科技与中文教学

Journal of Technology and Chinese Language Teaching

Volume 15 Number 1, June 2024
二〇二四年六月　第十五卷第一期

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

科技与中文教学

Journal of Technology and Chinese Language Teaching

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

Volume 15 Number 1, June 2024

Managing editor for this issue: Shijuan Liu

Articles

Columns

科技与中文教学

Journal of Technology and Chinese Language Teaching

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

Sponsor

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

Editorial board

Editorial staff

Contacts

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

Can ChatGPT Reliably and Accurately Apply a Rubric to L2 Writing Assessments? The Devil is in the Prompt(s)
(用 ChatGPT 评估二语写作的有效性研究：指令设计的重要作用)

Poole, Frederick J.
Michigan State University
(密西根州立大学)
poolefre@msu.edu

Coss, Matthew D.
Michigan State University
(密西根州立大学)
mattcoss@msu.edu

Abstract: This paper investigates the effectiveness of ChatGPT, a generative AI tool, in assessing second language (L2) writing. The study explores the practicality of employing ChatGPT as an assessment tool, focusing on the accuracy and reliability of the AI-generated scores compared to human raters. Various prompting strategies were tested to understand their impact on the effectiveness of ChatGPT in this context. The paper also examines the reliability of ChatGPT scores across different writing topics. The findings demonstrate that ChatGPT can serve as a valuable tool in L2 writing assessment, provided that it is used strategically with well-crafted prompts. The study contributes to the growing body of research on automated writing assessment tools, particularly in the realm of L2 learning, and offers insights into the practical application of such tools in educational settings.

摘要：本文研究了 ChatGPT 在评估二语（L2）写作中的有效性。研究探讨了使用 ChatGPT 作为评估工具的实用性，重点关注 AI 生成的评分与人工评分相比的准确性和可靠性。为了理解不同指令策略对 ChatGPT 在这一背景下有效性的影响，本研究测试了十种指令策略。本文还检验了 ChatGPT 在不同写作主题上的评分可靠性。研究结果表明，ChatGPT 可以作为 L2 写作评估中的一个有价值的工具，前提是要使用精心设计的指令策略。本研究为自动写作评估工具的研究提供了新的见解，特别是在 L2 学习领域，并提供了这些工具在教育环境中实际应用的见解。

Keywords: Artificial intelligence; automated essay scoring; second language writing; writing assessment; rubrics

关键词: 人工智能，自动作文评分，二语写作，写作评估，评分标准

1. Introduction

In the last year, ChatGPT and other generative Artificial Intelligence (AI) tools have taken the world by storm. ChatGPT was one of the fastest platforms to reach 1 million users and has continued to experience sustained growth and use since its release in November of 2022. Since then, numerous tools with similar functionalities have emerged including Gemini (Google), Claude (Anthropic), and Perplexity (Perplexity.ai), among others. These generative AI tools, also referred to as large language models (LLMs), make use of recent developments in deep neural networks called transformers to optimize text generation capabilities (Vaswani et al., 2017). For many language teachers, administrators, and researchers, the introduction of generative AI tools like ChatGPT into the educational landscape is both exciting and intimidating. These tools have incredible capabilities, making them appealing for a variety of efficiency-improvement purposes. However, uncertainty due to the complexity of these tools' technological underpinnings as well as their trustworthiness for educational purposes remain strong as well. The last year has seen countless examples of users experimenting with ChatGPT and other AI tools to explore their capabilities and limitations. One area that may be particularly appealing to language educators is the potential of using tools like ChatGPT for assessment purposes. Evaluating writing assessments in the L2 classroom can be both time consuming and taxing (Crusan et al., 2016). Asking AI tools to apply a rubric to automatically evaluate student essays would undoubtedly sound attractive to many educators. However, before these tools can be normalized for assessment purposes, it is important to explore approaches to ensure high levels of reliability and accuracy while also considering these tools' practical relevance for teachers and other end-users given their inherent complexity. In other words, while AI tools like ChatGPT offer exciting prospects for optimization in education, the extent to which they are usable and useful to teachers (among others) must be thoroughly explored before recommendations can be made.

Many scholars have noted the time-consuming nature of evaluating assessments (e.g., Crusan et al., 2016) and the difficulty of avoiding human-rater bias and error (e.g., Elder et al., 2007). AI tools seem to be able to assess large amounts of data, including additional language (L2) learner writing, accurately and reliably (e.g., Mizumoto & Eguchi, 2023), but whether such tools can be implemented in the classroom in a practical manner remains unexplored. In this paper, we assert that the primary affordance of ChatGPT as an assessment tool lies in its capacity to expand the analytical capabilities of language educators, assessment specialists, and other professionals by making advanced computational techniques more accessible, regardless of the user's prior technical experience. Thus, we set out to explore strategies for prompting ChatGPT to produce reliable, accurate, and interpretable results for L2 writing assessments. We focus on prompting strategies, as we argue that this is the most accessible and impactful strategy for language educators to employ ChatGPT as an automated assessment scoring tool.

In this study, we analyze a series of prompts to demonstrate how effective prompting can empower teachers, our primary stakeholders, to employ ChatGPT successfully, bypassing some of the technical knowledge required to extract usable information from assessment data in prior research (see Mizumoto & Eguchi, 2023). Based on our analysis of these prompts and their different levels of reliability, we offer language educators and other language program stakeholders a list of considerations to improve reliability of generative AI as assessment scoring tools, with important emphasis on how and when these tools should or should not be used.

2. Literature review

2.1 Automated writing assessment tools

There is a long history of research on developing automated writing assessment tools. Much of this research explores tools created by large testing or publishing companies such as *e-rater* by Educational Testing Service, *Intelligent Essay Assessor* by Pearson Education, or *Intellimetric* by Vantage Learning among others (Hussein et al., 2019). These systems typically include both an automated scoring system as well as an automated feedback system. Research exploring automated feedback in these systems tends to focus on student and teacher perceptions of feedback and the impact of the tool on writing quality (e.g. Link et al., 2022). In contrast, research exploring the automation of assessment scores focuses on how similar automated scores are to human raters. In this study we are primarily concerned with automated scoring which is often referred to as automated essay scoring (AES).

AES systems have been used primarily in high-stakes assessments due to the cost of developing them. The most common approach to developing AES systems involves first using human raters to evaluate essays. Then collecting numerous automatically generated indices of text quality, and finally applying statistical approaches to identify which combination of these indices correlate with human scores best (Attali, 2015). Through the years these AES tools have advanced by adding more complex indicators such as readability scores and other text features extracted with natural language processing techniques (e.g., cohesion scores, syntactic complexity), as well as more complex statistical approaches (e.g., Bayesian text classification, Deep Neural Networks) (Huawei & Aryadoust, 2023; Hussein et al., 2019).

Several systematic reviews illustrate that AES tools can be quite accurate, but results vary substantially (e.g., Hussein et al., 2019; Ramesh & Sanampudi, 2022). While most studies have found that AES tools tend to correlate strongly with human scoring (>.7), some studies have noted inaccuracies. For instance, Wang and Brown (2007) found that over 25% of students received failing scores for a writing placement test (for L1 speakers) by human raters, while only 2% received a similar score by the AES tool. Wang (2015) found that while EFL learners appreciated the quick feedback from an AES tool, *Criterion*, only 8% of students (n=53) who used the tool believed that it applied the writing rubric objectively and reliably to their writing. Furthermore, scholars have argued that AES tools may both misrepresent the writing construct and encourage a change in writing behavior to take advantage of weighted scoring systems (Deane, 2013). It is important to note that research on AES tools has been primarily (>90%) conducted with English language learners or L1 speakers of English (Huawei & Aryadoust, 2023), with few studies exploring other languages. Additionally, Reilly et al. (2014) noted in their study using an AES tool in an open online course that the AES tool was more accurate for L1 speakers of English than for L2 speakers of English. Qian et al., (2020) evaluated the *iWrite* system for L2 learners of English in China and concluded that the system failed to report accurate scores reliably. Thus, while these AES tools are continuing to improve, there is still some concern in terms of how accurately they are able to assess the written output of L2 learners.

Although much of the research has focused on the English language, there is a growing body of research on AES tools for the Chinese language. Yang et al. (2023) conducted a systematic review exploring such tools. In the 29 studies that they identified, 11 included data on language learners rather than L1 Speakers. The studies investigated corpora that ranged in size from 100 samples from a standardized L2 Chinese exam (the Hanyu Shuiping Kaoshi, HSK) to over 85,000 texts from L1 speakers of Chinese. The studies used a variety of metrics to evaluate the validity of scores produced from AES tools including Agreement Rate, Exact Agreement Rate, Pearson Correlation Coefficient, and Quadratic Weighted Kappa (QWK). The QWK scores ranged from 0.60 to 0.88, with the highest score for L2 learners reaching 0.714.

AES tools show great promise for L2 learners but to date they have been used with a very limited population for very specific purposes (e.g., mostly for English speakers on large scale, high stakes exams). As noted earlier, much of the research is dominated by large testing corporations who charge high prices for these assessments. Even when the costs are relatively low (~$4 per test) as is the case with ACTFL's new AES tool[1], testing groups of learners multiple times (i.e., the typical multiple assessments given in a language course or program) quickly increases the cost. This inevitably limits who can use AES tools and when and why they are applied. AES tools that are not developed and managed by large testing corporations often require high levels of technical and statistical expertise, which also limits who can use or develop these tools. In this study, we view the emergence of ChatGPT as a potential opportunity to explore a wider range of applications of AES for users with varying levels of technical and statistical expertise.

[1] https://www.actfl.org/news/actfl-and-lti-introduce-groundbreaking-automated-scoring-system-for-the-aappl-spanish-presentational-writing-component

2.2 ChatGPT and L2 writing assessment

Even in the first year since the release of ChatGPT, there have been many articles published on the applicability of using ChatGPT as an assessment tool. Most recently, Pfau et al. (2023) compared ChatGPT 3.5 Turbo's ability to identify errors with that of human raters using a corpus of essays at multiple proficiency levels produced by L1 Greek L2 English writers. They found that although ChatGPT did miss some errors, it was still strongly correlated with human raters (*r*=0.97). They note that even though human editing is still needed, ChatGPT greatly increases efficiency when identifying errors. Similarly, Jiang et al. (2023) also used ChatGPT in addition to three other AI tools to automatically identify errors in L2 Chinese writers. Similarly they found that AI models were very accurate with most of their models reaching around .8 accuracy levels. While being able to identify errors is important, it does not in itself lead to an assessment score.

In another study exploring the use of ChatGPT as an assessment tool for English language learners, Mizumoto and Eguchi (2023) used an IETLS TASK 2 rubric as the query (prompt) and used it to analyze 12,100 essays from the TOEFL11 test. The essays were previously rated by humans by separating them into either low, medium, or high levels on a five-point scale (following Blanchard et al., 2013), though little information is given on how these essays were scored. Mizumoto and Eguchi found that while ChatGPT had acceptable levels of reliability (quadratic weighted kappa~=0.38), a number of other statistical measures (e.g. GPT scores + Lexical measures + Syntactic complexity measures, + others) were needed to improve the scores to a QWK of .6. While this is promising, it again highlights the technical expertise needed to achieve accurate and reliable scores, thus undercutting a major affordance of tools like ChatGPT.

It is important to note that both studies only used and evaluated one prompt in their analyses and involved advanced English language learners (similar to other studies on AES tools). Further, there was no mention of the temperature parameters in either of these studies. These are not trivial points as they can impact the outcome of a query in ChatGPT significantly. Temperature in ChatGPT is a value between 0 and 1 that reflects the amount of variance or randomness that is allowed in a response to a prompt. The default setting is 0.7 which is argued to be the ideal setting for generating human-like text. This is somewhat problematic for assessments as scores given by ChatGPT will vary depending on the temperature level. For example, in Mizumoto and Eguchi's (2023) study, they noted that when running the same analysis twice their scores varied. Ultimately, they argued that this variance was acceptable, but if they had lowered or raised the temperature level, their reliability score between the two scores would undoubtedly follow suit. Thus, it is not unreasonable to assume that their results will vary at different temperature settings as it has in other studies (Coyne et al., 2023). While having a lower temperature may be ideal for returning numeric values, having a higher temperature may be needed when getting qualitative feedback or details on errors in a sentence as was the case in Pfau et al. (2023).

Coyne et al. (2023) also explored the use of ChatGPT as an assessment tool with English data that included errors. They were interested in exploring how well ChatGPT engaged in grammar correction. The authors identified 20 English sentences with errors and then explored how ten different prompts performed in identifying the grammar errors compared to human raters. They found that overall GPT-4 performed well in identifying errors and tended to perform better at lower temperatures. Equally important they illustrate that prompt engineering, the iterative process of developing effective prompts for generative AI, is an important factor in determining the effectiveness of ChatGPT as an assessment rating tool. With a temperature of .1, their prompts ranged in GLEU scores (a metric for error correction) from 0.31 to 0.582 across different prompts.

In this paper we argue that studies exploring ChatGPT should report both temperature and prompting strategies. But more importantly, we should explore the use of ChatGPT in a way that aligns with the affordances provided by the tool. Therefore, we argue that accuracy and reliability can be increased with effective prompting strategies. OpenAI has suggestions for improving prompting strategies, such as including more details in queries for relevant answers, asking ChatGPT to take on a role, using delimiters to indicate distinct parts of the prompt, specifying steps required to complete a task and asking the model to reflect on those, providing examples, and specifying desired output length (https://platform.openai.com/docs/guides/gpt-best-practices). In the next section we highlight the potential affordances of automated assessments and generative AI specifically as they do (and might) relate to L2 classrooms and discuss the practical implications of these tools for such contexts.

2.3 Evaluating ChatGPT for classroom-based assessments

A number of frameworks have been developed to assess and evaluate the use of AES tools (e.g. Williamson et al., 2012). These frameworks generally focus on constructing relevance and representation, accuracy of scores, generalization, extrapolation, and use of scores (e.g. Enright & Quinlan, 2010). Given that these areas of focus all depend on the use of score, and subsequently the consequences and impact of a score, it is reasonable to first explore this area and move backwards. Ferrara and Qunbar (2022) note when discussing validity claims for AES, we must explicitly delimit the scope of the claims to be made about an assessment. In other words, in order to make a claim about the appropriateness of the inferences derived from a particular assessment, one must first clarify the type and nature of the assessment.

In our study, we are specifically considering the use of ChatGPT for classroom-based assessments. Classroom-based assessments are, simply put, assessments that are conducted in a classroom setting by a teacher (as opposed to, for example, large-scale standardized assessments). Exploring the potential role of using automated assessments in the classroom setting requires that we first explore potential needs that such tools can fill. Classroom-based assessment includes weekly quizzes, unit tests, exit tickets, among others.

Although classroom-based assessments are usually described as either being formative or summative in nature, that is, *for* learning and *of* learning, respectively, Black and Wiliam (1998) argue that formative and summative are not properties of assessments *inherent* to the assessments themselves, but rather are properties of the *uses* of the information gathered from assessments. In other words, inferences, conclusions, and data can be *used* formatively or in a summative manner, even with the same assessment. Additional use of assessments in language learning programs include for diagnostic purposes and/or placement testing, but these usually occur outside of the classroom setting by a program coordinator or administrator.

Assessments that are used for formative purposes tend to involve more qualitative feedback rather than simply providing a learner with a score. This is because assessments that are used formatively aim to improve learning rather than simply measure it. This would suggest research involving the effectiveness of an automated assessment system that is targeting formative skills should focus on how well and relevant the feedback given by the system is. In contrast, summative assessments tend to have an accountability and/or administrative role in education. These assessments come at the end of an instructional unit or course and provide evidence of the extent to which learners have achieved established goals. Because information from summative assessments is often passed to other stakeholders (e.g. parents and administrators), quantitative evaluations are used for ease of communication and convenience. These assessments tend to involve higher stakes as scores usually impact learner grades and thus these assessments may have a gatekeeping effect (Winke, 2021). Thus, for automated scoring that is being applied to summative assessment data, the focus should be on reliability and accuracy of the tool's ability to generate a score.

In the present study, we are exploring ChatGPT's capacity to assess Chinese L2 writing samples reliably and accurately. We specifically consider how language educators may make use of this tool in their classroom settings and thus we explore approaches that are practical for in-class implementations. Given that we are focusing on primarily the accuracy of ChatGPT's ability to generate a proficiency score (a summative use of assessment), we are focusing on the potential use of this tool serve as a second rater or as a tool for learners to engage in self-assessment practices.

With this mind, we consider measurements for confirming accuracy and reliability of scores generated by automated assessments. Williamson et al. (2012) argued that for high stakes assessments at ETS, their threshold for accuracy using a quadratic weighted kappa measurement (QWK) was 0.7. Automated assessments at ETS include the GRE and TOEFL, among others. These are tests that usually cost individuals over $100 and have gatekeeping roles for graduate school (Winke, 2021). Compared to classroom-based assessments, these have significantly more impact on one's future and thus while 0.7 is a good benchmark for evaluation it is reasonable to consider a lower threshold for classroom-based assessments. It is also important to note that writing topic or task has also been shown to impact writing outcomes (James, 2008), and thus it is important to confirm that scores are reliable across writing tasks.

Finally, when considering relevance and representation, one must consider how scores are derived and how they map onto constructs that are being measured. In traditional automated assessment models score generation are quite intuitive. AES tools usually have a set of text metrics generated by Natural Language Processing techniques that represent parts of the writing construct. For example, Quinlan et al. (2009) provide a detailed overview of how 30 different indices (e.g. fragments, run-ons, proper nouns, etc.) map onto 8 subconstructs (e.g. Grammar, Usage, Mechanics, Style, Organization, Development, Lexical Complexity, and Topic-specific vocabulary usage) and further how those subconstructs are connected to writing standards. This is somewhat problematic with ChatGPT and other LLMs given that there is less transparency regarding how results are generated as they employ ‘black-box modeling approaches’ (Bauer & Zapata-Rivera, 2020, p. 24). In other words, one may ask ChatGPT to apply a rubric to a text (Mizumoto & Eguchi, 2023) or to generate similar metrics as found in other AES studies (e.g. count the number of fragments), but it is unclear how such metrics are actually calculated or how a rubric is applied (or not) to a text. While we cannot directly address this issue in this study, it is important to acknowledge when investigating the reliability and accuracy of ChatGPT as an assessment tool.

Thus, our study is guided by the following research questions:

1. How do prompting strategies affect the accuracy of ChatGPT generated scores compared to human raters?
2. Are ChatGPT scores reliable across different tasks?

3. Methods

3.1 Data set

Data from the present study were taken from a corpus of third semester university L2 Chinese learners (n=48) from a private university in the United States. As part of their regular coursework, these students completed a standardized L2 proficiency assessment of listening, speaking, reading, and writing during the final week of their semester. Students in this study ranged from a writing score of 4 (N=18) to 7 (N=6) on individual tasks (possible scores ranged from 1-9), corresponding to Intermediate Low and Advanced Low on the ACTFL proficiency scale, respectively. See Table 1 for a complete breakdown of students' scores on individual writing tasks by level.

Table 1 Frequency of Writing Scores by Human Raters

Score	ACTFL Proficiency Level	Counts

4	Intermediate Low	18
5	Intermediate Mid	57
6	Intermediate High	63
7	Advanced Low	6
Total		144

*Note each of the 48 students were scored on 3 writing tasks.

Data from the present study consists of each student's three writing tasks responses in this standardized assessment (n=144). The standardized assessment uses a computer-adaptive system, meaning that the difficulty level of writing task was determined based on their reading scores (computer-scored multiple-choice questions). There was a total of 9 possible tasks[2], 3 of which each targeted low-intermediate, intermediate, and advanced, respectively. Task level (intermediate-advanced) was determined by reading scores; task order was randomly assigned. The number of students who took each task at varying times (e.g. Time 1, Time 2, & Time 3) can be seen in Table 2. All students completed the tasks in the assigned order. Each writing task was scored holistically by one or two professional human raters and assigned a numeric score from 1-9, corresponding to Novice Low through Advanced High (CEFR levels A1 to C1) on the ACTFL scale. The present study, therefore, used the writing tasks, the students' responses, and the official assessment scores (from raters) to evaluate the efficacy of automated scoring using ChatGPT.

Table 2 Number of students assigned to each writing task

Prompt	Targeted Level	Time			
		1	2	3	Total
Newspaper	Intermediate	15	12	11	38
Lost in forest	Intermediate	12	14	12	38
Appliance	Intermediate	11	12	15	38
New pet	Low-Intermediate	4	2	2	8
Letter of appreciation	Low-Intermediate	3	1	4	8
Live anywhere	Low-Intermediate	1	5	2	8
Time in history	Advanced	1	1	0	2
Positive in hardship	Advanced	1	0	1	2
City council	Advanced	0	1	1	2

[2] Because the standardized test is a commercial test with copyright restrictions, the precise prompts cannot be shared here.

3.2 Data analysis

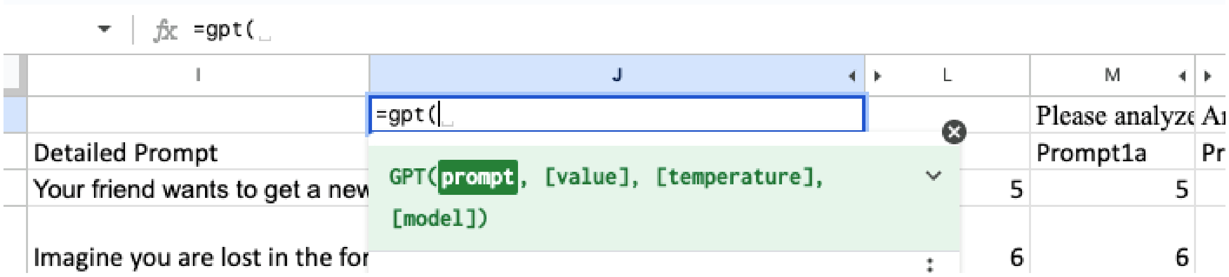
To assess the reliability of the scores generated by ChatGPT in this study, we use four reliability measurements including exact and adjacent agreement percentages, Pearson's correlation, and quadratic weighted kappa (QWK). Exact agreement percentage reflects the amount of exact agreement between the human rater and ChatGPT scores. Adjacent agreement percentages refer to scores by ChatGPT that were within 1 point (below or above) human rater scores. QWK is commonly used to quantify the degree to which measurements resemble each other (Williamson et al., 2012). Unlike correlation coefficients, QWK accounts for both correlation and agreement between measurements. In other words, while correlations may pick up on trends in similar directions, QWK also illustrates how close two scores are to each other. QWK is therefore more appropriate for assessing reliability than Pearson's *r* when there is systematic variability between raters or measurements for the same subject (Vanbelle, 2016). Another option for measuring interrater reliability is Cohen's kappa; however, this is limited to categorical ratings. Since the scores used in this study are ordinal numeric response options, QWK is more appropriate reliability indexes than Cohen's kappa. We report multiple metrics to ensure accuracy as suggested by recent studies (e.g. Doewes et al., 2023).

Additionally, to investigate fairness of ChatGPT in scoring these essays, we also use a mixed-effects regression to explore ChatGPT's scores across multiple writing tasks. Mixed-effects models are ideal when data are nested. In our study, we have participants who are scored on three different writing prompts at three different times. Given the likely effect of individual and time of writing (e.g. first writing task vs second or third writing task), we added these variables as random effect intercepts to the model. Additionally, we control for differences in proficiency and time spent on task by adding these variables as fixed variables. No interactions were added to this model. We first created a null model with only proficiency and time spent on the assessment entered into the model, and then we added a categorical variable for the writing topic. To make this variable more interpretable, we use effect coding which means that instead of having a reference variable with which to compare the effect of writing task, individual tasks are instead compared to a grand mean. These findings will be reviewed in the results section.

3.3 Technical considerations for analyzing text with ChatGPT

There are a few technical considerations that must be considered with ChatGPT. First, because we are analyzing 144 texts, it is not practical to use the browser-based platform for the analysis. Most users of ChatGPT simply navigate to chat.openai.com to submit a prompt. If we were to analyze our essays through the browser, we would need to copy and paste both a prompt and a text 144 times and then manually add scores to a database to be analyzed later. This would be a cumbersome process for us (and for any educator who is interested in using ChatGPT for assessment purposes). Additionally, for assessment purposes we want to adjust the temperature on ChatGPT. This cannot be done through the browser.

In contrast to using a browser to submit prompts, ChatGPT also be accessed by using the Application Programming Interface (API) through a programing language like Python. Google Sheets has an extension that also allows users to access ChatGPT through the API[3]. This extension allows a user to query ChatGPT from within a spreadsheet. Figures 1 illustrates how one can define a cell using a call to ChatGPT. In the image 'prompt' refers to the message that will be sent to ChatGPT, value is the text that is to be analyzed, temperature receives a value between 0 and 1, and model refers to the version of ChatGPT that one wants to use. For this study we used GPT-4 and set our temperature to 0.1 to reduce variability of responses. By using a Google Sheet, we can upload all data including the text to be analyzed into one sheet. This can greatly increase efficiency when it comes to applying ChatGPT to multiple texts.



```
fx =gpt(
```

I	J	L	M	
	=gpt(		Please analyze	A
Detailed Prompt			Prompt1a	Pr
Your friend wants to get a new		5	5	
Imagine you are lost in the for		6	6	

```
GPT(prompt, [value], [temperature], [model])
```

Figure 1 GPT in Google Sheets

It is also important to note that using the ChatGPT API in this way is not free and requires that users register with a credit card. GPT-3.5 Turbo costs $0.0015 (USD) per token for input, and $0.002 per token for output. While GPT-4 is significantly more at $0.03 per token for input, and $0.06 per token for output. Because our output is only 1 number, we are mainly focused on the cost of the input, which takes into account the length of the text the students write as well as the length of our prompt. Understanding the exact conversion from words to tokens is complicated because tokens are not directly related to letters or words, but rather to chunks of text. It is estimated that approximately 1000 tokens is equivalent to 750 words in English and about 1.7 tokens is equivalent to 1 character in Chinese. However, it is important to emphasize that these are estimates. For this reason, it is not possible to give an exact cost for each prompt analyzed, but to be transparent, we can report that we spent $76.03 to analyze 144 Chinese texts 10 times (for 10 prompts) with an average of 305 Chinese characters per text analyzed. Our prompts ranged from 298 characters to 6367 characters (including both English letters and Chinese characters) with an average of 1007 characters. This cost comes to approximately $7.60 per prompt or about $0.05 to analyze one text. Notably, OpenAI recently changed the cost of API use and has

[3] https://workspace.google.com/marketplace/app/gpt_for_sheets_and_docs/677318054654

reduced costs by half. The prices we report here reflect the pricing structure at the time of analysis (September-October 2023).

3.4 Prompt engineering

Similar to Coyne et al. (2023), we used 10 prompts (see Appendix) to explore how unique ChatGPT queries result in different outcomes for each student's test responses. In our first prompt, we start by asking ChatGPT to analyze student writings using the ACTFL scale without providing descriptions of the scale itself. We clarify that we only wanted a numeric value, returned. In our second prompt we become more detailed and provide simple descriptions for each individual proficiency level. In prompt three, we change to the AVANT descriptors (the developer and administrator of standardized assessment from which our data were collected). AVANT rubrics are based largely on ACTFL scales and descriptions, but they do use slightly different terminology. In prompt four, we apply a set of discrete rules that AVANT shared via presentation about their scoring procedures. This prompt relies on ChatGPT's ability to apply logical rules to essay scoring. In prompt five, we add the entire rubric from AVANT similar to what Mizumoto and Eguchi (2023) did in their study. In prompt six, we apply a specific strategy from OpenAI which suggests providing ChatGPT with a step-by-step procedure. In Prompts seven and eight we provide specific examples of what an essay at each level should look like. Prompt seven received one example and Prompt eight received two examples. Prompt nine is the same as prompt eight, except we used Chinese to prompt ChatGPT rather than English. Finally, prompt ten provides generic examples (e.g. not specific to the task) of each writing level.

Table 3 List and Descriptions of Prompts

Prompting Number	Prompting Strategy	Brief Description
1	Simple: No descriptions	Analyze using known knowledge about ACTFL scale
2	Simple: Apply Logic (ACTFL)	Add a description of each level
3	Simple: Apply Logic using AVANT descriptors	Add details from Avant
4	Simple: Rule-based: Avant	Apply clear cut off points
5	Complex: Complete Rubric from Avant	Complete Rubric
6	Complex: Detailed Step-by-Step Procedure	Step-by-step
7	Provide Examples: 1 Example	One-shot prompting
8	Provide Examples: 2 Examples	Two-shot prompting
9	Provide Examples: Same as P8 but in Chinese	Chinese Two-shot Prompting
10	Provide Examples: Generic Examples	Generic Examples

4. Results

Table 4 illustrates the findings from the ten prompts that we applied in our study. When using different prompts we found that correlations between ChatGPT and human rated scores ranged from 0.23 to 0.58. However, given the nature of the proficiency scales (i.e., an ordinal, nine-point scale), using the QWK is more appropriate for evaluating the accuracy of these prompts. The QWK scores range from 0.17 to 0.57 depending on the prompt used, with the most accurate scores coming from our 8th prompt. It is also important to explore the adjacent agreement given that these scores are on a nine-point scale. In other words, if a learner scores a 4 on the human rated assessments but receives a 5 from ChatGPT, the difference is between an Intermediate Low and an Intermediate Mid, this is not terribly concerning given that most students are assumed to be operating at a level above or below their proficiency level due to a number of factors (see Clifford, 2016, for discussion). In terms of adjacent agreement, we found a range between 74.3% and 97.2% with Prompt 2, 6, 7, 8, 9, and 10 all scoring over 90%.

Table 4 Similarity Measures

	Prompts									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Exact Agreement %	27.1	47.9	10.4	7.6	33.3	50.7	52.7	49.3	41.7	47.9
Adjacent Agreement %	74.3	97.2	45.8	40.3	86.8	92.4	96.5	93.8	95.8	95.8
Pearson's Correlation	0.23	0.45	0.42	0.53	0.54	0.42	0.49	0.58	0.50	0.45
Quadratic Weighted Kappa	0.17	0.44	0.18	0.18	0.37	0.38	0.48	0.57	0.42	0.45

We also provide a visual (See Figure 2) of the correlation and QWK scores to illustrate an issue with using correlation scores to assess automated scored. The scores are ordered from highest QWK to lowest. Prompt 5 and 4 both have significant correlation scores over 0.5, yet their QWK scores are much lower than their correlation coefficients. This suggests that there is some consistency with how ChatGPT is applying scores, but that the scores are not aligning with the scales being used (e.g. ACTFL's 1-9 scale).

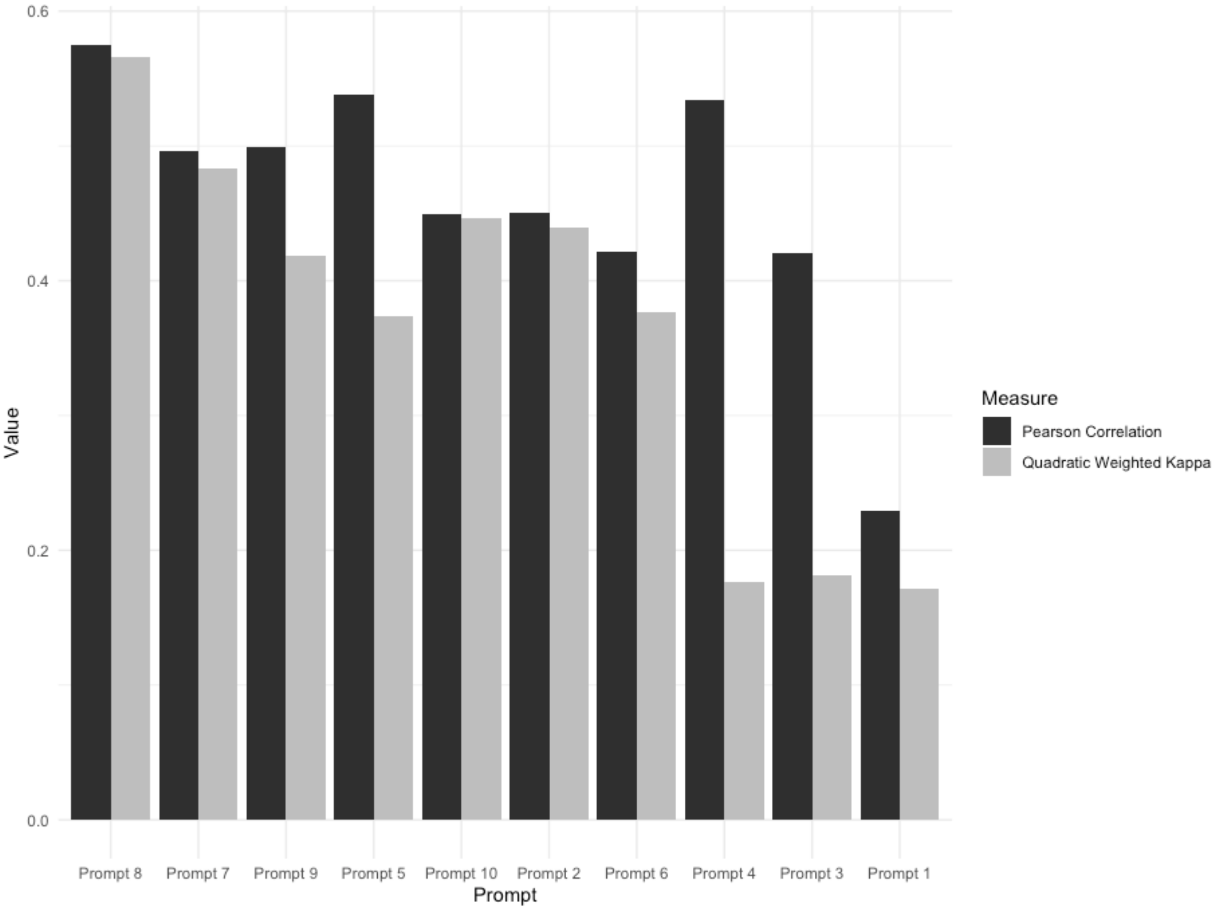
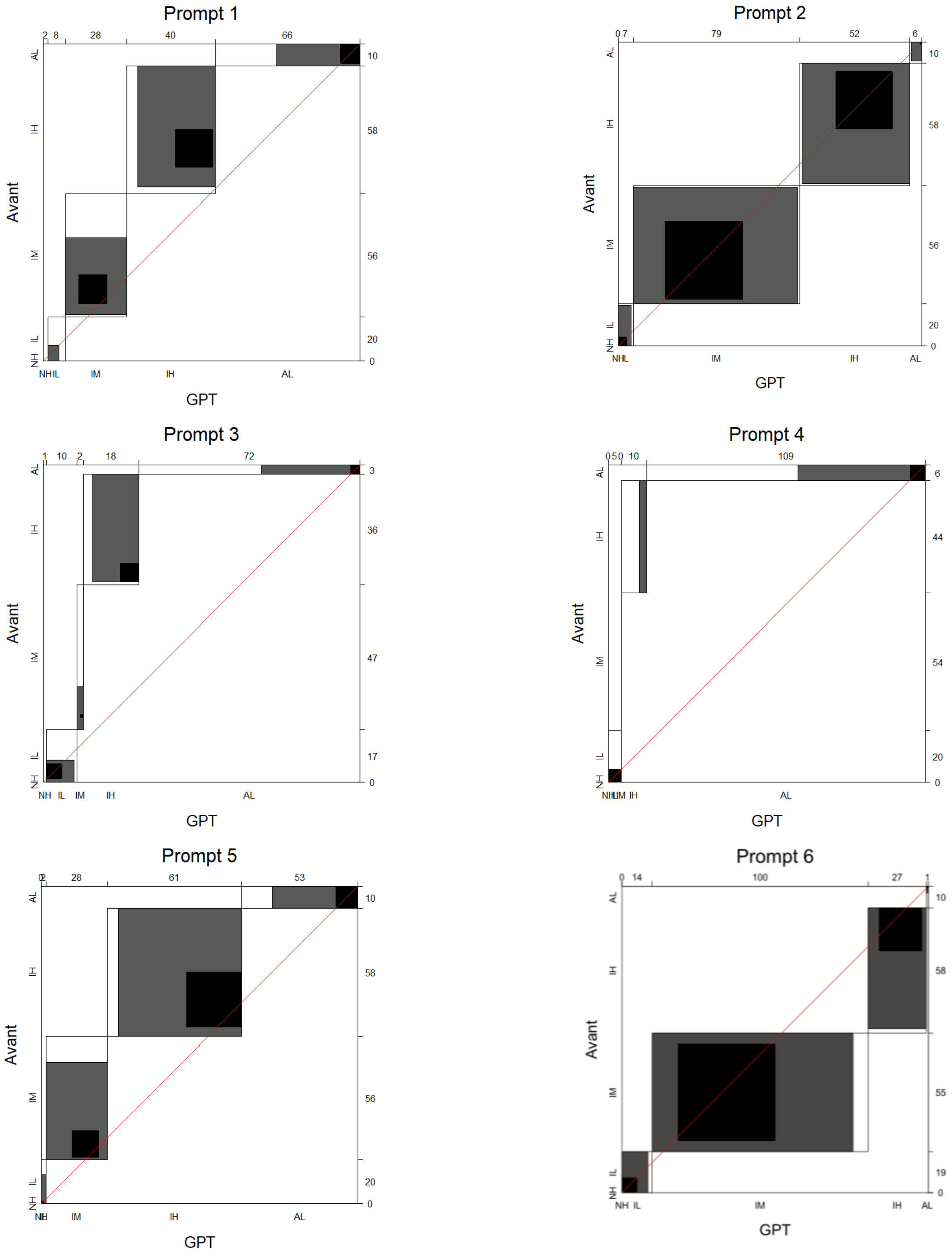


Figure 2 Comparison of Pearson Correlation and QWK for Each Prompt

To continue exploring these prompts visually, we generated a series of adjacency plots for each prompt. For the visuals (Figure 3), black boxes represent an exact match between human-rated and ChatGPT scores. Grey boxes represent examples of a 1-point difference between human-rated and ChatGPT scores. White boxes represent examples that have a larger than 1 point difference between human-rated and ChatGPT scores. Thus, we are looking for visuals with large black boxes, smaller grey boxes, and even smaller white boxes. Furthermore, prompts that have boxes centered on the diagonal represent ChatGPT scores that are more closely correlated with human-rated scores.


Prompt 1
2 8 28 40 66
10 58 56 20 0
AL IH IM IL NH
NHIL IM IH AL
Avant
GPT
Prompt 2
0 7 79 52 6
10 58 56 20 0
AL IH IM IL NH
NHIL IM IH AL
Avant
GPT
Prompt 3
1 10 2 18 72
3 36 47 17 0
AL IH IM IL NH
NH IL IM IH AL
Avant
GPT
Prompt 4
0 5 0 10 109
6 44 54 20 0
AL IH IM IL NH
NHILIM IH AL
Avant
GPT
Prompt 5
0 2 28 61 53
10 58 56 20 0
AL IH IM IL NH
NIL IM IH AL
Avant
GPT
Prompt 6
0 14 100 27 1
10 58 55 19 0
AL IH IM IL NH
NH IL IM IH AL
Avant
GPT

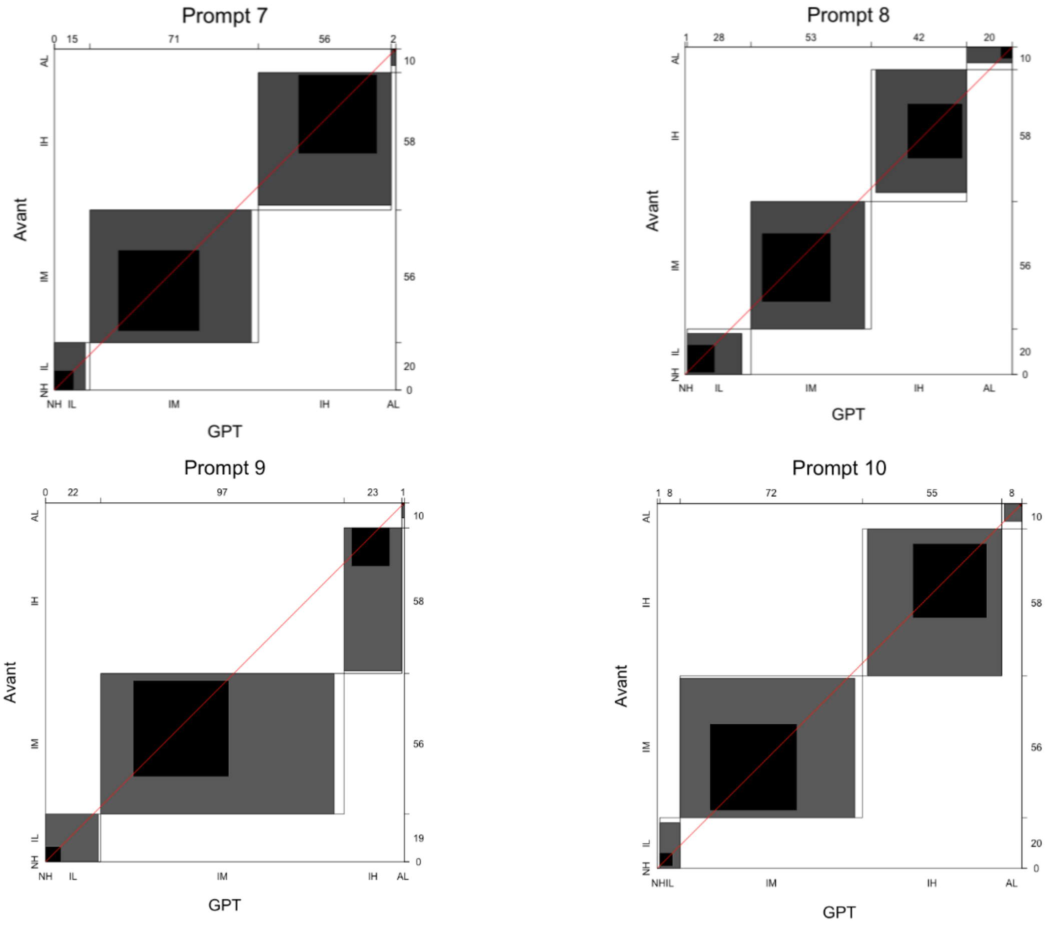


Figure 3 Exploring Prompt Performance via Adjacency Plots

Looking at Prompts 7, 8, it is clear that there are minimal examples of scores that diverge by more than two points, while prompts 3 and 4 are clearly problematic. Interestingly, Prompt 9, which is the same as Prompt 8 except it was written in Chinese performed worse.

To explore our second research question, we conducted a mixed-effects regression model to determine if ChatGPT scores are reliable across writing tasks. We used both individual participant and task order as random effects and compared the variance in random effects of individual and task between ChatGPT and human-rated scores. In both cases, individual differences account for large portions of the variance in scores with individual clusters accounting for ~26% of the variance in ChatGPT scores, and ~22% of the variance in human-rated scores. This is reasonable since we have varying proficiency levels in our data set. The variance associated with order of tasks is moderate in both cases at ~8% and ~6% respectively, but this does illustrate that task order plays a role in final scores. Further analysis shows that scores tend to decrease as order of task increases. This is likely due to a fatigue effect and further establishes the need for a mixed-effects model.

Table 5 Mixed-effects Regression Results

	ChatGPT (P8)		Human-Rated Score	
	Null	Full	Null	Full
Proficiency	0.677***	0.665***	0.772**	0.816***
	(0.146)	(0.195)	(0.111)	(0.151)
Time Spent on Assessment (minutes)	0.003	0.004	0.007	0.006
	(0.005)	(0.005)	(0.004)	(0.004)
Appliance		-0.558**		0.098
		(0.196)		(0.154)
Letter of Appreciation		-0.286		-0.030
		(0.325)		(0.258)
Live Anywhere		0.055		0.179
		(0.326)		(0.259)
Lost in Forest		-0.244		-0.002
		(0.196)		(0.154)
New Pet		-0.133		0.152
		(0.325)		(0.258)
Newspaper		0.261		-0.040
		(0.196)		(0.154)
Positive Hardship		0.658		0.742
		(0.511)		(0.413)
City Council		-0.258		-0.722
		(0.511)		(0.413)
Constant	1.457	1.630	0.764	0.510
	(0.857)	(1.105)	(0.650)	(0.853)
Observations	144	144	144	144
Log Likelihood	-187.748	-177.236	-148.214	-149.110
Akaike Inf. Crit.	387.497	382.472	308.427	326.220
Bayesian Inf. Crit.	405.316	424.050	326.246	367.797

Note: Topic is effect coded. *p**p***p<0.001

Table 5 demonstrates that when controlling for proficiency and time spent on task, writing task does predict outcomes for ChatGPT while it does not for human raters. Interestingly, the 'appliance' prompt was associated with more than a half-point lower score compared to other prompts. This is not the case for the human rated assessments. These findings will be explored further in the discussion section.

5. Discussion

In this paper, we set out to explore the effectiveness of ChatGPT to automatically apply a rubric to Chinese L2 writers. To date we are unaware of other studies that have explored the use of ChatGPT to assess L2 Chinese writers other than Jiang et al. (2023) which primarily focused on error detection. More importantly, we position our research as addressing the potential practicality of using these tools in classroom settings. With this in mind, we considered the time, technical expertise needed, and cost of implementing AES tools. In terms of technical expertise and time, we acknowledge that any approach that requires developing expertise in statistical measures and/or software is unlikely to be integrated into mainstream teaching practices. Thus, we focused on unique prompting strategies that can impact the accuracy of ChatGPT to assess writings, which we argue that any teacher would be readily able to implement without extensive training. More specifically, we applied rubrics specifically designed for the writing samples to student writings automatically with the help of ChatGPT. In our prompting strategies, we kept the prompts short to reduce costs while also adhering to best practices provided by OpenAI. In our series of prompts, we were detailed yet concise, we added logical steps for ChatGPT to follow, we tried prompts in both English and Chinese, and we tried prompts that included examples of performance at each level of the rubric, all of which teachers could be readily expected to do for classroom-based summative assessments.

To answer our first research question, we discovered that prompting strategies have a profound impact on scoring accuracy. Our results show that Pearson's *r* correlation scores ranged between 0.24 and 0.57 and QWK scores similarly ranged between 0.17 and 0.58. These are large differences and were primarily due to how ChatGPT was prompted. If generative AI tools are to be used widely, it is clear that training users on how to prompt ChatGPT for assessment purposes is needed. Further, steps to ensure reliability and accuracy are also needed. In our study, we found that using multiple examples in lieu of detailed descriptions of levels in a rubric performed the best, however, even with our best prompt we noticed some discrepancies between ChatGPT and the human raters. When we explored the performance of the prompts more closely, we noted that some students' scores were more closely aligned with human raters, while others diverged more. To better visualize this we plotted each individual's writing scores on three different writing prompts (see Figure 4). Purple shading represents writing tasks in which both human-raters and ChatGPT scored participants exactly equivalently. For example, participant #142 was given a 5 on all three writing prompts by both human raters and ChatGPT. Examples like this are most ideal for making robust validity claims about ChatGPT as an assessment tool. The colors blue (human-raters) and pink (ChatGPT) indicate scores that were not overlapping. Thus, Participant 135, for example, was given a 5 on three of their writing samples by human raters and then two of these samples were given 4 by ChatGPT while the third score was given a 6. For participant 144, both ChatGPT and human raters scored one writing sample as a 7, while two samples were given a 6 by ChatGPT and two were given a 5 by human raters.

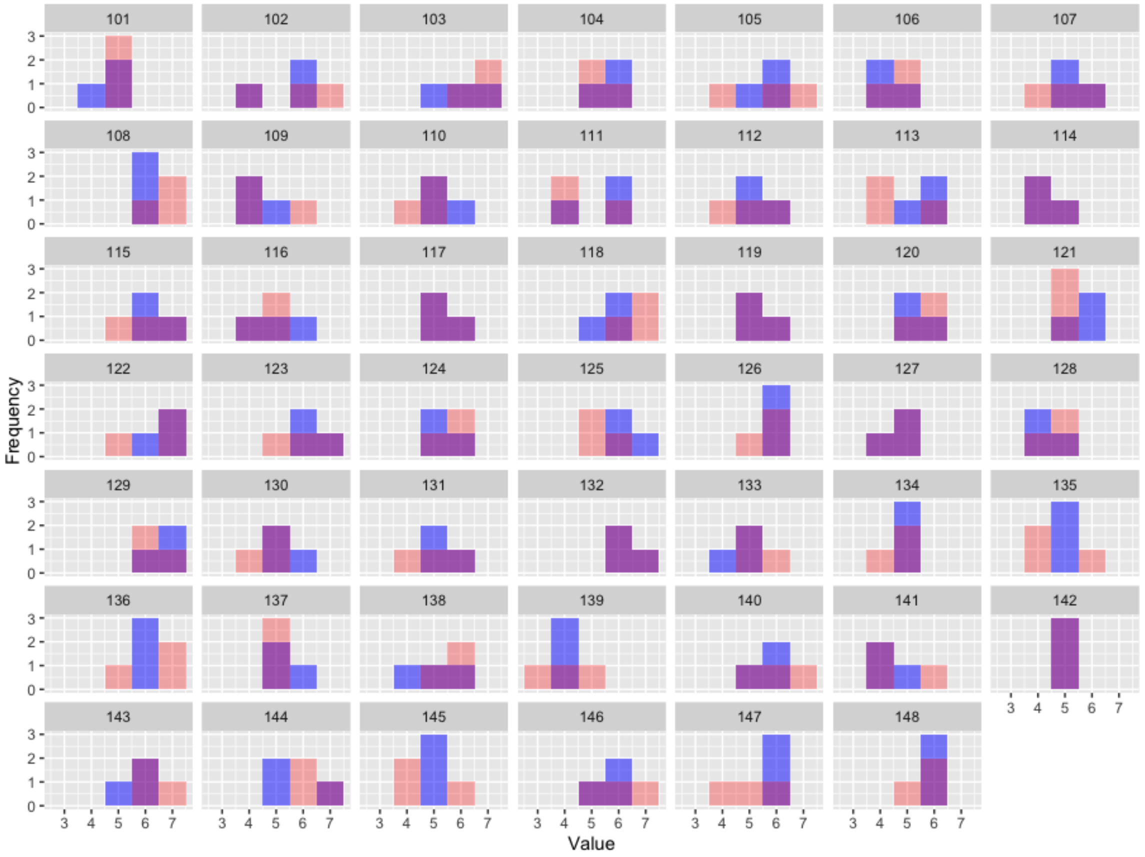


Figure 4 Rater and GPT Score Convergences and Divergences

We were not able to detect any trends between students who were scored poorly by ChatGPT in comparison to human raters. However with more data, identifying commonalities between students who were consistently scored incorrectly may be possible. Such data may provide insight into how ChatGPT is actually applying rubrics (i.e., help us dig further into its 'black box' mechanism).

Additionally, we explored the impact of writing topics on the reliability of our best prompt (prompt 8) visually. Figure 5 shows that most of the writing prompts came from either writing about an appliance that one finds to be useful (appliance), what one would do if they were lost in a forest (lost in forest) and one's perception about the relevance of newspapers in today's society (newspaper). The other topics had relatively fewer responses. Looking at the number of instances of exact agreement, ChatGPT seems to have performed better on the *newspaper* topic, while *appliance* and *lost in the forest* tended to have more diverging scores. However, statistically only the *appliance* writing topic showed scores that differed significantly from other writing topics (being systematically lower by about half a point on the 9-point scale). Similar to our discussion above on how individuals were scored, exploring performance on prompts can also lead to valuable insight into how ChatGPT applies rubrics. For example, future research may want to extract all of the misclassified essays from the *appliance* prompt to determine if any themes emerge.

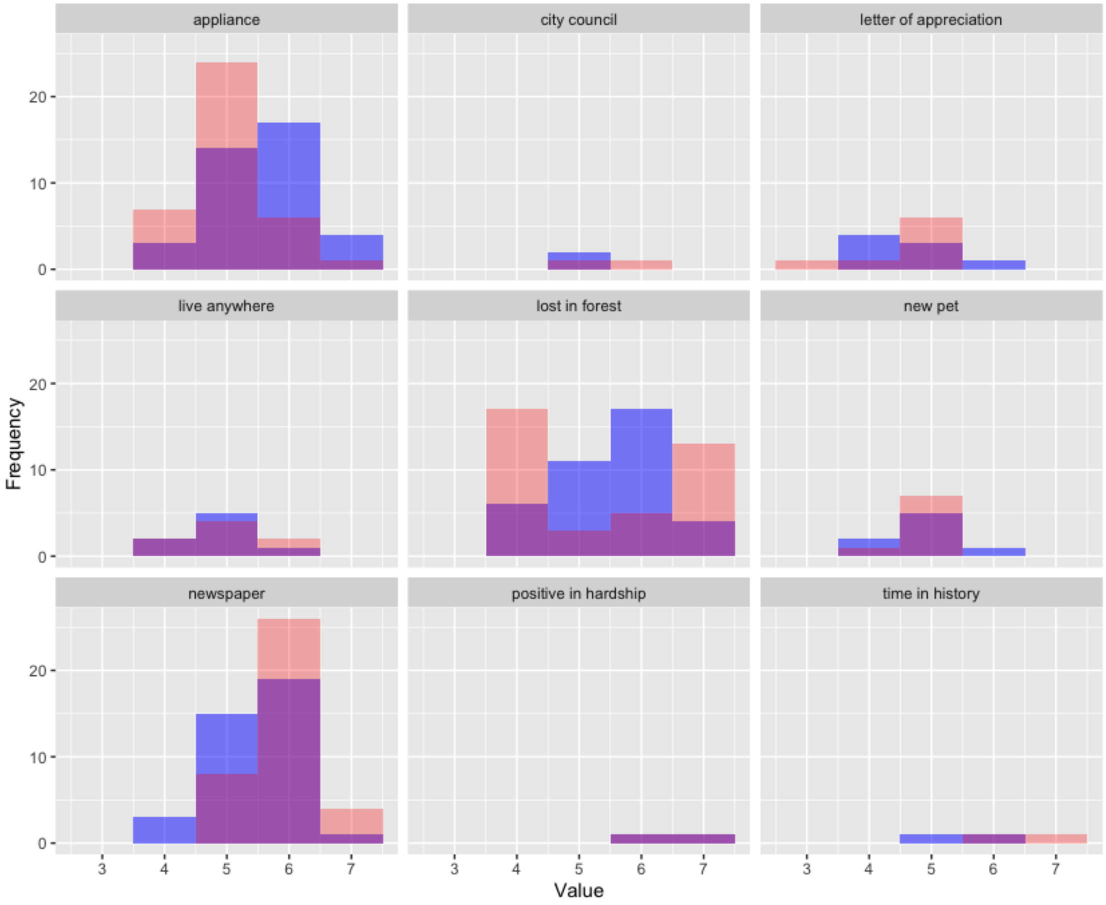


Figure 5 Rater and GPT Score Convergences and Divergences by topic

6. Conclusion

Our study did not find accuracy scores at levels reported in other AES studies. As noted earlier, many previous studies have generated scores that better approximate human-rated scores, with a number of studies finding QWK values over 0.7 (i.e., the threshold identified by ETS). That being said, many of the tools that have achieved or surpassed that threshold are either expensive or require technical expertise. Both of these caveats limit the widespread use of these tools. Additionally, it's important to note that many of the previous AES tools were created with a specific task and text type in mind. In our study, we applied one ChatGPT prompt to multiple writing tasks and found that scores were fairly reliable across tasks. This is an important consideration for a classroom teacher who likely will not be able to customize their tool to each writing assignment.

Although our study did not find that ChatGPT reached a desirable reliability threshold, we still argue that it can be used as an assessment tool for certain cases in classroom-based assessments. The first and most obvious use case is as a second coder.

ETS and other testing corporations often argue that AES tools should only be used as second coders (Ramineni & Williamson, 2018). Only a few testing companies rely primarily on an AES tool. Classroom teachers rarely have time to check scores or allow a second coder to check even a small portion of their graded papers (raising questions about reliability, especially for higher stakes classroom-based assessments like final course exams). Using ChatGPT as a second coder may help identify potential biases and/or errors for classroom-based assessments. As we noted in our study, many of our prompts were within 1-point of the human raters on a 9-point scale more than 90% of the time. As a second rater we argue that a 0.57 QWK with a +90% adjacent rater agreement is more than sufficient. For educators looking to use this tool, we suggest running an automated assessment with ChatGPT and then identifying any cases in which ChatGPT is more than 2 points off the human rater score. This does not automatically mean that the human rater was wrong, but it does provide a good starting point for reflecting on scores and further analyzing individual cases of highly divergent scores (including, possibly, prompting opportunities for further conceptual and analytical alignment within language programs or among colleagues).

In addition to using ChatGPT as a second coder, we also believe that it could be used as a self-assessment tool for language learners. Research has shown that writing in an L2 can benefit language learners (Polio & Park, 2016). However, teachers are often reluctant to assign writing without assessments. Using a self-assessment framework in which students write an essay, use ChatGPT to self-assess, and then reflect on the perceived accuracy of ChatGPT may not only increase the amount of writing that learners engage in but also it may support the development of metacognitive skills as well as digital literacy skills in relation to these new AI tools (Poole & Polio, 2024) as well as language proficiency literacy (see Coss and Van Gorp, forthcoming). Further, because this is used as a reflection tool rather than as a grading tool, any issue with accuracy is less concerning, as these can be mitigated by teacher-led or peer-to-peer discussion.

Regardless of how AI tools are used, our study highlights the importance of training teachers in how to best maximize both accuracy and reliability. The biggest takeaway that our study can offer at this point is that prompting matters. Luckily, there are easily-applied strategies that can greatly (relatively) enhance the reliability of ChatGPT-generated assessment score results. For example, the reliability scores in our study suggest that the best results come when a teacher uses past scored student examples or current examples to provide ChatGPT with an example of what writing looks like at each level. Prompts with examples, therefore, may be the optimal strategy for maximizing the reliability of ChatGPT for the uses we have discussed here.

6.1 Limitations

There are a few key limitations to our study. First, we had a limited range of scores on the ACTFL scale and only 48 participants. Ideally, we would have had an equal number of participants at each level of the ACTFL scale with an equal distribution across writing tasks. That being said, our participants did range four levels on the ACTFL scale, and our sample is likely to reflect that of a foreign language classroom in which this tool may be used (i.e., Intermediate-level courses). Nevertheless, future studies should also explore the reliability and accuracy of this tool for novice and advanced learners. Secondly, we only explored 10 ChatGPT prompts, there are undoubtedly other ways of prompting this tool which may lead to better outcomes. Recently OpenAi has released updates that allow 'Plus' members to create their own ChatGPT that is customized to their needs. Creating a custom ChatGPT that has a database of learners past writings with human-rated scores may prove to be more accurate, reliable, and practical for language educators. Finally, we only explored one language, Chinese. It is likely that ChatGPT will perform better on these assessment tasks with languages that are better represented in ChatGPT's training data (e.g., English). To confirm this, future studies should explore variation in assessment accuracy across multiple languages. Finally, our study was focused on more summative uses of assessment evaluation. Future studies should examine the extent to which ChatGPT and similar tools are able to offer formative or diagnostic feedback, and the extent to which these tools could be incorporated systematically into language classrooms for these important, recurring purposes. In this line of research, the perceptions of stakeholder (students, teachers, etc.) would be important to explore concurrently with the accuracy and reliability of ChatGPT.

References

Attali, Y. (2015). Reliability-Based Feature Weighting for Automated Essay Scoring. *Applied Psychological Measurement*, *39*(4), 303–313. https://doi.org/10.1177/0146621614561630

Bauer, M. I., & Zapata-Rivera, D. (2020). 'Cognitive Foundations of Automated Scoring', in Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). *Handbook of Automated Scoring*. CRC Press, pp. 13–28. https://doi.org/10.1201/9781351264808-2

Black, P., & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment*. GL Assessment.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series* (i-15). https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals*, *49*(2), 224–234. https://doi.org/10.1111/flan.12190

Coss, M. D., & Van Gorp, K. M. (forthcoming). *What proficiency levels do K-16 world language learners achieve? An ACTFL Research Brief.* ACTFL.

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. arXiv. https://doi.org/10.48550/arXiv.2303.14342

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, 28, 43–56. https://doi.org/10.1016/j.asw.2016.03.001

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Doewes, A., Kurdhi, N., & Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. *16th International Conference on Educational Data Mining*. Germany.

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37–64. https://doi.org/10.1177/0265532207071513

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317–334. https://doi.org/10.1177/0265532210363144

Ferrara, S., & Qunbar, S. (2022). Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. *Journal of Educational Measurement*, *59*(3), 288-313.

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*(1), 771–795. https://doi.org/10.1007/s10639-022-11200-7

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5, e208. https://doi.org/10.7717/peerj-cs.208

James, C. L. (2008). Electronic scoring of essays: Does topic matter?. *Assessing Writing*, *13*(2), 80–92. https://doi.org/10.1016/j.asw.2008.05.001

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages*, *8*(4), 247.

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2). https://doi.org/10.1016/j.resmal.2023.100050

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, *2*(3). https://doi.org/10.1016/j.resmal.2023.100083

Polio, C., & Park, J. H. (2016). Language development in second language writing. In Manchón, R. M. & Matsuda, P. (Eds.). *Handbook of Second and Foreign Language Writing*. de Gruyter, pp. 287–306. https://doi.org/10.1515/9781614511335-017

Poole, F. J., & Polio, C. (2024). From sci-fi to the classroom: Implications of AI in task-based writing. *TASK: Journal on Task-Based Language Teaching*, 3(2), 243-272.

Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's Automated Essay Scoring System iWrite. *Journal of Educational Computing Research*, *58*(4), 771–790. https://doi.org/10.1177/0735633119881472

Quinlan, T., Higgins, D., & Wolff, T. (2009). *Evaluating the Construct Coverage of the e-rater® Scoring Engine*. Research Report No. RR-09-01. Educational Testing Service.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series*, 2018(1), 1–31. https://doi.org/10.1002/ets2.12211

Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open and Distributed Learning*, *15*(5), 83-98. https://doi.org/10.19173/irrodl.v15i5.1857

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika*, 81, 399–410. https://doi.org/10.1007/s11336-014-9439-4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 31st Conference on Neural Information Processing Systems. CA, USA: Long Beach, (Available at) https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning and Assessment*, *6*(2).

Wang, P. L. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching*, *12*(1), 140–157.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: Issues and practice*, *31*(1), 2-13.

Winke, P. (2021). Foreword. In Mirhosseini, S. A. &De Costa, P. (Eds.). *The Sociopolitics of English Language Testing*. Bloomsbury, pp. vii–ix. https://doi.org/10.5040/9781350136025.0009

Yang, H., He, Y., Bu, X., Xu, H., & Guo, W. (2023). Automatic Essay Evaluation Technologies in Chinese Writing—A Systematic Literature Review. *Applied Sciences*, *13*(19). https://doi.org/10.3390/app131910737

Appendix

Prompts

Download the prompts here or from https://osf.io/y7utf/.

Volume 15 Number 1, June 2024
pp. 25-48

Predicting Chinese Language Learners' ChatGPT Acceptance in Oral Language Practices: The Role of Learning Motivation and Willingness to Communicate
(预测中文学习者在口语练习中对 ChatGPT 的接受度：学习动机和交流意愿的作用)

Li, Nuoen
(李诺恩)
The Chinese University of Hong Kong
(香港中文大学)
nuoenli@link.cuhk.edu.hk

Zhang, Lan
(张岚)
The Chinese University of Hong Kong
(香港中文大学)
1155128312@link.cuhk.edu.hk

Lau, Kit Ling
(刘洁玲)
The Chinese University of Hong Kong
(香港中文大学)
dinkylau@cuhk.edu.hk

Liang, Yu
(梁宇)
Beijing Language and Culture University
(北京语言大学)
liangyu@blcu.edu.cn

Abstract: Despite an increased interest in the potential benefits of ChatGPT for foreign language education, learners' intentions to use ChatGPT as a learning tool have so far received little research attention. This study aims at exploring Chinese language learners' acceptance of ChatGPT in oral language practices and its influencing factors based on the technology acceptance model (TAM). Data were collected from 375 Mongolian learners who learned Chinese as a foreign language (CFL) and analyzed by means of partial least squares structural equation modeling (PLS-SEM). The results indicated that learning motivation and willingness to communicate are critical antecedents of ChatGPT acceptance, and willingness to communicate has a critical mediating role on the link between the three motivational determinants (self-efficacy, utility value, and attainment value) and TAM variables. Prediction-oriented segmentation (POS) was further carried out and found unobserved heterogeneity among CFL learners' formation of ChatGPT acceptance rooted in the years of Chinese learning. The findings suggest the theoretical strengths of TAM in explaining CFL learners' adoption of AI-assisted language practices. Meanwhile, it underlies the importance to understand learners' psychological attributes before introducing technology-assisted speaking practices. Pedagogical insights into how to enhance ChatGPT acceptance among different learner populations were also offered.

摘要：尽管 ChatGPT 在外语教育中的潜在优势逐渐成为热点话题，但学习者将 ChatGPT 作为学习工具的意愿至今未受到充分的研究关注。

本研究旨在基于技术接受模型（TAM），探究中文学习者在口语练习中对 ChatGPT 的接受度及其影响因素。研究数据收集自 375 名蒙古中文学习者，并采用偏最小二乘结构方程模型（PLS-SEM）进行分析。结果表明，学习动机和交流意愿是影响 ChatGPT 接受度的关键前因，而交流意愿在三个动机因素（自我效能、实用价值与成就价值）与 TAM 变量之间的关联路径中发挥着重要的中介作用。预测导向分割（POS）分析进一步揭示了不同学习年限的中文学习者在 ChatGPT 接受度形成机制中存在异质性。上述研究结果证实了 TAM 在解释中文学习者采纳 AI 辅助语言实践方面的理论优势。同时，这也强调在引入技术辅助中文口语练习活动前了解学习者心理的重要性。此外，本文针对如何增强 ChatGPT 在不同学习者群体中的接受度提出了教学建议。

Keywords: ChatGPT, technology acceptance, oral language practices, learning motivation, willingness to communicate

关键词：ChatGPT、技术接受、口语练习、学习动机、交流意愿

1. Introduction

Artificial intelligence (AI) creates a novel paradigm for promoting the efficiency, effectiveness, and outcomes of teaching and learning across a wide range of educational settings. In foreign language (FL) education, chatbots can serve as a tireless language partner for learners to practice speaking with, or even an effective tutor or instructor to deliver extra knowledge that a language partner may not be able to provide due to their limited language proficiency level (Huang et al., 2022). Many studies have revealed the potential contributions that AI-powered chatbots could bring to oral language practices, such as improving language accuracy and fluency (Ruan et al., 2021), mitigating learners' anxiety (Hsu et al., 2023; Jeon, 2022), and enhancing engagement in speaking activities (Jeon, 2022; Ruan et al., 2021). Though these advantages have been generally acknowledged by researchers in the field, the integration of chatbots into FL learning practices greatly depends on learners' awareness of chatbots' practical value and their willingness to adopt them as regular learning tools. In this regard, the perceptions and acceptance of chatbots among FL learners warrant further research attention.

Previous research has sought to understand learners' acceptance of chatbots and its influencing factors. Several concerns, however, have appeared in the earlier investigations. First, there is a lack of empirical support for theoretical assumptions of information systems (IS) acceptance. The majority of relevant research was based on the technology acceptance model (TAM) (Davis, 1989), arguably the most popular yet parsimonious model in the IS field (Srite, 2006). However, certain theoretically conceptualized relationships between TAM variables have not been empirically validated yet (e.g., Liu & Ma, 2024), implying that further research is still necessary to determine the applicability of TAM in exploring chatbot acceptance. Second, the potential unobserved heterogeneity of learners' chatbot acceptance has received insufficient attention. According to Becker et al. (2013), unobserved heterogeneity captures situations where there is no clear theoretical account for heterogeneity in a certain population, in contrast to observed heterogeneity where prior knowledge about the group differences has been acquired. Existing literature has shown inconsistent findings regarding the relationships between TAM variables in the context of AI-assisted FL learning, which might be attributed to variations in participants' backgrounds across those studies. For instance, the hypothesized positive impact of perceived ease of use on attitudes failed to reach a statistically significant level in Liu and Ma (2024) with English language learners from various backgrounds in China, whereas it was supported by Belda-Medina and Calvo-Ferrer (2022) through a survey among college-level English language learners in Spain and Poland. Hence, the heterogeneity regarding the developing pattern of chatbot acceptance among learner populations with different backgrounds warrants further investigation. Third, little research effort was devoted to FL education, and there has been insufficient emphasis on learners' acceptance of chatbots for speaking practices. According to Petrović and Jovanović (2021), the most natural and effective application of chatbots is related to their fundamental nature—language practice. The capacity of chatbots could provide valuable learning opportunities, particularly for FL learners to practices their language either in text-based or oral-based manner, which requires special attention in the FL field. The recently released ChatGPT considerably expands the technological affordances of GenAI-powered chatbots in enabling better oral-based communication by providing customized feedback, answering follow-up questions, and generating more authentic and natural conversations (Kohnke et al., 2023; Tlili et al., 2023). Therefore, it is worthwhile to investigate FL learners' acceptance of ChatGPT, especially in oral language practices.

Motivation has been found to be an important source of users' acceptance and usage behavior of information technologies (Venkatesh, 2000; Venkatesh et al., 2003). To date, a range of motivational determinants has been identified as essential external variables for chatbot adoption and usage behaviors among FL learners, such as hedonic motivation (Strzelecki, 2023), perceived enjoyment (Chen et al., 2020), and perceived autonomy, relatedness, and competence (Xia et al., 2023). It has also been found that in FL learning, learners who have stronger motivation towards the language they learn are more inclined to actively seek out advantageous technology to optimize their learning experience (Hsu, 2017). Thus, understanding technology acceptance from an academic-learning motivational perspective would offer valuable insights into the matter. Furthermore, learners' willingness to communicate (WTC), which refers to their readiness to enter into discourse using a second or foreign language, is particularly crucial for their decisions to initiate communication as a volitional process (MacIntyre et al., 1998). When it comes to FL oral language practices, whether learners voluntarily commit to the advantages of technology in offering communication opportunities might also be greatly determined by their willingness to enter into discourse using the target language. Therefore, WTC should be taken into consideration as a critical individual difference factor influencing FL learners' adoption and usage, especially for oral-based interaction-enabling technologies.

In light of the above discussions, this study aims to explore CFL learners' acceptance of ChatGPT in oral language practices, and the role of learning motivation and WTC in affecting their acceptance by means of partial least squares structural equation modeling (PLS-SEM). Prediction-oriented segmentation (POS), a distance-based segment detection method in PLS path models, will be employed to investigate if there is any unobserved heterogeneity among learners' ChatGPT acceptance. The specific research questions that guide this study are:

1. Are the theoretical assumptions between TAM variables supported in the context of using ChatGPT in CFL oral language practices?
2. Do learning motivation and willingness to communicate have significant effects on CFL learners' ChatGPT acceptance in oral language practices?
3. Is there any unobserved heterogeneity among CFL learners regarding their ChatGPT acceptance in oral language practices?

2. Theoretical foundation and model development

2.1 Technology acceptance model

TAM is a well-established model that aims to explain and predict how users accept and use information technologies. According to TAM (Figure 1), the most proximal antecedent of technology use is behavioral intention, and whether an individual intend to use or reject the technology is determined by his/her attitude toward using the given technology. The attitude of the individual was considered to be affected by two key factors: (1) perceived ease of use (PEOU), which refers to ‘the degree to which a person believes that using a particular system would be free of effort’; and (2) perceived usefulness (PU), which refers to ‘the degree to which a person believes that using a particular system would enhance his/her job performance’ (Davis, 1989, p. 320). It also posits that PEOU has a significant positive effect on PU. In addition, Davis et al. (1989) suggested that an individual might form a strong behavioral intention towards a certain behavior they believe will increase their job performance, thus deriving the hypothesis regarding the direct positive effect of PU on behavioral intention. Hence, the following hypotheses were proposed on the basis of TAM's theoretical underpinnings:

H1: Perceived ease of use (PEOU) has a positive effect on perceived usefulness (PU).
H2: Perceived ease of use (PEOU) has a positive effect on attitude toward using (ATU).
H3: Perceived usefulness (PU) has a positive effect on attitude toward using (ATU).
H4: Perceived usefulness (PU) has a positive effect on behavioral intention to use (BIU).
H5: Attitude toward using (ATU) has a positive effect on behavioral intention to use (BIU).

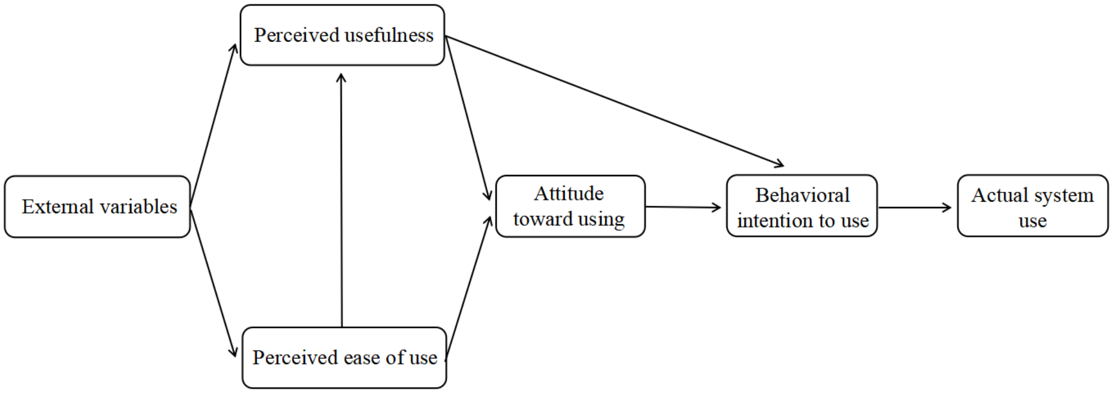


Figure 1 Technology acceptance model (Davis et al., 1989)

2.2 Relations between learning motivation and willingness to communicate

Learning motivation is an important individual difference variable that have been extensively investigated in FL research, which refers to an amalgamation of desires, attitudes, and efforts that encourage learners to learn the target language (Gardner, 1985). Among current motivation theories, expectancy-value theory (EVT) (Eccles-Parsons et al., 1983) demonstrated valuable promise in analyzing academic learning motivation with its overarching theoretical construct (Loh, 2019; Wang & Xue, 2022). Specifically, the conceptualized motivational determinants in EVT not only concerned learners' ability beliefs with *expectancy*, but also various motivational valences of the subjective task with *task values* in terms of shaping self-schema, achieving instrumentality, and offering enjoyment or pleasure.

In a meta-analysis of WTC with 64 studies, Elahi Shirvan et al. (2019) identified motivation as a key variable that influences foreign/second language learners' WTC. The predictive role of learning motivation in WTC has also been found with motivational determinants conceptualized in the EVT framework. In EVT, *expectancy* refers to individuals' beliefs about how they would do on upcoming tasks (Eccels-Parsons et al., 1983). It is highly related to *self-efficacy* that proposed in Bandura (1997), which comprises learners' beliefs on their competence to accomplish a certain task (Wigfield & Eccles, 2000). Therefore, self-efficacy has common be used as one important variable to measure expectancy component in the EVT in empirical research (Bai et al., 2020). The potential positive impact of self-efficacy on WTC has been theoretically reflected in MacIntyre et al.'s (1998) pyramid model of L2 WTC, which conceptualized learners' L2 self-confidence as a critical antecedent of WTC. Empirically, the positive influence of self-efficacy on WTC has also been supported in both traditional language classrooms (e.g., Yang & Lian, 2023) and digital language learning contexts (e.g., Soyoof, 2023; Zadorozhnyy & Lee, 2023).

Another essential aspect of EVT motivation, *task value*, refers to the incentives and reasons for choosing to do a certain work or activity (Eccles-Parsons et al., 1983). There are four components consist of task values: *utility value*, *attainment value*, *intrinsic value*, and *cost*. Utility value, or usefulness, has been defined as how well a particular task fits into individuals' present or future plans; attainment value is the importance of doing well on a given task; and intrinsic value refers to the enjoyment that one gains from doing a task (Eccles-Parsons et al., 1983; Wigfield & Eccles, 2000). Finally, cost is conceptualized as any negative aspect of engaging in a task (e.g., losing alternative opportunities, spending extra efforts, and causing negative emotions). Since cost is a multifaceted mechanism that greatly varies across individuals and still lacks detailed measures exhaustively listing its sources (Rosenzweig et al., 2019), this study solely focused on the former three types of task values.

Prior studies have linked the former three aspects of task values to WTC and revealed a significant relationship between the two constructs. Integrating utility value, attainment value, and intrinsic value as a composite variable, MacIntyre and Blackie (2012) found a significant relationship between task values and WTC among high school L2 French learners. Nagle (2021) also identified attainment value and intrinsic value as significant predictors of WTC with college-level L2 Spanish learners. Based on the comprehensive descriptive insights that EVT could offer into learners' motivational systems and the aforementioned empirical evidence about the influences of EVT motivational determinants on WTC, we formulated the following hypotheses:

H6: Self-efficacy (SE) has a positive effect on willingness to communicate (WTC).
H7: Utility value (UV) has a positive effect on willingness to communicate (WTC).
H8: Attainment value (AV) has a positive effect on willingness to communicate (WTC).
H9: Intrinsic value (IV) has a positive effect on willingness to communicate (WTC).

2.3 Relations between willingness to communicate and technology acceptance

Individual difference is an important sort of antecedent that determines learners' adoption and use of information technologies, which specifically influences PEOU and PU (Venkatesh & Bala, 2008). Previous studies have identified WTC as an important individual characteristic that affects learners' communication behaviors, either in in-class, out-of-class, or digital settings (e.g., Balouchi & Samad, 2021; Lee & Hsieh, 2019; Lee & Lee, 2020). Specifically, Lee and colleagues (Lee & Hsieh, 2019; Lee & Lee, 2020; Lee & Drajati, 2019) found that learners with higher levels of WTC were more likely to attach greater value to language communication, have positive perceptions about information technologies, and seek more opportunities to practice their language communicative skills in assistance with educational technologies. Such influencing links may arise from the nature of WTC as a final psychological step before actual language communication (Lee, 2020). The great value that communication-oriented learners attach to computer-assisted language interaction might also lead to their active cognitive involvement in the meaningful construction of the provided language input, the adaptation of communication strategies they used, and the close attention to functional features of assisted technologies in the speaking tasks (e.g., Mystkowska-Wiertelak, 2021), and as a result, foster positive perceptions about the usefulness and usability of communication-enabling technologies. Especially when it comes to oral language practices, a challenging task for many language learners that requires high cognitive engagement to understand spoken language and initiate communication immediately and effectively (Hsu et al., 2023), WTC might serve an even more crucial role in learners' perceptions towards chatbots and thereafter adoption decisions. Therefore, we proposed that:

H10: Willingness to communicate (WTC) has a positive effect on perceived ease of use (PEOU).
H11: Willingness to communicate (WTC) has a positive effect on perceived usefulness (PU).

2.4 The hypothesized model

Based on the above hypotheses, we developed a research model to predict FL learners' ChatGPT acceptance in oral language practices (Figure 2). Individual differences were explicitly targeted as the external variables, which include learning motivation and willingness to communicate. Learning motivation was further conceptually specified based on EVT, which consists of self-efficacy, utility value, attainment value, and intrinsic value. The technology acceptance construct was developed based on TAM, which consists of perceived ease of use, perceived usefulness, attitude toward using, and behavioral intention to use. In addition, Lee and Lu (2023) showed that learning motivation significantly predicted WTC in both classroom and digital learning environments, and learners with high WTC tend to enthusiastically seek opportunities for text-based or oral-based interactions with information and communication technologies. Thus, we also hypothesized that WTC has a mediating effect on the link between learning motivation and technology acceptance.

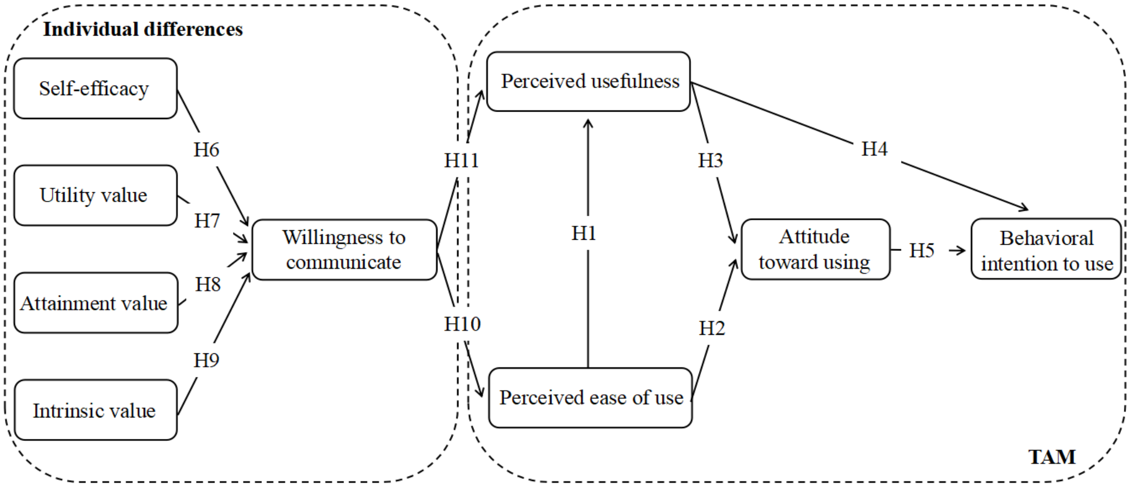


Figure 2 Proposed research model

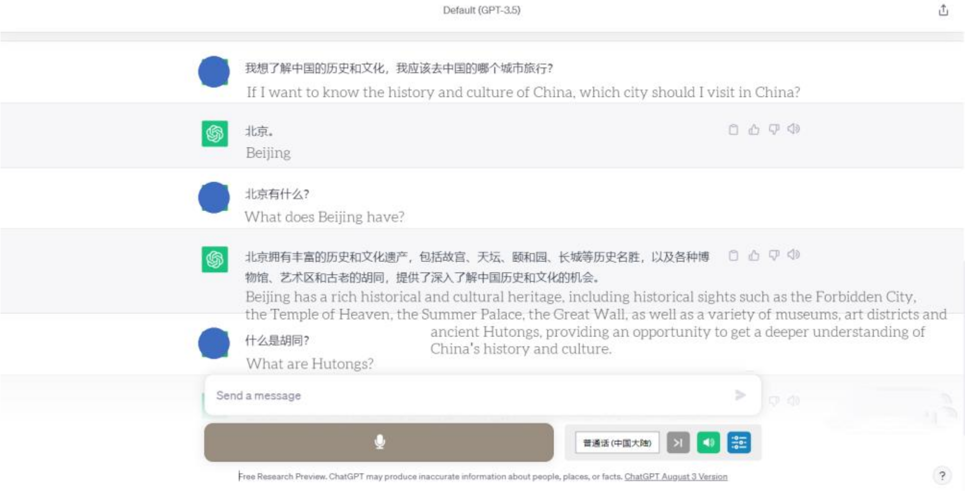
3. Research method

3.1 Research context and participants

This study was conducted at two comprehensive universities in Mongolia with the target to college-level learners who learned Chinese-as-a-foreign-language (CFL) as a compulsory course. To ensure all the potential participants formed clear perceptions towards ChatGPT, this study conducted group-based oral Chinese practice activities with ChatGPT-3.5 using the Chrome extension ‘Voice for Control ChatGPT’ in a total of 12 CFL classes prior to data collection, with the following steps: (1) learners were randomly divided into groups of three to five by their CFL instructor at first; (2) five minutes were then given to learners to discuss the topic they intended to speak about with ChatGPT, which was either based on personal interests or referenced the topics provided by their instructor. The oral practice topics that the instructor provided were designed according to learners' CFL textbooks and differentiated by learners' average level of language proficiency across classes, as shown in Table 1; (3) each group took turns participating in the discussion activity with ChatGPT (five to ten minutes) during class, and every learner in the group was required to take at least two conversational turns with ChatGPT in this process. A discussion example was illustrated in Figure 3, where the group of learners were curious about the best place to visit in China. Through the discussion with ChatGPT, learners in the group finally gained more knowledge about the Hutongs and reached an agreement to travel to Beijing; (4) the activities were repeated twice in one week with the above-mentioned group format and activity procedures. Learners were also encouraged to explore the use of ChatGPT as a chatbot in their extracurricular time to get a clearer understanding of the functions and features of ChatGPT.

Table 1 The samples of oral practice topic

Target learner	Topic	Sample initiating question
Beginner level	See doctors	What should I do if I am sick?
Intermediate level	Traveling	What is the best place to visit in China?
Advanced level	Chinese New Year	How do Chinese celebrate Chinese New Year?



我想了解中国的历史和文化，我应该去中国的哪个城市旅行?
If I want to know the history and culture of China, which city should I visit in China?

北京。
Beijing

北京有什么?
What does Beijing have?

北京拥有丰富的历史和文化遗产，包括故宫、天坛、颐和园、长城等历史名胜，以及各种博物馆、艺术区和古老的胡同，提供了深入了解中国历史和文化的机会。
Beijing has a rich historical and cultural heritage, including historical sights such as the Forbidden City, the Temple of Heaven, the Summer Palace, the Great Wall, as well as a variety of museums, art districts and ancient Hutongs, providing an opportunity to get a deeper understanding of China's history and culture.

什么是胡同?
What are Hutongs?

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

Figure 3 An example of group work in the discussion activities with ChatGPT

Following the completion of the activities, all CFL learners were invited to complete an unidentifiable questionnaire through the online questionnaire tool Wen Juan Xing. Only those learners who completed both the group-based oral practice activities and the questionnaire survey were regarded as final research participants in this study. The demographic information of the final 375 participants is presented in Table 2.

Table 2 Demographic information of research participants (*N* = 375)

	Category	Frequency	%
Age (*M*±*SD*)		20.53±2.15	—
Gender	Male	99	26.40%
	Female	276	73.60%
Years of Chinese learning	≤ 1 year	150	40.00%
	1-3 years	156	41.60%
	≥ 3 years	69	18.40%
Chinese language proficiency	Beginner level (level 1-2)	4	1.07%
	Intermediate level (level 3-4)	145	38.67%
	Advanced level (level 5-6)	119	31.73%
	Never participated	107	28.53%

Chinese language proficiency was referenced with the participants' passing level in Hanyu Shuiping Kaoshi (HSK, see Peng et al., 2020 for a detailed description of HSK).

3.2 Instruments

Two questionnaires were developed based on existing valid instruments: (1) the first questionnaire comprised 16 items that measured five variables of individual differences to reflect learners' EVT-based academic learning motivation and WTC. Items for self-efficacy (SE) were adapted from Shaaban & Ghaith (2000) to measure learners' general self-efficacy for Chinese speaking and self-efficacy for Chinese academic learning, while items for attainment value, utility value, and intrinsic value were adapted from Gaspard et al. (2017) to measure learners' overall task values in Chinese learning. Items concerning WTC were adapted from Lee and Lee (2020), which specifically focused on inside classroom situations; and (2) the second questionnaire consisted of 14 items that measured four variables in TAM to reflect learners' ChatGPT acceptance in oral language practices. Items for perceived ease of use and perceived usefulness were adapted from Davis (1989), while items for attitude toward using and behavioral intention to use were adapted from Venkatesh et al. (2003). The adaptation on the two instruments was mainly about phrasing the items within the CFL learning and ChatGPT-assisted oral language practice contexts. All the items were measured on a 5-point Likert scale (1: Strongly disagree to 5: Strongly agree).

3.3 Data analysis

PLS-SEM was implemented to assess the measurement model and the hypothesized structural model in SmartPLS 4.0. The reasons to use PLS-SEM rather than covariance-based structural equation modeling (CB-SEM) were: (1) little research have used CB-SEM to investigate FL learners' EVT-based motivation, willingness to communicate, and ChatGPT acceptance in oral language practices, therefore, the high level of statistical power of PLS-SEM would benefit the theory developing from a predictive standpoint (Hair et al., 2019); (2) PLS-SEM is not subject to data distribution restrictions, while CB-SEM can produce abnormal results with non-normal data (Hair et al., 2019); and (3) PLS-SEM offers better solutions with a small sample size (Hair et al., 2019; 2022). The inverse square root method proposed by Kock and Hadaya (2018) was used to determine whether our sample size was sufficient for PLS-SEM analysis. Given the anticipated effect size of 0.20 and the desired probability of 0.05, 155 samples would be required to detect the effect. Thus, 375 samples in this study met the minimum sample size requirements for PLS-SEM. Prediction-oriented segmentation (POS), a method for detecting unobserved heterogeneity that was specifically developed to fit PLS path modeling, was further employed to test whether the examined research model significantly differed among research participants. Compared to the traditional approach to segmentation in SEM by assigning samples to predefined segments on the basis of demographic variables, POS is especially beneficial in identifying potential heterogeneity in a case where there is a lack of ground rationale for distinguishing subgroups within a population, allowing for more efficient capturing of heterogeneity while avoiding under- or over-segmenting (Hair et al., 2016; Rigdon et al., 2010). Given that there has been little previous research on the disparities in the influencing relationships between learning motivation, WTC, and technology acceptance across different FL learner populations, we were thus conducting POS in an attempt to detect any unobserved heterogeneity from a predictive perspective. The demographic backgrounds of learners in different groups were also compared based on the POS results.

4. Results

4.1 The measurement model

The reliability and convergent validity were assessed through factor loadings, Cronbach' s alpha, composite reliability (CR), and average extracted variance (AVE). The factor loadings of each indicator ranged from 0.78 to 0.94 (Table 3). Both the Cronbach's alpha and CR (rho_c) for each latent variable were higher than the recommended value of 0.70, and the AVE for each latent variable exceeded the minimum requirement of 0.50, which corroborates the reliability and convergent validity of the measurement model.

Table 3 Reliability and convergent validity of the measurement model

Variable	Item	Factor loading	Mean	SD	α	CR	AVE
Self-efficacy (SE)	SE1	0.789	3.25	1.16	0.826	0.882	0.652
	SE2	0.784	3.29	1.12			
	SE3	0.848	3.54	1.04			
	SE4	0.807	3.84	1.07			
Utility value (UV)	UV1	0.901	4.24	0.98	0.859	0.914	0.780
	UV2	0.898	4.16	1.03			
	UV3	0.849	4.30	0.97			
Attainment value	AV1	0.934	4.30	0.97			

(AV)	AV2	0.931	4.21	1.00	0.920	0.949	0.862
	AV3	0.920	4.25	1.05			
Intrinsic value (IV)	IV1	0.941	3.90	1.11	0.924	0.951	0.867
	IV2	0.930	3.93	1.10			
	IV3	0.922	3.91	1.02			
Willingness to communicate (WTC)	WTC1	0.885	3.96	1.07	0.850	0.909	0.769
	WTC2	0.911	3.94	1.06			
	WTC3	0.834	3.78	1.09			
Perceived ease of use (PEOU)	PEOU1	0.904	3.84	0.96	0.917	0.941	0.800
	PEOU2	0.907	3.85	0.95			
	PEOU3	0.889	3.75	0.97			
	PEOU4	0.878	3.76	0.97			
Perceived usefulness (PU)	PU1	0.882	3.79	0.89	0.901	0.931	0.771
	PU2	0.889	3.85	0.83			
	PU3	0.872	3.81	0.90			
	PU4	0.870	3.79	0.97			
Attitude toward using (ATU)	ATU1	0.894	3.94	0.96	0.850	0.909	0.770
	ATU2	0.863	3.96	0.95			
	ATU3	0.874	3.94	0.95			
Behavioral intention to use (BIU)	BIU1	0.899	3.63	1.10	0.891	0.933	0.822
	BIU2	0.931	3.55	1.18			
	BIU3	0.889	3.65	1.10			

Discriminant validity was analyzed by using the Fornell-Larcker criterion (Fornell & Larcker, 1981) and the Heterotrait-Monotrait method (Henseler et al., 2015). The square root of AVE for each variable exceeded the correlations among latent variables (Table 4), which fulfilled the Fornell-Larcker criterion. The HTMT ratios of all latent variables were less than the criteria of 0.85 in Hair et al. (2019) (Table 5), demonstrating adequate discriminant validity of the measurement model. Furthermore, the variance inflation factor (VIF) values of all indicators ranged from 1.68 to 4.24, which is less than the suggested cut-off value of 5.0 in Hair et al. (2022), indicating high collinearity was not an issue in this study.

Table 4 Discriminant validity based on Fornell-Larcker criterion

	SE	UV	AV	IV	WTC	PEOU	PU	ATU	BIU
SE	0.807								
UV	0.163	0.883							
AV	0.069	0.480	0.928						
IV	0.003	0.396	0.564	0.931					
WTC	0.291	0.389	0.385	0.318	0.877				
PEOU	0.175	0.155	0.341	0.214	0.340	0.895			
PU	0.255	0.350	0.457	0.335	0.374	0.529	0.878		
ATU	0.205	0.143	0.331	0.263	0.330	0.506	0.515	0.877	
BIU	0.095	0.269	0.440	0.421	0.299	0.385	0.450	0.448	0.906

Table 5 Discriminant validity based on Heterotrait-Monotrait method

	SE	UV	AV	IV	WTC	PEOU	PU	ATU
UV	0.185							
AV	0.090	0.546						
IV	0.062	0.452	0.616					
WTC	0.333	0.451	0.430	0.351				
PEOU	0.204	0.173	0.369	0.231	0.382			
PU	0.292	0.401	0.501	0.364	0.428	0.577		
ATU	0.244	0.171	0.376	0.294	0.384	0.573	0.585	
BIU	0.112	0.311	0.486	0.461	0.340	0.426	0.500	0.515

4.2 The structural model

As the reliability and validity of the measurement model have been established, we examined the structural model to evaluate model quality and test the proposed hypotheses. The R^2 values of endogenous variables and the Stone-Geisser test (Q^2) were applied to ensure the predictive relevance of the model. According to Hair and Alamer (2022), R^2 values between 0 to 0.10, 0.11 to 0.30, 0.30 to 50, and > 0.50 indicate weak, modest, moderate, and strong explanatory power in L2 research. The R^2 values of endogenous variables in the structural model ranged between 0.12 to 0.34, indicating modest to moderate explanatory power. Q^2 values should be greater than zero for a particular endogenous variable to indicate predictive accuracy (Hair et al., 2022). The Q^2 values of all endogenous variables in the structural model were above zero, ranging from 0.09 to 0.26, thus establishing the predictive accuracy of the model.

The structural relationships between the latent variables are presented in Table 6. Eight hypotheses (H1-6, H10, and H11) were supported at a significant level of $p < .001$, and two hypotheses (H7 and H8) were supported at a significant level of $p < .01$. H9 was supported at a significant level of $p < .05$ but had a path coefficient below 0.02, and thus should be eliminated from the nested model. Indirect effects of the four motivational determinants on TAM variables were also examined under maximum likelihood estimation with 5,000 bootstrap samples (Table 7). Except for IV, all indirect paths from SE, UV, and AV to the four TAM variables reached significance, revealing the critical mediating role of WTC in the structural model.

Table 6. Path coefficients of the proposed research model

	Path	*β*	T statistics	f^2	*p*	Results
H1	PEOU -> PU	0.455	8.527	0.270	$< .001^{***}$	Supported
H2	PEOU -> ATU	0.325	5.196	0.115	$< .001^{***}$	Supported
H3	PU -> ATU	0.343	5.591	0.128	$< .001^{***}$	Supported
H4	PU -> BIU	0.298	5.172	0.089	$< .001^{***}$	Supported
H5	ATU -> BIU	0.295	5.623	0.087	$< .001^{***}$	Supported
H6	SE -> WTC	0.243	5.474	0.078	$< .001^{***}$	Supported
H7	UV -> WTC	0.204	3.370	0.040	$.001^{**}$	Supported
H8	AV -> WTC	0.200	2.830	0.033	$.004^{**}$	Supported
H9	IV -> WTC	0.123	2.024	0.014	$.043^{*}$	Supported
H10	WTC -> PEOU	0.340	6.211	0.131	$< .001^{***}$	Supported
H11	WTC -> PU	0.220	4.422	0.063	$< .001^{***}$	Supported

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 7 Indirect effects of learning motivation

	β	T statistics	Bias-corrected 95% CI		*p*
			Lower	Upper	
SE -> PEOU	0.083	3.833	0.046	0.128	$< .001^{***}$
SE -> PU	0.091	4.058	0.050	0.138	$< .001^{***}$
SE -> ATU	0.058	3.759	0.031	0.089	$< .001^{***}$
SE -> BIU	0.044	3.513	0.023	0.071	$< .001^{***}$
UV -> PEOU	0.069	3.174	0.032	0.118	$.002^{**}$
UV -> PU	0.076	3.238	0.033	0.125	$.001^{**}$
UV -> ATU	0.049	3.096	0.021	0.083	$.002^{**}$
UV -> BIU	0.037	2.992	0.015	0.064	$.003^{**}$
AV -> PEOU	0.068	2.377	0.019	0.132	$.018^{*}$
AV -> PU	0.075	2.452	0.021	0.142	$.015^{*}$
AV -> ATU	0.048	2.344	0.013	0.093	$.020^{*}$
AV -> BIU	0.036	2.249	0.010	0.075	$.025^{*}$
IV -> PEOU	0.042	1.895	0.002	0.089	.058
IV -> PU	0.046	1.933	0.002	0.096	.053
IV -> ATU	0.029	1.897	0.001	0.062	.058
IV -> BIU	0.022	1.861	0.001	0.048	.063

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

4.3 Prediction-oriented segmentation (POS)

PLS-POS was performed in an attempt to find any unobserved heterogeneity among the samples. The sum of all constructs weighted R^2 was chosen as the optimization criterion. Considering the above-mentioned minimum sample requirement, we opted for a 2-segment solution with 1000 iterations and a search depth of 375 to perform PLS-POS. The demographic information of learners in the two segments is displayed in Table 8, and the SEM results in segment 1 ($N = 200$) and segment 2 ($N = 175$) are presented in Figure 4 and Figure 5, respectively.

Table 8 Sample demographics in the two segments

	Category	Frequency			
		Segment 1 (*N* = 200)	% within group	Segment 2 (*N* = 175)	% within group
Age (*M*±*SD*)		20.50±2.07		20.56±2.25	
Gender	Male	57	28.5%	42	24.0%
	Female	143	71.5%	133	76.0%
Years of Chinese learning	≤ 1 year	86	43.0%	64	36.6%
	1-3 years	88	44.0%	68	38.9%
	≥ 3 years	26	13.0%	43	24.6%
Chinese language proficiency	Beginner level (level 1-2)	3	1.5%	1	0.6%
	Intermediate level (level 3-4)	77	38.5%	68	38.9%
	Advanced level (level 5-6)	65	32.5%	54	30.9%
	Never participated	55	27.5%	52	29.7%

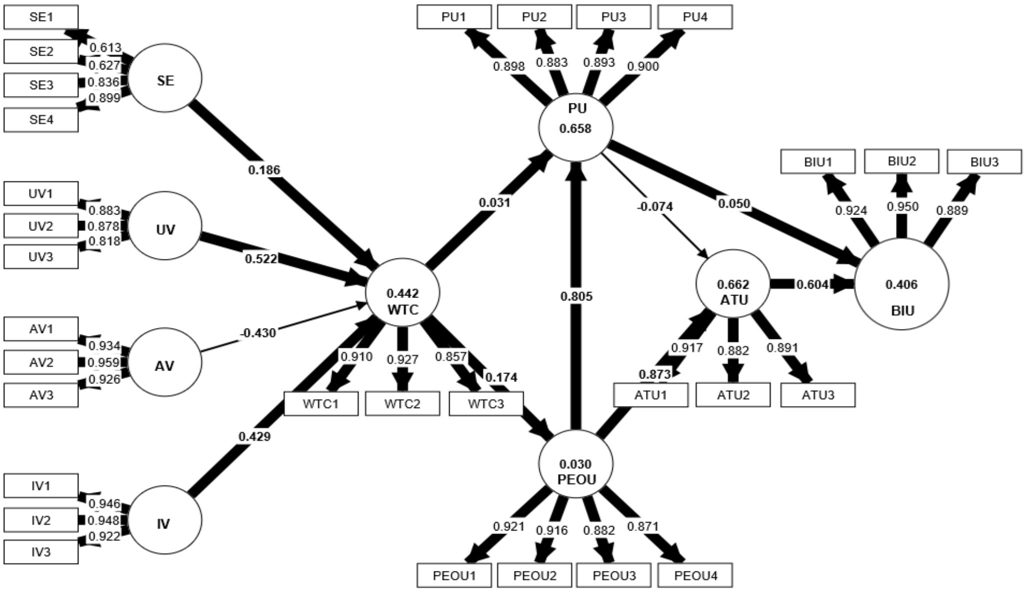


Figure 4 The structural model in segment 1

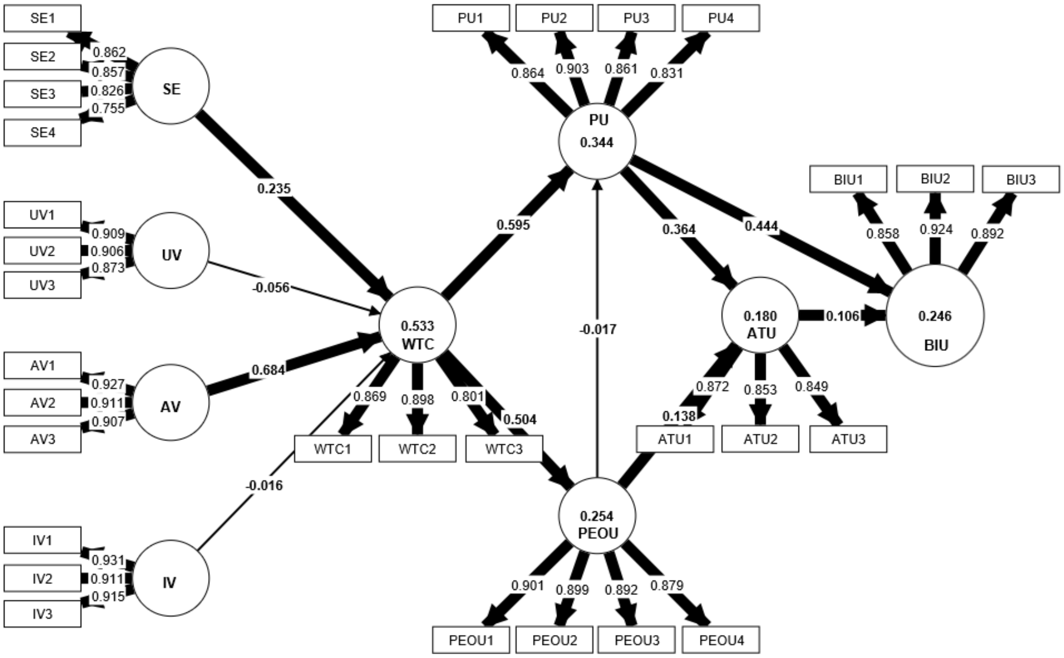


Figure 5 The structural model in segment 2

A multi-group analysis (MGA) was performed to test if there are any significant differences in the path coefficients across the two segmental models. The results indicated that all other paths in the model showed significant differences in their path coefficients, with the exception of the influence of SE on WTC (Table 9). Mann-Whitney U tests were further carried out to evaluate whether learners' age, gender, years of Chinese learning, and language proficiency differed across the two segments. The results indicated that the years of Chinese learning among learners in segment 2 was significantly longer than those among segment 1 (Z = 2.27, *p* = .023), but no statistically significant difference was found in other demographic background variables between the two segments.

Table 9 Differences in the path coefficients across the two segments

Path	Segment 1	Segment 2	Path coefficient difference	*p*
PEOU -> PU	0.805	-0.017	0.822	< .001***
PEOU -> ATU	0.873	0.138	0.735	< .001***
PU -> ATU	-0.074	0.364	-0.438	< .001***
PU -> BIU	0.050	0.444	-0.393	< .001***
ATU -> BIU	0.604	0.106	0.498	< .001***
SE -> WTC	0.186	0.235	-0.049	.598
UV -> WTC	0.522	-0.056	0.578	< .001***
AV -> WTC	-0.430	0.684	-1.114	< .001***
IV -> WTC	0.429	-0.016	0.445	.001**
WTC -> PEOU	0.174	0.504	-0.330	.003**
WTC -> PU	0.031	0.595	-0.564	< .001***

* $p < .05$, ** $p < .01$, *** $p < .001$.

5. Discussion

5.1 The effectiveness of TAM in predicting ChatGPT acceptance

To explore the effectiveness of TAM in the context of accepting ChatGPT as a learning tool in FL oral language practices, we developed a research model in which involved four TAM variables. The five hypotheses (H1-H5) between TAM variables were supported by the PLS-SEM results in this study, which were consistent with the theoretical assumptions in Davis et al. (1989). Specifically, FL learners' PEOU has a significant positive influence on PU, and both of the two positively affect BIU through ATU. The previous investigation on FL learners' ChatGPT acceptance based on TAM in Liu and Ma (2024) dissolved the positive influence of PEOU on ATU. However, in this study, both PEOU and PU were found to be significant predictors of ATU, forming a more holistic picture of TAM's theoretical strengths in predicting FL learners' ChatGPT acceptance. In addition to ATU, PU could also directly affect BIU, even having a stronger influence than ATU, which empirically supports the assumption in Davis et al. (1989) that 'people form intentions toward using computer systems based largely on a cognitive appraisal of how it will improve their performance' (p. 986).

5.2 The role of learning motivation and willingness to communicate

To explore the role of learning motivation and WTC in FL learners' ChatGPT acceptance, six hypotheses (H6-H11) were proposed in the research model. SE was found to be a significant predictor of WTC, which supports the theoretical assumption in MacIntyre et al.'s (1998) pyramid model of L2 WTC that learners' positive belief about their language ability is a critical antecedent of their willingness to communicate. Concurred with previous results in MacIntyre and Blackie (2012), AV and UV were also found as significant predictors of WTC. However, the positive effect of IV on WTC, though supported in PLS-SEM, failed to reach a sufficient effect size and thus cannot be accepted in this study. The eliminated influence of IV on WTC might result from the co-existence of other affective or emotional factors (e.g., L2 anxiety, shyness) as restraining forces for language communication (Pavelescu, 2023), which has especially been commonly reported among East Asian language learners under the influence of their cultural system and educational practices (for a detailed review, see Shao & Gao, 2016).

Furthermore, WTC significantly mediated the influences of SE, UV, and AV on the four TAM variables. In other words, FL learners with higher learning motivation are more willing to communicate in the target language and thus inclined to accept ChatGPT as a learning tool in oral language practices. This echoed Eccles-Parsons et al.'s (1983) argument with regard to learning motivation as a critical psychological antecedent of learners' subsequent academic task choices and achievement-related decision making. The findings further concretize the above argument in the context of ChatGPT-assisted oral language practices and highlight the significance of willingness to communicate in such academic decision-making process.

5.3 The unobserved heterogeneity among CFL learners

Due to the complexity of social and behavioral phenomena, heterogeneity in the samples is likely to exist (Becker et al., 2013). PLS-POS results in this study demonstrated that the formation pattern of ChatGPT acceptance is characterized by heterogeneity among CFL learners rooted in their years of Chinese learning. Two main findings can be drawn from the examined heterogeneity of ChatGPT acceptance between learners with different CFL learning experiences:

First, WTC has a greater impact on ChatGPT acceptance among learners with longer Chinese learning experiences compared to their counterparts, since the effects of the two paths from WTC to PEOU and to PU were significantly stronger in segment 2 than in segment 1. The reason for this might be that learners who had been learning Chinese for a longer time had more opportunities to speak the language while attributing their prior academic success to language communication (Wen & Piao, 2020), therefore attaching a greater value on language communication in FL learning and being more open to interacting with ChatGPT. Second, for learners with longer Chinese learning experiences, their BIU benefited more from PU, as the effect of PU on BIU was significantly higher in segment 2 than in segment 1. In contrast, for those learners with shorter Chinese learning experiences, their BIU was more influenced by PEOU through ATU, as the effects of the relevant two paths were significantly higher in segment 1 than in segment 2. Compared to CFL beginners, learners with longer learning experience may have already experimented with different educational technologies, acquired more effective technology-assisted learning techniques, and thus been able to interact with technologies more efficiently (Durndell & Haag, 2002; Luo, 2020). As a result, those long-term CFL learners may place more emphasis on ChatGPT's effectiveness for oral language practice than its efficiency. This finding also reveals that ChatGPT might play different roles among CFL learners. Given that the PU of ChatGPT is more important in forming acceptance for learners with longer Chinese learning experience, they may regard ChatGPT as a tutor or instructor with whom they expect to learn extra language knowledge; on the contrary, those beginners might consider ChatGPT simply as a convenient language partner to interact with because the PEOU of ChatGPT is more crucial for developing their behavioral intentions.

6. Implications, and limitations, conclusions

This study sought to predict CFL learners' acceptance of ChatGPT in oral language practices with learning motivation and willingness to communicate, as well as explore any potential heterogeneity of ChatGPT acceptance among CFL learners. The results of this study provide evidence on the effectiveness of TAM in investigating ChatGPT acceptance in the context of CFL oral language practices. TAM has recently been employed and validated in AI-assisted language learning, with a focus on automated writing evaluation (e.g., Li et al., 2019), intelligent tutoring systems (e.g., Ni & Cheung, 2023), and AI-powered chatbots (e.g., Belda-Medina & Calvo-Ferrer, 2022; Chen et al., 2020; Liu & Ma, 2024). This study further contributes to the TAM literature by concentrating on ChatGPT-assisted oral language practices. The technological tool examined in this study, ChatGPT, could be utilized for different learning purposes among foreign language learners, such as providing feedback for essays, generating assessment tasks, performing language translation, and recommending specific learning materials (Lo, 2023). The supported effectiveness of TAM in this study implies future IS research to contextualize measurements within specific learning purposes in order to accurately evaluate learners' technology acceptance, particularly when the targeted technology offers a variety of technical affordances.

Results also highlight the antecedental role of learning motivation and willingness to communicate in ChatGPT acceptance, which offer valuable practical insights from a pedagogical perspective. The findings demonstrated that situational analysis regarding learners' psychological attributes is necessary before delivering formal instructions with the aid of educational technologies in FL classrooms. With a thorough understanding of CFL learners' motivation towards academic learning and how potential sociocultural factors exert impacts in the learning contexts, teachers could utilize effective motivational strategies and learning activities as incentives to promote learners' acceptance of educational technologies (i.e., designing topics that learners are familiar with, incorporating cultural elements, and providing clear language structure for the scaffolding purpose), further leading to active engagement in technology-assisted language learning and producing meaningful educational outcomes. Moreover, while willingness to communicate has been extensively explored in traditional language learning contexts, it has received insufficient attention in technology-assisted language learning contexts. In our study, willingness to communicate was found to be a significant mediator between learning motivation and ChatGPT acceptance, which suggests future research focus more on learners' willingness to communicate and its impacts on technology adoption and usage, especially when the learning contexts require oral-based interaction in the target language. From a pedagogical standpoint, this finding also highlights the critical role of foreign language teachers in encouraging East Asian learners' willingness to communicate with effective pedagogical strategies and sufficient talking opportunities before implementing technology-assisted language practices.

The formation of ChatGPT acceptance appeared heterogeneity among CFL learners. This result offers possible explanations for why certain theoretically supported relationships between TAM variables had been dissolved in prior relevant investigations. To enhance ChatGPT acceptance among CFL learners, suitable instructional strategies should be carefully chosen when designing ChatGPT-assisted language learning activities, while different features of ChatGPT should be purposefully promoted throughout the process with consideration of learners' past learning experiences. When facing long-term CFL learners, more emphasis should be placed on linking ChatGPT-assisted oral language practices to their previous knowledge constructions, demonstrating the great potential of ChatGPT in enhancing their speaking performance. Whereas for those CFL beginners, teachers may start by providing more guidance on learner-technology interaction techniques that could be applied in AI-assisted language learning, supporting learners in generating operable and favorable interacting experiences, and thus developing positive attitudes towards technologies in oral language practices.

There are certain limitations in this study. First, due to the survey nature of our work and time constraints, the interaction time allotted to each learner in the speaking activities prior to the survey was relatively limited. To reach a more comprehensive understanding of learners' adoption of and interaction with chatbots, interventional or observational studies are thus advised for future research to reveal the interaction patterns and strategies utilized by learners with various levels of learning motivation and WTC. Second, our findings were solely based on self-reported survey data. Future research is expected to incorporate data from classroom observations or interviews to provide additional triangulation reference. Additionally, it is valuable to identify other individual-level, task-level, teacher-level, and organization-level influencing factors that may impact learners' acceptance of GenAI-powered chatbots with qualitative data. Last, the sample size in this study was somewhat small and limited to Mongolian CFL learners. Survey studies with larger sample sizes or include other CFL learner populations are thus recommended to enhance the generalizability of our findings and to improve the statistical power of the analysis on the intricate relationships between learning motivation, WTC, and TAM variables among different learner populations.

References

Bai, B., Nie, Y., & Lee, A. N. (2020). Academic self-efficacy, task importance and interest: Relations with English language learning in an Asian context. *Journal of Multilingual and Multicultural Development*, *43*(5), 438–451.

Balouchi, S., & Samad, A. A. (2021). The effect of perceived competence on second language communication frequency: The mediating roles of motivation, willingness to communicate, and international posture. *Education and Information Technologies*, *26*(5), 5917–5937.

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. Freeman.

Becker, J. M., Rai, A., Ringle, C. M., & Völckner, F. (2013). Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS Quarterly*, *37*(3), 665–694.

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as AI conversational partners in language learning. *Applied Sciences*, *12*(17), 8427.

Chen, H. L., Vicki Widarso, G., & Sutrisno, H. (2020). A chatbot for learning Chinese: Learning achievement and technology acceptance. *Journal of Educational Computing Research*, *58*(6), 1161–1189.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

Durndell, A., & Haag, Z. (2002). Computer self efficacy, computer anxiety, attitudes towards the internet and reported experience with the internet, by gender, in an East European sample. *Computers in Human Behavior*, *18*(5), 521–535.

Eccles-Parsons, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J., &

Midgley, C. (1983). Expectancies, values and academic behaviors. In Spence, J. T. (Ed.), *Achievement and Achievement Motives: Psychological and Sociological Approaches* (pp. 75–146). Freeman.

Elahi Shirvan, M., Khajavy, G. H., MacIntyre, P. D., & Taherian, T. (2019). A meta-analysis of L2 willingness to communicate and its three high-evidence correlates. *Journal of Psycholinguistic Research*, *48*(6), 1241–1267.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Gardner, R. C. (1985). *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*. Edward Arnold Publishers.

Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology*, *48*, 67–84.

Hair, J. F., & Alamer, A. (2022). Partial least squares structural equation modeling (PLS-SEM) in second language and education research: Guidelines using an applied example. *Research Methods in Applied Linguistics*, *1*(3), 100027.

Hair, J. F., Sarstedt, M., Matthews, L. M., & Ringle, C. M. (2016). Identifying and treating unobserved heterogeneity with FIMIX-PLS: part I – method. *European Business Review*, *28*(1), 63–76.

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, *31*(1), 2–24.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)* (3rd ed.). SAGE Publications.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135.

Hsu, L. (2017). EFL learners' acceptance of technology in a computer-assisted language learning (CALL) context: The role of intrinsic-extrinsic motivation in English learning. *International Journal of Information and Education Technology*, *7*(9), 679–685.

Hsu, M. H., Chen, P. S., & Yu, C. S. (2023). Proposing a task-oriented chatbot system for EFL learners speaking practice. *Interactive Learning Environments*, *31*(7), 4297–4308.

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.

Jeon, J. (2022). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, *31*(7), 4297–4308.

Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal*, *28*(1), 227–261.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550.

Lee, J. S. (2020). The role of grit and classroom enjoyment in EFL learners' willingness to communicate. *Journal of Multilingual and Multicultural Development*, *43*(5), 452–468.

Lee, J. S., & Drajati, N. A. (2019). Affective variables and informal digital learning of English: Keys to willingness to communicate in a second language. *Australasian Journal of Educational Technology*, *35*(5), 168–182.

Lee, J. S., & Hsieh, J. C. (2019). Affective variables and willingness to communicate of EFL learners in in-class, out-of-class, and digital contexts. *System*, *82*, 63–73.

Lee, J. S., & Lee, K. (2020). Affective factors, virtual intercultural experiences, and L2 willingness to communicate in in-class, out-of-class, and digital settings. *Language Teaching Research*, *24*(6), 813–833.

Lee, J. S., & Lu, Y. (2023). L2 motivational self system and willingness to communicate in the classroom and extramural digital contexts. *Computer Assisted Language Learning*, *36*(1–2), 126–148.

Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., & Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer Assisted Language Learning, 32*(7)*,* 784–804.

Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*, *18*(2), 125–138.

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410.

Loh, E. K. Y. (2019). What we know about expectancy-value theory, and how it helps to design a sustained motivating learning environment. *System*, *86*, 102119.

Luo, B. (2020). The influence of teaching learning techniques on students' long-term learning behavior. *Computer Assisted Language Learning*, *33*(4), 388–412.

MacIntyre, P. D., & Blackie, R. A. (2012). Action control, motivated strategies, and integrative motivation as predictors of language learning affect and the intention to continue learning French. *System*, *40*(4), 533–543.

MacIntyre, P. D., Dörnyei, Z., Clément, R., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*(4), 545–562.

Mystkowska-Wiertelak, A. (2021). The link between different facets of willingness to communicate, engagement and communicative behaviour in task performance. In Budzińska, K., & Majchrzak, O. (Eds.), *Positive Psychology in Second and Foreign Language Education* (pp. 95–113). Springer.

Nagle, C. (2021). Using expectancy value theory to understand motivation, persistence, and achievement in university-level foreign language learning. *Foreign Language Annals*, *54*(4), 1238–1256.

Ni, A., & Cheung, A. (2023). Understanding secondary students' continuance intention to adopt AI-powered intelligent tutoring system for English learning. *Education and Information Technologies*, *28*(3), 3191–3216.

Pavelescu, L. M. (2023). Emotion, motivation and willingness to communicate in the language learning experience: A comparative case study of two adult ESOL learners. *Language Teaching Research*. Advance online publication. http://doi.org/10.1177/13621688221146884

Peng, Y., Yan, W., & Cheng, L. (2020). Hanyu Shuiping Kaoshi (HSK): A multi-level, multi-purpose proficiency test. *Language Testing*, *38*(2), 326–337.

Petrović, J., Jovanović, M. (2021). The role of chatbots in foreign language learning: The present situation and the future outlook. In Pap, E. (Eds.), *Artificial Intelligence: Theory and Applications* (pp. 313–330). Springer.

Rigdon, E. E., Ringle, C. M., & Sarstedt, M. (2010). Structural modeling of heterogeneous data with partial least squares. In N. K. Malhotra (Ed.), *Review of Marketing Research* (volume 7, pp. 255–296). Sharpe.

Rosenzweig, E. Q., Wigfield, A., Eccles, J. S., Renninger, K. A., & Hidi, S. E. (2019). Expectancy-value theory and its relevance for student motivation and learning. In Renninger, K. A., & Hidi, S. E. (Eds.), *The Cambridge Handbook of Motivation and Learning* (pp. 617–644). Cambridge University Press.

Ruan, S., Jiang, L., Xu, Q., Liu, Z., Davis, G. M., Brunskill, E., & Landay, J. A. (2021). EnglishBot: An AI-powered conversational system for second language learning. *Proceedings of 26th International Conference on Intelligent User Interfaces*, 434–444. Association for Computing Machinery.

Shaaban, K. A., & Ghaith, G. (2000). Student motivation to learn English as a foreign language. *Foreign Language Annals*, *33*(6), 632–644.

Shao, Q., & Gao, X. (2016). Reticence and willingness to communicate (WTC) of East Asian language learners. *System*, *63*, 115–120.

Soyoof, A. (2023). Iranian EFL students' perception of willingness to communicate in an extramural digital context. *Interactive Learning Environments*, *31*(9), 5922–5939.

Srite, M. (2006). Culture as an explanation of technology acceptance differences: An empirical investigation of Chinese and US users. *Australasian Journal of Information Systems*, *14*(1), 5–25.

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2023.2209881

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*(1), 15.

Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, *11*(4), 342–365.

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478.

Wang, Q., & Xue, M. (2022). The implications of expectancy-value theory of motivation in language education. *Frontiers in Psychology*, *13*, 992372.

Wen, X., & Piao, M. (2020). Motivational profiles and learning experience across Chinese language proficiency levels. *System*, *90*, 102216.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81.

Xia, Q., Chiu, T. K., Chai, C. S., & Xie, K. (2023). The mediating effects of needs satisfaction on the relationships between prior knowledge and self-regulated learning through artificial intelligence chatbot. *British Journal of Educational Technology*, *54*(4), 967–986.

Yang, H., & Lian, Z. (2023). Ideal L2 self, self-efficacy, and pragmatic production: The mediating role of willingness to communicate in learning English as a foreign language. *Behavioral Sciences*, *13*(7), 597.

Zadorozhnyy, A., & Lee, J. S. (2023). Informal digital learning of English and willingness to communicate in a second language: Self-efficacy beliefs as a mediator. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2023.2215279

Appendix

Self-efficacy (SE)
I can speak Chinese fairly fluently.
I can communicate with Chinese speakers in Chinese.
I can receive a good grade from my Chinese course.
I can master the knowledge in my Chinese course.

Utility value (UV)
Being good at Chinese will bring me many benefits in my future daily life.
The things I learn in the Chinese course will be applicable in my future life.
In general, learning Chinese is practical for my future plans.

Attainment value (AV)
It is important to me to be good at Chinese.
Being good at Chinese means a lot to me personally.
In general, learning Chinese well is important to me.

Intrinsic value (IV)
I like learning Chinese.
I am fascinated by Chinese.
In general, learning Chinese is interesting to me.

Willingness to communicate (WTC)
I am willing to communicate in Chinese when I have a chance to talk freely in Chinese classes.
I am willing to communicate in Chinese when I have a chance to talk in front of other students in Chinese classes.
I am willing to communicate in Chinese when I have a group discussion in Chinese classes.
Perceived ease of use (PEOU)

It is easy to learn how to use ChatGPT to practice oral Chinese.

It is easy to become proficient in practicing oral Chinese using ChatGPT.
It is easy to orally interact with ChatGPT.
The interaction with ChatGPT is clear and understandable.

Perceived usefulness (PU)
Using ChatGPT could improve my oral Chinese learning performance.
Using ChatGPT could enhance my oral Chinese learning effectiveness.
Using ChatGPT could increase my Chinese language output in oral practices.
Using ChatGPT could help me complete oral Chinese practice tasks more quickly.
Attitude toward using (ATU)

I believe that using ChatGPT is a good idea.
I believe that using ChatGPT is advisable.
I agree with the practice of using ChatGPT for oral Chinese practices.

Behavioral intention to use (BIU)
I intend to use ChatGPT in oral Chinese practices in the future.
I intend to use ChatGPT regularly to practice my oral Chinese in the future.
I intend to use ChatGPT to practice on more topics in oral Chinese practices in the future.

Volume 15 Number 1, June 2024
pp. 49-69

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 B.多 C.大 D.重	A.强
	MASK2	A. 但是 B. 那么 C. 从而 D. 可见	D.可见
	MASK3	A. 吸收 B. 吸取 C. 吸引 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, we plan to fine-tune the LLMs to investigate their performance on current tasks.

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

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：{} 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}。 {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。 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} 1.重新仔细审题并重复题目 2.重点查看关键词所在的句子 3.重点查看句子对应的编号 4.阅读并重复你刚才的解释 5.根据{n}步的结果，检查你前面的解释中，是否存在错误 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。} 1.重新仔细审题并重复题目 2.重点查看关键词所在的句子 3.重点查看句子对应的编号 4.阅读并重复你刚才的解释 5.根据{1-4}步的结果，检查你前面的解释中，是否存在错误 6.告诉我你的错误并改正

4. Findings

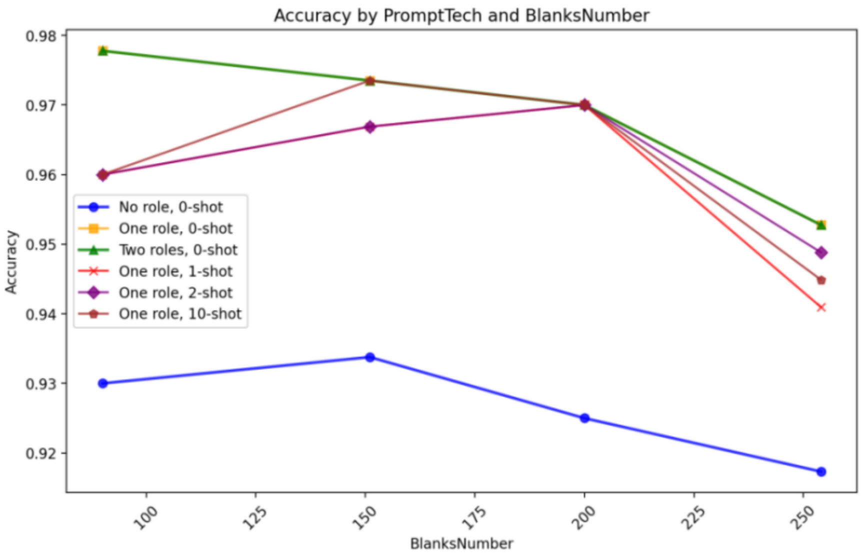
4.1 Experiment 1

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

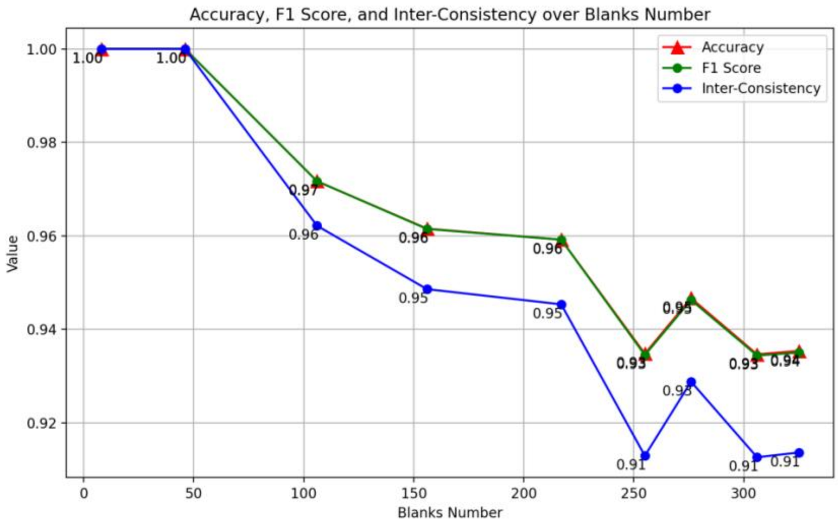
Metrics	GPT3.5 Turbo	ERNIE4.0	Baichuan2-13B
Accuracy	0.612	1	0.980
F1 Score	0.607	1	0.980
Consistency	0.484	1	0.973

* 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).

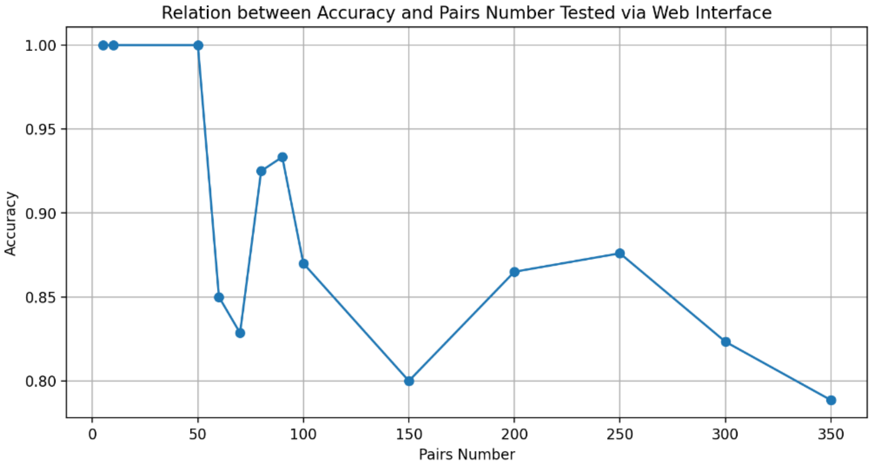


Figure 3 Variation of Accuracy with Pairs Number via Web Interface

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.

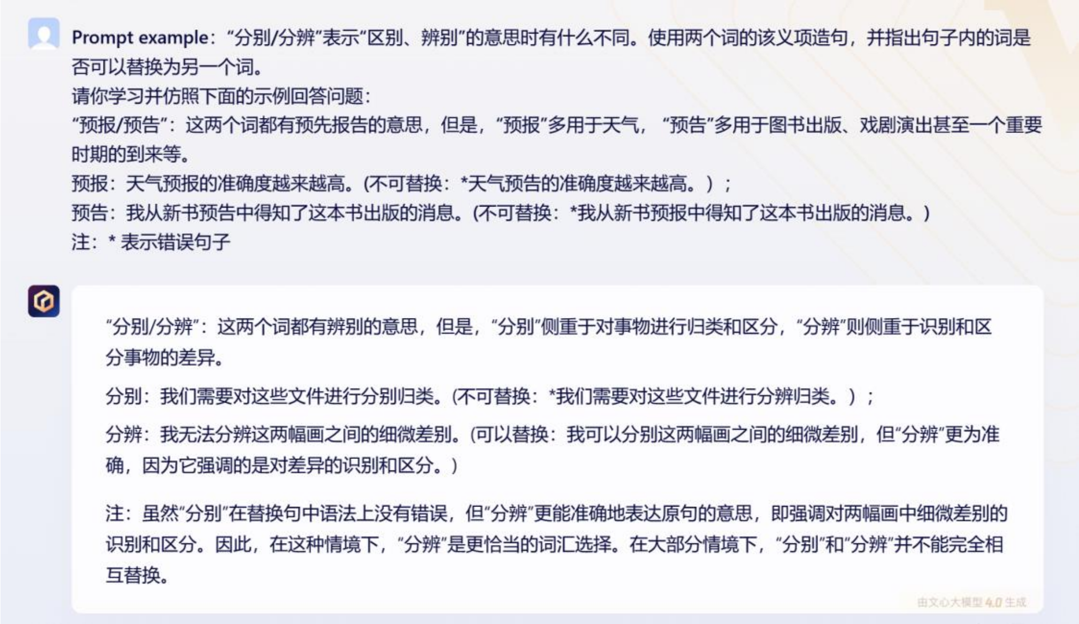
4.3 Experiment 3

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.



Prompt example：“分别/分辨”表示“区别、辨别”的意思时有什么不同。使用两个词的该义项造句，并指出句子内的词是否可以替换为另一个词。
请你学习并仿照下面的示例回答问题：
“预报/预告”：这两个词都有预先报告的意思，但是，“预报”多用于天气，“预告”多用于图书出版、戏剧演出甚至一个重要时期的到来等。
预报：天气预报的准确度越来越高。(不可替换：*天气预告的准确度越来越高。）；
预告：我从新书预告中得知了这本书出版的消息。(不可替换：*我从新书预报中得知了这本书出版的消息。）
注：* 表示错误句子

“分别/分辨”：这两个词都有辨别的意思，但是，“分别”侧重于对事物进行归类和区分，“分辨”则侧重于识别和区分事物的差异。

分别：我们需要对这些文件进行分别归类。(不可替换：*我们需要对这些文件进行分辨归类。）；

分辨：我无法分辨这两幅画之间的细微差别。(可以替换：我可以分别这两幅画之间的细微差别，但“分辨”更为准确，因为它强调的是对差异的识别和区分。）

注：虽然“分别”在替换句中语法上没有错误，但“分辨”更能准确地表达原句的意思，即强调对两幅画中细微差别的识别和区分。因此，在这种情境下，“分辨”是更恰当的词汇选择。在大部分情境下，“分别”和“分辨”并不能完全相互替换。

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.

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 B.多 C.大 D.重	A.强
	MASK2	A.但是 B.那么 C.从而 D.可见	D.可见
	MASK3	A.吸收 B.吸取 C.吸引 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：{} 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}。 {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。 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} 1.重新仔细审题并重复题目 2.重点查看关键词所在的句子 3.重点查看句子对应的编号 4.阅读并重复你刚才的解释 5.根据{n}步的结果，检查你前面的解释中，是否存在错误 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。} 1.重新仔细审题并重复题目 2.重点查看关键词所在的句子 3.重点查看句子对应的编号 4.阅读并重复你刚才的解释 5.根据{1-4}步的结果，检查你前面的解释中，是否存在错误 6.告诉我你的错误并改正

4. Findings

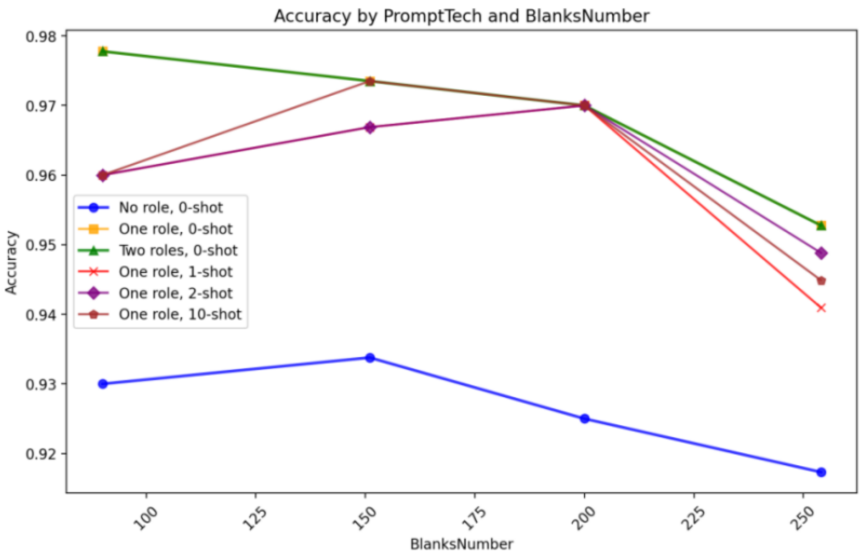
4.1 Experiment 1

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

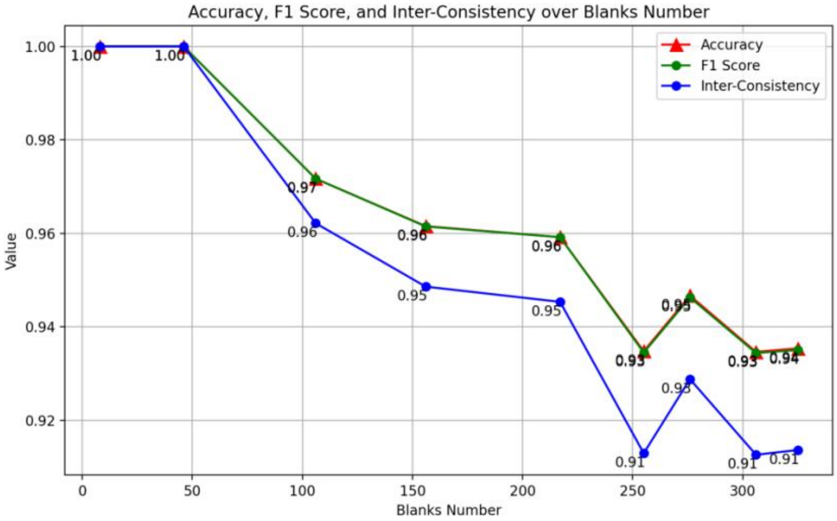
Metrics	GPT3.5 Turbo	ERNIE4.0	Baichuan2-13B
Accuracy	0.612	1	0.980
F1 Score	0.607	1	0.980
Consistency	0.484	1	0.973

* 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).

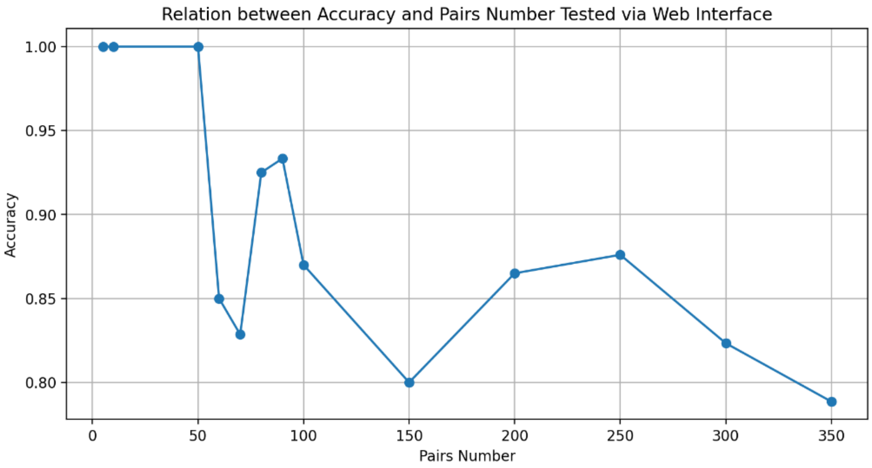


Figure 3 Variation of Accuracy with Pairs Number via Web Interface

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.

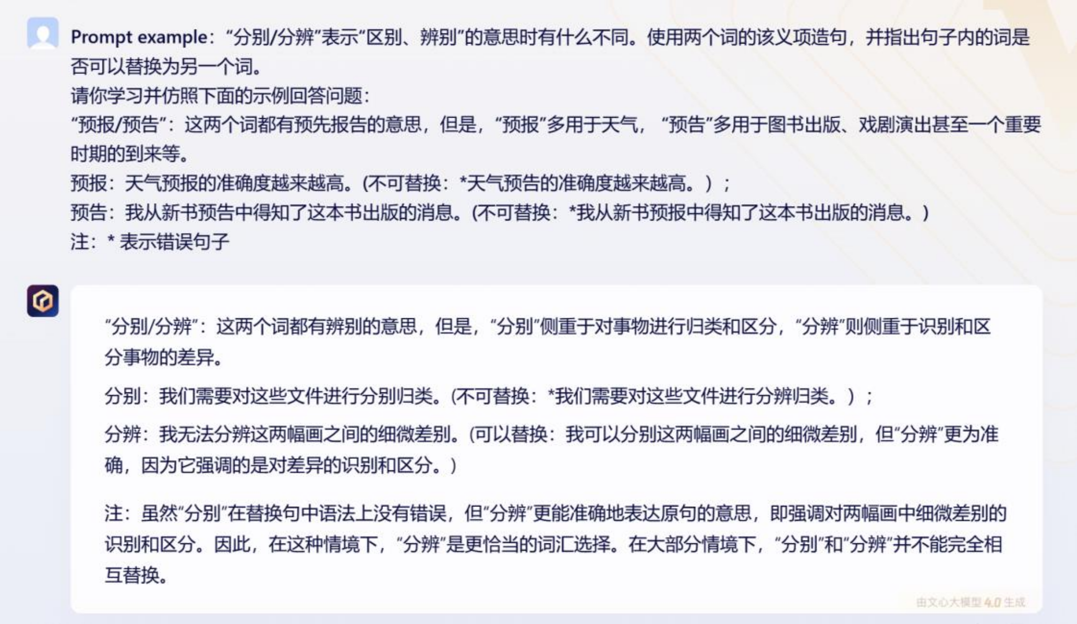
4.3 Experiment 3

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.



Prompt example：“分别/分辨”表示“区别、辨别”的意思时有什么不同。使用两个词的该义项造句，并指出句子内的词是否可以替换为另一个词。
请你学习并仿照下面的示例回答问题：
“预报/预告”：这两个词都有预先报告的意思，但是，“预报”多用于天气， “预告”多用于图书出版、戏剧演出甚至一个重要时期的到来等。
预报：天气预报的准确度越来越高。(不可替换：*天气预告的准确度越来越高。)；
预告：我从新书预告中得知了这本书出版的消息。(不可替换：*我从新书预报中得知了这本书出版的消息。)
注：* 表示错误句子

“分别/分辨”：这两个词都有辨别的意思，但是，“分别”侧重于对事物进行归类和区分，“分辨”则侧重于识别和区分事物的差异。

分别：我们需要对这些文件进行分别归类。(不可替换：*我们需要对这些文件进行分辨归类。)；

分辨：我无法分辨这两幅画之间的细微差别。(可以替换：我可以分别这两幅画之间的细微差别，但“分辨”更为准确，因为它强调的是对差异的识别和区分。)

注：虽然“分别”在替换句中语法上没有错误，但“分辨”更能准确地表达原句的意思，即强调对两幅画中细微差别的识别和区分。因此，在这种情境下，“分辨”是更恰当的词汇选择。在大部分情境下，“分别”和“分辨”并不能完全相互替换。

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.

基于 AI 技术的 CVC 中文视听语料库设计与应用
(The Design and Application of a Chinese Audio-Visual Corpus Based on AI Technology)

Wang, Tao
(王涛)
Beijing International Studies University
(北京第二外国语学院)
toddy@bisu.edu.cn

摘要：AI 赋能下的语言教学生态环境带来了新型人机协作关系，学习者获取信息与知识的方式变得更加场景化、智能化，教师需要利用多种资源媒介和 AI 技术手段引导学生形成主动探究问题、整合知识的能力。CVC 中文视听语料库通过采集来自汉语母语者生活中使用的视频语言材料，将教材内容、本体知识和视频语料数据相关联，为用户提供场景化的教学资源应用服务。基于 AI 技术的视听语料库应用可提供以视频内容检索为核心的资源型、智慧化语言教学手段。在服务语言教学同时，视频语料标注结果将有助于自然语言处理（ NLP ）和计算机视觉（ CV ）交叉领域的语言模型训练，在言语行为识别、多模态分析、情感分析等方面满足人工智能对多模态大数据的需求，反哺人工智能领域的未来发展进程。

Abstract: Cultivating an AI-powered language teaching ecosystem has introduced a new model of human-computer collaboration. Many learners are acquiring information and knowledge in a manner that is more contextualized and intelligent. Many educators are adaptive to explore a variety of media resources and AI technologies to guide learners in developing capabilities in problems resolving and knowledge integration. This article describes the design and application of a Chinese Audio-Visual Corpus (CVC) that collects visual language materials from native Chinese speakers in their daily lives to associate textbook content, ontological knowledge, and video materials data in order to provide learners with context-aware teaching resource applications. This AI-based audio-visual corpus offers resourceful and intelligent language teaching methods with the priority of video content retrieval. In addition to serving language teaching, the annotated results from this audio-visual corpus will aid in the training of language models in the interdisciplinary field of Natural Language Processing (NLP) and Computer Vision (CV). It meets the demand for multimodal big data in artificial intelligence for applications such as multimodal discourse analysis, speech act recognition, and sentiment analysis, thereby contributing to the future development of the field of artificial intelligence.

关键词：视听语料库；视频检索；人工智能；多模态；国际中文教育

Keyword: Audio-Visual Corpus, Video Retrieval, Artificial Intelligence, Multimodal, International Chinese Language Education

1. 引言

当前，AI 赋能下的语言教学生态环境带来了新型人机协作关系，学习者获取信息与知识的方式变得更加场景化、智能化。教师的角色也随之转变，从传统的知识提供者转化为引导者，利用多种资源媒介和 AI 技术手段引导学生形成主动探究问题、整合知识的能力。传统语言教学多采用文本材料，教师讲解语言知识规则的同时缺乏真实语言实例的使用，教学中难免存在一定的局限性，不利于第二语言习得效果。视频语料在呈现情景语境和社会语境的同时，可以还原真实交际过程中的语言要素和非语言要素，包含语音、文字、情境、表情、肢体动作、交际身份、文化背景等不同符号系统。Ginsburgh（1935）、Hendrix（1939）、Palomo（1940）、Kern（1959）和 Fallahkhair 等（2004）在相关研究中均提到真实视频在语言教学中的优势，指出有声电影、电视节目可以在有机语境中呈现目标语言，并推荐在外语教学中开展应用。冯惟钢（1995）、沈履伟（1995）、唐荔（1997）、王飙（2009）、张璐（2011）、刘立新、和邓方（2018）等国内学者也先后进行了影视资源在对外汉语教学中的应用研究，并就视听说教材的选材依据、编制理念进行了深入探讨。

CVC 中文视听语料库（www.chinafoucs.net.cn）（以下简称 CVC 语料库）采集汉语母语者生活中使用的真实语言材料，提供以视频节目内容检索为核心的资源型、智慧化语言教学工具。教师可针对教材词汇、语法等级大纲选取具有典型语境的视频语料展示语言功能实例，通过情景语境和文化语境促进学习者认知发展过程。通过结合情境设计练习活动，突出教学重点、解决教学难点、提高教学效率。将师生从单纯使用教材转向学材资源的开发、利用，变为教学活动的设计者和参与者，实现在用中学、以用促学、学用合一的目的。本文对 CVC 语料库的设计理念和使用方法进行详细说明。

2. 设计理念

近年来，随着现代外语教学理念和新文科建设的发展，语言学和语言教学也更为关注跨学科应用和话语研究转向，研究对象从平面媒体扩展到新媒体，从单一模态纸质出版物到多模态数字文化出版领域。研究方法从侧重静态语言形式结构描写到结合动态功能解释，从单一学科到多学科的交叉融合也是必然趋势。王涛（2012）提出视频语料库建设的必要性，并对视频语料采集、标注、检索实现过程进行了说明。2017 年，在北京第二外国语学院举办的“第一届汉语视听说教学理论与应用研讨会”上，王涛发表题为“多模态视角下的视听说教材立体化建设及教学创新”的主旨

报告，就视听说教材编写理念、教学模式创新进行了总体阐述。在“第二届汉语视听说教学理论与应用研讨会暨新媒体数字环境下的汉语教学创新研究学术会”上，王涛（ 2018 ）发表题为“视听说课程大纲设计与教学实践研究”报告，指出视听说课程和教学系统建立在视频语料库基础之上，教学内容全部来自真实语料，是汉语母语者生活中使用的语言。论文从系统功能语言学视角对视听语篇类型进行了阐述，进一步明确了课程开发与教学系统中视频语料库的作用（ 见图 1 ）。

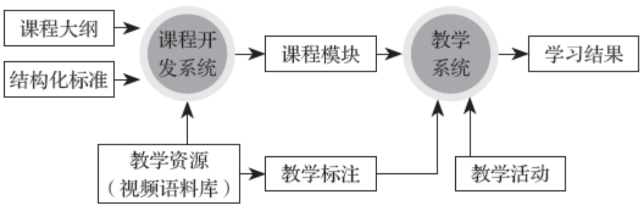


图 1 视听说课程开发与教学系统动力学模型

中文视听语料库研究项目正式被国际中文教学领域关注是在美国初中级中文教学兴趣组（ CLTA-SIG ）举办的线上讲座，王涛（ 2022 ）发表题为“中文视听语料库应用”的专题报告，详细介绍了视听语料库的核心理念及检索功能。讲座由耶鲁大学梁宁辉主持，相关资料整理发布在耶鲁大学初、中级中文教学交流网站[1]。在美国 TCLT《科技与中文教学》[2]期刊举办的科技教学系列讲座中，王涛（ 2024 ）发表题为“视听材料选取、教材编写及相关 AI 技术工具介绍”报告，介绍了 CVC 语料库的最新进展及教学应用成果。讲座由宾州印第安纳大学刘士娟主持，多位中外一线教师就语料库应用展开了深入探讨。

2.1 语料库设计

语料库语言学是国际中文教育领域的基础学科，通过对大规模口语或书面语真实语料统计分析，挖掘语言事实在意义和表达形式上的内在规律，为语言教学提供应用平台和实证性研究支持。根据建设目标、用途、语料来源、采集加工路线等不同方面，语料库存在多种类型。由于音视频介质语料采集、加工、存储成本较高，国内现有大型语料库多为文本形式，数据来源于古汉语、文学作品、新闻、报刊、社交媒体等书面语材料。詹卫东,郭锐等（ 2019 ）、荀恩东,饶高琦等（ 2016 ）研究结果显示，北京语言大学 BCC 语料库，口语语料主要来自新浪微博和影视字幕，占比为 6.3%。北京大学 CCL 语料库，现代文献中文学语料占比高达 92.15%，口语语料占比仅为 0.26%。普遍存在语料库采样不平衡、媒介形式单一、多模态语料库建设相对滞后等问题。另外，传统文本语料库查询系统采取的是一种基于字频、词频概率统计的科研量化手段，工具性地分析文本聚合关系并不能反映语言交际使用过程的全貌，无法满足智慧教育背景下语言学和语言教学研究对真实语料的需求。

CVC 语料库是一套面向国际中文教育领域的大规模音视频数据库，由北京第二

[1] 美国耶鲁大学初、中级中文教学交流网站 https://campuspress.yale.edu/exchange/

[2] 美国《科技与中文教学》期刊网址 http://www.tclt.us

外国语学院汉语学院王涛设计，北京视听说科技有限公司和自然语义（青岛）科技有限公司联合开发。该语料库依据《国际中文教育水平等级标准》（GF0025-2021）和《汉语视听说课程大纲的研发与应用案例》研究成果，结合词汇等级、平均语速、百字生词率等参数，通过优化算法实现语料自动标注及检索功能。系统整体架构由采集层、数据层、系统中台、用户管理层和应用接口层五部分组成：

1) 采集层由语料采集和多数据源语料对齐两大模块组成，包括视频对象文件存储、文本存储、同步管理模块。
2) 数据层主要由语料加工数据库和结构化元数据库两大模块组成，包括文本对象存储、文本纠错、原数据模块、分词切分、词性标注以及视频类型数据、语言形式数据、语篇类型数据、语体类型数据等元数据项组成。
3) 系统中台主要由算法池和应用程序两大模块组成，包括分词算法、词性标注算法、文本纠错算法、语言量化算法、语料维度、词汇图谱、多语种翻译以及接口管理等模型。自然语言处理（NLP）算法采用 HanLP 框架，是全球 NLP 开源领域（Github）用户量最大最受欢迎的 NLP 框架，具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。
4) 用户管理层包括用户基本信息管理、角色管理、个性化管理、应用管理、多维度检索功能、扩展管理功能等模块。
5) 应用接口层负责提供标准 API 接口，与其他系统进行集成交互。

2.2 语料采集

视频语料采集来自合作版权机构、公共媒体和网络视频节目，通过采集层字幕提取模块及语音识别技术转换为可针对语言内容检索的数据文本。用户输入关键词可以得到相应视频语料检索结果。视听语料库具有更强的交际互动性和功能阐释性，在语境、语用研究、多模态功能话语分析、互动语言学领域更符合语言学及语言教学需求。

2.3 语料分布

根据王涛（2018）视听说课程大纲研究，CVC 语料库分为电影、电视剧、情景剧、纪录片、综艺、访谈、辩论、朗诵、演讲、讲座、新闻、歌曲 12 类，将视频语料与语篇类型、语体程度以及语言表达形式关联，以便提供分类精准检索服务。截至 2024 年 6 月，语料分类及规模统计如下（见表 1）：

表 1 CVC 视听语料库规模分布统计

视频类型		字节数	百分比
电影		990651	6.96%
电视剧		9969426	70%
综艺		7080	0.05%
访谈		39189	0.28%
纪录片		2942924	20.66%

辩论		0	0.00%
朗诵		6339	0.04%
新闻		17511	0.12%
讲座		20727	0.15%
情景剧		55497	0.39%
演讲		166257	1.17%
歌曲		26626	0.19%
总计		**14242227**	**100%**

2.4 搜索引擎

CVC 语料库搜索引擎面向全球个人用户免费开放，支持通用检索、上下文全文检索、组合条件筛选检索三种模式，用户可通过以下两种方式登录：

1) 电脑 web 浏览器方式，输入网址 https://client.chinafocus.net.cn，使用微信或 WeChat 扫码登录（见图 2）。

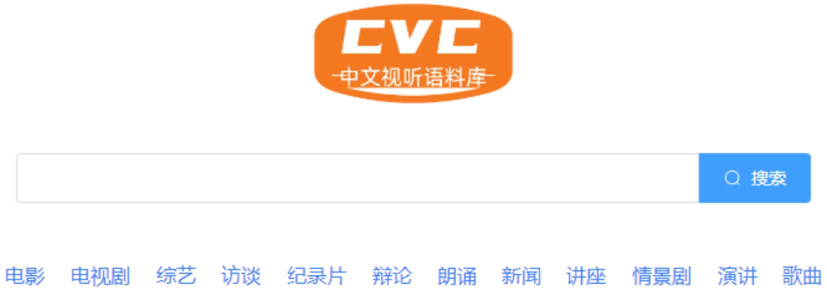
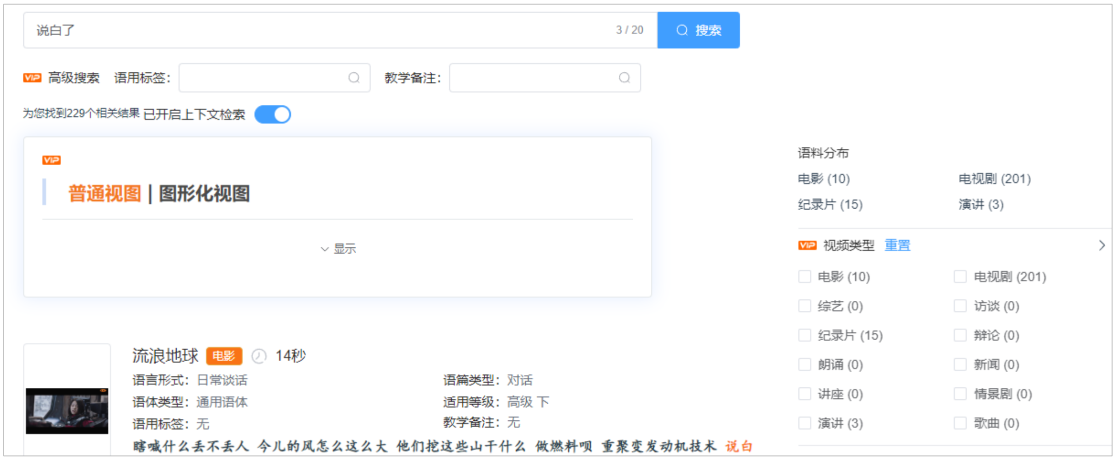


图 2 CVC 语料库首页

2) 手机移动端使用微信搜索“中文视听”公众号，点击“语料检索”登录。

CVC 语料库首页采用通用检索模式，支持词汇和常见构式检索，可识别超过 35 万条核心词库词汇。如搜索结果显示“抱歉没有找到相关的视频语料资源”，需点击开启“上下文检索”模式，可输入短语、关键词组合进行字符串全文匹配方式检索（见图 3）。



说白了 3 / 20 搜索

高级搜索 语用标签： 教学备注：

为您找到229个相关结果 已开启上下文检索

普通视图 | 图形化视图

显示

流浪地球 电影 14秒

语言形式：日常谈话　　语篇类型：对话

语体类型：通用语体　　适用等级：高级 下

语用标签：无　　教学备注：无

瞎喊什么丢不丢人 今儿的风怎么这么大 他们挖这些山干什么 做燃料呗 重聚变发动机技术 说白

语料分布

电影 (10)　　电视剧 (201)

纪录片 (15)　　演讲 (3)

视频类型 重置

□ 电影 (10)　　□ 电视剧 (201)

□ 综艺 (0)　　□ 访谈 (0)

□ 纪录片 (15)　　□ 辩论 (0)

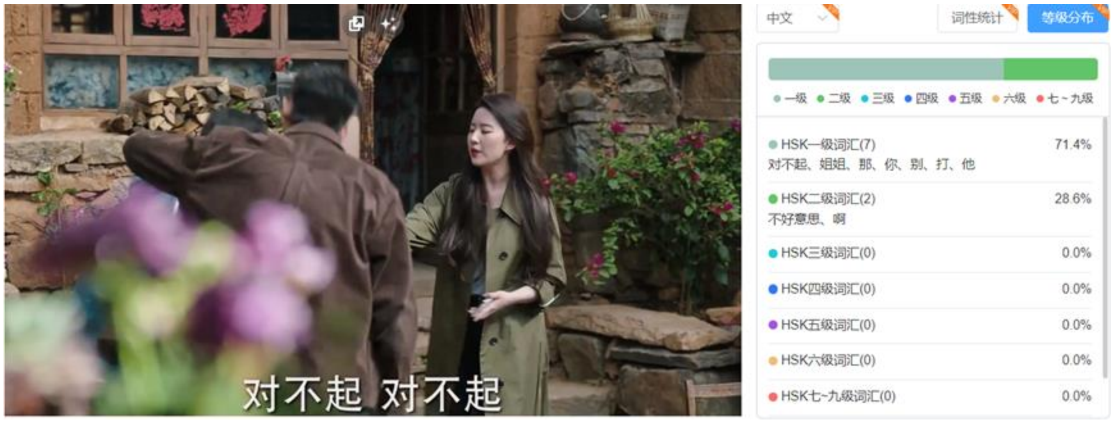
□ 朗诵 (0)　　□ 新闻 (0)

□ 讲座 (0)　　□ 情景剧 (0)

□ 演讲 (3)　　□ 歌曲 (0)

图 3 CVC 语料库上下文全文检索模式

高级用户模式仅向《中国微镜头》教材用户和合作机构院校教师开放，如有教学、科研用途需求可通过邮件发送至 mail@chinafocus.net.cn 申请 VIP 权限。VIP 授权用户可使用云计算软件服务，实现语料分布结果、视频类型、语言形式、语篇类型、语体类型、适用等级等自定义组合条件检索，并提供拼音标注、多语种机器翻译、词性标注、等级分布、语义图谱、语用标签、教材信息标注、视频课件编辑等 AI 语言辅助教学功能（ 见图 4 ）



● HSK一级词汇(7) 71.4%

对不起、姐姐、那、你、别、打、他

● HSK二级词汇(2) 28.6%

不好意思、啊

● HSK三级词汇(0) 0.0%

● HSK四级词汇(0) 0.0%

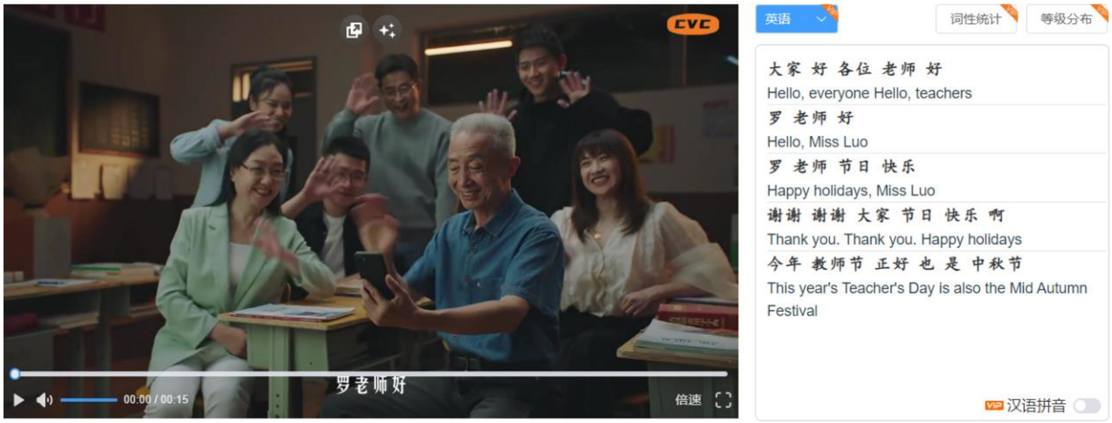
● HSK五级词汇(0) 0.0%

● HSK六级词汇(0) 0.0%

● HSK七~九级词汇(0) 0.0%

图 4 CVC 语料库 VIP 模式

CVC语料库采用基于神经网络的机器翻译模型，目前VIP模式支持语种包括英语、日语、韩语、德语、法语、俄语、西班牙语、阿拉伯语字幕翻译，可提供葡语、泰语、越南语、缅甸语等小语种机器翻译定制化服务（ 见图5 ）。

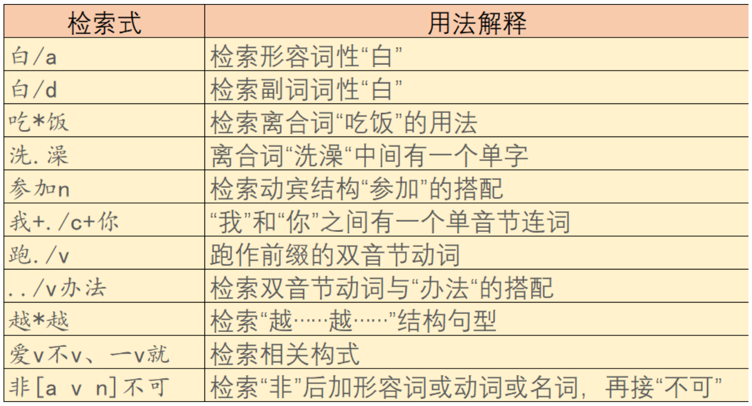


大家 好 各位 老师 好
Hello, everyone Hello, teachers
罗 老师 好
Hello, Miss Luo
罗 老师 节日 快乐
Happy holidays, Miss Luo
谢谢 谢谢 大家 节日 快乐 啊
Thank you. Thank you. Happy holidays
今年 教师节 正好 也 是 中秋节
This year's Teacher's Day is also the Mid Autumn Festival

图 5 CVC 语料库多语言机器翻译

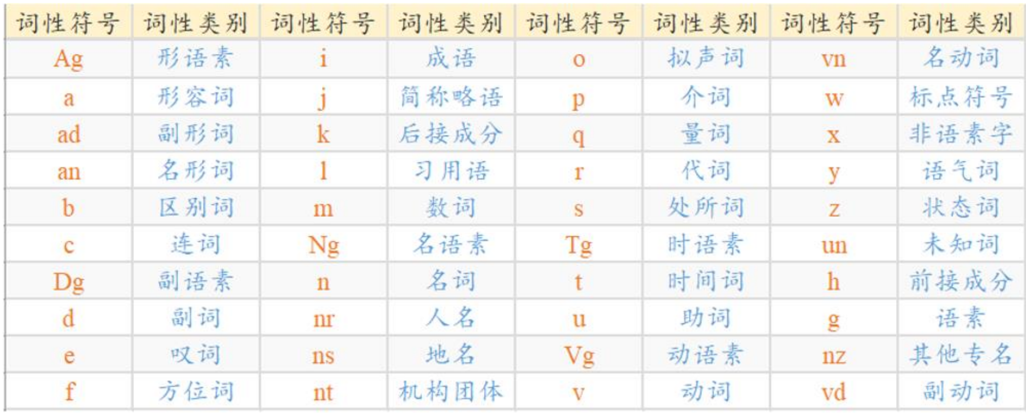
2.5 检索式

CVC 语料库提供中文语料库通用规则检索式，查询符合条件的语料结果。支持关键词、通配符、词性符号、空格或“+”搜索及常见语法构式检索。分词算法、词性标注算法和文本纠错算法采用 HanLP 框架算法模型，线上模型训练数据来自 9970 万字的大型综合语料库，覆盖新闻、社交媒体、金融、法律等多个领域。检索式规则说明及词性符号对照表如图 6、图 7 所示。



检索式	用法解释
白/a	检索形容词性“白”
白/d	检索副词词性“白”
吃*饭	检索离合词“吃饭”的用法
洗.澡	离合词“洗澡“中间有一个单字
参加n	检索动宾结构“参加”的搭配
我+./c+你	“我”和“你”之间有一个单音节连词
跑./v	跑作前缀的双音节动词
../v办法	检索双音节动词与“办法“的搭配
越*越	检索“越……越……”结构句型
爱v不v、一v就	检索相关构式
非[a v n]不可	检索“非”后加形容词或动词或名词，再接“不可”

图 6 CVC 语料库检索式规则说明



词性符号	词性类别	词性符号	词性类别	词性符号	词性类别	词性符号	词性类别
Ag	形语素	i	成语	o	拟声词	vn	名动词
a	形容词	j	简称略语	p	介词	w	标点符号
ad	副形词	k	后接成分	q	量词	x	非语素字
an	名形词	l	习用语	r	代词	y	语气词
b	区别词	m	数词	s	处所词	z	状态词
c	连词	Ng	名语素	Tg	时语素	un	未知词
Dg	副语素	n	名词	t	时间词	h	前接成分
d	副词	nr	人名	u	助词	g	语素
e	叹词	ns	地名	Vg	动语素	nz	其他专名
f	方位词	nt	机构团体	v	动词	vd	副动词

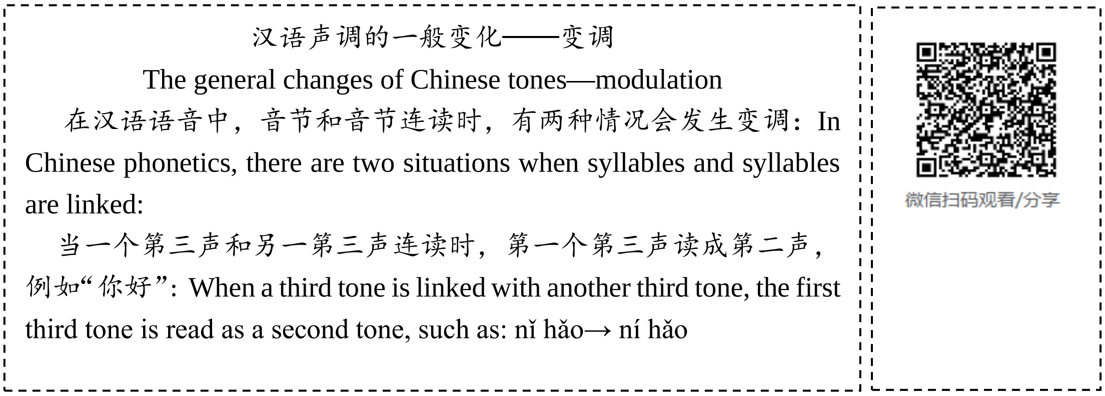
图 7 CVC 语料库词性符号对照表

3. 语料库应用

视频语料为学习者呈现了虚拟自然目的语的教学环境，有利于构建师生双方共享认知环境。从而在有效降低认知负荷前提下实现可理解性输入，为学习者创造有意义的输出机会。虞莉（ 2020 ）认为体演文化教学法通过身临其境的“体演”活动以及周密的课程设计，将语言教学与文化教学紧密结合，使语言学习者不仅能掌握语言技能，而且能获取跨文化交际能力，从而有效并得体地与母语者交流。CVC 语料库不仅可以应用于语音、词汇、语法、文化教学，还可以和体演法、任务型教学法相结合，丰富教学手段和教学设计，让课堂教学延伸到课前、课后环节，弥补常规课堂语言教学模式和平面教材的不足，进一步将“结构—功能—文化”为核心的教学理念场景化、话题化、实例化。教学应用示例如下：

3.1 语音教学

语音是中文教学的基础，传统听力技能教学内容普遍采用人工录制的教学语言，与母语者现实生活中使用的自然语言存在一定差异。CVC 语料库更为关注词汇、语法在真实口语交际活动中的表现形式，涉及音系-句法接口研究。即意义是如何通过音系层表达的，包括声调、音节组合、语调、节奏、轻重、停连等韵律结构特征。例如上声在自然语流中的调值多是半三或变调，学习者如果仅仅把注意力放在课文标注的声调符号上，容易忽视对语音本身的辨识。通过视频语料，学习者可以在真实情景中模仿正常语流中的发音，完成语音自然习得过程。用户在阅读教材注释时，可使用微信扫码观看视频语料，第一次扫码为登录语料库，以后扫码可直接观看（ 见图 8 ）。



汉语声调的一般变化——变调

The general changes of Chinese tones—modulation

在汉语语音中，音节和音节连读时，有两种情况会发生变调：In Chinese phonetics, there are two situations when syllables and syllables are linked:

当一个第三声和另一第三声连读时，第一个第三声读成第二声，例如“你好”：When a third tone is linked with another third tone, the first third tone is read as a second tone, such as: nǐ hǎo→ ní hǎo

微信扫码观看/分享

图 8. 语音教学示例

3.2 词汇教学

兼类词、虚词、副词是词汇教学的难点，CVC 语料库包含了大量真实语料数据，可以帮助学习者看到词汇在真实语境中的使用实例，增强学习者的自然表达能力和应用创造能力。值得注意的是，由于兼类词通常具有两种或两种以上不同语法特征，在不同语境中承担不同句法成分和词类属性，检索时需要标注词性符号以便提供精

准结果。例如根据《国际中文教育水平等级标准》，形容词词性“白”是 HSK1 级词汇，可通过“白/a”进行检索。副词“白”是 HSK3 级词汇，则需要输入“白/d”检索。另外，现有教学辞书注释中通常仅关注语义结构方面的描写，并没有给出主观情态功能的解释。例如语气副词“明明”，除了表示客观实事显然如此，还常用于表达自责或抱怨、指责别人的语气（ 见图 9 ）。

语气副词“明明”

《现代汉语词典》第 7 版中解释“表示显然如此或确实（下文意思往往转折）”

《现代汉语八百词》“表示显然这样。用‘明明’的小句前或后常有反问或表转折的小句。

微信扫码观看/分享

图 9 词汇教学示例

3.3 语法教学

演绎法是常见的语法教学方法，教师通常先展示语法形式结构，之后再进行讲解、操练。传统教学方法固然有助于形成规范的教学思路和教学模式，但也往往容易被教材和教学模式所约束，容易使讲练停留在机械性操练层面。CVC 语料库搜索引擎支持常见语法构式检索，以“把字句”教学为例，教师在讲解语法规则的同时，往往还需要将句法结构与功能、语境结合进行句型操练。如果直接输入“把”字搜索，目前可得到 9492 条结果，其中还包含了量词、动词结果；输入“把/p”检索，可以得到 9104 条介词词性结果；输入“把 n+v”可以缩小到 2119 条结果；输入“把 n+v+在”仅为 81 条结果。同理，输入“所＋v+的”可以得到构成“所……的”字短语作主语、宾语的例句；输入“所 v*的+n”可以得到“所”用于动词短语前作定语的用法。

与演绎法不同的是，归纳法、情境法、任务型教学法重视语言在交际中的实际使用，教学操作过程中可借助视频语料提供一些具有真实情境的语言用例，丰富教学手段和教学设计。例如疑问代词“怎么”表达主观情态意义的非疑问用法（ 见图 10 ）

疑问代词“怎么”的非疑问用法

形式结构：

1. np/vp+怎么+v
2. S+怎么+vp/ap

释义：用于质问、责骂，带有生气的语气。例句：

1．你怎么开车的！

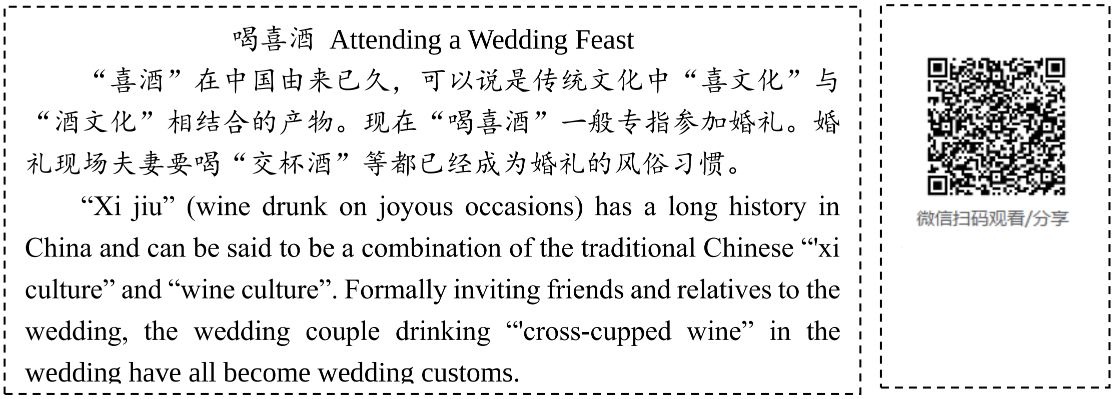
2．你怎么说话呢！

微信扫码观看/分享

图 10 语法教学示例

3.4 文化教学

CVC 视听语料库既包含饮食、茶、京剧、节日等中国传统元素，也涉及当代家庭、教育、婚姻、医疗、体育、贸易、一带一路、城市化等社会话题节目。视听材料可以通过融媒体视角聚焦中国社会发展进程，增强中华文化感召力和话语说服力，进一步完善中华优秀传统文化和中国式现代化进程的传播路径，推动中文国际话语体系多模态构建的故事化、形象化建设，发挥国际中文教育的“社会窗口”作用。例如《中国微镜头》视听说教材中级下《家庭篇》中介绍了现代年轻人的婚恋话题，课文的文化链接部分对“喝喜酒”和“交杯酒”进行了文化注释（见图 11）。



喝喜酒 Attending a Wedding Feast

“喜酒”在中国由来已久，可以说是传统文化中“喜文化”与“酒文化”相结合的产物。现在“喝喜酒”一般专指参加婚礼。婚礼现场夫妻要喝“交杯酒”等都已经成为婚礼的风俗习惯。

“Xi jiu” (wine drunk on joyous occasions) has a long history in China and can be said to be a combination of the traditional Chinese “'xi culture” and “wine culture”. Formally inviting friends and relatives to the wedding, the wedding couple drinking “'cross-cupped wine” in the wedding have all become wedding customs.

图 11 文化教学示例

4. 结语

智慧教育背景下的视听语料库研制，可以将教材内容、本体知识和视频语料相关联，为用户提供场景化的教学资源数据应用服务。探索人工智能技术与教育的融合创新，驱动教育理念、教学模式创新，培养学生自主学习能力，完成传统课堂教学向线上、线下混合模式的转变。从教学视角来看，视听语料库通过信息技术对教学资源进行重组，优化传统课堂教学流程，进一步拓展课前/课后学习环境，提高教学效率，提升教学质量，推动 ERP（Education Resource Planning）的实施；从学习者视角看，视听语料库通过数据驱动学习者主动构建认知过程，可以帮助学习者在真实情境中运用所学知识，发挥主动感知、思维、创新能力，促进深度学习方式的开展。

随着大数据和人工智能技术的不断发展，语料库建设及应用成果广泛运用于语言学、翻译、二语教学、融媒辞书编撰、教材开发等领域。CVC 语料库建设将有助于填补国际中文教育领域多模态平衡语料库研究空白。本文着重介绍了该语料库的设计理念和语料检索功能使用说明，后续将进一步结合具体案例进行教学应用经验总结分享。未来围绕视听语料库应用研究将从以下三个层面展开：宏观层面以语言学理论为基础，结合语用研究、互动语言学、多模态功能话语分析方向；中观层面可以结合任务型教学法、情境法、沉浸式教学法、体演文化教学法（PCA）、产出导向教学法（POA）、Backword Design、混合教学模式等；微观层面从视频语料在不

同课型中的使用入手，进行教学设计、立体化教材/学材开发、教学资源、教学评价等方面的应用研究。在服务语言教学同时，视频语料标注结果还将有助于自然语言处理和计算机视觉交叉领域的语言模型训练，在言语行为识别、多模态分析、情感分析等方面满足人工智能对多模态大数据的需求，反哺人工智能领域的发展进程。

致谢: 本项目获教育部中外语言交流合作中心 2023 年国际中文教学实践创新项目资助，《国际中文视听教材国际传播能力创新与实践研究》YHJXCX23-024。

参考文献

Feng, W. (1995). Audio-visual speaking teaching and the compilation of teaching materials. *Chinese Teaching in the World*, 4, 95-100. [冯惟钢. (1995). 视听说教学及其教材的编写. *世界汉语教学*, 4, 95-100.]

Liu, L., Deng, F. (2018). Compiling audio-visual-oral teaching materials with authentic data. *TCSOL Studies*, 3,31-37. [刘立新, 邓方. (2018). 基于“真实”材 料的视听说教材编制. *华文教学与研究*, 3, 31-37.]

Shen, L (1995). A brief discussion on the 'audio-visual speaking' teaching of Chinese as a foreign language. *Journal of Tianjin Normal University (Social Sciences)*, 1, 78-80. [沈履伟. (1995). 浅谈对外汉语的“视听说”教学. *天津师大学报(社会科学版)*, 1, 78-80.]

Tang, L. (1997). An initial exploration of Chinese 'audio-visual speaking' course teaching. *Journal of Open Learning*, 3, 20-23. [唐荔. (1997). 汉语“视听说”课程教学初探. *北京广播电视大学学报*, 3, 20-23.]

Wang, B. (2009). The audio-visual TCFL products published in Mainland China: A review and prospect. *Chinese Teaching in the World*, 2, 252-261. [王飙. (2009). 中国大陆对外汉语视听教材评述与展望. *世界汉语教学*, 2, 252-261.]

Wang, T. (2012). An initial exploration of the construction of a video corpus for teaching Chinese as a foreign language. *Series of International Research on Chinese Language(I)*, 1,175-181. [王涛. (2012). 对外汉语教学视频语料库建设初探.*国际汉语研究论丛（ 一 ）*, 1, 175-181.]

Wang, T. (2018). The development of a syllabus for the Chinese visual-audio-oral course and an application case. *Journal of International Chinese Teaching*, 4, 51-58. [王涛. (2018). 汉语视听说课程大纲的研发与应用案例. *国际汉语教学研究*, 4, 51-58.]

Xun, E., Rao, G., Xiao, X., & Zang, Ji. (2016). The construction of the BCC corpus in the age of big data. *Corpus Linguistics*, 1, 93-109. [荀恩东, 饶高琦, 肖晓悦, 臧娇娇. (2016). 大数据背景下 BCC 语料库的研制.语料库语言学, 1, 93-109.]

Yu, L. (2020). The performed culture approach: Intellectual history and core concepts. *Journal of International Chinese Teaching*, 2,42-49. [虞莉. (2020). 体演文化教学法: 渊源与核心. *国际汉语教学研究,* 2, 42-49.]

Zhan, W., Guo, R., Chang, B., Chen, Y., & Chen, L. (2019). The building of the CCL corpus: Its design and implementation. *Corpus Linguistics*, 1, 71-86. [詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙. (2019). 北京大学 CCL 语料库的研制. *语料库语言学*, 1, 71-86.]

Zhang, L. (2011). A Study on the selection of content and topics for audio-visual speaking materials in teaching Chinese as a foreign language. *Modern Chinese*, 1, 142-145. [张璐. (2011). 对外汉语视听说教材内容取材和话题选择研究. *现代语文*, 1,142-145.]

Use of Asynchronous Online Discussion in an Online Chinese Heritage Language Course
(异步在线讨论在一华裔线上中文课程中的应用)

Ji, Jingjing
(季晶晶)
Northwestern University
(西北大学)
jingjing.ji@northwestern.edu

Lin, Chuan
(林川)
University at Buffalo
(布法罗大学)
clin97@buffalo.edu

Abstract: With the increasing use of online teaching in schools, asynchronous online discussion (AOD) is becoming a common tool to facilitate interactions in online courses. However, very few studies explored using AOD in the context of Chinese language learning, including learning Chinese as a heritage language. To fill the gap, this article delineates the implementation and implications of AOD in an online Chinese heritage language course. A social learning platform named Yellowdig was adopted to conduct AOD, with two primary goals: community building and resource sharing. Students' reflections and feedback confirmed its social and educational benefits and indicated the promising utilization of AOD in other Chinese language courses of both in-person and online modes.

摘要：随着线上教学的增加，为促进网络课堂的互动，异步在线讨论的应用越来越普遍。然而，包括华裔中文教学在内的中文教学对于异步在线讨论的运用的研究依然较少。本文旨在讨论如何将异步讨论活动应用于一华裔中文线上中文课程。该异步讨论活动在名为Yellowdig 的教育社交平台上进行，以期达到两个目的：建立学习社群以及资源分享。学生的反馈肯定了该活动在社交、学习两大方面的益处，这也表明该活动可应用于其他线上或线下的中文课程。

Keywords: Asynchronous online discussion, social learning platforms, Chinese heritage language courses

关键字：异步在线讨论、教育社交平台、华裔中文课

1. Introduction

Asynchronous online discussion (AOD) may be merely an ancillary component in in-person courses. However, it is positioned as "a central hub" for online course activities (Dennen & Wieland, 2007). When it comes to designing an online Chinese language course specifically for Chinese heritage language learners (CHLLs), it should undoubtedly serve as one of the most critical components as well, given the learners' proficiency level in oral Chinese and the need to hone their reading and writing skills. Despite certain constraints that students may encounter, the benefits that students may reap from this type of online activity have been well documented in a plethora of studies. For instance, AOD could effectively reduce language learners' feeling of isolation and provide them opportunities to practice the language in a social environment (Comer & Lenaghan, 2013), which is much needed by language education in an online environment where meaningful face-to-face interaction could be limited.

Grounded in existing research findings, this article delineates how the AOD of an online Chinese heritage language (CHL) course was designed and implemented. A concrete example is used to present a more straightforward view. Student reflections are also discussed to provide further insight so that interested language instructors, administrators, or other stakeholders may make informed decisions regarding AOD in online teaching.

2. Literature review

Along with the growing popularity of online education, AOD has been widely adopted across many disciplines such as preservice teacher education (Ebrahimi et al., 2016; Im & Lee, 2003) and English as second/foreign language (ESL/EFL) education (Annamalai, 2017; Ware, 2004; Zhong & Norton, 2018). As the "beating heart" of online course activities (Sull, 2009), its value has been explored and confirmed by many pertinent studies.

Substantial evidence was presented in the extant studies to support the claim that the incorporation of AOD in online courses increased student interaction (Hammond, 2005). Particularly, introverted students or the students who used to be silent or peripheral participants in traditional classrooms tended to seize opportunities in AOD to voice their opinions (Alharbi, 2018; Arbaugh, 2000; Bolloju & Davison, 2003; Young, 2008). Hew and Cheung (2003) concurred that participants in online discussions feel more comfortable in expressing their thoughts more freely and descriptively (p. 13). Additionally, some relevant studies uncovered the other beneficial aspects of AOD in building a learning community, strengthening students' sense of belonging, and improving participants' critical thinking skills (Bendriss, 2014; Comer & Lenaghan, 2013; Liu, 2007).

Accompanying these encouraging findings, the existing studies also identified several factors that might affect the effectiveness of AOD in online courses. Fung (2004) found that students lacked interest in online discussions under the pressure of finishing required readings within a limited time. Therefore, she emphasized the significance of a reasonable timeframe and the relevance between the discussion questions and course topics. Some other studies highlighted the importance of explicit and theoretically informed discussion guidelines (Delahunty, 2018). In addition, timely response from peers was another major factor that affected students' participation in AOD (Cheung & Hew, 2004). Hew et al. (2010) conducted a comprehensive review of 50 empirical studies on AOD and revealed some other contributing factors, including not seeing the need to participate, other participants' behavior, student personality, and technical aspects.

Despite the abundant research in this area, very few studies explored the utilization of AOD in Chinese language learning. Qian and McCormick (2014) examined the utilization of an online discussion forum among novice Chinese language learners (CLLs), and the findings confirmed its positive impact, enhancing learners' sense of belonging and providing support to conquer difficulties in learning Chinese. Wang and Vásquez (2014) employed Facebook as the AOD forum, which was proven to present pedagogical potentials in second language (L2) literacy practice among intermediate CLLs. Relevant studies on CHLLs are strikingly scarce. Only one study involved this group of learners (Zhang, 2009), investigating the usage of essay writing in an online discussion board among Chinese heritage and L2 learners. Research findings indicated that the activity might facilitate creating a supportive learning community among different groups of Chinese learners. Among the studies, the book by Liu (2022) comprehensively discussed how the Chinese language has been taught in emergency remote learning, including examples from different parts of the world. However, there is little discussion that specifically addresses online heritage language teaching which presents unique challenges and needs due to the student group's distinctive language profiles compared to non-heritage students.

In sum, both the constraints and affordances of AOD in distant learning have been extensively discussed in various disciplines. However, relevant research in Chinese language education in general, and in CHL teaching in particular, remains scant. Despite the paucity, all the conducive and empirically proven findings in different fields serve as a great reference point for the AOD design in this article.

3. Overview of the course

The designing and implementation of the online CHL course in this article took place at an American private research university with a quarter system where dual-track Chinese language courses have been offered with a long history. The component of AOD was integrated into a second-year (intermediate level) Chinese language course for heritage learners. There were 28 students enrolled in this course, 13 in one section and 15 in the other. All the enrolled students were CHLLs whose proficiency levels ranged from intermediate-mid to intermediate-high according to ACTFL proficiency guidelines (ACTFL, 2012). It should be noted that this course was offered remotely only during the pandemic but has switched to the in-person mote in the post-pandemic era.

The course met four days a week, fifty minutes for each session. Most meeting days (i.e., three out of four) remained synchronous with one day being asynchronous when the course moved online in Spring. The asynchronous mode was adopted primarily to alleviate the stress experienced by students who were geographically dispersed in areas such as California, Chicago, and Hong Kong, as the affordances of asynchronous instruction allow for learning that breaks the temporal constraints. Furthermore, the asynchronous mode is normally arranged for the first teaching day of a new chapter, a great fit for students of different proficiency levels to self-study the basic vocabulary and grammar at their own pace and get ready for more meaningful practices in class. Consequently, the synchronous sessions may be devoted to task-oriented practices or discussions instead of drilling words and patterns that are tedious and less needed for CHLLs. Figure 1 illustrates the overall structure in which the two components were organized.

As indicated in the figure, this course used the textbook—Integrated Chinese (IC) Level 2 Part 2 along with the supplementary reading materials prepared by the instructor. Each quarter (i.e., ten weeks in total) covers six or seven lessons from IC. In spring quarter of 2020, the course started from Lesson 14 and finished the rest of the book. There were daily assignments due the next teaching day. For instance, Voicethread assignments were designed to guide students' asynchronous preview on Tuesdays. Besides these daily assignments, students were expected to carry out AOD on a designated platform, accounting for 15% of their final grades. How this component was designed and implemented will be outlined in detail in the following sections.

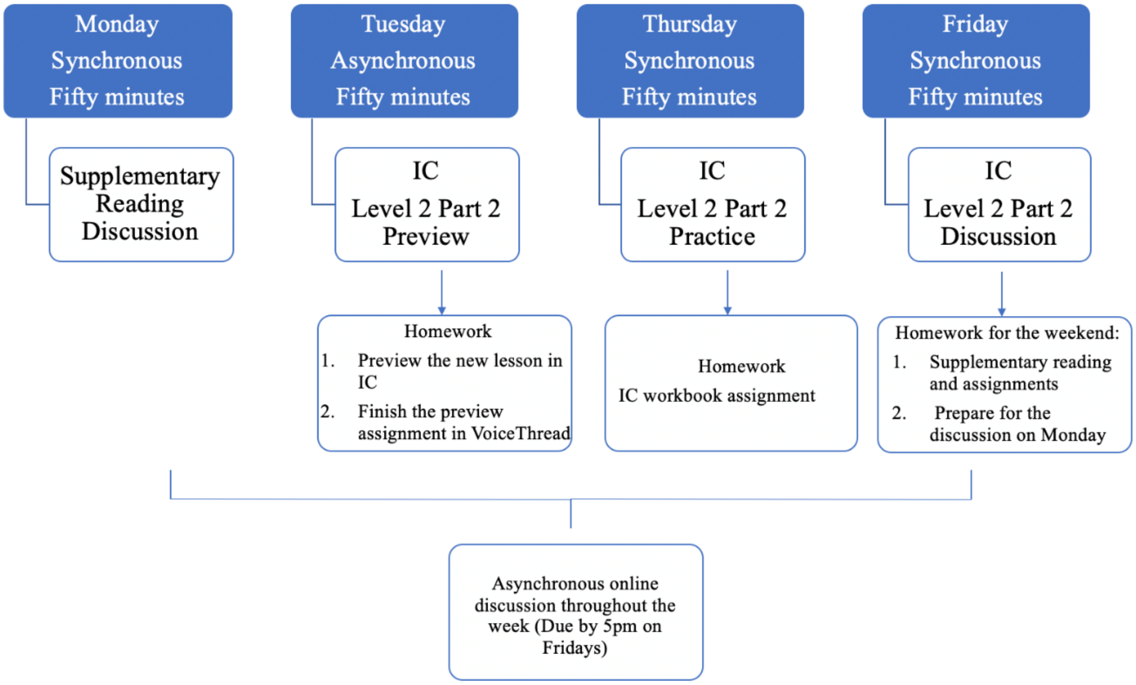

Monday
Synchronous
Fifty minutes
Tuesday
Asynchronous
Fifty minutes
Thursday
Synchronous
Fifty minutes
Friday
Synchronous
Fifty minutes
Supplementary Reading Discussion
IC Level 2 Part 2 Preview
IC Level 2 Part 2 Practice
IC Level 2 Part 2 Discussion
Homework
1. Preview the new lesson in IC
2. Finish the preview assignment in VoiceThread
Homework
IC workbook assignment
Homework for the weekend:
1. Supplementary reading and assignments
2. Prepare for the discussion on Monday
Asynchronous online discussion throughout the week (Due by 5pm on Fridays)

Figure 1 Course Structure

4. AOD design & implementation

The decisions about the various dimensions of AOD in this CHL course were deeply rooted in the prior empirical studies as well as considering the CHLLs' needs. To illustrate, the ensuing sections present the utilization of the AOD in this CHL course from the following angles: 1) Goals, 2) Platform, 3) Design, 4) Implementation, and 5) An example.

4.1 AOD goals

As discussed earlier, there are substantial benefits that students may gain from active participation in a well-designed AOD. However, the task per se or the platform that AOD is conducted on does not automatically lead to students' active and consistent participation. The elements affecting students' contribution to and learning outcomes from AOD should be factored into the design. Many studies suggested that curriculum designers not overload students within an online environment (Hammond, 2005). To put realistic and achievable expectations, the primary goals of the AOD in this course are two-fold:

- Community-building: As students use this space to interact with each other on a regular basis, it is hoped that a community could be built to provide social support, which seems to be especially important when classes are all remote.

- Resource-sharing: This space is intended to be where students share different types of outside-of-class resources relevant to the curriculum. Due to the high heterogeneity among CHLLs in terms of their linguistic and cultural repertoires, it is paramount to acknowledge and appreciate what each of them brings into the classroom, meaning that the curriculum should be built upon their "funds of Knowledge" (González et al., 2006). Additionally, this resource pool could be a venue for the instructor to know the students better before bridging the gap between in-class discussion and students' interests. In this sense, what students share in AOD will determine the content of the synchronous discussions.

4.2 AOD platform

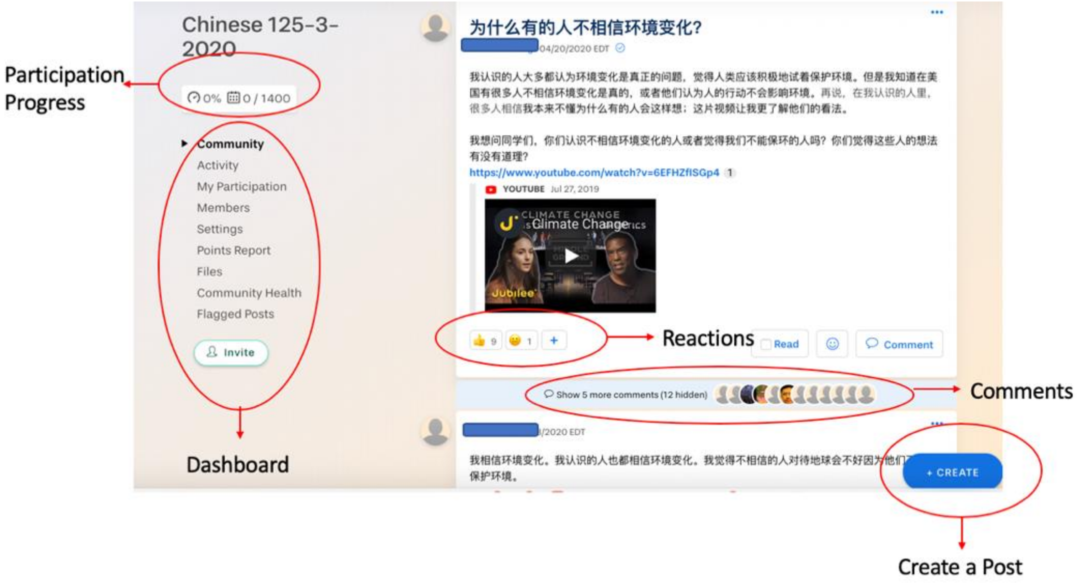
This course adopted a social learning platform named "Yellowdig" for the AOD component, primarily for four reasons.

First, Yellowdig has an interface similar to one of the most popular social networking websites—Facebook. Such similarity incorporates the communication that students are familiar and comfortable with into Chinese language learning. They intuitively know how to navigate the platform, how to make multimodal postings, such as texts, photos, emoticons, videos, and the like, and how to interact with each other (e.g., like and comment), which should reduce the learning curves that students might have otherwise. Further, these functions provide more lavish features for social learning (Huang & Chen, 2018) compared to the traditional threaded discussion boards (e.g., Canvas discussion board).

Second, Yellowdig is a social learning platform designed specifically for educational purposes and is inherently different from other social media tools such as Facebook and Twitter. As students prefer not to intertwine their academic studies and personal social lives (A & Gutsch, 2018; Jones et al., 2010), Yellowdig can serve as an ideal substitute that both inherits students' usual social habits and creates a separate social space for students to interact with each other.

Third, Yellowdig provides a very convenient and motivating grading system. It automatically grades students' participation according to the rubrics set up by the instructor in the system. Moreover, the platform may be seamlessly integrated into students' learning management systems (LMS), such as Canvas and Blackboard, so that the grades may be automatically synched in the LMS. Unlike traditional grading, Yellowdig intends to gamify the points-earning system, as students do not lose points but rather earn rewards for their contributions in the AOD. For instance, the instructor may design the rubrics in the system, allowing students to earn 100 points for a post, 80 points for a comment, 20 points for a "like" they receive from peers, and the like, with a weekly goal of 1000 points in total. The instructor may also require a minimum number of words in one post or comment. In addition to quantifying students' participation, the instructor may revoke the points if a post or comment is believed to be irrelevant, not well-thought-out, or does not contribute meaningfully to the conversation. This is to emphasize the quality of students' contributions to the AOD. Instructors indicated that the quality and quantity of students' posts in Yellowdig increased by more than 50% compared to other online discussion platforms (Gulinna & Gutsch, 2018, p.281).

Finally, the affordances of Yellowdig suit the aforementioned two goals of the AOD in this course. The utilization of Yellowdig could encourage students to be more actively engaged in participating in the AOD. The increased peer interaction is the premise for community building. Moreover, as maintained by Gulinna and Gutsch (2018), the layout of Yellowdig can promote learners to create a knowledge base for the entire class and utilize the shared resources in their future studies (p. 282), which is consistent with what this CHL course aimed to achieve. Figure 2 is a screenshot of Yellowdig that provides a look into the platform.



为什么有的人不相信环境变化?

[redacted] 04/20/2020 EDT

我认识的人大多都认为环境变化是真正的问题，觉得人类应该积极地试着保护环境。但是我知道在美国有很多人不相信环境变化是真的，或者他们认为人的行动不会影响环境。再说，在我认识的人里，很多人相信我本来不懂为什么有的人会这样想；这片视频让我更了解他们的看法。

我想问同学们，你们认识不相信环境变化的人或者觉得我们不能保环的人吗？你们觉得这些人的想法有没有道理？

https://www.youtube.com/watch?v=6EFHZflSGp4 1

[redacted] /2020 EDT

我相信环境变化。我认识的人也都相信环境变化。我觉得不相信的人对待地球会不好因为他们
保护环境。

Figure 2 Yellowdig interface

4.3 AOD design

As alluded to earlier, merely using new technology or a fancy tool does not automatically assure the expected learning outcome. Many other factors, especially the curriculum design and pedagogical decisions, exert a much more substantial impact on students' performance in AOD. This section details how Yellowdig was used in this CHL course to achieve the two objectives mentioned above.

A commonly adopted practice when using AOD in various disciplines is that the prescribed discussion questions are posted on threaded discussion boards by instructors or TAs after learning a new concept, unit, or chapter. Then students are required to answer these questions and respond to at least two peers' posts. The flow of AOD is from teachers to students. Although well-intended, authentic and meaningful communication among students might be hard to realize as students are probably not interested in these questions in the first place. On top of that, most students are forced to contribute under the pressure of losing points. In contrast, Yellowdig in this course is utilized backward from student to teacher to alleviate such concerns. Specifically, it functions in the following two ways.

First, the Yellowdig platform is open for four days, starting from 5 p.m. on Mondays to 5 p.m. on Fridays. Students are expected to share resources (e.g., videos, articles, songs, or anecdotes) related to the weekly class topics. They are also required to briefly explain the reasons for, the main points of, and their reactions to the shared materials. The resources could be either in English or Chinese; however, students' annotations and comments should be in Chinese.

Second, the instructor skims students' posts after the platform closes at 5 p.m. on Fridays for two purposes: 1) To identify students' common mistakes in using Chinese and the areas for improvement in discussion participation so that the instructor could provide the whole class feedback when meeting synchronously. Chiu and Hew (2018) underscored the importance of teacher feedback in AOD, stating that more constructive and timely feedback can encourage learners to participate more in the discussion forum (p. 18); 2) For the instructor to pinpoint the most intriguing, engaging, and thought-provoking topics based on the number of postings and comments. Subsequent supplementary reading materials are prepared based on such knowledge, in an attempt to bridge the gap between course materials and students' interests. Finding the appropriate reading materials is vital as selecting the right topics is one of the major contributing factors to the success or failure of any discussion (Bakar et al., 2013). Following that, students are expected to read the article(s) and complete the corresponding assignments so that they may readily participate in the synchronous discussions on the next teaching day.

It is noteworthy that the AOD interaction occurred primarily among students themselves while the instructor remained silent throughout the open period. In terms of the impact of instructor intervention in AOD, the research found that students participated less as an instructor's posts increased (Mazzolini & Maddison, 2007), and students interacted more with the instructor rather than with their peers (An et al., 2009). Participants expressed their preference not to have the instructor participate in their online discussions as teachers' omnipresent participation can be oppressive to certain students (Fauske & Wade, 2003). Consequently, Andresen (2009) explicitly pointed out that an instructor should "back off" and "spend his/her time preparing materials and carefully thought-out discussion questions and topics that relate to learning objectives" (p. 251). Meanwhile, the instructor's critical role in maintaining and facilitating students' AOD was also underscored by the pertinent studies (e.g., An et al., 2009; Zhu, 2006). Dennen (2005) maintained that it was an act of balancing in establishing instructor presence as the most favorable presence seemed to be letting students know that their messages were read without taking over the discussion (p. 142).

Drawing upon the research findings and out of pedagogical concerns, the instructor decided not to participate in students' AOD to avoid the negative impact of instructor presence indicated in prior studies. To cultivate and sustain students' discussion on Yellowdig, the instructor built her presence primarily outside AOD in lieu of during AOD, concentrating more on designing the discussion guidelines, reading students' posts, finding appropriate supplementary reading articles, and providing feedback.

4.4 AOD implementation

Prior to the start of the new quarter, an email was sent out to the enrolled students, introducing the discussion site—*Yellowdig*—and inviting them to get acquainted with each other and share their life and concerns about taking an online language course. Students were encouraged to explore the site and use the multimodal resources to make their self-introductions more visual and interactive. Although there were two sections for the course, only one community was created on Yellowdig as many students across sections took Chinese classes together in the previous two quarters and already knew each other. Additionally, a larger group might result in more resources shared in the community. Therefore, students have more options as to whom to interact with and what posts to read.

On the first day of the quarter, the instructor shared a document named "Yellowdig Discussion Guidelines" with all the students. The guidelines consisted of four components: 1) A brief introduction to Yellowdig and its weighted percentage; 2) Purpose of using Yellowdig AOD; 3) Yellowdig discussion protocols. In addition to laying out the expectations for the content of posts, the protocols also reminded students of the strategies of effective and civil communication online. For instance, it emphasized the importance of reacting to others' posts, which was not only an encouraging way to contribute to the community but also signified to the instructor what they were interested in. It also suggested students not wait until the last minute to post. The earlier they started posting, the higher chance they would get a reply as it provided ample opportunities for their peers to share their comments; 4) The overall rating scale of AOD, including four areas: quality, quantity, consistency, and etiquette. The quantity part was automatically measured by Yellowdig as discussed earlier. However, the rating scale reminded students that the instructor evaluates the other three aspects as well. For instance, the instructor observed whether students made steady and consistent contributions throughout the open days of Yellowdig to keep the conversation flowing. The quality fell into two sub-areas: language and content. The instructor assessed if there were errors in wording and whether the posts were logically organized and supported by details and examples. Etiquette was emphasized as well as students were expected to interact with each other respectfully, politely, and insightfully. Please see the complete content of the file in Appendix 1.

The first week was allocated for testing out the platform, the guidelines, and the reward-earning system in Yellowdig. Therefore, students' performance of that week was not counted into their final grades. An anonymous survey was administered among students over the first weekend so that the instructor could identify the problematic areas and make in-time adjustments accordingly. Overall, the piloting went smoothly, and students' participation in the community was satisfying. Surprisingly, one student even explored a new function on Yellowdig that neither the instructor nor other students had discovered—polling. This simple polling that student initiated among her peers, investigating their opinions of eating late-night snacks as the theme in the first week was health and lifestyle (IC Level 2 Part2, L14). Seventeen responses were received in total, building a foundation for further discussion in class.

Nevertheless, there were still two students who remained reticent in Yellowdig: one did not participate at all, and the other only reacted to two peers' postings with a smiling emoji. The instructor had foreseen such inactiveness when the institution announced during the spring break that all the undergraduate courses' gradings would be "pass or fail" to replace letter grades due to the pandemic's impact. Therefore, some students might feel much less motivated to make the greatest endeavor in their studies. To encourage these two students to be more engaged, the instructor sent out emails, inviting them to share more actively in the rest of the quarter.

The survey results also revealed a couple of problems and corresponding fine-tuning was made.

First, students reflected that Yellowdig counted words based on the number of spaces between the words, which apparently does not apply to the Chinese writing system. Hence, some students did not receive credits because the system erroneously considered that their postings were short of words although their postings met the requirement. Due to the flaw in the system per se, the instructor had to give up the requirement of a minimum word count starting from week two.

Second, students expressed that the weekly goal was a bit overwhelming as they were pressured to post as many as possible to earn the rewards; however, they neglected the quality of their contributions. To strike a better balance between the quality and quantity of students' posts and make the weekly goal more manageable, the instructor revised the reward-counting system from 1400 points in total to 800 points as reflected in Table 1.

Table 1 Rewarding System

Category	Rewards
A posting with 80 words minimum→ A posting	100 points
A comment with 50 words minimum→A comment	80 points
Receiving one comment	50 points
Receiving other reactions (e.g., emoticons)	10 points
Weekly Goal	1400 points→800 points (The surplus points may be accumulated for the following week.)

4.5 AOD—An example

This section uses Lesson 15—*Gender Equality*—from IC as a concrete example to present what students shared on Yellowdig and how the platform connected synchronous and asynchronous discussions.

Gender equality is never an easy topic. The textbook's content consists of two components: the story between Xuemei's (the character's name) uncle and his wife and a brief dialogue about Chinese men soccer. The text itself is not that difficult for CHLLs in this course. Evidently, they need supplementary materials to expand their readings and enrich relevant discussions. However, if not meticulously designed, the discussion questions could easily be too broad and general that students feel distant from such a topic and do not know what to say. Alternatively, the questions could be too challenging because language learners, especially the ones with lower proficiencies, do not have adequate words and grammar to articulate their ideas fully.

Throughout the four days that Yellowdig was open that week, students posted various types of materials about gender inequality, including relevant news articles, YouTube videos, and movie clips. The relevant topics that students submitted fell into a wide range as well. The best-received ones included 1) Gender inequality in Disney movies, 2) Social expectations for women, 3) Toys and gender roles, 4) Men's perceptions of gender issues, 5) Kids' perceptions of gender issues, and 6) Students' anecdotes. Examples of the posts could be found in Appendix 2.

Built upon students' AOD, the instructor eventually decided to adopt a news article titled "If I were a boy," which was about an online feminist campaign initiated by a website named Elite Daily. The article was selected because 1) this article only needed minimal adaption to better match the CHLLs' Chinese proficiency, and 2) the relevant discussions about this article allowed integrating many of the topics from Yellowdig. The questions (originally in Chinese, translated into English in this article) used in the subsequent synchronous session are listed below, which primarily stemmed from or were inspired by students' discussion on Yellowdig.

- 你遇到过男女不平等的情况吗？比如在你家、实习的时候、学校或者其他社交场合？如果愿意的话，请分享你的经历。
  Did you encounter gender inequality in your family life, internship, academic studies, or your social life? Please share if you feel comfortable.

- 如果你是男孩/女孩，你会跟现在不一样吗？为什么？请举例。

  Do you do things differently if you were a boy/girl? Why? Please give examples.
- 男女不平等常常让男性处于优势，那男女平等对男性有好处吗？他们也应该争取男女平等吗？为什么？

  Men are generally privileged in this society. Should they also strive for gender equality? Could they benefit from gender equality? Why?
- 你愿意做家庭主妇或者家庭煮夫吗？是浪费你的才能或高学历吗? 如果你是男性/女性，你会有不同的选择吗？

  Is it acceptable for you to be a housewife or a soccer dad? Is it a waste of your talents and diploma? If you were a man/woman, will you decide differently?
- 你能接受你的儿子玩芭比娃娃、你的女儿玩赛车吗？

  Is it acceptable to you if your son likes playing with barbie dolls and your daughter enjoys car-racing?
- 还有哪些性别刻板印象？比如在公司、学校、好莱坞电影里？请举例。

  What are the other gender stereotypes in different areas such as industry, academia, and Hollywood movies? Please give an example.

5. Students' reflections

Students were invited to submit a reflection on their Yellowdig discussions and participate in an interview with the instructor. To avoid conflicts of interest, both the reflections and interviews were scheduled at the end of the quarter after the instructor submitted all the grades. Specifically, they were encouraged to reflect on the aspects including but were not limited to 1) Their overall experience, 2) The beneficial aspects of Yellowdig discussion, 3) The drawbacks, and 4) Their suggestions. Eleven students in total submitted their reflections, and four students voluntarily participated in the individual interviews.

Two students indicated that their experience on Yellowdig was OK and candidly admitted that their participation was primarily driven by the weekly point requirements they needed to reach. Nevertheless, the rest reported rather favorable attitudes towards the AOD, confirming the social and educational benefits of participating in the Yellowdig discussions. As one summarized, "I think that during the course of online classes, yellowdig [sic] discussions can be a useful and productive way for students of class to interact with each other as well as practice their Chinese."

In the social aspect, students' reflections revealed that Yellowdig afforded space for increasing peer interaction while they were geographically dispersed, and they might learn more about other classmates in general, confirming research findings in prior studies (e.g., Hammond, 2005; Zhang, 2009). For instance, one student commented, "My overall experience with Yellowdig discussions was positive. I was able to interact with my classmates even though we did not see each other in person. Another student expressed, "Given the nature of online learning, I find it a nice way to interact with my peers."

Despite the overall positive social experience, one problematic aspect was identified in the students' reflection. Some peers' superficial comments made students' experience in the Yellowdig discussion less enjoyable. One student conveyed:

> Some classmates would post very thought-provoking discussion posts that I could tell actually showed that they cared; however, others (particularly in the replies), would leave brief comments just to say they did the assignment. It makes having genuine conversation difficult, and I hate that.

Students' feedback confirmed the findings by Hew et al. (2010) that peer behavior is one of the factors that affect students' learning experience in AOD. Although the course designers were attentive to this aspect when mapping out the guidelines for Yellowdig AOD, some students still put more weight on quantity compared to quality. In response to this problem, instructors may consider making the AOD activities much lower-stakes so that students would be less pressured to post copiously but more motivated to discuss thoughtfully.

Intriguing and resonating topics shared in Yellowdig discussions is another factor that contributed to students' positive social and learning experience. Students shared that the discussion-format style of Yellowdig gave them the chance to interact with interesting topics and concepts. One student reflected, "I think it was the right decision to have yellowdig [sic] posts focus on the topic of the lesson, because it gives the users something to post about." Students particularly appreciated the opportunities to make connections from the lesson texts to the world they live in and to things that are more relevant to them. As reflected in their comments, relating the materials learned in class directly to real-life events was intriguing. One student concurred and summarized in the reflection:

> Overall, I actually really liked the concept of the Yellowdig discussions because a lot of my classmates would bring up interesting questions, information, or viewpoints about the topics that we are currently covering in class and I think that it helped me make connections from our text to the world we live in and to things that are more relevant to us.

In addition to relevance, students' reflections and interviews indicated that the Yellowdig discussion expanded the scope of the topics as well. One student commented, "I think the Yellowdig discussion is quite interesting and can help promote exploration of topics that students might otherwise not be exposed to, while practicing Chinese at the same time." As conveyed in their reflections, students particularly enjoyed discussing topics surrounding Asian Americans with their peers. Students also appreciated the freedom and autonomy they had in Yellowdig, as one student reflected, "I felt Yellowdig discussions were a useful and interesting way to interact with classmates. I liked being also to freely choose what type of content we shared with each other and discussed." Another student added, "The ability to share articles and interesting findings with my classmates made Yellowdig more purposeful." Students' reflections above echoed the importance of topic selection in AOD emphasized in prior studies (e.g., Andresen, 2009; Bakar et al., 2003). When instructors are unsure of students' interests, giving them certain autonomy in topic selection could be a feasible and well-received method.

The supplementary reading articles based on students' Yellowdig discussions were overwhelmingly well received among students, which was another rather encouraging finding. This further confirmed the importance of selecting the appropriate topic and materials as discussed above. Many students brought up that the supplementary materials were really interesting and captivated their interests. They truly enjoyed reading these articles, putting thoughts together, and making insightful responses to the reading assignments' questions. One of the students even rated it as her favorite part of the course. Additionally, compared to the relatively short posts in Yellowdig discussions, students found supplementary readings and corresponding assignments afforded them a venue to elaborate their thoughts further. One student commented:

> I enjoy listening to/reading the supplementary material that laoshi finds and responding to it in an essay. This gives me more time to put some thought and effort into one response.

Additionally, the asynchronous nature of the Yellowdig discussion made the task more manageable for the students in different time zones. One student commented, "Yellowdig discussion was a good way to share ideas and communicate because I liked being able to view other people's content and respond at any time that worked for me."

The challenges of Yellowdig participation primarily rested in two areas. First, some students found Yellowdig very helpful to their Chinese learning as they had to constantly read and familiarize themselves with Chinese characters, which confirmed the pedagogical potential of AOD in Wang and Vásquez (2014). This is particularly useful for heritage language learners due to their relative weaker proficiency in reading and writing compared to their listening and speaking. However, the Yellowdig discussion posed some challenges to the students with relatively lower Chinese proficiency. Therefore, Google Translate was frequently used, which was energy-draining to them. Some students expressed that a lack of knowledge of many new words in Yellowdig discussions sometimes discouraged them from participating. Second, students expressed the difficulty they ran into in writing on discussion boards. As one student explicitly shared:

> I realized it's a lot more difficult than I thought to transfer between conversational Chinese (which I am already proficient in), to presentation/formal Chinese, which I am still struggling to learn.

One student indicated that it often took him/her a while to plan out and organize what she/he wants to say in a post or comment. Another student echoed that using Chinese to post made it more difficult to convey complex ideas. As pointed out in pertinent studies, the writing on discussion boards is a different genre of writing, a hybrid mode of spoken-like/written-like communication (Delahunty, 2018). The challenges that students met in this course necessitate more meticulously designed tasks that involve students in authentic online discussions in the target language community. Developing students' digital literacy in Chinese to appropriately communicate online should be an integral part of Chinese language education.

Second, the amount of schoolwork they received from other courses, along with other Chinese assignments, was the main deterrent that prevented them from participating. This is consistent with the findings in Fung (2004) that students usually lack interest in AOD due to the limitation of time. To encourage student participation, some students suggested making the point totals more achievable and the bar low enough that they could craft insightful responses. Despite the intended gamification of the grading system, many students still felt pressured by the weekly goal. Therefore, they sometimes just provided superficial comments to get the points. Another suggestion was listing posts based on categories, such as sports and home life, to make the discussion more organized and easier to follow. This may be realized by the “hashtag” function on Yellowdig, which curriculum designers and Yellowdig users should further explore. Moreover, it was recommended by some students to encourage participants to use different types of media on Yellowdig, such as polls, photos, and videos, to keep the discussion intriguing.

6. Conclusions

AOD has been a common feature of online education, while research on the utilization of AOD in Chinese language learning remains alarmingly scant. This article demonstrates an effort to integrate this component into an online CHL course. The social learning platform, Yellowdig, was selected to conduct the AOD out of pedagogical considerations, allowing the digital natives to discuss with each other in ways they are used to, as well as providing them a social space that is separate from their private social networking accounts. Decisions about various dimensions of AOD were premised on the empirically supported findings from prior studies. The students’ overall positive reflections confirmed that the Yellowdig discussion fulfilled its designated goals—community building and resource sharing—and indicated the promising utilization of AOD in other CHL courses or the advanced-level Chinese language courses in the non-heritage track. Though AOD was used in an online course, the findings could serve as useful references for in-person courses as well.

References

ACTFL, J. (2012). ACTFL proficiency guidelines 2012. https://www.actfl.org/uploads/files/general/ACTFLProficiencyGuidelines2012.pdf

Alharbi, M. A. (2018). Patterns of EFL Learners' and Instructor's Interactions in Asynchronous Group Discussions on Free Writing. *Journal of information technology education, 17*, 505. doi:10.28945/4143

An, H., Shin, S., & Lim, K. (2009). The effects of different instructor facilitation approaches on students’ interactions during asynchronous online discussions. *Computers & Education, 53*(3), 749-760. doi:10.1016/j.compedu.2009.04.015

Andresen, M. A. (2009). Asynchronous discussion forums: success factors, outcomes, assessments, and limitations. *Journal of Educational Technology & Society, 12*(1), 249-257. https://www.learntechlib.org/p/75172/

Annamalai, N. (2017). An Investigation into the Community of Inquiry Model in the Malaysian ESL Learners' Context. *Interactive Technology and Smart Education, 14*(3), 246. doi:10.1108/ITSE-07-2016-0021

Arbaugh, J. B. (2000). Virtual classroom versus physical classroom: An exploratory study of class discussion patterns and student learning in an asynchronous Internet-based MBA course. *Journal of Management Education, 24*(2), 213-233. doi:10.1177/105256290002400206

Bakar, N. A., Latiff, H., & Hamat, A. (2013). Enhancing ESL learners speaking skills through asynchronous online discussion forum. *Asian Social Science, 9*(9), 224-233. doi:10.5539/ass.v9n9p224

Barrett-Fox, R. (2020). Please do a bad job of putting your course online. https://anygoodthing.com/2020/03/12/please-do-a-bad-job-of-putting-your-courses-online/

Bendriss, R. (2014). Asynchronous online discussions: Perceptions on second language reading, writing, and critical thinking. Paper presented at the Third International Conference on E-Learning & E-Technologies in Education, Kuala Lumpur, Malaysia. https://www.proceedings.com/content/027/027185webtoc.pdf

Bolloju, N., & Davison, R. (2003). Learning through asynchronous discussions: experiences from using a discussion board in a large undergraduate class in Hong Kong. *eLearn, 2003*(6), 4. doi:10.1145/863928.863936

Cheung, W. S., & Hew, K. F. (2004). Evaluating the extent of ill-structured problem solving process among pre-service teachers in an asynchronous online discussion and reflection log learning environment. *Journal of Educational Computing Research, 30*(3), 197-227. doi:10.2190/9JTN-10T3-WTXH-P6HN

Chiu, T. K., & Hew, T. K. (2018). Factors influencing peer learning and performance in MOOC asynchronous online discussion forum. *Australasian journal of educational technology, 34*(4). doi:10.14742/ajet.3240

Comer, D. R., & Lenaghan, J. A. (2013). Enhancing discussions in the asynchronous online classroom: The lack of face-to-face interaction does not lessen the lesson. *Journal of Management Education, 37*(2), 261-294. doi:10.1177/1052562912442384

Delahunty, J. (2018). Connecting to learn, learning to connect: Thinking together in asynchronous forum discussion. *Linguistics and Education, 46,* 12-22. doi:10.1016/j.linged.2018.05.003

Dennen, V. P., & Wieland, K. (2007). From Interaction to Intersubjectivity: Facilitating online group discourse processes. *Distance education, 28*(3), 281-297. doi:10.1080/01587910701611328

Dennen, V. P. (2005). From message posting to learning dialogues: Factors affecting learner participation in asynchronous discussion. *Distance education, 26*(1), 127-148. doi:10.1080/01587910500081376

Ebrahimi, A., Faghih, E., & Marandi, S. S. (2016). Factors Affecting Pre-Service Teachers' Participation in Asynchronous Discussion: The Case of Iran. *Australasian journal of educational technology, 32*(3), 115. doi:10.14742/ajet.2712

Fauske, J., & Wade, S. E. (2003). Research to practice online: Conditions that foster democracy, community, and critical thinking in computer-mediated discussions.

Journal of Research on Technology in Education, 36(2), 137-153. https://files.eric.ed.gov/fulltext/EJ690928.pdf

Fung, Y. Y. (2004). Collaborative online learning: Interaction patterns and limiting factors. *Open Learning: The Journal of Open, Distance and e-Learning, 19*(2), 135-149. doi:10.1080/0268051042000224743

Gulinna, A, G., & Gutsch, S. (2018). Optimizing learner experience with intuitive asynchronous online discussion design. Paper presented at the The Annual Conference of the Association for Educational Communications and Technology, Kansas City, MO.

González, N., Moll, L. C., & Amanti, C. (2006). *Funds of knowledge: Theorizing practices in households, communities, and classrooms*: Routledge.

Hammond, M. (2005). A review of recent papers on online discussion in teaching and learning in higher education. *Journal of Asynchronous Learning Networks, 9*(3), 9-23. doi:10.24059/olj.v9i3.1782

Hew, K., Cheung, W., & Ng, C. (2010). Student contribution in asynchronous online discussion: a review of the research and empirical exploration. *Instructional Science, 38*(6), 571-606. doi:10.1007/s11251-008-9087-0

Hew, K. F., & Cheung, W. S. (2003). An exploratory study on the use of asynchronous online discussion in hypermedia design. *Journal of Instructional Science & Technology, 6*(1), 12-23. https://ascilite.org/archived-journals/e-jist/docs/Vol6_No1/hew.htm

Huang, T., & Chen, B. (2018). Uncovering the rich club phenomenon in an online class. In J. Kay & R. Luckin (Eds.), *Rethinking Learning in the Digital Age: Making the Learning Sciences Count* (Vol. 3). London, UK: International Society of the Learning Sciences, Inc.[ISLS].

Im, Y., & Lee, O. (2003). Pedagogical implications of online discussion for preservice teacher training. *Journal of Research on Technology in Education, 36*(2), 155-170. doi:10.1080/15391523.2003.10782410

Jones, N., Blackey, H., Fitzgibbon, K., & Chew, E. (2010). Get out of MySpace! *Computers & Education, 54*(3), 776-782. doi:10.1016/j.compedu.2009.07.008

Liu, S. (2007). Assessing online asynchronous discussion in online courses: an empirical study. *Proceedings of Technology, Colleges and Community Worldwide Online Conference*, USA, 24-32. http://hdl.handle.net/10125/69280

Liu, S. (Ed.). (2022). *Teaching the Chinese language remotely: Global cases and perspectives.* Springer Nature.

Mazzolini, M., & Maddison, S. (2007). When to jump in: The role of the instructor in online discussion forums. *Computers & Education, 49*(2), 193-213. doi:10.1016/j.compedu.2005.06.011

Qian, K., & McCormick, R. (2014). Building course cohesion: the use of online forums in distance Chinese language learning. *Computer Assisted Language Learning, 27*(1), 44-69. doi:10.1080/09588221.2012.695739

Sull, E. C. (2009). The (almost) complete guide to effectively managing threaded discussions.(Try This). *Distance Learning, 6*(4), 65.

Wang, S., & Vásquez, C. (2014). The effect of target language use in social media on intermediate-level Chinese language learners' writing performance. *CALICO Journal, 31*(1), 78-102. doi:10.11139/CJ.31.1.78-102

Ware, P. D. (2004). Confidence and competition online: ESL student perspectives on web-based discussions in the classroom. *Computers and Composition, 21*(4), 451-468. doi:10.1016/j.compcom.2004.08.004

Young, A. (2008). Structuring asynchronous discussions to incorporate learning principles in an online class: One professor's course analysis. *MERLOT Journal of Online Learning and Teaching, 4*(2), 218-224. https://jolt.merlot.org/vol4no2/young0608.pdf

Zhang, D. (2009). Essay writing in a Mandarin Chinese WebCT discussion board. *Foreign Language Annals, 42*(4), 721-741. doi:10.1111/j.1944-9720.2009.01051.x

Zhong, Q. M., & Norton, H. (2018). Educational affordances of an asynchronous online discussion forum for language learners. *TESL-EJ (Berkeley, Calif.), 22*(3). http://www.tesl-ej.org/wordpress/issues/volume22/ej87/ej87a1/

Zhu, E. (2006). Interaction and cognitive engagement: An analysis of four asynchronous online discussions. *Instructional Science, 34*(6), 451. doi:10.1007/s11251-006-0004-0

Appendix 1
Guidelines for Participation and Interaction on Yellowdig

Yellowdig (15%)

We are going to use Yellowdig for our asynchronous discussion. You may access the discussion platform through Canvas. In this course, Yellowdig is primarily used for our weekly out-of-class discussion among students. Please check the guidelines below for your participation and interaction with your peers on Yellowdig.

Purposes of Using Yellowdig

1) **Community-building:**

We would like to have a space to interact with the peers, which is especially important in these uncertain times when we have class remotely. Additionally, Laoshi would like to provide you a space to discuss topics of interest to you with your peers rather than topics imposed by Laoshi. Points will be assigned to you for acknowledging your contribution and social interaction. There is a built-in grading system in Yellowdig. Besides quantity, there are some other areas that Laoshi looks at when evaluating your participation. The rating scale is laid out in the last section of this guideline.

2) **Resource-sharing:**

Laoshi would like to provide you with a platform to share different types of outside-of-class resources relevant to our curriculum. Additionally, Laoshi would like to use your discussion posts as a topic-pool and foundation for our in-class discussion. That means Laoshi will read your posts, identify the topics that interest you most, and incorporate them into the supplementary reading and in-class discussion. In this sense, what you will share on Yellowdig will determine the content of our synchronous sessions.

Yellowdig Discussion Protocols

1. You are expected to share resources (e.g., videos, articles, songs, photos, your anecdotes) that are interesting and weekly theme related.
2. In addition to posting the resources, please also briefly explain the reasons you would like to share, the main points, and your reactions to what you share, just like what you normally do when you share something on the other social media such as Facebook or Twitter. The resources could be either in English or Chinese; however, your annotation and comments should all be in Chinese.
3. Don’t forget to check others’ posts and react (e.g., like it and comment). As mentioned previously, this is supposed to be a community where you share information, exchange opinions and conduct discussions. Additionally, you will not only help your peers earn points but also let Laoshi know what interests you.
4. Your points could be revoked. The Yellowdig point system encourages high-quality comments. Laoshi can revoke a student's points if Laoshi believes a comment is not relevant, well-thought-out, or does not contribute meaningfully to the conversation. (For example, points will be revoked if you simply put a comment without further explanation—你说的很有意思。)
5. Based on research results, the earlier you post, the higher chance you will get a reply as it provides ample opportunities for your peers to comment. Don’t wait until the last minute before the deadline. Normally, the discussion platform will be closed at 10 am on Fridays.

Please participate consistently throughout the open period. The deadline is marked in the weekly schedule as well.

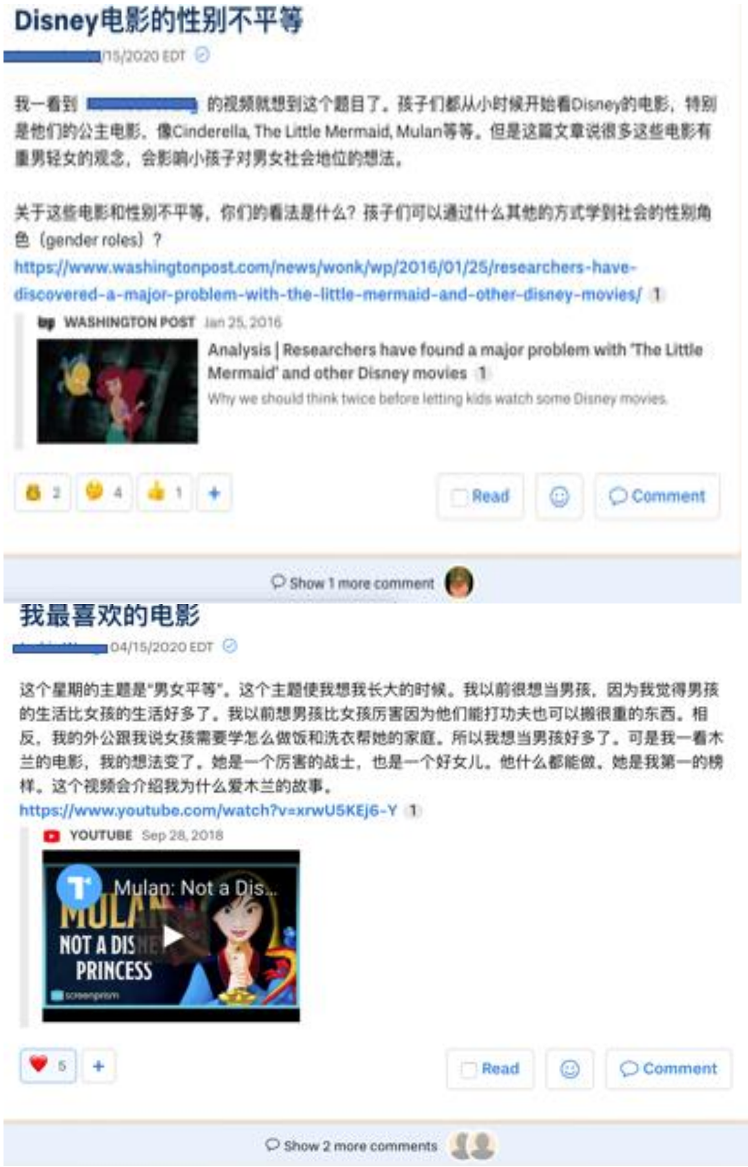
6. The first week of discussion will not be graded but for practice purposes. You will gain feedback that helps you prepare for future Yellowdig discussions.
7. You and Laoshi will conduct a reflection on Yellowdig activities together through anonymous surveys and open discussions and make the adjustments accordingly.

Rating Scale

	Quality		**Quantity**	**Consistency**	**Etiquette**
	Language	**Content**			
3	Minimal errors in wording A wide range of precise vocabulary and complex sentences Appropriate cohesive devices that link the ideas or information into a paragraph or paragraphs The posting is clearly presented and is easily understood by others	The posting is well and logically organized The posting is supported by details, examples, and/or evidence The posting is intriguing and inspiring to others	Actively participate in the conversations Frequently view peers’ posts Respond to diverse peers’ posts Very informative	Steady and consistent participation throughout the open days to keep the conversation flowing	Show appreciation (Acknowledge and appreciate your peers’ contribution.) Prompt response to peer posts Interact with others respectfully, politely and insightfully
2	Some errors in wording A range of general and specific vocabulary and some complex sentences Strings of sentences and occasionally a short paragraph with appropriate cohesive devices The posting is appropriately presented and is generally understood by others	The posting is adequately organized The posting is supported by some details The posting contributes ideas and somewhat facilitates conversations	Participate in the conversations Read most peers’ posts Attempt to respond to different peers’ posts Somewhat informative	Somewhat steady and consistent participation during the open time to facilitate the conversation	Show no sensitivity to others’ perspectives Show respect and sensitivity to peers’ backgrounds Respond to peer posts in a timely fashion
1	Many errors in wording General and sometimes specific vocabulary and simple sentences Strings of sentences without cohesive devices The posting is NOT clearly presented and is understood with some difficulty by others	The organization is problematic The posting is NOT supported by details The posting is a minor contribution to the conversation	Somewhat participate Read some peers’ posts Respond to a few peers’ posts Missing information	Inconsistent participation with little contribution to the conversation	Frequently not responding to peer posts Show little respect or sensitivity to peers’ views and backgrounds
0	Too many errors in wording Limited and general vocabulary Discrete and simple sentences The posting is understood with great difficulty by others	The posting is poorly organized The posting is irrelevant or simply a repetition of others’ statements There is no contribution to the conversations	Minimum participation Minimum effort to write a post	Last-minute posting or commenting	Show minimum effort to write a response (e.g., 我同意你的看法。谢谢分享。你的看法很有意思。) Show no respect or sensitivity to peers’ views and backgrounds

Adapted from A & Gutsch (2018)

Appendix 2
Examples: students' posts about gender equality



Disney电影的性别不平等

[redacted]/15/2020 EDT

我一看到 [redacted] 的视频就想到这个题目了。孩子们都从小时候开始看Disney的电影，特别是他们的公主电影，像Cinderella, The Little Mermaid, Mulan等等。但是这篇文章说很多这些电影有重男轻女的观念，会影响小孩子对男女社会地位的想法。

关于这些电影和性别不平等，你们的看法是什么？孩子们可以通过什么其他的方式学到社会的性别角色（gender roles）？
https://www.washingtonpost.com/news/wonk/wp/2016/01/25/researchers-have-discovered-a-major-problem-with-the-little-mermaid-and-other-disney-movies/ 1

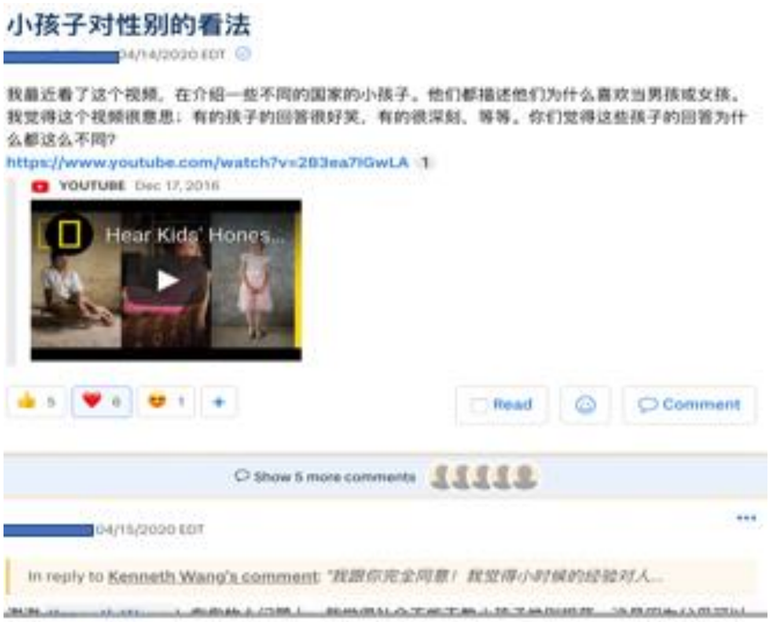
WASHINGTON POST Jan 25, 2016
Analysis | Researchers have found a major problem with 'The Little Mermaid' and other Disney movies 1
Why we should think twice before letting kids watch some Disney movies.

我最喜欢的电影

[redacted] 04/15/2020 EDT

这个星期的主题是"男女平等"。这个主题使我想我长大的时候。我以前很想当男孩，因为我觉得男孩的生活比女孩的生活好多了。我以前想男孩比女孩厉害因为他们能打功夫也可以搬很重的东西。相反，我的外公跟我说女孩需要学怎么做饭和洗衣帮她的家庭。所以我想当男孩好多了。可是我一看木兰的电影，我的想法变了。她是一个厉害的战士，也是一个好女儿。他什么都能做。她是我第一的榜样。这个视频会介绍我为什么爱木兰的故事。
https://www.youtube.com/watch?v=xrwU5KEj6-Y 1

YOUTUBE Sep 28, 2018



小孩子对性别的看法

[redacted] 04/14/2020 EDT

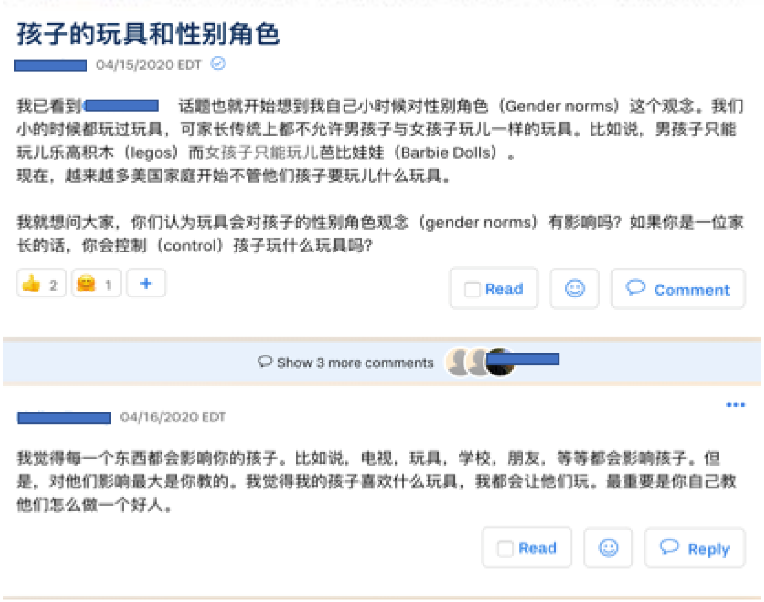
我最近看了这个视频。在介绍一些不同的国家的小孩子。他们都描述他们为什么喜欢当男孩或女孩。我觉得这个视频很意思：有的孩子的回答很好笑，有的很深刻，等等。你们觉得这些孩子的回答为什么都这么不同?

https://www.youtube.com/watch?v=2B3ea7lGwLA 1

YOUTUBE Dec 17, 2016

[redacted] 04/15/2020 EDT

In reply to Kenneth Wang's comment: *"我跟你完全同意！我觉得小时候的经验对人...*



孩子的玩具和性别角色

[redacted] 04/15/2020 EDT

我已看到[redacted] 话题也就开始想到我自己小时候对性别角色（Gender norms）这个观念。我们小的时候都玩过玩具，可家长传统上都不允许男孩子与女孩子玩儿一样的玩具。比如说，男孩子只能玩儿乐高积木（legos）而女孩子只能玩儿芭比娃娃（Barbie Dolls）。
現在，越来越多美国家庭开始不管他们孩子要玩儿什么玩具。

我就想问大家，你们认为玩具会对孩子的性别角色观念（gender norms）有影响吗？如果你是一位家长的话，你会控制（control）孩子玩什么玩具吗?

[redacted] 04/16/2020 EDT

我觉得每一个东西都会影响你的孩子。比如说，电视，玩具，学校，朋友，等等都会影响孩子。但是，对他们影响最大是你教的。我觉得我的孩子喜欢什么玩具，我都会让他们玩。最重要是你自己教他们怎么做一个好人。

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