overall, and the presence of a character list has a limited effect on model behavior.

Across all proficiency levels and experimental conditions, GPT-4.1 exhibits greater robustness to explicit sinographic constraints than GPT-4.1-mini, both in terms of lower mean out-of-level ratios and reduced variability across runs. For GPT-4.1-mini, providing a character list nonetheless yields a modest but consistent improvement, with an average reduction in instruction deviation of approximately 2.3 percentage points across levels.

6. Discussion

This section interprets the results presented above, discusses their pedagogical and methodological implications, and outlines the study's limitations.

6.1 Interpreting instruction deviation across levels

The strong effect of explicit EBCL-aligned character lists at the A1 and A1+ levels indicates that, at early stages of Chinese learning, LLMs benefit from concrete and exhaustive representations of sinographic constraints. With lists provided, GPT-4.1 shows substantial reductions in instruction deviation (4.6 percentage points at A1 and 3.2 at A1+), reflecting the fact that the limited size of the character inventories at these levels (250 characters at A1 and 320 at A1+) makes explicit constraints both manageable and effective.

These findings are consistent with prior work showing that chatbot-assisted learning tends to yield larger effects for beginner learners (Wang et al., 2025), and that system effectiveness depends on alignment with learner proficiency (Huang et al., 2022). At the A2 level, by contrast, instruction deviation remains low regardless of list provision, suggesting that as the character inventory expands (630 characters), models can rely more on internalized frequency distributions. This reduced sensitivity to explicit constraints aligns with pedagogical intuitions and with Zhao et al.'s (2024) observation that prompt-based lexical control becomes increasingly difficult as the constraint space grows.

6.2 Model capacity and constraint compliance

The systematic performance gap between GPT-4.1 and GPT-4.1-mini underscores the role of model capacity in constraint compliance. GPT-4.1 exhibits both lower instruction deviation and greater stability across runs, as well as perfect reliability, whereas GPT-4.1-mini shows higher variance and occasional task failures, particularly in interactive writing tasks at the A1 level.

These differences indicate that prompt-based control strategies cannot be assumed to generalize uniformly across model versions. This observation is consistent with findings in the broader Natural Language Processing (NLP) literature, which show that larger models are better able to handle multiple simultaneous constraints (Liu et al., 2023) and demonstrate improved instruction-following capabilities (Kalyan, 2024). From a pedagogical perspective, this raises equity concerns: as noted by Jeon and Lee (2023), the educational value of AI systems depends critically on output reliability, suggesting that users of smaller models may need additional safeguards or accept higher rates of deviation.

6.3 Qualitative analysis of constrained outputs

Qualitative inspection of model outputs complements the quantitative findings by revealing how instruction deviation manifests in practice. As illustrated in Appendix A, GPT-4.1 generally produces outputs that closely align with EBCL descriptors in terms of task structure and communicative intent, with deviations often limited to isolated lexical choices. Representative examples of model outputs under different constraint conditions are provided in Appendix A.

Such cases—where a character outside the target set is selected despite the availability of a compliant alternative—suggest that instruction deviation frequently results from probabilistic lexical selection rather than from a failure to interpret the constraint. This observation supports the use of instruction deviation as a diagnostic metric, capable of capturing fine-grained mismatches between intended and actual constraint enforcement. Similar patterns of “acceptable but suboptimal” outputs have been reported by J. Li et al. (2023) and by Jiang et al. (2024), who note a tendency for LLMs to privilege fluency and naturalness over strict constraint adherence.

More generally, instruction deviation should be understood as a measure of formal constraint compliance, not of pedagogical quality. A text composed exclusively of in-list characters may still be unnatural, communicatively poor, or developmentally inappropriate for a given learner, just as a text containing a few out-of-list characters may remain perfectly usable in the classroom. Constraint compliance is therefore a necessary but not a sufficient condition for pedagogical adequacy, and the present metric should be complemented in future work by teacher or learner judgments of naturalness, meaningfulness, and task appropriateness.

6.4 Pedagogical implications for CEFR–EBCL-aligned CFL

From a pedagogical standpoint, the results indicate that explicit character lists are particularly beneficial at beginner levels, where they help maintain alignment with carefully scaffolded sinographic progression. At higher levels, however, their diminishing impact suggests that teachers may reasonably prioritize communicative richness over strict character control, adjusting prompt constraints according to instructional objectives.

Crucially, the study demonstrates that meaningful alignment with CEFR–EBCL descriptors can be achieved without fine-tuning, using transparent prompt-based methods accessible to educators. This contributes to ongoing discussions on AI integration in language curricula (B. Li et al., 2024) and supports a hybrid view of "AI as tutor" and "AI as tool" (Labadze et al., 2023), in which LLMs are most effective when operating within well-defined pedagogical boundaries. The effectiveness of explicit lists also echoes the principle of controlled input emphasized in EBCL-aligned approaches (Guder, 2014; Bellassen, 2018).

At the same time, the practical usability of explicit character lists deserves attention. Inserting inventories of 250 to 630 characters into everyday prompts may appear cumbersome, and not all instructors will find it convenient to do so manually. In practice, however, this cost is incurred only once: the EBCL lists are publicly available and can be embedded in reusable prompt templates, stored as persistent custom instructions or project-level settings in mainstream chatbot interfaces, or shared within a teaching team. To support such uses, Appendix E provides a ready-to-use prompting procedure for teachers of A1 and A1+ learners, derived from the system prompts employed in this study.

Beyond classroom practice, the findings also carry implications for the designers of LLM-based educational tools. The results suggest that level-sensitive support for Chinese could be improved if systems were designed to respect explicit sinographic constraints natively, for example by embedding curated character inventories at the system level rather than relying on user-supplied prompts, especially for beginner learners. In this respect, the contribution of the study is not only pedagogical: it also informs the future design of educational AI systems intended to operate within externally defined proficiency frameworks.

6.5 Limitations and ethical considerations

Several limitations must be acknowledged. The study focuses exclusively on written output and does not address spoken production or learning outcomes. Moreover, instruction deviation reflects compliance with formal constraints rather than pedagogical effectiveness. In particular, compliance with a character inventory does not guarantee that the generated text is natural, meaningful, or developmentally appropriate; conversely, the assumption that EBCL inventories are pedagogically desirable, although grounded in the EBCL literature (see Section 3.2), remains open to discussion. These limitations mirror those identified in prior reviews of ChatGPT research in language education (B. Li et al., 2024).

In addition, the present study deliberately focuses on prompt-based constraint control and does not compare against algorithmic constrained decoding approaches such as grid beam search or energy-based decoding. While these methods can be effective, they require model-level access that is not available in closed-weight systems such as ChatGPT. This constraint motivates our prompt-centric approach and reflects the technical and pedagogical conditions under which teachers and learners currently operate.

A further methodological limitation concerns the statistical treatment of repeated generations. As noted in Section 4.4, the ten outputs produced for each condition are drawn from the same underlying system and are not independent observations in the strict sense, which may inflate the effective sample size assumed by Welch's t-tests. The significance levels reported in Section 5 should therefore be read as descriptive indications of within-condition variability rather than as formal population-level inferences.

Moreover, the scope of the conclusions must be clearly delimited. The study examines two closely related OpenAI models accessed through ChatGPT, currently the leading chatbot available to the general public. The findings, therefore, characterize the behavior of this specific model family and should not be generalized to large language models as a class, nor to open-weight or Chinese-specialized systems, whose constraint-compliance behavior remains to be established.

Ethical concerns related to over-reliance, academic integrity, and unequal access to more capable models remain salient (Adel et al., 2024; Cao et al., 2024; Crawford et al., 2023; Vaccino-Salvadore, 2023). While LLMs can serve as powerful support tools, they should complement rather than replace guided instruction. Finally, observed differences in model performance raise issues of equity, as access to more reliable models may not be uniformly available across learners and institutions.

7. Conclusions

This study investigated whether ChatGPT, currently the leading chatbot available to the general public, can be constrained by system prompts to generate written Chinese that aligns with CEFR–EBCL sinographic thresholds. Using instruction deviation as a quantitative metric, we compared prompt conditions with and without explicit character lists across two models (GPT-4.1 and GPT-4.1-mini) and three proficiency levels (A1, A1+, A2).

Three main findings emerge. First, explicit character lists significantly reduce out-of-level character production at beginner levels, with the strongest effects observed at A1 and A1+. Second, this effect diminishes at A2, where instruction deviation remains low regardless of the prompt condition, suggesting that strict sinographic control is most critical in the early stages. Third, model capacity matters: GPT-4.1 consistently outperforms GPT-4.1-mini in both compliance and reliability, indicating that pedagogical recommendations cannot be assumed to generalize across model versions.

These results demonstrate that prompt-based control, without fine-tuning, at least within the GPT-4.1 model family, can achieve meaningful alignment with CEFR–EBCL requirements—offering educators a transparent and accessible method for integrating LLMs into scaffolded Chinese language curricula. The study is limited to written production and does not assess actual learning outcomes, nor the naturalness, meaningfulness, or developmental appropriateness of the generated texts. Future research should extend this approach to other architectures, including open-weight models (LLaMA, Mistral) and Chinese-specialized LLMs (Qwen, Baichuan, Yi, DeepSeek), and examine how constraint compliance translates into measurable learner progress.

A complementary direction we are currently exploring involves models that enforce sinographic constraints by design. Such systems, however, would lack the visibility of mainstream platforms that students already use autonomously. Moreover, occasional exposure to out-of-threshold characters may itself be pedagogically valuable, as authentic language use inevitably involves navigating unfamiliar input—a core dimension of learner autonomy.

As generative AI becomes increasingly embedded in language education, empirical frameworks for evaluating and controlling model behavior will be essential for responsible pedagogical integration. Beyond classroom use, the constraint-compliance perspective developed here may also inform the design of future educational AI systems in which sinographic thresholds are supported natively rather than through user-side prompting.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Adel, A., Ahsan, A., & Davison, C. (2024). ChatGPT promises and challenges in education: Computational and ethical perspectives. *Education Sciences, 14*(8), 814. https://doi.org/10.3390/educsci14080814

Bellassen, J. (1989). *Méthode d'initiation à la langue et à l'écriture chinoises*. La Compagnie.

Bellassen, J. (2009). *Snowballing Chinese*. Sinolingua Press.

Bellassen, J. (2012). Is Chinese eurocompatible? Is the Common Framework common? The Common European Framework of Reference for Languages facing distant languages. In N. Tomimori, M. Furihata, K. Haida, N. Kurosawa, & M. Negishi (Eds.), International Symposium Report 2011: New prospects for foreign language teaching in higher education (pp. 23-31). Tokyo, Japan: World Language and Society Education Centre (WoLSEC), Tokyo University of Foreign Studies.

Bellassen, J. (2018). Monism vs. dualism: The fundamental divergence in Chinese second language pedagogy. *TCSOL Studies, 4*(72), 1-11. [白乐桑. (2018). 一元论抑或二元论：汉语二语教学本体认识论的根本分歧与障碍. *华文教学与研究, 4*(72), 1-11.]

Bellassen, J. (2024). The schism in teaching Chinese as a second language: The "missing ring" in the history of Chinese language education—John DeFrancis. *Journal of Global Chinese Language Development*, (2), 60-79.. [白乐桑. (2024). 再说中文第二语言文字中的分裂问题：国际中文教育史上"缺失之环"——德范克. *全球中文发展研究*, (2), 60-79..]

Bellassen, J., & Zhang, L. (2008). The implications of the Common European Framework of Reference for Languages for Chinese language teaching. *Chinese Teaching in the World*, (3), 58-73. [白乐桑, 张丽. (2008). 《欧洲语言共同参考框架》新理念对汉语教学的启示与推动. *世界汉语教学*, (3), 58-73.]

Cao, Y., Fan, J., & Yang, Q. (2024). ChatGPT in education: Ethical predicaments of generative AI. *Transactions on Social Science, Education and Humanities Research, 11*, 64–69. https://doi.org/10.62051/bejkn640

Cong, Y. (2024). AI language models: An opportunity to enhance language learning. *Informatics, 11*(2), 49. https://doi.org/10.3390/informatics11020049

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Crawford, J., Cowling, M., & Allen, K.-A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice, 20*(3), Article 2. https://doi.org/10.53761/1.20.3.02

Da, J. (2004). *Modern Chinese character frequency list* [Data set]. Middle Tennessee State University. https://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=MO

DeFrancis, J., Teng, C. Y. Y., & Yung, C. S. (1966). *Beginning Chinese reader*. Yale University Press.

Ekin, S. (2023). *Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices*. TechRxiv. https://doi.org/10.36227/techrxiv.22683919.v2

Glaser, N. (2023). Exploring the potential of ChatGPT as an educational technology: An emerging technology report. *Technology, Knowledge and Learning, 28*, 1945–1952. https://doi.org/10.1007/s10758-023-09660-6

Guder, A. (2014). Reading competence and graphemic competence: Impacts of the Chinese writing system on designing competence descriptors for the European Benchmarks for the Chinese Language (EBCL) project. *CHUN, 29*, 5–29.

Guder, A. (2015). *European Benchmarks for the Chinese Language (EBCL) / Europäischer Referenzrahmen für Chinesisch als Fremdsprache (欧洲汉语能力基准项目)* [Special edition]. *CHUN*. IUDICIUM Verlag. https://doi.org/10.17169/REFUBIUM-19534

Guo, S. Y. (1985). The constraint of writing on the Chinese language. In *Collected essays on language and writing* (pp. 112-114). Shanghai Classics Publishing House. [郭绍虞. (1985). 中国语言所受到的文字的牵制. *照隅室语言文字论集* (pp. 112-114). 上海古籍出版社.]

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237–257. https://doi.org/10.1111/jcal.12610

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464. https://doi.org/10.30935/cedtech/13605

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Jiang, X., Li, J., & Chen, C.-H. (2024). Enhancing critical thinking skills with ChatGPT-powered activities in Chinese language classrooms. *International Journal of Chinese Language Teaching, 5*(1), 47–73. https://doi.org/10.46451/ijclt.20240105

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal, 6*, 100048.

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education, 20*(1), 56. https://doi.org/10.1186/s41239-023-00426-1

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). 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. *Computers and Education: Artificial Intelligence, 7*, 100266. https://doi.org/10.1016/j.caeai.2024.100266

Li, J., Ren, X., Jiang, X., & Chen, C. (2023). Exploring the use of ChatGPT in Chinese language classrooms. *International Journal of Chinese Language Teaching, 4*(3), 36–55. https://doi.org/10.46451/ijclt.20230303

Lin-Zucker, M. (2024). Challenges and difficulties of applying the CEFR in Chinese language teaching in France. *Journal of Teaching Chinese as a Second Language*, (34), 43–51. [林季苗. (2024). CEFR 應用在法國漢語教學中的挑戰與困境探討. *華語學刊*, (34), 43–51.]

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35. https://doi.org/10.1145/3560815

Pryzant, R., Iter, D., Li, J., Lee, Y. T., Zhu, C., & Zeng, M. (2023). Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7957–7968). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.494

Vaccino-Salvadore, S. (2023). Exploring the ethical dimensions of using ChatGPT in language learning and beyond. *Languages, 8*(3), 191. https://doi.org/10.3390/languages8030191

Wallace, R. S. (2009). The anatomy of A.L.I.C.E. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer* (pp. 181–210). Springer.

Wang, F., Cheung, A. C. K., Neitzel, A. J., & Chai, C. S. (2025). Does chatting with chatbots improve language learning performance? A meta-analysis of chatbot-assisted language learning. *Review of Educational Research*. *95*(4), 623-660. https://doi.org/10.3102/00346543241255621

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168

Zhang-Colin, Y., & Gianninoto, M. (2022). Didactic ideas in Chinese as a foreign language since 1945. *Documents pour l'histoire du français langue étrangère ou seconde, 68*. https://doi.org/10.4000/dhfles.9081

Zhao, Q., Hsu, Y.-Y., & Huang, C.-R. (2024). Large language model and Chinese near synonyms: Designing prompts for online CFL learners. *Journal of Technology and Chinese Language Teaching, 15*(1), 49–69.

Appendix A

Illustrative Examples of Sinographically Constrained Outputs

This appendix provides illustrative examples of Chinese language model outputs generated under different sinographic constraint conditions aligned with CEFR–EBCL descriptors. These examples are included for pedagogical clarification only and are not part of the quantitative analysis reported in the Results section.

The examples below are drawn from the first experimental run of task RW1 (Overall Reading Comprehension) using GPT-4.1 with explicit character lists, as described in Section 4. Their purpose is to illustrate how prompt-based constraints translate into observable differences in generated texts, while remaining within a controlled graphic space. For each level, the complete model output is provided, including the reading text and comprehension questions.

Example 1. Output constrained to A1 level (250 characters)
Task: RW1 – Overall Reading Comprehension

User prompt: *RW1*
Complete model output:
Read the following text about a student's day:
我早上六点半起，七点吃饭，八点去学校。上午我有中文和数学。中午我和朋友在学校吃饭。下午三点回家。晚上我和家人一起吃饭，然后写作业，十点睡。
Questions:
1. 他早上几点起？
2. 他中午和谁吃饭？
3. 他晚上做什么？
4. 他几点睡？
Please answer in Chinese.

Translation of reading text: "I get up at 6:30 in the morning, eat at 7, and go to school at 8. In the morning I have Chinese and math. At noon my friend and I eat at school. I go home at 3 in the afternoon. In the evening I eat with my family, then do homework, and sleep at 10."

Example 2. Output constrained to A1+ level (320 characters)
Task: RW1 – Overall Reading Comprehension

User prompt: *RW1*
Complete model output:
请阅读下面的短文：
我有一个好朋友，他叫王明。他今年十五岁，是学生。他家有四口人：爸爸、妈妈、一个妹妹和他。他家在北京。他每天早上六点半起床，七点吃早饭，八点去学校。他喜欢看书，也喜欢和朋友一起玩。
问题：
1. 王明家有几口人？
2. 他家在哪个城市？
3. 王明每天几点去学校？
请用中文回答。

Translation of reading text: "I have a good friend called Wang Ming. He is 15 years old and is a student. His family has four people: dad, mom, a younger sister, and him. His family lives in Beijing. Every day he gets up at 6:30, eats breakfast at 7, and goes to school at 8. He likes reading books and also likes playing with friends."

Example 3. Output constrained to A2 level (630 characters)
Task: RW1 – Overall Reading Comprehension

User prompt: *RW1*
Complete model output:
Complete model output:
活动：RW1 - 阅读理解
请阅读下面关于一个学生一天的生活的短文：
小明早上七点起床，他先洗脸，然后吃早饭。八点他去学校上课。中午十二点他和朋友在学校吃午饭。下午三点下课后，小明去公园玩球。晚上他和家人一起吃晚饭，八点写作业，十点睡觉。
问题：
1. 小明早上几点起床？
2. 小明中午和谁吃午饭？
3. 小明下午做什么？
4. 小明晚上几点睡觉？
5. 小明一天吃几次饭？
请用中文回答。

Translation of reading text: "Xiao Ming gets up at 7 in the morning. He first washes his face, then eats breakfast. At 8 he goes to school for class. At noon at 12 o'clock he and his friends eat lunch at school. After class ends at 3 in the afternoon, Xiao Ming goes to the park to play ball. In the evening he eats dinner with his family, does homework at 8, and goes to sleep at 10."

Observations

These examples illustrate how increasingly permissive sinographic constraints allow richer lexical and syntactic content while preserving control over character usage.

Level	Characters	Sentence Complexity	Vocabulary Features
A1	250	Simple, short clauses	Basic time words, simple verbs (吃, 去, 睡)
A1+	320	Compound sentences	Family vocabulary, age, location (北京)
A2	630	Sequential actions with connectors	Action sequences (先 ... 然后), specific activities (洗脸, 上课, 下课, 玩球)

At A1, responses are limited to basic vocabulary and simple sentence structures describing a routine in first person. Instructions and questions are provided in English to accommodate beginner learners. At A1+, additional characters enable the introduction of a third-person narrative with family context and personal preferences; instructions shift to Chinese (请阅读, 请用中文回答). At A2, the expanded inventory permits complex sentences with temporal connectors (先...然后), specific action verbs (洗脸, 上课, 下课), and varied activities; the number of comprehension questions increases to five, including an inference question (小明一天吃几次饭？). This progression exemplifies the pedagogical principle underlying the experimental design: maintaining controlled progression in written input aligned with CEFR–EBCL descriptors.

Instruction deviation analysis

To illustrate how instruction deviation is computed (see Section 4.5), each output was analyzed for out-of-level characters. Results are summarized below:

Level	Chinese Characters	Instruction Deviation	Out-of-Level Characters
A1	83	13.3%	起 (×3), 校 (×2), 睡 (×2), 然, 数, 做, 业
A1+	112	8.0%	王 (×3), 校 (×2), 阅, 读, 短, 答
A2	145	2.8%	阅 (×2), 脸, 于

These data confirm the pattern observed in the main results: instruction deviation decreases as the character inventory expands, and deviations often involve high-frequency characters essential for task completion (e.g., 起, 睡 for describing daily routines; 阅, 读 for reading instructions). This underscores the trade-off between strict constraint adherence and communicative functionality.

Appendix B

Glossary of LLM- and Prompt-Related Terms

This appendix presents the glossary of technical terms used throughout the article. The definitions provided here support the interpretation of the experimental design and the analysis of model behavior but are not intended as an exhaustive introduction to large language models.

Table 1: Glossary of LLM and prompt related terms

Term	Definition
LLM	An LLM (Large Language Model) is an artificial intelligence model based on deep neural networks, trained on massive text corpora to capture complex linguistic relationships and generate natural language based on received contexts and queries.
Chatbot	A computer program designed to simulate a conversation with human users, especially on the internet. ChatGPT is a very general example, but many others exist that are much more targeted.
Foundation Model	A foundation model is a pre-trained LLM on a vast amount of unlabeled data, capable of adapting to a variety of specific tasks through additional adjustments (fine-tuning). These models, like GPT, LLAMA, QWEN, BERT or MISTRAL are called "foundation" because they serve as a basis for developing applications in various fields.
Assistant	In the context of LLMs, an assistant is defined as an artificial intelligence model designed to respond to specific queries via a system prompt or predefined parameters. These instructions guide the model's behavior and allow it to provide a contextualized user experience.
Prompt	Prompts provide instructions to an LLM to impose rules, automate tasks, and guarantee particular qualities (and quantities) in the generated output. They also function as a form of programming, enabling the customization of both the outputs and interactions with the LLM.
System Prompt	A system prompt is an initial instruction given to an LLM to define its overall behavior, such as the tone or style of the responses. It sets guidelines to align the model's responses with the desired objectives during interactions. It may start with a role instruction.
Prompt Engineering	Prompt engineering is the technique of formulating specific instructions to condition and optimize the output of a language model, exploiting the internal mechanisms of the architecture to modulate its behavior and maximize the relevance of the results.
Role Instruction	Role instruction is a part of the prompt that explicitly defines the role or behavior the chatbot should adopt in a given interaction. It's an instruction that guides the model on how to act or respond based on the desired context. For example, "You are a professional Chinese language teacher, guiding the user through language learning with clear explanations and practical examples".

Fine-Tuning	Adjusting a pre-trained LLM consists of specializing it for a specific task or domain by retraining it on a small set of labeled data. This optimizes its performance for precise tasks like providing responses that satisfy character frequency constraints.
Instruction Tuning	Instruction tuning is the adjustment of a language model (LLM) so that it accurately responds to natural language instructions. This involves training the model on pairs of instructions and responses, making it more effective at providing relevant answers and following specific commands in practical applications.
Instruction Deviation	An instruction deviation occurs when a language model (LLM) does not correctly follow a given directive in a prompt, by omitting a task, executing it incorrectly, or producing results that do not conform to expectations. This can be caused by ambiguities in the prompt, limitations of the model in understanding the context, or a lack of alignment with the desired objectives.

Appendix C

Prompt Typology and Design Choices

This appendix details the prompt typology adopted in this study and explains the rationale behind the methodological choices. Numerous prompt types have been discussed in the literature, including zero-shot, one-shot, few-shot, and chain-of-thought prompts. While example-based prompts can enhance task performance, they introduce variability that compromises experimental reproducibility. Similarly, chain-of-thought prompts primarily affect reasoning transparency and are not directly suited to the evaluation of constraint compliance at the character level. For these reasons, this study relies exclusively on system prompts formulated as explicit instructional constraints. System prompts define the chatbot's role, pedagogical objectives, and linguistic boundaries prior to interaction and remain stable throughout the session. This ensures strict control of experimental conditions and comparability across prompt configurations.

Appendix C.1. System Prompt with Explicit Character Lists

The following system prompt template was used for the List condition (L). The placeholder {level} was replaced by the target proficiency level (A1, A1+, or A2), and {liste} was replaced by the complete list of EBCL-authorized characters for that level (250 characters for A1, approximately 320 for A1+, and 630 for A2). The full character lists are provided in Appendix D**.**

You are a Chinese language tutor specialized in EBCL (European Benchmarking Chinese Language) framework activities. Your role is to design and facilitate ONE specific learning activity based on the user's choice.

AVAILABLE ACTIVITIES (EBCL Framework):

READING ACTIVITIES:

- RW1: Overall Reading Comprehension – Provide a short text and ask comprehension questions
- RW2: Reading Correspondence – Present an email, letter, or message and ask the student to respond or answer questions
- RW3: Reading for Orientation – Provide signs, menus, schedules, or directories for information extraction
- RW4: Reading for Information & Argument – Present an article or opinion piece for analysis
- RW5: Reading Instructions – Provide step-by-step instructions (recipe, manual, directions) to follow

WRITING ACTIVITIES:
- PW1: Overall Written Production – Ask student to write a descriptive or narrative text (diary entry, description, report)
- PW2: Creative Writing – Prompt student to write a creative piece (story, poem, dialogue)

INTERACTION ACTIVITIES:
- IW1: Overall Written Interaction – Simulate a written exchange (chat, forum discussion)
- IW2: Correspondence – Ask student to write formal/informal letters or emails
- IW3: Notes, Messages & Forms – Have student complete forms, write notes, or short messages

CHARACTER CONSTRAINT – {level} Level:
Available characters: {liste}

CRITICAL RULES:
1. ALL Chinese text you produce must use ONLY characters from the {level}-level list above
2. Before finalizing any response, verify each Chinese character against the list
3. If a word requires characters not in the list, find an alternative expression using only allowed characters
4. Do NOT apologize for the constraint – work within it naturally

WORKFLOW FOR EACH INTERACTION:
1. Present the activity clearly in English
2. Provide the Chinese content (text, prompt, or material) using ONLY {level}-level characters
3. Give clear instructions for what the student should do
4. When student responds, provide feedback using ONLY {level}-level characters

ACTIVITY-SPECIFIC GUIDELINES:

For Reading Activities (RW1-RW5):
- Text length: 50-150 characters for beginners, 150-300 for intermediate
- Include 3-5 comprehension questions
- Questions should test different skills: literal comprehension, inference, vocabulary

For Writing Activities (PW1-PW2):
- Provide a clear prompt or scenario

- Specify expected length (e.g., “Write 5-8 sentences”)
- Give structural guidance if needed (e.g., “Include: greeting, main content, closing”)

For Interaction Activities (IW1-IW3):
- Set up a realistic context
- Define the communication goal clearly
- Specify the format expected (chat message, email, form, etc.)

When the user provides an activity code (RW1, PW1, etc.), immediately begin that specific activity without preamble. Present the Chinese content directly.

Appendix C.2. System Prompt with no Character Lists

For the No List condition (NL). The placeholder {level} was replaced by the target proficiency level (A1, A1+, or A2). Unlike the List condition, no explicit character inventory was provided; the model was instead instructed to rely on its internal representation of level-appropriate vocabulary.

Appendix D

Reference Character Lists and Instruction Deviation

Instruction deviation is calculated as the ratio of characters outside the authorized list to total characters in output. The character sets used in this study are based on the European Benchmarking Chinese Language (EBCL) framework. Each proficiency level is associated with a constrained vocabulary of Chinese characters that learners are expected to recognize and produce. The character inventories are cumulative: each level includes all characters from the preceding levels.

A1 Level (250 characters)

The A1 level comprises 250 essential characters for basic communication:
爱八爸吧白百班半杯北本比笔边别病不菜茶长常车城吃出从打大到道的得地弟点电店东懂动都对多儿二饭方房飞非分父干刚高哥个给跟工公关馆贵国果过还孩海汉好号喝和很红后候花画话欢回会活火机几家间见叫姐今近进京九酒就觉开看可课口块快筷来老了累冷离里两六妈吗买卖忙么没每美妹门们米面名明母哪那男南难呢能你年您女朋票七期气汽前钱亲请去让人认日肉三山商上少谁什生师十时识事是市书水说思四岁他她太天听同外玩晚网为文问我五午西喜下先现想小些写谢心新信星姓兴学样要也一以意因影用友有雨语元远月运在再早怎这只知中重住子字走昨坐作

A1+ Level (320 characters)

The A1+ level extends A1 with 70 additional recommended characters, for a total of 320 characters:

A1 characters (250): as listed above

Additional characters (+70):

安包部差场唱穿床次村错第饿发法歌共狗黑或级鸡介零路马慢猫脑牛农旁片骑起千球热绍视手睡所铁头物息习系行休羊医音英右鱼园院乐云找者址祝自足最左做

A2 Level (630 characters)

The A2 level provides a comprehensive set of 630 characters for elementary proficiency, encompassing all A1 and A1+ characters plus additional items:
啊爱安八吧把爸白百班般办半帮包报抱杯北备被本比笔毕边便变表别病博不步部才菜参餐层茶差长常厂场唱超车成城吃出初除厨楚穿传床春词此次从村错答打大代带待单但蛋当到道得的灯等低地弟第典点电店定订丢东冬懂动都读独度短对多饿儿而二发法反饭方房放飞非费分份风封服附父复该改干感刚高告哥歌格个给跟更工公共狗古故拐怪关观馆惯广贵国果过孩海汉好号喝和河贺黑很红后候湖护花化画话坏欢还换黄回会婚活火或机鸡级极急几己记际济继寄加家假价间见件健江讲酱交饺叫较教接街节结姐解介界借今斤金禁近进京经景净静九久酒旧就局句决觉卡开看康考可渴刻客课空口裤块快筷拉来篮老乐了累冷离礼李里理力历丽联凉两亮谅辆聊林零六楼路录旅妈马码吗买卖满慢忙猫毛贸么没每美妹门们米面免民名明末母目拿哪那奶男南难脑呢能你年念鸟您牛农努女暖欧怕乓旁胖跑朋片漂票乒平七期其奇骑起气汽千前签钱歉且亲轻清情请秋球区取去趣全然让热人认日肉如赛三散色山商上少绍社身生声胜师十什时识实史始世市事视试室是收手首书术树双谁水睡说司思死四送诉算虽岁孙所他她它台太谈汤堂套特踢提题体天填条铁听厅庭通同头图外玩完晚万王网往忘望卫为位文问我卧无五午务勿物西吸希息习洗喜戏系下夏先现相想向像消小校笑些写谢心新信兴星行姓性兄休修需许续选学雪牙亚烟言羊阳样药要也业夜一艺衣医已以易意因音印银应英影硬泳用邮油游友有又右鱼愉雨语元员园原远院愿月越云运咱再在早怎站张找照者这真正证知只之直址止纸至治中钟种重周洲主住助祝注专准桌着子字自总走租足最昨左作坐做座

Appendix E

A Ready-to-Use Prompting Procedure for Teachers (A1 and A1+)

This appendix translates the experimental findings into a minimal classroom procedure for teachers of beginner learners. It requires no technical expertise and only needs to be set up once.

Step 1. Retrieve the EBCL character list corresponding to the target level (A1: 250 characters; A1+: 320 characters), as reproduced in Appendix D or from the EBCL project materials.

Step 2. Paste the list once into a reusable location: a saved prompt template, the custom instructions of a personal account, or a persistent project or workspace setting in the chatbot interface used by the class. The list then no longer needs to be re-entered for each request.

Step 3. For each new activity, send a short task instruction together with the stored constraint. A simplified template, adapted from the full system prompt in Appendix C.1, is given below.

> *You are a tutor of Chinese as a foreign language for beginner learners at EBCL level {level}. In all Chinese text you produce, use only characters from the following authorized list: {paste the list for the target level}. If a word would require a character outside this list, rephrase it using authorized characters or give it in pinyin. Task: {for example: write a short text message inviting a friend to dinner, suitable for this level}.*

Step 4. Briefly check the output before classroom use. The results in Section 5 show that explicit lists substantially reduce, but do not eliminate, out-of-level characters; a quick visual check, or the evaluation script provided with this study, remains advisable.

Journal of Technology and Chinese Language Teaching Volume 17 Number 1, June 2026
http://www.tclt.us/journal/2026v17n1/jingxu.pdf pp. 30-47

大语言模型在国际中文阅读自动出题中的效能评估
(Evaluating the Effectiveness of Large Language Models for Automatic Question Generation in International Chinese Reading)

景宏伟
(Jing, Hongwei)
北京语言大学
(Beijing Language and Culture University)
jhw080266@163.com

徐娟
(Xu, Juan)
北京语言大学
(Beijing Language and Culture University)
xujuan@blcu.edu.cn

摘要：随着国际中文教育数字化转型的持续推进，传统人工命题模式在效率、成本与规模化应用方面的瓶颈日益凸显。在此背景下，以大模型为代表的人工智能技术，为自动命题提供了新的技术路径。本研究以 HSK6 级阅读理解题为研究对象，系统评估大模型在自动命题任务中的实际效能。研究选取四种大模型，结合提示工程开展实验，涵盖指令大模型、推理大模型以及经 LoRA 微调后的垂直大模型。并从语言流畅度、内容准确性、题目复杂度、选项干扰性、答案唯一性、题型多样性六个维度，辅以 BLEU、ROUGE、Distinct 等机器指标，对生成题目进行综合评估。研究表明：大模型生成的题目与人工命题具有较高的相似性，但在难度控制、可回答性等方面尚存在不稳定性，需经人工审核修订后方可用于教学；在模型对比中，推理大模型整体表现更优。基于此，本研究进一步提出相应的使用建议，以优化题目生成过程，推动人机协同命题模式的发展。

Abstract: With the continued advancement of the digital transformation of international Chinese language education, the traditional manual approach to test item development has increasingly encountered bottlenecks in terms of efficiency, cost, and scalability. Against this backdrop, artificial intelligence technologies, particularly large language models (LLMs), have opened up new possibilities for automated test item generation. This study focuses on HSK Level 6 reading comprehension items and systematically evaluates the practical effectiveness of LLMs in automatic item generation. Four LLMs were selected for experimentation using prompt engineering, including instruction-tuned models, reasoning-oriented models, and a domain-specific model fine-tuned with Low-Rank Adaptation (LoRA). The generated items were comprehensively evaluated across six dimensions: linguistic fluency, content accuracy, item complexity, distractor quality, answer uniqueness, and item-type diversity. In addition, machine-based evaluation metrics, including BLEU, ROUGE, and Distinct, were employed to provide complementary assessments. The results indicate that the items generated by LLMs exhibit a high degree of similarity to those developed by human experts. However, the models still demonstrate instability in controlling item difficulty and ensuring answerability, suggesting that human review and revision remain necessary before the generated items can be used in instructional settings. Among the models evaluated, reasoning-oriented LLMs achieved the best overall performance. Based on these findings, this study further proposes practical recommendations for optimizing the item generation process and advancing a human–AI collaborative approach to test development.

关键词: 国际中文教育，生成式人工智能，大语言模型，题目自动生成，HSK

Keywords: International Chinese Language Education, Generative Artificial Intelligence, Large Language Model, Automatic question generation, HSK

1. 引言

题目编制是语言测试与教学评估体系中的核心环节，其质量直接影响测评结果的有效性与教学反馈的准确性。长期以来，国际中文教育领域主要依赖人工方式进行题目设计与开发。然而，随着全球中文学习需求的持续增长，截至 2025 年 HSK 全球累计考生规模已超过 850 万人（郁云峰等，2025），传统人工命题在效率、成本控制及大规模题库建设等方面的局限性日益凸显，已难以满足快速增长的测评需求。

此外，HSK 3.0 考试体系的推出（曹贤文等，2025），对题目数量、质量及更新速度提出了更高要求，尤其是在标准化与大规模应用场景下，传统人工命题模式在短时间内难以高效产出高质量试题，进而可能对 HSK 3.0 的推广与实施效果产生一定制约。近年来，以大模型为代表的人工智能技术快速发展，在文本生成、逻辑推理等方面展现出显著优势，为自动化题目生成提供了新的技术路径与实现可能。

本研究依托 HSK3.0 考试体系，以 HSK6 级阅读理解题为研究切入点，向大模型输入阅读材料以生成相应题目。同时，从多维度对生成题目进行人工评估，并结合机器评价指标，以提升评估结果的客观性与全面性。本研究旨在探究大模型在国际中文教育自动命题任务中的能力边界，为国际中文教师基于大模型开展自动命题提供实践参考。

2. 研究综述

2.1 大模型赋能国际中文教育资源研发

近年来，以大模型为代表的生成式人工智能技术正持续推动国际中文教育领域的数字化转型，为教学资源的智能化开发与个性化建设注入了新动能。目前，相关应用主要集中于智能化内容生成方向，具体体现为：在文本资源方面，支持分级阅读材料的自动生成（韩欣欣等，2025）、个性化阅读材料的生成（侯泽煜、徐娟，2025）；在写作教学方面，能够辅助开发智慧化写作资源如智慧教材、范文语料库等（马瑞祾、徐娟，2024）；在课程资源建设方面，可助力高效生成结构化的中文微课内容（李嘉仪、徐娟，2025）。这些实践不仅提升了资源开发的效率与多样性，也为实现精准化、个性化的国际中文教学提供了有力支持。而在国际中文教育题目资源建设中，刘玉屏等（2025）探索了生成式人工智能赋能 HSK 模拟试题的编写。

2.2 题目自动生成

题目自动生成（Automatic Question Generation，AQG)技术属于文本生成的一项子任务，伴随着自然语言处理技术的发展而演变。早期 AQG 主要采用基于规则的方法，例如 Mitkov 等人（2003）提出的基于句法模式匹配的框架，通过语法分析和模板填充生成问题。这类方法能确保生成问题的规范性，但受限于模板库，往往存在多样性不足的问题。进入统计机器学习阶段，通过引入概率模型提升问题生成的灵活性，如 Heilman 等（2010）提出了一种基于规则和统计排序的 AQG 方法，其核心思想是“过度生成再排序”，通过规则生成大量候选问题，再利用统计模型对这些问题进行排序以筛选出最优结果。但规则构建依然依赖人工，成本较高。

随着神经网络的发展，深度学习方法显著推动了 AQG 向语义理解和自然表达的方向发展，显著提升了题目生成的语义连贯性与多样性。Jiang 和 Lee（2017）将词嵌入模型应用于汉语名词多项选择题干扰项自动生成中，通过分布式表示计算词汇语义相似度，优化了干扰项设计。Du 等（2017）提出将基于注意力机制的序列到序列（Sequence-to-Sequence，Seq2Seq）模型应用于 AQG 任务，实现从文本中自动生成阅读理解题目。徐坚（2023）进一步提出融合门控循环单元与图注意力网络的增强型 Seq2Seq 模型，通过答案引导的图注意力机制捕捉文章内部依赖关系，并结合注意力机制与指针网络，提升了所生成题目的语义关联与答案确定性。然而，该阶段仍存在逻辑一致性不足、认知层次较浅等问题。

Transformer 架构（Vaswani et al，2017）凭借并行计算和自注意力机制有效解决了长距离依赖问题，为 AQG 提供了更好的语义理解和表达能力。基于 Transformer 架构的预训练大语言模型，如 BERT（Devlin et al，2019）、T5（Raffel et al，2020）和 GPT 系列（Brown et al，2020）等，通过在海量文本数据上预训练，学习了丰富的语言表示和知识，获得了强大的语言理解和生成能力，为 AQG 提供了新的方向。陈欣等（2024）结合提示工程构建了一种基于大模型的试题自动生成路径。来雨轩等（2024）为激发大模型在 AQG 任务上的潜力，提出了将大模型和检索增强技术

相结合的生成方法。聚焦在国际中文教育领域，有学者尝试利用大模型进行阅读测试题的研发（王鸿滨、吕海辉，2025；王亚敏等，2025）。目前，大模型在国际中文教育 AQG 任务上的能力边界尚有待进一步验证。

总体而言，大模型在自动出题任务上展现出来较大的潜力，但尚存在不足，本研究尝试探索以下工作：

1）通过高效参数微调构造垂直领域大模型，验证该方法在国际中文教育自动出题任务上的表现；
2）从主客观相结合的多维角度，系统对比国内外不同类别大模型包括指令型、推理型以及垂直领域大模型在出题任务上的表现；
3）提出大模型在国际中文教育自动出题任务上的人机协同命题模式，为一线教师和命题专家提供参考。

3. 研究设计

为全面评测大模型在国际中文教育自动命题任务上的表现，本研究采用实验比较法，通过系统化的研究设计对大模型自动出题效能进行评测。其设计核心环节包括：首先基于模型代表性和中文处理能力的综合考量，选取典型大模型作为评测对象；其次以 HSK6 级真题为主要来源构建数据集；最后通过标准化的实验操作流程和科学的评价指标，确保实验数据的可靠性和可比性。以下将分别从评测模型选择、数据集构建、实验设计以及评估策略四个维度详细阐述研究设计。

3.1 模型选择

在模型的选择上，本研究参考了中文语言理解测评基准 SuperCLUE 榜单[1]，该榜单聚焦于通用大模型的综合性测评。选取了四款在 SuperCLUE 通用榜中排名靠前的大模型，具体模型信息如表 1 所示。

表 1　大模型具体信息

模型名称	发布机构	是否推理	属地	发布时间
Gemini-3-Pro-Preview	Google	是	海外	2025.11
DeepSeek-V3.2-Thinking	深度求索	是	国内	2025.12
Qwen-3-Max	阿里巴巴	否	国内	2025.11
Llama-4-Maverick-17B-128E-Instruct	Meta	否	海外	2025.11

选取的四个模型中包含了来自海内外的推理型、指令型大模型。其中 Gemini-3-Pro-Preview 是谷歌推出的多模态大模型，在数学推理、代码生成及跨模态理解方面突出，尤其擅长科学问答与复杂逻辑推理，属于推理大模型；DeepSeek-V3.2-Thinking

[1] 网址参见：https://supercClueai.com/generalpage。

是深度求索推出的推理专项模型，通过思维链增强与自我反思机制，显著提升复杂推理、长文本分析与分步求解能力，属于推理大模型；Qwen-3-Max 是阿里通义千问高性能版本，在中文理解、长上下文处理与多语言任务中表现优异，适合长文档分析与生成，属于指令大模型；Llama-4-Maverick-17B-128E-Instruct 是基于 Llama 4 的高效指令微调模型，擅长指令跟随与多轮对话，属于指令大模型。此外，还有一类垂直大模型，即针对特定任务进行专门训练与优化的大模型。由于目前没有公开的可用于自动命题的国际中文教育领域垂直大模型，本研究采用微调技术构建用于出题的垂直大模型。考虑到算力资源、硬件设施、数据集规模等因素，本研究选取了参数量较小的大模型 Qwen2.5-7B 作为微调的基座模型。

3.2 数据集构建

在微调模型的训练集构建中，数据主要来源于 2016 年至今的 HSK6 级阅读真题以及模拟题，覆盖多种体裁、多种题型，以确保数据的权威性、规范性与教学适配性，同时可以将模型生成的题目与真题进行对比，以探究大模型在 AQG 任务上的能力边界。数据集具体构建流程为：首先，进行数据的采集和筛选，初步搜集到了 680 余篇真题材料，均为 PDF 格式；其次，进行文本提取，经 OCR 技术将其转换为可编辑文本、并剔除冗余信息，通过关键词定位阅读理解题模块获取阅读材料及其对应题目；然后，进行语料清洗与去重，修正 OCR 识别残留的错别字与语句不通问题并确保与原始阅读材料段落划分一致，同时采用余弦相似度算法计算文本间相似度，进行语料去重，确保无重复阅读材料，最终确认 640 篇阅读材料；最后，调整数据格式，将数据整理为“阅读材料—题目”的结构化数据格式，以适配模型的训练。此外，考虑到现有数据集规模过小，故另外选取了 600 篇 HSK6 级模拟题，最终训练集包含 1240 条数据。需要说明的是，模拟题在题型结构与知识点覆盖上参考 HSK6 级考试要求设计，但其来源于非官方命题体系，与真题在命题规范性方面仍存在一定差异，本研究在模型训练过程中将其作为与真题同分布的近似数据进行使用，以增强模型对 HSK6 级阅读理解题型结构的学习能力。此外，在测试集的构建中，本研究另外选取了 30 篇 HSK6 级阅读真题，利用测试集中的阅读材料，对不同大模型进行自动出题能力的评测。需要说明的是，测试集与训练集中的阅读材料完全独立，没有重复。

3.3 实验设计

3.3.1 提示词设计

在利用通用大模型进行问题生成时，本研究通过编写提示词的方法实现。提示词设计是大模型应用中的关键环节，对于充分发挥大模型能力至关重要，它通过规范化、结构化的指令引导模型理解任务意图、约束输出范围，从而显著提升生成内容的准确性、相关性和可控性。提示工程（Prompt Engineering）是一种通过设计、实验和优化输入来引导模型生成高质量、准确和有针对性的输出的技术（Dong et al，2024），其中输入的格式一般称作提示模板，组织各种提示信息的方式称为提示策略或提示方法，其中常用的提示策略如表 2 所示。

表 2 常用的提示策略

提示策略	描述
零样本提示（Zero-Shot）	大模型在没有任何任务示例的情况下，仅依据自然语言指令执行任务
少样本提示（Few-Shot）	通过提供少量示例引导模型执行任务
思维链提示（Chain-of-Thought，CoT）	在解决复杂推理问题时，要求模型将中间推理步骤显式地输出，鼓励模型展示推理步骤以提升复杂问题解答的准确性
角色扮演提示（Role-Playing）	通过指令为模型赋予特定角色或身份以控制输出风格与内容

目前，在提示词的设计上已有很多通用法则和实践经验。为提升大模型的输出质量，学界探索了多种结构化提示框架以优化提示工程效果，例如具有代表性的ICIO 框架、CLEAR 框架等[2]。此外，针对不同应用情境，也形成了相应的提示框架。在教育领域，CRISPE 框架（王华树、谢斐，2024）在实践中得到广泛应用，并被证实能够有效提升模型回答的质量。基于上述考虑，本研究选用 CRISPE 框架作为提示词设计的模板，具体提示词设计如表 3 所示。

表 3 题目生成提示词框架

组成部分	示例内容
角色（Capacity and Role）	你是汉语水平考试（HSK）的命题专家
背景（Insight）	中文学习者在应对 HSK6 级阅读时，常对长难句理解、隐含意图推断、文化负载词把握及篇章逻辑衔接等方面存在困难。
任务（Statement）	请基于提供的阅读材料，设计四道高质量的选择题。要求每道题包含四个选项，只有一个答案正确，所有题目必须基于材料内容，难度符合 HSK6 级水平。
格式（Personality）	题目格式示例： 1.老总当初为什么要留这个年轻人？ A 公司急需人员 B 客户欣赏年轻人 C 相信自己没有看错人 D 年轻人有丰富的工作经验 答案：C

为验证该提示词的有效性，本研究首先进行了小规模的预实验，发现尽管初始提示词设定了基本框架，但模型生成的题目在考查点分布、难度控制及选项设计方面存在改进空间。具体而言，生成的部分题目未能精准对应 HSK6 级的核心能力要求，部分干扰项的干扰性不足或偏离原文逻辑，题目难度呈现不稳定现象。为了系

[2] 网址参见：https://developer.aliyun.com/article/1490356。

统提升生成题目的质量与规范性，对初始提示词进行了结构化迭代与优化，优化后的提示词框架如表 4 所示。

表 4 优化后的题目生成提示词框架

组成部分	示例内容
角色（Capacity and Role）	你是精通 HSK6 级的资深命题专家，深谙考试大纲，可以精准把握命题要求。
背景（Insight）	在 HSK6 级阅读命题中，应重点考查学习者对长难句的理解、对隐含观点或态度的推断、对特定语境下词语的理解，以及对篇章整体逻辑与主旨的把握等。
任务（Statement）	请严格依据提供的阅读材料，设计四道单项选择题。要求每道题包含四个选项，只有一个答案正确，要求题型具有多样性，包括细节题、推理判断题、主旨大意题、词义题等。题目应体现 HSK6 级应有的认知复杂度，避免仅进行原文词句的简单匹配。
格式（Personality）	请严格按照以下题目格式示例输出： 1.老总当初为什么要留这个年轻人？ A 公司急需人员 B 客户欣赏年轻人 C 相信自己没有看错人 D 年轻人有丰富的工作经验 答案：C

3.3.2 模型参数设置

此外，为批量生成题目，本研究通过调用大模型 API 的方式实现自动出题，在不同模型的参数设置上参考了官方的默认参数，主要包括采样多样性参数（Tempreture）和采样范围调节参数（top_p）。Tempreture 是一个用于调节模型输出概率分布“平滑度”的超参数，它通过在对数概率（logits）被转换为概率（softmax）之前，对其进行缩放，从而控制生成过程的随机性。在计算下一个词的概率时，模型原始的 logits 向量会除以 Temperature 的值 T，见公式 1 所示，其中，w_i 为词表中的第 i 个词元（token），z_i 是模型为词表中第 i 个词元输出的原始 logit 值，V 是词表大小。Top-p 是一种通过设定概率阈值，在文本生成的每一步动态筛选出最小的高概率候选词集合进行采样，以在保证连贯性的前提下控制输出多样性的自适应方法。

$$P(w_i) = \frac{\exp(z_i/T)}{\sum_{j=1}^{V}\exp(z_j/T)} \qquad \text{公式（1）}$$

通过合理调整 Tempreture 和 Top-p，可以引导大模型在生成文本的“创造性”和“可控性”之间找到最佳平衡。在本研究中，各个模型的参数设置参考了官方的默认参数设置以及在实际应用中的效果，具体参数信息如表 5 所示。

表 5 模型参数设置

模型	Tempreture	Top-p
Gemini-3-Pro-Preview	1.0	0.95
DeepSeek-V3.2-Thinking	1.0	0.95
Qwen-3-Max	0.8	0.9
Llama-4-Maverick-17B-128E-Instruct	0.8	0.9

3.3.3 垂直领域大模型构建

在构造自动出题的垂直大模型时，本研究采用有监督微调（Supervised Fine-Tuning，SFT），在模型训练方式的选择上，采取了 LoRA 微调（Low-Rank Adaptation）方法。LoRA 是一种针对大型预训练语言模型的高效微调技术，它旨在解决全参数微调所带来的计算和存储成本问题，其核心思想是冻结预训练模型的原始参数，并通过引入少量可训练的低秩矩阵来模拟参数更新。这样在微调过程中，只需要优化这些低秩矩阵的参数，而不需要修改原始模型的参数，从而大大减少了需要训练的参数量（Manakul et al，2023）。本研究选取了 LLaMA-Factory 作为微调工具，这是一个专为大模型微调而设计的低代码训练框架，它提供了一套完整的工具和接口，以简化和加速大模型的训练、微调和部署过程。微调时的超参数配置如表 6 所示。

表 6 微调超参数设置

参数	取值
训练轮数	3
学习率	3e-4
批处理大小	16
权重衰减系数	0.01
学习率调整策略	Linear
LoRa 秩值	8

3.4 评估策略

在评估大模型生成题目的质量时，为确保评估的全面性和有效性，本研究采取客观指标与主观评价相结合的评估方法。

3.4.1 客观指标

题目自动生成作为文本生成的一项子任务，因此本研究参考了文本生成任务中常用的一些机器指标来评估题目质量。选取了 BLEU（Bilingual Evaluation Understudy）（Papineni et al，2002）、ROUGE（Recall-Oriented Understudy for Gisting Evaluation）（Lin，2004）以及 Distinct（Li，2016）三个指标，取值范围均为 0 至 1，通常用百分比表示，其中利用 BLEU、ROUGE 指标来判断大模型生成题目与参考题目（真题）的表面相似性，利用 Distinct 指标来判断题目的多样性。BLEU 主要用来评估自动出题和参考题目之间的 n-gram 的重叠程度（即相似度），本研究分别将 n 设置为

1、2、3、4，然后求四个结果的均值，计算公式见式（2），其中 BP 为惩罚因子，避免因为题目过短而给出过高分数，计算公式见式（3），lr 表示最短的参考题目长度，lc 为模型生成的题目长度；ROUGE 运用 n-gram 上的召回率（Recall）来衡量自动生成的题目与参考题目之间的相似度，计算公式见式（4），n 通常取值为 1、2 和 L，本研究使用 ROUGE-L 值（最长公共子序列的匹配度）来进行评估；Distinct 是评估文本生成多样性的指标，通过统计生成文本中不重复的 n-gram 的比例来衡量词汇多样性，包括 macro-distinct 和 micro-distinct，其中 macro-distinct 关注单个文本，而 micro-distinct 关注生成的全部文本，本研究使用 micro-distinct 来评估生成题目的多样性，其中 n 设置为 2，计算公式见式（5）。

$$\mathrm{BlEU} = \mathrm{BP} \times exp\left(\sum_{\mathrm{n}=1}^{N} \mathrm{W_n} \times \mathrm{logP_n}\right) \qquad \text{公式（2）}$$

$$\mathrm{BP} = \begin{cases} 1，\mathrm{lr} < \mathrm{lc} \\ exp（1 - \mathrm{lr}/\mathrm{lc}），\mathrm{lr} \geq \mathrm{lc} \end{cases} \qquad \text{公式（3）}$$

$$\mathrm{ROUGE} - \mathrm{N} = \frac{\sum_{\mathrm{S}}\sum_{gram_N} count_{\mathrm{match}}（gram_N）}{\sum_{\mathrm{S}}\sum_{gram_N} count（gram_N）} \times 100\% \qquad \text{公式（4）}$$

$$\mathrm{Distinct} - \mathrm{N} = \frac{count(uningram)}{count(ngram)} \qquad \text{公式（5）}$$

3.4.2 主观评估

在进行人工评价时，参考韩雨婷等（2025）梳理的基于大模型的题目自动生成系统专家审核维度体系，并结合对 HSK 试题的特点分析，从“语言流畅度”“内容准确性”“题目复杂度”“选项干扰性”“答案唯一性”“题型多样性”六个维度出发对生成的题目进行系统评估。其中，“题型多样性”指标是针对同一篇阅读材料所生成的四道题目，在考查形式（如细节题、主旨题、推断题、词义理解题等）上的分布，其余五个维度均针对单道题目进行评价。

在评估人员选择方面，为提升评分效度，所有参与的评分人员均需具备 HSK 命题经验、国际中文教育相关背景或丰富的一线教学经历。本研究共邀请 4 位有经验的命题员参与，包括 1 位北京语言大学国际中文学院讲师（教龄 3 年，多次参与 HSK 命题），以及 3 位北京语言大学评价院语言测试方向博士（均具备 HSK6 级教学经验并多次参与汉语考试命题）。

在具体评估流程上，首先针对测试集中的 30 篇阅读材料，使用不同大模型逐篇生成题目，并与真题进行混编。混编后形成的评估题库共包含 180 篇阅读材料，每篇材料附有四道题目，题目来源涵盖真题与不同模型生成的题目。该题库以盲评形式分发给评估人员，即人工在评分时不知题目具体来源，仅依据上述六个维度进行逐题评估。

在评分方式上，“语言流畅度”“内容准确性”“题目复杂度”“选项干扰性”“题型多样性”五个指标均采用李克特五级量表进行评分，分值越高代表评估人员对该项指标越认同，其中 1 表示很不满意、2 表示不满意、3 表示一般满意、4 表示比较满意、5 表示很满意。其中，“答案唯一性”为定类变量，分为“唯一”与“不唯一”两类，数值 1 代表答案唯一，数值 0 代表答案不唯一包括不存在正确答案和存在多个正确答案两种情况。

4. 实验结果与分析

4.1 生成题目示例

以测试集中的一篇阅读材料为例，不同模型生成的题目对比如表 7 所示，该材料来源于 H61330[3]真题中的第一篇阅读理解题。

表 7 不同模型生成题目对比示例

题目来源	题干	选项
Gemini-3-Pro-Preview	根据文中提到的历史演变，唐朝时面条被称作：	A 煮饼 B 汤饼 C 水引 D 冷淘
	关于山西面食，下列哪项正确？	A 只有蒸和煮两种主要制法 B 有据可查的面食种类超过 280 种 C 所有的面食都统称为“记心火烧” D 家庭主妇通常只会做三四种面食
	孩子入学第一天吃“记心火烧”，寄托了父母怎样的愿望？	A 岁月延绵，年年有余 B 健康长寿，一生平安 C 能够增长学问和心眼 D 能够学会制作各种面食
	最适合做上文标题的是：	A 面食的制作工艺 B 山西面食文化 C 丰富多彩的面食名称 D 中外面食的差异
DeepSeek-V3.2-Thinking	根据材料，面条在唐朝被称作什么？	A 煮饼 B 汤饼 C 水引 D 冷淘
	材料中提到，山西有据可查的面食有多少种？	A 数十种 B 约 200 种 C280 种 D 超过 300 种
	关于山西面食所承载的文化寓意，下列哪一项说法正确？	A 过年吃“接年面”寓意长寿 B 孩子上学吃“记心火烧”是希望孩

[3] H61330 是 HSK6 级考试试卷的编号。

		子专心学习 C 过生日吃拉面是希望年年有余 D“接年面”是在生日时食用的
	外国友人评价“世界面食在中国，中国面食在山西”，旨在说明什么？	A 山西面食的历史最为悠久 B 山西面食的种类最为丰富 C 山西是面食文化的杰出代表 D 山西面食的烹饪方法最复杂
Qwen-3-Max	根据文章，唐朝时期面条的称呼是什么？	A 煮饼 B 汤饼 C 水引 D 冷淘
	文中提到“记心火烧”是在什么场合食用的？	A 过生日时 B 过年时 C 孩子第一天上学时 D 家人远行时
	从文中可以推断出，山西面食之所以种类繁多，主要是因为：	A 山西地理环境适合种植多种粮食作物 B 当地人对面食有深厚的情感和文化认同 C 外国游客的大量需求推动了面食创新 D 政府大力推广面食制作技艺
	文中“娇儿宠称多”这句话的意思最接近于：	A 孩子被宠爱时会得到很多昵称 B 被珍视的事物往往拥有多种称呼 C 面食名称复杂，令人难以记住 D 山西人喜欢给食物起新奇的名字
Llama-4-Maverick-17B-128E-Instruct	第 1 段中的画线词语“推陈出新”是什么意思？	A 翻新 B 更新 C 增加 D 更换
	下列哪一项是山西面食被誉为中国五大面食之一的代表？	A 拉面 B 刀削面 C 记心火烧 D 接年面
	孩子第一天去上学时，山西人会让孩子吃“记心火烧”，其寓意是？	A 希望孩子长寿安康 B 希望孩子多学知识 C 希望岁月延绵不绝 D 希望家中年年有余
	下列哪一项最适合作为本文的标题？	A 山西面食的历史渊源 B 名扬海外的刀削面 C 面食之乡——山西 D 山西面食的制作方法
微调后的模型	山西的刀削面被誉为中国五大面食之一，据材料可知山西有据可查的面食有多少种？	A 数十种 B 两百种 C280 种 D 五百种
	关于山西面食，下列哪项正	A 制法单一

	确？	B 花样繁多 C 以牛肉为主 D 多吃有利于健康
	根据上文，山西人为什么喜欢把面食叫做“精神食粮”？	A 面食营养丰富 B 面食品种繁多 C 面食能治病 D 面食寄托了人们的感情
	下列对文章内容的理解，不正确的一项是？	A 山西刀削面是中国五大面食之一，享誉海外 B 山西面食的称谓变化与时间、地域因素有关 C 山西面食仅能作为充饥的食物，没有其他价值 D 外国友人对山西面食的地位给予了高度认可

4.2 题目质量评估

本研究从客观指标和人工评估两个方面对大模型生成的题目进行全面评价，以确保评估的全面性和准确性。

4.2.1 客观指标

BLEU、ROUGE-L 以及 Distinct 指标的计算结果如表 8 所示。从整体上看，不同模型生成题目的 BLEU 值和 ROUGE-L 值均不高，这主要是由于生成的题目均为选择题，且题干和选项内容均较短，经分词处理后在机器指标上的表现并不好。

表 8 客观机器指标

模型	BLEU	ROUGE-L	Distinct
Gemini-3-Pro-Preview	19.76%	27.42%	66.67%
DeepSeek-V3.2-Thinking	20.64%	27.26%	64.18%
Qwen-3-Max	15.48%	24.64%	65.45%
Llama-4-Maverick-17B-128E-Instruct	16.26%	22.78%	52.08%
微调后的模型	13.24%	20.36%	53.99%

从内容质量指标 BLEU、ROUGE-L 上来看，通用大模型整体优于特定的微调大模型。其中，Gemini-3-Pro-Preview 与 DeepSeek-V3.2-Thinking 表现最为突出，BLEU 分数分别达到 19.76%与 20.64%，ROUGE-L 分数均超过 27%，说明二者在词汇匹配与语义覆盖方面与参考题目具有相对较高的相似性，生成的题目在内容上更贴近人工命题风格；在多样性指标 Distinct 上，Gemini-3-Pro-Preview 与 Qwen-3-Max 分别取得 66.67%与 65.45%的最高值，表明其生成题目的用词变化丰富，避免了重复与模板化表达。而微调后的模型虽然在 Distinct 上略高于 Llama-4-Maverick，但仍显著低于其它模型，反映出其生成文本的多样性相对有限。总的来说，通用大模型特别

是 Gemini-3-Pro-Preview、DeepSeek-V3.2-Thinking 两个推理大模型在题目生成的内容相关性与语言多样性方面均表现更佳；而专门微调的模型并未显示出预期优势。

4.2.2 人工评估

人工评分结果如表 9 所示，以真题得分为基准，重点考察各模型与真题的接近程度，差值越小代表模型表现越佳，越接近人工命题水平。其中括号中的数据表示模型在不同维度上得分与真题的差距，高于真题为“+”，低于为“-”。

表 9 评分结果

题目来源	语言流畅度	内容准确性	题目复杂度	选项干扰性	题型多样性
真题	4.58	4.65	3.96	3.92	3.98
Gemini-3-Pro-Preview	**4.56（-0.02）**	4.55（-0.10）	3.26（-0.70）	**3.55（-0.37）**	**4.02（+0.04）**
DeepSeek-V3.2-Thinking	4.61（+0.03）	**4.62（-0.03）**	**3.34（-0.62）**	3.46（-0.46）	4.10（（+0.12）
Qwen-3-Max	4.50（-0.08）	4.38（-0.27）	3.08（-0.88）	3.28（-0.64）	3.92（-0.06）
Llama-4-Maverick-17B-128E-Instruct	4.55（-0.03）	4.24（-0.41）	3.21（-0.75）	3.15（-0.77）	3.74（-0.24）
微调后的模型	4.48（-0.10）	4.08（-0.57）	3.28（-0.68）	2.90（-1.02）	3.25（-0.73）

从整体上看，所有模型在语言流畅度与内容准确性两个基础维度上表现最佳，与真题得分差距极小，表明大模型在语言规范性与内容忠实性上较为接近人工命题水平。然而，在体现命题专业重要能力的题目复杂度与选项干扰性维度上，所有模型均与真题存在显著差距，显示出大模型在高阶认知考查与精细选项设计方面存在明显短板。具体而言：

在语言流畅度上，大模型生成题目的流畅度与人类较为接近，在大部分题目中评估人员没有感受到人与机器的明显差异，显示了大模型强大的自然语言生成能力，生成的文本具有较好的拟人度。

在内容准确性上，DeepSeek-V3.2-Thinking 和 Gemini-3-Pro-Preview 的表现最好，与真题的差距更小，展现出其强大的上下文理解能力，生成题目与原文具有较高的一致性。

在题目复杂度上，大模型生成题目与真题存在明显差距，不能很好地控制题目难度，其中 DeepSeek-V3.2-Thinking 的表现与真题差距最小，经评估人员反馈，大模型生成的部分题目存在难度过大的情况，选项中出现难度过大的成语（如“病入膏肓”）以及难以分辨的近义词（如“翻新”和“更新”）。

在选项干扰性上，Gemini-3-Pro-Preview 相对表现最好，部分题目错误选项的干扰性较小，不能有效设计出基于典型错误、具有合理迷惑性的选项，如词语辨析、固定词语搭配等。

在题型多样性上，除微调后的模型表现较差外，其它模型表现与真题较为接近，覆盖了细节题、推理题、词义题、主旨大意题等多种题型。

不同大模型生成题目的答案唯一性占比如表 10 所示。易知，通用大模型生成题目具备较好的可回答性，答案唯一性占比均超过 90%；而经微调后的模型在该指标上表现显著落后，所生成的题目存在“无正确选项”和“多个正确选项”的情况，不符合单项选择题的基本设计要求。这一结果说明，答案设计的逻辑自洽性是影响大模型命题质量的关键因素，后续可通过加强答案定位验证机制、严格检验选项唯一性等方法，进一步提升生成题目的可靠性与可用性。

表 10 答案唯一性指标

题目来源	答案唯一	答案不唯一
Gemini-3-Pro-Preview	92.26%	7.74%
DeepSeek-V3.2-Thinking	92.15%	7.85%
Qwen-3-Max	93.64%	6.36%
Llama-4-Maverick-17B-128E-Instruct	91.35%	8.65%
微调后的模型	82.47%	17.53%

此外，从不同类别大模型的表现差异来看：首先，以 DeepSeek-V3.2-Thinking、Gemini-3-Pro-Preview 为代表的推理型大模型综合表现最优。这很可能得益于其内部的思维链推理机制，使其在题目生成过程中能够更好地模拟人工命题的认知流程，逐步理解文本、定位考查点、构思干扰项并确保答案唯一性，从而在内容准确性、选项干扰性等重要维度上更接近真题；其次，以 Qwen-3-Max 为代表的指令型大模型在语言流畅度与题型多样性等维度上表现良好，说明其能够较好遵循生成指令与格式要求。然而，其在题目复杂度与选项干扰性等需深层文本理解与逻辑设计的指标上略逊于推理型模型；最后，本研究所采用的微调垂直大模型整体表现不佳，可能在当前训练数据规模与微调策略下，模型未能充分融合领域知识并保持原有生成能力，可见，在优质训练数据受限的情况下，微调并非是有效提升大模型执行具体教学任务能力的手段。

5. 讨论

总的来说，大模型在题目生成任务上具有显著潜力，但仍在一定程度上存在题目难度控制、题目可回答性等问题，未来应建立人机交互的题目自动生成机制，重点发挥大模型在题目生成中的效率优势，并通过人工审核来提高生成题目的可靠性。为帮助教师减负增效，本研究提出以下两条使用建议。

5.1 优化提示设计与模型遴选，实现精准生成

教师应依据具体的教学目标与考查重点，采用结构化、精细化的提示设计并选择适配的大模型，以提升生成题目的质量与适用性。在提示设计中，应明确题目考查的认知层次、题型及难度要求，并可提供少量示例作为格式与风格的参考，从而有效引导模型输出。此外，可通过迭代优化提示内容，逐步形成稳定、高效的结构化提示模板。在模型选择上，可依据任务特性进行遴选：若侧重语言规范与格式准确性，可优先选用指令遵循能力强的模型如 Qwen-3-Max；若需加强题目的逻辑深度与高阶思维考查，则建议选用具有显式推理能力的模型如DeepSeek-V3.2-Thinking。通过结构化的提示设计与模型匹配，能够显著提升生成题目与教学情境的契合度与可用性。

5.2 建立“生成—审核—修订”的人机协同流程，实现闭环优化

教师可将大模型作为题目的初步批量生成工具，随后基于专业判断对生成结果进行审核。审核应重点关注题目与教学目标的契合度、难度是否合理、题目是否可回答。对于未达标的题目，教师可进行针对性修订或提供明确修改指令，重新投入生成环节。通过多次“生成—审核—修订”的迭代，形成持续优化的闭环流程。此机制既能发挥大模型在快速大量生成上的效率优势，又能确保最终题目经过严格的专业把关，从而在提升命题效率的同时保障题目的科学性与适用性。

6. 结语

本研究以 HSK6 级阅读理解题为例，探索了国内外不同类别大模型自动出题的效果，为大模型赋能教育领域自动出题的智能化转型提供实践参考。研究表明，大模型在自动出题任务上具有显著潜力，不同类别模型表现存在差异，其中推理大模型的整体表现更优，但在选项设计、难度控制、可回答性等方面与人工命题尚存在明显差距，无法直接用于教学实践，因此尚不能替代专业命题教师，而更适合作为辅助生成工具，在人工审核与修订的基础上投入教学使用。未来，可以进一步探索多模态大模型在图文题、听力题等多种题型上的应用效果，推动国际中文教育在资源建设上的数智化转型。

致谢：本文受教育部中外语言合作交流中心国际中文教育研究课题重大项目“面向国际中文教育的生成式人工智能（AIGC）应用研究”（24YH03A）；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）“融合LLM和GraphRAG的自动命题方法研究”（26YCX040）的资助。徐娟为本文通讯作者。

参考文献

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

Cao, X. W., Feng, L. P., Wu, Z. W., Lu, G., Wang, J. M., Zhang, Y. L., Zhu, Y., Liang, X., & Zhang, X. S. (2025). Discussion on the new HSK 3.0 syllabus and exam promotion. *Journal of Yunnan Normal University (Teaching and Research of Chinese as a Foreign Language), 23*, 1-8.[曹贤文，冯丽萍，吴中伟，路广，王佶旻，张艳莉，朱勇，梁霞，& 张新生. (2025). "HSK3.0 新考纲及考试推广"大家谈. *云南师范大学学报(对外汉语教学与研究版), 23*, 1-8.]

Chen, X., Li, M. R., Zhou, Y. Q., Zhou, T., & Zhang, F.(2024). Research on the automatic generation path of test questions based on large language models. *China Examinations, 2024*, 39-48. [陈欣，李蜜如，周悦琦，周同，& 张峰. (2024). 基于大语言模型的试题自动生成路径研究. *中国考试, 2024*, 39-48.]

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics.

Dong, Q., Li, L., Dai, D., Xu, C., Zhu, Y., Sun, G., Sun, C., Jiang, S., Jia, Y., Sui, Z., & Chang, B. (2023). *A survey on in-context learning.* arXiv. https://doi.org/10.48550/arXiv.2301.00234

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1342-1352). Association for Computational Linguistics.

Han, X. X., Ma, R. L., & Xu, J. (2025). Exploring the technical path of DeepSeek empowering international Chinese teaching resource construction: A case study of graded reading text generation. *International Chinese Language Teaching Research, 2025*, 30-40. [韩欣欣，马瑞祾，& 徐娟. (2025). DeepSeek 赋能国际中文教学资源建设的技术路径探索——以分级阅读文本生成为例. *国际汉语教学研究, 2025*, 30-40.]

Han, Y. T., Wang, W. X., Liu, H. Y., & You, X. F. (2025). Technological innovations and practical challenges in automatic question generation. *Advances in Psychological Science, 33*, 1766-1782. [韩雨婷，王文轩，刘红云，& 游晓锋. (2025). 题目自动生成的技术革新与现实挑战. *心理科学进展, 33*, 1766-1782.]

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 609-617). Association for Computational Linguistics.

Hou, Z. Y., & Xu, J. (2025). Research on personalized generation of international Chinese reading materials based on large language models. *International Chinese Language Education (Chinese and English), 10*, 32-44. [侯泽煜，& 徐娟. (2025).

基于大语言模型的国际中文阅读材料个性化生成研究. *国际中文教育(中英文), 10,* 32-44.]
Jiang, S., & Lee, J. S. Y. (2017). Distractor generation for Chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 143-148). Association for Computational Linguistics.
Lai, Y. X., Wang, Y. D., & Wang, L. (2024). Subject test question generation method based on large language model and retrieval enhancement. *Journal of Chinese Information Processing, 38*, 148-158. [来雨轩，王艺丹, & 王立. (2024). 基于大语言模型与检索增强的学科试题生成方法. *中文信息学报, 38*,148-158.]
Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 110-119). Association for Computational Linguistics.
Li, J. Y., & Xu, J. (2025). Development and evaluation of international Chinese micro-courses based on generative artificial intelligence: A case study of elementary level sentence patterns. *Journal of Gannan Normal University, 46*, 53-61. [李嘉仪, & 徐娟. (2025). 基于生成式人工智能的国际中文微课开发与评价——以初等水平句式为例. *赣南师范大学学报, 46,* 53-61.]
Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* (pp. 74-81). Association for Computational Linguistics.
Liu, Y. P., Ou, Z. G., & Wu, X. Q. (2025). Effectiveness evaluation of generative artificial intelligence empowering international Chinese teaching: A case study of instructional design, HSK mock test question writing, and essay scoring. *Journal of Ethnic Education Research, 36* 156-166. [刘玉屏，欧志刚, & 武晓琴. (2025). 生成式人工智能赋能国际中文教学的效果测评——以教学设计、HSK 模拟试题编写及作文评分为例. *民族教育研究, 36,* 156-166.]
Ma, R. L., & Xu, J. (2024). Innovative research and development of intelligent teaching resources for international Chinese writing in the digital intelligence era. *International Chinese Language Teaching Research, 2024*, 13-23. [马瑞祾, & 徐娟. (2024). 数智时代国际中文写作智慧教学资源创新研发. *国际汉语教学研究, 2024,* 13-23.]
Manakul, P., Liusie, A., & Gales, M. J. F. (2023). *SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models.* arXiv. https://doi.org/10.48550/arXiv.2303.08896
Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing* (pp. 17-22). Association for Computational Linguistics.
Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30,* 5998-6008.

Wang, H. B., & Lyu, H. H. (2025). Research on automatic generation of Chinese reading test questions based on large language models. *International Chinese Language Teaching Research, 2025*, 41-54. [王鸿滨, & 吕海辉. (2025). 基于大语言模型的中文阅读测试题自动生成研究. *国际汉语教学研究, 2025,* 41-54.]

Wang, H. S., & Xie, F. (2024). Research on innovation of translation education practice model driven by large language model technology. *Chinese Translators Journal, 45*, 70-78. [王华树, & 谢斐. (2024). 大语言模型技术驱动下翻译教育实践模式创新研究. *中国翻译, 45,* 70-78.]

Wang, Y. M., Bin, S., & Zhao, Y. (2025). Research on the development and application of large language models empowering Chinese as a second language reading tests. *Language Teaching and Linguistic Studies, 2025*, 1-12. [王亚敏，宾帅, & 赵杨. (2025). 大语言模型赋能中文二语阅读测试的研发与应用研究. *语言教学与研究, 2025,* 1-12.]

Xu, J. (2023). Automatic generation of reading comprehension questions supported by semantic graphs. *Journal of Intelligent Systems, 19*, 420-428. [徐坚. (2023). 语义图支持的阅读理解型问题的自动生成. *智能系统学报, 19,* 420-428.]

Yu, Y. F., Qian, X. H., Zhong, Y. H., Li, Y. M., Song, M. S., Xie, X. Q., Han, B. C., Ma, X. N., Gu, C. Y., Bai, L. M., Bai, J. H., Hao, Q. X.,Li, D. G., & Ru, S. (2025). HSK3.0 and the new ecology of international Chinese education: Standard innovation, technological empowerment and global development—A multi-perspective discussion. *International Chinese Language Education, 10*, 1-17. [郁云峰，钱旭红，钟英华，李宇明，宋明顺，谢小庆，韩宝成，马西尼，古川裕，白罗米，白建华，郝清新，李登贵, & 茹丝. (2025). “HSK3.0 与国际中文教育的新生态：标准革新、技术赋能与全球发展”多人谈. *国际中文教育(中英文), 10,* 1-17.]

Journal of Technology and Chinese Language Teaching Volume 17 Number 1, June 2026
http://www.tclt.us/journal/2026v17n1/huangkashiwagikang.pdf pp. 48-65

Comparing Automatic Speech Recognition and Teacher Assessments of Japanese Learners' Mandarin Chinese Pronunciation: Accuracy, Agreement, and Pronunciation Difficulty Detection
(自动语音识别与教师对日本汉语学习者普通话发音评估的比较：准确性、一致性及发音困难识别)

Huang, Weihsun
(黄暐勋)
Kobe University
(神户大学)
223c302@stu.kobe-u.ac.jp

Kashiwagi, Harumi
(柏木治美)
Kobe University
(神户大学)
kasiwagi@kobe-u.ac.jp

Kang, Min
(康敏)
Kobe University
(神户大学)
kang@kobe-u.ac.jp

Abstract: Computer-assisted pronunciation training (CAPT) increasingly incorporates automatic speech recognition (ASR) to provide pronunciation assessment and feedback. However, the extent to which ASR systems evaluate non-native Mandarin Chinese speech in a manner comparable to human teachers remains unclear. This study compares the assessments generated by three ASR systems—Whisper, Azure, and Gladia—with ratings provided by native Chinese-speaking teachers for the word-level Mandarin Chinese pronunciation of 31 Japanese learners. Two research questions are addressed: (1) To what extent do these ASR systems assess learner pronunciation comparably to teachers? (2) Can ASR assessments help identify learners' pronunciation difficulties? A three-point scoring scheme was developed to evaluate learners' productions of 20 Mandarin Chinese words. Comparative analyses were conducted from the perspectives of learner proficiency and pronunciation characteristics. The results showed that all three ASR systems generally underestimated learner performance relative to teacher ratings, although Whisper produced assessments that were most consistent with those of the teachers. The agreement between ASR and teacher assessments also varied according to learner proficiency. Furthermore, ASR performance was strongly influenced by initial–final combinations, suggesting that ASR assessments can help identify specific pronunciation difficulties. These findings support the potential of ASR as a complementary tool for pronunciation assessment in Mandarin Chinese CAPT.

摘要：随着计算机辅助发音训练（Computer-Assisted Pronunciation Training, CAPT）的发展，自动语音识别（Automatic Speech Recognition, ASR）系统日益广泛地应用于发音评估与反馈。然而，ASR 系统对非母语者普通话发音的评估能否达到与教师相近的水平，仍缺乏充分的实证研究。本研究比较 Whisper、Azure 和 Gladia 三种

ASR 系统与中文母语教师对 31 名日本学习者 20 个汉语词语发音的评估结果，以探讨：（1）ASR 系统在词语层面的发音评估与教师评分具有多大程度的一致性？（2）ASR 评估是否有助于识别学习者的发音困难？本研究建立三级评分标准，对学习者发音分别进行 ASR 评分与教师评分，并从学习者水平和发音特征两个层面比较两者的评估结果。研究结果显示，三种 ASR 系统均倾向于低估学习者的发音表现，其中 Whisper 与教师评分的一致性最高。此外，ASR 与教师评分的一致程度会因学习者水平而有所不同。进一步分析发现，ASR 评估结果受到声母—韵母组合的显著影响，表明 ASR 评估有助于识别学习者的具体发音困难。本研究结果支持 ASR 作为普通话计算机辅助发音训练中辅助发音评估工具的应用潜力。

Keywords: Automatic speech recognition, ASR Assessment, Japanese learner of Chinese, Teacher assessment, Computer-assisted pronunciation training (CAPT)

关键词: 自动语音识别，ASR 评估，日本汉语学习者，教师评估，计算机辅助发音训练（CAPT）

1.Introduction

Mandarin Chinese courses are offered as required electives at many universities in Japan. First-year students who begin studying Chinese as a foreign language are typically required to take two classes per week and continue their studies for 15 weeks during both the spring and fall terms. Owing to large class sizes, students often have limited opportunities for pronunciation practice, a challenge common to foreign language classroom instruction (Gao, 2025; Chen, 2011).

Computer-assisted pronunciation training (CAPT) has been recognized as an effective pedagogical approach for improving learners' pronunciation, particularly in the context of English as a foreign language (EFL) (Fouz-González, 2015). CAPT systems frequently incorporate automatic speech recognition (ASR) technology, which is used to detect phonetic errors, provide corrective feedback, ultimately enhance learners' pronunciation and potentially improve awareness of grammatical features (Ehsani & Knodt, 1998; Burleson, 2007; Neri et al., 2008; Eskenazi, 2009; Wang & Young, 2014; Tsai, 2019; McCrocklin, 2019; Dai & Wu, 2023; Issa & Hahn-Powell, 2025). ASR-based CAPT has also been applied to the learning of Chinese as a foreign language (CFL) (Da, 2015; Zhao et al., 2019; Watanabe et al., 2019, Li et al., 2024).

Despite these developments, the accuracy of ASR in evaluating non-native speech remains a matter of ongoing concern (Ehsani & Knodt, 1998, Derwing et al, 2000; Sunaoka, 2018; McCrocklin et al., 2019; Inceoglu et al., 2023; Hirai & Kovalyova, 2024). Eskenazi (2009) emphasized that CAPT systems employing ASR should be capable of detecting individual pronunciation errors and assessing fluency in a manner comparable to human experts. Similarly, O'Brien et al. (2018) highlighted the need to identify ASR-derived measure that align closely with human judgements. Even with recent technological advances, the question of how accurately ASR systems recognize learner speech remains critical. A recent study investigating the performance of five speech-to-text applications for EFL learners has shown that recognition accuracy is influenced not only by the systems' technical capabilities but also by characteristics of the learners' utterances (Hirai & Kovalyova, 2024).

In the domain of CFL, Sunaoka (2018) analyzed the recognition accuracy of non-native speech in a Chinese long-distance group discussion using the ASR function integrated in Google Translation, arguing that teachers must verify ASR evaluations to compensate for technological limitations. Nevertheless, it remains unclear how accurately contemporary ASR technologies recognize CFL learners' speech, and which ASR-derived assessments most closely reflect human judgements. To address these gaps, the present study compares assessments of Japanese learners' speech generated by three ASR systems with evaluations provided by native Chinese-speaking teachers.

2. Literature Review

2.1 Computer-Assisted Pronunciation Training Systems with ASR

A substantial number of CAPT systems incorporating ASR have been developed for various EFL learning purposes. Burleson (2007) employed ASR to improve segmental errors produced by non-native speakers. Five Mandarin-speaking learners of English underwent pronunciation training targeting six phonemic contrasts. The ASR was used to recognize their productions and provide feedback, while native English listeners evaluated pre- and post-training recordings using forced-choice minimal pair tasks. The results demonstrated a significant improvement in learners' segmental intelligibility. Wang and Young's (2014) ASR-based iCASL system further examined the presentation of corrective feedback and demonstrated the effectiveness of a three-level feedback scheme. Windows Speech Recognition (WSR), a built-in speech recognition tool in Microsoft Windows, has also been integrated into sentence dictation practice, with findings suggesting that it can serve as a useful complement to face-to-face pronunciation instruction (McCrocklin, 2019). More recently, Issa and Hahn-Powell (2025) reported the use of a fine-tuned speech model within a CAPT system to investigate the effectiveness of ASR corrective feedback on the pronunciation of Arabic. These studies illustrate the diversity of ASR technologies employed in CAPT systems.

In the field of CFL, Da (2015) introduced Google's ASR, embedded in Chrome browser, into classroom pronunciation practice for ten non-native learners, suggesting that meaningful or frequently used expressions may be more suitable for Pinyin activities than isolated syllables groups. Zhao et al. (2019) developed "KoToToMo", a smart phone-based system for read-aloud practices that utilizes operating-system speech recognition. In addition to repetition and shadowing tasks, the system allows learners to conduct pronunciation "trials", enabling them to confirm their performance based on recognition results and receive feedback. Watanabe et al. (2019) proposed the "ST-lab" system, which integrates both ASR and text-to-speech (TTS) technologies via the Web Speech API. In its "Reading Aloud Practice" module, the ASR component evaluates learners' pronunciations and displays the recognition results, allowing learners to repeat items until a "correct answer" is achieved or optionally skip them. As with EFL, a wide range of ASR technologies has been integrated into CAPT systems for CFL (Da, 2015; Wei & Zhang, 2018; Watanabe et al., 2019; Zhao et al., 2019).

2.2 ASR Assessment Accuracy of Learner Speech

The accuracy of ASR remains a critical issue in its deployment within CALL environments, largely because most commercial speech recognizers are trained on standard native pronunciations (Ehsani & Knodt, 1998). Derwing et al. (2000) evaluated the effectiveness of a widely used ASR package in providing corrective feedback based on two criteria: whether the ASR system recognizes ESL speech at an acceptable level and whether the ASR system has the potential to identify production difficulties. They argued that the usefulness of ASR depends on how closely its assessments of ESL speech approximate those of native listeners, and they emphasized the need for careful evaluation of ASR applications according to these criteria.

McCrocklin et al. (2019) examined the accuracy rates of Windows Speech Recognition (WSR) and Google Voice Typing in recognizing the speech of advanced non-native English speakers and found that Google's system outperformed WSR. Inceoglu et al. (2023) compared the assessments generated by Google Assistant for four non-native speakers' accented English to those of native listeners. Their findings showed that the consistency between ASR evaluations and native listener judgments varies depending on the speaker and the type of oral production.

A more recent study evaluated the accuracy of America English transcriptions produced by five speech-to-text applications: Google Docs Voice Typing, Apple Dictation, Windows 10 Dictation, Dictation.io and "Transcribe" by comparing them with human-generated transcriptions (Hirai & Kovalyova, 2024). Thirty non-native speakers completed four speaking tasks, including reading a short passage and answering freely to questions. Consistent with the findings of Inceoglu et al. (2023), the study revealed that accuracy is shaped not only by the ASR systems' recognition capabilities but also by the type of speech and the influence of learners' L1 on their L2 productions. These results suggest that different ASR systems may yield varying levels of accuracy and, consequently, differing degrees of instructional effectiveness for non-native speakers.

Despite extensive research on ASR accuracy in EFL contexts, this issue has received little attention in the field of CFL. How effectively ASR systems assess the pronunciation of CFL learners remains largely unexplored.

2.3 Pronunciation Difficulties

As noted earlier (Derwing et al., 2000), it is essential to determine whether an ASR has the potential to identify learners' production difficulties. Below, we summarize pronunciation challenges commonly observed among Japanese learners of CFL.

The pronunciation difficulties of Japanese learners primarily arise from fundamental differences between the phonological systems of Chinese and Japanese. Through contrastive analysis, Lin (2019) observed that the Chinese phonetic inventory is more complex than that of Japanese, containing retroflex consonants not found in Japanese, lacking the vowel /e/, and presenting challenges in distinguishing the nasal finals /n/ and /ng/. Using the NTNU Chinese Learner Corpus of Interlanguage Phonology, Fang et al. (2015) conducted a systematic error analysis and reported that, for initial consonants, the labiodental fricative /f/ was often realized as the Japanese bilabial aspirated sound /フ/. Errors involving aspirated consonants were also common, particularly when aspirated sounds occurred in word-final positions. Regarding final errors, the high rounded vowel /u/ showed the highest error rate among all finals. For the rounded vowel /ü/, Japanese learners tended to struggle due to the lack of rounded front vowels in Japanese and their unfamiliarity with lip rounding in this context.

These findings suggest that if an ASR system can evaluate learners' utterances in a manner comparable to human teachers, it could greatly assist teachers in providing targeted corrective feedback and enable learners to address errors promptly during CAPT activities.

In this study, we investigate how accurately three ASR systems assess the Mandarin Chinese speech of Japanese learners by comparing their assessments with those of native Chinese-speaking teachers. We address this issue by examining the following research questions:

1. To what extent do these ASR systems assess learners' word-level utterances in a manner comparable to teachers?
2. Do these ASR systems have the potential to identify pronunciation difficulties?

3. Methodology

3.1 Recognizers and Words to Pronounce

3.1.1 Automatic Speech Recognition Systems

This study employed three automatic speech recognition (ASR) systems: Whisper by OpenAI, Azure by Microsoft, and Gladia by Gladia Inc. All three systems support multiple languages and demonstrate high recognition accuracy, having been widely adopted in commercial transcription services. Whisper, in particular, is open source, providing substantial flexibility for research and development. In the present study, a downloaded desktop version of Whisper was used (Whisper Desktop, 2023), while the other two ASR systems were accessed via their online platforms (Microsoft, 2024; Gladia, 2024).

3.1.2 Word Selection

Based on Fang et al.'s (2015) findings that Japanese learners of Mandarin primarily struggle with distinguishing aspirated and unaspirated consonants, producing the vowel /e/ (which does not exist in Japanese), articulating the labiodental fricative /f/, realizing retroflex consonants such as /zhi/, /chi/, /shi/, /ri/, and /er/, and pronouncing nasal finals including /an/, /ang/, /en/, and /eng/, this study selected 20 two-syllable words that contain these pronunciation features as speech materials for a preliminary investigation. Based on our teaching experience, we considered two-syllable words to be relatively easy for beginning learners to pronounce. The selected items were aligned with the learners' instructional progression. The target words are listed in Table 1.

Table 1 Words to Pronounce

客气, 学校, 放心, 词典, 听懂, 注意, 好吃, 老师, 日本, 二十
咖啡, 工作, 你家, 买菜, 北京, 便宜, 很远, 常常, 告诉, 别走

3.2 Scoring Criteria

3.2.1 Speech Intelligibility and Comprehensibility

Foreign language learner pronunciation assessment can be approached from the perspective of native listener comprehension, which is typically divided into two dimensions: *intelligibility* and *comprehensibility* (Munro & Derwing, 1999). *Intelligibility* refers to the extent to which listeners can accurately identify the linguistic content produced by the speaker (e.g., phonemes and words), emphasizing objective phonetic recognition. *Comprehensibility*, on the other hand, concerns the degree of effort required for listeners to understand the speaker's intended meaning, representing a more subjective, global evaluation.

In the context of ASR assessment, intelligibility constitutes the primary evaluative dimension (Inceoglu et al., 2023), as ASR systems rely predominantly on acoustic features and lack the ability to process higher-level linguistic, contextual, or pragmatic information. However, in real classroom settings, teachers often find learners' speech easier to understand than native listeners without teaching experience, particularly those unfamiliar with non-native speech patterns. Since listener comprehension can influence learners' motivation to engage in pronunciation practice, comprehensibility remains an important perspective in assessing foreign language pronunciation. Although *intelligibility* and *comprehensibility* are theoretically distinct constructs, they may be treated similarly in classroom practice for pedagogical purposes.

In other words, if speech recognizers could "understand" learner speech in a manner similar to teachers, learners might be more motivated to practice pronunciation. Therefore, this study compares ASR recognition outcomes (intelligibility) with teacher evaluations of comprehensibility to examine the degree of alignment between ASR systems and human instructors in assessing learners' word-level Mandarin pronunciation. Through this comparison, we aim to explore the potential for ASR systems to provide teacher-like assessments.

3.2.2 Scoring Learner Speech

Table 2 Pronunciation Scoring Criteria

Score	ASR Criteria	Teacher Criteria
0	Both characters unrecognizable	Word meaning incomprehensible
1	One character correctly recognized	Pronunciation unclear but meaning comprehensible
2	Both characters correctly recognized	Pronunciation clear and meaning comprehensible

Since ASR systems output only recognition results, each recognized Chinese character was regarded as correct if it corresponded to the target character shown in Table 1. This study adopted a three-level scoring scheme for each word of the 20 two-syllable words: 0 points if neither character was recognized, 1 point if only one character was correctly recognized, and 2 points if both characters were correctly recognized. This scoring method corresponds to the Character Error Rate (CER) evaluation framework, which is calculated based on three error types—substitution, deletion, and insertion. As the focus of this study is on pronunciation at the word level, homophonic outputs generated by the ASR systems were treated as correct.

Teacher scoring was conducted by five native Mandarin-speaking instructors with extensive experience teaching Chinese as a foreign language in Japan. This suggests that the instructors are familiar with Japanese learners' pronunciation and may therefore be more tolerant when evaluating learner speech. The raters listened to the learners' recorded utterances and assigned scores on a three-point comprehensibility scale: 0 points if the word's meaning was completely incomprehensible, 1 point if the pronunciation was unclear but the meaning remained interpretable, and 2 points if the pronunciation was clear and the meaning fully comprehensible. The five teachers rated the samples independently without consultation. This multi-rater design reduces the influence of individual subjective tendencies and enhances the overall reliability of the scoring. The scoring criteria are summarized in Table 2.

3.3 Speakers and Procedure of Recording Utterances

The learner speech data consisted of word-level utterances produced by 31 university students who were taking introductory Chinese courses for the first time. These students received two 90-minute Chinese classes per week and had completed phonetic instruction prior to the recording sessions. The target words, along with their pinyin and Japanese translations, were provided for in-class pronunciation practice, and additional practice was assigned as homework.

Students were instructed to make their recordings in quiet environments using their smartphones and to submit the audio files via the learning management system at the university. They made the recordings individually. The researcher then converted the submitted audio files into formats compatible with each ASR system and scored each learner's pronunciation of each word based on the criteria shown in Table 2.

4. Results

4.1 Interrater Reliability

To address RQ1, we summed the scores of all target words rated by the three ASR systems and the five teachers for each student. Each total score ranged from 0 to 40. Although the data were quantitative, they did not follow a normal distribution.

Before using the mean teacher scores as the representative measure of human assessment, we first examined interrater reliability among the five teachers. The intraclass correlation coefficient (ICC) was calculated based on the total scores, and Fleiss's kappa coefficients were computed for each individual word. All statistical analyses were conducted using IBM SPSS Statistics 30.0.

The results indicated substantial agreement among the teachers (ICC = 0.63) (Landis & Koch, 1977). Across the 20 words, Fleiss's kappa values ranged from 0.20 (slight agreement) to 0.63 (substantial agreement). Specifically, one word showed slight agreement (北京: $\kappa = 0.20$, $z = 3.69$, $p < .001$), 11 words showed fair agreement ($\kappa = 0.21$–0.40), seven words showed moderate agreement ($\kappa = 0.41$–0.60), and one word showed substantial agreement (别走: $\kappa = 0.64$, $z = 11.93$, $p < .001$). Among the ASR systems, a fair level of agreement was observed (ICC = 0.55). Following Inceoglu et al. (2023), we consider the teacher ratings to be sufficiently reliable and therefore use the mean teacher scores as the representative measure of teacher assessment.

4.2 Rating Results

The values shown in Table 3 represent the mean scores and standard deviations of students' total scores as assessed by the three ASR systems and the teachers. To maintain consistency with the scoring criteria described in Section 3.2.2, the total scores were divided by 20 so that the resulting values correspond to the mean and standard deviation per word. T1–T5 denote the five teachers.

Table 3 Rating Results by ASR and teachers

Rater	Whisper	Azure	Gladia	T1	T2	T3	T4	T5	Teacher
M	1.46	1.25	1.42	1.69	1.56	1.57	1.45	1.38	1.53
SD	0.27	0.30	0.28	0.17	0.24	0.24	0.34	0.32	0.24

The results show that the mean teacher score was 1.53 (SD = 0.24). Among the three ASR systems, Whisper achieved the highest performance (M = 1.46, SD = 0.27), followed by Gladia (M = 1.42, SD = 0.28), while Azure demonstrated the lowest performance (M = 1.25, SD = 0.30).

The individual teacher scores (T1–T5) indicate that, although the overall standard deviation (0.24) was lower than those of the ASR systems, noticeable variation remained across teachers, with standard deviations ranging from 0.17 (T1) to 0.34 (T4). These findings suggest that even among experienced Chinese language instructors, individual differences persist in evaluating learner pronunciation. This underscores the challenge of establishing fully standardized human assessment and highlights the potential utility of ASR-based evaluation.

The Wilcoxon signed-rank test results indicated that the difference between Whisper scores and teacher scores was not statistically significant, whereas the scores from Azure and Gladia differed significantly from teacher scores. These findings suggest that Whisper provided a more teacher-like assessment of the 20-word utterances produced by the 31 learners compared with Azure and Gladia.

To present the differences between ASR systems and teacher ratings in a more intuitive manner, we calculated percentage difference scores using the formula

(ASR Score − Teacher Score) / Teacher Score × 100%.

The results are summarized in Table 4.

Table 4 Percentage Differences

ASR System	Percentage Difference
Whisper	-4.32%
Azure	-18.57%
Gladia	-7.09%

On average, the results suggest that Azure had the most difficulty recognizing the learners' Chinese word-level utterances compared with the other two systems.

4.3 Influence of Learner Proficiency

Based on the five-number summary of the teacher average scores, the 31 students were divided into three proficiency groups:

1. Low-proficiency group (0.99–1.45): 10 learners (Learners 1–10)
2. Medium-proficiency group (1.48–1.58): 10 learners (Learners 11–20)
3. High-proficiency group (1.64–1.96): 11 learners (Learners 21–31)

Learner 1 received the lowest teacher score (0.99), whereas Learner 31 received the highest (1.96). In comparison, the ASR system ratings yielded the following distributions:

- Whisper classified 5 learners as low-, 4 as medium-, and 10 as high-proficiency;

- Azure classified 6 learners as low-, 4 as medium-, and 7 as high-proficiency;
- Gladia classified 7 learners as low-, 5 as medium-, and 8 as high-proficiency.

These results suggest that the consistency between ASR systems and teacher assessments increased for learners in the high-proficiency group.

Regarding extreme values, Whisper assigned the lowest score to Learner 4 (0.60), Azure to Learners 3 and 16 (0.70), and Gladia to Learner 5 (0.75). Whisper's highest score was for Learner 22 (1.90), Azure's for Learner 27 (1.85), and Gladia's for Learners 22, 24, and 28 (1.80). The variation in highest and lowest scores across teachers and ASR systems indicates that ASR assessments are influenced by learner proficiency levels.

4.4 Extremum Cases and Pronunciation Characteristics

According to the percentage difference values, the ASR scores for Learner 1 were as follows: Whisper overestimated the learner's performance by 36% (1.35), Azure underestimated it by 14% (0.85), and Gladia overestimated it by 46% (1.45).

Several characteristics emerged from Learner 1's pronunciation analysis conducted by the authors. The pronunciations of "老师" and "日本" were clearly inaccurate, and most teachers assigned scores of 0 or 1. However, both Whisper and Gladia correctly recognized these utterances. Conversely, Whisper and Azure failed to recognize "客气", whereas Gladia succeeded. These findings suggest that ASR assessments for low-proficiency learners are influenced not only by the learners' pronunciation characteristics but also by system-specific recognition criteria In this case, Azure's score was the closest to the teacher assessment, and the difference was not statistically significant according to the Wilcoxon signed-rank test.

For the highest-proficiency learner (Learner 31, teacher score = 1.96), the ASR scores were as follows: Whisper underestimated the performance by 18% (1.61), Azure by 13.27% (1.70), and Gladia by 11% (1.74). Pronunciation checks conducted by the authors revealed that Learner 31's pronunciations were highly native-like and contained no noticeable segmental errors. However, the duration of each pronunciation was relatively long, which may have caused recognition difficulties for the ASR systems. Sunaoka (2018) notes that excessive pronunciation length can negatively affect recognition accuracy, potentially leading to insertions, omissions, and other errors. In this case, Gladia's assessment was the closest to the teacher evaluation, and the difference was not statistically significant based on the Wilcoxon signed-rank test.

5. Analysis of Pronunciation Characteristics

5.1 Rating Results of Six Pronunciation Categories

To address RQ2—whether these recognizers have the potential to identify pronunciation difficulties—we examined the ASR ratings from the perspective of pronunciation characteristics. The 20 target words were classified into six major categories based on the initial consonant of the first character: bilabial, apical, velar, palatal, retroflex, and alveolar. Because of the small sample size in each category, the following findings should be regarded as exploratory and interpreted as case-study evidence. Table 5 presents the mean ASR percentage difference scores for each category.

Table 5 ASR Percentage Difference Scores Based on Pronunciation Categories

Pronunciation Category	Whisper	Azure	Gladia
Bilabial	1.9%	-17.0%	5.0%
Apical	-1.8%	-10.7%	-7.7%
Velar	-9.4%	-23.3%	-12.6%
Palatal	17.6%	5.7%	17.6%
Alveolar	-34.2%	-53.0%	-17.1%
Retroflex	-9.8%	-18.0%	-23.3%

The results indicate that ASR ratings varied across pronunciation categories. All ASR systems tended to underestimate learner performance relative to teacher ratings, with the exception of palatal sounds. For apical sounds, the percentage differences across the systems were relatively small, whereas alveolar sounds exhibited much larger variation. The mean teacher score for apical sounds was1.69, while alveolar sounds received a lowerr mean score of1.17. These findings suggest that the degree of consistency between ASR assessments and teacher evaluations is strongly influenced by the specific pronunciation characteristics of each sound category.

5.2 Bilabial Sound Words

Table 6 ASR and Teacher Scores of Bilabial Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teacher
便宜	p bilabial/zero initial	ian/i	1.74	1.03	1.74	1.37
别走	b bilabial/z alveolar	ei/ou	1.48	1.45	1.55	1.64
北京	b bilabial/j palatal	ei/ing	1.87	1.65	1.87	1.75
买菜	m bilabial/c alveolar	ai/ai	1.74	0.97	1.74	1.72
放心	f labiodental/x palatal	ang/in	1.26	1.48	1.45	1.45

Table 6 presents the scores of the bilabial sound words assigned by the ASR systems and the teachers. The results of the Wilcoxon signed-rank tests indicate significant differences between teacher ratings and ASR scores for “便宜” (Whisper: $p = .003$; Gladia: $p = .045$), “北京” (Whisper: $p = .028$; Gladia: $p = .020$), and “买菜” (Azure: $p < .001$), whereas no significant differences were observed for “别走” or “放心.”

For “便宜,” which involves the aspirated consonant /p/ combined with the compound final /ian/, Whisper and Gladia assigned higher scores than the teachers. Azure’s ratings did not significantly differ from teacher assessments. Student pronunciation analysis conducted by the authors confirmed that while learners could distinguish between aspirated /p/ and unaspirated /b/, both the strength and duration of aspiration were insufficient, which may have contributed to the different recognition outcomes across systems.

For "买菜," Azure assigned an exceptionally low score, with recognition errors frequently occurring in the second syllable—for example, producing "在" or "开." Learner pronunciation checks revealed that although students' /c/ pronunciations were distinguishable, insufficient aspiration strength may have caused Azure's performance to diverge from that of the other two ASR systems.

Although Japanese lacks the /f/ sound and previous research by Fang et al. (2015) reported that /f/ is often confused with /h/ by Japanese learners, such confusion was not observed for "放心" in this study. A possible explanation is that the vowel /a/ exists in Japanese, and when combined with /f/, the overall phonetic structure becomes easier for learners to produce, reducing the likelihood of confusion.

5.3 Apical Sound Words

Table 7 ASR and Teacher Scores of Apical Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teacher
听懂	t apical/d apical	ing/ong	1.48	1.29	1.58	1.72
你家	n apical/j palatal	i/ia	1.74	1.65	1.26	1.68
老师	l apical/sh retroflex	ao/zero	1.74	1.58	1.84	1.68

There were three apical-initial words in the learner dataset, as shown in Table 7. The results of the Wilcoxon signed-rank tests indicate that a significant difference between ASR and teacher scores occurred only for "老师" (Gladia: $p = .023$). This suggests that Gladia may overestimate learners' productions of apical sounds, whereas the assessments provided by Whisper and Azure were consistent with teacher ratings.

5.4 Velar Sound Words

Although six words contained velar-initial sounds, the Wilcoxon signed-rank test results showed no significant differences between ASR and teacher scores for five of the items. The only exception was "客气", for which significant differences were found across all systems (Whisper: $p < .001$; Azure: $p < .001$; Gladia: $p = .003$).

Table 8 ASR and Teacher Scores of Velar Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teachers
客气	k velar/q palatal	e/i	0.13	0.16	0.52	1.06
告诉	g velar/s alveolar	ao/u	1.45	1.32	1.26	1.60
工作	g velar/z alveolar	ong/uo	1.94	1.48	1.94	1.79

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teachers
很远	h velar/zero initial	en/yuan	1.71	1.23	1.29	1.57
好吃	h velar/ch retroflex	ao/zero	1.64	1.55	1.39	1.52
咖啡	k velar/f bilabial	a/ei	1.90	1.87	1.74	1.84

The ratings for "客气" indicate that students' primary difficulties with velar sounds involve the /e/ final and aspiration control. Learner pronunciation checks showed that only a few of the 31 students produced "客气" with relative accuracy. The pronunciation errors observed fell into three main categories: (1) insufficient aspiration of /k/, (2) errors in the /e/ final, and (3) tone errors. ASR recognition outputs included forms such as "各级," "各起," and "课题," reflecting these deviations.

In contrast, ASR recognition of "咖啡," which shares the same initial /k/, exhibited high consistency with teacher ratings. This disparity suggests that ASR recognition errors for velar-initial words vary according to initial–final combinations. In other words, such contrasts may help elucidate specific pronunciation difficulties among learners.

5.5 Palatal Sound Word and Alveolar Sound Word

Table 9 ASR and Teacher Scores of Palatal Sound and Alveolar Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teacher
学校	x palatal/x palatal	ue/iao	1.87	1.68	1.87	1.59
词典	c alveolar/d apical	i/ian	0.77	0.55	0.97	1.17

Only one palatal-initial word and one alveolar-initial word were included in the learner data. The Wilcoxon signed-rank test results showed significant differences between teacher and ASR scores for "学校" (Whisper: $p = .018$; Gladia: $p = .010$) and for "词典" (Whisper: $p = .002$; Azure: $p < .001$).

For the palatal-initial word "学校," ASR systems tended to overestimate learner performance. Learner pronunciation checks indicated that students generally produced the palatal fricative /x/ accurately, whereas the final /u/ was often omitted or realized as "xie." These deviations may have been insufficiently penalized by the ASR systems, leading to higher scores than those assigned by teachers.

For the alveolar-initial word "词典," ASR systems tended to underestimate performance relative to teacher ratings. Recognition outputs frequently included "ji dian"–type errors (e.g., "寄典," "机点"), suggesting that learners often substituted the aspirated alveolar affricate /c/ with the unaspirated palatal affricate /j/. This substitution reflects insufficient aspiration in producing /c/, and ASR systems appeared highly sensitive to this cue, resulting in lower scores.

5.6 Retroflex Sound Words

Retroflex consonants constitute a phonetic category unique to Mandarin and are entirely absent from the Japanese phonological system; thus, they represent one of the greatest pronunciation challenges for Japanese learners. Both ASR and teacher ratings for retroflex-initial words tended to be low. The Wilcoxon signed-rank test showed a significant difference only for Gladia's score for "二十" ($p < .001$).

Learner pronunciation checks indicated that the primary difficulty with "常常" stemmed from the retroflex affricate /ch/. Many students failed to produce sufficient tongue-tip retroflexion, resulting in ASR outputs such as "江" and "强." Additionally, because "二十" contains the retroflex final /er/, this likely contributed to Gladia's substantially lower score relative to teacher ratings.

Table 10 ASR and Teacher Scores of Retroflex Sound Words

Vocabulary	Initial	Final	Whisper	Azure	Gladia	Teachers
常常	ch retroflex/ch retroflex	ang/ang	1.10	0.65	0.71	1.10
注意	zh retroflex/zero initial	u/i	1.16	0.84	1.61	1.21
二十	zero initial/sh retroflex	er/zero	1.10	1.55	0.23	1.41
日本	r retroflex/b bilabial	zero/en	1.45	1.32	1.55	1.59

Overall, the word-level analyses preliminary demonstrate that ASR assessments are strongly affected by specific pronunciation characteristics, and that initial–final combinations play a crucial role in determining the degree of alignment between ASR and teacher evaluations. Individual initials or finals may yield high recognition accuracy when paired with certain syllables, as in "咖啡," or low accuracy when combined differently, as in "客气." Such contrasts in acoustic features suggest that ASR outputs may help identify learner pronunciation difficulties based on systematic patterns across syllable structures.

6. Conclusion and Future Work

In this study, we examined how effectively three ASR systems—Whisper, Azure, and Gladia—evaluate Japanese learners' Mandarin word-level pronunciation by comparing ASR-generated scores with ratings provided by experienced teachers. Two research questions were addressed: (1) To what extent do ASR systems assess learner pronunciation in a manner consistent with teachers? (2) Do ASR systems have the potential to identify learner pronunciation difficulties? A scoring scheme was developed to quantitatively evaluate the utterances of 31 Japanese learners producing 20 Chinese words, and ASR–teacher comparisons were conducted from the perspectives of learner proficiency and pronunciation characteristics.

Analyses based on learner-level scores showed that although all three ASR systems tended to underestimate learner performance relative to teachers, Whisper provided the most teacher-like assessments overall. With regard to individual proficiency levels, Azure aligned most closely with teacher ratings for the lowest-proficiency learner, whereas Gladia showed the highest alignment for the highest-proficiency learner. These patterns suggest that ASR performance is influenced by learner proficiency. These results are partly consistent with those reported by Hirai and Kovalyova (2024), who also observed variability in ASR performance depending on phonetic features for English non-native speakers.

Word-level analyses preliminary revealed that ASR assessments were strongly affected by specific pronunciation features, particularly the interaction between initials and finals. Velar-initial words showed the highest overall consistency with teacher ratings, except for "客气," whereas alveolar-initial words exhibited the largest discrepancies. Single-syllable initials or finals may yield high or low recognition accuracy depending on their syllabic combination, as demonstrated by the contrast between "咖啡" and "客气." These contrasts indicate that ASR outputs may be used to identify pronunciation difficulties by analyzing systematic acoustic deviations.

The findings suggest that ASR systems have the potential to provide teacher-like assessments of learners' pronunciation. However, learner proficiency and pronunciation features should both be taken into account when implementing ASR systems for pronunciation learning and practice.

Additionally, the standard deviations among the five professional teachers ranged from 0.17 to 0.34, demonstrating that even trained raters exhibit individual differences when assessing learner pronunciation. This finding underscores the need for objective scoring methods and highlights the potential value of ASR systems as auxiliary tools in computer-assisted pronunciation training (CAPT). While ASR systems cannot fully replace human evaluators, they can provide consistent, phoneme-level assessment and reduce teacher workload, particularly in large-scale or formative assessment contexts.

This study has several limitations. Future research should expand the word set to encompass a wider range of phonetic features and include learners with more diverse proficiency levels. Further work should also examine ASR performance at the sentence and paragraph levels and incorporate additional scoring criteria to support the development of more comprehensive and objective assessment frameworks.

In summary, the findings highlight both the potential and the current limitations of applying ASR technologies to CAPT for Mandarin Chinese and provide insights for designing more effective pronunciation assessment and feedback systems.

References

Burleson, D. F. (2007). Improving intelligibility of non-native speech with computer-assisted phonological training. *Indiana University Linguistics Club Working Papers, 7*(1), 1-18. https://scholarworks.iu.edu/journals/index.php/iulcwp/article/view/25801

Chen, J. (2011). Application of VoiceThread in Chinese teaching and learning: Some examples. *Journal of Technology and Chinese Language Teaching, 2*, 81-94. [蘇芳儀. (2011). VoiceThread 應用於中文教學的幾個例子. *科技与中文教学, 2*(1), 81-94.] http://www.tclt.us/journal/2011v2n1/chenj.pdf

Da, J. (2015). The application of speech recognition technology in Chinese language learning: What can be learned from a Pinyin lab session. *Journal of Technology and Chinese Language Teaching, 6*(1), 16-24. [笪骏. (2015). 语音识别技术在中文教学中的应用：一堂汉语拼音练习课的启示. *科技与中文教学, 6*, 16-24.] http://www.tclt.us/journal/2015v6n1/da.pdf

Dai, Y. & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning, 36*, 861-884. https://doi.org/10.1080/09588221.2021.1952272

Derwing, T. M., Munro, M. J. & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly, 34(3)*, 592-603. https://doi.org/10.2307/3587748

Ehsani, F. &Knodt, E. (1998). Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm. *Language Learning & Technolog*y, *2*(1), 54-73. https://doi.org/10.64152/10125/25032

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*(10), 832–844. https://doi.org/10.1016/j.specom.2009.04.005

Fang, S., Chen, C., Wang, C., Yang, H., & Chen, H. (2015). An error analysis on Japanese learners' Chinese pronunciation with the aid of Chinese learners' oral corpus. *Journal of Chinese Language Teaching, 12*(3), 93-123. [方淑華, 陳慶華, 王敬淳, 楊惠媚, &陳浩然. (2015). 藉學習者口語語料庫探究日籍生常見的華語語音偏誤與教學建議. 華語文教學研究, *12*(3), 93-123.]

Fouz-González, J. (2015). Trends and directions in computer-assisted pronunciation training. In J. A. Mompean (Ed.), *Investigating English pronunciation: Trends and directions* (pp. 314–342). Palgrave Macmillan. https://doi.org/10.1057/9781137509437_14

Gao, F. (2025). A study on the learning motivations, goals, difficulties, and expectations of Japanese university students learning Chinese as a foreign language. *Journal of Aichi Shukutoku University*, *15*, 49-62. [高飛. (2025). 外国語として中国語を学習している日本人大学生の学習動機、目的、困難点及び期待. *愛知淑徳大学論集—交流文化学部篇— 第15 号*, 49-62.] https://aska-r.repo.nii.ac.jp/records/2000370

Gladia. (2024). *Gladia* [Audio transcription software]. https://www.gladia.io/

Hirai, A., & Kovalyova, A. (2024). Speech-to-text applications' accuracy in English language learners' speech transcription. *Language Learning & Technology, 28*(1), 1–21. https://doi.org/10.64152/10125/73555

Inceoglu, S., Chen, W. H. & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL, 35*(1), 89–104. https://doi.org/10.1017/S0958344022000192

Issa, E., & Hahn-Powell, G. (2025). Computer-assisted pronunciation training for foreign language learning of grammatical features. *Language Learning & Technology, 29*, 1–20. https://doi.org/10.64152/10125/73622

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Li, N., Zhang, L., Lau, K. L., & Liang, Y. (2024). Predicting Chinese language learners' ChatGPT acceptance in oral language practices: The role of learning motivation and willingness to communicate. *Journal of Technology and Chinese Language Teaching, 15*(1), 25-48. [李诺恩, 张岚, 刘洁玲, & 梁宇. (2024). 预测中文学习者在口语练习中对ChatGPT的接受度：学习动机和交流意愿的作用. *科技与中文教学, 15*, 25-48.] http://www.tclt.us/journal/2024v15n1/lizhanglauliang.pdf

Lin, C. (2019). The phonetic problems of Japanese Chinese language learners and teaching suggestions. *TCSL Forum,* 27, 9-37. [林嘉惠. (2019). 日籍華語學習者的語音問題與其教學建議. *華語學刊*, 27, 9-37.] https://www.airitilibrary.com/Article/Detail/P20151202001-201912-202005290011-202005290011-9-37

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc

McCrocklin, S., Humaidan, A., & Edalatishams, E. (2019). ASR dictation program accuracy: Have current programs improved? *In J. Levis, C. Nagle, & E. Todey (Eds.), Proceedings of the 10th pronunciation in second language learning and teaching conference* (pp. 191–200), Iowa State University. https://www.iastatedigitalpress.com/psllt/article/id/15376/

Microsoft. (2024). *Azure AI Speech* [Audio transcription software]. https://azure.microsoft.com/products/ai-services/ai-speech

Munro, M. J. & Derwing, T. M. (1999) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*(s1), 285-310. https://doi.org/10.1111/0023-8333.49.s1.8

Neri, A., Cucchiarini, C. & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL, 20*(2), 225-243. https://doi.org/10.1017/S0958344008000724

O'Brien, M. G, Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I, Strik, H., Levis, J. M., Munro, M. J., Foote, J. A. &Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation, 4*(2), 182 – 207. https://doi.org/10.1075/jslp.17001.obr

Sunaoka, K. (2018). Using automatic speech recognition technology to reverse analyze communication strategies between non-native speakers in a Chinese long distance group discussion. *Journal of Technology and Chinese Language Teaching, 9*(2), 61-82. [砂冈和子. (2018). 以语音识别技术逆向分析汉语远场群体讨论中非母语者的交互策略. 科技与中文教学, *9*(2), 61-82.] http://www.tclt.us/journal/2018v9n2/sunaoka.pdf

Tsai, P. (2019). Beyond self-directed computer-assisted pronunciation learning: A qualitative investigation of a collaborative approach. *Computer Assisted*

Language Learning, 32(7), 713-744. https://doi.org/10.1080/09588221.2019.1614069

Wang, Y. & Young, S. S. (2014). A study of the design and implementation of the ASR-based iCASL system with corrective feedback to facilitate English learning. *Educational Technology & Society, 17*(2), 219-233. https://www.jstor.org/stable/jeductechsoci.17.2.219

Watanabe, Y., Omae, T. & Odo, S. (2019). Investigating the effect of chinese pronunciation teaching materials using speech recognition and synthesis functions. *Journal of Technology and Chinese Language Teaching*, *10*(2), 102-124. http://www.tclt.us/journal/2019v10n2/watanabeomaeodo.pdf

Wei, W. & Zhang, J. (2018). An intelligent Chinese pronunciation teaching app and the preliminary result of a teaching experiment. *Journal of Technology and Chinese Language Teaching*, *9*(2), 83-97. [魏巍 & 张劲松. (2018). 一款汉语智能语音教学App及教学实验初步结果. *科技与中文教学, 9*(2), 83-97.] http://www.tclt.us/journal/2018v9n2/weizhang.pdf

Whisper Desktop. (2023). *Whisper* (small version 17.4.4.) [Audio transcription software]. https://github.com/const-me/whisper

Zhao, X., Tomita, N., Konno, F., Ohkawa, Y. & Mitsuishi, T. (2019). Development and practice of review material KoToToMo for use on smartphones in blended learning by beginning learners of Chinese in university. *Transactions of Japanese Society for Information and Systems in Education, 36*, 131-142. [趙秀敏，冨田昇，今野　文子, 大河雄一 & 三石大. (2019). 大学初修中国語ブレンディッドラーニングのためのスマートフォン利用復習教材「KoToToMo」の開発と実践. *教育システム情報学会誌. 36*, 131-142.] https://www.jstage.jst.go.jp/article/jsise/36/2/36_360211/_article/-char/ja/

生成式人工智能辅助中文教学视频开发的实践探索：教师—人工智能协同模式的构建与评价

(Exploring Generative AI-Assisted Development of Chinese Language Instructional Videos: Construction and Evaluation of a Teacher–AI Collaborative Model)

方淳
(Fang, Chun)
北京语言大学
(Beijing Language and Culture University)
fangchun_blcu@163.com

马瑞祾
(Ma, Ruiling)
云南师范大学
(Yunnan Normal University)
maruiling_blcu@163.com

摘要：在中文作为第二语言教学中，视频是最常用的多模态资源之一，教学需求量大，但却面临着制作成本较高、耗时费力等困境。本研究旨在探讨生成式人工智能（GAI）在中文二语教学视频制作中的实践路径，以词汇为例进行生成探究与应用测评，构建并评估了一个 GAI 赋能中文教学视频生成的完成流程；并邀请了中文教师与学习者对 GAI 生成教学视频进行评价。结果显示，GAI 显著降低了视频制作的门槛与时间成本，生成内容基本可用，但在语言内容的解释深度和例句择取等方面仍有不足。最后，本文提出探讨了 GAI 赋能中文教学视频制作的三条启示，以期推动 GAI 更有效地赋能国际中文教学。

Abstract: Video-based instruction is one of the most widely used forms of multimodal resources in Chinese as a second language (CSL) teaching. Despite the high demand for instructional videos, their production is often constrained by substantial time, labor, and financial costs. This study explores the practical application of generative artificial intelligence (GAI) in the development of CSL instructional videos. Using vocabulary instruction as an illustrative case, the study investigates the process of AI-assisted video generation and evaluates its pedagogical applicability. A complete workflow for GAI-enabled Chinese language instructional video production was developed and assessed, and both Chinese language teachers and learners were invited to evaluate the AI-generated videos. The results indicate that GAI can substantially reduce the technical barriers, time investment, and production costs associated with video creation. While the generated videos were generally usable for instructional purposes, limitations remained in areas such as the depth of linguistic explanations and the selection of pedagogically appropriate example sentences. Based on these findings, the study proposes three implications for the use of GAI in Chinese language instructional video development, with the aim of promoting more effective integration of GAI into international Chinese language education.

关键词：生成式人工智能；中文二语教学；教学；视频

Key words: Generative artificial intelligence; International Chinese education; Teaching; Video

1. 引言

在第二语言教学中，教学视频是学习者进行课前预习、课堂辅助与课后复习的重要资源。视频材料被视为促进可理解性输入与情境建构的有效媒介，视觉与语言信息的整合能够提升学习效果（Baek et al., 2026）。在中文作为第二语言（中文二语，Chinese as a Second Language, CSL）等二语教学中，视频能够提供直观、生动且情境化的语言输入，是促进语言习得与文化理解的重要工具。二语教学中常用的视频可分为两类：一类是真实视频，多来源于影视作品、纪录片或社交媒体，教师从中截取可用于教学的片段。此类视频的优势在于提供真实的交际语境和丰富的文化信息，有助于学习者沉浸式地感知语言（Huang, 2025）；另一类是教学视频，由教师为讲解特定语言点或文化知识而专门设计制作，多用于课堂教学或课后复习。此类视频以教学为导向，目标清晰，通常包含“知识讲解—操练—检测”等环节，是课堂教学中较为重要的多模态支架材料。

尽管教学视频在教学中具有重要价值，其应用却受到诸多限制。真实视频的内容往往与教学重点不直接相关，视频中的词汇难度、语速、口音及背景噪声等因素有可能会增加学习者的理解负荷。教师若要在课堂中使用适配的视频，往往需要从大量资源中筛选、剪辑并加以教学化改造，不仅耗时费力，还会受到版权等因素制约。教学视频的制作同样面临类似困境：一个精心制作的教学视频需要经历脚本撰写、素材搜集、拍摄录制、后期剪辑等多个环节，对多数教师而言，时间与技术门槛相对过高。因此，在中文二语等二语教学中，教学视频的供需矛盾突出，针对特定教学重难点的有效视频资源常常供给不足。

生成式人工智能（Generative Artificial Intelligence, GAI）的发展为教学视频的开发提供了一个新的技术路径。GAI 具备多模态内容生成与逻辑推理能力，能够快速生成文本、图像与视频等素材，展现出高效、低成本、可扩展的潜力（Fernandez-Espinosa et al., 2025）。已有研究探讨了 GAI 在中文二语教学资源生成中的应用，如分级阅读文本生成（韩欣欣等，2025）、教案协同生成（丁安琪 & 蒙小凤，2025）、个性化阅读材料生成（侯泽煜 & 徐娟，2025）以及 AI 微短剧制作（丰迪等，2026）。然而，据作者所知，关注 GAI 用于中文二语教学视频制作的研究还较少。同时，GAI 生成内容在语言准确性、教学逻辑与文化适切性等方面的教学效度仍有待验证（Ouyang et al., 2022），这也是教学中采用 GAI 生成视频的主要障碍。

本文以词汇教学视频为切入点，探究 GAI 在中文二语教学视频制作中的可行性与应用边界，并在此基础上构建可复制、可推广的 GAI 辅助教学视频生成流程，以期为中文二语教学资源的智能化开发提供实践参考。

2. 视频在中文二语教学中的应用与研究

相较于教学文本和静态图片，视频能够提供近乎真实的交际语境，并且通过协调视觉与听觉信息降低语言和知识理解难度，激发学习者的兴趣与动机，在中文二语教学中是丰富语言输入的材料和辅助教学的工具。

2.1 视频在中文二语教学中的应用范围

教学视频凭借其多模态（视觉与听觉）的呈现方式，能够为第二语言学习，特别是中文二语教学创设丰富的语言知识输入语境。

2.1.1 语言知识教学

教学视频能够将抽象的语言知识置于具体、动态的语境中，帮助学习者在理解语言形式的同时掌握其使用场景，从而提升教学效果。在中文二语教学中，视频在语音、汉字、词汇、语法、语用五个方面的应用各有侧重。

（1）语音教学

语音教学通常以教师发音示范、学习者模仿开始，在课堂中教师的发音示范往往是一瞬间的，学生难以充分观察和模仿，通常需要课后借助音视频等进一步学习操练。相较于单一的音频资源，视频的优势不仅在于可反复播放，还可将发音方法、发音部位等语音知识可视化表征出来，促进认知加工。例如，在区分[i]和[ü]时，两者的区别在于唇形的圆展，视频可以通过特写镜头持续展示唇形变化，便于学生观察和跟读。再如，送气音与不送气音的教学中，视频可以通过纸片浮动的方式直观呈现气流强弱，帮助学生感知发音特征，这类演示在课堂上往往（因教学时间的限制）难以重复呈现。因此，视频更适合作为语音教学的辅助材料，配合教师的讲解与示范，能够达到更好的教学效果。

（2）汉字教学

汉字的字形结构复杂，笔画顺序和部件组合是中文教学的难点。视频可以通过动态演示逐笔展示汉字的书写过程，清晰呈现笔顺、笔画走向及部件之间的位置关系。例如，在讲解“妈”字时，视频可以先展示“女”字旁，再逐步添加“马”部，帮助学习者理解形声字的构成逻辑。相比于静态的图片或板书，视频能够将汉字的书写过程动态化、可视化，便于学习者建立字形认知。

（3）词汇教学

视频能够提供比图片和文字更丰富的词汇学习情境。以“购物”主题为例，教师可以展示一段超市购物的视频，将“收银台”“购物车”“打折”等词汇与具体场景对应起来，帮助学习者在情境中理解词义并掌握搭配与使用环境。对于“爱”“喜欢”“恨”等抽象词汇，静态的文本或图片释义往往不够直观，而视频可以通过多模态的方式呈现这些词语的使用情境，帮助学习者在真实交际语境中理解其含义。

（4）语法教学

视频适用于讲解复杂的语法结构，如某些特殊句式和补语。传统语法教学多依赖文字例句，学习者通过例句归纳句式，但难以快速理解其语用功能；且教师也常借助实物和动作演示来导入语法点。但动作演示的适用范围有限，一些抽象或不便在课堂中实际操作的情境难以立刻呈现，如教授“上来”“下去”等趋向补语时，往往力有不逮。相比较而言，视频能够在保留动作演示直观性的基础上，克服上述局限。教学视频还能够呈现超越课堂实物范围的场景，如虚拟人物在餐厅、超市等真实交际情境中使用“把字句”完成特定动作。同时，视频可以配合字幕、提示和回放，便于学习者反复观看和理解。

（5）语用教学

在中文中，“道歉”“感谢”“拒绝”等言语行为在不同语境下有不同的表达方式。仅依靠对话文本，学习者难以把握其中所蕴含的语气差异。视频可以模拟真实的社交场景，呈现不同场合下的语言使用方式。例如，可以通过视频对比正式场合与非正式场合中“请求帮助”的不同表达，帮助学习者培养语用得体性。此外，对于“好的”“好吗”“好呀”“好吧”“好呢”等语气词，视频可以通过具体情境展现不同选项所传递的语气差异，帮助学习者在实际交际中做出恰当选择。

2.1.2 文化知识教学

语言是文化的载体。由于文化具有较强的依附性，文化教学需要借助实物、图像或视频等具体媒介，使教学内容更加直观、可感（李泉 & 孙莹，2023）。视频能够将文化内容与语言知识融为一体，是实现这一目标的理想工具。视频在文化教学中的应用主要体现在以下两个方面。

第一，呈现多元文化主题。教师可以制作涵盖传统手工艺、中华美食、传统节日、当代社会生活、流行文化及科技发展等内容的视频，用于课堂教学。例如，一段关于春节的简短视频，不仅可以帮助学习者掌握“春联”“年夜饭”“拜年”等词汇，还能使其直观感受节日的氛围与文化内涵。

第二，进行文化对比与内涵解读。视频能够直观展示中外文化的异同，引导学

习者进行文化对比，增进对文化多样性的理解。例如，关于中国餐桌礼仪的视频可以让学习者对比本国的餐饮习惯，深化对文化差异的认识。对于一些抽象的文化概念，如“面子”“关系”等，可以通过剧情短片展现具体的人物互动与故事情节，帮助学习者理解这些概念在实际交际中的含义与功能。

2.2 视频在中文教学中的应用现状

早期，教师多选用课堂实录或影视素材进行中文二语教学。随着自媒体资源的丰富，视频素材的来源进一步拓展。近年来，GAI 的发展革新了教学视频的生产方式，使视频内容可以实现自动化生成与智能化呈现，既能够根据教学需求定制内容，也降低了教师制作视频的技术门槛，使不具备视频制作技能的教师也能够便捷地开发教学视频。

关于视频应用于中文二语教学的现有研究主要从两个方向展开。一方面，学者们探讨了视频在中文教学中的应用，并结合教学实践验证了视频应用于语言教学的实际效果，研究涉及短视频用于课堂教学的机遇与困境（苏放 & 温向明，2025）、微视频教学短片的教学价值（周洋，2016）、交互式微视频资源在课堂中的建设与应用（张屹等，2013；范福兰等，2012）、利用新闻视频提升学习者的学业表现（Wen et al., 2021）、运用幽默视频与回声法促进词汇教学（Do et al., 2022）、关注越南学习者使用 YouTube 辅助中文学习（Wang et al., 2022），以及视频讲座作为第二语言学习方式的可行性（Zhang et al., 2023）；另一方面，也有学者对视频资源在教学中的定位进行了分类，将其视为数字化教学的重要资源形态（王洪梅等，2017），或国际中文智慧教育的软件微形态之一（马瑞祾 & 徐娟，2025）。总体上看，现有研究多关注视频在教学中的使用效果，如多媒体学习认知理论（Cognitive Theory of Multimedia Learning, CTML）的应用，或视频对学习者注意力与认知负荷的影响，而较少系统探讨教学视频的开发模式。换言之，“如何有效使用视频”的研究相对集中，但“如何高效开发优质教学视频”这一前置性问题尚未得到充分关注。

为解决上述问题，已有部分研究开始探讨了 GAI 在视频生成中的应用，如基于生成对抗网络（Generative Adversarial Network, GAN）技术生成手语视频（Sreemathy et al., 2024），以唐诗情境视频为例探索 AI 在教学多模态资源生成中的潜力（Chen & Wu, 2024），以及对 GAI 影响视频生产（姜博，2025）和短视频文本生成价值（李雨昊 & 高迎刚，2024）的讨论。这些研究表明，人工智能技术正在逐步渗透教学资源的开发领域。但基于 GAI 的中文二语教学视频制作研究尚处在起步阶段，仍需深入探索。

总之，教学视频的应用研究已较为充分，但将 GAI 用于其制作的研究仍待探索和深化。技术上的可行性并不等同于教学上的有效性，GAI 生成的教学视频是否符合中文二语教学的实际需求，也有待验证。因此，本研究以中文教学视频开发为对象，通过实际制作流程考察 GAI 生成视频的可行性，为中文二语教学资源的智能化建设提供实证依据。

3. 研究设计

3.1 研究问题

GAI 的发展为中文二语教学资源的建设提供了新的技术路径。与传统教学视频制作中技术门槛高、开发周期长等问题相比，GAI 能够实现分镜脚本、图像与视频的自动化生成，在效率上具有明显优势。然而，技术上的可行性并不等同于教学中的适用性。中文二语教学视频还需保证语言知识的准确性和文化内容的恰当性，这是 GAI 生成内容能否真正服务于教学的关键。基于此，本研究以词汇教学视频为案例，通过设计提示词生成视频脚本，并进一步生成教学视频，考察 GAI 在中文二语教学视频制作中的应用效果与实践边界。具体包括以下三个研究问题：

RQ1：中文教师如何与 GAI 协同开发适用于中文教学的教学视频？

RQ2：中文教师与学习者如何评价 GAI 生成的中文教学视频的整体质量？

RQ3：生成式人工智能辅助中文教学视频开发过程中面临哪些问题与挑战？

3.2 研究方法

本研究旨在展示一个由 GAI 赋能的中文教学视频开发流程，并邀请中文教师与学习者对选取教学视频示例进行质量评估。本研究拟探索出人智协同、动态优化的中文教学视频资源创制模式，并据此开展实践。

“人智协同”是指教师与 GAI 共同合作完成教学视频的制作。在该过程中，教师需要发挥自己的专业素养，对画面脚本进行审核，比如先出现对应的生词，再讲解释义，最后进行练习，这需要结合教学经验来抉择。而 GAI 主要负责完成效率型的任务，比如快速生成画面脚本，按照画面脚本提示词生成视频内容，辅助教师快速开发教学视频。整个过程中教师需占主导地位，GAI 执行教师发出的指令，生成对应的画面脚本和视频。

“动态优化”是指在教学视频开发过程中，教师依据教学目标与学习者需求，根据实际情况调整提示词，进而改变 GAI 生成内容。例如针对不同国别和语别的学习者，容易出现词语混淆使用的情况，此时，教师可优化提示词，生成画面脚本，并加入对应词语混淆的讲解。GAI 生成内容高度依赖提示词，面对生成内容存在教学重点不突出等问题的时候，通过提供多种类型的提示词，不断调整角色设定、任务要求、输出结构和质量约束，通过优化提示词提升生成内容质量。除此之外，GAI 生成的脚本和视频是否可以最终在教学中使用，还是待加工的教学半成品，需要教师结合学科知识进行审核与修正。例如，教师可以对例句进行修改，使其更符合真实交际场景，增强视频的教学针对性。

3.3 案例选择依据

词汇是语言学习的基础，也是中文二语教学中的重点和难点（李如龙 & 吴茗，2005）。学习者主要通过有意学习和语境中无意习得两种途径掌握词汇（江新，1998），且大部分词汇来源于语境中的自然接触（Nagy et al., 1987）。因此，借助视频提供的语境来促进词汇学习具有明显优势（Yanto & Elih Sutisna, 2018；Minalla & Amir Abdalla, 2024）。本研究选择中文二语教学中的离合词（Separable Verbs）为例，具体选取“见面”“帮忙”“睡觉”等高频常用词。选择理由如下：首先，所选的离合词均为学习者在日常语言交际中高频使用的词语；其次，这些词可代表离合词不同维度的教学难点与常见偏误，“见面”体现了离合词不能带宾语的要求，学生常犯“见面我的朋友”的错误，“帮忙”展示了离合词可扩展但扩展成分受限的特性，可以说“帮他的忙”，但“帮忙”本身不能带宾语，只能以“离”的形式引出关涉对象，“睡觉”作为日常词汇，常用程度较高，便于创设生活化情境（王瑞敏，2005）；最后，这三个词的讲解难度适中，能有效检验 GAI 生成教学脚本质量。

此外，选择视频作为离合词教学的呈现载体具有明显优势。离合词兼具词汇与短语双重属性，其使用过程往往涉及离析、扩展、搭配等动态变化，仅依靠静态文字说明或纯语音讲解，学习者难以直观理解其具体用法与交际功能。离合词在交际中常依附于特定的情境，其语义的理解与使用规则都依赖具体语境的支撑。教学视频能够通过人物动作、场景转换和情境对话等多模态信息，将离合词的使用过程动态呈现出来。教师利用视频将抽象的语言规则转化为可视化、情境化的教学内容，使学习者在真实或拟真的交际情境中理解离合词的结构特点与语用功能，视频是更直观且易于理解的语言输入的支架。

3.4 模型选择

生成中文教学视频及内含的各类图片都需设计相应的脚本，而 GAI 可帮助教师形成脚本底稿。脚本底稿的质量又与基座模型的能力密切相关，故教师需选用合适的模型。本研究参考中文语言理解测评基准“SuperCLUE”榜单[1]进行模型选择，该榜单 2026 年 5 月测评结果如图 1 所示。

[1] SuperCLUE 榜单 5 月测评网址参见：https://www.superclueai.com/generalpage。

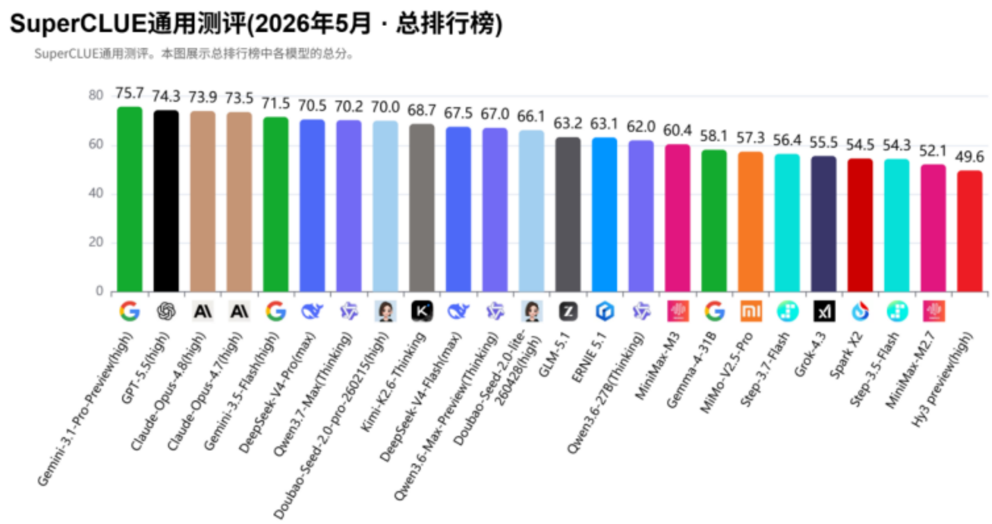


图 1 SuperCLUE 模型 2026 年 5 月测评排行榜

在 5 月份的测评榜单中，列举模型如表 1 所示，由表 1 可知，“DeepSeek-V4-Pro”位居开源模型的榜首，且“幻觉控制”“科学推理”“精确指令遵循能力”得分均在前列，因此本研究选择该模型作为视频分镜脚本生成模型。

表 1　模型 5 月测评榜单

模型名称	发布机构	是否推理	是否开源	属地
Gemini-3.1-Pro-Preview(high)	Google	是	否	海外
GPT-5.5(high)	OpenAI	是	否	海外
Claude-Opus-4.8(high)	Anthropic	是	否	海外
Claude-Opus-4.7(high)	Anthropic	是	否	海外
Gemini-3.5-Flash(high)	Google	是	否	海外
DeepSeek-V4-Pro(max)	深度求索	是	是	国内
Qwen3.7-Max(Thinking)	阿里巴巴	是	否	国内
Doubao-Seed-2.0-pro-260215(high)	字节跳动	是	否	国内
Kimi-K2.6-Thinking	月之暗面	是	是	国内

这里需要提及的是，教师在选择模型生成脚本时推荐参考相关的模型测评榜单，一般可以选择开源模型与具备推理能力的模型，这样既能节约生成脚本的经济成本，也能提高生成脚本的质量。当然，教师也可以不考虑经济成本，以模型能力作为唯一的选择依据，同样只需要将提示词输入模型，等待输出结果即可。

3.5 提示词设计

与大模型对话时，为了让大模型准确理解用户（可以是教师，也可是学习者）

的意图与需求，尽可能地生成符合教学需求的素材，需要使用规范化、结构化的提示词明确生成任务，充分发挥大模型的内容生成能力。提示工程（Prompt Engineering）是通过设计、实验和优化输入来引导模型生成高质量、准确和有针对性的输出内容（Dong et al., 2024）。无论是脚本、图片还是视频，都需要选择合适的提示词框架，提示词框架就如同用户与 GAI 沟通的对话指南，帮助用户把复杂的需求拆解成 GAI 能够理解的结构化指令，即通过明确角色设定、任务目标、结构要求与质量约束四个要素，增强 GAI 生成内容的教学适配性，这样既可以保障语言准确性又可以兼顾文化敏感性，使技术工具真正成为教师专业能力的延伸而非替代。表 2 展示了一个中文教学视频脚本生成的常用提示词框架。

表 2 中文教学视频脚本生成常用提示词框架

组成	提示词示例
角色	你是一位经验丰富的国际中文教师，精通词汇/语法教学，擅长脚本编写。
任务	请为{hsk_level}的中文学习者编写一个关于{Content}的词汇/语法教学视频的脚本，时长为{minutes}。
内容	生成脚本需要包括以下 3 部分：（1）内容讲解；（2）偏误类型；（3）典型例句。
质量	脚本所用词语需要确保例句语言符合日常表达，文化得体。

需要说明的是，仅仅使用一次性整体输入的提示词往往难以覆盖所有的教学细节，因此，教师可以使用提示词框架进行多轮交互对话，生成多个内容以供参考与备用，也可以进行整合，尽可能实现教学内容的覆盖，减轻教师编写的负担。

3.6 评价方法

为评价 GAI 生成的教学视频的质量，本研究参考 Xu 等（2025）和欧志刚等（2024）开发了一张《GAI 生成中文二语教学视频评估量表》，该量表包含三个维度、九个观测点（如表 3 所示）。研究采用混合评估方式，面向中文教师与中文学习者发放问卷，收集他们对视频的评估意见。该量表采用 Likert5 度量表（1=非常不同意，2=比较不同意，3=既不同意也不反对，4=比较同意，5=非常同意）进行量化评分，并设有“总体评价”的开放式问题，用于收集质性反馈，后期结合师生反馈优化提示词。

表 3 GAI 生成中文二语教学视频效度评估量表

维度	指标	描述语
设计	激励性	生成资源能提升学习者的兴趣
	清晰性	生成资源在视听方面的清晰度
	可用性	生成教学内容易于理解与使用
内容	准确性	生成教学资源没有错误与歧义
	充足性	生成资源能总体反映教学内容
	恰当性	生成资源能被师生理解与接受
技术	流畅性	生成视频画面清晰、流畅，运行时不存在卡顿现象
	同步性	音频与画面在时间上精确对齐，口型与发音匹配
	生动性	使用数字人形象及资源生动形象、自然

4. GAI 生成流程

秉持“人智协同”与“动态优化”的研发思路，本文将中文教学视频的生成过程划分为“确定教学内容”“生成分镜脚本”“生成图像与视频”三个阶段，并以离合词为例，具体呈现各阶段的生成流程。

4.1 确定教学内容

在二语教学中，教学内容的设计是关键的一环。教师如何讲解、讲解的深度如何，这些都需要提前规划。不同国别和语别的学习者在学习同一知识点时，会出现不同类型的偏误。以离合词为例，常见的偏误包括“插入成分偏误”“重叠形式偏误”“宾语使用偏误”三类（王瑞敏，2005）。偏误讲解是教学内容的重点，教师需明确指出存在的偏误类型。

然而，据笔者观察，GAI 虽能大致分析并指出教学内容中可能存在的偏误类型，但时常会出现错误。例如，在离合词脚本生成中，如果教师输入的提示词为

> 你是一名国际中文教师，擅长脚本编写，结合留学生在学习离合词时容易出现的偏误类型及其解决策略，依托学生一天的学习生活以‘睡觉、见面、聊天、唱歌、跑步、请客’这几个离合词为内容设计一份微课脚本，内容之间要有连贯性，时间 5～7 分钟，转换成分镜脚本，以表格的形式呈现。

通过诊断 GAI 生成微课脚本的思维链（见图 2）发现，由于未在提示词中明确离合词的具体偏误类型，生成的内容存在两处不当之处：（1）混淆了介宾短语的宾语与动宾式离合词的宾语，将“宾语使用不当”表述为“介词缺失”，这一表述不恰当，可能误导学习者；（2）“重叠形式的偏误”提供了正确示例，而另外三类偏误提供了错误示例，缺乏教学一致性，不利于学习者有效学习。

因此，教师应将教学内容的核心知识作为预设内容放入提示词，要求 GAI 严格按照此内容输出，以避免知识性错误。具体操作上，教师可先依据《国际中文教育中文水平等级标准》[2]确定教学内容，然后在“HSK 动态作文语料库”[3]与“全球中介语语料库”[4]中，对教学内容的偏误类型进行检索与分析（具体使用说明可参看语料库使用说明），对提示词进行补充。更为简洁有效的方式是直接使用对应的工具书，如《国际中文教育中文水平等级标准语法教学手册》[5]《国际中文教育中文水平等级标准·词汇速记速练手册》[6]等，这些工具书依据《国际中文教育中文水平等级标准》中的词汇和语等内容编写而成，能较为全面的覆盖教学中可能遇到的词汇知识与语法知识。因此，教师为避免检索语料库查询偏误这一耗时费力的工作，可以选择参考成熟的教学资料，并按照手册中给出的思路讲解，提高效率。

首先，我得确定用户的核心需求。他们希望微课内容有连贯性，并且以学生的日常活动为背景。这可能意味着需要将每个离合词自然地嵌入到一天的不同时间段，比如早上跑步、上课见面、午休聊天、下午唱歌、晚上请客、睡觉等。这样的安排不仅连贯，还能帮助学生将学习内容与实际生活联系起来，提高记忆效果。

接下来，我需要考虑学生常见的偏误类型。比如，离合词后误加宾语（如“见面朋友”），介词缺失（如“聊天老师”），补语位置错误（如“跑步了半小时”），以及重叠形式错误（如“唱唱歌”）。解决策略包括对比正确和错误用法，视觉化结构，互动练习等。在分镜设计中，需要将这些策略融入各个场景，确保每个偏误都有对应的教学点。

然后，我需要构思如何将一天的时间轴与离合词结合起来。比如，早晨起床、跑步，上午上课见面，午休聊天，下午活动唱歌，晚上请客，最后睡觉。每个场景对应一个离合词，并展示正确和错误的用法，通过连贯的故事线让学生跟随主角的一天，逐步学习各个离合词的正确用法。

分镜脚本方面，需要保持每个镜头的连贯性和逻辑性。从早晨开始，展示主角的活动，遇到问题，学习正确用法，练习，最后总结。每个镜头的时间控制在合理范围内，总时长5-7分钟。可能需要调整之前的表格结构，确保每个步骤都有明确的画面描述、配音字幕、时长和备注。

此外，用户可能希望脚本不仅展示正确用法，还要有互动环节，比如填空或判断对错，以增强参与感。同时，作业部分需要结合实际应用，比如写日记，巩固学习内容。需要检查之前的亮点是否依然适用，比如偏误针对性、认知策略、记忆强化和文化融入，确保新脚本同样具备这些优点。

最后，要确保所有离合词（睡觉、见面、聊天、唱歌、跑步、请客）都被涵盖，每个词的教学点清晰，并且符合一天的时间顺序。可能需要调整部分离合词的出现顺序，使其更符合日常活动的逻辑。例如，跑步通常在早晨，见面在上午上课时，聊天在午休，唱歌在下午活动，请客在晚餐，睡觉在晚上结束。

总结下来，我需要将离合词融入学生一天的生活场景，每个场景对应一个离合词，通过故事线连接，设计互动和反馈机制，确保教学策略有效，并且保持微课的连贯性和教育性。

图 2 DeepSeek 生成离合词微课脚本思维链

4.2 生成分镜脚本

在利用 GAI 生成教学视频的过程中，生成内容主要依赖脚本提示词。大模型根据脚本提示词可先生成图片再生成视频，也可直接生成视频。因此，脚本生成是关

[2] 《国际中文教育中文水平等级标准》下载地址参见：http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202103/t20210329_523304.html。

[3] HSK 动态作文语料库网址参见：hsk.blcu.edu.cn/Login。

[4] 全球中介语语料库网址参见：https://qqk.blcu.edu.cn/#/login。

5 参见：https://www.blcup.com/PInfo/index/11228。

6 参见：https://www.blcup.com/PInfo/index/12574

键环节，而脚本质量又取决于提示词的设计。已有研究表明，CRISPE 框架在国际中文教学资源生成中表现最优（史金生 & 葛星辰，2025）。

CRISPE 框架包含五个组成部分。CR（Capacity and Role）指“能力与角色”，即明确告知 GAI 应扮演的身份。角色越具体、与任务越相关，GAI 生成的内容就越专业、也越符合预期。I（Insight）指“背景洞察”，提供背景信息和上下文，帮助 GAI 理解任务目的，一般包括受众群体、目标预期、约束条件和背景信息，设计提示词时需要明确这些内容。S（Statement）指“任务陈述”，即对生成任务进行详细说明，让 GAI 理清逻辑和先后顺序，避免遗漏，也便于检查生成过程。P（Personality）指“个性风格”，教师可根据使用场景和受众群体自定义输出风格，如“轻松愉快型”或“简洁明了型”，使 GAI 以预期的方式回答。E（Experiment）指“实验”，核心在于要求 GAI 提供多个答案，避免思维定势，同时为教师提供更多选择空间。本研究以 CRISPE 框架为参考，设计了中文二语教学视频脚本生成的提示词框架，具体见表 4。

表 4　中文二语教学视频脚本生成提示词框架

组成部分	示例内容
能力与角色（Capacity & Role）	你是一位[角色身份，如：有 20 年经验的儿童汉语启蒙教师/商务汉语培训师/HSK 备考专家]的资深词汇/语法教学专家，尤其擅长教学法，例如：全身反应法教学/任务型教学/汉字故事化讲解]。
背景洞察（Insight）	教学对象是[母语、年龄、水平，如：英语母语、6-8 岁、零基础]的中文学习者，教学目标是掌握[词汇/语法，例：离合词、趋向补语]的内容，视频时长约[X]分钟，其常见偏误是{偏误内容}，需要对每类偏误进行针对性讲解。 其他约束[不使用学生母语、禁止出现拼音/需同时展示拼音等]
任务描述（Statement）	请生成一份完整的词汇/语法教学视频分镜脚本，需要按照“导入—偏误示例—内容讲解—练习—总结”的结构模块讲解词汇/语法，每个部分需要给出教师用语（中文）、配图、效果、字幕、时长（秒），练习需要预留学习者思考时间。
个性风格（Personality）	整体视频风格为[风格类型，例：亲切活泼/简洁清晰/幽默轻松]，教师语速[偏慢/适中/偏快]，语气[例：充满鼓励、富有节奏感、自然交谈]，教学中多使用[例：夸张的表情提示、手势指令、重复强化]。
多样化实验（Experiment）	请生成[X]个不同构思的[需要多样化的部分，如：导入设计/操练形式/结尾]脚本，并简要说明每个脚本的适用场景和优缺点。

本研究选用 DeepSeek-V4-Pro[7]的专家模式生成离合词的分镜脚本，并对生成的分镜内容进行细致修改。分镜脚本的具体内容见附录 1。需要说明的是，教师需结合自己的教学经验对 GAI 生成的脚本加以审核和修订，不宜直接使用 GAI 生成的脚本，脚本也可作为后续生成视频的文本材料。在中文二语教学中，技术主要发挥减负增效的工具价值，中文教师依旧占据主导地位，所有的教学设计与活动都要围绕学习者展开。

4.3 生成图像与视频

教师使用 GAI 生成分镜脚本，经修改后形成教学材料。随后，教师可利用脚本中的提示词生成对应的图片或视频素材。图像与视频的生成方式与脚本生成类似，均通过设计提示词实现，GAI 根据用户需求和描述输出相应素材。相比文本生成，图片生成更注重图像内容、质量、风格与背景等要素，具体要求如表 5 所示。

表 5 图像生成要素特征

元素	要求
内容	明确需要的素材对象
环境	明确主体背景或活动场景
风格	明确图片风格，如卡通、古风、写实与清新等
色彩	明确生成的素材为哪种颜色
视角	明确画面布局添加远景、近景等指令
比例	依据需求指定图像的大小和比例
细节	依据实际灵活增加细致要求

除了简单的图像与视频之外，GAI 的一个优势就是能够生成虚拟数字人形象，这样能减轻教师需真人出镜录制视频的压力，有效提高视频的制作效率，数字人形象一般可分为拟真数字人（如图 3 所示）与卡通数字人（如图 4 所示）两类，教师可根据学习者的认知风格做出选择。一般而言，面向成人学习者，可优先选择拟真数字人，面向少儿学习者则可选择卡通数字人。数字人形象的要求如表 6 所示，需要说明的是，数字人教师可不出现在视频中，视频直接呈现 PPT 的内容；或者，数字人也可仅在部分需要的环节出现。中文教师可根据教学设计灵活安排数字人教师出场与否以及何时出现。

[7] DeepSeek-V4-Pro 网址参见：https://chat.deepseek.com/。

表 6 数字人形象要求

维度	要素	内容
形象	服饰	整洁、正装
	神情	和蔼、严肃
语音	语速	适中、较慢
	语调	平调、升调、跌宕起伏
	音量	适中、较大
状态	静态	口型一致
	动态	口型、手势、转身

图 3 拟真数字人形象

图 4 卡通数字人形象

生成分镜脚本后，教师使用 GAI 对各镜头的内容进行细化，生成每个镜头对应的脚本素材。随后，教师利用这些素材直接生成图像和视频，并将生成的视频片段拼接为完整视频。各镜头细化的具体内容见附录 2。在视频模型选择方面，综合考虑制作成本和技术可行两个方面，本研究选用 Dou bao-Seedance-2.0（下文简称豆包）

作为视频生成模型，后续图像和视频的生成均使用该模型。

在生成图像和视频时，目前单个免费的 GAI 工具通常无法直接生成超过 10 秒的视频，因此最终的教学视频常需由多个短视频拼接而成。生成视频有两种思路，教师可酌情选择。

第一种是文生视频模式。该模式依据详细的画面脚本，GAI 读取画面提示词后逐一生成各画面的视频，并对视频编号保存。最后，教师将视频导入剪辑软件进行合成。该模式的关键在于画面脚本，教师需明确教学活动的展开方式、讲解顺序、各环节涉及的语言知识（如偏误类型出现的位置）、交互式教学活动设计、视频元素、例句选择及其呈现方式等。该模式对教师专业素养要求较高，教师需根据分镜脚本改编成画面脚本。

以“洗脸”为例，对应的画面脚本如表 7 所示。由于豆包生成的单个视频一般在 10 秒左右，因此表 7 将两次输入的画面脚本提示词合在一起，如果视频中出现了人物，需要在同一个对话框中进行视频的生成，需要特别提示“人物和画面与前一个视频保持一致”，确保前后人物不变。

表 7 文生视频模式画面脚本示例

画面 1：老师先微笑着展示女孩洗脸的图片，屏幕上出现“洗脸”。 画面 2：老师说“大家好！今天我们学习‘洗脸’”。 画面 3：老师用手势比画解释洗脸的词义。 画面 4：视频中通过动画把“洗脸”拆开，依次插入“个”“了”“完”，变成“洗个脸”“洗了脸”“洗完脸”，让学生看清离合词的用法，不需要拼音。 画面 5：屏幕上出示例句“我每天早上洗脸。”和“我洗了脸。”，并配有早上起床的图片来帮助理解。 画面 6：屏幕上出现“洗洗脸”三个字，下面箭头指向“AAB”，说明离合词的重叠形式是“AAB”形式。 画面 7：老师微笑着挥手说“你也用‘洗脸’说个句子吧！再见！” 讲解时需要语速稍缓，用词清晰准确，尽量讲解慢一点，让学生能听懂。依次完成下面内容的生成。

根据表 7 中的画面脚本生成的视频画面示例如图 5 所示，2 张图片为先后生成的 2 个视频中的截图，人物形象基本上保持了稳定，呈现的示例清晰，文字没有错误，可以作为教学视频使用。

图 5 文生视频模式生成视频内容示例

第二种是图生视频模式。该模式先依据分镜脚本生成各画面的图片，再将两张图片分别设为起点和终点，通过提示词补充中间过渡画面，形成流畅视频，最后按顺序拼接合并。

具体步骤如下：首先，准备首帧和尾帧两张图片（关键帧），在豆包中选择首尾帧模式，并上传两张图片。其次，将教学中需讲解的内容作为提示词输入，说明从首帧到尾帧之间应呈现的内容及其呈现方式。最后，豆包根据提示词自动补帧。

相比文生视频模式，图生视频模式对教师要求更低，便于控制生成过程和细节，并能及时调整错误。通过设定首尾帧，教师引导生成方向，豆包则自动完成视频生成。

以“见面”为例，提示词可设计为：“风格为扁平化教育动画；角色为青年女教师（齐肩黑发、圆框眼镜、浅蓝衬衫、深蓝长裤），保持造型一致，比例为 16:9，请你根据上述描述绘制老师在教室的图像。”将上述提示词输入豆包，即可生成教师在教室的图片。再输入下一画面的提示词：“以这张图为基础，图片风格为扁平化教育动画，比例为 16:9，生成画面左侧出现一个男孩头像图标，旁边标‘见（see）’；右侧出现女孩头像图标，标‘面（face）’的画面。”生成的图片如图 6 所示。

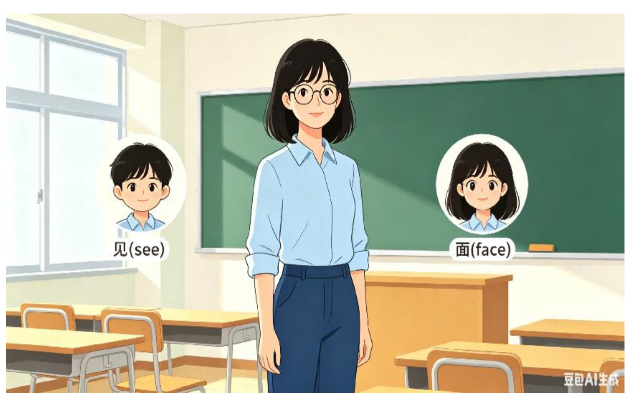


图 6 图生视频模式生成图片示例

将图 5 的 2 张图片示例导入豆包的首尾帧，输入以下提示词：“以这 2 张图片为基础，生成一个 10 秒的教学视频，比例为 16:9；文字动效（手写描边出现，气泡缓入缓出）；音频（老师原声讲解，清晰无杂音，略带中文教学特有的和缓节奏）；无背景音乐。镜头为统一扁平化动画风格，干净米白色的背景。画面左方出现一个微笑男孩头像，下方弹出英文‘见（see）’；右方出现微笑女孩头像，下方弹出‘面（face）’。两个头像相向滑动，在画面中央会合，变为两人握手的小图标，上方渐显汉字‘见面’，并浮现柔和光晕。画外音来自女教师，语气活泼。镜头固定，平视，无角色出镜，只展示图形文字动画。场景/动作：画面左侧出现一个男孩头像图标，旁边标‘见（see）’；右侧出现女孩头像图标，标‘面（face）’。两人向中间移动，面对面时图标合并，出现握手符号和汉字‘见面’。”教师画外音解说。台词：‘见’是 see，‘面’是 face，两个人 face to face，就是‘见面’——meet。”，进行内容的生成，视频内容截图如图 7 所示。生成视频示例已由哔哩哔哩（bilibili）用户“ai 学中文”[8]在平台发布，可供读者观看与参考。（ai 学中文 UID:3707016988068229）

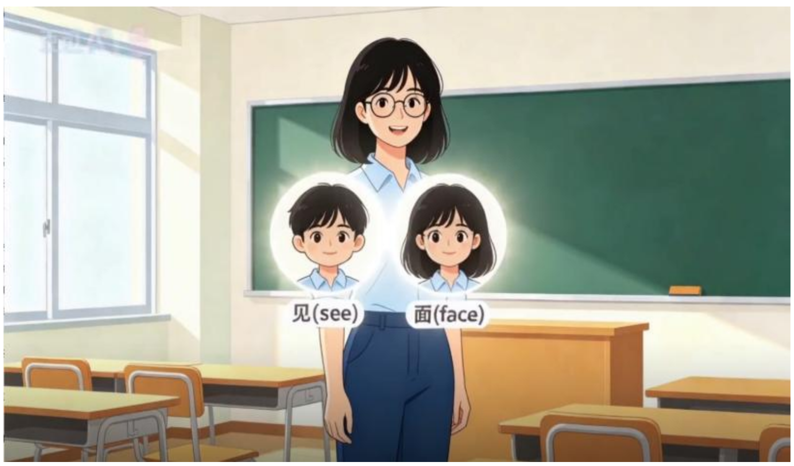


图 7 图生视频模式生成视频内容示例

4.4 生成过程总结

在整个生成实践过程中，教师的教学准备（包括教学资料、偏误分类等）构成了向 GAI 发出精准指令的提示词基础。提示词中需融入教学原则与方法，以确保 GAI 生成的脚本与教学需求相匹配。教师对脚本进行修改后，基于分解后的画面生成图片与视频，并将视频片段合成，即可用于教学。此外，教师可在教学实践中收集学习者的反馈，例如学习者是否偏好数字人教师出镜，以及拟真数字人与卡通数字人哪一种更符合其需求。这些反馈可为教学视频的改进提供参考。

为方便教师使用 GAI 工具生成脚本、图片与视频，本研究汇总了当前常用的 GAI 工具及其网址，便于教师直接访问。具体汇总结果见表 8。

[8] 生成视频示例参见： https://b23.tv/PamfKDu。

表 8 常用 GAI 工具汇总表

类别	模型名称	官网链接	是否收费
脚本生成	DeepSeek-V4	https://www.deepseek.com	免费（网页版）； API 收费
	Kimi K2.7	https://kimi.moonshot.cn	部分免费；API 收费
	GLM-5.2	https://www.zhipuai.cn	API 收费
	豆包	https://www.doubao.com	免费
	通义千问 Qwen	https://www.qianwen.com/	部分免费；API 收费
	Gemini	https://gemini.google.com	部分免费
	ChatGPT	https://chat.openai.com	部分免费； Plus 收费
文生图片	通义万相	https://tongyi.aliyun.com/wan/	部分免费
	即梦 AI	https://jimeng.jianying.com	部分免费
	Midjourney	https://www.midjourney.com	收费
	Stable Diffusion	https://stability.ai	开源免费 （可本地部署）
	文心一言	https://chat.baidu.com/	开源免费
	智小象	https://www.hidreamai.com	部分收费
	Agnes-Image	https://www.agnes.ai	免费开放 API
文生视频	豆包	https://www.doubao.com/	API 收费
	可灵	https://kling.kuaishou.com	部分免费
	海螺	https://www.hailuoai.com	部分免费
	通义万相	https://tongyi.aliyun.com/wan/	部分免费
	腾讯混元	https://aistudio.tencent.com/	部分免费

类别	模型名称	官网链接	是否收费
	Vidu	https://www.vidu.cn	部分免费
图生视频	通义万相	https://tongyi.aliyun.com/wan/	API 收费
	豆包	https://www.doubao.com/	API 收费（0.023 元/千 tokens）
	可灵	https://kling.kuaishou.com	部分免费
	即梦 AI	https://jimeng.jianying.com	部分免费

5. 结果分析

5.1 师生评价分析

本文采用混合研究方法，通过向中文教师与中文学习者发放量表问卷收集数据（量表见表 3，问卷见附录 3），并辅以开放式问题收集质性反馈。本次研究共回收师生问卷 75 份，其中教师问卷 30 份，学习者问卷 45 份。本研究采用 Cronbach's α 系数评估问卷的内部一致性，《GAI 生成中文二语教学视频评估量表》总体 Cronbach's α 为 0.823，表明问卷具有良好的信度，可以进行后续数据的统计分析。数据结果见表 9。

表 9 GAI 生成教学视频素材师生评分表

维度		教师评分		学习者	
		平均分	标准差	平均分	标准差
设计	激励性	3.30	0.78	4.20	0.75
	清晰性	4.10	0.30	4.27	0.68
	可用性	4.10	1.04	4.20	0.75
内容	准确性	3.80	1.08	4.20	0.83
	充足性	3.60	0.92	4.33	0.79
	恰当性	4.10	0.70	4.27	0.77
技术	流畅性	3.90	0.83	4.20	0.75

	同步性	3.30	0.98	4.33	0.60
	生动性	3.20	1.22	4.20	0.65

（1）设计维度

在设计维度，学习者的评分从高到低依次为清晰性（4.27）、激励性（4.20）、可用性（4.20），均在 4.20 以上，说明学习者认为视频的视听呈现、学习兴趣激发和教学使用便利性均达到较高水平。教师的评分从高到低依次为清晰性（4.10）、可用性（4.10）、激励性（3.30）。教师对清晰性的评分最高（4.10），与学习者差距最小（差值 0.17），说明双方均认可视频的视听效果；对激励性的评分最低（3.30），与学习者差距最大（差值 0.90），说明教师认为 GAI 生成的视频在激发学习兴趣方面仍有明显不足，而学习者则认为视频具有一定的吸引力。

教师对激励性的评分显著低于其他指标（3.30 vs 清晰性 4.10、可用性 4.10），标准差为 0.78，说明教师群体普遍认为当前 GAI 生成的视频在内容趣味性和参与度方面需要改进。与之相对，学习者在激励性维度给出 4.20 分，表明学习者对视频的接受度较高。这一差异说明教师与学习者对"趣味性"的评判标准存在不同——教师更关注教学内容的呈现方式是否具有教育意义上的吸引力，学习者则更关注视频是否能够引起持续观看的兴趣。

（2）内容维度

在内容维度，教师评分从高到低依次为恰当性（4.10）、正确性（3.80）、充足性（3.60）。恰当性得分最高，说明教师认为词汇发音与图像素材基本符合教学群体的认知水平；充足性得分最低，表明教师认为 GAI 生成视频的信息量尚不足以充分支撑教学内容。学习者评分从高到低依次为充足性（4.33）、恰当性（4.27）、正确性（4.20），三项均高于 4.20，说明学习者对生成内容整体认可度较高。

教师与学习者在正确性（3.80 与 4.20）、充足性（3.60 与 4.33）两个指标上差距较大，差值分别为 0.40 和 0.73，说明教师对内容的准确性和信息覆盖度要求更高。两者在恰当性（4.10 与 4.27）上差距最小（差值 0.17），表明师生均认可视频素材与教学群体的匹配度。教师对充足性的评分最低，而学习者对充足性评分最高，反映出教师期望生成视频包含更充分的教学信息，学习者则认为现有内容已基本满足学习需求。

（3）技术维度

在技术维度，师生评分差距较为显著。学习者的评分从高到低依次为同步性（4.33）、流畅性（4.20）、生动性（4.20），三项均高于 4.20，说明学习者认为视频的音画同步、画面流畅度和数字人形象自然度均达到较高水平。教师的评分从高到低依次为流畅性（3.90）、同步性（3.30）、生动性（3.20），均低于学习者。其

中，教师对生动性的评分最低，且标准差达到 1.22，说明教师普遍认为当前 GAI 生成的数字人形象在拟真度、自然度和表现力方面仍有不足，且教师群体内部对此感受差异较大。学习者对同步性的评分最高（4.33），标准差仅 0.60，说明学习者对视频的音画同步效果认可度较高，评价较为一致。教师对流畅性的评分（3.90）相对接近学习者（4.20），说明双方均认可视频的播放流畅度和画面清晰度。

这一差异反映出教师与学习者对技术质量的关注点不同。学习者作为终端观看者，对视频的流畅性和音画同步更为敏感，这两项指标直接影响了观看体验；而教师作为教学资源的制作者和设计者，对数字人形象的自然度和表现力有更高期待，认为其直接关系到教学视频能否有效吸引学习者注意力并营造沉浸感。教师对生动性的低评分也提示当前 GAI 工具在数字人生成的人物细节、表情丰富度和肢体自然度等方面仍有较大改进空间。

（4）小结

综合三个维度，学习者的评分在九个指标上均高于教师。其中技术维度师生差距最大，尤其是生动性和同步性两个指标；设计维度中激励性的差距次之；内容维度中充足性的差距也较为明显。教师评分中，生动性（3.20）、激励性（3.30）和同步性（3.30）低于 3.50，是教师群体评价最低的三项，说明当前 GAI 工具在数字人形象自然度、学习兴趣激发和音画同步方面仍有较大改进空间。从标准差来看，教师评分的离散程度普遍高于学习者，尤其在生动性（教师 1.22，学习者 0.65）和可用性（教师 1.04，学习者 0.75）两个指标上最为明显，说明教师群体内部对 GAI 工具的评价存在较大分歧，而学习者的评价则更为一致。这一差异可能与教师对技术工具的操作经验、教学理念以及对生成内容质量要求的不同有关。

总体而言，GAI 生成的教学视频在学习者层面已获得基本认可，但在教师层面尚未达到他们期待的教学使用的理想标准。

5.2 资源创制的可行性分析

本研究验证了 GAI 在提升教学视频制作效率方面的优势。根据作者的经验，在传统流程中，一位中文教师独立完成一个 3 分钟的教学视频，需经历脚本撰写（约 1 小时）、素材搜集与 PPT 制作（约 3 小时）、视频录制与剪辑（约 1 小时），总计约 5 小时。而采用本研究设计的 GAI 工作流，教师完成提示词设计后，由 GAI 生成教学脚本，平均耗时 5 分钟；再利用文生图或文生视频工具制作讲解视频，平均耗时 30 分钟。从开始到完成视频制作可在 1 小时以内完成，加上必要的人工修订与视频重制时间，总耗时控制在 2 小时以内。这表明 GAI 是一种具备可扩展性和低成本特点的教学视频开发工具。

然而，关注效率提升的同时不能忽视生成内容的可信度。未经教师审核的生成内容存在教学风险，如脚本文化偏差、图片人物服饰不当、例句不够典型、语境脱离实际交际等。经过教师主导的人工验证后，GAI 生成内容的教学可靠性得以提高，

更能胜任教学任务。因此，GAI 生成内容应定位为教学半成品或初稿，教师的角色从内容创造者转变为内容设计者、审核者与优化者（Dennison et al., 2025）。对 GAI 生成内容的信任应建立在系统性验证的基础之上，而非盲目接受。

6. 讨论与启示

基于上述实践流程与师生评价结果，本研究围绕三个研究问题展开讨论，并提炼对中文二语教学资源开发与教师发展的启示。

6.1 回应 RQ1：人机协同开发资源的实践模式

RQ1 关注的是中文教师如何与 GAI 协同开发教学视频，本研究通过全过程实践，总结出一条可参考、可复制的操作路径。

该路径包含三个关键环节。第一，教师设计语言教学内容。教师需先根据课程标准和工具书明确教学要点，再将相关知识嵌入提示词，以避免 GAI 的误判。第二，用结构化的提示词驱动脚本生成。本研究采用 CRISPE 框架设计提示词，生成初步分镜脚本后，教师需对例句选择、讲解顺序和练习设计进行审核修改。修改后的脚本既作为教学材料，也作为后续视频生成的画面依据。第三，两种视频生成模式的灵活选用。文生视频模式要求教师具备清晰的画面构思，适用于有明确设计思路的场景；图生视频模式通过设定首尾帧让 GAI 自动补帧，操作要求较低，适用于希望控制生成细节的场景。

上述环节中，教师负责教学设计、内容审核和质量把关，GAI 负责脚本草拟、图像生成和视频合成。教师负责全过程主导，而 GAI 主要作为辅助工具发挥作用。

6.2 回应 RQ2：师生对资源的评价存在差异

RQ2 关注的是中文教师与学习者如何评价 GAI 生成教学视频的整体质量。评价数据显示，学习者在设计、内容、技术三个维度共九个指标上的评分均高于教师，反映出师生评价视角的显著差异。

通过对设计、内容、技术的三维分析，本研究发现，学习者的评价主要基于直观的视频观看体验，更加关注视频是否易于理解、画面是否清晰、内容是否有趣，因此对 GAI 生成视频的整体接受度较高。中文教师的评价则基于教学专业性，更能察觉 GAI 生成内容在教学逻辑严谨性、例句典型性、文化内涵深度等方面的不足，特别是对数字人的表现力更加关注。

上述差异说明，学习者的评价反映的是视频的“可接受性”——是否能够满足观看和学习的基本需求；教师的评价反映的是视频的“教学适切性”——是否能够在教学环节中有效发挥作用，两者共同构成对 GAI 生成教学视频的完整评估。因此，

在评估 GAI 生成教学视频时，不能仅以学习者满意度作为唯一标准，教师的专业评判同样不可或缺。

6.3 回应 RQ3：GAI 生成视频的主要问题与挑战

RQ3 关注的是 GAI 辅助中文教学视频开发过程中面临的问题与挑战。通过全过程实践追踪与师生反馈，本研究识别出两类典型问题与挑战。

第一类是事实性错误，GAI 在生成脚本时可能出现知识性错误，如 4.1 节所述将“宾语使用不当”误判为“介词缺失”。这类错误涉及概念混淆，若不纠正可能误导学习者。第二类是教学层面的偏差，GAI 生成的例句虽语法正确，但存在“不典型”或“脱离真实交际语境”的问题，降低了教学视频的实际效果。

上述两类风险均无法通过升级模型或增加参数量来自动消除，需要中文教师借助学科知识、教学经验和跨文化敏感度进行识别和修正。

7. 结语

本研究以词汇教学视频为例，探讨了 GAI 在中文二语教学视频制作中的可行性。研究发现，GAI 能够显著提升教学视频的制作效率，降低开发成本，生成内容结构完整，具备初步使用潜力。本研究具有以下价值：第一，为 GAI 在中文二语教学视频生产中的应用提供了基于具体词汇内容的实证证据，验证了其可行性与应用边界。第二，根据已有的研究综合构建了一个 GAI 生成中文二语教学视频的评估量表，为后续研究提供了一个开发与完善测量工具的基础。第三，提出了以教师专业判断为核心的、可复制的 GAI 辅助教学视频设计的人机协同框架，为教师提供了生成实践指南。

GAI 工具迭代迅速，本研究所用工具与提示词仅能视为初始版本，随着技术发展，工具与提示词需持续更新和改进。因此，GAI 并非完全可靠的教学视频生产者，而是需要教师深度介入的半自主设计助手，教师的专业素养与审核验证是人机协作的基础。教师与学习者应在实践中深化中文教学的智慧化（马瑞祾 & 徐娟，2026），推进国际中文教育人机协同的深入发展（丰迪 & 徐娟，2026）。

未来研究可从以下方向深化：第一，开展更大规模、覆盖更多课型的教学实验，在一线教学中检验 GAI 生成视频的实际助学效果；第二，探索更有效的提示词设计方法，提升 GAI 输出内容的教学深度和适用性；第三，关注师生在 GAI 辅助资源开发与使用中的身份认同、情感态度、学习动机与学习体验，构建更为协调的人机协同共创与共学模式。

致谢：本文系教育部中外语言合作交流中心国际中文教育研究课题重大项目“面向国际中文教育的生成式人工智能（AIGC）应用研究”（24YH03A）；北京语言大学研究生创新基金项目“国际中文教育智能体评价指标体系的构建与应用探索”（项目编号：26YCX072）的阶段性成果。马瑞祾为本

文通讯作者。

参考文献

Baek, I., Kim, Y., & Kwak, N. (2026). PedaCo-Gen: Scaffolding pedagogical agency in human-AI collaborative video authoring. *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3772363.3798741

Chen, X., & Wu, D. (2024). Automatic generation of multimedia teaching materials based on generative AI: Taking Tang poetry as an example. *IEEE Transactions on Learning Technologies, 17*, 1327-1340.

Ding, A., & Meng, X. (2025). From imitation to reconstruction: The impact of AI-generated lesson plans on instructional design ability of novice teachers in international Chinese education. *Journal of Guizhou Normal University (Social Sciences), 2025*(05), 42-52. [丁安琪, &蒙小凤. (2025). 从模仿到重构: AI 生成教案对国际中文教育新手教师教学设计能力的影响. *贵州师范大学学报（社会科学版）*, *2025*(05), 42-52]

Do, H.-D., Wen, J.-M., & Huang, S. K. (2022). Impact of humorous Chinese-teaching videos and Echo method on teaching Chinese language: A case study of Vietnamese students. *The Asia-Pacific Education Researcher, 31*(6), 725-738.

Dong, Q., Li, L., & Dai, D. (2024). A survey on in-context learning. arXiv. https://arxiv.org/abs/2301.00234

Fan, F., Zhang, Y., & Bai, Q., Lin, L. (2012). Analysis of the application effect of teaching mode based on interactive micro-video teaching resources. *Modern Educational Technology, 22*(06), 24-28. [范福兰,张屹,白清玉,林利 (2012). 基于交互式微视频教学资源教学模式的应用效果分析. *现代教育技术*, *22*(06), 24-28.]

Feng, D., Li, J., & Xu, J. (2026). AI micro-drama in international Chinese education: Theoretical construction, technical paths and application practice. *Journal of Hainan Normal University (Social Sciences), 39*(02), 124-134. [丰迪，李嘉仪, & 徐娟.(2026). 国际中文教育 AI 微短剧：理论建构、技术路径与应用实践. *海南师范大学学报（社会科学版）, 39*(02), 124-134.]

Feng, D., & Xu, J. (2026). Human-AI collaborative teaching in international Chinese education: Theoretical framework and practical approach. *Journal of Yunnan Normal University (Teaching and Research of Chinese as a Foreign Language), 24*(01), 75-83. [丰迪, & 徐娟.(2026). 国际中文教育人智协同教学：理论框架与实践进路. *云南师范大学学报（对外汉语教学与研究版）*, *24*(1), 75-83.]

Fernandez-Espinosa, M., Zhang, K., Bendarkawi, J., Ponce, A., Mata, S.-C., Aliu, A., Zhang, L., Fernandez Medina, F., Mangione-Lora, E., Monroy-Hernandez, A., & Gomez-Zara, D. (2025). Practicing a second language without fear: Mixed reality agents for interactive group conversation. arXiv. https://arxiv.org/abs/2510.08227

Han, X., Ma, R., & Xu, J. (2025). Exploring technical pathways for DeepSeek-enabled international Chinese teaching resource construction: A case study of graded reading text generation. *Journal of International Chinese Teaching,* (1), 30-40.[韩

欣欣，马瑞祾, & 徐娟.(2025).DeepSeek 赋能国际中文教学资源建设的技术路径探索——以分级阅读文本生成为例. *国际汉语教学研究*, (1), 30-40.]

Hou, Z., & Xu, J. (2025). Research on personalized generation of international Chinese reading materials based on large language models. *International Chinese Language Education (Chinese and English), 10*(04), 32-44.[侯泽煜, & 徐娟.(2025). 基于大语言模型的国际中文阅读材料个性化生成研究. *国际中文教育（中英文）*, *10*(04), 32-44.]

Huang, Y. (2025). Practice and reflections on generative artificial intelligence in teaching Chinese as a foreign language: A case study of the Confucius Institute at Cambodia-China University of Applied Sciences and Technology. *Oriental Renaissance: Innovative, Educational, Natural and Social Sciences, 5*(26), 978-983.

Jiang, B. (2025). Narrative practices of telling Chinese stories through domestic animated film and television works. *Editors' Friend*, (8), 89–97. [姜博. (2025). 国产动画影视作品讲好中国故事的叙事实践研究. *编辑之友*, (8), 89-97.]

Jiang, X. (1998). Vocabulary acquisition research and its implications for language teaching. *Language Teaching and Linguistic Studies*, (3), 65–73. [江新. (1998). 词汇习得研究及其在教学上的意义. *语言教学与研究*, (3), 65-73.]

Li, Q., & Sun, Y. (2023). New approaches to Chinese culture teaching: Localization of content and storytelling of methods. *Applied Linguistics*, (1), 33-44. [李泉, &孙莹. (2023). 中国文化教学新思路：内容当地化、方法故事化. *语言文字应用*, (01), 33-44.]

Li, R., & Wu, M. (2005). On two principles of vocabulary teaching in teaching Chinese as a foreign language. *Language Teaching and Linguistic Studies*, (02), 41-47. [李如龙, &吴茗. (2005). 略论对外汉语词汇教学的两个原则. *语言教学与研究*, (02), 41-47.]

Laufer, B. (1986). Possible changes in attitude towards vocabulary acquisition research. *International Review of Applied Linguistics*, *24*(1), 69 – 75.

Ma, R., & Xu, J. (2025). DeepSeek-empowered “five transformations” in smart international Chinese education. *Journal of Yunnan Normal University (Teaching and Research on Chinese as a Foreign Language Edition)*, *23*(2), 1–10. [马瑞祾, &徐娟. (2025). DeepSeek 赋能国际中文智慧教育“五件”创变. *云南师范大学学报（对外汉语教学与研究版）*, *23*(2), 1-10.]

Ma, R., & Xu J. (2026). Generative AI empowering the smartification of international Chinese teaching and learning: Reflections based on empirical research. *Chinese Teaching in the World*, *40*(01), 129-142. [马瑞祾, & 徐娟.(2026). 生成式人工智能赋能国际中文教与学的智慧化——基于实证研究的思考. *世界汉语教学*, *40*(01), 129-142.]

McLaughlin, B. (1978). The monitor model: Some methodological considerations. *Language Learning*, *28*(2), 309-332.

Minalla, A. A. (2024). Enhancing young EFL learners' vocabulary learning through contextualizing animated videos. *Theory and Practice in Language Studies*, *14*(2), 358-364.

Nagy, W. E., Anderson R. C., & Herman P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, *24*(2), 237-270.

Ou, Z., Liu, Y., Qin, K., Wang, Y., & Li, X.. (2024). Generation and evaluation of AI multimodal teaching resources: Based on the application of AIGC in international Chinese education. *Modern Educational Technology*, *34*(09), 37-47.[欧志刚，刘玉屏，覃可，王艺, &李欣燚.(2024).人工智能多模态教学资源的生成与评价——基于 AIGC 在国际中文教育的应用.*现代教育技术*, *34*(09), 37-47.]

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., & Mishkin, P. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

Shi, J., & Ge, X. (2025). Application of generative artificial intelligence in international Chinese language teaching based on prompt engineering strategies. *Journal of Shaanxi Normal University (Philosophy and Social Sciences Edition)*, *54*(5), 135-145. [史金生, & 葛星辰. (2025). 基于提示词策略的生成式人工智能在国际中文教学中的应用研究. *陕西师范大学学报(哲学社会科学版)*, *54*(5), 135-145.]

Sreemathy, R., Chordiya, P., Khurana, S., & Turuk, M. (2024). Sign language video generation from text using generative adversarial networks. *Optical Memory and Neural Networks*, *33*(4), 466-476.

Su, F., & Wen X. (2025). Opportunities, problems and countermeasures of applying short videos in college teaching. *China University Teaching*, (03), 77-82. [苏放, &温向明. (2025). 短视频应用于高校教学的机遇、问题和对策. *中国大学教学*, (03), 77-82.]

Wang, G., Zhou G., & Li Z. (2022). Research on the application of Chinese teaching based on social media video platforms. *International Journal of Distance Education Technologies*, *20*(1), 1-16.

Wang, H., Wang, Y., Ding, C., Tang, L. (2017). 3D video resources: A new form of digital educational resources. *Modern Educational Technology*, *27*(04), 19-24. [王洪梅，王运武，丁超, & 唐丽.(2017). 3D 视频资源:数字化教育资源的新形态. *现代教育技术*, *27*(04), 19-24.]

Wang, R. (2005). Analysis of errors in the use of Chinese separable words by international students. *Applied Linguistics*, (S1), 27-29. [王瑞敏. (2005). 留学生汉语离合词使用偏误的分析. *语言文字应用*, (S1):27-29.]

Wen, J. M., Dung, H., Liu, E. Z. F., Lin, C. H., & Huang, S. K. (2021). Learning Chinese in a role as news broadcaster: Is this a worth-trying teaching method? *International Journal of Online Pedagogy and Course Design (IJOPCD)*, *11*(1), 15-35.

Xu, T., Liu, Y., Jin, Y., Qu, Y., Bai, J., & Zhang, W. (2025). From recorded to AI-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology, 56*(4), 1463-1487.

Yanto, E. S., & Nugraha S. I. (2018). Video viewing as a mediation of learning content-based vocabulary: Assisting students in understanding disciplinary vocabulary in context. *Indonesian Journal of Applied Linguistics*, *8*(2), 316-324.

Hang, Y., Huang, X., Xie, H., Bai, Q. (2013). Development and application of interactive micro-video teaching resources: Taking "Research Methods in Educational Technology" as an example. *e-Education Research*, *34*(05), 48-54. [张屹，黄欣，谢浩泉，& 白清玉.(2013). 交互式微视频教学资源的研发与应用——以“教育技术学研究方法”为例. *电化教育研究*, *34*(05), 48-54.]
Zhang, Z., Li, H., & Zhou, J. (2023). Teaching with social context in instructional video facilitates second language vocabulary learning. *Heliyon, 9*(3), 12.
Zhou, Y. (2016). Micro-video teaching clips: Value, characteristics and creation process. *Modern Educational Technology*, *26*(08), 69-75. [周洋. (2016). 微视频教学短片:价值、特征及创作流程. *现代教育技术*, *26*(08), 69-75.]

附录 1 汇微课制作的完整提示词

DeepSeek-V4-Pro 生成分镜脚本

镜号	时长	画面描述	台词/字幕	镜头运动
1	10s	扁平化动画风格，明亮教室，浅米色背景，柔光。一位年轻中国女教师，齐肩黑发，戴圆框细金属眼镜，穿浅蓝色棉质衬衫、深蓝长裤、白色运动鞋，正面站立，面带亲切微笑，右手轻轻挥手。她说话时，头顶上方浮现楷体汉字“离合词”，随即“见面”两个大字跳至画面中央，带有轻微弹跳动画，线条干净，色彩以浅蓝、暖黄为主。	大家好，今天我们一起来学习：离合词——“见面”。	中景镜头，固定机位，无镜头运动。
2	10s	教学引入： 同一扁平化动画风格，干净米白背景。在同一黑板左方出现一个微笑男孩头像，下方弹出英文“见（see）”；右方出现微笑女孩头像，下方弹出“面（face）”。两个头像相向滑动，在画面中央会合，变为两人握手的小图标，上方渐显汉字“见面”，并浮现柔和光晕。	“见”是 see，“面”是 face，两个人 face to face，就是“见面”——meet。	镜头固定，平视，只展示图形文字动画。
3	17s	偏误示例与讲解： 扁平化教育动画风格，角色外观完全保持一致：女教师，齐肩黑发，圆框细金属眼镜，浅蓝色棉质衬衫，深蓝长裤，白色运动鞋。教师中近景，	“见”是看，“面”是脸。见面本来就是	镜头固定，平视。

镜号	时长	画面描述	台词/字幕	镜头运动
		站在画面左侧，右侧是一块黑板。她边说边用右手轻触黑板：白板上先弹出汉字“见”配一只卡通眼睛图标，“面”配一张微笑面孔图标。两者向中心滑动，融合成“见面=看见对方的脸”的文字条，并有两个小人面对面挥手。紧接着，“面”字被一个放大镜图标圈住，放大闪烁，弹出标签“宾语在里面！”。随后画面下方淡入一组对比：左边是一碗热饭图标和红色叉号卡“❌ 吃饭米饭”，右边是两个小人图标和红色叉号卡“❌ 见面他”。两组叉号同时抖动两次后碎裂消失，分别化为绿色对勾卡“✅ 吃饭”和“✅ 和他见面”。教师看向镜头，自然微笑点头。动画线条圆润，配色以暖黄、浅蓝、白色为主，画面清爽无杂物。文字动效采用手写描边出现，气泡缓入缓出。	“看见对方的脸”，“面”已经是见的对象了！就像“吃饭”，饭就是吃的东西，不能说“吃饭米饭”。同样，见面也不能再说“见面他”，要说“和他见面”。	
4	10s	偏误示例与讲解： 同一扁平化风格，保持女教师角色完全一致（黑发、圆眼镜、浅蓝衬衫）。教师近景，面对镜头讲解。她右手向左侧划动，左侧出现一张卡片“❌ 我见面你。”，卡片呈淡红色，微微抖动；老师再向右划，卡片翻转变为“✅ 我和你见面。”并打勾。随后下方滑入“❌ 昨天我见面过他。”，再换为“✅ 昨天我见过他一面。”。当说到“见过面”时，汉字“见”和“面”之间飞入一个红色“过”字，短暂发光。	不能说“我见面你”，要说“我和你见面”。也不能说“昨天我见面过他”，要说“昨天我见过他一面”。	镜头小幅推近至特写，后拉回，画面清爽，字体圆润卡通。
5	13s	偏误示例与讲解+练习： 同一扁平化风格，女教师角色外观完全一致。教师半身景，身边依次浮现三个半透明气泡：“见	见面中间还能加很多词：见了一	镜头短暂摇向题目卡再

镜号	时长	画面描述	台词/字幕	镜头运动
		了一次面”“见过面”“见见面”。然后镜头切为画面居中显示一张题目卡“我明天见面他。❌”，教师的手从画面右侧伸入，做出“划掉”手势，卡片翻转为“我明天跟他见面。✅”。	次面，见过面。那“我明天见面他”对不对？错！要说“我明天跟他见面”。	回到教师，平滑过渡。
6	10s	总结： 同一扁平化风格，女教师外观保持不变。教师中景，双手自然向两侧打开，掌心朝上。在她胸前浮现三个圆形图标，分别用简约图形表示：1.人物前置箭头（后面不加对象），2.红色“过”字插入缝隙（中间加“过”），3.数字“1”和沙漏（数量/时间）。三图标顺时针轻微旋转，最后聚拢形成一个圆环。教师微笑点头，右手挥手道别，画面底部浮现手写体“下次见！”。	见面是离合词，记住这三条：后面不加对象，过放中间，能加动作数量时间。你肯定能用对！下次见！	镜头缓缓拉远，定格在全景，光线柔和。

附录 2 视频脚本详细版

图像与视频详细镜头

【镜头 1】开头 0-10s

场景/动作：女教师正面中景，半侧身站在黑板右侧，微笑挥手。她说话时，头顶动态弹出汉字“离合词”【离合词释义】，随后“见面”二字放大居中，笔画微微弹跳。

离合词=可以“分开”也可以“合在一起”用的动词。

合：两个字连在一起，像一个普通动词。

离：两个字中间可以插入别的词（如“了”“过”“时间”“数量”等）

台词：大家好，今天我们一起来学习：离合词——“见面”。

豆包视频提示词：扁平化动画风格，明亮教室，浅米色背景，柔光。一位年轻中国女

教师，齐肩黑发，戴圆框细金属眼镜，穿浅蓝色棉质衬衫、深蓝长裤、白色运动鞋，正面站立，面带亲切微笑，右手轻轻挥手。她说话时，头顶上方浮现楷体汉字“离合词”，随即“见面”两个大字跳至画面中央，带有轻微弹跳动画，线条干净，色彩以浅蓝、暖黄为主。中景镜头，固定机位，无镜头运动。

【镜头 2】导入 10—20s

场景/动作：

同一黑板左侧出现一个男孩头像图标，旁边标“见(see)”；右侧出现女孩头像图标，标“面(face)”。两人向中间移动，面对面时图标合并，出现握手符号和汉字“见面”。教师画外音解说。

台词：“见”是 see，“面”是 face，两个人 face to face，就是“见面”——meet。

豆包视频提示词：同一扁平化动画风格，干净米白背景。在同一黑板左方出现一个微笑男孩头像，下方弹出英文“见（see）”；右方出现微笑女孩头像，下方弹出“面（face）”。两个头像相向滑动，在画面中央会合，变为两人握手的小图标，上方渐显汉字“见面”，并浮现柔和光晕。画外音来自女教师，语气活泼。镜头固定，平视，无角色出镜，只展示图形文字动画。

【镜头 3】讲解 1 20—37s

场景/动作：场景/动作：教师侧身站在画面左侧，同一黑板占据左侧三分之二。她轻点黑板，先蹦出两个大字：“见”旁边出现一只眼睛图标，“面”旁边出现一张笑脸面孔图标。两者向中间靠拢，形成“见面=看见对方的脸”，两个小人挥手。随后“面”字高亮闪烁，被一个放大镜圈住。画面一转，下方滑入一组对比：左半边是一碗热饭图标 + “❌ 吃饭米饭”，右半边是两个小人图标 + “❌ 见面他”。红叉闪烁两下后，两道叉号同时碎裂，替换为正确形式：“✅ 吃饭”和“✅ 和他见面”。教师微笑着点头，画面干净收束。

台词：“见”是看，“面”是脸。见面本来就是“看见对方的脸”，“面”已经是见的对象了！就像“吃饭”，饭就是吃的东西，不能说“吃饭米饭”。同样，见面也不能再说“见面他”，要说“和他见面”。台词：“见”是看，“面”是脸。见面本来就是“看见对方的脸”，“面”已经是见的对象了！就像“吃饭”，饭就是吃的东西，不能说“吃饭米饭”。同样，见面也不能再说“见面他”，要说“和他见面”。

豆包视频提示词：扁平化教育动画风格，角色外观完全保持一致：女教师，齐肩黑发，圆框细金属眼镜，浅蓝色棉质衬衫，深蓝长裤，白色运动鞋。教师中近景，站在画面左侧，又侧是一块黑板。她边说边用右手轻触黑板：白板上先弹出汉字“见”配一只卡通眼睛图标，“面”配一张微笑面孔图标。两者向中心滑动，融合成“见面=看见对方的脸”

的文字条，并有两个小人面对面挥手。紧接着，“面”字被一个放大镜图标圈住，放大闪烁，弹出小标签“宾语在里面！”。随后画面下方淡入一组对比：左边是一碗热饭图标和红色叉号卡“❌ 吃饭米饭”，右边是两个小人图标和红色叉号卡“❌ 见面他”。两组叉号同时抖动两次后碎裂消失，分别化为绿色对勾卡“✅ 吃饭”和“✅ 和他见面”。教师看向镜头，自然微笑点头。镜头固定，平视，动画线条圆润，配色以暖黄、浅蓝、白色为主，画面清爽无杂物。

【镜头 4】讲解 2　37—47s

场景/动作：教师再次出镜，左侧弹出错误例句卡“❌ 我见面你。”，卡变红色闪动；随即被替换为正确卡片“✅ 我和你见面。”；下方再弹“❌ 昨天我见面过他。”，接着变为“✅ 昨天我见过他一面。”。说到“过”时，“见”与“面”中间插入一个红色的“过”字特效。

台词：不能说“我见面你”，要说“我和你见面”。也不能说“昨天我见面过他”，要说“昨天我见过他一面”。

豆包视频提示词：同一扁平化风格，保持女教师角色完全一致（黑发、圆眼镜、浅蓝衬衫）。教师近景，面对镜头讲解。她右手向左侧划动，左侧出现一张卡片“❌ 我见面你。”，卡片呈淡红色，微微抖动；老师再向右划，卡片翻转变为“✅ 我和你见面。”并打勾。随后下方滑入“❌ 昨天我见面过他。”，再换为“✅ 昨天我见过他一面。”。当说到“见过面”时，汉字“见”和“面”之间飞入一个红色“过”字，短暂发光。镜头小幅推近至特写，后拉回，画面清爽，字体圆润卡通。

【镜头 5】讲解 3+练习　47—60s

场景/动作：教师身旁出现两个气泡：“见了一次面”“见过面”。随后画面转为快问快答，浮现句子“我明天见面他。”，配上红叉，教师伸手将其划掉，弹出“我明天跟他见面。”并打勾。

台词：见面中间还能加很多词：见了一次面，见过面。那“我明天见面他”对不对？错！要说“我明天跟他见面”。

豆包视频提示词：同一扁平化风格，女教师角色外观完全一致。教师半身景，身边依次浮现三个半透明气泡：“见了一次面”“见过面”“见见面”。然后镜头切为画面居中显示一张题目卡“我明天见面他。❌”，教师的手从画面右侧伸入，做出“划掉”手势，卡片翻转为“我明天跟他见面。✅”。无其他角色，教师的手为同一风格插画手型。镜头短暂摇向题目卡再回到教师，平滑过渡。

【镜头 6】总结　60—70s

场景/动作：教师双手展开，画面汇聚出三个关键词图标：“后面不加对象”“中间加

‘过’”“能加时间、动作、数量”，围成一个圆环。教师微笑点头，挥手道别，字幕出现“下次见！”。

台词：见面是离合词，记住这三条：后面不加对象，过放中间，能加动作数量时间。你肯定能用对！下次见！

豆包视频提示词：同一扁平化风格，女教师外观保持不变。教师中景，双手自然向两侧打开，掌心朝上。在她胸前浮现三个圆形图标，分别用简约图形表示：1. 人物前置箭头（后面不加对象），2. 红色“过”字插入缝隙（中间加“过”），3. 数字“1”和沙漏（数量/时间）。三图标顺时针轻微旋转，最后聚拢形成一个圆环。教师微笑点头，右手挥手道别，画面底部浮现手写体“下次见！”。镜头缓缓拉远，定格在全景，光线柔和。

附录 3 调查问卷完整版

AIGC 生成教学视频调查表

本问卷旨在了解您对 GAI 生成的中文二语教学视频的使用体验与评价。所有数据仅用于学术研究，请根据您的真实感受填写。感谢您的参与！

第 1 题 您的身份[单选题]

○ 教师
○ 学生

第 2 题 您的汉语水平等级[单选题]

○ HSK1
○ HSK2
○ HSK3
○ HSK4
○ HSK5
○ HSK6

第 3 题 内容评价[矩阵量表题]

请根据您的观看结果选择最符合的项：1→5 表示非常不满意→非常满意

题目	1	2	3	4	5
准确性：教学资源没有错误与歧义	○	○	○	○	○
充足性：教学资源能总体反映教学内容	○	○	○	○	○
恰当性：教学资源能被师生理解与接受	○	○	○	○	○

第 4 题 设计评价[矩阵量表题]

请根据您的观看结果选择最符合的项：1→5 表示非常不满意→非常满意

题目	1	2	3	4	5
激励性：生成资源能提升学习者的兴趣	○	○	○	○	○
清晰性：生成资源在视听方面的清晰度	○	○	○	○	○
可用性：生成资源易于理解与使用	○	○	○	○	○

第 5 题 技术评价[矩阵量表题]

请根据您的观看结果选择最符合的项：1→5 表示非常不满意→非常满意

题目	1	2	3	4	5
流畅性：视频画面清晰、流畅，不存在卡顿现象	○	○	○	○	○
同步性：音频与画面在时间上精确对齐，口型与发音匹配	○	○	○	○	○
生动性：数字人形象及资源生动自然	○	○	○	○	○

第 6 题 您对 AIGC 生成资源辅助教学的态度[单选题]

○ 有助于教学
○ 需要客观评价
○ 可用性不足

第 7 题 您对本次教学视频的其他意见或建议[填空题]

e·Chinese Plus: An Open-Access Online Platform for Spanish-Speaking Learners of Chinese
(e·Chinese Plus：面向西班牙语学习者的中文练习在线开放平台)

Casas-Tost, Helena
Universitat Autònoma de Barcelona
Helena.Casas@uab.cat

Vargas-Urpí, Mireia
Universitat Autònoma de Barcelona
Mireia.Vargas@uab.cat

Abstract: In an era in which digital transformation is reshaping the landscape of language learning, innovative and inclusive approaches are becoming crucial. This paper presents e·Chinese Plus, an initiative of a team of lecturers from the Universitat Autònoma de Barcelona (UAB). e·Chinese Plus is a comprehensive, Moodle-based, open-access platform for enhancing the acquisition of Chinese as a foreign language (CFL) specifically addressed to Spanish-speaking learners. The paper outlines the platform's pedagogical rationale and describes the methodological principles guiding the design and development of its activities, as well as its distinctive features and potential use as a complement to formal instruction. It further examines how generative artificial intelligence (GenAI) has been selectively integrated into the design of specific activity types, while emphasizing the central role of teaching experience in addressing the needs of Spanish-speaking learners, particularly through activity design and in-task feedback. Finally, the paper discusses future development plans and ongoing challenges, highlighting the potential of this initiative to inspire other CFL teachers interested in creating similar digital learning resources.

摘要： 在数字化转型重塑语言学习格局的背景下，兼具创新与包容的教学方法尤显关键。本文介绍了巴塞罗那自治大学（Universitat Autònoma de Barcelona）教师团队开发的 e·Chinese Plus 综合开源学习平台项目。该项目基于 Moodle 系统，面向西班牙语为母语的中文学习者，旨在提升他们的中文习得效果。本文阐述了平台的教学理念、练习设计与开发的方法论原则，展示了其特色功能及作为课堂补充工具的应用潜力，并探讨了生成式人工智能（GenAI）在特定练习活动设计中的针对性整合，同时强调了教学经验在回应西班牙语母语学习者需求中的核心作用，尤其是在练习设计与任务内反馈方面。最后，本文概述了平台的未来发展规划与持续建设所面临的挑战，以期为有意开发类似中文数字化学习资源的对外汉语教师提供参考借鉴。

Keywords: Chinese as a foreign language; technology-enhanced language learning; generative artificial intelligence; open education resources

关键词: 对外汉语，技术增强语言学习，生成式人工智能, 开放教育资源

1. Introduction

This paper introduces e·Chinese Plus[1] (Casas-Tost et al., 2024-26), a Moodle-based, open-access platform offering interactive digital activities with automated correction and feedback. The platform provides a wide range of activities to develop both receptive and productive language skills and is specifically designed to meet the needs of Spanish-speaking learners of Chinese as a foreign language (CFL). This targeting is most evident in pronunciation, grammar, and translation activities, and, specifically, in the detailed feedback, as explained in this article. More broadly, it seeks to democratize access to high-quality online CFL resources and expand opportunities for active learning, primarily targeting Spanish speakers while remaining accessible to Chinese-language learners worldwide, including those from less-resourced language communities. By promoting inclusive access to language learning resources, the platform aims to empower learners from diverse backgrounds, regardless of their instructors' digital literacy levels—an approach that aligns with the principles of the Universal Design for Learning framework (CAST, 2024), which advocates flexible educational environments that accommodate individual learning.

The development of the platform coincided with the proliferation of generative artificial intelligence (GenAI), which has been integrated into the design of many of the activities. This paper aims to present the platform, its rationale, its specificities—especially considering its main target users—and its potential application as a supplement to formal teaching. Special emphasis is placed on the methodology used to create the activities and the role of GenAI in their design, to offer an example of how to effectively incorporate GenAI into CFL instruction—particularly relevant at a time when language education is actively exploring GenAI-enhanced tools and methods (see Casas-Tost et al., under review, on the specific case of CFL).

2. The Rationale for e·Chinese Plus

According to official figures, "[a]s of May 2023, more than 180 countries have conducted Chinese-language teaching programs, and 81 countries have incorporated the language into their national education system" (Xinhua, 2024). This global interest in the study of Chinese is also evident in Spain, where the number of HSK exam candidates has increased significantly, placing the country among the top five in Europe in recent years (Chen et al., 2021). This trend has been accompanied by a notable rise in the availability of digital resources for learning Chinese.

[1] c.f. https://dtieao.uab.cat/gelea2lt/echineseplus/

While this expansion may appear encouraging, it is well established that not all digital resources are of equal quality, nor do they adapt equally well to different teaching approaches or learning styles. In the specific context of open educational resources (OER) for CFL, Zhang (2022) emphasizes the importance of understanding which resources improve specific language skills, how learners engage with them, and how teachers can integrate preferred materials into curricula to optimize learning outcomes. In this regard, Zhang (2022: 28) also stresses the need to be aware of what resources are currently available and to have quality criteria for assessing their usefulness.

To address this need, in 2021, a team of language teachers and researchers from the Universitat Autònoma de Barcelona (UAB) created e·Chinese Tools (Rovira-Esteva et al., 2021–2026), an open-access database of digital resources for learning Chinese, which now contains over 450 resources. This previous project allowed us to conduct a prospective analysis of the kind of digital resources currently available. A preliminary analysis of the database's content (see Rovira-Esteva et al., 2022) revealed, on the one hand, that vocabulary is the most frequently addressed skill—in fact, it is particularly prominent on Instagram, as Rovira-Esteva & Vargas-Urpí (2024) note— while listening, cultural knowledge, and graphemic skills also feature heavily. On the other hand, writing remains underrepresented due to the difficulty of providing automated correction.

This analysis also identified several shortcomings among the current available resources: first, a limited number of resources tailored to Spanish-speaking learners, who remain underrepresented in the global Chinese language education arena, where English is still the predominant language of instruction. Second, we also noted an imbalance between receptive and productive skills, with a greater focus on the former. In fact, while multimodality is well-featured (most digital resources offer video content), only around 20% involve active practice, mainly via apps, revealing a shortage of opportunities to develop more productive skills. This pattern is further confirmed by studies on digital resources for specific skills, such as Casas-Tost's (2026) analysis of tools for pronunciation practice or Gay-Punzano & Vargas-Urpí (2025) about cultural competence. Third, this preliminary analysis also revealed reduced accessibility because many of the highest-quality resources are not free.

The predominance of learning apps as a means of delivering active practice also entails certain limitations. Wang (2024) analyzed four widely used applications (Duolingo, LingoDeer, SuperChinese, and HelloChinese) and observed that their fixed learning paths align well with microlearning principles and the needs of self-directed learners, but offer limited flexibility for students who wish to use them to complement formal instruction. Moreover, in light of ongoing debates on smartphone addiction, particularly among young people, and its potential impact on learning (see, for example, the meta-analysis by Sunday et al., 2021), the exclusive availability of some learning resources via mobile devices may itself be considered a drawback. Finally, most language-learning apps rely on a "freemium" model, which restricts free access to only part of the available content and functionalities, thereby preventing learners from fully benefiting from the app's pedagogical affordances.

Other studies of the application of digital resources in the teaching and learning of CFL have focused on their effectiveness or on the analysis of specific tools. For example, regarding character learning, Mason & Zhang (2017), based on a survey of 140 CFL learners, observed that nearly all (94%) used at least one mobile app to support their character learning. Pleco was the most relied-on app, even though not all its features were equally used. Sun (2022) demonstrated that the use of Wordwall in teaching Chinese characters at the YCT1 level significantly improved students' character recognition and comprehension and had a positive effect on learner motivation. Ma (2024) studied Quizlet's impact on 60 CFL students, finding significant gains in listening, speaking, and reading skills but not in writing, aligning with previous research; and that, importantly, Quizlet enhances student engagement through self-paced and tailored practice. Finally, Gay-Punzano & Vargas-Urpí (2025) examined how digital resources address cultural and intercultural competences and noted biases, recurring topics, and a mostly passive approach.

Previous studies have been valuable not only in confirming the affordances and strengths of existing digital tools, but also in identifying their limitations, which informed the development of our new digital resource. In this context, the main rationale for creating e·Chinese Plus was to design a platform that specifically addresses the needs of Spanish-speaking learners. This includes, for example, activities targeting linguistic features that are particularly challenging for this group, as well as feedback grounded in contrastive linguistics to enhance pedagogical relevance. In addition, the platform was conceived as fully open-access and designed to allow learners to select activities that best suit their individual learning needs, deliberately avoiding fixed or predetermined learning paths. Its design as a platform rather than a mobile app allows it to be used across multiple devices. The following sections describe the platform in greater detail.

3. e·Chinese Plus Features

e·Chinese Plus is a Moodle-based, open-access platform that provides activities for practicing CFL. Teachers and students who are familiar with Moodle environments will therefore recognize the types of activities it offers. As already mentioned, the platform is not conceived as a course but as a bank of exercises that can supplement formal learning undertaken through regular courses or informal learning for self-directed learners. Activities are structured into four Moodle courses corresponding to proficiency levels according to the Common European Framework of Reference for Languages (CEFR), ranging from basic (A1) to Intermediate (B2), as B2 is typically the highest level of Chinese taught at universities in Spain and, in general, at other institutions. Learners first enroll in the course or courses that match their level or learning needs (see Figure 1).

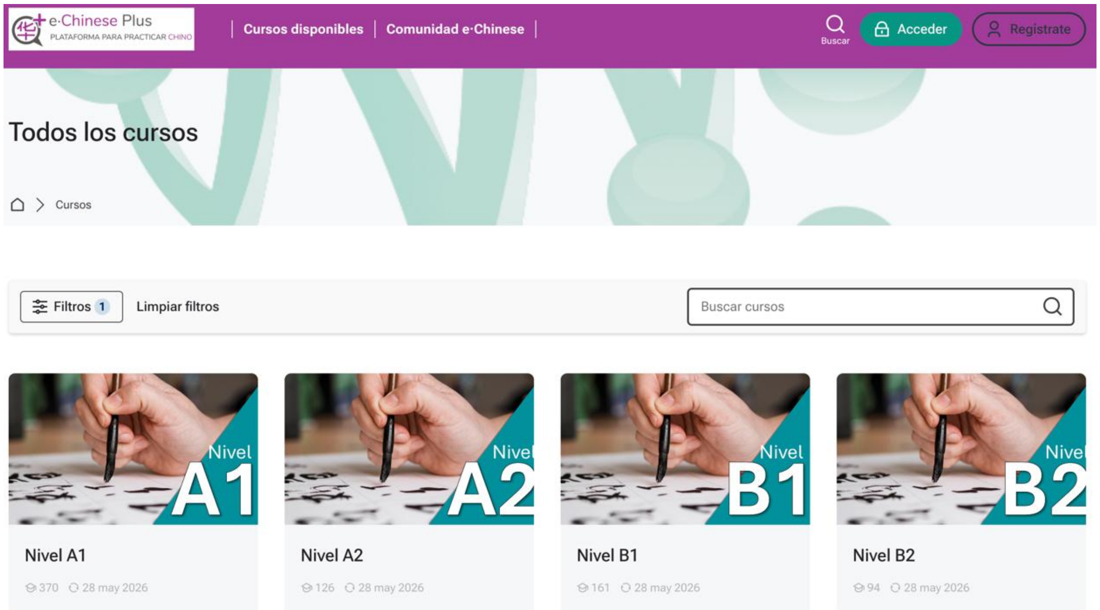


Figure 1 Screenshot of the platform

Within each course, they can then use two filters to select activities: language skill and topic (see Figure 2). Students can complete any activity as many times as they wish, in whatever order they prefer, and track their progress in each skill. This structure allows learners to focus on the level most relevant to them while retaining the flexibility to choose activities based on their immediate pedagogical needs.

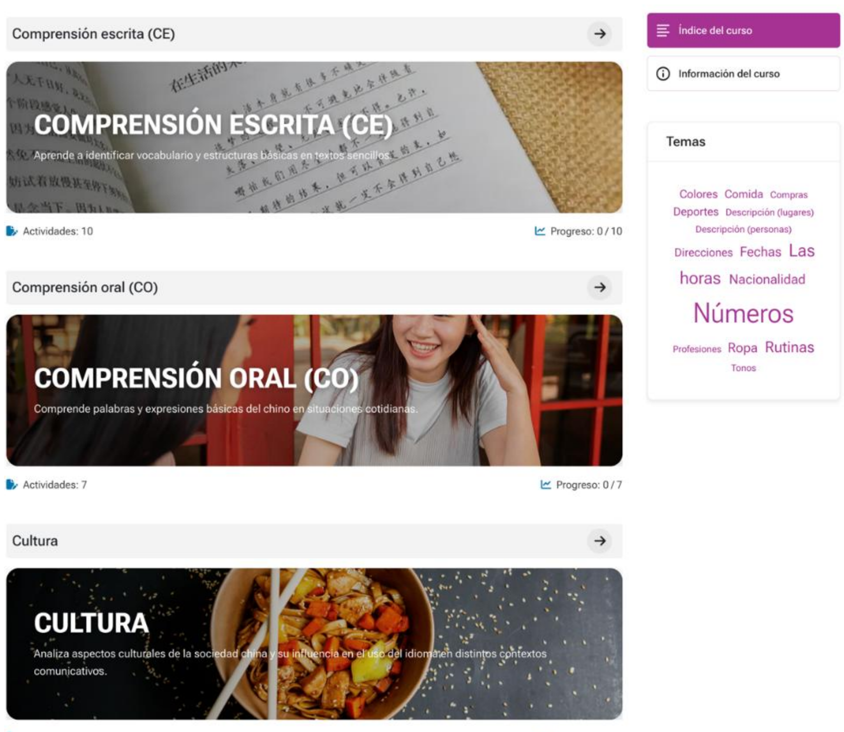


Figure 2 Screenshot of the skill and topic filters

The skill filter covers a wide range of areas, including reading, listening, speaking, vocabulary, grammar, pronunciation, translation, cultural awareness, and knowledge of Chinese characters. Thus, the platform goes beyond the four traditional language skills to offer exercises that target areas particularly relevant to Chinese language learners. Some, such as character writing and the use of the pinyin transcription system to practice pronunciation, are unique to the Chinese writing system and require specific attention. Many activities, such as translation, pronunciation, grammar and culture-focused tasks, explicitly take into account contrastive aspects between Chinese and Spanish (e.g. how to translate Chinese gender indeterminacy into a gender-marked language such as Spanish). In response to user feedback, a new tag for sequenced activities has recently been introduced to create short learning paths composed of several activities that, taken together, integrate diverse skills around a single topic.

Rather than a technical limitation, this design choice reflects a deliberate prioritization of usability and pedagogical relevance. Although many activities naturally integrate multiple skills (e.g., listening, reading, speaking, and vocabulary), assigning one principal tag facilitates more precise filtering and enables users to quickly identify activities that best match their immediate learning needs. To this end, the project team—comprising seven Chinese language teachers and researchers—engaged in thorough discussion to determine and prioritize the most appropriate tags, with the aim of enhancing both the usability of the platform and the practical usefulness of each activity for its intended users.

The topic filter allows activities to be grouped according to theme. The list of topics is open-ended and includes family, food, nationalities, sports, etc. This tag is optional, as not all activities can be meaningfully assigned a topic (e.g. pronunciation or character-writing activities). When choosing one of them, the platform offers multiple activities revolving around the same topic. For example, the topic *color* includes four activities. At level A1, there is a vocabulary activity in which students must select the words that do not correspond to a given color, making it ideal for visually oriented learners (Figure 3), and a speaking activity in which learners name clothing items, state their color, and supply the appropriate measure word (Figure 4), combining visual and auditory modes. At level A2, one activity requires students to listen to descriptions of combinations of two colors and identify the resulting color (Figure 5, while another asks them to read the names of two colors and say the resulting color aloud (Figure 6), thus engaging both read/write and auditory learning preferences. These variations support differentiated instruction and multimodal learning, while also adapting to both teaching and learning needs.

Figure 3. Example of vocabulary activity (reading)

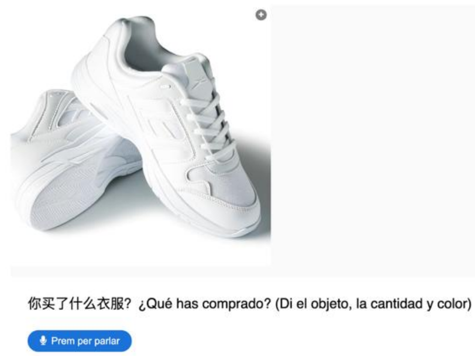


Figure 4. Example of vocabulary activity (speaking)

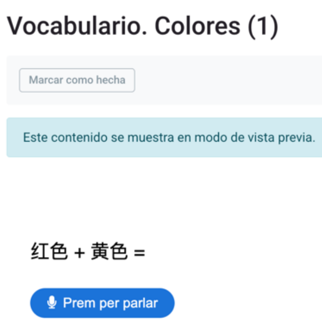


Figure 5 Example of vocabulary activity (reading + speaking)

Figure 6 Example of vocabulary activity (listening + reading)

In addition, each level includes two forums: a student forum, where learners enrolled in the same course can interact with one another, and a dedicated forum through which they can contact the platform creators. Beyond the level-specific courses, the platform also includes a general community space with several forums where all members, regardless of their proficiency level, can exchange ideas, share information about scholarships and learning opportunities, ask questions, and contribute resources of potential interest to the wider e·Chinese Plus community.

3.1. Workflow and methodology for the design of activities

The creation of the activities follows a standardized workflow designed to ensure pedagogical consistency, technical reliability, and overall quality control across the platform. First, a team member creates an activity. Next, it is peer-reviewed by at least two fellow teachers, who provide feedback on all aspects of it (content, level, in-activity feedback, and technical considerations) using a spreadsheet. The activity is then revised, based on the reviewers' feedback, by its creator, who subsequently publishes it on the platform. A student with a B2 proficiency level pilots the activity and provides further feedback, which is used to finetune it, if necessary. If the two reviewers are unsure about anything, a third team member reviews the activity. Any disagreements are discussed at the team's periodic meetings. The role of the project manager has organically grown from creating activities and handling the technical aspects of the platform to include reviewing all activities to ensure consistency in their format and overall quality.

In addition to more traditional computer-assisted language learning technologies, the team has used GenAI systems when creating activities. Following Tolstykh & Oshchepkova's (2024) classification, GenAI has been applied in four main areas, each with specific advantages and constraints:

1. Text generation. We have used GenAI systems such as ChatGPT 4.0 or above to generate, correct, and modify or refine texts for reading and listening comprehension activities. This accelerates the drafting process and provides varied linguistic input. However, the generated texts sometimes contain inaccuracies or unsuitable constructions or vocabulary, so all outputs are systematically reviewed, edited and approved by native-speaking team members, in keeping with Tolstykh & Oshchepkova (2024).
2. Image generation. Many of the images used in vocabulary, listening and speaking activities have been generated using the Dall·E AI system (integrated into ChatGPT 4.0 and above). This system has proven useful for certain images (e.g. pieces of clothing in Figure 4), but adds more details than required in some cases, and produces images unsuitable for the purpose of the exercises in others (e.g. exercises in which students have to describe locations). These inconsistencies often made images unsuitable without substantial human intervention. For this reason, image creation has relied on a hybrid approach that combines GenAI output with the work of a professional illustrator to ensure accuracy, clarity and pedagogical adequacy.
3. Text-to-speech production. Synthetic voices allow us to quickly produce audio materials. We create a text or use a GenAI system to create a text, which is then proofread and corrected by native Chinese team members before being entered into a text-to-speech tool (e.g. TTS Maker)[2] to create an audio file with synthetic voices. This process is practically identical to that described in Tolstykh & Oshchepkova (2024: 12). Still, human recordings remain preferable for pronunciation-based tasks, where accuracy, prosody and speed are critical.
4. Oral speech analysis. H5P's "Speak the Words" feature uses the Annyang speech-recognition engine to convert learner speech into text and match it against predefined answers, a process that requires extensive teacher input to anticipate multiple correct variants. This enables a degree of automated evaluation, but the system's accuracy varies depending on accent, prosody and speed—limitations particularly relevant for CFL learners. Since Annyang is a general-purpose Automatic Speech Recognition (ASR) system not trained on Mandarin spoken by second-language learners, its baseline performance is uneven. Consequently, no activities requiring fine-grained tone discrimination for very short words have been created, as the system is not sufficiently reliable for such tasks. ASR-based activities are, therefore, treated as guided practice rather than as high-stakes assessment.

[2] c.f., https://ttsmaker.com/

Although GenAI can, in theory, produce entire activities (texts, questions and distractors), our experience—and recent research—show that this approach is pedagogically insufficient. Automatically generated questions frequently rely on superficial lexical overlap rather than deeper comprehension (Casas-Tost et al., under review; Thornburn, 2024). In CFL contexts, GenAI-paraphrased questions may also introduce vocabulary beyond students' level. Despite experimenting with enhanced prompting and chain-of-thought techniques (Sánchez-Gijón & Palenzuela-Badiola, 2023), this method rarely saved time and often produced materials misaligned with course objectives. For this reason, GenAI is being used as a support tool that provides drafts which teachers then refine, restructure or replace, depending on pedagogical goals, rather than as an autonomous activity generator. This hybrid model maintains teacher control over difficulty, linguistic accuracy, and CFL-specific sequencing.

GenAI is therefore used strategically where it adds value (e.g., drafting, audio, visual assets) and avoided where it may compromise pedagogical accuracy, such as in grammar explanations or productive-skills tasks requiring nuanced evaluation. This human-in-the-loop approach maximizes GenAI's effectiveness while mitigating its current limitations. The result is that most of the platform's activities have been fully created by members of the research team, while 24% have involved GenAI input for the creation of text, audio or images. Furthermore, all in-activity feedback has been prepared by team members. Our experience shows that GenAI's grammar explanations are not always accurate or relevant. Furthermore, our knowledge of mistakes frequently made by Spanish-speaking learners of Chinese, as well as the small body of research analyzing this topic (e.g. Liu, 2019 and Wang, 2025) has been crucial to providing tailored feedback in many exercises (e.g. pronunciation, grammar, translation, etc.).

3.2. Technical aspects of activities

Activities are built using either H5P or Moodle questionnaires. Formats include multiple choice, drag-and-drop, true/false questions, matching tasks, memory games, image selection, image pairing, gap-filling, paragraph sorting, short-answer writing and speaking prompts. In designing the platform, we have intentionally diversified both the formats and the types of input used in order to move beyond the most traditional exercise types (although these are also included, as they are still very useful). Our aim has been to create multimodal activities that respond to the needs of different learner profiles and make practice more challenging and engaging, as shown in Figures 3 to 6.

All activities must provide automated correction and feedback, and due to technical constraints, fully open-ended tasks (e.g., free description or extended writing) are not feasible. This is particularly challenging for speaking and writing tasks, where teachers must anticipate and encode all acceptable correct answers. For instance, the writing activity "One day in Xiao Ming's life" requires multiple formulations of "Xiao Ming eats alone" to be listed as correct, such as:[3]

[3] The full stop at the end of each sentence is omitted because we identified issues with punctuation in H5P open-ended responses. The activity's instructions warn students not to use full stops.

- 小明一个人吃早饭
- 小明一个人吃
- 他一个人吃早饭
- 一个人吃
- 小明一个人吃早餐
- 他一个人吃早餐
- 一个人吃早餐

Speaking activities face similar constraints, as Annyang—the speech recognition engine used by H5P—matches learner utterances to predefined acceptable answers. As with writing tasks, teachers must anticipate and enter a comprehensive list of correct variations in advance. This is reflected in Figure 7, taken from an activity in which students have to say the time shown on the clock; answers that the system has interpreted as incorrect are displayed in red, while all the possible correct answers entered by the teacher appear in green.

Figure 7 Example of a speaking activity and its feedback

The technical limitations involved—particularly the dependence on predefined answers in open-ended tasks—significantly constrain the quantity and complexity of the speaking and writing activities available on the platform. Consequently, the overall number of these activities is markedly lower than that of other types of tasks that are not subject to the same restrictions.

3.3. Feedback provided in the activities

Since the platform is designed to foster learner autonomy and the corrections and feedback the activities provide are automated, it is essential that the feedback be of very high quality. Accordingly, in addition to indicating whether or not answers are correct, detailed written feedback in Spanish is provided when they are incorrect. This feedback aims to explain, as clearly as possible, why an answer is incorrect or, through hints, to guide students towards the correct response. The latter approach can be seen in Figure 8, where the learner is asked if they missed an aspirated sound in any of a word's syllables.

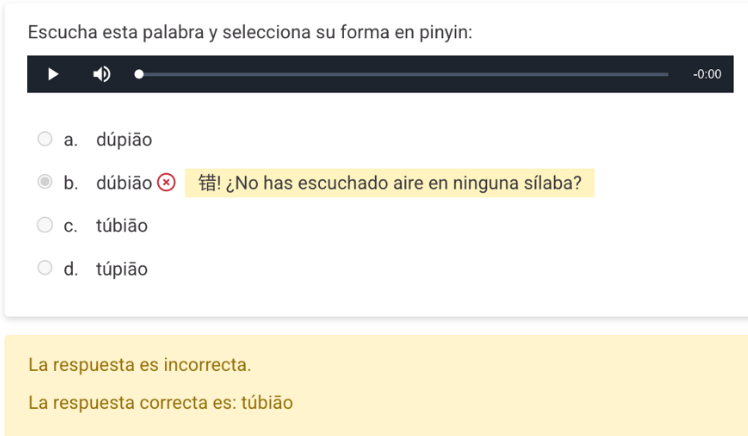


Figure 8 Example of feedback in a pronunciation activity

Feedback is also crucial in grammar activities, where offering clear explanations or relevant examples helps students understand language rules and patterns, correct their mistakes, and reinforce their learning. For instance, as shown in Figure 9, the feedback clarifies which particle should be used and the reason why.

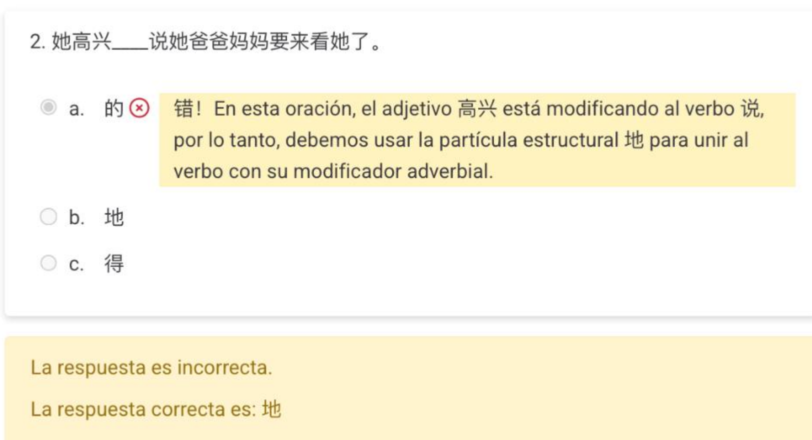


Figure 9 Example of feedback in a grammar activity

Translations into Spanish and detailed grammar explanations are also used in the feedback for certain complex activities. For instance, in the example of Figure 10, students need to place words in the correct order to form complete sentences in which interrogative words are used with an absolute meaning. The feedback provides the translation of the intended sentence into Spanish and emphasizes that "什么时候" does not mean "when" in this case, but "always", due to its combination with the adverb "都". In addition, the feedback draws learners' attention to the prototypical sentence structure associated with this construction.

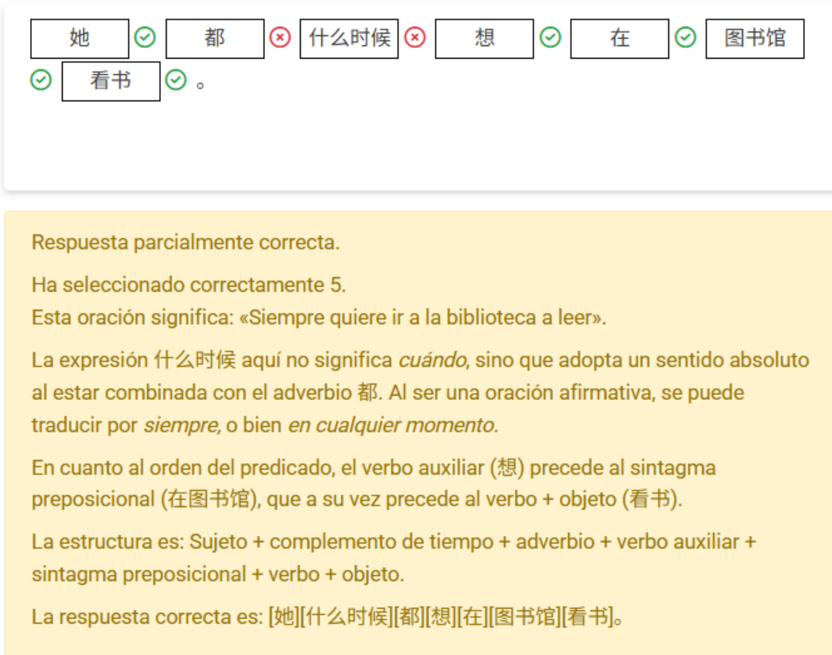


Figure 10 Example of detailed feedback in a B1 grammar activity

From a technical perspective, activities created with H5P tend to be more visual and interactive, but the tool does not support highly elaborated, individualized feedback. Moodle questionnaires (for example, multiple-choice or true/false formats) are therefore preferred when detailed, explanatory feedback is required. In cases where detailed feedback cannot be provided due to the nature and format of the activity (e.g. matching exercises in listening tasks), alternative options are implemented (e.g. providing an audio transcript).

4. Activities currently available on e·Chinese Plus

e·Chinese Plus is an ongoing project that is frequently updated. At the time of writing (February 2026), it offers approximately 170 activities covering four levels and ten skills, including seven sequenced activities, and has more than 525 registered users.

Figure 11 shows a breakdown of the platform's activities by level. Just over half the activities belong to levels A1 and A2. Level A1 has the highest number of activities, accounting for 38% of the total, while level B2 has the lowest, representing only 16%. This slight imbalance somewhat reflects the distribution of students by level, as well as the fact that the team's members teach more lower-level classes than intermediate ones. Over time, however, the goal is to offer a greater balance of activities across levels.

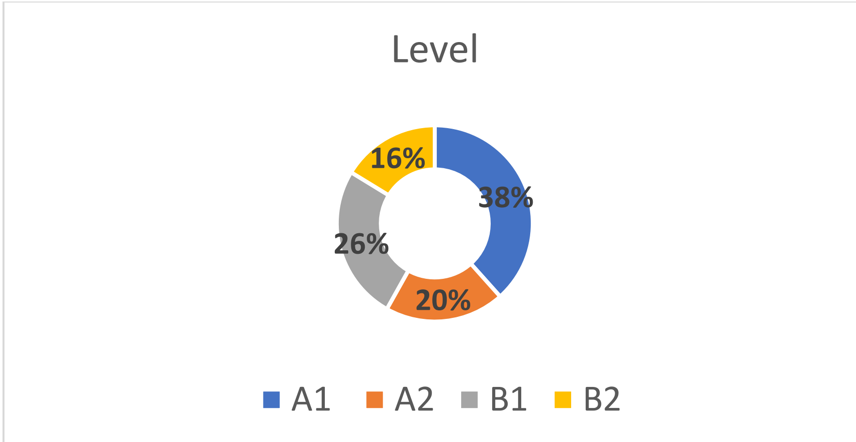


Figure 11. Activities by level

Figure 12 illustrates the distribution of activities by skill. Reading and vocabulary are the skills with the highest number of activities (30), followed by listening (22), and grammar (21). In contrast, there are fewer activities that focus on skills like translation (7), speaking and writing (6 each, making them the least represented skills at the moment). There are currently only seven sequenced activities, as this is a newly created category.

This distribution highlights a clear emphasis on receptive skills, such as reading and listening, along with vocabulary and grammar, which are foundational for language acquisition and well suited to closed interactive activities with feedback. In contrast, productive skills, such as writing and speaking, feature less prominently. This is partly due to the greater difficulty entailed in designing closed activities that automatically provide corrections and meaningful feedback for productive skills, owing to the technical limitations of current formats. We are working to address this imbalance by creating more activities for developing productive skills. Pronunciation, characters, and culture—skills particularly relevant to Chinese—fall in the middle of the distribution and are developed through activities that approach them from multiple perspectives.

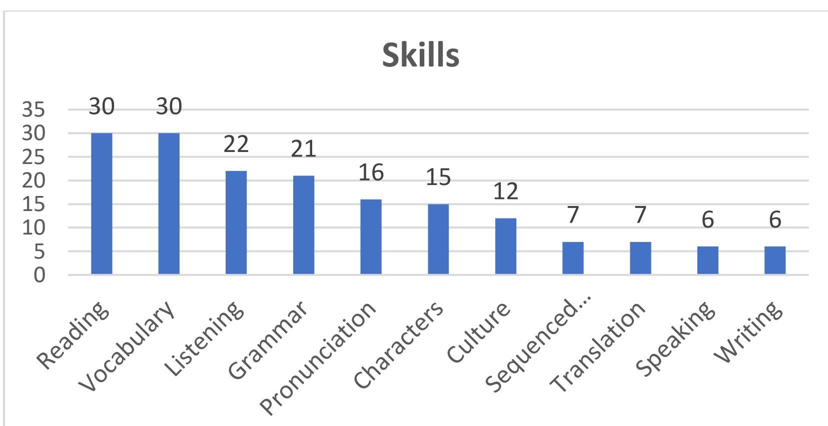


Figure 12. Activities by language skill

There is a noticeable correlation between proficiency level and the types of skills targeted in the activities, as can be seen in Figure 13. In general, receptive and form-focused skills are more evenly distributed across levels, while productive skills tend to be clustered at the beginner levels. More specifically, productive skills, such as speaking and writing, are primarily concentrated at level A1, where it is easier to anticipate all possible correct responses. As learners' range of vocabulary and grammatical structures grows, accounting for all possibilities becomes increasingly challenging.

Nonetheless, we are addressing the current imbalance with the aim of achieving a more even distribution of activities across levels and skills, taking the specific pedagogical needs of each proficiency level into account at all times. For instance, pronunciation and character writing are especially important at the beginner levels and require focused practice in the early stages. However, these skills remain relevant at higher levels, where they can be approached through different types of activities that suit learners' evolving needs and abilities. Accordingly, we have also developed activities targeting pronunciation and character writing for more advanced learners, such as those at levels B1 and B2. For example, the platform currently has two B2 level activities for the skills in question: one involving a tongue twister (for pronunciation) and another posing a riddle about writing characters.

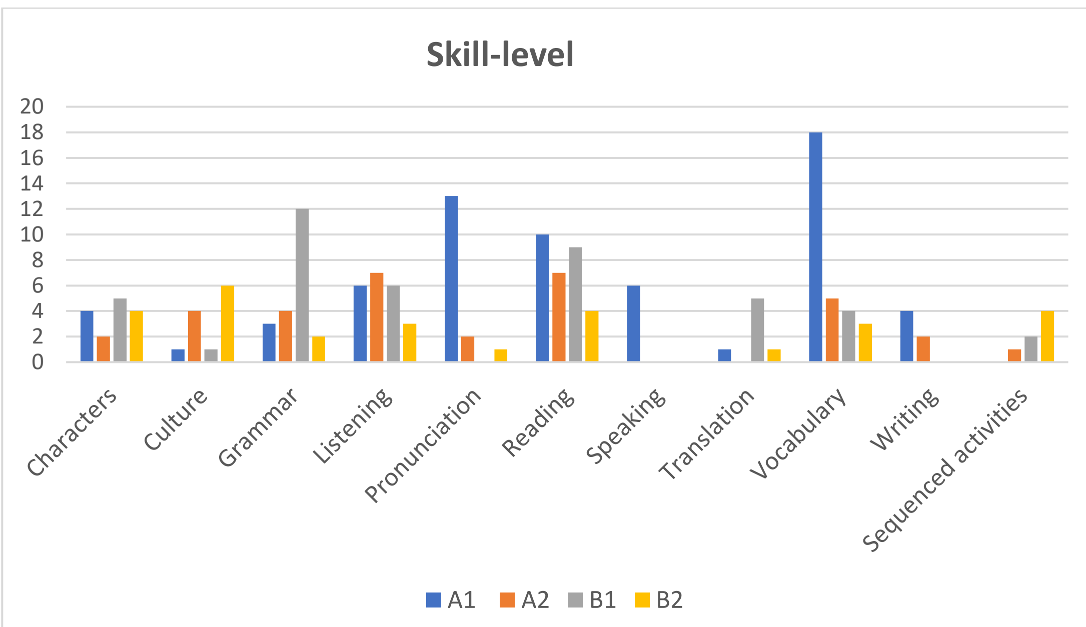


Figure 13 Activities by language skill and proficiency level

Finally, as regards learning styles, it should come as no surprise that 95% of the activities include text, requiring students to either read or write. However, it is worth noting that 60% of the activities combine two or even three different input types. Thus, the majority of the activities can be considered multimodal, catering for a broader range of learning preferences.

5. Potential applications of the platform

e·Chinese Plus is primarily intended for students who wish to supplement their regular Chinese classes, but it can also be seamlessly integrated into classroom-based instruction. After completing a simple, free registration process, learners gain full access to all activities on the platform and may complete them in any order, as many times as they wish. This flexibility allows students to incorporate practice into their daily routines and to focus on the skills, topics or activity types that best match their current needs.

The platform interface is in Spanish, and activity instructions are also provided in Spanish; in intermediate-level activities, instructions are additionally offered in Chinese. Feedback is available exclusively in Spanish and is particularly rich in grammar and receptive pronunciation activities, where learners are asked to discriminate between similar sounds. In these cases, feedback goes beyond indicating whether an answer is correct or incorrect, often pointing to specific segments or features that may have caused difficulty. By contrast, feedback in other activities, such as speaking, is less detailed, as the system can only display the target answers and the words interpreted by the ASR engine.

In addition to individual student use, teachers can incorporate e·Chinese Plus into their courses by recommending or assigning specific activities to be completed in class or as homework, either to reinforce material covered in lessons or to provide targeted extra practice on particular skills or topics. For example, if a teacher is working at the A1 level (beginner level) and students are learning how to tell the time in Chinese, four activities can be recommended as homework tasks: a vocabulary drag-and-drop exercise (Figure 14); a reading comprehension activity in which students sequence a story based on a video (Figure 15); a listening exercise requiring students to select the correct time based on audio input (Figure 16); and an oral production activity in which students state the time according to visual prompts (Figure 17). Together, these activities illustrate the platform's multimodal approach and its attention to different learning styles, as input is provided through a variety of formats—including images of analogue and digital clocks, video, audio, and written text—and learners engage with diverse task types.

Figure 14 Vocabulary exercise to practice telling the time

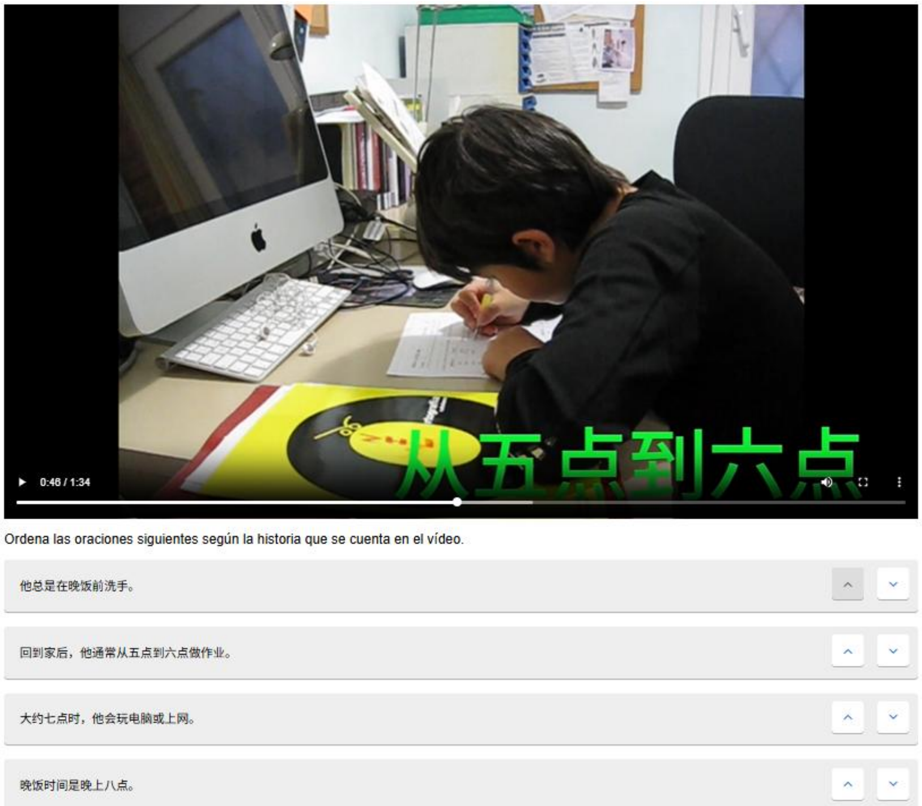


Figure 15 Reading comprehension activity to practice telling the time

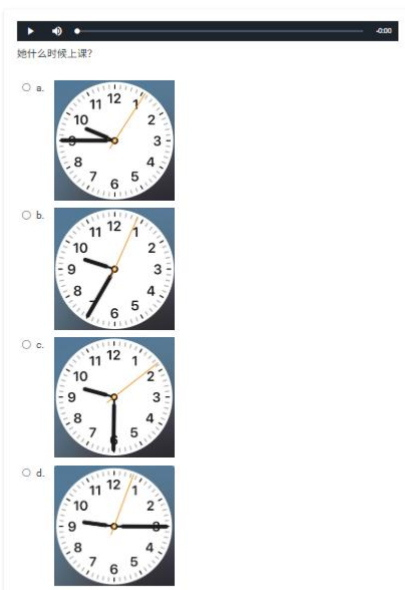


Figure 16. Listening to practice telling the time

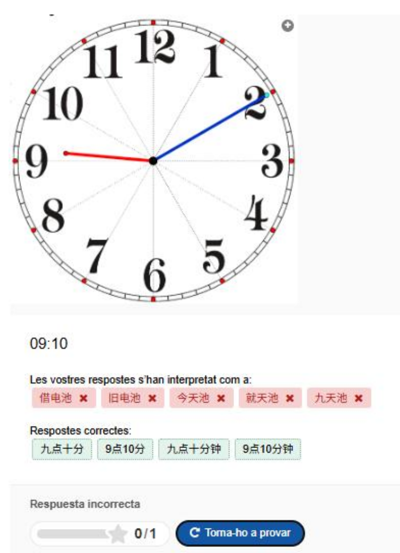


Figure 17. Oral expression activity to say the time

Finally, the platform may also serve as a source of inspiration for teachers who wish to design their own activities using Moodle-based environments. In particular, activities aimed at developing cultural competence remain relatively scarce in existing resources, which tend to adopt a more passive approach (Gay-Punzano & Vargas-Urpí, 2025). By contrast, e·Chinese Plus increasingly incorporates tasks that promote active engagement and allow learners to apply their skills. Figure 18 illustrates an example of a culture-focused activity in which students are required to recognize hand gestures used to represent numbers in Chinese.

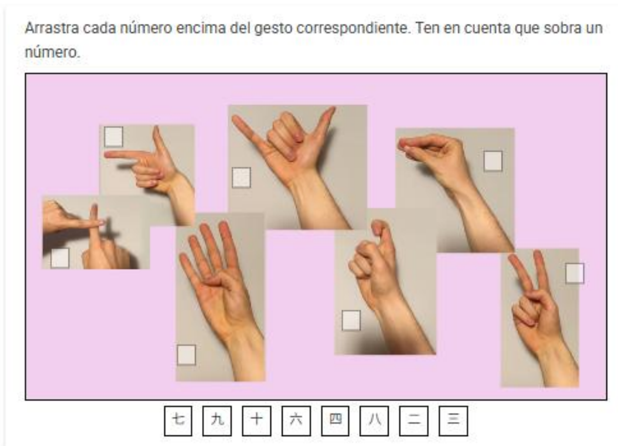


Figure 18 Example of a culture activity

Taken together, these design choices align closely with the principles of Universal Design for Learning, particularly in their emphasis on providing multiple means of representation, engagement, and expression. By offering varied input modalities, flexible task types, and non-linear learning paths, the platform seeks to accommodate diverse learner profiles and preferences while supporting learner autonomy. In this way, e·Chinese Plus not only addresses identified pedagogical gaps in existing digital resources but also adopts an inclusive design framework that facilitates access and meaningful learning for Spanish-speaking students, while simultaneously offering a valuable repository of pedagogical ideas for CFL teachers.

6. Conclusions and future directions

e·Chinese Plus addresses a critical need in the landscape of Chinese language learning for Spanish-speaking students. The project has created a wide range of engaging activities that accommodate individual learning differences, aligning with the Universal Design for Learning framework and differentiated instruction, and incorporating new technologies such as GenAI. Our ongoing review of existing digital resources, together with a cyclical evaluation of the platform's own strengths and weaknesses—particularly with regard to underrepresented activity types or skills—enables us to continuously refine the platform while responding to identified pedagogical needs.

The platform is designed to supplement formal instruction and, thanks to its multimodal, multiformat approach, to better adapt to students' learning styles. In the specific context of CFL, Zhang (2022) emphasizes the role of open educational resources (OER) as valuable complements to formal instruction, stressing the importance of knowing which tools enhance specific skills, how learners use them, and how teachers can integrate them into curricula. These considerations have informed the development of e·Chinese Plus. Furthermore, since the activities do not require teacher supervision, students can easily integrate them into their routines, completing them as often as they wish. This flexibility supports self-paced learning and targeted practice, as also observed by Ma (2024) in her study on the advantages of Quizlet.

e·Chinese Plus overcomes a limitation of many CFL apps, that of providing only fixed learning paths in which students cannot repeat activities or freely select the type of exercises they need, as pointed out by Wang (2024). The platform's clear classification of activities by proficiency level, skill, and topic enhances its user-friendliness and adaptability to learners' evolving needs. Studies such as the review by Lyu & Qi (2020: 158) have identified a gap between in-class and out-of-class learning as a major issue in technology integration for CFL. However, e·Chinese Plus addresses this by offering a structure and filtering system that enables teachers to create and select complementary tasks and students to independently select activities appropriate to each stage of their learning process.

Our current aim is to create more activities to fill current gaps in skills, topics, and levels and achieve a more balanced distribution across levels and skills. Additionally, we want to explore creating more sequenced activities that serve as learning paths, with multiple activities centered on a single topic. We believe such paths will guide students through activities thoughtfully organized around sound pedagogical principles, while also allowing many of the individual activities within these learning paths to be used independently, depending on learners' needs and learning contexts. Future work will also focus on collecting feedback from learners and instructors, and gathering longitudinal data to better assess content quality, user experience and learning effectiveness, which are common indicators to assess digital resources (Zhang et al., 2025). We also aim to explore new GenAI-supported functionalities—particularly for productive skills—once they become reliably compatible with Moodle's pedagogical and technical constraints. At the time of writing, the integration of AI into Moodle is primarily achieved through plugins that support text and image generation and text summarization. However, future developments may make it possible to incorporate AI-based correction tools designed to provide feedback on production activities. In such cases, thorough prior testing will be essential to ensure that the feedback is pedagogically appropriate and does not automatically suggest corrections or alternatives that exceed learners' proficiency levels.

Finally, our experience confirms that GenAI integration has expanded our professional competencies and enhanced our digital literacy as teachers, echoing the insights of Tolstykh & Oshchepkova (2024). GenAI has enabled us to design activities that go beyond traditional formats, such as filling in gaps or answering multiple-choice questions. However, the activities currently available on e·Chinese Plus are not as personalized, flexible, or interactive as those generated by GenAI chatbots, a limitation that is particularly relevant in the case of productive skills, which remain the most underrepresented on the platform. At the same time, the limitations of GenAI have reconfirmed the ongoing importance of the human factor in the teaching and learning process.

Acknowledgement: This research has received financial support from the Universitat Autònoma de Barcelona (UAB) and the Fundació Institut Confuci de Barcelona.

The authors wish to express their sincere gratitude to all members of the e·Chinese Plus project for their support, collaboration and insightful input while this project was being carried out. Special thanks are also due to the colleagues and reviewers who generously took the time to read earlier drafts of this article and provide constructive feedback, which contributed to improving the clarity of the final version.

GenAI use disclosure statement: The authors of this work acknowledge having used ChatGPT-4o to improve its readability and language. The authors have subsequently reviewed and edited the text as necessary and take full responsibility for the publication's content.

References

Casas-Tost, H. (2026). Digital resources for learning Chinese pronunciation: e-Chinese Tools as a case study (Recursos digitales para el aprendizaje de la pronunciación del chino: e· ChineseTools como estudio de caso). *Bellaterra Journal of Teaching and Learning Languages and Literature, 19* (1) e1390. https://doi.org/10.5565/rev/jtl3.1390

Casas-Tost, H., González-Torrents, I., Rovira-Esteva, S., & Vargas-Urpí, M. (under review). Using GAI for reading comprehension activities in the teaching and learning of Chinese as an additional language.

Casas-Tost, H., Gay-Punzano, L., Guo, X., Nie, L., Paoliello, A., Rovira-Esteva, S., & Vargas-Urpí, M. (2024-2026). e·Chinese Plus: Plataforma para practicar chino. https://dtieao.uab.cat/gelea2lt/echineseplus

CAST (2024). Universal Design for Learning Guidelines version 3.0. https://udlguidelines.cast.org

Chen, C.; Li, Q., & Yang, Q. (2021). Chinese language education in Spain: Current landscape and future directions (西班牙中文教育发展现状与前瞻). *Tianjin Shifan Daxue Xuebao Shehui Kexueban (天津师范大学学报社会科学版*), 3,16–23.

Gay-Punzano, L., & Vargas-Urpí, M. (2025). The representation of Chinese culture in digital resources to learn Chinese as an additional language (Representaciones culturales en los recursos digitales para aprender chino como lengua adicional). *Revista Internacional de Lenguas Extranjeras. International Journal of Foreign Languages*, 24, 255–279.https://doi.org/10.17345/rile24.4098

Liu, S. (2019). *Main problems in acquiring Chinese as a foreign language based on an analysis of errors made by Spanish students.* [Liu, S. (2019). *Principales problemas en la adquisición del chino como lengua extranjera a partir del análisis de errores de estudiantes españoles.*] [Doctoral dissertation, Universitat Autònoma de Barcelona]. http://hdl.handle.net/10803/670110

Lyu, B., & Qi, X. (2020). A review of research on technology-assisted teaching and learning of Chinese as a second or foreign language from 2008 to 2018. *Frontiers of Education in China*, *15*(1), 142–163.

Ma, X. (2024). Enhancing language skills and student engagement: investigating the impact of Quizlet in teaching Chinese as a foreign language. *Language Testing in Asia* 14, 5 https://doi.org/10.1186/s40468-024-00275-3

Mason, A., & Zhang, W. (2017). An exploration of the use of mobile applications to support the learning of Chinese characters employed by students of Chinese as a foreign language. In Q. Kan & S. Bax (Eds), *Beyond the language classroom: researching MOOCs and other innovations* (pp. 99–112). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.mooc2016.674

Rovira-Esteva, S., Vargas-Urpí, M; Casas-Tost, H., & Paoliello, A. (2021–2026). *e·Chinese Tools: Tecnologías para la enseñanza y el aprendizaje del chino*. https://dtieao.uab.cat/txicc/echinese. https://doi.org.10.5565/ddd.uab.cat/259978 (v.5)

Rovira-Esteva, S., Vargas-Urpí, M., Casas-Tost, H., & Paoliello, A. (2022). e·Chinese Tools: The Pan Gu of digital resources for teaching and learning Chinese as a foreign language (e·Chinese Tools: el Pan Gu de los recursos digitales para el aprendizaje y la enseñanza del chino como lengua extranjera). *Sinología Hispanica. China Studies Review*, *14*(1): 53–78. https://revpubli.unileon.es/index.php/sinologia/article/view/7379

Rovira-Esteva, S., Vargas-Urpí, M. (2024). From the traditional classroom to mobile microlearning: Analysing the potential of Instagram for Chinese language learning (Del aula tradicional al microaprendizaje móvil: análisis del potencial de Instagram para aprender chino). *Sinología Hispanica. China Studies Review*, *16*(1), 1–26. https://doi.org/10.18002/sin.v16i1.8245

Sánchez-Gijón, P., & Palenzuela-Badiola, L. (2023). Analysis and evaluation of ChatGPT-Induced HCI shifts in the digitalised translation process. In C. Orasan, R. Mitkov, G. Corpas Pastor & J. Monti (Eds.), *International Conference Human-informed Translation and Interpreting Technology (HiT-IT 2023). Proceedings* (pp. 227–267). Incoma Ltd. https://doi.org/10.26615/issn.2683-0078.2023_021

Sun, X. (2022). *Activities to teach YCT1 Chinese characters using digital resources: the example of Wordwall*. [Sun, X. (2022). *Actividades para la enseñanza de caracteres chinos de YCT1 con el uso de recursos digitales - tomando el Wordwall como ejemplo.*] [Master's tesis, Universitat Autònoma de Barcelona]. https://ddd.uab.cat/record/265379

Sunday, O. J., Adesope, O. O., & Maarhuis, P. L. (2021). The effects of smartphone addiction on learning: A meta-analysis. *Computers in Human Behavior Reports*, 4, 100114. https://doi.org/10.1016/j.chbr.2021.100114.

Thornburn, R. (2024). Creating Reading Comprehension Activities with AI. *Hong Kong TESOL Blog.* https://hongkongtesol.com/blog/creating-reading-comprehension-activities-ai

Tolstykh, O., & Oshchepkova, T. (2024). "Beyond ChatGPT: roles that artificial intelligence tools can play in an English language classroom". *Discover Artificial Intelligence*, *4* (1), 60. https://doi.org/10.1007/s44163-024-00158-9

Wang, B. (2024). *The Potential of Apps for Microlearning Chinese as a Foreign Language: Four Case Studies*. [Wang, B. (2024). *El potencial de las aplicaciones para microaprendizaje del chino como lengua extranjera: cuatro estudios de caso*.] [Master's tesis, Universitat Autònoma de Barcelona]. https://ddd.uab.cat/record/300063

Wang, Z. (2025). *The use of social media in language teaching and learning: A proposal for written expression in Chinese*. [Wang, Z. (2025). *El uso de las redes sociales en la enseñanza-aprendizaje de lenguas: Una propuesta para la expresión escrita en chino*.] [Doctoral dissertation, Universitat Autònoma de Barcelona]. http://hdl.handle.net/10803/695618

Xinhua. (2024, June 13). *Why the Chinese language is gaining global appeal?* People's Daily Online. http://en.people.cn/n3/2024/0613/c90000-20181008.html

Zhang, S. (2022). Intermediate-level language learners' use of online accessible resources to supplement learning: An exploratory study. *Journal of Technology and Chinese Language Teaching*, *13*(1), 26–45. http://www.tclt.us/journal/abstractshow.php?id=133

Zhang, K., Hou, S., Song, J., & Xiao, R. (2025). Construction of an Evaluation Indicator System for Chinese Learning Apps Based on BERT-LDA. *Journal of Technology and Chinese Language Teaching*, *16*(2), 23-47. http://www.tclt.us/journal/2025v16n2/zhanghousongxiao.pdf [张邝弋, 侯尚余, 宋靖雯, & 肖锐. (2025). 基于 BERT-LDA 的中文学习 APP 评价指标体系构建研究. *科技与中文教学*, *16*(2), 23-47.]

URL (网址): http://www.tclt.us/journal
Email (电子邮件): editor@tclt.us