# Teaching Methods of Large Language Model on L2 Text Conversation: The Role of AI Situational Dialogue Applications

Jiamin Wu

College of Foreign Languages, Capital Normal University, Beijing, 100089, China

1232071014@cnu.edu.cn

**Abstract.** Situational dialogue driven by large language models (LLMs) has received wide attention in second language (L2) teaching. However, most of the existing research focuses on a single model. This article analyzes the application of artificial intelligence-driven situational dialogue in L2 conversation teaching. This study adopts a quasi-experimental design and uses five artificial intelligence models to explore the differences between situational dialogue and traditional teaching practices, the performance differences between different LLMs, and how teachers can use different artificial intelligence systems in teaching. Analysis shows that these models differ in emotional expression, human-like degree, and information integration. Learners' preferences are influenced by factors such as the tone of AI, information integrity, and feedback form. Based on these findings, this article puts forward the following suggestions: the use of artificial intelligence in the classroom should be dominated by teachers, artificial intelligence as an auxiliary tool, and the appropriate model should be selected according to the teaching objectives. At the same time, educators should pay close attention to the structure and emotional design of AI feedback, and strengthen ethical and privacy supervision to ensure the applicability and safety of AI in the field of education.

**Keywords:** Large Language Model; Situational Dialogue; AI-Assisted Teaching; Second Language Learning.

## 1. Introduction

In the context of rapid globalization, learning a second language is crucial for improving personal chances and promoting cross-cultural communication. However, a lot of L2 learners often lack experience and have significant nervousness when speaking, which further impedes their ability to communicate. Therefore, it is very important for educators to explore how to use appropriate skills to improve the communication skills of second language learners. In recent years, the emergence of AI and LLMs has provided new ideas for solving this problem. By training a large amount of language data, LLM provides a stable and continuous text dialogue platform for second language learners, and can provide real-time and personalized feedback on learners' answers, helping them practice and correct mistakes in time [1]. In addition, the high flexibility of LLM enables teachers to customize exercise materials for each student, thus greatly reducing the teaching burden of teachers and improving work efficiency. The data shows that the timely feedback and practice opportunities provided by AI chatbots can significantly improve the language output performance of learners [2]. Other studies have also found that training with AI dialogue systems in the classroom can help students maintain continuous oral practice and gain a positive learning experience [3]. On the technical side, LLM-based situational dialogue models have combined open-ended conversation with specific situational tasks and show good generalization to unseen topics, allowing learners to practice language on a wider range of topics [4]. However, AI technology also has inherent limitations such as data bias, system instability, and potentially biased content. These negative factors pose challenges to the effective integration and broader adoption of AI in real classroom teaching. Therefore, based on the above background, this study aims to compare the effects of different mainstream large language models on L2 learners' situational dialogue practice under the same tasks and training mechanisms. The research seeks to examine the role of LLM-driven AI situational dialogue in improving learners' dialogic skills, analyze causes of performance differences among models, and draw implications from experimental data for L2 teaching practice, so as to provide references for teachers using AI-assisted dialogue instruction.

# 2. Method

## 2.1. AI Applications in Teaching

The rapid development of AI and LLM technologies has greatly advanced their application depth and feasibility in education. The fast iteration of LLMs has noticeably improved their computational efficiency and generalization ability, enabling them to more precisely analyze learners' still-developing language expressions and to generate customized learning tasks that match individual cognitive levels [5]. At the same time, their immediate feedback and positive reinforcement mechanisms help strengthen students' confidence and learning motivation, and can effectively make up for the limits of traditional classroom teaching when teacher resources are scarce [6]. AI writing tools have developed from simple grammar and spelling checks to comprehensive writing assistance based on LLM, which are used in writing teaching. By interacting with chatbots, learners can confirm terms, improve their understanding of meanings, and receive personalized grammatical error correction in a real language environment, thus promoting a deeper understanding and practical application of vocabulary and grammar [7]. Educational feedback research shows that AI can provide comprehensive, logical, and process-oriented evaluation. A considerable number of students are more inclined to use AI, and their preferences for manual input and AI input are also similar [8]. Moreover, the rational nature of AI decision-making helps reduce human cognitive biases and improve the objectivity and fairness of assessment [9,10]. Research on AI-assisted conversation instruction demonstrates that students who use chatbots in both official and casual settings not only increase their language proficiency but also develop a deeper comprehension of cross-cultural communicative norms, which improves their capacity for cultural adaptation [11]. In addition, Lee et al. demonstrated on the Poe AI platform the feasibility of building task-oriented dialogue agents, offering researchers a convenient and effective tool to quickly construct and test LLM chatbots [12].

In summary, the wide application of artificial intelligence in education offers educators an opportunity to use its technical features for dialogue teaching. However, existing educational studies mainly focus on traditional auxiliary teaching functions and assignment assessment, such as grammar and vocabulary correction, real-time feedback mechanisms, and automatic exercise generation, while discussion of dialogue-based application modes is insufficient. Second, the current literature in the field of dialogue teaching research frequently analyzes difficulties from the perspectives of language theory or AI program design, and it has not completely examined the deeper educational effectiveness of AI dialogue teaching. In addition, most studies take ChatGPT as the sole research subject, and the educational value of other AI models in classroom scenarios still lacks systematic examination. Therefore, this study focuses on the mechanisms by which AI affects situational text-based dialogue teaching for L2 learners, and raises the following research questions:

RQ1: What is the impact of large language model-driven contextual dialogue on the conversational outcomes of second language learners?

RQ2: What differences exist among various large language models in enhancing learners' situational text-based dialogue?

RQ3: How do L2 learners perform when practicing dialogue with different AI dialogue systems, and which factors are most influential on their performance?

## 2.2. Research Design

### 2.2.1 Purpose

The goal of the current experiment is to implement task-based text-dialogue training and assess its usefulness in enhancing learners' conversational English skills. A similar preliminary study was carried out by Xu, Qin, Chen, et al., who suggested an efficient situational dialogue model. By optimizing the existing AI, the model effectively combines task-based oriented training with open dialogue, thus encouraging students to practice conversation. Their research results provide strong evidence for the feasibility and benefits of using existing LLMs in conversational learning. Based on these results, this experiment made targeted adjustments to the situational dialogue text proposed by

Xu et al. and the system prompts in the framework, thus ensuring the effectiveness of AI-driven dialogue training in the educational environment. In addition, Chang, Wang, Wang, et al. summarized the multi-dimensional evaluation indicators of LLM based on existing research in their systematic review, including Natural Language Processing Tasks; Robustness; Ethics, Bias, and Trustworthiness; Social Science; Natural Science and Engineering; Medical Applications; Agent Applications; and Other Applications [10]. This article offers accurate and practical evaluation criteria, which provide a solid theoretical foundation for the quantitative evaluation system used in this study. Under this standard framework, the selection of experimental evaluation indicators is closely compatible with the teaching method and the characteristics of L2 teaching. Through systematic screening and adjustment, the scoring criteria of this experiment were determined.

### 2.2.2 Design

This research adopts a quasi-experimental design and combines multi-platform comparative hybrid methods. By selecting different AI platforms, this study has built an intelligent dialogue environment that supports real-time interaction and feedback, simulating multi-round text dialogue scenarios in message reply mode. The study collected the conversation records of each participant, the quantitative data of AI feedback, and the participants' subjective evaluation of the AI platform under unified situational tasks and evaluation standards. The research aims to systematically compare the differences between different platforms in terms of interaction quality and teaching effect.

Five learning platforms were used in this experiment: ChatGPT 5, Claude Sonnet 4.5, Gemini 2.5 Flash, DeepSeek V3, and Doubao 9.0.0. In order to improve the performance of AI in scenario construction, multi-round interaction, and overall evaluation, and to ensure that the output results can achieve the expected teaching goals, the experiment sets a similar reasoning depth within the functional limits of each platform. In addition, in order to ensure the unity and comparability of the experimental process, all platforms are loaded with the same prompt template before testing.

### 2.2.3 Procedure

The formal test consists of three main stages and is carried out in a fixed order. The first is the stage of situational dialogue. Participants need to use the daily text message style to have a conversation with AI on a given topic. AI will play the preset role in the prompt and conduct multiple rounds of follow-up questions according to the details of the participants' answers. When AI judges that enough information has been obtained, it will automatically end this stage and enter the next stage. Next is the stage of teacher evaluation and question-and-answer. AI will switch to the role of teachers, provide comprehensive feedback on the performance of participants from multiple dimensions such as vocabulary, grammar, style, and socio-cultural knowledge, and put forward targeted suggestions for improvement. Participants can respond to the feedback and ask follow-up questions. The end of this stage is up to each participant. Finally, there is the stage of subjective scoring and qualitative feedback. Participants use the pre-provided scoring criteria to score the performance of artificial intelligence in the interaction process on a 1-5-point system (1 point for very dissatisfied, 5 points for very satisfied), and provide supplementary qualitative comments, such as the depth of follow-up questions or the practicality of feedback. During the whole experiment, all the dialogue records and the evaluation texts of AI were kept intact.

The evaluation dimension and its definition are as follows. Sentiment analysis: evaluate the appropriateness of emotional expression in the dialogue process and the degree of consideration of emotional factors by AI when providing feedback. Semantic understanding: Evaluate whether AI can fully understand the deep meaning of the dialogue and provide a response in line with the context. Social knowledge understanding: aims to evaluate the ability of AI to properly apply daily common sense and cultural background knowledge in dialogue. Summarization: Evaluate whether the comments or summaries of AI are logically coherent and whether they are of substantial help to learners. Questioning: Evaluate the accuracy and reasonableness of AI in answering participants' questions in the second stage. Factuality: Evaluate the authenticity and accuracy of AI answers.

### 2.2.4 Participants

Seven Chinese college students were recruited to participate in this study. All participants speak English as a second language, and their English level is between B2 and C1 in the Common European Framework of Reference for Languages (CEFR). All participants have not lived or studied in English-speaking countries or regions. In order to ensure the effectiveness of the test, all participants have experience in using AI, and the difficulty of the test materials is set according to their current academic level and existing experience in using AI. Before the formal test, all participants received a detailed explanation of the test process and practiced once or twice.

This study follows the standard ethical procedures, and the participants are completely voluntary. Participants are free to withdraw at any time without any penalty.

## 3. Result

After the experiment, the researcher calculated the average score according to the collected scoring table. All scoring processes are carried out under strict supervision, and the data is considered valid. The summary results are shown in Table 1.

Data analysis shows that Claude has the highest overall score, followed by Gemini, which is better than Claude in terms of social knowledge understanding indicators. ChatGPT's performance is at a moderate level, but it maintains a comparative advantage in terms of questions and answers and factuality. Doubao's overall ranking is low, but all its index scores are higher than the median (3), especially in terms of social knowledge understanding. The score of DeepSeek is below average, and the score of each sub-dimensional is lower than that of the other four AI models.

In the qualitative feedback of participants choosing their favorite AI models, Claude got 4 votes (4/7), Gemini and ChatGPT got 1 vote (1/7) each, while DeepSeek and Doubao did not get any votes.

**Table 1.** Mean ratings of the five AI models across six user-experience dimensions.

| | Sentiment Analysis | Semantic Understanding | Social Knowledge Understanding | Summarization | Question Answering | Factuality |
|---|---|---|---|---|---|---|
| ChatGPT | 3.17 | 3.67 | 3.17 | 3.67 | 3.67 | 3.67 |
| Claude | 4.00 | 4.17 | 3.83 | 4.33 | 4.00 | 4.00 |
| Gemini | 3.33 | 4.17 | 4.17 | 4.00 | 3.33 | 3.83 |
| DeepSeek | 2.37 | 2.33 | 2.50 | 2.50 | 2.33 | 2.83 |
| Doubao | 2.83 | 3.33 | 3.67 | 2.83 | 3.33 | 3.17 |

### 3.1. AI Model Comparison

According to the experimental results, LLM-driven situational dialogue has several advantages, particularly in terms of boosting learner engagement and immersion, increasing practice opportunities and output variety, and offering quick, focused feedback. AI demonstrates a stable ability to perform role-play and can effectively integrate situational materials to enhance the realism of conversations, thereby stimulating learner motivation and promoting the use of communicative strategies. However, this study also identified several factors that limit teaching effectiveness, including some models' non-human tone, follow-up questions that are weakly tied to instructional goals, occasional slow responses from individual models, and instability in output format. These issues can undermine interaction fluency in class and distract learners' attention.

The five models all show significant differences in emotional expression and humanization, information integration ability, language coherence, and response speed. Claude showed obvious advantages in terms of emotional expression and humanization. Its reply usually contains encouraging emotional language, which effectively improves the humanization of the dialogue. In the feedback stage, Claude usually gives a positive overall impression first, and then ends with additional encouragement and a list of points; this structure seems to enhance learners' confidence and motivation to practice. In contrast, although ChatGPT is logically clear, its tone is relatively bland, which sometimes makes the interaction appear programmatic. Doubao lacks emotional connection in

some aspects; for example, it jumps directly from the first stage to the evaluation stage without a summary, resulting in a fault in the learning process. In terms of information integration, Gemini has the most outstanding performance. It greatly improves the contextual closeness of task-based dialogues, such as asking for ways or arranging itineraries. Models that rely less on external information are relatively more suitable for pure language structure training. In terms of language coherence, ChatGPT's response is direct and logically structured, and excels in syntax and structural inspection. In terms of response speed, most AI models respond quickly. However, DeepSeek is the slowest to respond, which obviously limits its application for real-time classroom interaction.

### 3.2. Learner's Engagement and Feedback

The experimental results show that the AI model generally exceeded the initial expectations of the participants in the L2 conversation practice, but the learning results were affected by many factors.

First of all, the emotional presentation directly affects the participation and evaluation tendency of learners. For example, Claude's humanized tone, appropriate encouragement, and infectious connecting phrases significantly enhanced the participants' emotional commitment and self-confidence; therefore, some participants showed a more tolerant evaluation attitude. In contrast, demonstrators with a bland tone or weak speech connection, although the information transmission is accurate, find it difficult to stimulate the motivation to practice. This shows that emotional factors have an amplifying effect on learning participation.

Secondly, the completeness and authenticity of the information have a significant impact on the training effect. Some testers pointed out that Gemini's integration of external information improves the realism of the situation and the completion of the task, which in turn enhances the enthusiasm of learners. At the same time, the two domestic Chinese models enhance the participants' understanding and immersion through local expressions and language in line with the cultural context, thus prompting participants to give more detailed feedback.

Third, response speed and interaction rhythm are key technical factors affecting the classroom experience and acceptance. Due to the long response time, DeepSeek received consistent negative feedback from all participants. This long waiting not only reduces the effect of instant practice but also systematically reduces the final score of the model, which shows that even if the content of the tool is valuable, the response delay will significantly weaken its educational availability.

Finally, individual preferences lead to significant differences in the acceptance of the same model output by different learners. For example, some participants obviously prefer the models they are already familiar with. The diversity of feedback forms has also caused differences of opinion. This shows that the effectiveness of AI assistive tools depends not only on the performance of the model, but also largely on the learners' usage habits and cognitive preferences.

## 4. Discussion

### 4.1. Implications for Model Performances

The study confirms that different models show a certain degree of human-like reasoning ability in the test. However, some participants pointed out that all AI behaviors are essentially imitations of human behaviors and lack real awareness and intention, so they cannot be fully trusted. Ben der, Gebru, McMillan-Major, et al. also believe that AI is essentially piecing together language sequences without coherent communication intentions, so the generated text is not based on real communication purposes; therefore, learners' sense of understanding may be a kind of self-deception [13]. This interactive model shows that teachers must carefully interpret the meaning contained in AI responses, and further explore how to avoid learners falling into learning traps due to adaptive adjustment.

Experiments have observed that the information integration ability of some AI models is specifically reflected in the ability to connect with external applications, which makes them show significant advantages in environmental perception dialogue. However, this behavior has also raised ethical concerns about the boundaries of business content in the field of education. For education,

especially underage learners, full commercialization may bring advertising implantation, data security risks, and potential addiction problems. Therefore, it is urgent to clarify the boundaries between educational AI and fully functional commercial versions, and to carry out more in-depth research to define the commercial restrictions and regulatory standards of educational AI. Existing regulations in the fields of games and video platforms may provide useful references for the protection of minors and the control of commercial content.

Rapid response is one of the core advantages of the AI model. In educational applications, the emphasis should be placed on maintaining a quick response, and instant feedback should be taken as a key performance indicator. In addition, when conducting educational supervision and effect assessment, technical factors that interfere with interaction should be minimized to ensure the smooth operation of the feedback process.

### 4.2. Implications for L2 Teaching Practice

The research results of this experiment are of great practical significance to teachers in the field of L2 education.

First of all, in view of the fact that the application of AI in the field of education has become a foregone conclusion and has practical value, teachers should actively integrate AI into teaching practice. Teachers need to learn how to use AI for teaching and think creatively about how to integrate AI into curriculum design. For example, in exercises focusing on emotional situational dialogue, teachers can choose an LLM with richer emotional expression, and in tasks that require clear logical reasoning, they can choose a more objective platform. The selection and management of AI models based on language learning models should be accurately in place to maximize their teaching benefits and promote effective L2 communication and learning among students.

Secondly, teachers play a crucial and leading role in AI-assisted L2 teaching. Teachers must be responsible for determining the learning content, supervising the communication process, and adjusting the parameters of AI dialogue to avoid deviation from the learning goal when AI dominates the interaction. In addition, educators have the responsibility to remind students of social prejudices and technical risks that may exist in the application of AI from an early age [14]. The fundamental goal of AI-assisted L2 teaching should be human-computer complementarity: AI provides instant initial feedback on basic language elements such as vocabulary and grammar, while educators verify and correct this feedback, and add creative or more in-depth insights to jointly promote learners in the continuous progress in language ability.

Third, teachers should pay special attention to the emotional output of AI, because emotional factors are crucial to language development. In order to ensure a smooth learning process, teachers must identify and correct any negative emotions in the classroom in time and adjust the emotional expression and task difficulty of AI according to the unique characteristics and needs of each learner.

Crucially, AI raises a series of ethical issues, including privacy and data protection, language bias and stereotypes, and the risk of learners' over-reliance on AI systems [15]. Teachers must be extra careful when introducing AI, because many students are still minors, which is a critical period for ethical development. Teachers must clearly point out that human beings play a guiding and role model role in ethical education, and ensure that AI will not spread wrong ethical concepts to students.

Finally, the study observed a fluctuating phenomenon of LLM output, that is, AI will adjust its language to adapt to the conversational context, thus affecting the subsequent performance of the model and the reaction of learners. Some studies have explored this problem; for example, Xie, Wang, Feng, et al. proposed to maintain the accuracy of model judgment by synthesizing high-quality preference data [16]. However, there could still be some uncertainty in the model output, and it may not fully meet the teaching goals of teachers. Therefore, teachers must carry out real-time supervision and guide students on how to critically evaluate and deal with unstable or misleading AI-generated information.

## 5. Conclusion

This study confirms the significant advantages of AI-based situational dialogue models in L2 teaching, reveals the performance differences between different language learning model platforms, and the far-reaching impact of these differences on teaching practice. Research results show that compared with traditional language practice methods, AI-driven situational dialogue can significantly improve the participation of learners. The AI dialogue system provides real-time and personalized feedback and rich context simulation, making the learning process more immersive and autonomous.

The research results further confirm that different AI models have significant differences in feedback style, language interaction, and human-likeness, and each model emphasizes different tone and structural characteristics in its output. From the perspective of learners, feedback tone, interaction fluency, and personalization are key factors affecting the learning effect and user preferences. Learners obviously prefer AI feedback that is friendly, encouraging, and fluent. In addition, the higher degree of personalization in AI response is associated with stronger learner participation and satisfaction. Therefore, in the actual classroom application, teachers should choose or customize the AI dialogue system according to the specific teaching goals and the characteristics of learners. At the same time, teachers must maintain a dominant position in AI-assisted teaching and ensure proper control and supervision of AI functions to maximize their positive teaching potential.

However, there are certain limitations in this study. The relatively small sample size and short observation period limit the comprehensive evaluation of the long-term stability and migration effect of different models. Future research should expand the sample size and extend the longitudinal observation period to explore the long-term impact and universality of AI-assisted learning. In addition, follow-up research can include learners of different ages, genders, and cultural backgrounds to test the adaptability of AI dialogue systems in different learning situations, so as to provide more universal theoretical and practical insights for L2 education.

## References

[1] An X, Chai C S, Li Y, et al. Modeling students' perceptions of artificial intelligence-assisted language learning. Computer Assisted Language Learning, 2025, 38(5-6): 987-1008.

[2] Wiboolyasarin W, Wiboolyasarin K, Tiranant P, et al. AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. Ampersand, 2025, 14: 100224.

[3] Ericsson E, Johansson S. English speaking practice with conversational AI: Lower secondary students' educational experiences over time. Computers and Education: Artificial Intelligence, 2023, 5: 100164.

[4] Xu S, Qin L, Chen T, et al. Large language model-based situational dialogues for second language learning. arXiv preprint arXiv:2403.20005, 2024.

[5] Naveed H, Khan A U, Qiu S, et al. A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 2025, 16(5): 1-72

[6] Polakova P, Klimova B. Implementation of AI-driven technology into education–a pilot study on the use of chatbots in foreign language learning. Cogent education, 2024, 11(1): 2355385.

[7] Zhang Z, Huang X. The impact of chatbots based on large language models on second language vocabulary acquisition. Heliyon, 2024, 10(3).

[8] Dai W, Lin J, Jin H, et al. Can large language models provide feedback to students? A case study on ChatGPT. 2023 IEEE international conference on advanced learning technologies (ICALT). IEEE, 2023: 323-325.

[9] Hagendorff T, Fabi S, Kosinski M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 2023, 3(10): 833-83

[10] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. ACM transactions on intelligent systems and technology, 2024, 15(3): 1-45.

[11] Vanjani M, Aiken M, Park M. USING A MULTILINGUAL CHATBOT FOR FOREIGN LANGUAGE PRACTICE. Issues in Information Systems, 2020, 21(2).

[12] Lee J H, Shin D, Hwang Y. Investigating the capabilities of large language model-based task-oriented dialogue chatbots from a learner's perspective. System, 2024, 127: 103538.

[13] Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021: 610-623.

[14] Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 2023, 103: 102274.

[15] Mohamed A M. Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: perceptions of EFL Faculty Members. Education and Information Technologies, 2024, 29(3): 3195-3217.

[16] Xie Q, Wang Z, Feng Y, et al. Ask Again, Then Fail: Large Language Models' Vacillations in Judgment. arXiv preprint arXiv:2310.02174, 2023.