AI Chatbots as a Hypothesis-Testing Ground: Enhancing Oral Accuracy for Chinese EFL Undergraduates Through Task-Based Learning

Xinyu Fu

Beijing Foreign Studies University, Beijing, 10089, China
yuki160736@qq.com

Abstract. This review examines the role of AI chatbots in enhancing oral accuracy among Chinese EFL undergraduates from the combined perspectives of the Output Hypothesis and Task-Based Language Teaching (TBLT). In Chinese EFL contexts, the development of accurate spoken English is constrained by limited speaking opportunities, high speaking anxiety, and teacher-centered, exam-oriented instruction. Synthesizing empirical and theoretical studies published between 2020 and 2025, the paper conceptualizes AI chatbots as a "hypothesis-testing ground" that supports learners' formulation, testing, and refinement of linguistic hypotheses through iterative cycles of output, feedback, and modification.

The review identifies three key mechanisms through which chatbot-mediated tasks facilitate oral accuracy development: (1) the creation of low-stakes, non-judgmental environments that reduce anxiety and increase willingness to communicate; (2) the provision of immediate and adaptive feedback that promotes hypothesis revision; and (3) the affordance of sustained and repeatable task-based interaction that is difficult to achieve in large classrooms. The paper also addresses key limitations, including technological constraints, variability in learner affective responses, and limited interactional authenticity.

Based on the synthesis, the review outlines major research gaps and future directions, including micro-level process analyses, systematic evaluation of feedback quality, and longitudinal research in Chinese EFL contexts. Pedagogical implications are discussed with respect to hypothesis-testing-oriented task design, learner metacognition, and blended learning models integrating chatbot-mediated practice with teacher instruction.

Keywords: AI chatbots; oral accuracy; Output Hypothesis; Task-Based Language Teaching (TBLT); Chinese EFL undergraduates

1 Introduction

Oral accuracy remains a persistent challenge for Chinese EFL undergraduates despite prolonged formal instruction, a phenomenon often described as "mute English". Traditional classroom conditions, including large class sizes, exam-oriented curricula, and teacher-centered pedagogy, limit opportunities for sustained oral output and contribute

S. Garcia-Esteban et al. (eds.), *Proceedings of the 2026 5th International Conference on Social Sciences and Humanities and Arts (SSHA 2026)*, Advances in Social Science, Education and Humanities Research 1014,
https://doi.org/10.2991/978-2-38476-577-5_63

to heightened speaking anxiety. From the perspective of the Output Hypothesis, oral accuracy develops through repeated cycles of output, feedback, and modification that enable learners to test linguistic hypotheses [1]. However, such low-risk, high-frequency opportunities are scarce in Chinese classrooms, often resulting in fossilized interlanguage.

Recent advances in Artificial Intelligence (AI), particularly Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) [2], have enabled conversational chatbots to function as non-judgmental interlocutors for meaning-focused interaction [3, 4]. This review, therefore, examines how AI chatbots, embedded within a Task-Based Language Teaching (TBLT) framework, support the hypothesis-testing function of output in Chinese EFL contexts.

2 Theoretical Framework

This study draws on Swain's [5-7] Output Hypothesis, which positions language production as central to second language development. Focusing on the hypothesis-testing function, output enables learners to test and revise provisional hypotheses through interactional feedback, a process closely linked to oral accuracy development. TBLT provides the pedagogical framework for eliciting such output through meaning-focused, goal-oriented tasks [8]. Pre-task–task–post-task cycles and interactional processes promote negotiation of meaning and output modification [9, 10]. Within these frameworks, AI chatbots can function as stable task interlocutors that amplify output-driven learning processes.

3 AI Chatbots as a Hypothesis-Testing Ground in TBLT

Building upon the Output Hypothesis and TBLT, recent studies (2020–2025) conceptualize AI chatbots as a dynamic "hypothesis-testing ground" for oral output development. While the reviewed studies often involve learners from East Asian settings (e.g., Korea), their collective findings offer critical insights that can inform pedagogical applications for Chinese EFL undergraduates.

3.1 Creating the Low-Stakes Environment for Testing

A central mechanism identified in the literature is the creation of a low-stakes, psychologically safe interactional environment that lowers affective barriers to speaking. Foreign language speaking anxiety (FLSA) has long been recognized as a major constraint on oral participation [11-13], particularly in teacher-centered classrooms where fear of negative evaluation can suppress output [14]. Consistent with earlier findings on computer-mediated communication [15], empirical studies demonstrate that chatbot-mediated interaction reduces anxiety and enhances learners' willingness to communicate, confidence, and motivation [16-19]. Evidence from Taiwan and Korea further shows that sustained task-based chatbot use can simultaneously lower anxiety and improve

speaking performance [18, 20], creating the psychological safety necessary for learners to articulate and test interlanguage hypotheses without fear of embarrassment [21].

3.2 Providing Feedback for Hypothesis Revision

Beyond affective support, AI chatbots function as critical feedback providers that enable hypothesis revision [4], a core component of the Output Hypothesis. Immediate and context-sensitive feedback allows learners to identify mismatches between intended meaning and linguistic form, prompting reformulation. Recent research highlights the growing sophistication of such feedback, particularly in generative AI systems. kim et al. [18], for example, found that ChatGPT-generated feedback in pragmatics role-play tasks facilitated both noticing and deeper understanding through combined sociopragmatic and pragmalinguistic explanations, resulting in improved performance on complex speech acts. These findings align with earlier work on affect-sensitive and context-aware conversational agents [22, 23], suggesting that chatbot feedback can move beyond surface error correction to support metalinguistic reflection and interlanguage restructuring by confirming, disconfirming, or refining learners' hypotheses.

3.3 Enabling Iterative Practice within Tasks

When embedded within TBLT, chatbots further operationalize the iterative cycle of output-feedback-modification through sustained and repeatable task-based interaction. Task-oriented AI chatbots have been shown to compensate for limited speaking opportunities by engaging learners in meaning-focused dialogues that require negotiation of meaning and functional language use, including real-world tasks such as taking orders at a restaurant, shopping for shoes, and reserving a party room [24]. Because chatbots are always available and infinitely patient, learners can repeatedly test linguistic hypotheses across similar or varied task scenarios. This level of interactional density is challenging to achieve in large classes. The potential of generative AI in this regard is notable, as Liu and Reinders [25] found that it enhanced motivation and supported self-regulated learning compared to pre-scripted systems. As always-available interlocutors, AI chatbots promote negotiation of meaning and output central to both TBLT and the Output Hypothesis through repeated practice.

3.4 Current Limitations and Challenges

However, the literature also identifies important constraints. Technological limitations, including inaccurate speech recognition, shallow conversational depth, and occasional feedback breakdowns, may result in flawed or unclear responses that hinder effective hypothesis revision [24, 26]. Furthermore, affective benefits are not universal, with some learners reporting increased anxiety, and the limited authenticity of chatbot interaction may reduce the transferability of practice to real-world communication [27]. These challenges suggest that the effectiveness of chatbot-mediated hypothesis testing depends not merely on technological access but on pedagogically informed task design, appropriate scaffolding, and guided interpretation of feedback.

4 Research Gaps and Future Directions

Although studies published between 2020 and 2025 support the potential of AI chatbots as a task-based "hypothesis-testing ground", several gaps limit a comprehensive understanding of their effectiveness for Chinese EFL undergraduates.

4.1 Limited Focus on Chinese EFL Undergraduates

Existing empirical work is heavily concentrated in other East Asian contexts, particularly Korea and Japan [17, 22, 24], a pattern also noted in recent systematic reviews [28, 29]. The distinct socio-educational landscape of China, characterized by large class sizes, intense testing pressure, and specific L1 transfer issues, remains underexplored. Future research should directly investigate whether affective and communicative gains reported elsewhere (16, 18) are replicable and equally robust in Chinese university settings.

4.2 Insufficient Micro-Level Evidence on the Hypothesis-Testing Process

Although the macro-level benefits of chatbot interaction, such as increased conversation turns and task success rates, are well-documented [24], the micro-process of how learners form, test, and revise linguistic hypotheses during these interactions remains poorly understood. We lack fine-grained analyses of the specific moments where a learner's output is pushed, the nature of the cognitive comparison they make upon receiving chatbot feedback, and how this leads to immediate or delayed modifications in their interlanguage.

Future research could employ methodologies such as stimulated recall interviews, conversation analysis of interaction logs, and eye-tracking to illuminate these internal mechanisms. This would move beyond simply measuring outcomes to explaining the process of acquisition facilitated by chatbots, directly addressing the core of the Output Hypothesis.

4.3 Lack of a Systematic Framework for Evaluating Feedback Quality

While chatbots are recognized as feedback providers [4], there is a notable absence of a coherent, empirically grounded framework for evaluating the pedagogical quality of this feedback. Studies report on feedback in a fragmented manner, ranging from implicit recasts in conversation [24] to explicit sociopragmatic explanations. It remains unclear which types of feedback (e.g., implicit vs. explicit, pragmalinguistic vs. metalinguistic) are most effective for facilitating hypothesis revision for different error types and proficiency levels.

Future research must systematically categorize AI chatbots' feedback and investigate its impact on the accuracy and durability of learners' hypothesis revisions. Establishing such a framework is crucial for guiding the development of more pedagogically

intelligent chatbots that can provide feedback optimally supporting interlanguage restructuring.

4.4 Shortage of Longitudinal Studies on Long-Term Development

Most studies rely on short-term interventions, capturing immediate gains but providing limited evidence on the sustainability of oral accuracy or the long-term development of hypothesis-testing behaviors [25, 29, 30]. It therefore remains unclear whether chatbot-mediated practice leads to durable learning or transfer to real-world human-to-human communication. Longitudinal, mixed-method studies tracking learners over extended periods are needed to evaluate sustained effects.

5 Pedagogical Implications

Based on the synthesis of empirical findings and theoretical foundations reviewed in this paper, the following pedagogical implications are proposed to optimize the integration of AI chatbots as tools for oral output practice within a TBLT framework for Chinese EFL undergraduates, who often learn in large, exam-oriented classrooms with limited individual speaking opportunities.

5.1 Designing "Hypothesis-Testing-Driven" Tasks

Teachers should design tasks that prioritize meaningful communication and linguistic experimentation rather than form-focused drills, thereby creating regular opportunities for output and hypothesis testing [31, 32]. Tasks should involve clear, goal-oriented outcomes grounded in learners' everyday or academic contexts to ensure an authentic need to communicate [24]. At the same time, task complexity should be carefully adjusted to stretch learners' interlanguage without causing cognitive overload, particularly in mixed-ability classrooms.

A pre-task–task–post-task cycle offers a practical structure for implementation. The pre-task phase prepares key language knowledge, the task phase engages learners in chatbot-mediated interaction as the primary site for iterative practice, and the post-task phase supports reflection and teacher-led feedback to consolidate accuracy and address persistent errors.

5.2 Fostering Metacognition and Feedback Literacy

To maximize the benefits of chatbot interaction, learners require explicit guidance on using these tools strategically for autonomous practice. Instruction should cultivate metacognitive awareness of the output-feedback-revision cycle [7], encouraging students to formulate hypotheses, attend to chatbot feedback, and refine their language accordingly. Rather than passively accepting chatbot responses, learners should be trained to actively set focused goals, monitor their performance, and reflect on recurring errors.

Such feedback literacy shifts chatbot use from simple task completion to deliberate language improvement, supporting self-regulated learning and reducing the risk of fossilized output [33, 25]. Through structured reflection and guided strategy use, chatbot-mediated practice can more effectively promote sustained gains in oral accuracy.

5.3 Proposing a Blended Learning Model

To address limited speaking opportunities in large Chinese EFL classrooms, a blended model that connects in-class instruction with out-of-class chatbot practice is recommended. During the in-class pre-task phase, teachers introduce and scaffold tasks by clarifying goals, key language points, and strategies to prepare learners for independent interaction [34].

Learners then complete core tasks with AI chatbots outside class, where the low-stakes and always-available environment supports repeated practice and hypothesis testing, enabling more extensive output than is typically possible in large classrooms [24]. Task repetition further promotes fluency and consolidates accuracy development [35].

Learners finally return to class for reporting, peer discussion, and teacher-led feedback. This stage supports attention to form, targeted error correction, and authentic, socially complex negotiation of meaning, all of which remain difficult for chatbots to fully replicate [27]. By complementing out-of-class chatbot practice with teacher-guided interaction, this integrated model consolidates learning and facilitates the transfer of gains to real-world communication, thereby promoting sustained oral development.

6 Conclusion

This review has examined the role of AI chatbots in supporting oral accuracy development among Chinese EFL undergraduates through the combined perspectives of the Output Hypothesis and TBLT. It argues that, when embedded within a TBLT framework, AI chatbots offer a promising environment for operationalizing the hypothesis-testing function of output. By providing a low-stakes and psychologically safe interactional space, chatbots can reduce FLSA and encourage linguistic risk-taking. Moreover, their capacity to deliver immediate responses and to sustain repeated practice within meaningful tasks supports the output-feedback-modification process that is central to interlanguage development and accuracy gains.

Drawing on this synthesis, three pedagogical implications emerge for the Chinese EFL context. First, instruction should prioritize hypothesis-testing-driven tasks that are meaning-oriented, contextually relevant, and structured around a clear pre-task, task, and post-task cycle. Second, learners' metacognition and feedback literacy should be explicitly developed so that chatbot interaction promotes conscious hypothesis formation and revision rather than superficial practice. Third, a pragmatic blended learning model is recommended, in which chatbots are employed for extensive out-of-class oral

practice while classroom time is reserved for deeper negotiation of meaning and teacher-mediated feedback that current AI systems cannot yet fully provide.

Despite this potential, important constraints remain. Technological limitations, particularly in speech recognition accuracy and conversational depth, may undermine the reliability of feedback and disrupt hypothesis testing. Moreover, the current evidence base remains constrained not only by a shortage of longitudinal and context-specific research in Chinese university settings, but also by limited micro-level accounts of how learners engage in hypothesis testing and by the absence of systematic frameworks for evaluating the pedagogical quality of chatbot feedback. Future studies could adopt more context-sensitive, process-oriented, and longitudinal designs to better understand the mechanisms and impacts of chatbot-supported task-based practice on sustained oral development. Addressing these issues will help position AI chatbots not as substitutes for teachers, but as complementary tools that expand opportunities for meaningful output and support learners' progression from classroom learning to functional language use.

References

1. Swain, M.: The output hypothesis: Just speaking and writing aren't enough. *Canadian Modern Language Review* 50, 158–164 (1993)
2. Adamopoulou, E., Moussiades, L.: Chatbots: History, technology, and applications. *Machine Learning with Applications* 2, 100006 (2020)
3. Fryer, L., Carpenter, R.: Bots as language learning tools. *Language Learning & Technology* 10(3), 8–14 (2006)
4. Huang, W., Hew, K.F., Fryer, L.K.: Chatbots for language learning—are they really useful? *Journal of Computer Assisted Learning* 38(1), 237–257 (2022)
5. Swain, M.: Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In: Gass, S., Madden, C. (eds.) *Input in Second Language Acquisition*, pp. 235–253. Newbury House (1985)
6. Swain, M.: Three functions of output in second language learning. In: Cook, G., Seidlhofer, B. (eds.) *Principle and Practice in Applied Linguistics*, pp. 125–144. Oxford University Press (1995)
7. Swain, M.: The output hypothesis: Theory and research. In: Hinkel, E. (ed.) *Handbook of Research in Second Language Teaching and Learning*, pp. 495–508. Routledge (2005)
8. Prabhu, N.S.: *Second Language Pedagogy*. Oxford University Press (1987)
9. Long, M.H.: A role for instruction in second language acquisition. In: Hyltenstam, K., Pienemann, M. (eds.) *Modelling and Assessing Second Language Acquisition*, pp. 77–99. Multilingual Matters (1985)
10. Willis, J.: *A Framework for Task-Based Learning*. Longman (1996)
11. Horwitz, E.K., Horwitz, M.B., Cope, J.A.: Foreign language classroom anxiety. *Modern Language Journal* 70(2), 125–132 (1986)
12. MacIntyre, P.D., Gardner, R.C.: The subtle effects of language anxiety on cognitive processing. *Language Learning* 44, 283–305 (1994)
13. Woodrow, L.: Anxiety and speaking English as a second language. *RELC Journal* 37, 308–328 (2006)
14. Liu, M., Jackson, J.: Chinese EFL learners' unwillingness to communicate. *Modern Language Journal* 92, 71–86 (2008)

15. Arnold, N.: Reducing foreign language communication apprehension. *System* 35(4), 469–486 (2007)
16. Ding, D., Muhyiddin B Yusof, A.: AI-powered conversation bots and speaking anxiety. *Humanities and Social Sciences Communications* 12(1), 1223 (2025)
17. Jeon, J.: AI chatbot affordances in the EFL classroom. *Computer Assisted Language Learning* 37(1–2), 1–26 (2024)
18. Kim, A., Su, Y.: AI chatbot and willingness to communicate. *System* 122, 103256 (2024)
19. Shikun, S., Grigoryan, G., Huichun, N., Harutyunyan, H.: AI chatbots in EFL classroom. *Arab World English Journal* 1(1), 292–305 (2024)
20. Hsu, M.-H., Chen, P.-S., Yu, C.-S.: Task-oriented chatbot system for EFL learners. *Interactive Learning Environments* 31(7), 4297–4308 (2023)
21. Young, D.J.: Creating a low-anxiety classroom environment. *Modern Language Journal* 75, 426–437 (1991)
22. Ayedoun, E., Hayashi, Y., Seta, K.: Communicative and affective strategies in conversational agents. *International Journal of Artificial Intelligence in Education* 29(1), 29–57 (2019)
23. Tu, J.: AI-powered chatbot for language tutoring. *Journal of Physics: Conference Series* 1693(1), 012216 (2020)
24. Yang, H., Kim, H., Lee, J.H., Shin, D.: AI chatbot as English conversation partner. *ReCALL* 34(3), 327–343 (2022)
25. Liu, M., Reinders, H.: Do AI chatbots impact motivation? *System* 128, 103544 (2025)
26. Alrajhi, A.S.: Artificial intelligence pedagogical chatbots. *Cogent Education* 11(1), 2327789 (2024)
27. Annamalai, N., Rashid, R.A., Munir Hashmi, U., Mohamed, M., Harb Alqaryouti, M., Eddin Sadeq, A.: Chatbots for English language learning. *Computers and Education: Artificial Intelligence* 5, 100153 (2023)
28. Du, J., Daniel, B.K.: Systematic review of AI-powered chatbots. *Computers and Education: Artificial Intelligence* 6, 100230 (2024)
29. Şahin Kızıl, A., Klimova, B., Pikhart, M., Parmaxi, A.: Systematic review of chatbots in language education. *Journal of Computer Assisted Learning* 41(2), e70001 (2025)
30. Hwang, G.-J., Chang, C.-Y.: Opportunities and challenges of chatbots in education. *Interactive Learning Environments* 31(7), 4099–4112 (2023)
31. Ellis, R., Skehan, P., Li, S., Shintani, N., Lambert, C.: *Task-Based Language Teaching: Theory and Practice*. Cambridge University Press (2019)
32. Long, M.H.: The role of the linguistic environment in SLA. In: Ritchie, W.C., Bhatia, T.K. (eds.) *Handbook of Second Language Acquisition*, pp. 413–468. Academic Press (1996)
33. Ellis, R.: Current issues in grammar teaching. *TESOL Quarterly* 40(1), 83–107 (2006)
34. Van den Branden, K.: The role of teachers in TBLT. *Annual Review of Applied Linguistics* 36, 164–181 (2016)
35. Bygate, M.: Effects of task repetition on oral language. In: Bygate, M., Skehan, P., Swain, M. (eds.) *Researching Pedagogic Tasks*, pp. 23–48. Pearson (2001)