# Large Language Model and Chinese Near Synonyms: Designing Prompts for Online CFL Learners
# (大语言模型与汉语近义词：针对二语学习者线上学习的提示设计)

Zhao, Qun
(肇群)
The Hong Kong Polytechnic University
(香港理工大學)
qun.zhao@connect.polyu.hk

Hsu, Yu-Yin
(許又尹)
The Hong Kong Polytechnic University
(香港理工大學)
yu-yin.hsu@polyu.edu.hk

Huang, Chu-Ren
(黃居仁)
The Hong Kong Polytechnic University
(香港理工大學)
churen.huang@polyu.edu.hk

**Abstract:** We propose a novel approach of applying large language models (LLMs) to better identify the Zone of Proximal Development (ZPD) of learners of Chinese as a foreign language (CFL). In particular, we designed prompts that assist LLMs in identifying the correct ZPD for CFL learners in order to provide more effective scaffolding. This study utilizes near synonyms to actuate this scaffolding procedure. By beginning with a base prompt and optimizing it in iterative instances, the models are better able to identify proper use-cases for the nuances of each near synonym, leading to more accurate and practical feedback responses. In three experiments, we used different prompts to test the capability of LLMs to understanding and differentiating near synonyms. We found that prompts containing explanations and guidance of reasoning can significantly improve the performance of these models. We attribute this improvement to the addition of interactive learning in prompt design. Adopting the scaffolding framework of learning, we propose the "Zone of Proximal Development Prompts" that can help LLMs to properly identify the correct ZPD of the CFL learners.

**摘要：**本研究提出了一种创新性的方法，来更好地应用大语言模型识别汉语作为外语学习者的最近发展区以提高学习效果。具体来说，我们通过设计提示来帮助大语言模型识别学习者的正确最近发展区，以提供更有效的学习支架。我们以近义词学习任务为本创新性方法的研究先导，首先给出基础提示，进而使用迭代的方法优化提示，促使大语言模型更好地识别近义词之间的细微差别，进而引导模型给出更为准确且实用的反馈。我们通过三个实验测试了大语言模型在不同提示下对近义词的理解和使用能力，并发现包含解释和思考指引的提示能显著提高模型的表现。我们将这一提高归因于在提示设计中融入了互

动学习。采用支架式学习的理论框架，我们提出了“最近发展区提示”，这有助于大语言模型识别汉语学习者的最近发展区。

**Keywords:** Large language models, prompt engineering, Chinese as a foreign language, AI-assist learning, zone of proximal development, scaffolding theory of learning

**关键词：**大语言模型；提示工程；汉语作为外语学习；AI 辅助学习；最近发展区；支架式学习

## 1. Introduction

Near synonyms are words that have highly similar but nonidentical meanings (Lyons,1995). It is common for many dictionaries, such as the Modern Chinese Dictionary (7th edition), to use near synonyms like 方便 fāngbiàn / 便利 biànlì, and 珍惜 zhēnxī / 爱惜 àixī, to define each other (Chief et al., 2000; Li, 2023). In the field of teaching and learning Chinese as a Foreign Language (CFL), the discrimination and collocation of near synonyms are some of the most challenging issues to be explored (Zhang, 2007; Xing, 2013; Li, 2023).

Large language models (LLMs) can be an instructional scaffolding device (Shin et al., 2022). To be specific, LLMs can significantly enhance learning and teaching by generating learner-centric materials, facilitating interaction, and providing personalized feedback in second language (L2) teaching and learning (Bonner et al., 2023; Dai et al., 2023; Moussalli & Cardoso, 2020). In addition, LLMs can be considered as an efficient way to link multiple data-sources, hence can be considered as a natural extension of the linked-data approach to language learning (Huang et al. 2022). Based on these reasons, we propose that LLMs can be an effective tool for CFL learners to learn and discriminate near synonyms. However, a challenge arises as many CFL learners face difficulties in effectively using LLMs due to their limited Chinese proficiency and communication skills (Cai, 2023). To resolve this challenge, it is crucial to guide learners on how to interact with LLMs (Liu et al., 2023).

Prompts are the main channel of communication between the user and LLMs. They elicit LLMs to produce responses that are in line with the user’s intentions. The quality of the prompts directly affects the quality of the generated responses (Ekin, 2023). In other words, a poorly crafted prompt for LLMs “may lead to unsatisfactory or erroneous responses” (Ekin, 2023, p. 3). Prompt engineering fine-tunes the input prompts given to LLMs, optimizing their performance to achieve desired outcomes (Wang et al., 2023). This study focuses on prompt engineering for CFL learners to learn near synonyms; specifically, we explore two key questions: (1) What factors in prompts affect LLMs’ performance in

distinguishing near synonyms? (2) What kind of prompts are most suitable for CFL learners to use to self-study near synonyms using LLMs?

Based on *The Input Hypothesis* (Krashen, 1984), *Error Analysis* (Lu,1994), *The Module-Attribute Representation of Verbal Semantics* (MARVS) *Theory* (Huang et al., 2000), and the characteristics of Chinese grammatical structures, we iteratively optimize prompts in three experiments: The cloze test (4.1), discrimination of near synonyms (4.2), and sentence construction of near synonyms (4.3). This causes LLMs to generate accurate word usage, applicable examples, and explanations for learners. We will show that LLMs' performance does not consistently improve with the addition or replacement of prompt skills—such as the few-shot technique that gives a few demonstrations of the task to LLMs (Brown et al., 2020)—and that more examples in prompts do not necessarily improve accuracy, but well-explained examples can boost performance. By utilizing the scaffolding learning framework, we introduce "Zone of Proximal Development Prompts" that assist LLMs in pinpointing the appropriate Zone of Proximal Development for CFL learners, which initially trains LLMs by providing background information, examples, and explanations for LLMs, and then uses LLMs as teachers, providing more effective scaffolding support to CFL learners. This study presents an innovative approach that optimizes using LLMs as CFL teachers for self-directed learners.

## 2. Literature review

### 2.1 Near synonyms for Chinese language teaching and learning

For CFL learners, misusing near synonyms in terms of meaning and collocation often coexists (Li, 2022). Xing (2013) observed that L2 vocabulary acquisition entails a shift from semantic comprehension to practical application, a challenging transition. Yang (2004) proposed that distinguishing Chinese near synonyms should begin with basic, connotative, and stylistic meanings. Resources such as "Business Chinese Dictionary" (Lu & Lv, 2006), "1700 Groups of Frequently Used Chinese Synonyms" (Yang & Jia, 2007), and "HSK Standard Course" (Jiang et al., 2015) provide important learning materials for learners of Chinese. However, some researchers assert that corpora beyond dictionaries and grammar books are the most dependable linguistic knowledge repositories (Feng, 2010). Corpus-based studies on Chinese near synonyms have provided theoretical support for learning them as a second language, such as Huang et al.'s (2000) Model-Attribute Representation of Verbal Semantics (MARVS) theory. Utilizing the MARVS theory, Cheng (2018) categorized the meanings of the stative verb "大/dà (big)" by consulting the Sinica Corpus, WoNef, and various dictionaries, conducted a detailed and precise analysis of lexical sense classification, offering insights for vocabulary instruction and textbook revision in CFL. Additionally, resources built upon extensive corpora like the Chinese Collocation Knowledge Bases for CFL learners (Hu & Xiao, 2019) and the Chinese Near Synonyms Knowledge Base (Li, 2022) can serve as auxiliary tools for learners.

LLMs are trained on vast amounts of corpus data. In recent years, the role of generative Artificial Intelligence (AI) in assisting L2 learning has been increasingly

proposed and validated (Moussalli & Cardoso, 2020; Cai, 2023; Zaghlool & Khasawneh, 2023). We believe that LLMs will become an important source of learning materials and an assistant for future CFL learning. Therefore, this study explores their ability to differentiate and use Chinese near synonyms, investigates factors affecting LLMs' performance in this context for self-study by learners of Chinese near synonyms, and designs suitable prompts.

### 2.2 Scaffolding and Zone of Proximal Development: An interactive and supportive learning environment

Lantolf and Aljaafreh (1995) established that L2 learners require feedback that falls within their "zone of proximal development (ZPD)" to improve their L2 proficiency towards target levels. The ZPD is the gap between what a learner can accomplish functioning alone (i.e., actual level of development) and what that person is capable of in collaboration with other, more expert individuals (i.e., potential level of development) (Vygotsky, 1978).

Scaffolding is the support rendered by an educator or peer with greater expertise, empowering the learner to undertake tasks they could not complete alone (Cappellini, 2016). This support is most effective when applied within the learner's ZPD (Palinscar & Brown, 1984). The scaffolding process involves three critical steps: initially, the teacher evaluates the learner's present developmental stage; subsequent support and direction are provided; and ultimately, the scaffolding is incrementally removed (Van Der Stuyf, 2002). Scaffolding transforms a language learner from a passive recipient of linguistic knowledge into an active participant or contributor, fostering autonomous engagement in the learning process with diminishing oversight required (Betts, 2004). Studies emphasized that scaffolding underpins learner autonomy in foreign language acquisition (Smith & Craig, 2013; Chen, 2021).

In digital settings, scaffolding is universally accessible and offers broad-based support for learners' educational needs (Wood et al., 1976). Recent studies suggest that LLMs show potential as a scaffolding instrument in instruction (Shin et al., 2022). However, careful prompting is crucial when integrating LLMs into L2 education (Caines et al., 2023), and it is vital to scaffold learners' interactions with LLMs appropriately (Liu et al., 2023).

### 2.3 Prompt engineering of LLMs

In the field of natural language processing, prompt engineering has gained prominence as an innovative approach. It offers a more efficient and cost-effective way to leverage LLMs (Wang et al., 2023). Essentially, prompt engineering fine-tunes the questions or commands given to AI models, optimizing their performance to achieve desired outcomes (Wang et al., 2023). This process enhances the model's ability to provide accurate and contextually appropriate answers for downstream tasks (Lo, 2023). LLMs significantly benefit from meticulous prompt engineering, which can be done either manually (Reynolds & McDonell, 2021) or automatically (Shin et al., 2020).

In recent studies, scholars have explored various prompt methods, including gradient-based approaches (Lester et al., 2021), 0-shot techniques (Reynolds & McDonell, 2021), one-shot strategies (Ekin, 2023), few-shot paradigms (Brown et al., 2020), and the Chain of Thought (CoT) method (Wei et al., 2022). Additionally, frameworks such as the CRISPE framework (Nigh, 2023), OpenPrompt (Ding et al., 2021), and DifferentiAble pRompT (DART) (Zhang et al., 2022) have demonstrated successful prompt engineering. However, while specific domain studies are being conducted (Heston & Khun, 2023; Meskó, 2023), research in the field of education and L2 teaching remains relatively scarce, particularly in the context of CFL.

## 3. Methodology

We adopted an empirical research paradigm and quantitative methodologies for data analysis. We conducted three experiments: The cloze test, discrimination of near synonyms, and sentence construction with near synonyms, which evaluate the ability of LLMs to recognize and understand near synonyms from distinct perspectives.

To be specific, the cloze test is a part of the Reading (阅读) task in the HSK5 Test (汉语水平考试五级). This part contains four short texts, each containing 3-4 cloze blanks for filling a word or a clause; participants need to select the right answer from four options (as seen in Table 1). We elicit LLMs to select the best answer for each blank under different prompts in experiment 1. In the discrimination of near synonyms test (experiment 2), we ask LLMs to choose a better sentence from a sentence paired with near synonyms. For example, to discriminate the near synonyms pair 安静 ānjìng ‘quiet’ and 清净 qīngjìng ‘tranquility; peacefulness’, we elicit LLMs to choose the one in the sentence pair in (1) that better expresses “The children have all fallen asleep quietly.”

1) a. 孩子-们 都 已经 安静-地 入睡 了。
Háizi-men dōu yǐjīng ānjìng-de rùshuì le.
‘The children have all fallen asleep quietly.’
b. 孩子-们 都 已经 清静-地 入睡 了。
Háizi-men dōu yǐjīng qīngjìng-de rùshuì le.
‘The children have all fallen asleep quietly.’

For sentence construction with the near synonyms test (experiment 3), we evaluate the sentences LLMs make under different prompts. For instance, we initially give a prompt as shown in (2), interactively optimize prompts afterward (see details in the following section), and evaluate the outputs to verify the effectiveness of most craft prompts.

2) Prompt:
“用[分别 fēnbié /分手 fēnshǒu] 造句
‘Make sentences with [separation/breakup]’

### 3.1 Date collection and preprocessing

The dataset for experiment 1 includes over 320 blanks collected from the HSK5 Test. Each short text contains 3-4 cloze blanks, which will be recorded as individual items along with their corresponding standard answers (Table 1).

**Table 1 Sample of the Cloze Test Data**

| Text | Blanks | Options | Standard Answers |
|---|---|---|---|
| 土豆会令人发胖吗？做法不当的话，当然会。做过“土豆烧肉”的人都知道，土豆的吸油能力很[MASK1]。据测定，一只中等大小的不放油的“烤土豆”仅含 90 千卡热量，而同一个土豆做成炸薯条后所含的热量能达 200 千卡以上。[MASK2]，令人发胖的不是土豆本身，而是它[MASK3]的油脂。 | MASK1 | A.强<br>B.多<br>C.大<br>D.重 | A.强 |
| | MASK2 | A.但是<br>B.那么<br>C.从而<br>D.可见 | D.可见 |
| | MASK3 | A.吸收<br>B.吸取<br>C.吸引<br>D.吸纳 | A.吸收 |

The dataset for experiment 2 consists of 400 sentence pairs collected from the “1700 Groups of Frequently Used Chinese Synonyms (1700 对近义词用法对比) (Yang & Jia, 2007) and the Global Chinese Interlanguage corpus (GCI corpus; 全球汉语中介语语料库[1]). Each pair comprises a good sentence and a bad sentence with near synonyms marked as “x” and “y” individually to facilitate LLMs processing (as shown in Table 2).

[1] 全球汉语中介语语料库 URL: http://qqk.blcu.edu.cn

**Table 2 Sample of Discrimination of Sentences with Near Synonyms Data**

| x (Good sentence) | y (Bad sentence) |
|---|---|
| 孩子们都已经**安静地**入睡了。 | 孩子们都已经**清静地**入睡了。 |
| 我**被迫**无奈才答应跟他去。 | 我**被动**无奈才答应跟他去。 |
| 听到爷爷去世的消息，她**暗暗**伤心。 | 听到爷爷去世的消息，她**偷偷**伤心。 |

Given the importance of addressing common errors in Chinese language learning, this study utilizes a total of 30 pairs of misused synonyms of real student data from the GCI corpus for experiment 3. We organize high-error-rate words and their corresponding near synonyms into a dataset as near synonyms pairs. For instance, "分别 fēnbié" is the word with the highest frequency of misuse in the corpus. We manually screened for errors caused by misunderstandings of near synonyms. In the sentence as shown in (4)" (For ease of reading, other errors in the original sentence have been corrected), the appropriate word to use is "分辨 fēnbiàn", but the student incorrectly used "分别 fēnbié". Therefore, the near synonyms pair "分别/分辨" as shown in (3) was entered into the dataset.

3) 分别/分辨
fēnbié/ fēnbiàn
'distinguishing; individually; and parting/distinction; discrimination'

4) 首先 要 谈 中国 汉字 发音，有 四个 声调，
Shǒuxiān yào tán Zhōngguó hànzì fāyīn, yǒu sìge shēngdiào,
最难 【分别】 [Cb分辨] 的 是 第一和第四 声。"
zuìnán【fēnbié】[Cb fēnbiàn] de shì dìyī hé dìsì shēng.
'First, let's talk about the pronunciation of Chinese characters. There are four tones, and the most difficult part is to distinguish the first and fourth tones.'

For the GCI corpus data, each collected sentence that contains errors is manually cleaned in five steps (as seen in Table 3). First, correct other errors in the sentences (according to the annotations) but retain the near synonyms error. Second, delete other parts (if necessary) that do not affect the independent meaning of the clause, as there might be ambiguous expressions that could affect the experiment's validity. Third, record the sentence that was preliminarily corrected but still contains a near synonym error, such as y (bad sentence) in the dataset. Fourth, correct the near synonym errors in the sentence. Fifth, record the corrected sentence as x (good sentence).

**Table 3 An Example of Data Cleaning in Experiment 2**

| Procedures | Cleaned Sentences |
|---|---|
| Original Data with Annotations | 在南京，我常常【利用】[Cb 坐]地铁【**还是**】**[Cb 或]**公共汽车，公用汽车【的】[Cd]费，比韩国，【很】[Cd]便宜。 |
| Step 1: Correct Unrelated Errors and Annotations | 在南京，我常常坐地铁**还是**公共汽车，公用汽车的费，比韩国，很便宜。 |
| Step 2: Delete Ambiguous Part | 在南京，我常常坐地铁**还是**公共汽车。 |
| Step 3: Record Incorrect Sentence | y: 在南京，我常常坐地铁**还是**公共汽车。 |
| Step 4: Correct Near Synonym Error | 在南京，我常常坐地铁**或**公共汽车。 |
| Step 5: Record the Correct Sentence | x: 在南京，我常常坐地铁**或**公共汽车。 |

* 在南京，我常常坐地铁或公共汽车。

Zài Nánjīng, wǒ chángcháng zuò dìtiě huò gōnggòngqìchē.

'In Nanjing, I often take the subway or the bus.'

Additionally, it is worth noting that due to the limited amount of data, to ensure the reliability, validity, and generalizability of the experiments as much as possible, each time the model is tested via API access in experiment 1 and experiment 2, the *random shuffle* function is used to randomize the data. When testing via the web interface, Research Randomizer is utilized for random sampling to select data for testing.

### 3.2 Large Language Models selection

In this study, we tested three LLMs, ERNIE4.0, Baichuan2-13B, and GPT3.5 Turbo, based on the SuperCLUE benchmark. The SuperCLUE (Xu et al., 2023) is a comprehensive Chinese large language model benchmark, which is an extension and development of a popular benchmark named The Chinese Language Understanding Evaluation (CLUE) (Xu et al., 2020). The datasets for SuperCLUE's tests include language understanding data, long text data, role-playing data, and generation and creation data (Xu et al., 2023), which are highly relevant to the tasks of this study. In the six tests conducted from August 2023 to February 2024[2], ERNIE4.0 ranked first three times, and Baichuan2-13B ranked first once in the leaderboard of China's LLMs, and both models can be accessed via APIs and web interfaces. Meanwhile, we also selected GPT3.5 Turbo from OpenAI, a world-leading company in the field. GPT3.5 Turbo is a much lower-cost and more feasible option than GPT4 on current and future study, although GPT4 ranked at the top of the SuperCLUE list for now. Specifically, given the limited data size and computing power available for this study, prompt engineering has proven to be an effective method for enhancing the performance of LLMs (Wang et al., 2023). However, in future research,

[2] SuperCLUE report URL: https://www.cluebenchmarks.com/superclue_2404

we plan to fine-tune the LLMs to investigate their performance on current tasks. Consequently, we will be able to compare the outcomes of prompt engineering with those of fine-tuning.

### 3.3 Evaluation

The evaluation metrics for experiment 1 and experiment 2 include accuracy, F1 score, and internal consistency. These three metrics are crucial aspects of assessing the performance of language models. They reflect the model's accuracy, predictive power, and the coherence and consistency of the predictive results from different perspectives. Specifically, accuracy represents the proportion of correct predictions made by the model out of the total number of predictions. The F1 score is the harmonic mean of precision and recall, used to measure the model's predictive ability for positive classes. Internal consistency is an important indicator for evaluating the reliability and robustness of a model. A model with internal consistency can provide more trustworthy predictive results. We ran each task three times on each model in experiments 1 and 2, and the median of the three runs was recorded as the result. After identifying the model that performs the best under the same prompt through comparison, we conducted additional prompt-optimizing tests (including experiment 3) on that model.

For the sentence construction task, we invited three CFL teachers to score the sentences provided by the no-technique prompt (pre-test) and the technique prompt (post-test) using a 5-point Likert scale respectively. As learners often misuse near synonyms due to their easily confused senses, the model's output sentences should be grammatically correct and illustrate the nuanced differences and easily confused senses between near synonyms. We used three scoring standards to measure the suitability of the model's sentences for self-study of near synonyms: 1. The sentences have no grammatical and pragmatic errors; 2. The sentences are constructed with an easily confused sense of near synonyms; 3. When the grammar and semantics are correct, whether the target word in the sentence can be replaced with a corresponding near-synonym, and whether the model explains. The experiment used the average score of three Chinese teachers as the final score for analysis.

Accessing LLMs via API with Python code can result in accuracy, F1 score, and internal consistency. However, because of the emergent abilities of LLMs (Wei et al., 2022), the outputs generated by LLMs can be not only a simple option like an answer as "A", it can give users some analysis and reasons for their choice. Therefore, we access LLMs via the web interface in this situation, as well as for experiment 3.

### 3.4 Prompt optimizing

Given that both the instructional and target languages are Mandarin Chinese, the prompts used in this study will also be in Mandarin (Table 4). Although auto-prompting provides efficiency (Shin et al., 2020), we adopted manually designed prompts that are more likely to match tasks at the initial stage of the study due to the varying nature of CFL learning tasks and learners. This method ensures that the prompts align precisely with each

task's specific requirements, thereby guiding LLMs to produce more accurate and contextually appropriate content. The formulation of these prompts adheres to the Capacity and Role, Insight, Statement, Personality, and Experiment (CRISPE) framework (Nigh, 2023), which encapsulates five fundamental parts: Capacity and Role, Insight, Statement, Personality, and Experiment. This study utilizes and tests various prompt techniques such as 0-shot techniques (Reynolds & McDonell, 2021), one-shot strategies (Ekin, 2023), few-shot paradigms (Brown et al., 2020), and the Chain of Thought approach (Wei et al., 2022). In addition, we leverage the input Hypothesis (Krashen, 1984), Error Analysis (Lu,1994), The Module-Attribute Representation of Verbal Semantics (MARVS) theory (Huang et al., 2000), and the characteristics of Chinese lexical, grammatical, and pragmatic structures.

We analyze the relationship among prompt techniques, the number of questions, and the performance of LLMs using statistical description, t-test, and simple linear regression. This analysis helps us understand how different factors influence the performance of LLMs and guides us in optimizing the prompts.

**Table 4 Examples of Tested Prompts**

| Templates | Examples |
|---|---|
| 你是汉语语言专家，请你根据搭配频率，判断 {"x"}和{"y"}哪句更好。从搭配、语义轻重、使用习惯、语体、语法等方面分析句子中关键词的细微差别。 | 你是汉语语言专家，请你根据搭配频率，判断 "孩子们都已经安静地入睡了。"和"孩子们都已经清静地入睡了。"哪句更好。从搭配、语义轻重、使用习惯、语体、语法等方面分析句子中关键词的细微差别。 |
| 区分动词近义词的一种方法是分析与其搭配的对象、范围、程度等的不同。例如：{}<br>请你根据词语搭配对象、范围、程度的不同思考并回答：{x:/y:}哪句更好？ | 区分动词近义词的一种方法是分析与其搭配的对象、范围、程度等的不同。例如：{查阅/查看}。<br>{查阅}的对象范围小，只包括文件等；{查看}的对象范围大，包括文件、物体等。因此，{x: 警察查看了事故发生现场。/y: 警察查阅了事故发生现场。}，x 句较好。<br>请你根据词语搭配对象、范围、程度的不同思考并回答：{x:由于信号受到打扰，电视总不清楚。/y:由于信号受到干扰，电视总不清楚。}哪句更好？ |

| | |
|---|---|
| 按照下面的步骤反思你刚才关于{word/sentence1}和{word/sentence2}的答案和解释：{E}<br>1.重新仔细审题并重复题目<br>2.重点查看关键词所在的句子<br>3.重点查看句子对应的编号<br>4.阅读并重复你刚才的解释<br>5.根据{n}步的结果，检查你前面的解释中，是否存在错误<br>6.告诉我你的错误并改正 | 按照下面的步骤反思你刚才关于{“受”}和{“挨”}的答案和解释：{“挨”和“受”在某些方言中可互换，但普通话中更常用“挨”，且“挨”在某些表达中含有一种经历或忍受的意味，所以选 x。“受贿”是固定搭配，所以选 y。}<br>1.重新仔细审题并重复题目<br>2.重点查看关键词所在的句子<br>3.重点查看句子对应的编号<br>4.阅读并重复你刚才的解释<br>5.根据{1-4}步的结果，检查你前面的解释中，是否存在错误<br>6.告诉我你的错误并改正 |

## 4. Findings

### 4.1 Experiment 1

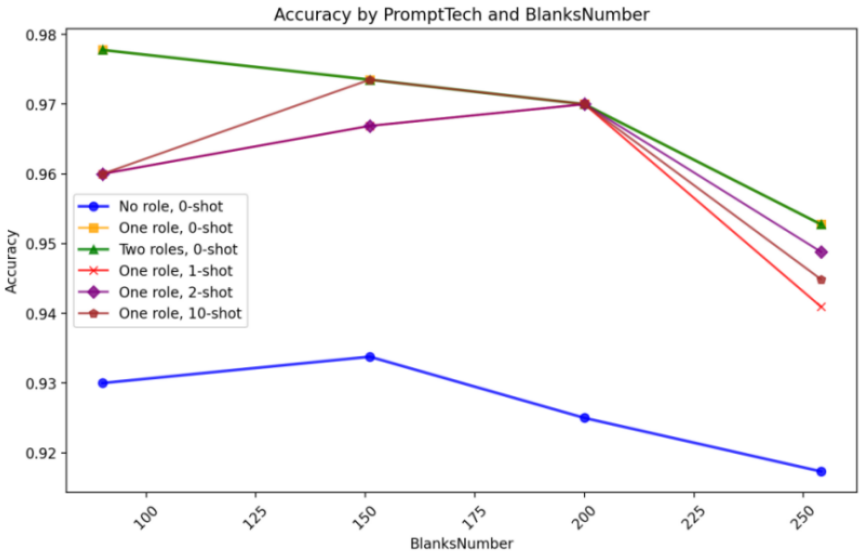
The experiment initially accessed three models via API and randomly selected 13 texts, comprising a total of 49 blanks, from the dataset. The same prompt (zero-shot, expert role) was used to test the accuracy, F1 score, and internal consistency of the three models on the same task. Each model was run three times for the task, and the median of the three results was adopted. The experimental results showed that ERNIE4.0 scored the highest (as shown in Table 5), so the subsequent tests in this experiment will be conducted using ERNIE4.0.

**Table 5 The Performance of Three LLMs on the Cloze Test Task**

| Metrics | GPT3.5 Turbo | ERNIE4.0 | Baichuan2-13B |
|---|---|---|---|
| Accuracy | 0.612 | 1 | 0.980 |
| F1 Score | 0.607 | 1 | 0.980 |
| Consistency | 0.484 | 1 | 0.973 |

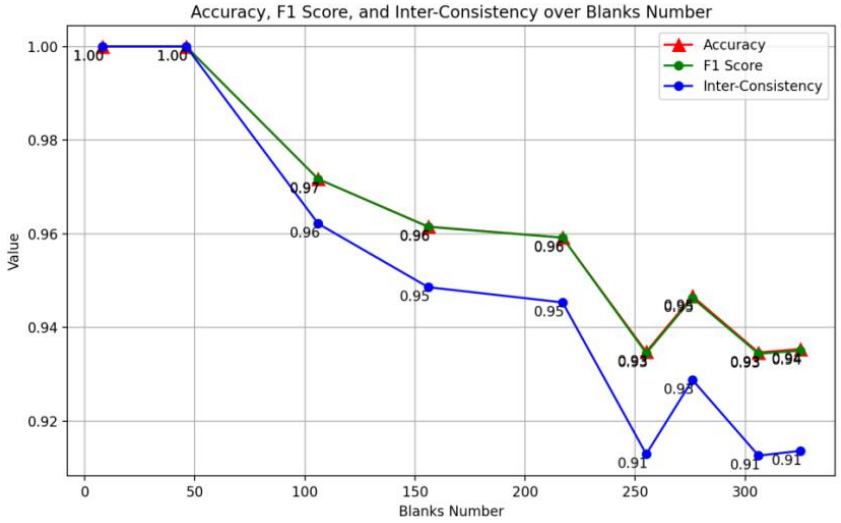
* The results were kept to three decimal places in the count.




* The results were kept to two decimal places in the count

**Figure 1 Accuracy of Prompt Techniques and Number of Blanks**




* The results were kept to two decimal places in the count.

**Figure 2 Comparative Analysis of Accuracy, F1 Score, and Inter-Consistency across Varying Blanks Numbers**

Subsequently, we tested different prompt techniques on ERNIE4.0 (Figure 1). Compared to zero-shot, few-shot (Brown et al., 2020) did not significantly improve the model's answer accuracy when k=1, k=2, and k=10. The "role-playing" (Ladousse, 1987) and the "CoT" (Wei et al., 2022) guide the model's thinking and emphasize the display of the analysis and thinking process in the answer, significantly increasing the accuracy. Specifically, when we tested 20 blanks, which were randomly selected from the dataset three times on the Web interface, the mean accuracy of the answer without techniques and not showing the thinking process was 0.93. However, when we used the above techniques

and emphasized the analysis and thinking process, informing the model of the key points of problem-solving, the mean accuracy of the answer to the same question reached 1. Interestingly, when guiding reflection, having the model use two roles (teacher and student) to check and question each other did not significantly improve the accuracy of the results.

In addition, we also found that the number of questions inputted at once may affect the model's performance. As can be seen from Figure 2, overall, as the volume of questions increases, the accuracy, F1 score, and internal consistency all exhibit a downward trend. In other words, the more questions given at once, the lower the potential performance score of the model. It is worth noting in this test that when the number of questions given at once is less than 250, the accuracy and F1 score are greater than 0.95. However, when the test data included 254 questions, the accuracy and F1 scores dropped below 0.95. This represents a significant change.

### 4.2 Experiment 2

In the beginning, we randomly selected 50 sentence pairs to test three LLMs using the same prompt (zero-shot, expert role). ERNIE4.0 performed the best with an accuracy of 0.980, F1 score of 0.990, and internal consistency of 0.960 (as shown in Table 6). Therefore, subsequent tests will be conducted exclusively using ERNIE4.0.

**Table 6 The Performance of Three LLMs on Sentence Pairs Judgement**

| Metrics | GPT3.5 Turbo | ERNIE4.0 | Baichuan2-13B |
|---|---|---|---|
| Accuracy | 0.620 | 0.980 | 0.960 |
| F1 Score | 0.765 | 0.990 | 0.980 |
| Internal Consistency | 0.510 | 0.960 | 0.918 |

* The results were kept to three decimal places in the count.

Similar to experiment 1, using the "role-playing" (Ladousse, 1987) paradigm and the CoT technique (Wei et al., 2022) in the prompt improved the model's answer accuracy. Specifically, without using "role-playing" (Ladousse, 1987) and CoT techniques (Wei et al., 2022), ERNIE4.0's accuracy of 10 and 50 pairs of judgments was 0.6 and 0.74, respectively. However, the highest accuracy reached 1 with techniques.

An interesting finding is that asking LLM to display its thinking process and analysis helps increase accuracy. For 50 sentence pairs, the accuracy can reach 1 when we instruct as shown in (5). In contrast, the accuracy is 0.98 (as shown in Table 6) without guiding LLM to display its thinking process instruction as shown in (6).

5) Prompt:
逐步分析和思考后给出答案和分析过程。
'Provide the answer and analysis process after gradually analyzing and thinking.'

6) Prompt:
不要展示分析过程，只告诉我你的答案。
'Do not show the analysis process; just tell me your answer.'

We also tested ERNIE4.0's performance with different numbers of sentence pairs: 5, 10, 50, 60, 70, 80, 90, 100, 150, 200, and 250 input at once. These tests were conducted under the same prompt (zero-shot, expert role, display think process) via the web interface. We found that when no more than 50 sentence pairs were given at once, the model's accuracy could reach 1. However, the accuracy quickly dropped when more than 50 pairs were given (as shown in Figure 3).

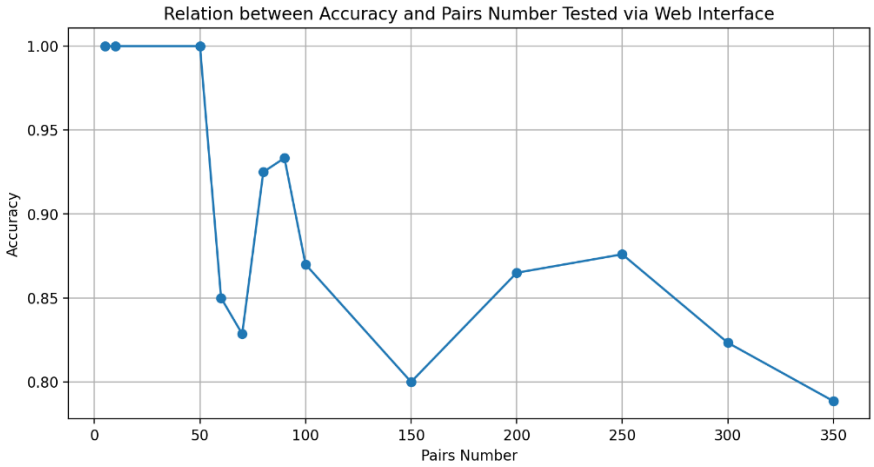



**Figure 3 Variation of Accuracy with Pairs Number via Web Interface**

Meanwhile, a simple regression analysis showed a significant impact of the number of sentence pairs on accuracy. The model was statistically significant with an F-statistic of 10.95 and a Prob (F-statistic) of 0.00697, indicating a significant impact of the number of sentence pairs on accuracy. The model's intercept was 0.9530, which is highly significant, with a t-value of 38.497. The coefficient of pair number was -0.0005, significant with a t-value of -3.309, suggesting a negative correlation between the number of pairs and accuracy. Thus, the more sentence pairs input at once, the lower the LLMs' accuracy.

### 4.3 Experiment 3

Without a specially designed prompt (pre-test), the model generally produces sentences with correct usage, but the sentences may not involve the easily confused sense of near synonyms. As the instance we state in section 3.1: 分别 fēnbié and 分辨 fēnbiàn. From the sentence extracted in the corpus as shown in (4), we see that the easily confused sense is "distinction, distinguishing" while the "parting" sense of "分别 fēnbié" does not confuse with "分辨 fēnbiàn". Yet, without special guidance, the model's sentences with "分别" do not involve the easily confused sense, as shown in (7) and (8).

7) 我们即将分别，踏上各自的人生道路，但我们的友谊将永远长存。
Wǒmen jíjiāng fēnbié, tàshàng gèzì de rénshēng dàolù, dàn wǒmen de yǒuyì jiāng yǒngyuǎn chángcún.
'We are about to part ways and embark on our own paths in life, but our friendship will last forever.'

8) 小孩子的视力还没有完全发育，有时候难以分辨远处的物体。
Xiǎoháizi de shìlì hái méiyǒu wánquán fāyù, yǒushíhou nányǐ fēnbiàn yuǎnchù de wùtǐ.
'Children's vision is not fully developed yet, sometimes making distinguishing objects in the distance hard.'

To elicit LLMs to generate sentences accurately according to the learner's confusion, we adopt three approaches to prompting (post-test). The first approach is to provide sentences with errors and let the model actively identify and learn the focus of the current task. The second approach involves giving a warning about the usage of easily confused senses in near synonyms when the learner does not have sentences with errors, which requires the learner to point out their points of confusion. The third approach is used when the learner does not have specific confusion; we ask the model to analyze and construct sentences for each sense of the near synonyms and the easily confused senses. Figure 4 shows an example of the outputs generated by ERNIE4.0 under our craft prompt.

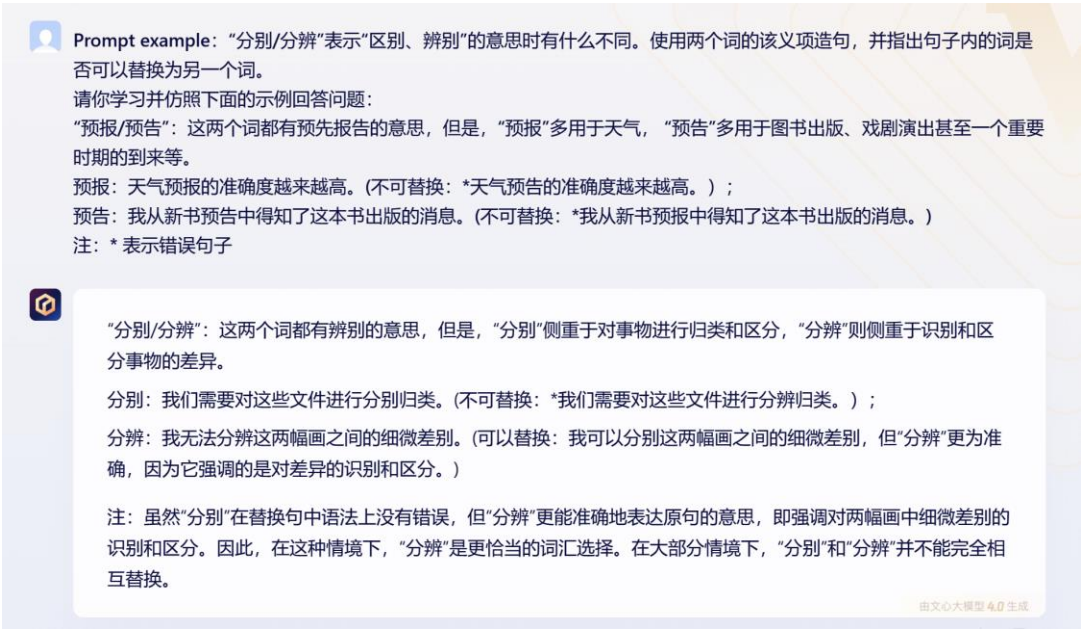



**Figure 4 An example of the Outputs under Craft Prompt**

A paired-sample t-test was conducted to compare pre-test and post-test scores. There was a significant difference in scores for pre-test (M=4.49, SD=0.46) and post-test (M=4.95, SD=0.09) conditions; t (29) = -5.85, $p < .001$ (two-tailed). The results suggest a statistically significant increase from pre-test to post-test scores, indicating that our technique prompt significantly improves the model's performance.

Since the ideal input should be comprehensible to learners (Krashen, 1984), sentences output by the model using higher-level vocabulary and grammar beyond learners' language proficiency may cause additional understanding burdens. Therefore, we suggest assigning the model the identity of a CFL learner and their Chinese level, limiting the sentence's grammar difficulty and length, and asking the model to follow the i+1 principle (Krashen, 1984) to provide sentences matching learners' Chinese level. After the model receives clear vocabulary and grammar level restrictions, there is some improvement in language difficulty matching.

## 5. Discussion and interpretation of the results

Through three experiments, we discovered that different LLMs perform differently on the same tasks. ERNIE4.0 tends to provide detailed explanations without requests and achieves the highest accuracy and F1 score. When provided with professional instruction, it excels at recognizing, explaining, and demonstrating nuances of near synonyms from semantic and pragmatic perspectives.

Regarding the factors that influence the model's performance, we found that both the number of questions given at once and the prompt techniques play a role. Specifically, the number of questions given at once can affect the performance of LLMs. In our experimental data, the model's performance significantly decreases when more than 50 or even 250 questions are given at once. Therefore, we do not recommend giving too many questions at once when using LLMs.

For the design of the prompt, we first agree that the language of the prompt should convey the requirements clearly and specifically (Ekin, 2023; OpenAI, n.d.), and the "role-playing" paradigm (Ladousse, 1987) applies to three tasks. At the same time, we also found that simply increasing the examples may not improve the model's performance. However, providing examples while giving the model appropriate guidance, such as guidance on the order of thinking and the parts that need to be focused on, can help the model first understand our needs, arouse the model's corresponding knowledge reserves, and usually elicit the model to give answers that are more in line with user expectations.

We believe that "role-playing" (Ladousse, 1987) and providing guidance on steps of learning and key learning points in prompts incorporate the element of interactive support of learning. That is, following the scaffolding framework of education (Wood et al., 1976), support and interaction are crucial to effective learning. In other words, LLM cannot directly interact with the learners. However, designing the prompts to incorporate the interactive supporting elements could provide effective scaffolding to the CFL learners. We refer to this prompt pattern as the "Zone of Proximal Development Prompts" (ZPDP), which helps LLMs to identify the correct ZPD (Lantolf & Aljaafreh, 1995) of the CFL learners involved. The ZPDP model first learns the user's information (identity, Chinese language level), the user's learning goals, the current task mode, the solution ideas of the current task, etc., so that the model can provide the relevant knowledge and is most

supportive of learning. Then, the model uses its knowledge and the information just learned to generate answers for users, to achieve the purpose of assisting learners in learning Chinese. The advantage of ZPDP is that it does not need to consume a lot of computing power to retrain the model, but activates the existing knowledge and abilities of the LLMs to improve the performance of the language model in the downstream task of Chinese language knowledge tutoring, and well-motivated by the scaffolding theory of learning (Wood et al., 1976).

## 6. Implication and limitation

Intelligent Computer-Assisted Language Learning (ICALL) has been at the forefront of learning technology for decades. The recent emergence of generative AI and LLMs brings both possibilities and challenges to this field. The current study focuses on better leveraging LLMs to assist language learning and aims to help learners obtain answers from LLMs through optimized prompts. These personalized answers are generated to address specific learners' queries, aiding them in real-world problem-solving. This research substantiates the viability of the First Principles of Instruction framework (Merrill, 2002) for ICALL by demonstrating its applicability in assisting CFL learners to self-study near synonyms using LLMs. In addition, it fills the research gap related to using prompt engineering with LLMs for CFL.

In addition, the ZPDP model is reusable and generalizable for CFL learners. When learners use it, they only need to fill in their specific conditions and needs in the blanks of the pattern to get a more accurate answer. It improves learners' efficiency using LLMs and reduces their learning costs. It is expected to solve the dilemma of many learners who cannot learn anytime and anywhere from Chinese human teachers. As long as learners have a device that can access the internet, they can turn LLMs into their personal portable Chinese teachers.

Note that the performance of LLMs in the current study could be unstable due to both the dynamic nature of LLM and constraints on data and computing power. Given such constraints, perplexity should be an appropriate metric for evaluating performance, but we cannot access the function of the three LLMs through API. Additionally, near synonyms learning is one of many challenging learning tasks for L2 learners. Our future research directions include how to use LLMs for more learning tasks and how to implement better evaluation measures such as perplexity.

## References


Betts, G. (2004). Fostering autonomous learners through levels of differentiation. *Roeper Review*, *26*(4), 190-191.

Bonner, E., Lege, R., & Frazier, E. (2023). Large Language Model-based Artificial Intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, *23*(1), 23-41.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language Models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cai, W. (2023). Learning and teaching Chinese in the ChatGPT context. *Language Teaching and Linguistic Studies*, *4*, 13-23. [蔡薇. (2023). ChatGPT 环境下的汉语学习与教学. *语言教学与研究*, *4*, 13-23. ]

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., ... & Buttery, P. (2023). On the application of Large Language Models for language teaching and assessment technology. *CEUR Workshop Proceedings. v. 3487*, 173-197. https://ceur-ws.org/Vol-3487/paper12.pdf

Cappellini, M. (2016). Roles and scaffolding in teletandem interactions: A study of the relations between the sociocultural and the language learning dimensions in a French–Chinese teletandem. *Innovation in Language Learning and Teaching*, *10*(1), 6-20.

Chen, C. (2021). Using scaffolding materials to facilitate autonomous online Chinese as a foreign language learning: A study during the COVID-19 pandemic. *Sage Open*, *11*(3). https://doi.org/10.1177/21582440211040131

Cheng, Y. (2018). *Sense analysis of stative verb by French learners- A study of polysemy "Big"* [Master's thesis, National Taiwan Normal University]. National Digital Library of Theses and Dissertations in Taiwan. [鄭語箴. (2018). *法籍學習者之狀態動詞語義分析研究-以[大]之多義性為例* [學位論文, 臺灣師範大學]. 臺灣博碩士論文知識加值系統. ] https://hdl.handle.net/11296/qegrg5

Chief, L. C., Huang, C. R., Chen, K. J., Tsai, M. C., & Chang, L. L. (2000). What can near synonyms tell us. *International Journal of Computational Linguistics and Chinese Language Processing*, *5*(1), 47-60.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. Can Large Language Models provide feedback to students? A case study on ChatGPT. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323-325. DOI:10.1109/ICALT58122.2023.00100.

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H., & Sun, M. (2022). OpenPrompt: An Open-source Framework for Prompt-learning. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 105–113. DOI: 10.18653/v1/2022.acl-demo.10

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

Feng, Z. (2010). Mining knowledge & extracting information from corpus. *Foreign Languages and Their Teaching, 4*, 1-7. [冯志伟. (2010). 从语料库中挖掘知识和抽取信息. *外语与外语教学, 4*, 1-7.]

Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education, 2*(3), 198-205.

Hu, R., & Xiao, H. (2019) The construction of Chinese Collocation Knowledge Bases and their application in second language acquisition. *Applied Linguistics, 1*, 135-144. [胡韧奋, 肖航. (2019). 面向二语教学的汉语搭配知识库构建及其应用研究.*语言文字应用, 1*, 135-144.]

Huang, C. R., Ahrens, K., Chang, L. L., Chen, K. J., Liu, M. C., & Tsai, M. C. (2000). The Module-Attribute Representation of Verbal Semantics: From semantic to argument structure. *International Journal of Computational Linguistics and Chinese Language Processing*, *5*(1), 19–46.

Huang, C. R., Li, Y. L., Zhong, Y., & Zhu, Y. (2022). A Linked Data Approach to an Accessible Grammar of Chinese for Students. *Chinese Language Learning and Technology*, *2*(1), 1-29. https://doi.org/10.30050/CLLT.202206_2(1).0001

Jiang, L. (2014). *HSK standard course*. Beijing Language and Culture University Press. [姜丽萍. (2014). *HSK 标准教程*. 北京语言大学出版社.]

Krashen, S. D. (1984). Principles and practice in second language acquisition. Pergamon Press.

Ladousse, G. P. (1987). *Role play* (Vol. 3). Oxford University Press.

Lantolf, J. P., & Aljaafreh, A. (1995). Second language learning in the zone of proximal development: A revolutionary experience. *International Journal of Educational Research, 23*(7), 619-632.

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for Parameter-Efficient Prompt Tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* 3045–3059. DOI:10.18653/v1/2021.emnlp-main.243

Li, J. (2022). A study on the Construction of Chinese Near Synonyms Knowledge Base. *Acta Scientiarum Naturalium Universitatis Pekinensis, 1*, 106-112. [李娟. (2022). 汉语近义词辨析知识库构建研究. *北京大学学报 (自然科学版), 1*, 106-112.]

Li, J. (2023). Exploring the differentiation of near synonyms in Smart-Technologies Framework. *2023 International Conference on Asian Language Processing (IALP)*, 370–376. https://doi.org/10.1109/IALP61005.2023.10337004

Liu, L., Shi, Z., Cui, X., Da, J., Tian, Y., Liang, X.,… Hu, X. (2023).The opportunities and challenges of ChatGPT for international Chinese language education: View summary of the Joint Forum of Beijing Language and Culture University and the Chinese Language Teachers Association of America. *Chinese Teaching in the World, 3*, 291-315. [刘利,史中琦,崔希亮,笪骏,田野,梁霞, ... 胡星雨. (2023). ChatGPT 给国际中文教育带来的机遇与挑战——北京语言大学与美国中文教师学会联合论坛专家观点汇辑. *世界汉语教学, 3*, 291-315.]

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship, 49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720

Lu, J., & Lv, W. (2006). The compilation of a monolingual learner's dictionary of Chinese as a foreign language: A venture and some considerations. *Chinese Teaching in the World, 1*, 59-69. [鲁健骥，吕文华. (2006). 编写对外汉语单语学习词典的尝试与思考——《商务馆学汉语词典》编后. *世界汉语教学, 1*, 59-69.]

Lu, J. (1994). Chinese grammar errors analysis of foreign learners. *Language Teaching and Linguistic Studies, 1*, 49-64. [鲁健骥. (1994). 外国人学汉语的语法偏误分析. *语言教学与研究, 1*, 49-64.]

Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge University Press.

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research & Development, 50*(3), 43–59.

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research, 25*, e50638. DOI: 10.2196/50638

Moussalli, S., & Cardoso, W. (2020). Intelligent personal assistants: Can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning, 33*(8), 865–890.

Nigh, M. (2023, June 24). ChatGPT3 prompt engineering. Retrieved from: https://github.com/mattnigh/ChatGPT3-Free-Prompt-List

OpenAI. (n.d.). *Best practices for prompt engineering with the OpenAI API: How to give clear and effective instructions to OpenAI models.* https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and instruction, 1*(2), 117-175.

Reynolds, L., & McDonell, K. (2021). Prompt programming for Large Language Models: Beyond the few-shot paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3411763.3451760

Shin, D., Lee, J. H., & Lee, Y. (2022). An exploratory study on the potential of machine reading comprehension as an instructional scaffolding device in second language reading lessons. *System, 109*, 102863.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting knowledge from Language Models with automatically generated prompts. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 4222–4235. DOI: 10.48550/arXiv.2010.15980

Smith, K. M., & Craig, H. (2013). Enhancing learner autonomy through CALL: A new model in EFL curriculum design. *CALICO Journal, 30*(2), 252.

Van Der Stuyf, R. R. (2002). Scaffolding as a teaching strategy. *Adolescent learning and development, 52*(3), 5-18.

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., & Yin, M. (2024). Unleashing ChatGPT's power: A case study on optimizing information retrieval in Flipped Classrooms via prompt engineering. *IEEE Transactions on Learning Technologies, 17*, 629-641. DOI: 10.1109/TLT.2023.3324714.

Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L. (2023). What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. *Computers & Education*, *194*, 104703. https://doi.org/10.1016/j.compedu.2022.104703

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Fedus, W. (2022). Emergent abilities of Large Language Models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... Zhou, D. 2022. Chain-of-Thought prompting elicits reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *17*(2), 89-100.

Xing, H. (2013). Collocation knowledge and second language lexical acquisition. *Applied Linguistics, 4*, 117-126. [邢红兵. (2013). 词语搭配知识与二语词汇习得研究. *语言文字应用, 4*, 117-126. ]

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., ... Lan, Z. (2020). CLUE: A Chinese language understanding evaluation benchmark. *Proceedings of the 28th International Conference on Computational Linguistics*. 4762-4772. https://aclanthology.org/2020.coling-main.419

Xu, L., Li, A., Zhu, L., Xue, H., Zhu, C., Zhao, K., ... Lan, Z.(2023). SuperCLUE: A comprehensive Chinese Large Language Model benchmark. *arXiv*. https://doi.org/10.48550/arXiv.2307.15020

Yang, J. (2004). How to compare the usage of near synonyms in class. *Chinese Teaching in the World*, *3*,96-104. [杨寄洲. (2004). 课堂教学中怎么进行近义词语用法对比. *世界汉语教学, 3*, 96-104.]

Yang, J., & Jia, Y. (2007). 1700 groups of frequently used Chinese synonyms. *Beijing Language and Culture University Press*. [杨寄洲, 贾永芬. (2007). 1700 对近义词语用法对比: *北京语言大学出版社*.]

Zaghlool, D. Z. D., & Khasawneh, D. M. A. S. (2023). Incorporating the impacts and limitations of AI-driven feedback, evaluation, and real-time conversation tools in foreign language learning. *Migration Letters*, *20*(7), Article 7. https://doi.org/10.59670/ml.v20i7.4863

Zhang, B. (2007). Synonym, near-synonym and confusable word: A perspective transformation from Chinese to interlanguage. *Chinese Teaching in the World, 3*, 98-107+3. [张博. (2007). 同义词、近义词、易混淆词: 从汉语到中介语的视角转移. *世界汉语教学, 3*, 98-107+3.]

Zhang, N., Li, L., Chen, X., Deng, S., Bi, Z., Tan, C., Huang, F., & Chen, H. (2022). Differentiable prompt makes pre-trained language models better few-shot learners. *CoRR, abs/2108.13161*. https://arxiv.org/abs/2108.13161

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.