

# 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)	Bellassen, Joël (白樂桑)	Zucker, Jean-Daniel (諸葛梁)
University of Jean Moulin Lyon 3 miao.lin-zucker@univ-lyon3.fr	Institut National des Langues et Civilisations Orientales joel.bellassen@inalco.fr	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.

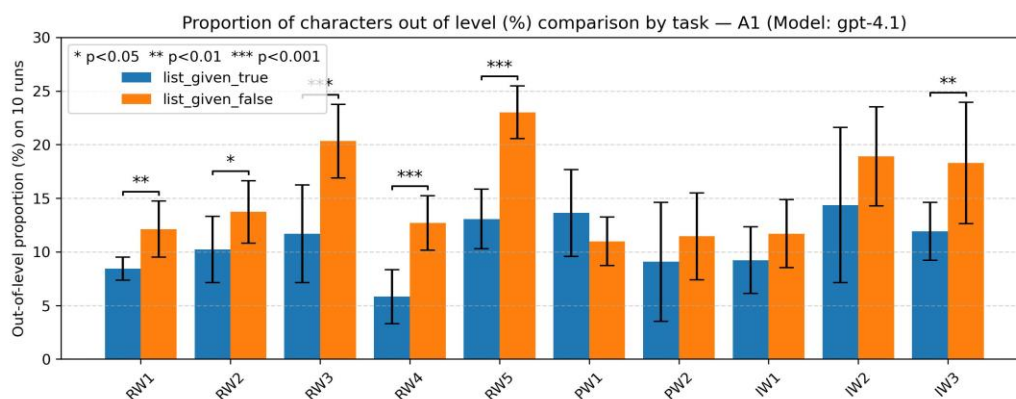
### 5.1 Overview of experimental data

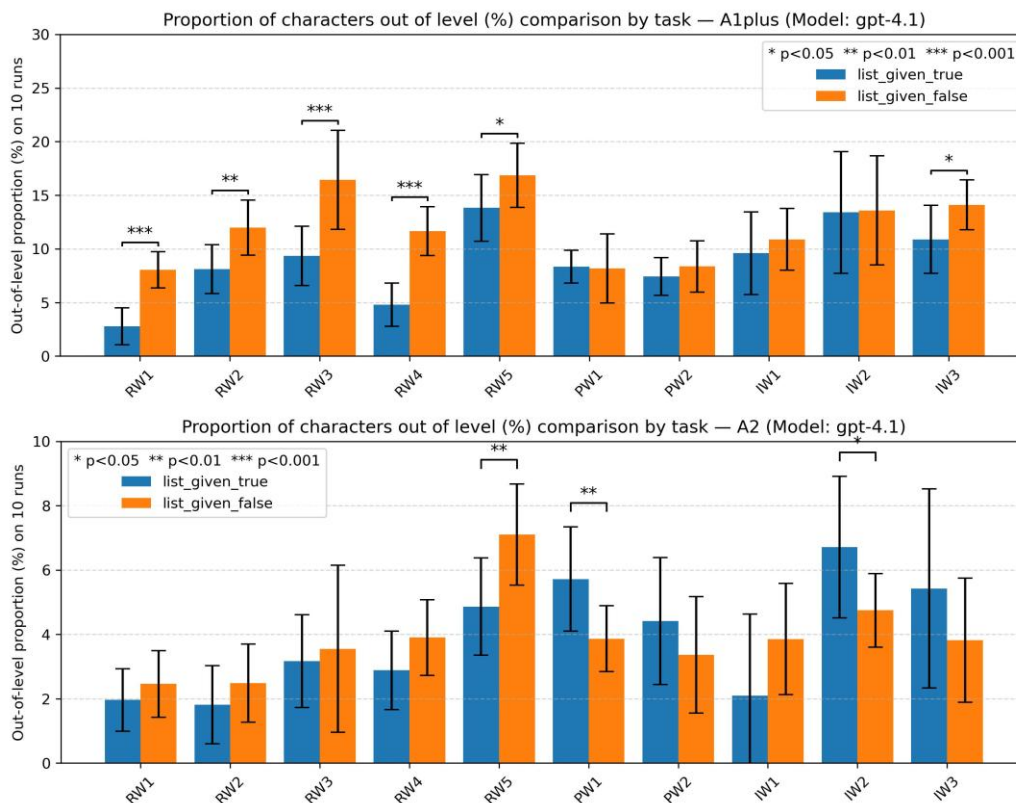
The experimental design crossed four factors: model (GPT-4.1, GPT-4.1-mini), proficiency level (A1, A1+, A2), task type (ten EBCL-aligned written tasks), and prompt condition (with explicit character list vs. without list). For each combination, ten generations were produced (10 runs per condition), 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

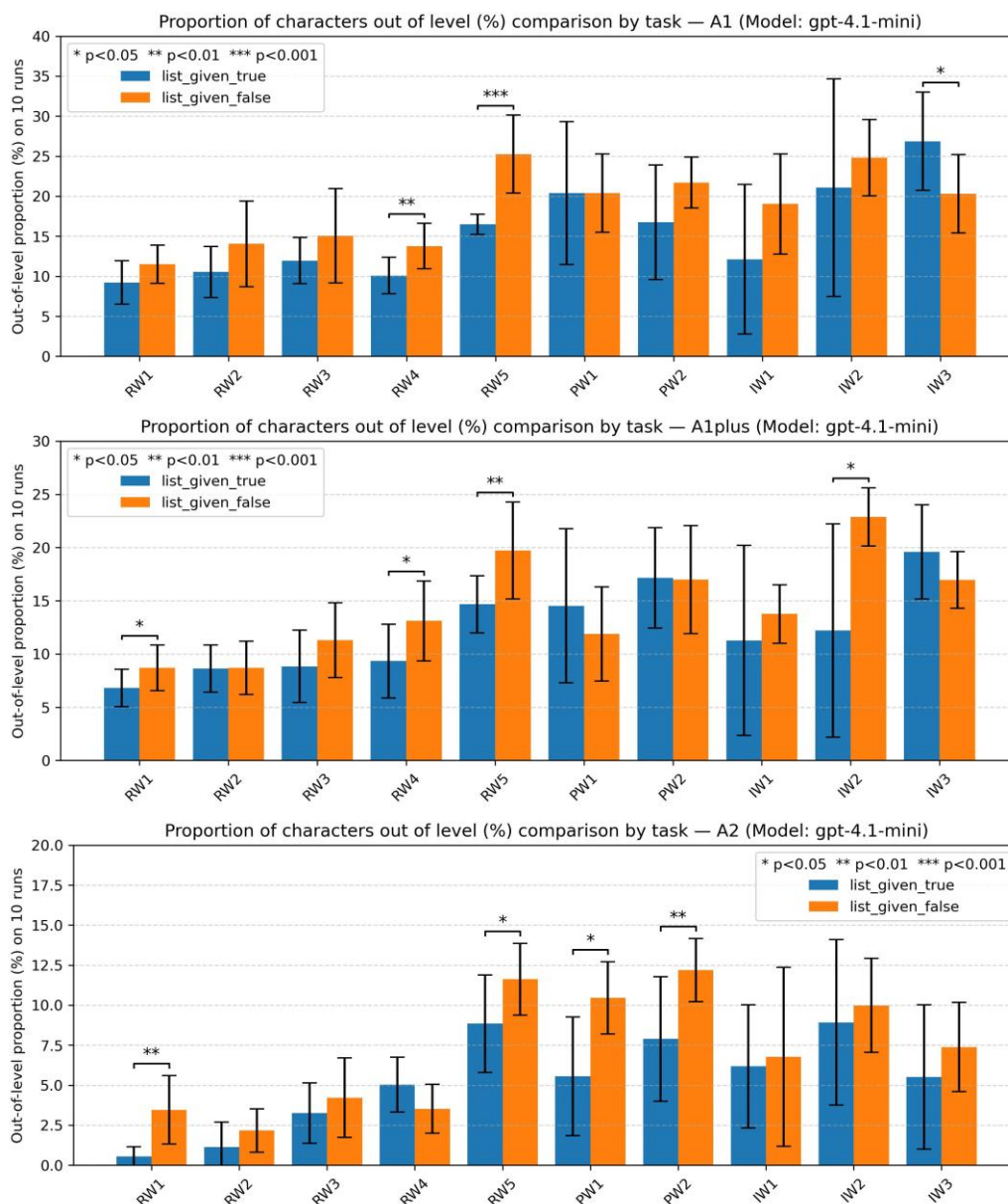
For the GPT-4.1 model (Figure 1 below), the inclusion of a character list is associated with a marked reduction in instruction deviation at levels A1 and A1+. With the character list, GPT-4.1 achieves a mean out-of-level ratio of 10.74% at A1 level (vs. 15.32% without list,  $\Delta=4.6$  pp), 8.85% at A1+ level (vs. 12.01% without list,  $\Delta=3.2$  pp), and 3.91% at A2 level (vs. 3.92% without list,  $\Delta\approx 0$  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.

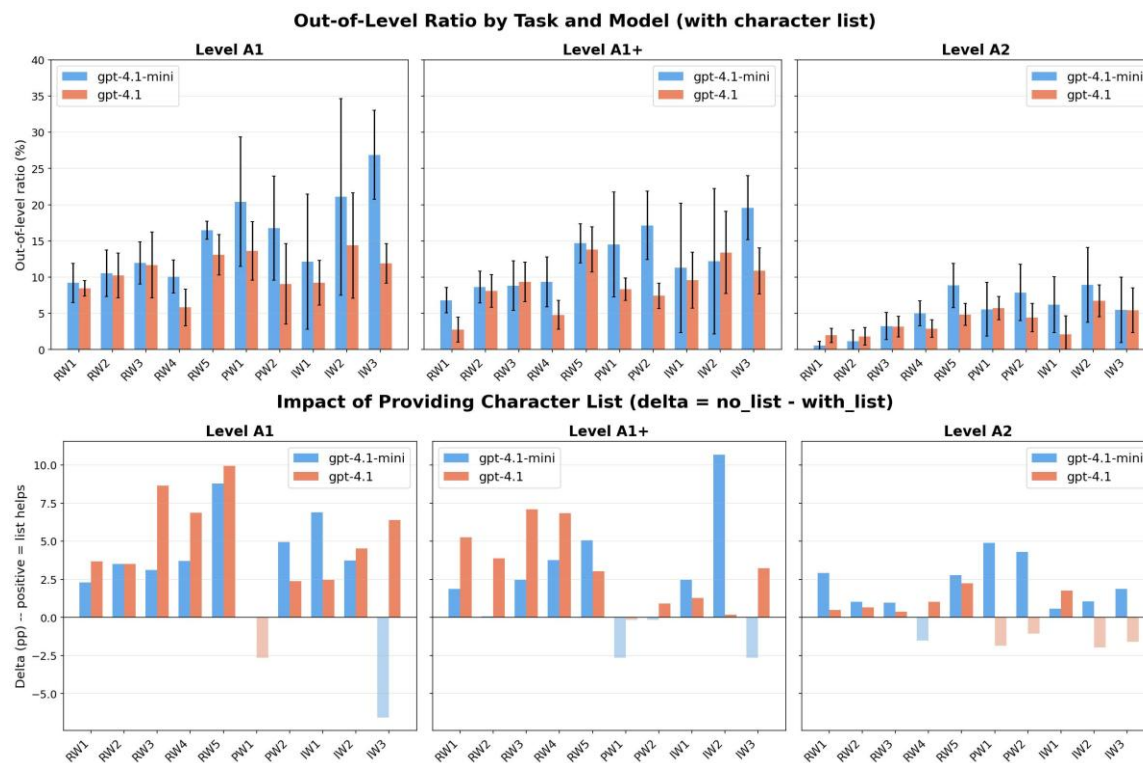


**Figure 2** Effect of providing an EBCL-aligned character list on the proportion of out-of-level characters generated by the GPT-4.1-mini model across tasks and proficiency levels. For each task (RW1–RW5, PW1–PW2, IW1–IW3), results compare conditions with and without an explicit character list in the 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 ( $\text{delta} = \text{no\_list} - \text{with\_list}$ ), 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., Iyer, 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买忙么没每美妹们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份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慢猫毛么没每美妹们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兵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嬉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  
衣医己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.