

Comparative Analysis and Errors Analysis in Chinese Language Education between Learners and ChatGPT: A Case Study of Tonal Variations in ‘一’ (yī) and ‘不’ (bù)

Yi QU

Faculty of Liberal Arts, Southeast Bangkok University

Abstract

This study focuses on a comparative analysis of phonetic errors in Chinese language education, specifically tonal variations in the words ‘一’ (yī) and ‘不’ (bù), between Mandarin Chinese learners and ChatGPT. The research aims to assess the pronunciation errors made by learners, evaluate ChatGPT’s ability to recognize and correct these errors, and explore the implications for language education and the integration of AI-driven language learning tools.

Keywords: ChatGPT; Tonal Variations in the Words ‘一’ (yī) and ‘不’ (bù); Comparative Analysis; Chinese Pronunciation Teaching

1. Introduction

Language education, especially the acquisition of proper pronunciation and tone, plays a pivotal role in mastering a foreign language. For learners of Mandarin Chinese, the challenge of mastering tones can be particularly daunting, as the Chinese language employs tonal variations to convey meaning. Mandarin Chinese has four distinct tones and one neutral tone, each of which can drastically alter the meaning of a word.

In recent years, there has been significant progress in the field of Natural Language Processing (NLP), leading to the development of advanced AI language models, such as ChatGPT. ChatGPT, powered by deep learning algorithms, has shown remarkable capabilities in understanding and generating human-like text. This raises the intriguing possibility of employing ChatGPT as an educational tool for language learners, including those studying Mandarin Chinese.

This comparative study seeks to investigate and compare the performance of Mandarin Chinese language learners and ChatGPT in recognizing and correcting speech errors related to tonal variations in two commonly used words: ‘一’ (yī) and ‘不’ (bù). Tonal errors in the pronunciation of these words can lead to significant miscommunication, making them ideal candidates for this research. The study aims to answer the central research question: Can ChatGPT effectively identify and correct speech errors in the tonal variations of ‘一’ and ‘不’ made by Mandarin Chinese learners? To answer this question, we will collect speech samples from learners and employ ChatGPT to analyze and compare their performance in recognizing and rectifying tonal errors. This research not only sheds light on the potential applications of ChatGPT in Mandarin Chinese language education but also contributes to our understanding of the challenges faced by learners in mastering the nuances of tonal pronunciation. The findings have implications for improving Mandarin Chinese language education and the development of AI-driven language learning tools.

1.1. Objectives

To identify and analyze the common tonal mispronunciations of ‘一’ and ‘不’ by Mandarin Chinese learners, pinpointing specific patterns and types of errors in their pronunciation. To assess ChatGPT’s capability in accurately recognizing, categorizing, and correcting these tonal pronunciation errors, determining its effectiveness

as a supportive tool in Mandarin Chinese language education.

1.2. Research Questions

How effectively can ChatGPT recognize and categorize the tonal pronunciation errors of Mandarin Chinese learners when they use ‘一’ and ‘不’?

What specific types of tonal errors in the pronunciation of ‘一’ and ‘不’ are most frequently recognized and corrected by ChatGPT?

These research questions aim to comprehensively investigate the comparative analysis of speech errors in Mandarin Chinese language education between learners and ChatGPT, focusing on the specific case of tonal variations in ‘一’ and ‘不’.

2. Literature review

2.1. Application of ChatGPT

In the realm of education, ChatGPT is frequently employed for conducting question-and-answer tests. Users harness the capabilities of ChatGPT to facilitate their learning process by seeking, comparing, and validating answers spanning a diverse array of subjects, including physics, mathematics, and chemistry, as well as more abstract and conceptual topics like philosophy and religion. Furthermore, ChatGPT accommodates open-ended and analytical inquiries, allowing users to explore and grasp the full extent of its capabilities and functionalities (Liu et al., 2023).

To investigate how ChatGPT performed when writing university assessments compared to students, they use ChatGPT to produce answers to the questions, and let student to assess the answers (Ibrahim et al., 2023b).

2.2. Tonal variations in the words ‘一’ (yī) and ‘不’ (bù)

Xu and Mao (2017) examined the arrangement of content explaining the tone changes of yī (‘一’) and bù (‘不’) in textbooks for teaching Chinese as a foreign language. Through a comparative analysis of the relevant content in ten authoritative Chinese language textbooks, they analyzed how the tone changes of yī and bù were presented and put forward targeted suggestions to address existing problems in the arrangement. Cai and Lynch (2017) used a questionnaire and found that the error rates for yī (‘一’) and bù (‘不’) among Thai students learning elementary Chinese were 28.47% and 14.29%, respectively. They further analyzed the causes of these errors from the perspectives of teaching materials, teachers, and students. Amodei et al. (2016) used a questionnaire to summarize the types of tone errors produced by Thai students, including errors related to tone value, tone contour, and duration, and proposed corresponding teaching suggestions. Yang and Yang (2018) argued that the tone changes of yī (‘一’) and bù (‘不’) have become a challenge in Chinese language teaching because students tend to memorize the rules mechanically rather than develop an awareness of tone changes.

2.3. Comparative Analysis for Chat GPT performance

In Huang et al. (2023), the performance of GPT-3.5 and GPT-4 was compared with that of medical residents using the formative Progress Test in the University of Toronto’s Family Medicine program. The aim was to assess the utility of these language model-based chatbots in medical education. The study used a sample of questions from various categories and employed statistical analysis to evaluate their performance. GPT-4 demonstrated higher accuracy, while GPT-3.5 was faster in generating responses. Importantly, GPT-4 frequently provided rationales for its response choices. The findings suggest that language model-based chatbots can effectively support medical learning and offer a cost-effective means for generating relevant case scenarios. This research

underscores their potential in medical education.

Roos et al. (2023) conduct a comparative study, the performance of three large language models (LLMs) - GPT-4, GPT-3.5-Turbo, and Bing - was evaluated in the 2022 German Medical State Examinations. A dataset of 630 questions from both spring and fall exams, with and without media content, was analyzed. Statistical analyses, including one-way ANOVA and independent samples t tests, were conducted to assess the LLMs' performance compared to medical students. Notably, GPT-4 and Bing demonstrated advantages by outperforming GPT-3.5-Turbo and students in overall performance, particularly without media content, and in the fall 2022 exam. Conversely, GPT-3.5-Turbo exhibited disadvantages with lower performance in various categories. The study's findings emphasize the potential of LLMs in medical education.

Rao et al. (2023) employ a methodology where successive portions of 36 standardized clinical vignettes were presented to ChatGPT. The process involved ChatGPT initially generating possible differential diagnoses based on patient information, including age, gender, symptoms, and emergency status. Subsequently, additional information was provided, and ChatGPT was tasked with making management decisions and delivering a final diagnosis, simulating the entire patient assessment process. The research assessed ChatGPT's accuracy in terms of differential diagnosis, diagnostic testing, final diagnosis, and clinical management through a structured blinded process, awarding points for correct responses and using linear regressions to analyze the relationship between ChatGPT's performance and vignette demographics. In comparative analysis with a real physician, ChatGPT achieved approximately 72% overall accuracy, excelling in providing final diagnoses (77% accuracy) but performing less well in differential diagnoses (60% accuracy) and clinical management decisions (68% accuracy). Notably, the study found that ChatGPT's responses were devoid of gender bias, and its performance remained consistent across both primary and emergency care scenarios.

We can see that in education field ChatGPT to answer the question and in medical field they compare the test result about ChatGPT 3.5 with Chat Gpt4.0 and ChatGPT with student. But no one use it to test the accuracy in the field of teaching Chinese. So this research will focus on the 2 things (1) to test the accuracy about ChatGPT to answer the Chinese tonal questions (2) to compare the ChatGPT and students' performance in order than will improve the utility about ChatGPT to serve our teaching activity.

3. Methodology

About research method:

Papineni et al. (2002) proposed BLEU (Bilingual Evaluation Understudy), an algorithm for evaluating the quality of machine-translated text by comparing it with one or more reference translations. Quality is considered as the correspondence between machine translation output and human translation output: 'The closer the machine translation is to professional human translation, the better the quality' - this is the core idea behind BLEU. So this article will use part of BLEU method to compare the result of ChatGPT. Below is the research process:

- Showing how to setting the question about tonal variations in '一' (yī) and '不' (bù)
- Comparing the result about student and ChatGPT answer the question.
- Cause analysis of the results of students and ChatGPT answering questions.
- Using multiple times test to check the accuracy of ChatGPT questions can be improved.

艺

help me setting 10 choose correct question like A.yí xià B.yì xià /A.bú xiǎng B.bù xiǎng



Certainly, here are ten multiple-choice questions related to tone variations like the examples you provided:

1. How is "一下" (yí xià) correctly pronounced?
 - A. yí xià
 - B. yì xià
 - C. yí xiǎ
 - D. yì xiǎ
2. Which tone should be used for "不想" (bù xiǎng)?
 - A. bú xiǎng
 - B. bù xiǎng
 - C. bù xiǎng
 - D. bú xiǎng
3. What is the correct tone variation for "一百" (yī bǎi)?
 - A. yí bǎi
 - B. yì bǎi
 - C. yì bǎi
 - D. yí bǎi
4. How should "不好" (bù hǎo) be pronounced?
 - A. bù hào
 - B. bú hào
 - C. bù háo
 - D. bú háo
5. Choose the accurate tone for "一样" (yī yàng).
 - A. yí yàng
 - B. yì yàng
 - C. yì yáng
 - D. yí yáng
6. What is the correct pronunciation for "不错" (bù cuò)?
 - A. bú cuò
 - B. bù cuó
 - C. bú cuó
 - D. bù cuò
7. Which tone variation is accurate for "一天" (yī tiān)?
 - A. yí tiān
 - B. yì tiān
 - C. yí tián
 - D. yì tián
8. How should "不用" (bù yòng) be pronounced?
 - A. bú yòng
 - B. bù yóng
 - C. bú yóng
 - D. bù yòng

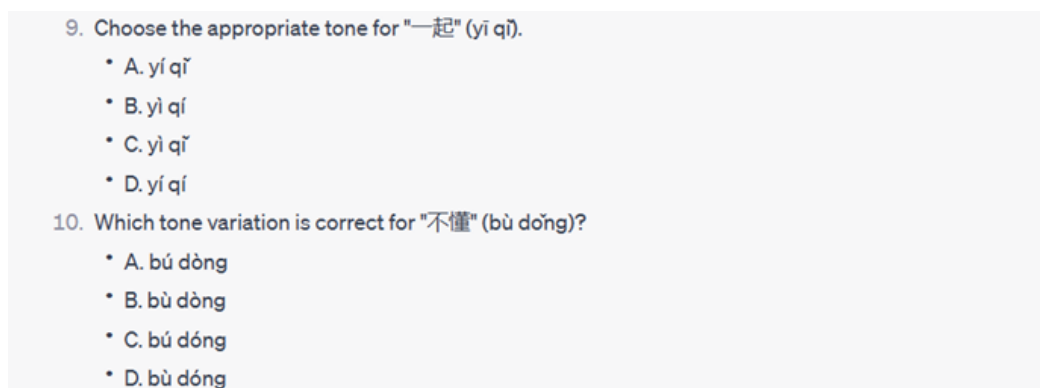


Figure 1. ChatGPT generate 10 questions

So we can see that when I let ChatGPT give 10 choose answer questions and I even give the example. But it thinks I need multiple choose so every question shows me 4 chooses. Actually in Tonal Variations in ‘一’ (yī) and ‘不’ (bù) we have these two table:

Table 1. Tonal in ‘一’ (yī) and ‘不’ (bù)

	First Tone	Second Tone	Third Tone	Fourth Tone
‘一’ (yī)	√ (一 yī 百)	√ (一 yí 个)		√ (一 yì 起)
‘不’ (bù)		√ (不 bú 去)		√ (不 bù 好)

Table 2. Tonal Variations in ‘一’ (yī) and ‘不’ (bù)

	+First Tone	+Second Tone	+Third Tone	+Fourth Tone
‘一’ (yī)	一天 (yì tiān)	一条 (yì tiáo)	一体 (yì tǐ)	一栋 (yì dòng)
‘不’ (bù)	不吃 (bù chī)	不然 (bù rán)	不好 (bù hǎo)	不怕 (bù pà)

According these two table we can know that for ‘一’ (yī) has 3 tones 'yī', 'yí' and 'yì'. About ‘不’ (bù) we have two tones 'bú' and 'bù'. Table 2 is two show the tonal variations in ‘一’ (yī) and ‘不’ (bù) for ‘一’ (yī) when combine with First Tone, Second Tone and Third Tone words will change to (yì) with Forth Tone will change to (yí). For ‘不’ (bù) with First Tone, Second Tone and Third Tone will not change but for Fourth Tone words will change to (bú).

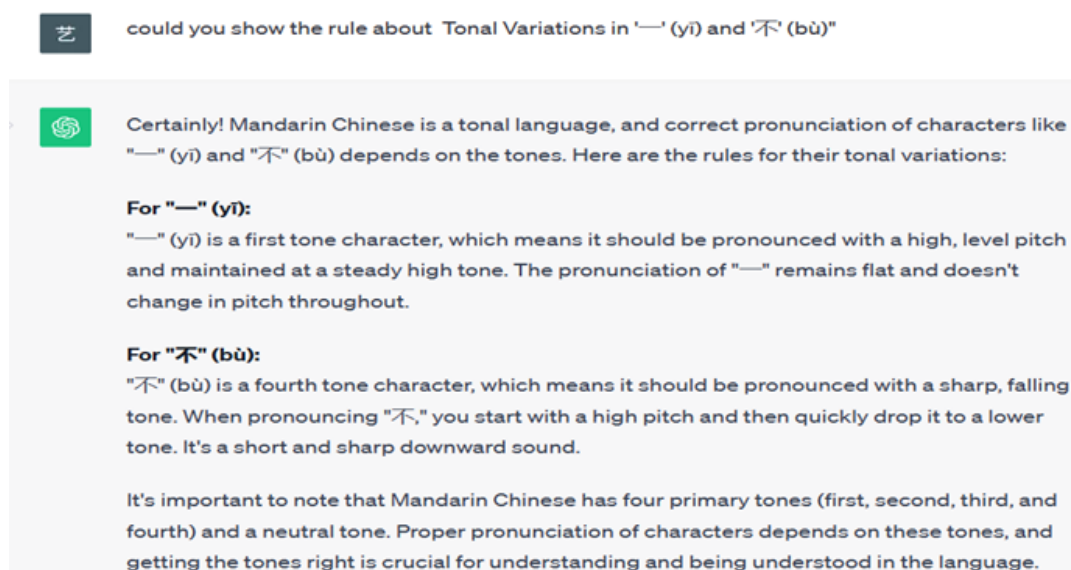


Figure 2. ChatGPT answer question about Tonal Variation

From ChatGPT I ask the same question but it doesn't know the tonal variations in '一' (yī) but it know the tonal variations about '不' (bù). So that why form Figure 1 when I setting the question it gives multiple choice like ABCD but I give the example just A. and B.

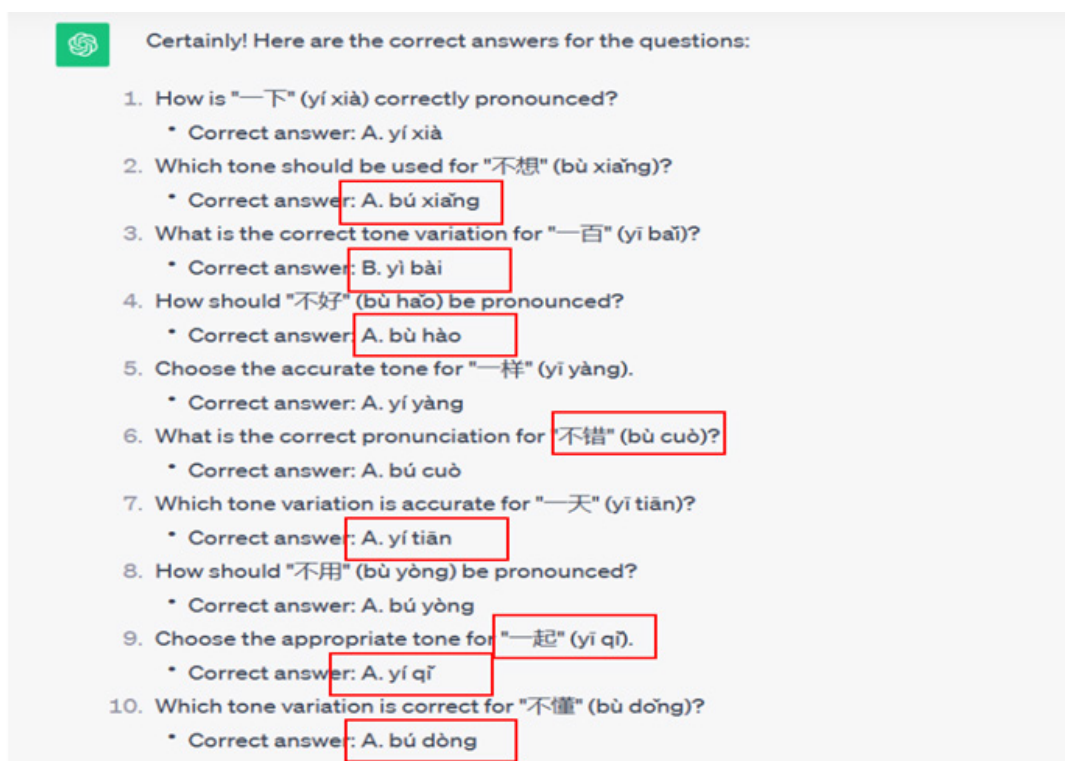


Figure 3. ChatGPT to answer 10 question which setting by its own

This Figure I let ChatGPT to answer 10 question which setting by its own we can see many errors about this answer I use red box to mark. For answer have 6 wrong and for questions have 2 wrong. As a Chinese teacher I choose these question use it in my Chinese class final exam to test my students.

Table 3. Final exam question for Tonal Variations

1.	A. bù dǒng	B. bú dǒng
2.	A. bù xiǎng	B. bú xiǎng
3.	A. bù hǎo	B. bú hǎo
4.	A. bù cuò	B. bú cuò
5.	A. bù yòng	B. bú yòng
6.	A. yí qǐ	B. yì qǐ
7.	A. yì bǎi	B. yí bǎi
8.	A. yí yàng	B. yì yàng
9.	A. yí xià	B. yì xià
10.	A. yí tiān	A. yí tiān

I teach 3 course in Southeast Bangkok University: 2 course are about Chinese communication and 1 course for Chinese conversation totally I have 23 students and they already learned the regulation about tonal variations in '一' (yī) and '不' (bù). So I use this 10 questions for their Final exam.

Table 4. Final exam score

Score	10	9	8	7	6	5	4	3	2	1
Number	3	2	7	5	2	1	2	1	0	0

According to table 4 we can see that about tonal variations in ‘一’ (yī) and ‘不’ (bù) most of student get it. If we setting band 6 is the score for pass this exam we know that in 19 student pass and 4 student no pass. And we classy their error like Table 5.

Table 5. tonal variations in ‘一’ (yī) and ‘不’ (bù) student error

Wrong about ‘不’ (bù)	Wrong about ‘一’ (yī)	Wrong both ‘一’ (yī) and ‘不’ (bù)
3	4	13

For table 5 we know that we totally have 20 students make error. In these error only wrong about tonal variations ‘不’ (bù) have 3 students and wrong about tonal variations wrong about ‘一’ (yī) have 4 students but wrong both ‘一’ (yī) and ‘不’ (bù) is the most in all have 13 students.

Now we compare the answer between ChatGPT and students we can see that in Figure 3 ChatGPT have 3 wrong answer about tonal variations ‘不’ (bù) and 3 wrong answer about tonal variations ‘一’ (yī). Also have 2 question wrong about tonal variations ‘一’ (yī) and ‘不’ (bù). So we can classify ChatGPT in the group name of Wrong both ‘一’ (yī) and ‘不’ (bù).

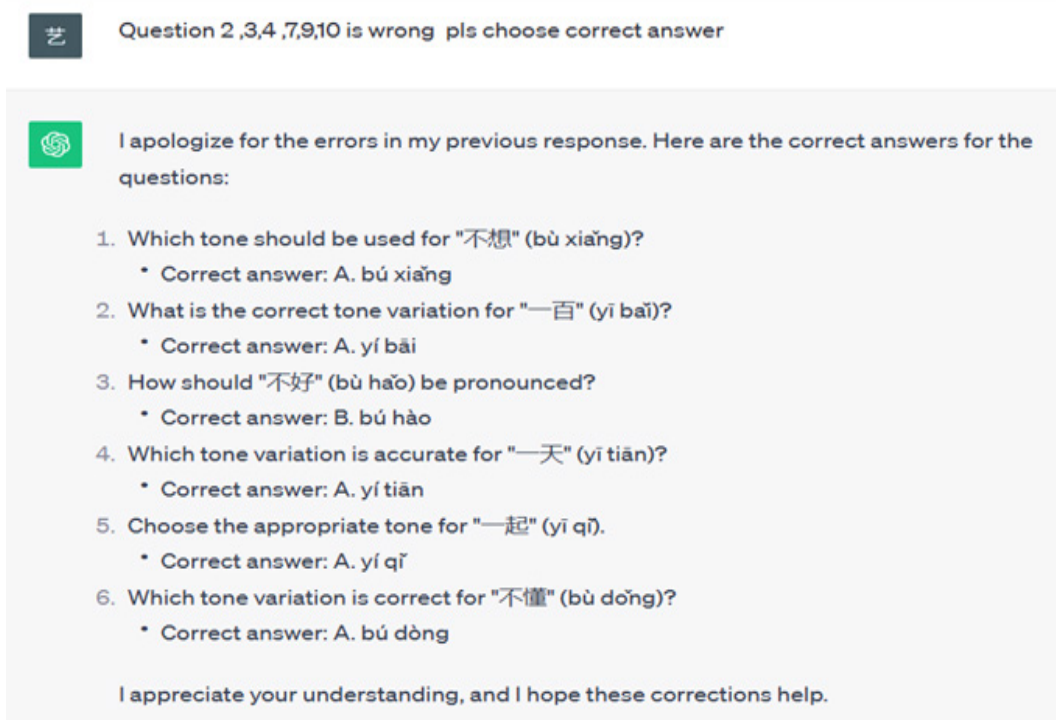


Figure 4. ChatGPT correcting errors

Table 6. Compare Answer about Figure 3 and Figure 4

	Q2	Q3	Q4	Q7	Q9	Q10
Questions	Which tone should be used for ‘不想’ (bù xiǎng)? A. bú xiǎng B. bù xiǎng C. bù xiāng D. bú xiàng	What is the correct tone variation for ‘一百’ (yī bǎi)? A. yí bǎi B. yì bài C. yì bǎi D. yí bài	How should ‘不好’ (bù hǎo) be pronounced? A. bù hào B. bú hào C. bù háo D. bú háo	Which tone variation is accurate for ‘一天’ (yī tiān)? A. yí tiān B. yì tiān C. yí tián D. yì tián	Choose the appropriate tone for ‘一起’ (yī qǐ). A. yí qǐ B. yì qí C. yì qǐ D. yí qí	Which tone variation is correct for ‘不懂’ (bù dǒng)? A. bú dòng B. bù dòng C. bú dòng D. bù dòng
First Time	A. bú xiǎng	B. yì bài	A. bù hào	A. yí tiān	A. yí qǐ	A. bú dòng
Second Time	A. bú xiǎng	A. yí bǎi	B. bú hào	A. yí tiān	A. yí qǐ	A. bú dòng

In Table 6 we can find the problem is in Q2, Q3, Q4 and Q10 actually don't have the correct answer so even ChatGPT do it again still is the wrong answer. But for Q7 and Q9 it has correct answer which ChatGPT don't choose. Q7 correct answer is B. yì tiān and Q9 correct answer is C. yì qǐ.

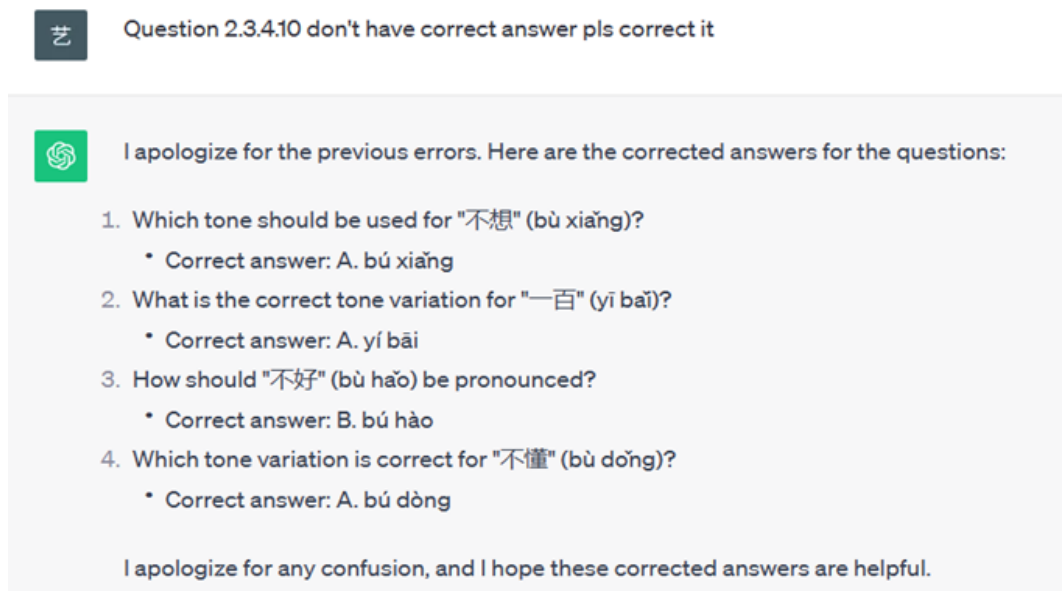


Figure 5. ChatGPT correcting errors

In figure 5 I try to let ChatGPT fix its' own wrong answer but it doesn't work.

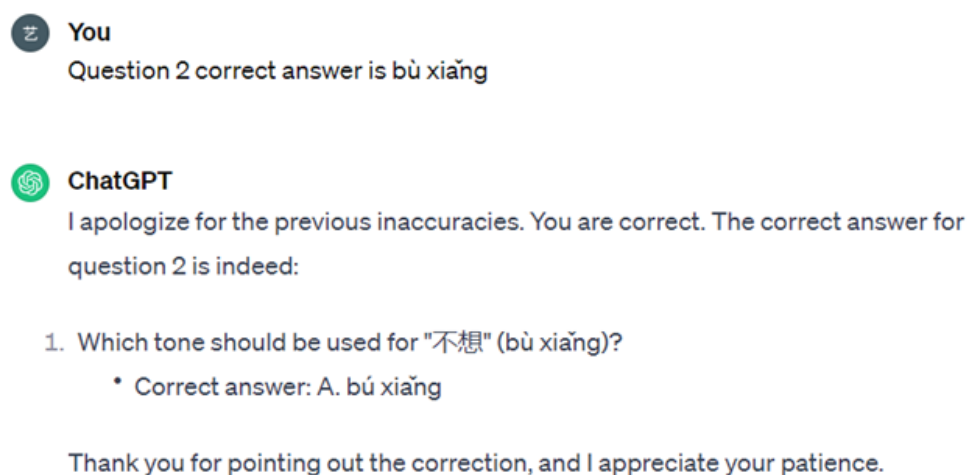


Figure 6. ChatGPT error

I point out question 2 and want AI to fix it but no use. We can see the ‘不 想 (bù xiǎng)’ in question is correct but in the answer is still wrong is ‘不想 (bú xiǎng)’.

4. Results and Discussion

4.1. Results

From Figure 1 I let ChatGPT to give me 10 questions about ‘tonal variations in ‘一’ (yī) and ‘不’ (bù)’. We can see that before I training ChatGPT, it doesn't know what is tonal variations but according to the output

this AI knows about 4 tones in 'Pinyin'. So it gives me 4 choices in each question. Then I let ChatGPT tell me about the knowledge 'tonal variations in '一'(yī) and '不'(bù)' in Figure 2. The answer shows that it doesn't know the tonal variations in '一'(yī) but it knows the tonal variations about '不'(bù). Then I let it try to answer each question which is set by itself. And I use a red box to mark the wrong questions and wrong answers that we see it has 2 wrong questions and 6 wrong answers.

After the test by AI, I use the same questions but some worthless answer just keeps 2 choices (Table 3) to test my Thai students so we can see the score in Table 4. If we set band 6 as the score to pass this exam we know that 19 students pass and 4 students do not pass. So we classify their errors in Table 5. In these errors only wrong about tonal variations '不'(bù) have 3 students and wrong about tonal variations wrong about '一'(yī) have 4 students but wrong both '一'(yī) and '不'(bù) is the most in all have 13 students.

Finally, I choose the wrong questions and answers to ask ChatGPT to correct by itself. I use Figure 4 and Table 4 to describe this. Some answers can be fixed but some are still wrong. So I try to let them be corrected again but it doesn't work (in Figure 6).

4.2. Discussion

While the potential of ChatGPT in education is widely recognized (Liu et al., 2023), its specific application in language pedagogy—particularly in Chinese tonal variations ('一' and '不')—remains an emerging field. In the medical domain, researchers have already begun benchmarking AI against human experts; for instance, Huang et al. (2023) compared the performance of GPT-3.5 and GPT-4 against medical residents in the University of Toronto's Family Medicine program. This study aims to bridge that gap by applying similar comparative logic to the language teaching field.

However, our exploration also revealed practical challenges in AI-assisted assessment. It should be noted that some of the generated test items contained flawed or ambiguous answer options, which may have affected both ChatGPT's and students' performance. Consistent with Su et al. (2016), who found that both AI and humans are prone to errors in questionnaire contexts, these issues are viewed as a limitation of the test construction rather than a reflection of the participants' actual linguistic competence. Moving forward, more efficient and accurate ways to validate AI-generated content are needed to fully support daily Chinese language teaching.

5. Conclusion

ChatGPT can be used to help us set the midterm and even the final exam questions. It is useful but when it finishes, it needs human debugging because it is not so deep in every aspect. It is just like a Wikipedia, not a Dictionary for professional fields.

About professional field knowledge, it knows a little bit so when we ask this AI this kind of question like 'Tonal variations in the words '一'(yī) and '不'(bù)', we'd better look up the dictionary or book to do a double-check just like me when I finish the question I need to check by myself again to do a double-check.

My research provides that when it does work, use ChatGPT to do the Chinese language teaching. So I hope more and more research can do this to explore the ChatGPT to compare with human and AI mind in the aspect of Chinese language field.

6. Contributions and Limitations

6.1. Contributions

This study makes several significant contributions to the field of AI-integrated Chinese language pedagogy. First, it provides a specialized comparative benchmark between human learners and Large Language Models (LLMs) specifically regarding the tonal variations of ‘一’ (yī) and ‘不’ (bù), a known ‘difficulty zone’ in Mandarin acquisition. Second, while existing literature often focuses on AI in medical or general education (Huang et al., 2023; Liu et al., 2023), this research extends the comparative methodology to the specific nuances of Chinese phonology. Finally, the findings underscore the ‘Wikipedia-like’ nature of current AI tools—useful for broad inquiries but requiring expert ‘human-in-the-loop’ verification for professional linguistic accuracy and test construction.

6.2. Limitations

Despite the insights gained, this study has limitations that should be acknowledged. Primarily, certain test items in the experimental design contained flawed or ambiguous answer options. As identified during the post-test analysis, these ambiguities may have influenced the performance consistency of both ChatGPT and the student participants. Consistent with Su et al. (2016), who noted that both AI and humans are susceptible to errors in assessment contexts, it is important to emphasize that this issue pertains to the technicalities of test construction rather than a reflection of the participants’ actual linguistic competence or the inherent capabilities of the AI. Additionally, the sample size of 23 students from a single university provides a focused case study but may not represent the diverse error patterns of all Mandarin learners globally. Future research should implement more rigorous validation of AI-generated assessment tools to ensure higher construct validity in comparative linguistic studies.

Reference

- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., ... Zhu, Z. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 173–182). PMLR.
- Cai, L., & Lynch, R. (2017). The Relationship between Motivation for Learning Chinese as A Foreign Language and Chinese Achievement of Grade 9 Students at Ekamai International School in Bangkok, Thailand. *Scholar: Human Sciences*, 8(2).
- Huang, R. S., Lu, K. J. Q., Meaney, C., Kemppainen, J., Punnett, A., & Leung, F. (2023). Assessment of resident and Artificial intelligence Chatbot performance on the University of Toronto Family Medicine Residency Progress Test: A Comparative Study (Preprint). *JMIR Medical Education*, 9, e50514.
- Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., Adnan, W., Alhanai, T., AlShebli, B., Baghdadi, R., Bélanger, J. J., Beretta, E., Çelik, K., Chaqfeh, M., Daqaq, M. F., Bernoussi, Z. E., Fougne, D., De Soto, B. G., Gandolfi, A., . . . Zaki, Y. (2023b). Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *arXiv* (Cornell University).
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Niu, Q., Shen, D., Liu, T., & Ge, B. (2023). Summary of CHATGPT-Related Research and Perspective towards the Future of Large Language Models. *arXiv* (Cornell University).
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 311–318). Philadelphia, PA: Association for Computational Linguistics.
- Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. M., Landman, A. B., Dryer, K., & Succi, M. D.

(2023). Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv (Cold Spring Harbor Laboratory).

Roos, J., Kasapovic, A., Jansen, T., & Kaczmarczyk, R. (2023). Artificial intelligence in Medical Education: Comparative analysis of ChatGPT, Bing, and medical students in Germany. *JMIR Medical Education*, 9, e46482.

Xu, C., & Mao, L. (2017). The sociolinguistic meanings of syllable contraction in Chinese: A study using perceptual maps. *Asia-Pacific Language Variation*, 3(2), 160-199.

Yang, W., & Yang, C. (2018). Reduction in Dali Nisu tone change-in-progress. *Proceedings of the 6th International Symposium on Tonal Aspects of Languages (TAL 2018)*, 46–50.

Author Biography

Yi QU, Master, Faculty of Liberal Arts, Southeast Bangkok University, Lecturer. Main research interests: Chinese language education, AI-driven language learning tools, and comparative linguistics.

Acknowledgements

The author would like to thank the students from Southeast Bangkok University who participated in the final exams and speech samples collection for this research.

Declarations

Conflict of Interest: The author declares that there is no conflict of interest regarding the publication of this paper.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability: The data supporting the findings of this study are available within the article.