AI Feedback vs. Teacher Feedback: A Comparative Study of English-Speaking Practice among L1 Mandarin EFL Learners

Stella Li

A thesis submitted to
Auckland University of Technology
in partial fulfilment of the requirements for the degree of
Master of Creative Technologies (MCT)

2026

School of Future Environments

Table of Contents

List of Figures ..... *iv*

List of Tables ..... *v*

Abstract ..... *ix*

Chapter One: Introduction *1*

Chapter Two: Literature Review *4*

2.1 EFL Speaking Development and L1 Mandarin Learners **4**

2.1.1 Speaking as a Complex, Real-Time Skill 4

2.1.2 Speaking Challenges for L1 Mandarin EFL Learners 5

2.2 Feedback in EFL Speaking Practice **7**

2.2.1 The Role of Feedback in Oral Language Development 7

2.2.2 Teacher Feedback: Affordances and Constraints 8

2.2.3 AI-Generated Feedback: Affordances and Concerns 10

2.3 AI-Supported Speaking Practice and the Comparison Gap **12**

2.3.1 Recent Empirical Work on AI-Supported Speaking 12

2.3.2 Research Gap and Rationale for the Present Study 13

Chapter Three: Methodology *16*

3.1 Research Design **16**

3.2 Research Setting, Participants, and Recruitment **17**

3.2.1 Research Context 17

3.2.2 Participants 18

3.2.3 Recruitment 18

3.3 Programme Design and Procedures **19**

3.3.1 Overall Programme Structure 19

3.3.2 Shared Baseline Session and Group Allocation 20

3.3.3 Session Design 21

3.3.4 Cohort Delivery and Attendance 22

3.4 AI Tool and Prompt Design **22**

3.5 Data Collection Instruments and Sources **25**

3.5.1 IELTS Score Data 25

3.5.2 Likert-Scale Exit Survey Items 25

3.5.3 Open-Ended Exit Survey Responses 26

3.6 Data Analysis **26**

3.6.1 Quantitative Analysis 27
3.6.2 Qualitative Analysis 28
3.6.3 Integration of Quantitative and Qualitative Findings 29
3.7 Researcher Role 29
3.8 Ethical Considerations 30
3.9 Chapter Summary 31
Chapter Four: Findings 33
4.1 Participant Overview 33
4.2 IELTS Score Changes 34
4.2.1 Descriptive Pattern of Change 34
4.2.2 Within-Group Comparisons 35
4.2.3 Between-Group Comparisons 36
4.2.4 Patterns Across Individual IELTS Skills 38
4.3 Likert Survey Responses 38
4.3.1 Agree and Strongly Agree Responses 38
4.3.2 Item-Level Group Comparisons 39
4.4 Open-Ended Responses 40
4.4.1 Shared Themes 41
4.4.2 AI-Supported Group: Distinctive Themes 43
4.4.3 Teacher-Led Group: Distinctive Themes 44
4.5 Integration of Quantitative and Qualitative Findings 45
4.6 Chapter Summary 47
Chapter Five: Discussion 48
5.1 Speaking Gains, Repetition, and Feedback Source 48
5.2 Mixed Methods Convergence: Speaking Scores and Learner Development 50
5.3 Mixed Methods Divergence: Different Affordances of AI and Teacher Feedback 52
5.4 Pedagogical Implications 55
5.5 Limitations and Future Research 57
5.6 Chapter Summary 60
Chapter Six: Reflections on the Use of AI for Language Learning and Research 61
6.1 My Use of AI as a Teacher 61

6.1.1 Positioning Myself as a Teacher-Researcher 62
6.1.2 Designing AI-Supported IELTS Speaking Practice 63
6.1.3 Observing Participants' Interaction with AI 66
6.1.4 What I Learned as a Teacher 68
6.2 My Use of AI as a Researcher 70
6.2.1 Literature Searching and Reading 70
6.2.2 AI as a Process Tutor 74
6.2.3 AI as a Writing Support Tool 77
6.2.4 Risks and Limitations of Using AI in Research 79
6.3 Key Takeaways 81
6.3.1 For Language Learners 81
6.3.2 For Teachers 81
6.3.3 For Researchers 82
6.3.4 AI Prompt Design Guidelines 82
6.4 Chapter Summary 83
Chapter Seven: Conclusion 85
References 87
Appendix A Participant Information, Eligibility Screening, and Consent Materials 91
Appendix B Group Allocation Instructions 101
Appendix C Exit Survey 104
Appendix D AI Feedback Prompt Used in the Study 111

List of Figures

Figure 1 Overall Programme Structure 20
Figure 2 IELTS Speaking Part 1 Sample Questions 21
Figure 3 IELTS Speaking Part 2 Sample Questions 21
Figure 4 Pre- and Post-Test IELTS Speaking Scores by Group 35
Figure 5 IELTS Gain Scores by Skill and Group 37
Figure 6 Percentage of Participants Who Selected Agree or Strongly Agree by Category and Group 39
Figure 7 Prompt Development Process 63
Figure 8 AI Prompt Revision Cycle 65
Figure 9 AI-Supported IELTS Speaking Part 2 Practice Cycle 66
Figure 10 Anonymised AI Interaction Record 67
Figure 11 Example of AI-Supported Answer Expansion 68
Figure 12 Example of AI Support for Identifying Literature Search Terms 71
Figure 13 Example of AI Support for Locating and Checking Relevant Literature 72
Figure 14 Example of AI-Supported Initial Screening of a Journal Article 73
Figure 15 Example of AI Support for Learning Data-Analysis Procedures: Part 1 76
Figure 16 Example of AI Support for Learning Data-Analysis Procedures: Part 2 77

List of Tables

Table 1 Summary Comparison of AI Platforms Tested During the Preparatory Phase 23
Table 2 Cohort Composition and Size 33
Table 3 Baseline IELTS Scores by Group 34
Table 4 Descriptive Summary of Pre- to Post-Test IELTS Score Changes 35
Table 5 Within-Group Pre- to Post-Test Comparisons by Group 36
Table 6 Between-Group Comparisons on IELTS Gain Scores 37
Table 7 Likert Survey Items and Abbreviated Labels Used in Chapter 4.3 38
Table 8 Likert Survey Responses by Group 40
Table 9 Themes Identified in the Open-Ended Responses 41
Table 10 Integrated Evidence Across IELTS Scores, Survey Responses, and Open-Ended Responses 45
Table 11 Anatomy of the Final AI Feedback Prompt 65
Table 12 Different Affordances of AI Feedback and Teacher Feedback 69
Table 13 AI-Supported Literature Searching and Reading Workflow 74
Table 14 Researcher Learning From AI-Supported Writing 78
Table 15 Risks and Safeguards in AI-Supported Research 80
Table 16 AI Prompt Design Guidelines for Researchers, Teachers, and Learners 83

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature:

Artificial Intelligence Declaration

Specify the chapter number(s) in your research proposal/research component where AI was used.	
Chapter number(s):	1-7
Briefly explain why the AI tool was used (e.g., idea generation, summarization, data analysis, image creation, etc.).	
Purpose of AI Use:	AI tools were used as learning and research-support tools during the thesis process. Uses included language polishing and readability checking; support for literature screening and translation; explanations of statistical analysis, document formatting, and proofreader's comments; and delivery of structured feedback in the AI-supported speaking practice condition in the research programme.
Clearly state the AI tools used.	
AI Tool(s) Used:	ChatGPT; Claude; Doubao-Seed-2.0; Google Translate
Prompts or queries entered into the AI tool (only the instructive portion of the prompt is required). Include all relevant variations.	
Prompts Used:	Examples included: check grammar and readability; make the wording more concise without changing the meaning; improve sentence clarity; identify possible overclaiming; summarise the abstract or method and indicate its possible relevance; explain this statistical test, formatting procedure, or supervisor/proofreader comment; translate this Chinese response into English; and suggest a clearer organisation for this paragraph or table without changing the findings. For the AI-supported speaking sessions in the research programme, Doubao-Seed-2.0 was instructed to provide feedback using teacher-designed IELTS Speaking criteria, identify strengths and areas for improvement, suggest spoken reformulations, and ask follow-up practice questions.
A summary or description of the AI-generated output. (Optional: attach output in an appendix.)	
Output Received:	Outputs included grammar and readability suggestions, concise wording options, explanations of statistical and formatting procedures, literature-screening summaries, translation drafts, organisational suggestions for paragraphs and tables, and structured speaking feedback for participants in the AI-supported condition.
Describe how the AI output was used, edited, or transformed.	
Post-AI Processing Methods:	All AI outputs were reviewed critically and used only as support. The researcher made the final decisions about thesis content, structure, argument, analysis, interpretation, conclusions, and wording. Suggestions were revised, rewritten, accepted, or rejected as appropriate. References and claims were checked against original sources; statistical statements against the dataset and JASP outputs; and translations and quotations against the original Mandarin responses. For the research programme, the AI prompt was designed and tested in advance and used within a teacher-guided lesson structure. AI outputs were not treated as authoritative sources.
You may use this free text box to explain the intent for the use of the AI generated content. *For example, rather than stating what prompt(s) were used, you may articulate the purpose for which the prompt was used (supported by examples) here:*	
AI tools were used as learning and research-support tools, not as substitutes for my own work. ChatGPT and Claude supported language refinement, literature screening, and explanations of research procedures; Google Translate supported initial translation of Chinese survey responses; and Doubao delivered structured feedback in the AI-supported speaking condition through a teacher-designed prompt. The submitted thesis is substantively my own work, and I remained responsible for all research decisions, analysis, interpretation, conclusions, and final wording.	

STUDENT DECLARATION					
By signing you are confirming that the AI use stated in the table(s) above are accurate and follow AUT's recommended guidelines detailed in the AI Hub and Postgraduate Handbook.					
Student Name:	**Stella Li**	**Signature:**		**Date:**	**10/07/2026**

Acknowledgements

I am deeply grateful to my supervisor, Associate Professor Ann Morrison, for her guidance throughout the past year. Under her supervision, I learned far more than how to complete a thesis. I developed a professional researcher's toolkit and strengthened my academic judgement. Her feedback continually encouraged me to examine my assumptions and approach my work with greater care and rigour. I was also deeply influenced by the positivity she brought to every supervision meeting. Her encouraging and forward-looking attitude gave me a steady sense of support throughout the research process. I feel very fortunate to have had her as my supervisor and honoured to have worked with her over the past year.

I would also like to thank my proofreader, Agnes Terraschke, for her professional, thoughtful, and exceptionally detailed comments. Her careful attention to language, reasoning, and clarity added considerable polish to the final work.

Ethical approval for this study was granted by the Auckland University of Technology Ethics Committee (AUTEC; reference 25/429) on 5 February 2026.

My sincere thanks go to Zhang Yu for her valuable advice on the design and use of AI feedback in the speaking sessions. I am especially grateful to all the participants who took part in this study. Their time, trust, and honest contributions made this research possible.

I would also like to thank Samra Jabeen Hanif, Sophia Rose Lawler-Dormer, Dylan Verdonkschot and Stefan Marks for sharing ideas, helping me work through practical problems, and supporting me throughout the research process. I am also grateful to Auckland University of Technology for providing the academic environment and resources that enabled me to undertake this research and develop as an independent researcher.

Finally, I would like to thank my family and friends for their support throughout this year. I am deeply grateful to my parents for always believing in me and supporting my decisions, even from afar. I would also like to thank my partner, Kevin, for his patience, understanding, care, and constant encouragement. My thanks also go to Xiaotao, Sydni, Seaya, Jay, Brad, Sun Haonan, Xue Fuli, Ma Junnan, and Liu Xi for their friendship, companionship, and encouragement.

This journey was more difficult than I had imagined, but it was also more rewarding. It tested my patience, resilience, and persistence, and shaped me both as a researcher and as a person. I hope to carry the knowledge, skills, and perspectives I have gained into the future and continue exploring new possibilities with confidence, curiosity, and perseverance.

Abstract

Speaking is a major challenge for many L1 Mandarin learners preparing for IELTS. This study compared AI-supported and teacher-led feedback in an eight-session online IELTS Speaking programme. Using a mixed methods design, the study involved 24 participants, with 13 receiving AI-supported feedback through Doubao-Seed-2.0, a generative AI chatbot, and 11 receiving teacher-led feedback. Outcomes were examined through official IELTS Speaking scores and learner survey responses. While both groups showed significant within-group gains in IELTS Speaking, the between-group comparison did not detect a significant difference in speaking gain. Participants in both groups reported greater confidence and clearer expression. AI-supported participants particularly valued answer expansion, alternative expressions, and repeated practice, while teacher-led participants valued pronunciation correction, live explanation, and classroom interaction. The findings suggest that the two feedback modes may offer different but complementary forms of support. A reflection chapter further examines the researcher's experiences of using AI in both teaching and research, highlighting the importance of teacher judgement, careful prompt design, and critical AI use.

Keywords: generative artificial intelligence; AI feedback; teacher feedback; EFL speaking development; speaking practice; IELTS Speaking; L1 Mandarin learners

Chapter One: Introduction

Many Chinese learners study English in the hope of improving their educational and employment opportunities both at home and abroad. In English as a Foreign Language (EFL) context, the International English Language Testing System (IELTS) is a widely recognised English proficiency test used for study, work, and migration purposes. As a result, many learners take IELTS when applying to overseas universities, seeking employment, or pursuing migration pathways. However, learning another language is difficult, and some skills are harder to develop than others. For many Chinese learners of English, speaking is a particular challenge. This is reflected in IELTS Academic test-taker performance data for 2024–2025, where candidates whose first language was recorded as Chinese achieved a mean speaking band score of 5.5, the lowest among their four skills (IELTS, 2025b). Speaking can be difficult to improve because, beyond vocabulary and grammar, the IELTS Speaking test requires candidates to develop ideas in real time, sustain fluency and coherence, and use pronunciation and discourse organisation in ways that are understandable to examiners, as reflected in the official IELTS Speaking criteria (IELTS, 2025a). These demands require more than lecture-focused classes; learners also need opportunities for interaction, speaking practice, feedback, and revision. In this thesis, “IELTS Speaking” is capitalised when referring to the official test component, while “speaking” is used in lowercase when referring to speaking more generally.

Speaking development depends on opportunities to produce language in interaction, receive feedback on actual spoken performance, and try again. Research on second language speaking and instructed second language acquisition (SLA) has suggested that interaction gives learners opportunities to produce spoken output, negotiate meaning, notice communication problems, and refine their language through further practice (Goh & Burns, 2012; Loewen & Sato, 2018). Feedback is important in this process because it can draw learners’ attention to specific problems in their spoken language, such as unclear pronunciation, grammatical errors, limited vocabulary, or unclear organisation. Oral corrective feedback in speaking practice can also create opportunities for repair, learner uptake, and further practice (Li & Vuono, 2019; Lyster et al., 2013). For IELTS candidates, this means that preparation needs to go beyond memorising template answers. They need opportunities to answer questions aloud, develop and organise ideas, and receive feedback on how clearly and fluently their responses are communicated.

Feedback is therefore central to speaking instruction. In speaking practice, feedback can help learners understand how their spoken performance is received and where it could be improved. This may include feedback on pronunciation, grammar, vocabulary, fluency, coherence, and the development of content (Li & Vuono, 2019; Lyster et al., 2013). In IELTS Speaking preparation, feedback also needs to be connected to the demands of the speaking test, such as whether an answer is clear, sufficiently developed, coherently organised, and understandable to the listener

(IELTS, 2025a). In conventional speaking instruction, teachers usually take responsibility for providing this support. They can listen to learners' responses, identify specific problems, explain why an answer may be unclear or underdeveloped, and guide learners towards more effective responses.

In many conventional EFL classrooms, however, teacher feedback is not always available to the extent learners need. Teachers often have to divide limited class time among many students, which means each learner may receive only a small number of speaking turns and limited individualised feedback. Zhang et al. (2024) provided an example of this constraint: the teacher in their study found it difficult to provide individualised feedback on every learner's oral task because of class size and limited time. For the present study, this highlights a practical gap between what speaking learners need and what classroom instruction can provide.

Generative artificial intelligence (AI) has recently attracted attention as a possible source of additional speaking practice and feedback. In language learning, generative AI can provide immediate responses, reformulations, explanations, and repeated opportunities for practice (Kohnke et al., 2023; Law, 2024; Yang & Li, 2024). Research has also reported potential benefits for speaking development and affective outcomes such as willingness to communicate, confidence, and anxiety reduction (Du & Daniel, 2024; Fathi et al., 2024; Zhang et al., 2024).

However, AI-generated feedback may also be inaccurate, overly general, or insufficiently sensitive to learners' intended meaning and learning context (Kohnke et al., 2023; Law, 2024). Despite growing interest in AI-supported speaking practice, direct comparisons between AI-generated feedback and teacher feedback remain limited, particularly in structured IELTS Speaking preparation.

The present study therefore compares AI-generated feedback and teacher feedback in an eight-session online IELTS Speaking programme for first language (L1) Mandarin EFL learners. The AI-supported group received feedback from Doubao (2023), while the teacher-led group received feedback from the teacher. A mixed methods design was used to examine both changes in IELTS Speaking scores and learners' experiences of the two feedback modes. The study addressed the following research questions:

1. How do AI-supported feedback and teacher-led feedback compare in supporting IELTS Speaking development among L1 Mandarin EFL learners?
2. How do learners experience the two feedback modes?

Comparable or different support may be reflected not only in changes to the speaking scores, but also in how learners experience response development, pronunciation support, repeated practice, and interaction.

The findings are intended to contribute to current understanding of how AI-supported and teacher-led feedback compare in IELTS speaking practice, both in relation to speaking development and learners' experiences.

This thesis is organised into seven chapters. Chapter One introduces the research background, problem, aim, and structure of the study. Chapter Two reviews literature on EFL speaking development, feedback in speaking instruction, AI-supported language learning, and the need to compare AI feedback with teacher feedback. Chapter Three explains the research design, participants, intervention, data collection, and analysis procedures. Chapter Four presents the quantitative and qualitative findings. Chapter Five discusses the findings in relation to previous research and the research questions. Chapter Six presents a researcher-teacher reflection on prompt design, feedback delivery, classroom interaction, teacher judgement, and responsible AI use. Chapter Seven concludes the thesis by summarising the main contributions, limitations, and implications for EFL speaking classrooms and future research.

Chapter Two: Literature Review

This chapter reviews the literature that informs this thesis. It is organised into three sections. The first section discusses English as a Foreign Language (EFL) speaking development, first explaining speaking as a complex real-time skill and then examining the challenges faced by L1 Mandarin learners preparing for the International English Language Testing System (IELTS), including limited speaking opportunities, willingness to communicate, affective factors, and communicative explicitness. The second section examines feedback in EFL speaking practice. It first outlines the role of feedback in oral language development, then compares the affordances and constraints of teacher feedback and artificial intelligence (AI) feedback. The final section reviews recent empirical work on AI-supported speaking practice and identifies the research gap addressed by this study: how AI-generated feedback compares to teacher feedback within the same structured IELTS Speaking programme.

2.1 EFL Speaking Development and L1 Mandarin Learners

For many language learners, developing speaking ability is an important but challenging part of learning English. This section considers speaking as a complex skill before examining challenges that may be relevant to L1 Mandarin learners preparing for IELTS Speaking.

2.1.1 Speaking as a Complex, Real-Time Skill

Speaking proficiency is often treated as a visible marker of overall English ability. However, attaining speaking fluency when learning another language can be difficult as speaking is not a single, simple skill. Goh and Burns (2012) described speaking as a complex ability that involves linguistic knowledge and cognitive processing. This means that successful speaking requires more than knowing grammar and vocabulary. Speakers need to listen to their interlocutor while simultaneously analysing what is being said, identifying the intended meaning, formulating an appropriate response, selecting suitable expressions, organising grammar, managing pronunciation, and maintaining fluency and intonation as the conversation continues. Kormos (2006) similarly explained second language speech production as a process that requires learners to conceptualise meaning, formulate language, and articulate speech under real-time pressure. These accounts help explain why learners may perform well in receptive or written tasks but still struggle when they need to speak spontaneously.

The complexity of speaking is also reflected in the IELTS Speaking assessment (IELTS, n.d.–a). The official IELTS Speaking criteria assess candidates through four areas: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation (IELTS, 2025a). These criteria show that IELTS Speaking is not simply a standardised test that can be prepared for through memorised answers. Candidates are assessed on how well they can develop ideas, sustain coherent speech, use language flexibly, and make themselves understood to an examiner.

A prepared answer may be limited if it is too short, unclear, underdeveloped, or poorly connected to the question. IELTS Speaking requires learners to turn language knowledge into real-time communication, which means it offers a clear context for spoken language performance.

Fluency forms part of the IELTS Speaking criterion of fluency and coherence (IELTS, 2025a). It depends partly on learners' ability to manage language production in real time (Foster, 2020; Kormos, 2006). Foster (2020) argued that oral fluency is shaped by cognitive, linguistic, social, and individual factors. In IELTS Speaking, candidates need to access language, organise ideas, and maintain communication while responding to unfamiliar questions. The challenge is therefore not only whether learners know enough English, but whether they can use that knowledge to produce clear and developed responses during the speaking task.

Interaction is another important part of speaking development. Long's (1996) Interaction Hypothesis emphasises that language learning can be supported through interaction, especially when learners negotiate meaning and receive signals that their message needs adjustment. Loewen and Sato (2018) also positioned interaction as central to instructed second language acquisition because it connects input, output, negotiation, and feedback. In speaking practice, learners are therefore not only producing language; they are testing whether their language is clear and meaningful to another listener. Through interaction, learners may become aware that an answer is too short, that a word choice is unclear to the listener, or that further explanation is needed. These moments can become useful for learning when learners receive feedback and have opportunities to modify their output. The value of this process depends on whether learners have regular access to responsive speaking practice and feedback. This is particularly relevant for IELTS learners, who need repeated opportunities to practise spoken responses and receive feedback.

For IELTS candidates, the need for repeated and responsive speaking practice is closely linked to the structure of the test. The IELTS Speaking test requires candidates to respond to familiar questions in Part 1, speak at length on a given topic in Part 2, and discuss more general and abstract issues in Part 3 (IELTS, n.d.–a). These tasks require candidates to organise ideas, use appropriate language, and explain opinions clearly to an examiner (IELTS, 2025b, n.d.–a). Research on task repetition suggests that repeated oral tasks can help learners gain greater control over spoken language performance (Bygate, 2001). Therefore, IELTS Speaking preparation needs to move beyond memorised model answers. Learners need opportunities to practise speaking, receive feedback, revise their responses, and try again.

2.1.2 Speaking Challenges for L1 Mandarin EFL Learners

Recent IELTS Academic test-taker performance statistics for 2024–2025 show that candidates whose first language was recorded as Chinese achieved a mean Speaking band score of 5.5, the lowest of their four skill scores and joint 39th out of 40 listed first-language groups (IELTS, 2025b).

These data do not represent all Chinese-speaking learners, and individual learners may perform very differently from this overall pattern. However, they do show a relevant pattern within this IELTS test-taker group: Speaking appears to be relatively weaker than Listening, Reading, and Writing. This pattern helps justify the focus of the present study on L1 Mandarin learners preparing for IELTS Speaking.

Research on Chinese and East Asian EFL learners suggests that speaking challenges are not only related to linguistic knowledge, but also to learners' access to opportunities for communication and their willingness to communicate. Peng (2012) showed that Chinese EFL learners' willingness to communicate (WTC) was shaped by the interaction of learner beliefs, motivation, cognitive and linguistic factors, affective factors, and classroom environment. Similarly, Shao and Gao (2016) reviewed research on East Asian language learners' reticence and WTC, suggesting that learners' classroom participation should be understood in relation to individual and contextual conditions, such as self-confidence, topic familiarity, task design, interlocutor familiarity, group size, medium of communication, and classroom environment. Together, these studies suggest that speaking difficulty is not simply a matter of insufficient English knowledge, but is also shaped by the learning context, including classroom conditions and opportunities to use English in spoken interaction.

Recent Chinese EFL research has examined the affective side of speaking participation in greater depth. Bai (2023) found that foreign language anxiety and boredom were negatively associated with WTC among Chinese struggling EFL learners. Chen (2024) showed that speaking anxiety and expectancy-value beliefs were related to Chinese undergraduates' spoken English proficiency. In this context, expectancy-value beliefs refer to learners' expectations of success and the value they attach to English learning. Lin et al. (2025) further found that foreign language anxiety was negatively related to WTC, while academic buoyancy, meaning learners' capacity to cope with everyday academic challenges, was positively related to WTC. These studies suggest that learners' speaking participation is closely connected to affective and motivational factors, especially anxiety, confidence, perceived value, and resilience.

Another challenge concerns communicative explicitness in IELTS Speaking. Hall (1976) introduced the distinction between high-context and low-context communication, where meaning relies more on shared background and implicit understanding in high-context communication, and more on direct verbal expression in low-context communication. A simple example is the difference between implying a request by saying that a room is cold and explicitly asking someone to close the window; in the former, the listener must infer the intended request from the context, whereas in the latter, the intended meaning is stated more directly (House & Kasper, 1981). Hall and Hall (1990) are often cited for associating Japanese communication with relatively high-context communication and Anglo-American communication with relatively low-context communication. Later studies extended this framework to other East Asian contexts. For example, Kim et al. (1998) compared Chinese, Korean, and American cultures and found patterns consistent with the view that Chinese

and Korean communication contexts may be relatively higher-context than American communication contexts. This distinction is relevant to IELTS Speaking because the test requires candidates to communicate with an examiner who does not share their personal background knowledge. The IELTS Speaking criteria reward clear, coherent, and sufficiently developed responses, so candidates need to state their ideas explicitly, support them with reasons or examples, and organise their responses for the listener. For some L1 Mandarin and East Asian EFL learners, this may create an additional challenge, especially if their previous speaking experience has not required extended explanation in English.

However, this framework needs to be used cautiously. It should not be taken to mean that all L1 Mandarin learners are indirect, unclear, or culturally predisposed to communicate in the same way. Kittler et al. (2011) cautioned that high-context and low-context classifications can be overgeneralised when national or cultural groups are treated as fixed and internally uniform. Park et al. (2012) also showed that preferences for direct communication can vary across individuals as well as cultures. In this study, Hall's (1976) framework is therefore used only as a broad lens for understanding communicative explicitness in IELTS Speaking, not as a fixed cultural explanation of L1 Mandarin learners' performance. The focus is on the test-oriented need to make ideas clear, developed, and accessible to the listener, and on how feedback may help learners practise this form of spoken performance.

Overall, the literature presented in this section suggests several practical challenges for IELTS Speaking preparation. Thus, learners with fewer opportunities to use English in spoken interaction may have fewer chances to build confidence, fluency, and response development. Learners who feel anxious, less confident, or less willing to communicate may also be less likely to take part in the speaking practice that could help them improve (Bai, 2023; Lin et al., 2025; Peng, 2012). In addition, IELTS Speaking requires candidates to produce clear, explicit, and sufficiently developed responses for an examiner who does not share their personal background knowledge. These challenges are shaped by speaking opportunities, affective factors, test demands, and learners' previous learning experiences.

2.2 Feedback in EFL Speaking Practice

Feedback is an important part of speaking practice because it helps learners understand how their spoken performance can be improved. This section first examines the role of feedback in oral language development and then discusses the affordances and constraints of teacher feedback and AI-generated feedback.

2.2.1 The Role of Feedback in Oral Language Development

Feedback is central to oral language development because it can help learners notice problems in their language use and respond to them. Schmidt's (1990) Noticing Hypothesis suggests that learners need to notice relevant language features before these features can become part of their

developing language system. Long's (1996) Interaction Hypothesis also emphasises the role of interaction and negotiation of meaning in creating opportunities for second language development. Loewen and Sato (2018) further positioned feedback within this wider interactional process, where input, output, negotiation, and feedback work together to support instructed second language acquisition. From this perspective, feedback is not simply correction after performance; it is a mechanism that helps learners become aware of gaps between what they intend to say and how successfully they express it. Corrective feedback has been shown to support second language learning when it is connected to learner output, allowing learners to identify problems in their language use, consider how these might be addressed, and modify their language production in a subsequent attempt (Li & Vuono, 2019; Lyster et al., 2013). In this thesis, the Noticing Hypothesis provides a useful theoretical lens for understanding both AI-generated and teacher feedback. Although the two feedback modes differ in how feedback is delivered, both may support learning by drawing learners' attention to gaps in their spoken performance and giving them an opportunity to revise their responses in a subsequent attempt.

In IELTS speaking practice, feedback needs to address more than discrete language errors. The IELTS Speaking criteria assess fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation (IELTS, 2025a). This means that feedback needs to focus not only on grammar or vocabulary, but also on whether an answer is clear, sufficiently developed, coherent, and understandable to the listener. A learner may produce an answer with few obvious errors but still be too brief, vague, poorly organised, or disconnected from the question. Feedback can help learners recognise these communicative limitations and target the areas identified in the feedback for their revisions.

The value of feedback also depends on whether learners can act on it. Hattie and Timperley (2007) argued that effective feedback helps learners understand the learning goal, their current progress, and the next steps needed for improvement. In speaking practice, this suggests that feedback should ideally lead to another attempt, not simply an end-point comment. For IELTS learners, a comment on pronunciation, coherence, or answer development is more useful when it gives them a clearer direction for repeating the same speaking task after feedback. Feedback in oral practice is therefore closely connected to repetition: it gives learners a reason and a focus for trying again.

2.2.2 Teacher Feedback: Affordances and Constraints

Considering the key role of feedback in oral language development, the source of feedback also matters. Teacher feedback and AI-generated feedback can both support speaking practice, but they differ in how feedback is produced and delivered. This section examines the affordances and constraints of teacher feedback before the following section considers AI-generated feedback.

Teacher feedback has value in speaking practice because teachers can respond to both language form and communicative meaning. Since speaking involves linguistic, cognitive, affective, and social dimensions, feedback in speaking lessons often needs to go beyond surface-level correction (Goh & Burns, 2012). In IELTS-focused lessons, a teacher may comment on pronunciation, grammar, vocabulary, fluency, coherence, idea development, or how easy a response is to follow, all of which are connected to the demands of the IELTS Speaking criteria (IELTS, 2025a). Teacher feedback can therefore help learners understand not only what is inaccurate, but also why a response may sound unclear, underdeveloped, or weakly organised.

Teacher feedback is also adaptive. Hattie and Timperley (2007) argued that effective feedback helps learners understand the learning goal, their current progress, and the next steps needed for improvement. In speaking practice, this means that teachers can respond at different levels: correcting a pronunciation error, suggesting a clearer way to organise an answer, prompting a learner to extend an idea, or encouraging a learner who lacks confidence. This flexibility is important because learners do not all need the same type of feedback at the same time. Some may need direct correction, while others may benefit more from reformulated sentences, alternative word choices, or encouragement to keep speaking.

Teacher feedback may also provide interpersonal support that AI-generated feedback cannot fully reproduce. In speaking practice, learners often respond not only to the content of feedback, but also to how it is delivered. Research on foreign language emotions shows that teacher and classroom variables are linked to learners' anxiety and positive learning experiences (Dewaele et al., 2018). In a Chinese EFL context, Yang et al. (2024) also found that teacher support was associated with learners' willingness to communicate through positive language-learning emotions. These studies suggest that teacher feedback may support learners not only through correction, but also through reassurance, encouragement, and interpersonal support.

Moreover, teacher feedback involves judgement about timing and focus. Oral corrective feedback research shows that different feedback types work differently across learners and classroom contexts (Lyster et al., 2013). In IELTS speaking practice, this judgement is particularly important because performance is assessed across fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation (IELTS, 2025a). A teacher listening to a learner speak therefore needs to attend to several aspects of performance at the same time. As a result, teacher feedback is often selective, focusing on the issue that is most urgent or most useful for the learner's next attempt rather than every possible problem.

For IELTS-focused speaking practice, teachers' decisions about feedback focus help learners connect language choices with communicative effect. A teacher can explain why an answer needs more detail, why a point is not clear to the listener, or how pronunciation affects understanding. Teacher feedback can therefore support both language development and the ability to produce clearer, more developed spoken responses.

However, teacher feedback is not always available to the extent learners need. In many classroom settings, teachers must divide limited lesson time among multiple students. This can restrict the number of speaking turns each learner receives and the amount of individualised feedback that can be provided. Zhang et al. (2024) noted this issue in a Chinese EFL speaking context, where class size and limited time made it difficult for the teacher to provide detailed oral feedback to every learner. This limitation does not reduce the pedagogical value of teacher feedback; rather, it highlights the difficulty of providing enough individualised feedback within the context of conventional group instruction.

Oral feedback is also time intensive. Useful feedback requires teachers to identify the most relevant issue, explain it clearly, and ideally give the learner a chance to act on it. Shute (2008) emphasised that formative feedback should be specific and usable, but in a group speaking lesson this level of feedback is difficult to provide for every student. If feedback is too brief, learners may not know how to improve; if it is too detailed, other students may lose speaking time. The issue, then, is not whether teacher feedback is valuable, but whether learners can receive enough of it under normal classroom conditions.

For IELTS preparation, this creates a practical problem. Learners need repeated opportunities to practise, receive targeted feedback, and revise their spoken responses. One-to-one teacher feedback can provide this support, but it is difficult to scale and may not be available to learners on a regular basis. Group classes can provide structured teaching and opportunities for speaking practice, but they may not provide enough repeated and individualised feedback for every learner. This mismatch between the feedback learners need and the feedback teachers can realistically provide creates space for examining additional feedback sources.

2.2.3 AI-Generated Feedback: Affordances and Concerns

AI-generated feedback offers a different way of organising speaking practice, but it needs to be defined carefully. In this thesis, AI feedback refers to feedback produced through Doubao during IELTS-style speaking practice. Doubao is connected to ByteDance's Seed model family, which includes general-purpose models with LLM and multimodal capabilities (ByteDance Seed, n.d.). In the AI-feedback sessions, learners used Doubao's voice function to practise spoken responses and receive generated feedback. The tool was guided by a researcher-developed prompt that instructed Doubao to take the role of an IELTS speaking tutor and provide feedback with reference to the IELTS Speaking criteria (see Chapter 6.1.2 for a more detailed discussion of how the prompt was developed). This feedback included comments on the learner's response, suggestions for improvement, and possible reformulations. It was used as a practice support tool rather than as a validated measure of IELTS Speaking performance.

This use of AI feedback can be located within a longer history of computer-assisted language learning (CALL), especially research on dialogue-based systems for language practice. Earlier

CALL research had already examined how learners practise language through computer-based dialogue systems. Bibauw et al. (2019) described dialogue-based CALL as involving different degrees of openness, interactional design, modality, and feedback. Recent large language model (LLM)-based tools extend this tradition because they can generate more open-ended responses to learner prompts, including follow-up questions, reformulations, and feedback (Kohnke et al., 2023; Law, 2024; Yang & Li, 2024). However, this flexibility also means that the quality of feedback depends on the task design, the prompt, and the learner's ability to judge the AI response critically.

The clearest benefit of AI feedback is access. Generative AI tools can respond outside scheduled lesson time, allowing learners to practise without waiting for a teacher or speaking partner. Law (2024) identified immediacy, personalisation, and learner autonomy as key advantages of generative AI in language learning. In EFL speaking research, Du and Daniel's (2024) systematic review also showed that AI-powered chatbots have been used to support speaking practice, boost confidence, improve pronunciation, and reduce anxiety. This does not mean that AI feedback is automatically better than teacher feedback. Rather, it suggests that AI may increase the number of practice-and-feedback opportunities available to learners when teacher time is limited.

Beyond access, AI feedback may also make revision easier to repeat. In IELTS Speaking preparation, AI tools may support this revision process by offering reformulations, examples, or prompts that learners can use to revise a spoken response (Kohnke et al., 2023; Law, 2024). For example, a learner may compare an original answer with an AI-suggested version and notice how the response has been extended, clarified, or reorganised. This comparison can make revision more visible, but it does not guarantee learning. The value depends on whether learners understand the feedback and try to apply it in a new response, rather than simply copying the improved version.

However, AI-generated feedback should not be taken at face value. Kohnke et al. (2023) cautioned that AI-generated responses may be inaccurate, overly general, or poorly matched to learners' needs. This is a concern in IELTS Speaking preparation because learners may not always know whether an AI suggestion is natural, appropriate, or aligned with assessment expectations. Additionally, if an AI tool provides a polished answer that is too advanced or too scripted, the learner may produce a better-looking response without developing the ability to speak more independently.

AI feedback also differs from teacher feedback in its lack of human and pedagogical judgement. A teacher can respond to a learner's hesitation, confidence, learning history, and classroom context, while AI feedback is generated from prompts and language patterns. AI may suggest fluent wording or more detailed content, but it may not know whether the learner can use that language in spontaneous speech. It may also miss signs that a learner is becoming over-reliant on prepared answers. For this reason, teacher guidance remains important when AI is used for speaking practice (Law, 2024).

These concerns have shaped the design of this thesis. In the AI-supported condition, AI feedback was not treated as a fully independent replacement for the teacher. The teacher still organised the lesson, managed the speaking tasks, and could clarify or correct AI feedback when necessary. This design reflects a cautious view of AI-supported speaking practice: AI may expand access to feedback and repetition, but teacher oversight remains important for keeping the feedback pedagogically appropriate.

Viewed together, teacher feedback and AI-generated feedback share some functions but differ in how support is provided across the four IELTS Speaking assessment criteria (IELTS, 2025a). Both can draw learners' attention to problems, suggest alternative ways of expressing ideas, and support revision of spoken responses. Teacher feedback can respond selectively across fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation through live listening and interaction. AI feedback can also support aspects such as language choice, coherence, and answer development through reformulations, examples, and repeated opportunities for revision, although support for pronunciation and other features of real-time spoken delivery may be more limited. These similarities and differences suggest that feedback source may shape not only what support learners receive, but also which aspects of speaking performance are most directly addressed during practice.

2.3 AI-Supported Speaking Practice and the Comparison Gap

As AI tools become more common in language learning, their use in speaking practice has attracted increasing attention. This section reviews recent empirical research on AI-supported speaking before identifying the gap that provides the rationale for the present study.

2.3.1 Recent Empirical Work on AI-Supported Speaking

Although generative AI has been widely discussed in language education, speaking-specific research remains more limited than broader work on text-based learning and feedback. Reviews of generative AI in second language learning show that these tools have been used for feedback, learner support, and language practice (Law, 2024; Yang & Li, 2024). Dialogue-based CALL research also provides useful background: Bibauw et al.'s (2022) meta-analysis of effectiveness studies on dialogue-based CALL found that dialogue systems can support second language development, although effects vary across tool design, learning context, and outcome measure. For this thesis, the key question is how AI tools can support spoken interaction and oral feedback, rather than whether generative AI is useful for language learning in general.

Recent reviews have focused more directly on EFL speaking. Du and Daniel's (2024) systematic review shows that AI-powered chatbots have been used in speaking practice in many areas, such as pronunciation practice, confidence building, and anxiety reduction. However, the evidence base in this area is still emerging. As generative AI tools and their uses in language learning continue to develop, findings need to be interpreted in relation to the specific tools, tasks, intervention designs,

and outcome measures examined. A careful interpretation is therefore that AI can create useful conditions for speaking practice, but its effects need to be examined in relation to specific tasks, feedback designs, and learner groups.

Empirical studies suggest that the effects of AI-supported speaking practice depend partly on what AI is compared with and how the technology is used. Fathi et al. (2024) reported improvements in speaking skills and willingness to communicate when AI-mediated interaction was compared with face-to-face peer interaction. In a different comparison, Zhang et al. (2024) compared an AI chatbot with an intelligent tutoring system and found stronger improvement in willingness to communicate in the chatbot condition. Together, these studies show that AI-supported speaking cannot be treated as a single, uniform intervention: outcomes may vary according to the comparison condition, tool design, and type of interaction. Neither study, however, directly compared AI-generated feedback with teacher feedback under matched classroom conditions.

Recent studies also differ in the outcomes they emphasise. Wang et al. (2024) focused on affective and self-perceived outcomes, including willingness to communicate, speaking anxiety, and self-perceived communicative competence, while Cha et al. (2024) examined personalised feedback for oral presentation practice, and Ding and Yusof (2025) investigated AI-powered conversation bots in an IELTS-related speaking context. This range of outcomes illustrates the growing breadth of AI-supported speaking research, but it also makes direct comparison across studies difficult. Improvements in confidence or willingness to communicate, for example, do not necessarily indicate measurable improvement in speaking performance. For the present study, this distinction is important because both score changes and learner experiences are examined rather than treated as equivalent outcomes.

Overall, the existing evidence is promising but difficult to compare directly because studies vary in their tools, comparison conditions, outcome measures, and research designs. This variation provides the basis for the more specific research gap discussed in the following section.

2.3.2 Research Gap and Rationale for the Present Study

The studies reviewed above show that AI-supported speaking practice has potential, but they do not fully answer how AI feedback compares with teacher feedback. Existing research has examined AI-supported speaking from several angles, including learner interaction, speaking performance, affective outcomes, tool design, and learner perceptions. However, direct classroom-based comparisons between AI feedback and teacher feedback remain limited, especially in IELTS-style speaking practice. Reviews of generative AI in language education also emphasise the need for teacher guidance and critical use, suggesting that AI feedback should be examined in relation to classroom teaching rather than only as an independent practice tool (Kohnke et al., 2023; Law, 2024; Yang & Li, 2024). The existing literature therefore supports the promise of AI-supported speaking practice, but it does not yet show how AI feedback works when placed beside

teacher feedback within the same teaching programme, using the same curriculum, speaking tasks, and learning context.

This distinction matters because feedback source may shape both learning outcomes and learner experience. AI feedback can make practice more available and repeatable, but it may also be inaccurate, too generic, or difficult for learners to evaluate without guidance (Kohnke et al., 2023; Law, 2024). Teacher feedback, by contrast, can draw on classroom context, learner history, and human judgement, but it is limited by class time and the number of learners who need individual attention. The key issue is therefore how AI feedback and teacher feedback may support learners in comparable or different ways when used within the same teaching programme. This question is important in test-preparation contexts, where learners need both repeated practice and feedback they can trust.

Moreover, most studies in this area do not isolate feedback source as the planned difference. If an AI group and a comparison group use different tasks, materials, practice time, or classroom structures, it becomes difficult to know whether the outcome is related to the AI tool itself or to other features of the intervention. This is a key issue for IELTS Speaking preparation, where task type, repetition, feedback timing, and teacher support can all influence performance. A stronger comparison therefore needs to keep the curriculum and speaking tasks as consistent as possible while varying the source of feedback.

This study addresses this gap by comparing AI-generated feedback and teacher feedback in an eight-session online IELTS Speaking programme for L1 Mandarin EFL learners. Both groups followed the same curriculum and IELTS-aligned speaking tasks, with feedback source as the only planned difference. The AI-supported group used Doubao, while the teacher-led group received feedback from the teacher. Doubao is included as the AI tool used in this study, but the main focus is not the platform itself. Rather, the focus is on how AI feedback, delivered through a tool accessible to the target learners, compares with teacher feedback in the same structured learning context.

This design also avoids treating AI feedback as a fully independent replacement for the teacher. In the AI-supported condition, the teacher still organised the lesson, managed the speaking tasks, and could clarify or correct AI feedback when necessary. This reflects the concerns raised in the literature: AI may expand access to feedback and repetition, but teacher oversight remains important for keeping the feedback pedagogically appropriate.

By using a mixed methods design, this thesis examines both quantitative data in the form of score change and qualitative data on learner experience. Pre- and post-programme IELTS Speaking scores were used as one measure of speaking development, while survey responses showed how learners experienced the two feedback modes. This is essential because two feedback sources may lead to similar score outcomes but different learner experiences.

This study thus offers insights into how AI and teacher feedback compare under the same learning conditions in terms of content, learner experience, and ultimately test outcomes.

Chapter Three: Methodology

This chapter explains the methodological design of the study. It first positions the researcher in relation to the study and then describes the overall research design, research setting, participants, recruitment process, programme structure, AI tool and prompt design, data collection instruments, and data analysis procedures. The chapter also discusses the researcher's role and the ethical considerations involved in conducting a small-scale online IELTS Speaking programme.

The design and delivery of the programme drew on the researcher's five years of experience teaching IELTS Speaking in face-to-face and online settings, direct experience with IELTS Speaking as a test taker who achieved band 9.0, the highest IELTS band, and regular experience using AI tools for language learning and teaching support. This experience helped the researcher design IELTS-style speaking tasks, give feedback using IELTS Speaking criteria, assess the usefulness of AI-generated feedback, and manage the AI-supported sessions in class. The researcher's dual role as teacher and researcher is discussed further in Section 3.7.

3.1 Research Design

This study compared two feedback modes in an online IELTS Speaking programme for L1 Mandarin Chinese learners of English. The final sample included 24 participants, all of whom followed the same online eight-session curriculum. Eleven participants received feedback from the teacher-researcher, while thirteen received AI-generated feedback from Doubao-Seed-2.0 (ByteDance Seed, n.d.) alongside teacher guidance in real time. The lesson content, task format, session length, and teacher-researcher were kept consistent across the two conditions, meaning that the source of feedback was the main planned difference.

A convergent mixed methods design was adopted (Creswell & Creswell, 2018; Creswell & Plano Clark, 2018) in which quantitative and qualitative data were collected during the same phase of the study, analysed separately, and then brought together at the interpretation stage. This approach was appropriate because the study was concerned not only with measurable change after the programme, but also with how participants experienced the two feedback modes and what might help explain differences or patterns in the findings. The mixed methods design therefore allowed the study to examine both performance change and learner experience.

In design terms, the study used a standard comparative experiment with a pretest–posttest structure. Participants' IELTS scores before and after the programme were compared to examine change over time, while the two feedback conditions were compared to explore whether different feedback sources were associated with different outcomes or learner experiences.

The study was delivered across two cohorts, each consisting of two groups—one teacher-led and one AI-supported.

Both cohorts followed the same overall design and used the same materials, session structure, and procedures. The second cohort was added to increase the overall sample size rather than to test a revised version of the programme. For this reason, data from both cohorts were combined for the main analysis. Three sources of data supported the design: pre- and post-programme IELTS scores, nine Likert-scale exit survey items, and three open-ended exit survey questions. These data sources are described in detail in Section 3.5.

3.2 Research Setting, Participants, and Recruitment

Research on IELTS speaking practice needs to take account of both the learning context and the learners involved. This section describes the research setting, the participants, and the process through which they were recruited.

3.2.1 Research Context

The study ran as a free supplementary online IELTS Speaking programme for adult IELTS learners whose first language was Mandarin Chinese. It was not part of a formal course, did not carry institutional credit, and was not designed to replace a full IELTS preparation course. Instead, it provided additional speaking practice aligned with IELTS Speaking preparation. The sessions used IELTS-style Speaking Part 1 and Part 2 tasks, followed the IELTS Speaking assessment criteria, and focused on feedback and immediate re-performance. The programme targeted adult IELTS learners, particularly university students or recent graduates planning to apply for master's programmes overseas.

Sessions were delivered online in small groups through Tencent Meeting (Tencent Meeting, n.d.), a video-conferencing platform developed by Tencent. Tencent Meeting was chosen because it was accessible and familiar to learners based in mainland China, where the participants were located. It was also practical for the researcher's online teaching context, as the researcher had previously used it for online classes and found its interface suitable for classroom interaction with learners in and outside mainland China.

To ensure that the one-hour sessions remained manageable, each feedback condition was planned for approximately eight participants, or 16 for each cohort. This group size was chosen because each session included a short IELTS Speaking Part 1 warm-up, followed by repeated IELTS Speaking Part 2 practice cycles. In each cycle, participants prepared a response, completed a speaking turn, received individual feedback, and then made a second attempt. If the groups had been larger, fewer participants would have had time to speak, receive feedback, and use that feedback within the same session.

Sessions were scheduled in the evening, usually between 6:30 p.m. and 8:30 p.m. China Standard Time, to accommodate participants' daytime study or work commitments.

3.2.2 Participants

The target participants were adult L1 Mandarin Chinese learners based in China who are preparing for the IELTS test. This prerequisite ensured that participants shared the same first language background and had recent IELTS test experience. The shared background was also needed to make the pre- and post-programme comparisons more meaningful, because large differences in participants' backgrounds or test history could have made the results harder to interpret (Dörnyei, 2007; Phakiti, 2014; Shadish et al., 2002).

To be eligible, participants needed to:

(a) be aged 18 or above;

(b) have Mandarin Chinese as their first language;

(c) hold an official IELTS result issued within the previous six months;

(d) plan to sit another IELTS test within one month after completing the programme;

(e) be willing and able to attend the online sessions.

Current students of the researcher were not eligible. This exclusion was made to minimise any sense of coercion or perceived obligation to participate, as students might have felt pressured to take part because of the existing teacher–student relationship. This issue is discussed further in Sections 3.7 and 3.8.

3.2.3 Recruitment

The recruitment poster was shared through WeChat (Tencent, n.d.) and Rednote/Xiaohongshu (Xiaohongshu, n.d.), two social media platforms commonly used by the target participant group in mainland China. For readers less familiar with these platforms, WeChat is a social networking platform similar to Facebook and was used to share information about the study within the researcher's personal network and relevant groups, while Rednote/Xiaohongshu functions more as a public posting and discussion platform and can be compared to Reddit or X (formerly Twitter). The poster briefly introduced the study as a free supplementary online IELTS Speaking programme, outlined the eligibility criteria and time commitment, and included a Qualtrics link/QR code for registration. Interested individuals accessed the study directly through this Qualtrics link, where they viewed the Participant Information Sheet and completed the eligibility screening and consent process. The participant information, eligibility screening, and consent materials are provided in Appendix A.

Each consenting participant was assigned a unique Study ID (in the range S000001–S999999) through the Qualtrics survey flow, which was then used to organise the research data and link information across stages of the study. Group allocation was later carried out using the Qualtrics Survey Flow Randomizer (Qualtrics, n.d.).

Participants were expected to attend all eight sessions. To ensure that the analysis was based on participants who had actually engaged with the programme, an attendance threshold of 75% (six of eight sessions) was applied; participants below this threshold were excluded from the final dataset.

In Cohort 1, recruitment ran from 5 to 10 February 2026 and 16 participants initially enrolled; 11 met the attendance criterion and were retained (6 AI-supported, 5 teacher-led). In Cohort 2, recruitment ran from 18 to 23 February 2026 and again 16 participants enrolled, of whom 13 were retained (7 AI-supported, 6 teacher-led). The final analysis dataset therefore contains 24 participants: 13 in the AI-supported condition and 11 in the teacher-led condition. The two conditions were originally allocated equal numbers (16 each); the eventual difference in group sizes reflects withdrawal for personal or scheduling reasons.

3.3 Programme Design and Procedures

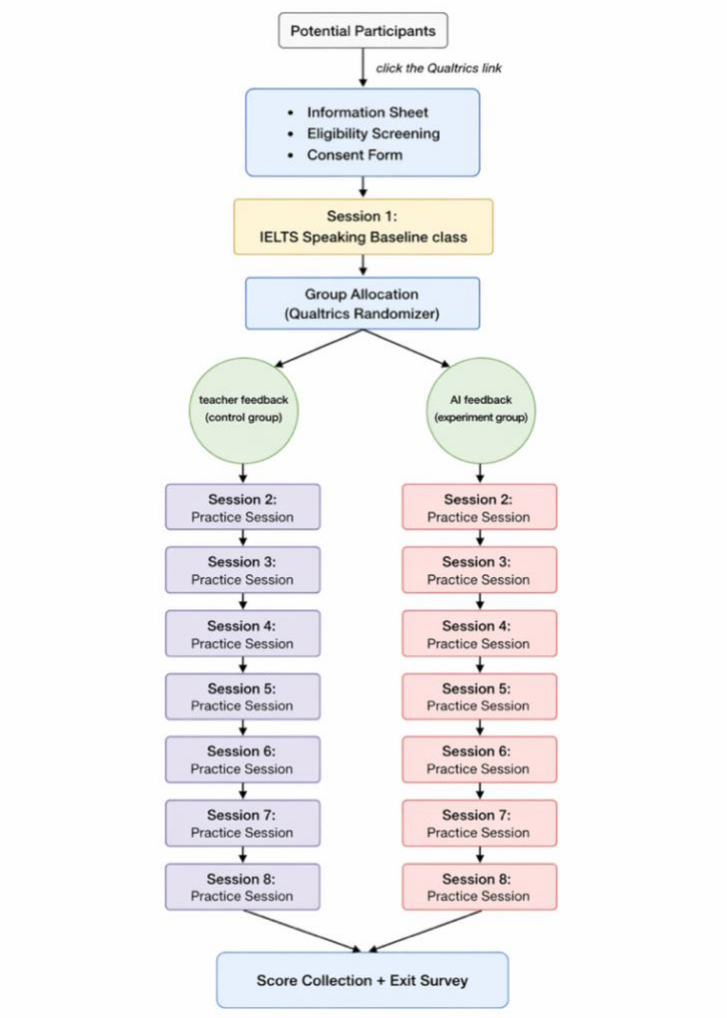
The programme was designed to provide a consistent speaking-practice experience across the two feedback conditions. This section explains the overall programme structure, group allocation, session design, and the delivery of the two cohorts.

3.3.1 Overall Programme Structure

Figure 1 summarises the overall programme structure. The intervention procedure was identical across the two cohorts. Each cohort completed one shared baseline session followed by seven practice sessions in the two separate groups— eight sessions in total, over approximately four weeks. The full sequence consisted of eligibility screening and consent, the shared baseline session, group allocation, seven practice sessions, and an exit survey.

Figure 1

Overall Programme Structure



3.3.2 Shared Baseline Session and Group Allocation

After enrolling, all participants attended Session 1 together as a shared baseline session. This session introduced the IELTS Speaking test, reviewed the four speaking assessment criteria, and outlined the structure of the following seven sessions. This process ensured that every participant entered the intervention phase with the same foundational understanding of the test. No condition-specific practice took place during this session.

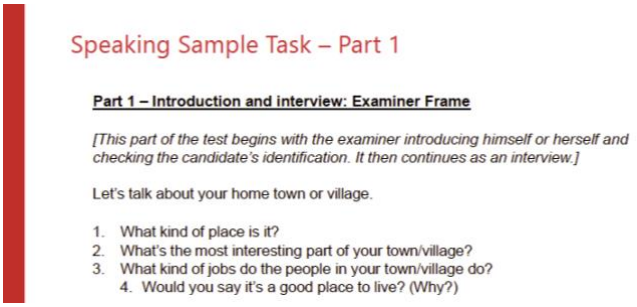
At the end of Session 1, participants entered their Study ID into Qualtrics and were randomly assigned to either the teacher-led group or the AI-supported group using the Qualtrics Randomizer. The randomizer was set to allocate participants as evenly as possible between the two conditions. After allocation, participants were notified of their assigned condition and received the Tencent Meeting link. No further teaching session took place immediately after allocation; participants used the link to attend Sessions 2–8 at the scheduled times. The group allocation instructions shared with participants are provided in Appendix B

3.3.3 Session Design

Each practice session lasted one hour. The first part was a short warm-up using IELTS Speaking Part 1 questions (see Figure 2) which lasted for 10 minutes. Participants were asked general questions on familiar topics such as home, family, work, studies, and interests (British Council, n.d.). This structure reflected a common IELTS Speaking preparation approach, where learners begin with shorter Part 1 responses before moving into longer Part 2 tasks. Unlike a full IELTS speaking course, the programme focused mainly on repeated speaking practice, feedback, and second attempts.

Figure 2

IELTS Speaking Part 1 Sample Questions



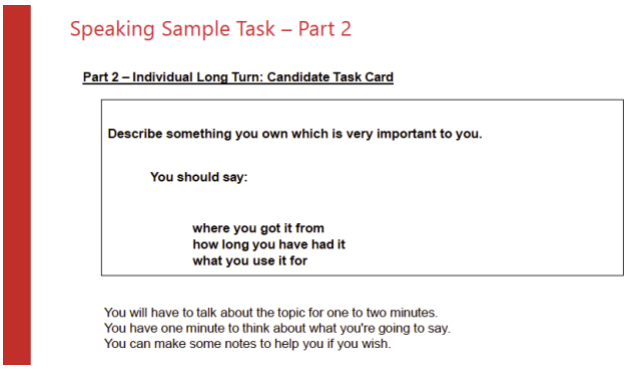
Note. From IELTS Speaking Sample Tasks (p. 3), by IELTS, 2023, https://ielts.org/cdn/ielts-sample-tests/ielts-speaking-sample-tasks-2023.pdf Copyright 2023 by IELTS.

Then the main part of the session focused on IELTS Speaking Part 2. In IELTS Speaking Part 2 (see

Figure **3**), participants were given a task card containing a particular topic and key points that they should talk about (IELTS, 2023).

Figure 3

IELTS Speaking Part 2 Sample Questions



Note. From IELTS Speaking Sample Tasks (p. 5), by IELTS, 2023, https://ielts.org/cdn/ielts-sample-tests/ielts-speaking-sample-tasks-2023.pdf Copyright 2023 by IELTS.

In the practice session, students repeated speaking cycles built around one minute of preparation for the speaking task, up to two minutes of speaking, around three minutes of feedback, and a

second attempt after the feedback. In practice, many students spoke for around 1.5 to 2 minutes, so timing was adjusted slightly where needed. The repeated cycle was designed to create immediate opportunities for noticing, revision, and re-performance within one session.

The second attempt was included so that feedback could be used immediately, not just received as a comment to review later. In each practice cycle, students first produced a response, received feedback on that specific response, and then tried the same task again with a clearer focus on what to improve. In line with the research on task repetition and feedback discussed in Chapter Two, a second attempt was included in each practice cycle to allow learners to apply feedback to the same speaking task (Wisniewski et al., 2020; Zhang et al., 2023).

The speaking activities took place in the main Tencent Meeting room rather than in breakout rooms. Participants took turns responding to the speaking tasks, and other participants in the same feedback condition could listen to both the responses and the feedback while waiting for their own turn. The only planned difference was the source of feedback. In the teacher-led condition, feedback was provided directly by the teacher. In the AI-supported condition, Doubao-Seed-2.0 generated the feedback in class. The teacher remained present to guide students in working with the AI output, clarify it where needed, manage the pace of the activity, and correct the AI output when it was unclear or inaccurate. The AI feedback process and prompt design are described in Section 3.4.

3.3.4 Cohort Delivery and Attendance

Cohort 1 attended sessions from 10 February to 12 March 2026, and Cohort 2 attended sessions from 23 February to 18 March 2026. The second cohort was added to increase the overall sample size, not to test a revised version of the programme.

Attendance was recorded across the eight sessions. Participants were expected to attend all sessions, but a 75% attendance threshold was applied for inclusion in the final dataset. Participants who attended fewer than six of the eight sessions were excluded from the analysis.

3.4 AI Tool and Prompt Design

The AI-supported condition used Doubao-Seed-2.0 to generate feedback during IELTS speaking practice. Doubao was not treated as an official IELTS scoring system or as a replacement for teacher judgement. In the AI-supported sessions, students did not work independently with Doubao at the same time. Instead, they took turns completing the speaking task within the class. After each student's response, the spoken response was submitted to Doubao, and the AI-generated feedback was discussed before the student made a second attempt. The teacher remained present throughout the sessions to guide students, explain the feedback, manage the pace of the activity, and correct the AI output when needed. Students may have had access to

Doubao outside class, but any independent use outside the scheduled sessions was not part of the intervention and was not monitored or analysed.

The choice of the AI tool was based on its practical fit with the study context. In technology-enhanced language learning, tools should be selected according to the learner group, teaching context, and learning task, rather than general capability alone (Chapelle, 2001; Hubbard, 2011). Because participants were based in mainland China, reliable access was an important requirement. ChatGPT was not selected because it was not readily accessible for this participant group and would have required additional access arrangements, such as a virtual private network (VPN). The selected tool, therefore, needed to be accessible in mainland China without a VPN. It also needed to be usable on mobile devices, fast enough for live online lessons, and suitable for generating IELTS-style speaking feedback.

Before the intervention, three platforms were tested: ChatGPT, DeepSeek, and Doubao. Preparatory testing used an earlier version of Doubao, while the AI-supported practice sessions used Doubao-Seed-2.0 through the Doubao web interface, with voice input. Each platform was tested with the same IELTS Speaking Part 2 prompts and sample learner responses. The comparison focused on whether the feedback was specific, clear, linked to IELTS Speaking criteria, useful for spoken response improvement, and workable within a one-hour Tencent Meeting class. Practical issues such as access, response speed, input mode, and ease of classroom use were also considered. Table 1 summarises this preparatory comparison. The comparison is specific to this study context and is not intended as a general ranking of the three platforms.

Table 1

Summary Comparison of AI Platforms Tested During the Preparatory Phase

Dimension	ChatGPT	DeepSeek	Doubao
Accessibility for mainland China-based learners	Limited without VPN access	Accessible	Accessible
Suitability for spoken-response practice	Text and voice: input and output available	Text-based, less suitable for speaking interaction	Text and voice: input and output available
Classroom integration	Less suitable because of access constraints	Moderately suitable	More suitable for Tencent Meeting-based sessions
Feedback structure	Strong	Less consistent for the planned workflow	Strong
Overall decision	Not selected	Not selected	Selected for the AI-supported condition

Note. AI = artificial intelligence; VPN = virtual private network. The comparison is based on preparatory testing conducted by the researcher using the same IELTS Speaking Part 2 prompts and sample responses across the three platforms.

Doubao was selected because it was the most workable option for this participant group and teaching format. It was accessible to learners in mainland China and produced feedback in a

format that could be used during live speaking practice. A usable structure meant feedback that was brief, clearly organised around IELTS Speaking criteria, linked to the learner's response, and practical enough for a second attempt in the same class. Feedback was less usable when it was too long, too general, or difficult to apply immediately.

The design of the AI feedback prompt was treated as part of the intervention design. In LLM-based tasks, prompts help define the task, guide the response format, and shape the output generated by the model (Liu et al., 2023; White et al., 2023). In this study, the prompt refers to the instruction given to Doubao to generate feedback on a learner's spoken response; it does not refer to the IELTS Speaking task itself. The prompt therefore needed to do more than ask Doubao to "give feedback". It needed to specify the AI's role, the IELTS Speaking focus, the feedback structure, and the kind of language support that would be useful during a live speaking lesson. This was also important because educational uses of LLMs require a clear pedagogical purpose and human oversight (Kasneci et al., 2023; Yan et al., 2024).

The AI feedback prompt was developed by the researcher and Zhang Yu, an experienced IELTS teacher. It was tested and revised using IELTS Speaking Part 2 mock-test tasks and sample learner responses. During testing, the output was checked for whether it was specific to the learner's response, aligned with IELTS Speaking criteria, clearly organised, natural in spoken English, and short enough to use within the feedback time available in class. The prompt was revised when the output was too general, too long, or not useful for a second speaking attempt. The full AI feedback prompt used in the study is provided in Appendix D.

The final prompt asked Doubao to act as a supportive IELTS speaking coach and provide feedback in a consistent format. It instructed the AI to consider all four IELTS Speaking criteria while prioritising the most important issues in the learner's response. The output included two strengths, two grammar-focused issues, two spoken rewrites, and two follow-up questions, supported by short quotations from the learner's response. It did not provide an overall band score.

In the AI-supported sessions, the feedback process followed the same speaking cycle as the teacher-led condition. After a learner completed a speaking task, they used Doubao's voice input function to submit their spoken response. Doubao then generated feedback using the prepared prompt. The teacher checked that the AI feedback responded to the learner's specific speaking attempt rather than offering generic advice. The teacher then helped students focus on the most useful parts of the AI output, clarified vague wording, and corrected the feedback if it was inaccurate. Students then used the feedback to produce a second response to the same task. In this way, AI supported part of the feedback process while the teacher remained responsible for the lesson and for the pedagogical use of the feedback.

3.5 Data Collection Instruments and Sources

The study drew on three data sources: pre- and post-programme IELTS scores, Likert-scale exit survey responses, and open-ended exit survey comments. Each source addressed a different part of the research aim. IELTS scores provided a standardised, externally assessed measure of English language performance before and after the programme. Likert-scale responses captured participants' perceptions of feedback usefulness, confidence, clarity, and engagement in a structured form that allowed comparison between groups. Open-ended responses added explanatory detail by allowing participants to describe, in their own words, which aspects of the programme supported their speaking and what remained limited.

3.5.1 IELTS Score Data

The first source was participants' IELTS scores from before and after the programme. IELTS is an internationally recognised English language proficiency test with well-documented validity and reliability as an outcome measure in applied linguistics research (IELTS, n.d.–c). After completing the eight sessions, sitting another IELTS test, and receiving their results, participants reported both their earlier and their new scores in the exit survey.

To support the accuracy and credibility of the score data, participants were also required to provide evidence of their IELTS results, such as an official Test Report Form or an electronic screenshot. The pre- and post-programme Speaking scores were obtained from participants' official IELTS test results and therefore reflect the assessment of an IELTS examiner, not the researcher. IELTS maintains the reliability of examiner-rated sections through examiner training and certification, retraining and recertification every two years, and ongoing face-to-face and distance monitoring. For 2024–2025 performances, IELTS reported an inter-rater reliability coefficient of .90 for Speaking based on comparisons between original ratings and subsequent re-ratings (IELTS, n.d.–b). The IELTS examiners were external to the present study, and the researcher did not provide them with information about participants' group allocation. No additional study-specific inter-rater reliability check was conducted because the researcher did not have access to recordings of participants' official IELTS Speaking tests for independent rescoring.

This score data produced pre- and post-test values for the overall band score and for each of the four section scores: Speaking, Listening, Reading, and Writing. The Speaking score served as the primary outcome of the study; the overall band and the other three section scores were analysed as supplementary indicators.

3.5.2 Likert-Scale Exit Survey Items

The second data source was the set of nine Likert-scale items included in the post-intervention exit survey, which was delivered through Qualtrics. Each item was rated on a five-point scale, from *Strongly Disagree* (1) to *Strongly Agree* (5). This format is widely used in applied linguistics survey

research (Dörnyei, 2007). Following Boone and Boone (2012), the Likert-scale items were treated as ordinal data and reported using frequencies, percentages, medians, and interquartile ranges rather than as continuous data. The implications of this decision for the inferential analysis are addressed in Section 3.6.1.

The items focused on aspects of learner experience that IELTS scores could not capture. They covered two broad areas: perceived speaking development and feedback clarity, including clarity of expression, directness, confidence, feedback usefulness, and awareness of areas for improvement; and programme experience and engagement, including session structure, workload and pace, comfort in the online format, and willingness to recommend this type of speaking practice. These groupings were used only to organise the reporting of results; no composite Likert-scale scores were calculated.

The nine Likert-scale items were developed specifically for this study rather than adopted from an existing validated instrument. Their content was informed by the research questions and intervention design, and the items were discussed during development with Zhang Yu, an EFL educator with postgraduate study in linguistics, to review their relevance and clarity. The survey was not formally piloted as a separate instrument before the study. To support clarity for L1 Mandarin participants, the participant survey was provided bilingually in English and Chinese, with attention to maintaining consistent meaning across the two language versions.

3.5.3 Open-Ended Exit Survey Responses

The third source was participants' responses to three open-ended exit survey questions. These questions asked what participants found most helpful for their speaking development, what they thought could be improved, and whether they had any other comments. The open-ended responses were included because closed survey items could show how participants rated the programme but could not fully explain the reasons behind those ratings. They allowed participants to describe the feedback and learning experience in their own words, including what they found useful, limited, or difficult. These responses therefore provided qualitative detail to help interpret the IELTS score changes and Likert-scale patterns, and formed the basis of the qualitative analysis described in Section 3.6.2 (Creswell & Plano Clark, 2018; Dörnyei, 2007). Both the Likert-scale items and open-ended responses were self-reported data and were used to represent participants' perceptions of their experience rather than as independent measures of learning outcomes.

3.6 Data Analysis

This section explains how the three data sources were analysed. The IELTS score data and Likert-scale survey items were analysed quantitatively, while the open-ended exit survey responses were analysed qualitatively. The two strands were then brought together at the interpretation stage, following the convergent mixed methods design described in Section 3.1 (Creswell & Creswell, 2018; Creswell & Plano Clark, 2018).

3.6.1 Quantitative Analysis

The quantitative analysis examined two forms of data: participants' IELTS score change from pre- to post-programme, and their responses to the nine Likert-scale exit survey items. Because the two cohorts followed the same study design, materials, and procedures, their data were pooled for the main analysis. This produced a single dataset of 24 participants: 13 in the AI-supported condition and 11 in the teacher-led condition.

The choice of statistical procedures was shaped by the size and structure of the dataset. First, the sample was small, which limited statistical power and made individual outliers more influential (Larson-Hall, 2016; Mackey & Gass, 2022). Second, the score and survey data had limited scale ranges. IELTS band scores are reported in fixed 0.5-band increments rather than as continuous raw scores, and the Likert-scale items used a five-point ordinal response format. For this reason, medians and interquartile ranges were reported alongside means and standard deviations for the IELTS score data, while Likert-scale responses were summarised using frequencies, percentages, medians, and interquartile ranges. Third, Shapiro–Wilk tests were used to assess whether the score-change variables met the normality assumption required for parametric tests. Overall, the main quantitative analysis used descriptive statistics and non-parametric inferential tests. Given the small sample, statistical significance was therefore interpreted cautiously and considered alongside effect sizes and descriptive patterns, rather than being treated on its own as evidence for or against a meaningful group difference.

Descriptive statistics were calculated for the pre- and post-programme IELTS scores in each group. Baseline IELTS scores were summarised using means and standard deviations. Pre- and post-programme scores were summarised using medians, while gain scores were reported separately as the median of participants' individual score changes. These summaries were presented for the overall IELTS band score and each of the four section scores: Speaking, Listening, Reading, and Writing. Baseline equivalence between the two groups was checked at pre-test using Mann–Whitney U tests, so that any initial group differences could be considered when interpreting the post-programme results.

Within-group change from pre- to post-programme was examined using Wilcoxon signed-rank tests, run separately for the AI-supported and teacher-led groups. Between-group differences in gain scores, calculated as post-test score minus pre-test score, were examined using Mann–Whitney U tests. Effect sizes were reported as matched rank-biserial correlations for Wilcoxon signed-rank tests and rank-biserial correlations for Mann–Whitney U tests, including baseline, gain-score, and Likert item-level comparisons. Effect sizes were reported as absolute values when interpreting their magnitude. These were interpreted using Plonsky and Oswald's (2014) field-specific benchmarks for second-language research, where $r \approx .25$ is small, $\approx .40$ is medium, and $\geq .60$ is large.

The Likert-scale items were summarised using counts, percentages, medians, and interquartile ranges. Between-group differences were examined item by item using Mann–Whitney U tests. Because running nine separate tests increases the chance of at least one statistically significant finding occurring by chance, Section 4.3.2 reports the uncorrected p-values, rank-biserial correlations, and whether the results remained significant after applying a Holm–Bonferroni correction. Following Plonsky's (2013) recommendations for quantitative reporting in second-language research, descriptive statistics, inferential test results, and effect sizes are reported together throughout Chapter 4. The primary quantitative analyses were carried out in JASP (JASP Team, 2026), with rank-biserial correlations calculated from the corresponding Mann–Whitney U statistics where required.

3.6.2 Qualitative Analysis

The qualitative analysis focused on participants' responses to the three open-ended exit survey questions. Because these data consisted of short written comments rather than interview transcripts, the analysis used a descriptive thematic categorisation. The aim was to identify the main categories of learner experience reported by participants and to use representative comments to help explain the quantitative results. This approach was informed by thematic analysis but was applied in a descriptive way to fit the short survey-response format (Braun & Clarke, 2006; Dörnyei, 2007).

The open-ended responses from all 24 participants were exported from Qualtrics and compiled into an Excel spreadsheet. The participant-facing exit survey was bilingual, and participants could answer the open-ended questions in either Chinese or English. All 24 participants provided their responses in Mandarin Chinese. These responses were initially translated into English using Google Translate (Google, n.d.) and then checked and revised by the researcher against the original Mandarin responses. Participant quotations are presented in English translation in the findings chapter, while the original Mandarin responses were retained in the research dataset for checking and interpretation.

The researcher read through the responses several times and grouped comments according to their main meaning. For example, comments about pronunciation correction, answer development, confidence, repeated practice, feedback usefulness, workload, or requests for more practice time were grouped into separate categories. The number of comments in each category was then used to show how commonly each issue appeared across participants' responses. These counts were used descriptively only; they were not treated as statistical results.

Representative quotations were selected to illustrate each category and to show how participants described their learning experience in their own words. The qualitative analysis therefore combined simple category counts with participant comments. This made it possible to show not only what

patterns appeared in the open-ended responses, but also how participants explained the usefulness and limitations of each feedback condition.

To support consistency in the qualitative categorisation, the researcher returned repeatedly to the original responses when reviewing category boundaries and retained the original Mandarin responses alongside the English translations for checking. Category groupings and counts were recorded in the working spreadsheet so that analytical decisions could be revisited during the analysis. The developing categories and interpretations were also discussed in supervision meetings, which provided an additional opportunity to question and refine the analysis. However, no formal member checking, independent second coding, or structured peer debriefing was conducted.

3.6.3 Integration of Quantitative and Qualitative Findings

The quantitative and qualitative findings were brought together at the interpretation stage, in line with the convergent mixed methods design (Creswell & Creswell, 2018). In this study, integration meant reading the IELTS score changes, Likert-scale responses, and open-ended comments together rather than treating them as three separate sets of results. This follows the view that mixed methods integration should help develop a more complete interpretation of the research problem, rather than simply placing quantitative and qualitative findings side by side (Fetters et al., 2013).

The integration focused on what each data source added to the interpretation. IELTS scores showed whether participants' test performance changed after the programme. Likert-scale responses showed how participants rated different aspects of the programme, such as feedback usefulness, confidence, clarity, workload, and comfort in the online format. Open-ended comments were then used to explain these patterns in more detail by showing what participants found helpful or limited in each feedback condition.

By combining these findings, the study was able to move beyond a simple comparison of score change. The integrated analysis helped identify how the AI-supported and teacher-led conditions appeared to support speaking practice, where participants' experiences were similar, and where the two feedback modes seemed to offer different kinds of support.

3.7 Researcher Role

The overall study design was developed by the researcher in consultation with the supervisor, and the researcher delivered all teaching sessions across both conditions and both cohorts.

The researcher's professional background informed the design and delivery of the programme. The researcher had experience teaching English in both face-to-face and online settings, as well as direct experience with IELTS Speaking as both a teacher and a high-level test taker. This background supported the design of IELTS-style speaking tasks, the interpretation of feedback

quality, and the decision to keep the AI-supported condition teacher-mediated rather than fully independent.

This dual role as both teacher and researcher required reflexive attention. The researcher was not an outside observer, but an active participant in the learning environment being studied. To reduce variation between conditions, the same lesson structure, task materials, session timing, and overall teaching procedures were used across both groups and both cohorts. Using the same teacher-researcher across all sessions also helped keep instructional voice, pacing, and classroom management consistent. However, this did not mean that the two conditions were identical in every interaction. Small differences in explanation, emphasis, or classroom response may have occurred.

The researcher's direct involvement also provided practical insight into how the two formats operated from the teacher's perspective. Delivering both conditions made it possible to observe how teacher feedback and AI-supported feedback differed in classroom management, feedback delivery, pacing, and the amount of teacher intervention required. This was relevant to the study's interest in the feasibility of using AI-supported feedback in structured speaking practice. These observations were not treated as a separate data source. The researcher's practitioner perspective is therefore used only as contextual background for interpreting the study, especially when discussing the classroom use of AI-supported feedback.

3.8 Ethical Considerations

Ethical approval for the study was granted by the Auckland University of Technology Ethics Committee (AUTEC Reference 25/429) on 5 February 2026. Several aspects of the study design were adjusted during the ethics review process to reduce possible risks to participants.

One key ethical issue concerned the relationship between the researcher and potential participants. The original plan had allowed for the possibility of recruiting from the researcher's current students, but this was changed because current students might have felt implicit pressure to participate or to remain in the study. Current students of the researcher were therefore excluded, and recruitment was limited to learners with no existing relationship with the researcher.

Participant contact was also carefully managed. The researcher did not collect participants' email addresses or social media contact details. Instead, the Qualtrics study link was posted publicly on WeChat and Xiaohongshu, and all scheduling information and session links were delivered through the Qualtrics flow after consent. The exit survey link was shared at the end of the final session through Tencent Meeting, with a reminder posted publicly on the same social media platforms. Participants who wanted access to the findings were informed that a summary would be made available through the supervisor's website.

The use of AI also required ethical consideration. The AI-supported sessions were conducted under active teacher supervision rather than leaving students to interact with the tool

independently. The researcher was present throughout all AI-supported sessions, guided students in how to use and interpret the AI output and corrected or clarified the feedback when necessary. This meant that participants were not expected to act on AI-generated feedback without teacher support.

Participation was voluntary, and all participants completed the consent process before joining any sessions. Information about the study was provided through an online Participant Information Sheet and a digital consent form, both hosted in Qualtrics under the researcher's AUT account. These materials were available in both English and Chinese to reduce the risk of language-related misunderstanding. Only participants who gave electronic consent were enrolled. Participants were informed that they could withdraw without giving a reason at any time before their data were de-identified, and that they could skip any open-ended survey question they did not wish to answer.

The study involved limited disclosure during participation. At the consent stage, participants were told that the programme included two practice conditions and that allocation would be random. After allocation, each participant was informed only of their own assigned condition; details about the activities or format of the other condition were not shared during the programme. This was intended to reduce expectancy effects related to between-group comparison. After the final session and completion of the exit survey, participants were debriefed about the purpose of the comparison and the two feedback approaches used in the study.

Confidentiality was managed with the use of Study IDs. After consent, each participant was assigned a unique Study ID through Qualtrics, which was used to label all study data and link information across the different stages of the project. The analytic dataset used for results and reporting did not rely on real names, and identifiable records, such as consent records, were stored separately from the main research dataset.

Because the study was delivered in small online groups through Tencent Meeting, confidentiality among participants was necessarily limited, as participants could see and hear one another during class. This limitation was explained in the Participant Information Sheet. To support privacy, participants were asked to use their Study ID in class activities, and no full names or IELTS scores were shared during the sessions. In the AI-supported condition, participants were also instructed not to enter unnecessary personal identifiers into the AI tool.

Research data were collected through Qualtrics and stored on secure AUT systems. Access was limited to the researcher and the supervisor. All data were stored on AUT OneDrive/Teams, and consent forms were stored separately. Files will be retained for six years and then permanently deleted from AUT storage and Qualtrics.

3.9 Chapter Summary

This chapter has explained how the study was designed and conducted. The methodology was designed to compare AI-supported and teacher-led feedback within the same structured IELTS

Speaking programme while also examining participants' experiences of the two feedback modes. A convergent mixed methods design was used so that IELTS score changes, Likert-scale responses, and open-ended comments could be interpreted together. The chapter also outlined the practical and ethical decisions made to support a small-scale online speaking programme, including random allocation, consistent lesson procedures, teacher mediation of AI feedback, participant confidentiality, and the researcher's dual role as teacher and researcher. The next chapter presents the findings.

Chapter Four: Findings

This chapter presents the findings of the four-week, two-cohort study comparing two forms of online IELTS speaking feedback: an AI-supported group (n = 13) and a teacher-led group (n = 11). It first describes the participant groups and their baseline IELTS scores, before reporting pre- to post-test changes in IELTS performance. The chapter then presents participants' survey responses and open-ended comments, followed by an integration of the quantitative and qualitative findings.

As outlined in the Methodology (see Chapter Three), the analysis focused mainly on non-parametric statistics because of the small sample size and the ordinal nature of IELTS band scores. Wilcoxon signed-rank tests were used to examine pre- to post-test changes within each group, and Mann–Whitney U tests were used to compare gain scores between the two groups. Descriptive statistics, survey patterns, and qualitative themes are presented together to show how the two forms of speaking practice were experienced by participants.

4.1 Participant Overview

In total, 32 participants initially enrolled across the two cohorts. After attendance and data screening, 24 participants completed the four-week intervention and were included in the final analysis. As shown in Table 2, Cohort 1 included 11 valid participants, with 6 in the AI-supported group and 5 in the teacher-led group. Cohort 2 included 13 valid participants, with 7 in the AI-supported group and 6 in the teacher-led group. Across both cohorts, the final sample therefore consisted of 13 participants in the AI-supported group and 11 participants in the teacher-led group.

Table 2

Cohort Composition and Size

	Initially enrolled	Final included	AI-supported group	Teacher-led group
Cohort 1	16	11	6	5
Cohort 2	16	13	7	6
Total	32	24	13	11

Note. Only those participants were included who attended the required sessions, completed the exit survey, and provided usable pre- and post-test data.

All participants were Mandarin Chinese L1 speakers preparing for the IELTS exam. They all held a recent official IELTS result before the intervention and planned to retest within one month after the programme. Baseline IELTS scores were similar across the two groups, as shown in Table 3.

Mann–Whitney U tests found no significant pre-test differences between the AI-supported and teacher-led groups for the Overall, Speaking, Listening, Reading, or Writing scores (all $p > .25$). The corresponding rank-biserial correlations ranged from $r = .00$ to $r = .28$, indicating generally

small baseline differences between the two groups. This baseline similarity reduces the likelihood that later group differences were due to pre-existing proficiency differences.

Table 3

Baseline IELTS Scores by Group

IELTS score	AI-supported group M (SD)	Teacher-led group M (SD)	U	p	r
Overall	6.15 (0.72)	6.05 (0.61)	79.5	.655	.11
Speaking	5.73 (0.81)	5.64 (0.67)	77.0	.767	.08
Listening	6.46 (0.92)	6.05 (0.76)	91.5	.252	.28
Reading	6.46 (1.14)	6.36 (0.95)	72.5	.976	.01
Writing	5.77 (0.56)	5.77 (0.47)	71.5	1.000	.00

Note. p values were calculated from Mann–Whitney U tests comparing the two groups at pre-test (r = rank-biserial correlation, reported as an absolute value. M = mean; SD = standard deviation). No comparison reached statistical significance.

4.2 IELTS Score Changes

This section reports changes in IELTS scores from pre-test to post-test. The results are presented in four parts: descriptive patterns of change, within-group pre-to-post-test comparisons, between-group comparisons on gain scores, and patterns across the individual IELTS skills.

4.2.1 Descriptive Pattern of Change

Table 4 summarises the median pre-test and post-test scores for each IELTS outcome together with the difference between the pre- and post-test medians for each group. Median scores are emphasised because IELTS band scores move in fixed 0.5 increments and are therefore better treated as ordinal data.

Speaking was the main focus of the intervention. The AI-supported group's median Speaking score increased from 6.00 to 6.50, while the teacher-led group's median remained at 6.00. Overall median scores were unchanged in the AI-supported group and increased from 6.00 to 6.50 in the teacher-led group. The largest descriptive change was in Listening, where the AI-supported group's median increased from 6.50 to 7.00 and the teacher-led group's median increased from 6.00 to 7.00.

Table 4

Descriptive Summary of Pre- to Post-Test IELTS Score Changes

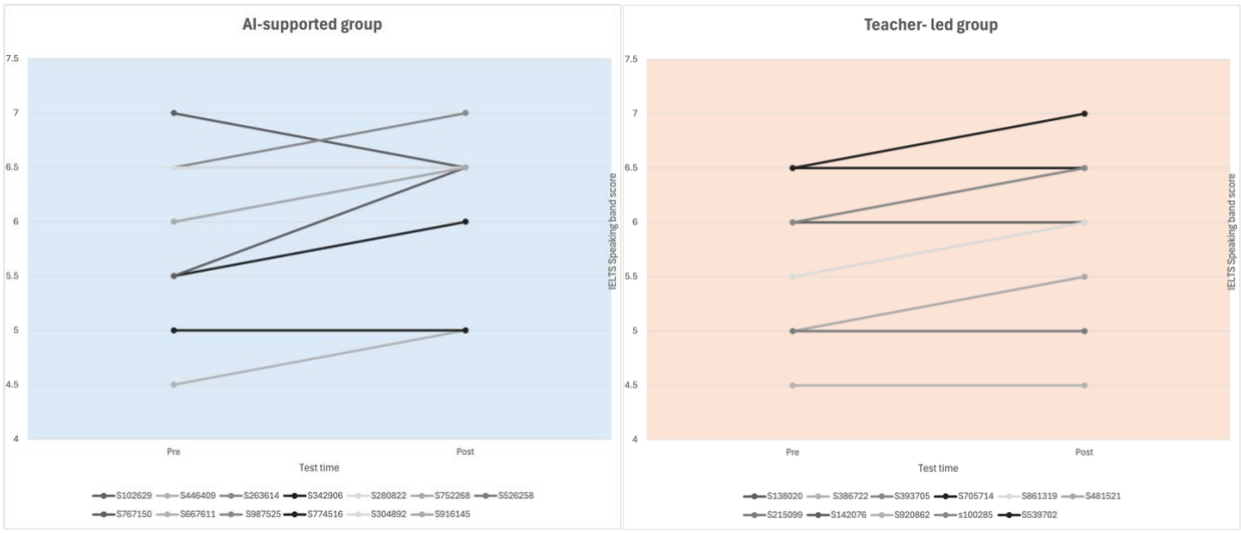
IELTS score	AI pre-IELTS Mdn	AI post-IELTS Mdn	AI difference in medians	Teacher pre-IELTS Mdn	Teacher post-IELTS Mdn	Teacher difference in medians
Overall	6.50	6.50	0.00	6.00	6.50	0.50
Speaking	6.00	6.50	0.50	6.00	6.00	0.00
Listening	6.50	7.00	0.50	6.00	7.00	1.00
Reading	6.50	6.50	0.00	6.50	6.00	-0.50
Writing	6.00	6.00	0.00	6.00	6.00	0.00

Note. Difference in medians = post-test group median minus pre-test group median (Mdn = median).

Figure 4 shows the individual pre- to post-test trajectories for IELTS Speaking. This figure is included because Speaking was the targeted skill in the intervention. The figure shows that most participants in both groups either improved or maintained their speaking score, with no clear visual evidence that one group improved more than the other.

Figure 4

Pre- and Post-Test IELTS Speaking Scores by Group



Note. Each line represents one participant's pre- to post-test speaking score. Higher scores indicate higher IELTS Speaking performance.

4.2.2 Within-Group Comparisons

Wilcoxon signed-rank tests were used to examine whether each group made statistically significant gains from pre-test to post-test. As the data presented in Table 5 suggest, significant gains were found in the AI-supported group for Speaking (W = 4.50, z = −2.13, p = .025, r = .80) and Overall band score (W = 0.00, z = −2.20, p = .020, r = 1.00). No significant changes were found for

Listening, Reading, or Writing. The Writing result showed a positive direction of change, but this did not reach statistical significance. The effect sizes for the significant Speaking and Overall gains were large.

In the teacher-led group, significant gains were found for Speaking (W = 0.00, z = −2.20, p = .020, r = 1.00), Listening (W = 0.00, z = −2.37, p = .021, r = 1.00), and Overall band score (W = 0.00, z = −2.02, p = .037, r = 1.00). No significant changes were found for Reading or Writing. The effect sizes for the significant Speaking, Listening, and Overall gains were also large.

The most important finding from the within-group tests is that both groups improved significantly in Speaking, the skill directly targeted by the intervention, with large effect sizes in both groups. The teacher-led group also showed a significant improvement in Listening with a large effect size.

Table 5

Within-Group Pre- to Post-Test Comparisons by Group

IELTS score	AI W	AI z	AI p	AI r	Teacher W	Teacher z	Teacher p	Teacher r
Overall	0.00	−2.20	.020*	1.00	0.00	−2.02	.037*	1.00
Speaking	4.50	−2.13	.025*	0.80	0.00	−2.20	.020*	1.00
Listening	13.00	−1.13	.276	0.42	0.00	−2.37	.021*	1.00
Reading	16.50	−0.71	.501	0.27	30.00	0.26	.824	0.09
Writing	7.50	−1.47	.143	0.58	3.50	−1.47	.129	0.67

Note. Wilcoxon signed-rank tests were used to compare pre-test and post-test scores within each group (r = matched rank biserial correlation, reported as an absolute value). *p < .05.

4.2.3 Between-Group Comparisons

The next analysis examined whether one group improved more than the other. Gain scores were calculated by subtracting each participant's pre-test score from their post-test score. A positive gain score therefore indicates improvement, a score of zero indicates no change, and a negative gain score indicates a lower post-test score. Mann–Whitney U tests were used to compare the gain scores of the AI-supported and teacher-led groups across Overall, Speaking, Listening, Reading, and Writing. The results are shown in Table 6 and visualised in Figure 5.

None of the between-group comparisons reached statistical significance, indicating that no significant difference in gain scores was detected between the two conditions across the five IELTS outcomes. For Speaking, both groups had the same median gain of 0.50 (U = 77.0, p = .744, r = .08). The corresponding rank-biserial correlation was very small, indicating that the magnitude of the between-group difference in Speaking gain was minimal. Across the five IELTS outcomes, absolute rank-biserial correlations ranged from r = .01 to r = .28, indicating generally small between-group differences.

Figure 5 provides a visual summary of the gain score distributions. The boxplots show that IELTS Speaking gains were centred above zero in both groups, while Overall scores changed little. Some variation was visible in the other skills. For example, the teacher-led group had a higher median Listening gain than the AI-supported group, while the AI-supported group had a higher median Reading gain. However, these differences were not statistically significant and should therefore be interpreted with caution.

In short, over the four-week intervention, both groups improved in IELTS Speaking, but the between-group analysis did not detect a significant difference in speaking gain between the AI-supported and teacher-led groups, and the corresponding effect size was very small.

Table 6

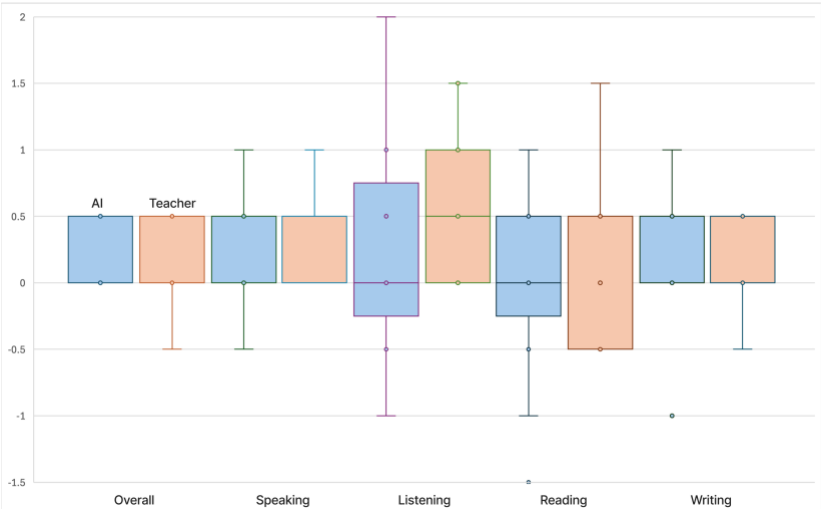
Between-Group Comparisons on IELTS Gain Scores

Variable	AI median gain	Teacher median gain	U	z	p	r
Overall	0.00	0.00	72.0	+0.03	1.000	0.01
Speaking	0.50	0.50	77.0	+0.32	.744	0.08
Listening	0.00	0.50	51.5	−1.16	.245	0.28
Reading	0.50	−0.50	87.0	+0.90	.366	0.22
Writing	0.50	0.00	79.5	+0.46	.634	0.11

Note. (Gain = post-test score minus pre-test score. r = rank-biserial correlation, reported as an absolute value) No comparison reached statistical significance.

Figure 5

IELTS Gain Scores by Skill and Group



Note. Gain scores were calculated as post-intervention score minus pre-intervention score. Scores above zero indicate improvement, while scores below zero indicate a decrease. Boxes show the interquartile range, the horizontal line shows the median, and whiskers show the spread of scores.

4.2.4 Patterns Across Individual IELTS Skills

Across the individual IELTS skills, the clearest pattern was found for Speaking. Both groups made significant Speaking gains, and the between-group comparison showed no significant difference in Speaking gain. This indicates that both AI-supported and teacher-led practice were associated with improvement in the skill directly targeted by the intervention.

Listening showed a different pattern. The teacher-led group made a significant Listening gain, while the AI-supported group did not. Reading and Writing did not show significant change in either group, which is consistent with the focus of the intervention on speaking practice.

4.3 Likert Survey Responses

After the intervention, participants completed a nine-item Likert-scale survey about their experience of the speaking course. Responses were rated on a five-point scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). The items covered perceived improvement in Speaking, confidence, feedback usefulness, clarity about what to improve, the structure and pace of the sessions, comfort during practice, and willingness to recommend the programme. As Likert responses are ordinal, the results are reported using medians and the percentage of participants who selected *Agree* or *Strongly Agree* as an indication of overall agreement with each item.

4.3.1 Agree and Strongly Agree Responses

Table 7 presents the full Likert survey items and the abbreviated labels used in Figure 6. The items are listed in the same order as they appear in the figure.

Table 7

Likert Survey Items and Abbreviated Labels Used in Chapter 4.3

Abbreviated label in Chapter 4.3	Full Likert survey item
Recommend	I would recommend this type of speaking practice to other IELTS learners.
Comfortable	I felt comfortable participating in the online sessions.
Reasonable pace	The pace and workload were reasonable for me.
Well-structured	The practice activities were well structured and easy to follow.
Know what to improve	After each session, I understood what I needed to improve.
Feedback useful	The feedback I received was useful for improving my speaking.
Confidence	I feel more confident speaking in IELTS-style tasks.
Direct/specific	I became more direct and specific when expressing opinions or making requests in English.
Express clearly	The sessions helped me express my ideas more clearly in English.

Note. Participants responded to each item on a five-point Likert scale, from 1 = Strongly Disagree to 5 = Strongly Agree. Figure 6 uses abbreviated labels for readability.

Overall, both groups responded positively to the intervention. As shown in Figure 6 and Table 8, the percentage of participants selecting *Agree* or *Strongly Agree* was high across most items in

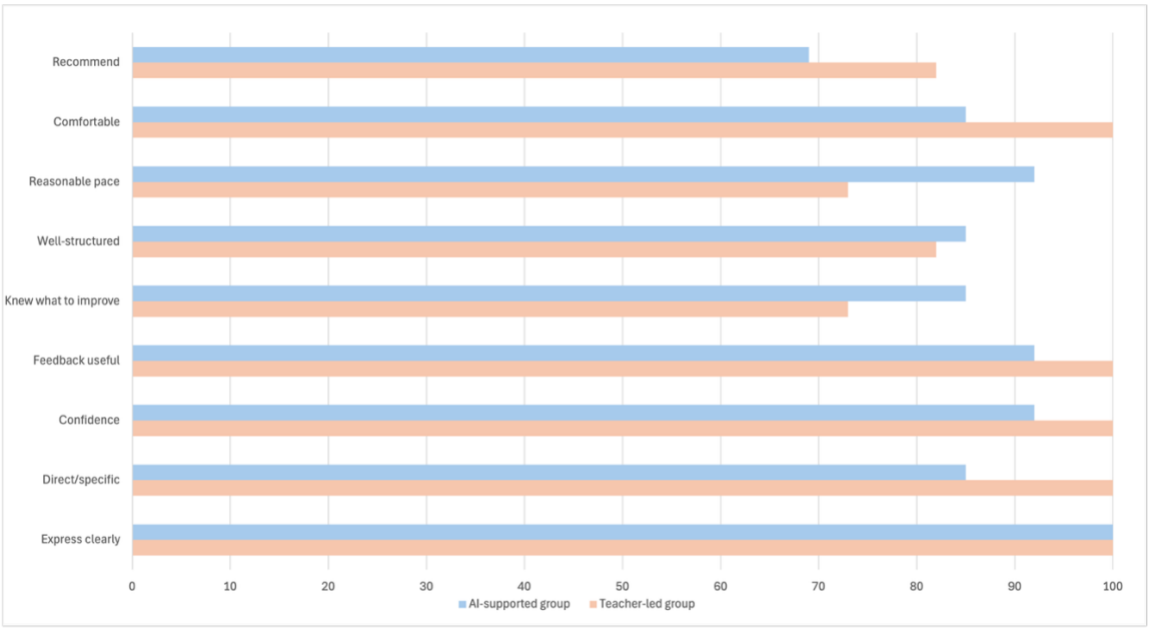
both groups. Because responses were generally positive, Figure 6 focuses on *Agree* and *Strongly Agree* percentages rather than showing the full response distribution. This provides a clearer visual comparison between the AI-supported and teacher-led groups.

In the AI-supported group, the highest-rated item was “Express clearly”, with 100% of participants selecting *Agree* or *Strongly Agree*. This was followed by “Confidence”, “Feedback useful”, and “Reasonable pace”, each at 92%. In the teacher-led group, several items received 100% agreement, including “Express clearly”, “Direct/specific”, “Confidence”, “Feedback useful”, and “Comfortable”.

Some differences were also visible. The AI-supported group gave higher ratings for “Reasonable pace” (92%) than the teacher-led group (73%). The teacher-led group gave higher ratings for “Direct/Specific”, “Feedback useful”, “Comfortable”, and “Recommend”.

Figure 6

Percentage of Participants Who Selected Agree or Strongly Agree by Category and Group



Note. Percentages indicate the proportion of participants selecting 4 (Agree) or 5 (Strongly Agree) on each Likert item. Item labels are abbreviated for readability and are defined in Table 7.

4.3.2 Item-Level Group Comparisons

Mann–Whitney U tests were used to compare the AI-supported and teacher-led groups on each survey item. The results are shown in Table 8. Only one item showed a statistically significant group difference at the uncorrected .05 level, namely that the course was presented at a “Reasonable pace” ($U = 101.5$, $p = .040$, $r = .42$). The AI-supported group reported higher agreement on this item (92% Agree or Strongly Agree) than the teacher-led group (73%). The corresponding rank-biserial correlation indicated a medium-sized difference, suggesting that AI-supported participants may have perceived the pace and workload as more manageable.

This result should be interpreted cautiously. Because nine item-level comparisons were conducted, a Holm-Bonferroni correction was applied to account for multiple testing. After correction, the difference noted for the “Reasonable pace” item no longer remained statistically significant. Therefore, despite the medium effect size, it is treated as a tentative pattern rather than a confirmed group difference.

No statistically significant differences were found for the other eight items. Across the nine items, absolute rank-biserial correlations ranged from r = .00 to r = .42, with “Reasonable pace” showing the largest effect. Overall, the Likert responses suggest that participants in both groups reported similar experiences of the speaking practice and that these experiences were broadly positive. The possible difference in perceived pace is examined further in the open-ended responses in Section 4.4.

Table 8

Likert Survey Responses by Group

Item	AI Mdn	AI Agree%	Teacher Mdn	Teacher Agree%	U	p	r
Recommend	4	69%	4	82%	67.0	.805	.06
Comfortable	4	85%	4	100%	66.0	.735	.08
Reasonable pace	4	92%	4	73%	101.5	.040*	.42
Well-structured	4	85%	5	82%	54.0	.285	.24
Know what to improve	4	85%	4	73%	79.5	.636	.11
Feedback useful	4	92%	5	100%	52.0	.210	.27
Confidence	5	92%	4	100%	80.0	.597	.12
Direct/specific	4	85%	4	100%	71.5	1.000	.00
Express clearly	5	100%	4	100%	83.0	.461	.16

Note. Mdn = median rating. Agree% = percentage of participants selecting 4 (Agree) or 5 (Strongly Agree). *p* values are from Mann–Whitney U tests comparing the two groups. *r* = rank-biserial correlation, reported as an absolute value. *p* < .05 before correction for multiple comparisons; this result did not remain significant after Holm–Bonferroni correction.

4.4 Open-Ended Responses

The post-intervention survey also included three open-ended questions: (1) “What was the most helpful part of the study for your speaking?” (2) “What could be improved?”, and (3) “Any other comments?”. All participants responded in Mandarin Chinese, their first language. The responses were initially translated into English using Google Translate and then checked and revised by the researcher against the original Mandarin responses, which were retained for reference. All 24

participants answered the first question, 19 answered the second question, and 7 answered the third question.

The responses were analysed using descriptive thematic categorisation informed by Braun and Clarke (2006), as described in Section 3.6.2. The analysis identified eight themes: four shared themes across both groups, two themes distinctive to the AI-supported group, and two themes distinctive to the teacher-led group. Table 9 summarises the themes and the number of participants whose responses contributed to each theme. This section describes these themes and illustrates them using quotes from the exit surveys.

Table 9

Themes Identified in the Open-Ended Responses

Theme	AI-supported group (n = 13)	Teacher-led group (n = 11)
Shared themes		
Growing confidence in speaking English	4	5
Pronunciation correction as a valued component	2	4
Wish for more practice time	6	6
Movement toward more direct and detailed speaking	7	3
AI-supported group distinctive themes		
AI as a novel and engaging classroom element	5	0
Interest in self-directed and cross-skill AI use	3	0
Teacher-led group distinctive themes		
Value of live human interaction and peer presence	0	3
Concerns about class pace	0	2

Note. Counts indicate the number of participants whose responses contributed to each theme. One participant could contribute to more than one theme.

4.4.1 Shared Themes

Four themes appeared across both groups. These shared themes show that several aspects of the speaking-practice experience were similar regardless of feedback source.

Theme 1: Growing Confidence in Speaking English

A widely shared experience across both groups was an increased sense of confidence in using English in spoken contexts. Participants in both conditions described this as the most helpful outcome of the intervention.

The following quotes from the qualitative survey answers illustrate this sense of improvement.

"I feel more confident when I speak English."

(Participant S446409, AI group)

"My mispronunciations were corrected, and I can now speak English with confidence."

(Participant S263614, AI group)

"I feel more confident when speaking English."

(Participant S138020, Teacher group)

"My confidence in speaking English has improved, and I liked the classroom interaction."

(Participant S100285, Teacher group)

Theme 2: Pronunciation Correction as a Valued Component

In both groups, several participants singled out pronunciation correction as one of the most useful elements of the practice. In the Teacher group, this was the most frequently named helpful feature; in the AI group, it was mentioned alongside requests for "stricter" correction.

The following comments illustrate how participants described this improvement in their own words.

"My mispronunciations were corrected, and I can now speak English with confidence."

(Participant S263614, AI group)

" More rigorous correction of mispronunciations."

(Participant S752268, AI group)

"After my pronunciation was corrected, it became more standard, and I am now able to express myself more confidently in English."

(Participant S386722, Teacher group)

"The pronunciation correction was very useful."

(Participant S142076, Teacher group)

Theme 3: Wish for More Practice Time

When asked what could be improved, the most consistent response in both groups was a request for more practice time, higher class frequency, longer sessions, or one-on-one practice opportunities. This was the dominant suggestion across both conditions.

This theme can be exemplified by the following excerpts from the participants' feedback.

"The frequency of classes is not high enough."

(Participant S102629, AI group)

"The frequency of lessons could be higher, and I would also like to see one-on-one lessons with more practice time."

(Participant S446409, AI group)

"There is not enough time for speaking practice; the lesson could be extended by one hour."

(Participant S138020, Teacher group)

"The current course time is limited. Extending it would give more time for teachers and students to communicate, which would make the learning more effective."

(Participant S861319, Teacher group)

Theme 4: Movement Toward More Direct, Detailed Speaking

Participants in both groups described a shift in how they constructed their spoken responses: away from short, surface-level answers and toward more elaborated, more direct expressions. The two

groups expressed this shift in subtly different ways: AI participants more often described learning to "add details" or "expand answers"; Teacher participants more often described speaking "more directly" or moving away from templates.

The following quotes represent a selection of responses that were coded into this theme.

> *"My score improved. I learned how to give details and not just very short answers."*
>
> (Participant S526258, AI group)
>
> *"I have built the habit of saying more details when I speak English. I know how to expand my answers, and I no longer feel I have nothing to say."*
>
> (Participant S667611, AI group)
>
> *"This was the first time I learned the concept of low-context and high-context. When I speak English now, I know how to unfold my answers."*
>
> (Participant S774516, AI group)
>
> *"I can express myself more directly now."*
>
> (Participant S393705, Teacher group)
>
> *"It helped me learn how to express what I want to say more directly, rather than going around or blindly memorising templates, which had not been giving me good scores."*
>
> (Participant S861319, Teacher group)
>
> *"My pronunciation and fluency improved, and I express my meaning more directly than before."*
>
> (Participant S539702, Teacher group)

4.4.2 AI-Supported Group: Distinctive Themes

Two themes appeared only in the AI-supported group. These themes relate to participants' experience of AI as a classroom tool and their interest in using AI.

Theme 5: AI as a Novel and Engaging Classroom Element

A theme unique to the AI group was the perceived novelty and enjoyment of having AI integrated into the classroom. Several participants described the AI-supported sessions as more interesting or fresher than their previous experience of English classes, and one participant noted a shift in the teacher-student dynamic, framing the AI as something that made the relationship feel more collaborative rather than evaluative.

The following quotes from the survey serve to illustrate this theme.

> *"My speaking score improved a lot. My fluency improved this time, and I spoke with more detail. The addition of AI made the class very interesting — it felt like the teacher and we were no longer on opposite sides."*
>
> (Participant S767150, AI group)
>
> *"Using AI in class felt very novel. … It was my first time experiencing AI used in class, and I found it quite interesting."*
>
> (Participant S667611, AI group)
>
> *"I learned how to tell a story. The exercises in class were interesting, and having AI added made it feel a bit fresher than ordinary classes."*
>
> (Participant S987525, AI group)

Theme 6: Interest in Self-Directed and Cross-Skill AI Use

A second AI-distinctive theme reflected participants' curiosity about extending AI use beyond the classroom. Some asked for guidance on using AI for independent practice and raised questions about how to judge the reliability of AI-generated feedback. Others noted that skills they learned for speaking transferred to writing and asked whether AI could support other IELTS components.

Representative quotes:

> *"I hope the teacher can guide us on how to use AI to practise English independently, and how to judge whether the feedback AI gives is correct."*
>
> (Participant S916145, AI group)
>
> *"Apart from speaking, can other subjects also be improved with AI?"*
>
> (Participant S916145, AI group)
>
> *"I found that what I learned for speaking could also be used for writing. I am glad that my writing score also improved."*
>
> (Participant S987525, AI group)
>
> *"It would be even better if both the teacher's comments and the AI's comments could be obtained."*
>
> (Participant S667611, AI group)

4.4.3 Teacher-Led Group: Distinctive Themes

Two themes appeared only in the teacher-led group. These themes relate to the value of human interaction and the pace of the sessions.

Theme 7: Value of live human interaction and peer presence

A theme unique to the Teacher group was the value placed on live human interaction—both with the teacher and with peers. Participants described peer presence as motivating and identified in-class communication with the teacher as a key part of effective learning. This theme did not appear in the AI group's responses.

The following quotes from the survey responses serve to illustrate this theme.

> *"The classroom interaction was very good. Having other students there gave me the motivation to participate in the practice."*
>
> (Participant S215099, Teacher group)
>
> *"My confidence in speaking English has improved, and I liked the classroom interaction."*
>
> (Participant S100285, Teacher group)
>
> *"Extending the course time would increase the time for teacher–student communication and feedback, which would make the learning more effective."*
>
> (Participant S861319, Teacher group)

Theme 8: Concerns About Class Pace

A small but consistent theme in the Teacher group was that the pace of the classes felt too fast. This theme echoes the Likert finding reported in Section 4.3.2, where Teacher participants gave lower ratings to "Reasonable pace" than AI participants. None of the AI participants raised pace as

a concern in the open responses, although one mentioned a desire for a slower one-on-one option.

Only two participants commented on this issue. Their quotes are presented below:

> *"[Improvement] The pace of the class could be a little slower."*
>
> (Participant S481521, Teacher group)
>
> *"[Improvement] The pace could be a little slower."*
>
> (Participant S215099, Teacher group)

The open-ended responses show both overlap and contrast between the two groups. Both groups reported increased confidence, valued pronunciation correction, wanted more practice time, and described movement toward more direct and detailed speaking. At the same time, AI-supported participants highlighted the novelty of AI and its possible use for independent learning, while teacher-led participants emphasised live interaction, peer presence, and concerns about pace. These qualitative patterns are integrated with the IELTS score results and Likert survey findings in the next section.

4.5 Integration of Quantitative and Qualitative Findings

This section brings together the three sources of data: IELTS score changes, Likert survey responses, and open-ended responses. The purpose is to show how the quantitative and qualitative findings relate to each other in addressing the research questions, with particular attention to how AI-supported and teacher-led feedback compare in supporting IELTS Speaking development and how learners experienced the two feedback modes. The analysis identified four main patterns, which are summarised in

Table ***10***.

Table 10

Integrated Evidence Across IELTS Scores, Survey Responses, and Open-Ended Responses

Integrated finding	IELTS score evidence	Likert survey evidence	Open-ended response evidence
Comparable IELTS Speaking gains	Speaking Mdn gain: AI +0.50; Teacher +0.50. AI p = .025; Teacher p = .020; between-group n.s.	Most items: n.s. between groups	Both groups: confidence; clearer and more developed responses
More direct and detailed speaking	Speaking: significant gain in both groups	High agreement: clearer expression; direct/specific English	AI: expansion and details. Teacher: directness; fewer templates
Possible pacing difference	Not measured	AI rated pace higher; not significant after correction	Teacher: 2 requested slower pace. AI: 0 pace concerns
Possible value of live interaction	Listening: Teacher Mdn +0.50, *p* = .021; AI Mdn 0.00, *p* = .276	Not measured	Teacher: interaction; peer presence; real-time explanation

Note. The table summarises how the main integrated findings were supported across the three data sources. AI = AI-supported group; Teacher = teacher-led group; Mdn = median; n.s. = not significant.

The strongest point of convergence was that both the quantitative and qualitative findings pointed to improvement across the two conditions. The IELTS results showed gains in learners' speaking scores in both the AI-supported and teacher-led groups, while the survey and open-ended responses indicated generally positive learner experiences, including increased confidence and clearer spoken expression. Together, these findings suggest that both conditions supported learners' speaking practice during the four-week intervention.

The second integrated finding concerns the nature of the speaking improvement. Both groups reported becoming clearer, more specific, and more detailed in their spoken responses. This was visible in the Likert survey, where both groups gave high ratings for expressing ideas more clearly and using more specific expressions. The open-ended responses added more detail. AI-supported participants often described learning to add details and expand their answers, while teacher-led participants more often described expressing meaning directly and moving away from memorised templates. These descriptions are closely related to the study's focus on helping learners develop more explicit and sufficiently detailed spoken communication for IELTS.

The third integrated finding concerns the different affordances of the two formats. The AI-supported group rated "Reasonable pace" more positively than the teacher-led group, although this difference did not remain significant after correction for multiple comparisons. The open-ended responses showed a similar pattern: two teacher-led participants asked for a slower pace, while no AI-supported participant identified pace as a concern. This may point to a possible pacing difference between the two conditions, which is discussed in more detail in Chapter Five and Six.

The teacher-led group also showed a distinctive pattern. It was the only group to make a significant Listening gain, and teacher-led participants described live classroom interaction as valuable. However, the between-group comparison of Listening gain scores was not statistically significant, so these findings do not establish a greater Listening benefit from teacher-led practice. This pattern may partly relate to incidental listening exposure through teacher instructions, follow-up questions, and peer responses. However, outside IELTS preparation, test-day variation, individual differences, or chance variation within the small sample may also have contributed. This was not the primary target of the intervention, but it appears to be an important feature of the teacher-led experience.

Overall, the integrated findings suggest that AI-supported and teacher-led practice were associated with broadly comparable IELTS Speaking development in this study. At the same time, the two formats offered different learning conditions. AI-supported participants perceived the pace as more manageable, while teacher-led participants valued live interaction and peer presence. These findings are discussed further in Chapter Five.

4.6 Chapter Summary

This chapter presented the findings of a four-week mixed methods study comparing AI-supported and teacher-led IELTS speaking practice for Mandarin Chinese L1 learners. The results showed that both groups made significant gains in IELTS Speaking, with no significant between-group difference in Speaking gain scores.

The survey and open-ended responses supported this pattern. Participants in both groups reported positive experiences, including greater confidence, clearer expression, and movement toward more direct and detailed speaking. These findings relate to the study's focus on developing more explicit spoken communication.

The two formats also showed different affordances. AI-supported practice appeared to offer a more manageable pace for some learners and encouraged interest in independent AI use. The teacher-led group also showed a significant Listening gain. This pattern may partly relate to the live teacher input, peer responses, and additional listening exposure available in the teacher-led sessions. However, outside IELTS preparation, test-day variation, and individual differences may also have contributed to the result.

The findings suggest that the participants' IELTS Speaking scores of both groups improved after completing the course, with no significant difference detected between the two conditions. Learners' survey and open-ended responses were generally positive in both groups, although the two feedback modes appeared to support speaking practice in different ways. The implications, limitations, and future directions are discussed in Chapter Five.

Chapter Five: Discussion

This chapter discusses the findings in relation to the research questions and the literature reviewed in Chapter 2. It focuses on what the comparison between AI-supported and teacher-led IELTS speaking practice suggests, how the quantitative and qualitative findings relate to each other, and how the findings should be read within the limits of a small-scale study.

The discussion is organised around six sections. Section 5.1 considers the speaking gains and the role of repeated practice, while Section 5.2 discusses the convergence between the score data, survey responses, and open-ended comments. The chapter then examines the divergence between the two feedback modes in Section 5.3, focusing on the different affordances of AI-supported and teacher-led practice, followed by a consideration of the pedagogical implications in Section 5.4. The chapter closes with a discussion of the limitations of the current study and suggestions for future research in Section 5.5 and a summary in Section 5.6.

5.1 Speaking Gains, Repetition, and Feedback Source

The main quantitative finding was that both groups showed significant within-group gains in IELTS Speaking. Based on the pre-test and post-test data collected from the participants, the AI-supported and teacher-led groups made similar median gains in Speaking. The between-group comparison did not detect a significant difference in speaking gain between the two groups. This suggests that both feedback modes were associated with speaking improvement in this small-scale study. However, it does not necessarily show that the two feedback modes were equivalent or that one mode was better than the other. The absence of a significant between-group difference can be interpreted in several ways. The observed effect size for Speaking gain was small ($r = .08$), but the sample of 24 participants also limited the study's statistical power to detect modest group differences. The four-week intervention may also have been too short for differences between feedback modes to translate into detectable increases in IELTS Speaking scores. In addition, the shared practice cycle and the teacher-guided nature of the AI-supported condition may have reduced the contrast between the two groups. The null result should therefore be interpreted cautiously and not as evidence that the two feedback modes were equivalent.

Nevertheless, a median gain of 0.5 band in IELTS Speaking over the four-week intervention is still notable, given the demands of the test. To improve, learners need to work on several aspects of spoken performance at the same time, including idea organisation, vocabulary retrieval, grammatical control, pronunciation, and real-time fluency (Goh & Burns, 2012; Kormos, 2006). They also need to respond to the IELTS Speaking criteria, which assess fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation (IELTS, 2025a). A gain in Speaking score therefore suggests meaningful development in a complex performance skill, where

improvement requires learners to coordinate several aspects of spoken language at the same time.

The shared practice cycle provides an important context for interpreting this result. Both groups practised IELTS-style speaking tasks, received feedback, revised their responses, and made second attempts. The teacher-led group received direct teacher feedback, while the AI-supported group received AI-generated feedback within a teacher-guided routine. In other words, even though the feedback source differed, both conditions gave learners opportunities to act on feedback. This may help explain why both groups improved. Thus, considering the Speaking gains made by both groups, the results suggest that feedback was useful when it was connected to repeated practice and revision.

This explanation is consistent with research on feedback, noticing, interaction, and task repetition. Learners may benefit when feedback helps them notice a gap and when they have opportunities to modify their output and perform the task again (Bygate, 2001; Li & Vuono, 2019; Long, 1996; Lyster et al., 2013; Schmidt, 1990). In the present study, both groups had access to this cycle of production, feedback, revision, and repeated performance. The shared structure of the programme therefore needs to be considered alongside the two feedback sources.

The AI-supported condition also needs to be defined carefully. Learners did not simply use the generative AI chatbot Doubao on their own. The teacher designed the programme, developed the AI feedback prompt, controlled the timing, and guided learners through the AI-supported sessions. The AI tool provided feedback, but the learning routine remained teacher guided. For this reason, the findings need to be understood as evidence about guided AI-supported speaking practice within a structured programme, rather than independent chatbot use.

This aligns with recent research on generative AI in language learning. Kohnke et al. (2023) argued that ChatGPT and related tools can offer useful affordances for language teaching and learning, but they also require digital competence and careful use. Yang and Li (2024)'s review also showed that the value of generative AI depends on task design, learning objectives, and appropriate integration. The present study supports this view. AI appeared useful when it was embedded in a clear pedagogical routine, rather than treated as a stand-alone replacement for teacher feedback.

This explanation helps position the present findings in relation to recent AI-supported speaking studies. Some studies have reported that AI-mediated or chatbot-based practice can support speaking performance, willingness to communicate, and speaker confidence as well as help alleviate anxiety (Du & Daniel, 2024; Fathi et al., 2024; Zhang et al., 2024). The present study partly supports these findings because the AI-supported group showed significant within-group gains in IELTS Speaking. However, since the between-group comparison did not show that the AI-supported group improved significantly more than the teacher-led group, the findings suggest that

AI-supported speaking practice did not provide an advantage over teacher-led feedback in this small-scale study.

5.2 Mixed Methods Convergence: Speaking Scores and Learner Development

The quantitative and qualitative findings converged on the same broad pattern. Both groups made significant within-group gains in IELTS Speaking, and learners in both groups reported positive changes in their speaking experience. The survey and open-ended responses point to an increase in confidence, clearer expression, and more developed spoken responses. These qualitative findings help explain what the score gains meant from the learners' perspective.

Confidence was one of the clearest shared themes. Participants from both groups reported feeling more confident when speaking English after the programme. For example, one AI-supported participant wrote, "I feel more confident when I speak English" (Participant S446409), while a teacher-led participant similarly reported, "My confidence in speaking English has improved, and I liked the classroom interaction" (Participant S100285). These comments suggest that gains in confidence were not limited to one feedback mode.

Another key aspect of learners' speaking development is their ability and willingness to use that knowledge in real time. Research on willingness to communicate and speaking anxiety shows that confidence, classroom environment, anxiety, and perceived competence can affect whether learners participate in spoken English practice (Bai, 2023; Chen, 2024; Lin et al., 2025; Peng, 2012). Although this study did not directly measure anxiety or willingness to communicate, the learner comments suggest that confidence was an important part of their reported speaking development.

The programme structure may help explain this confidence gain. Learners knew that each speaking task would be followed by feedback and a second attempt. The practice routine was therefore predictable. Instead of giving one answer and moving on, learners had the chance to notice problems, revise their response, and try again. This process may have made IELTS-style speaking feel more manageable. The study did not directly measure anxiety, so no claim can be made about anxiety reduction. However, the confidence reported by both groups suggests that structured speaking practice helped learners feel more able to participate in IELTS-style speaking tasks.

A second point of convergence was participants' perception that their speaking had become clearer and more developed after the programme. When describing the gains they had experienced in the exit survey, AI-supported participants most frequently mentioned that they had learned to add details, expand their answers, and avoid very short responses. Participants from the teacher-led group more often commented that they had learned to speak more directly and

move away from memorised templates. Although the two groups emphasised different aspects of learning, these comments suggest a shared sense that their speaking had become clearer and better structured following the programme. More generally, the comments suggest that learners had developed a better awareness of how IELTS Speaking responses need to be presented and structured, and how they could improve their own responses.

The participant comments make this shift more concrete. One AI-supported participant wrote, “I have built the habit of saying more details when I speak English. I know how to expand my answers, and I no longer feel I have nothing to say” (Participant S667611). In the teacher-led group, one participant explained that the programme helped them “express what I want to say more directly, rather than going around or blindly memorising templates” (Participant S861319). In this comment, speaking more directly appeared to mean expressing ideas more clearly, with less reliance on indirect wording or memorised answer templates. These comments show that even though the two groups did not describe improvement in exactly the same way, both moved toward clearer and more purposeful spoken responses.

This point connects to the study’s use of the high-context and low-context framework. The framework informed the AI prompt design, especially the focus on helping learners make their ideas clearer, more direct, and more fully developed in line with IELTS Speaking requirements. In practice, this meant encouraging learners to move from short or implicit answers toward responses with clearer explanations, reasons, and examples. This focus is consistent with the IELTS Speaking criteria (IELTS, 2025a).

A further shared theme was the wish for more practice time. Participants in both groups asked for more opportunities to speak, longer sessions, higher class frequency, or more one-on-one practice. One AI-supported participant suggested that “the frequency of lessons could be higher” and asked for “one-on-one lessons with more practice time” (Participant S446409). A teacher-led participant similarly commented that “there is not enough time for speaking practice” and suggested that the lesson could be extended (Participant S138020). These comments point to a central problem in speaking instruction: learners often need more supported speaking time than a short group course can provide.

The desire for additional practice strengthens the pedagogical argument for AI-supported work. Teachers cannot provide unlimited individual speaking turns and detailed feedback in group settings. AI does not solve this problem completely, but it can extend opportunities for rehearsal and revision outside limited class time. Its value may therefore be strongest where the main bottleneck is the amount of individual practice available to learners.

The survey results support these qualitative patterns. Both groups gave high ratings to their perceived ability to express ideas more clearly in English, confidence when speaking in IELTS-style tasks, the usefulness of feedback, and the structure of the programme. Most item-level comparisons did not show significant between-group differences, so the survey data should not be

used to claim that one condition was clearly better. Instead, the item-level comparisons are more useful as evidence of a broadly positive learner experience across both conditions.

Overall, the convergent findings show shared benefits rather than a single winning feedback mode. The quantitative results show that speaking scores changed, and the qualitative findings explain what learners felt they gained from the process. Together, the two strands suggest that structured IELTS speaking practice, feedback, and opportunities to revise responses were useful in both conditions.

5.3 Mixed Methods Divergence: Different Affordances of AI and Teacher Feedback

Alongside this convergence, the findings also showed divergence. Thus, while both formats were associated with positive speaking-practice outcomes, they appeared to offer different kinds of support. AI-supported practice was more closely associated with answer expansion, alternative expressions, novelty, repeated practice, and a more manageable pace. Teacher-led practice was more closely associated with pronunciation correction and real-time diagnosis. The contrast suggests that the two feedback modes should not be understood simply as better or worse versions of the same support. Instead, the data suggest that participants perceived different affordances in the two feedback modes.

One of the clearest perceived affordances of AI-supported feedback was support for response development. In IELTS speaking practice, learners may have ideas but still need support with expanding their responses, choosing suitable wording, and expressing the same idea in different ways. AI can quickly generate several ways to express the same idea, which gives learners options for expanding and refining their responses. They can compare wording, choose expressions that fit their own level, and build a more usable expression bank. Participants appeared to find AI feedback useful for response development because it could quickly provide additional content, examples, and alternative wording in a single response.

The open-ended comments from the AI-supported group support this interpretation. Learners described adding details, expanding answers, and learning how to say more. One participant wrote, “My score improved. I learned how to give details and not just very short answers” (Participant S526258). Another participant connected this directly to the study’s communication focus: “This was the first time I learned the concept of low-context and high-context. When I speak English now, I know how to unfold my answers” (Participant S774516). These comments suggest that participants perceived AI feedback as helpful in showing how a short answer could be developed into a fuller response.

Rather than using this result to make a broad claim that AI improves speaking, it is more useful to link it to the specific affordances of generative AI. Previous literature has suggested that generative

AI can provide immediate responses, reformulations, explanations, and repeated opportunities for practice (Kohnke et al., 2023; Law, 2024). The present findings add a more specific classroom-based observation: in this IELTS Speaking programme, AI feedback seemed especially useful for helping learners expand short responses and access alternative expressions. However, this does not mean that AI feedback was superior to teacher feedback overall. Its strength appeared to lie mainly in content development and wording support.

One reason for this pattern is that AI is well suited to language-generation tasks such as reformulating, extending, and rephrasing learner responses. A short or vague spoken response can be turned into a fuller answer or expressed in several alternative ways. The learner can then compare these options and decide which expressions to use in their own speaking. Teacher feedback can also provide reformulations, but in a live group session it is difficult for one teacher to generate multiple individualised wording options for every learner. AI may therefore be particularly useful when learners need fast, repeated support with expression building.

From a noticing perspective, AI-generated reformulations may have helped learners see the gap between a short or vague answer and a more developed IELTS-style response. The feedback would have also highlighted possible expansions on the topic. This point also connects back to the high-context and low-context framework by Hall (1976) discussed in Chapter 2. In this study, the framework informed the AI prompt design by encouraging feedback that made learners' ideas more explicit and developed. For example, when a learner gave a short answer, AI feedback could show how the answer might be expanded by adding a reason, an example, or a clearer explanation. In this way, AI feedback helped learners see how an implicit or underdeveloped response could be turned into a fuller IELTS-style answer.

The findings also suggest that the two feedback modes may have supported more explicit spoken communication in different ways. Participants in the AI-supported group often described learning to expand short answers, add details, and make their ideas more developed. Teacher-led participants more often referred to expressing meaning more directly and moving away from memorised templates. Viewed through the high-context and low-context framework, these patterns suggest that AI-supported feedback may have helped learners make implicit ideas more explicit through expansion and elaboration, while teacher-led feedback may have supported more direct expression through immediate clarification and live feedback. These interpretations should be treated cautiously, as the study did not directly measure changes in communication style.

The teacher-led group identified a different pattern of strength. Teacher-led participants more often highlighted pronunciation correction, live explanation, and real-time identification of problems. In their exit surveys, one teacher-led participant noted that "the pronunciation correction was very useful" (Participant S142076), while another reported that after pronunciation correction, their speech became "more standard" and they were able to express themselves more confidently (Participant S386722). These comments point to a form of feedback that depends on live listening

and immediate human judgement. In a live speaking class, the teacher can notice whether a learner's answer is unclear, whether a pronunciation issue affects understanding, or whether the learner is relying too heavily on memorised templates. The teacher can then adjust the feedback immediately.

Overall, teacher feedback was more closely aligned with interactional feedback in real time. The teacher could interrupt, clarify, give pronunciation feedback, or respond to learners' pausing and hesitation as they occurred. In this sense, both modes appeared to support noticing and answer revision, but they did so through different routes. IELTS Speaking is not only about what learners say, but also about how they say it. A teacher can respond to hesitation, pronunciation, intonation, rhythm, repair, and uncertainty as they occur. These features are difficult for a general-purpose chatbot to diagnose fully, especially when the tool is not designed as a specialised pronunciation or fluency assessment system.

Another perceived benefit highlighted by participants was feeling less pressured or nervous when using the AI tool than when speaking directly to a teacher. Students found it intriguing to use new technology for their speaking development. One participant wrote that "the addition of AI made the class very interesting" and that it "felt like the teacher and we were no longer on opposite sides" (Participant S767150). The comment suggests that participants may have experienced feedback differently when AI was involved. Instead of receiving all correction directly from the teacher, learners worked with the teacher to interpret AI-generated feedback. Feedback may therefore have felt less judgemental and personal.

The possible pacing difference should not be interpreted as a difference in actual lesson time. Both groups followed the same session length and practice cycle. Rather, it refers to learners' perceived experience of the lesson pace. This point comes from the survey item on whether the pace of the lessons was reasonable, where the AI-supported group gave slightly more positive ratings, and from the open-ended comments, where some teacher-led participants requested more time or a slower pace. This suggests that the AI-supported format may have felt more manageable for some learners, although the study did not directly measure cognitive load, anxiety, or time perception. Research on human-chatbot interaction suggests that response time and interaction design can shape users' perceived latency and overall experience (Abbas et al., 2021; Gnewuch et al., 2022). In this study, the written, structured format of AI-generated feedback may have allowed learners to reread suggestions while reviewing their responses. This may have influenced their perceptions of the lesson pace, although both groups practised publicly in the main meeting room. Although the study did not directly measure anxiety, cognitive load, or time perception, this may help explain why the AI-supported pace was experienced as more manageable.

Teacher-led feedback in turn offered peer presence and interpersonal interaction. Some teacher-led participants valued listening to others, speaking in front of classmates, and receiving feedback from a real teacher. One participant wrote, "The classroom interaction was very good. Having other

students there gave me the motivation to participate in the practice" (Participant S215099). The social dimension also matters because speaking development is not only an individual process. Learners need opportunities to speak to real listeners, respond to clarification, and manage the pressure of live communication. The teacher-led condition appeared to provide this interactional dimension more strongly.

The divergence between the two feedback modes can therefore be understood as a difference in pedagogical function. AI-supported feedback was perceived as particularly useful for repeated support with expansion, wording, and alternative expressions. Teacher-led feedback was more often valued for pronunciation, fluency, live diagnosis, and interpersonal interaction. For IELTS Speaking preparation, both forms of support are useful. AI can help learners develop what they want to say, while teacher feedback can help them understand how their spoken performance is heard and evaluated in real time.

The findings therefore support a complementary interpretation. AI-supported and teacher-led feedback did not simply differ in delivery format; they directed learners' attention to different aspects of speaking. AI made expansion and alternative wording more available. Teacher feedback made pronunciation, fluency, and interactional problems more visible. From a feedback-processing perspective, this complementarity may reflect different ways of supporting noticing and revision. AI-generated feedback may make lexical choices, answer development, and alternative formulations more visible through repeated written reformulation, while teacher feedback may draw attention more immediately to pronunciation, fluency, and interactional problems as they occur in live speech. Both modes may therefore support noticing and subsequent revision, but through somewhat different feedback-processing routes. This suggests that AI feedback may be most useful when it is used to extend teacher-led instruction, rather than as a replacement for teacher judgement.

5.4 Pedagogical Implications

The findings suggest that guided AI-supported practice may be useful as a complement to teacher-led IELTS Speaking instruction. The strongest implication is not that AI should replace teachers. Instead, AI may extend the amount and type of practice available to learners. In this study, its clearest value was in answer expansion, alternative expressions, repeated practice, and post-feedback revision. These are areas where learners often need frequent individualised support, but where teacher time is limited.

The implication is especially relevant for EFL speaking contexts, where learners may have limited opportunities to use English in everyday life outside the classroom. In group classes, teachers cannot listen to every learner repeatedly, provide detailed feedback on every response, and allow each learner to practise again as often as needed. AI can help address this practical constraint by giving learners additional opportunities to rehearse, revise, and consolidate speaking responses

outside the limited class time. Used carefully, it may help learners prepare more developed answers before class and continue practising after class.

Because the AI-supported condition in this study was guided and teacher-informed, the pedagogical implications should be understood in relation to structured AI-supported practice rather than independent AI use. In other words, the findings suggest the value of AI when it is used within a clear learning routine, with teacher support for task design, prompt use, and feedback interpretation. The introduction of AI supported practice would require teachers to show learners how to write useful prompts, judge feedback accuracy, select appropriate expressions, and adapt AI suggestions into their own voice. Learners would also need to understand the limits of AI feedback, especially for pronunciation, fluency, and live communicative performance. It would be important for teachers to emphasise to learners that AI should only be used as a practice tool and that not every AI suggestion should be accepted as correct.

These complementary affordances suggest a reason to explore hybrid models. Although the present study did not test a hybrid condition, future research could examine a teacher-informed model that combines AI-supported rehearsal and response development with teacher feedback on pronunciation, fluency, interaction, and judgement.

One possible classroom model could use AI before, during, and after class. Before class, learners could use AI for rehearsal and idea development. They could practise a response, ask AI for alternative wording, and bring revised answers to class. During class, teachers could focus on areas where human feedback is especially important, such as pronunciation, fluency, coherence, exam relevance, and real-time clarification. After class, learners could use AI again to consolidate expressions and practise similar questions at their own pace. This would help address the problem of limited teacher time, because learners could continue speaking practice outside class without waiting for individual teacher feedback.

AI could also support teachers' post-class feedback work. For example, teachers could use AI to organise lesson notes, summarise common problems, and prepare short individual feedback reports for learners. These reports could remind learners of their main issues, useful expressions, and suggested practice tasks after class. This may reduce the need for learners to take detailed notes during the lesson and allow them to focus more on listening, speaking, and participating in the class. In this model, AI would be used to support repetition, language expansion, and post-class consolidation, while the teacher remains responsible for judgement, correction, interaction, and feedback quality.

AI should also be positioned carefully in relation to assessment. Doubao was used as a feedback and practice tool in this study, not as a validated IELTS scoring system. Learners may benefit from AI suggestions on content and wording, but they still need teachers or trained assessors to help them understand IELTS performance standards. AI can support preparation, but it should not be treated as an official substitute for IELTS assessment or expert teacher judgement.

More broadly, these findings contribute to ongoing discussions about the role of AI in technology-enhanced education. Rather than reducing the importance of teachers, the results suggest a possible shift in the teacher's role towards designing learning tasks, guiding the use of AI, evaluating feedback quality, and providing forms of support that automated systems may not provide reliably. This may be particularly relevant in high-stakes test preparation, where learners need frequent practice but also require expert judgement about performance standards. In this context, AI may extend access to practice and feedback, while human expertise remains important for interpreting performance and maintaining feedback quality.

Overall, the pedagogical implication is that AI-supported practice is most useful when it is structured, guided, and connected to clear learning goals. AI can increase opportunities for rehearsal and response development, while teachers remain central to feedback quality, pronunciation support, interaction, and critical evaluation. This suggests that AI-supported practice may be most useful as an extension of teacher-led speaking instruction, rather than as a separate or unguided form of practice.

5.5 Limitations and Future Research

The main limitation of this study was the small sample size. The final sample included 24 participants, with 13 in the AI-supported group and 11 in the teacher-led group. This limited the statistical power of the between-group comparisons and means that the quantitative findings should be treated as indicative rather than conclusive. While the study suggests clear trends in terms of how AI-supported feedback compares to teacher-led feedback in the context of an IELTS speaking course, a larger scale study is needed to confirm these findings.

Participant attrition also limits the interpretation of the findings. Of the 32 participants initially enrolled, 24 met the attendance and data requirements for inclusion in the final analysis. The findings therefore reflect participants who met these requirements and may not represent the experiences or outcomes of all those initially enrolled. Excluding participants after random allocation may also have introduced selection bias.

The programme was also relatively short. It lasted four weeks and included eight sessions. This was sufficient for examining short-term score changes and learner perceptions, but it cannot show how speaking development might unfold over a longer period of instruction. A longer course would make it possible to examine whether learners continued to develop with more sustained practice. Future studies could also include a follow-up survey or delayed speaking measure to explore whether learners' reported confidence, response development, and speaking performance were maintained after the programme ended.

The use of official IELTS scores also had limitations. Official IELTS results provided a validated and consistent outcome measure because they reflected participants' real test performance. However, IELTS band scores are broad and move in 0.5 increments, so small changes in speaking

performance may not be fully captured. IELTS Speaking is also assessed through four criteria: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. In this study, participants' available results showed only the final speaking band score, so the analysis could not identify which specific criterion had improved most. Although the Speaking scores were produced within the official IELTS examiner quality-assurance system, the present study did not include an additional independent second rating or study-specific inter-rater reliability check. This was because the researcher did not have access to recordings of participants' official IELTS Speaking tests for independent rescoring. Score changes may also have been affected by topics, test-day conditions, and independent preparation outside the programme. For these reasons, the IELTS score data was interpreted together with the survey and open-ended findings.

One unexpected pattern was the teacher-led group's significant Listening gain. This result should be interpreted with particular care because Listening was not directly targeted by the intervention. It may reflect live teacher input, peer responses, outside IELTS preparation, test-score fluctuation, individual differences, or chance variation within the small sample. For this reason, it is treated as a secondary pattern rather than as evidence that the intervention improved receptive skills. The possible relationship between Speaking course mode and Listening performance could be explored in future research.

A further limitation is that participants' independent use of Doubao outside the scheduled sessions was not monitored. Although such use was not part of the intervention, learners in the AI-supported group may have accessed additional AI feedback between sessions, which could have influenced their outcomes. In this study, learners used Doubao within a teacher-designed lesson structure and with an IELTS Speaking prompt. Independent learners may use AI in different ways, rely too heavily on AI-generated expressions, or accept feedback without judging whether it is accurate or appropriate. Further research is needed to better understand how guided AI-supported practice and independent AI self-study affect speaking development.

Furthermore, the findings are only applicable to one specific AI platform. During the tool selection and prompt design stages, Doubao was found to be capable of giving detailed feedback that was suitable for the tasks required in this study. It could respond to IELTS-style Speaking answers, comment on content and wording, and provide suggestions for expansion and alternative expressions. However, in this sense, Doubao was used deliberately as a general-purpose generative AI chatbot, not as a specialised pronunciation or fluency assessment tool. This means that it could not fully diagnose features such as pronunciation accuracy, hesitation, intonation, rhythm, or tone in the same way that a teacher could respond to live spoken performance. Other AI systems, especially tools designed specifically for speaking assessment or pronunciation training, may produce different results.

The absence of audio recordings also limited the analysis of spoken performance. The study analysed official IELTS scores, survey responses, and learners' open-ended comments, but it did

not directly analyse changes in pronunciation, pausing and hesitation, fluency, intonation, interaction, response organisation, or learners' uptake of feedback. This does not weaken the value of the existing findings, but it does mean that the study could not show exactly which spoken features changed during the programme. Future studies could include audio or video data to examine speaking development in more detail and to compare AI-generated feedback with teacher feedback at a finer linguistic level.

The survey findings also relied on learner self-report and should therefore be interpreted as participants' perceptions of their experience. Because the researcher was also the teacher during the programme, some participants may have felt more inclined to respond positively. Cultural expectations around politeness and respect for teachers may also have made some participants less comfortable expressing strong criticism. These influences cannot be confirmed from the available data, but they may have affected both the Likert-scale ratings and the open-ended comments.

The researcher's dual role should also be acknowledged. The researcher designed the AI prompt and delivered the teacher-led sessions. This helped maintain consistency across the programme, but it also means that the researcher's pedagogical decisions shaped both conditions. Participants' responses and classroom behaviour may also have been influenced by their awareness of the researcher's role. A larger research team in future studies could help reduce this limitation by separating the roles related to teaching, feedback delivery, data collection, and data analysis.

Taken together, these limitations mean that the findings should be interpreted cautiously. The observed patterns cannot be attributed solely to the feedback mode, as other features of the programme and participants' wider IELTS preparation may also have influenced the outcomes.

These limitations point to several directions for future research. Larger samples would provide stronger statistical power and help reduce the influence of individual differences. A larger research team and a larger number of learners would also make it easier to examine whether the patterns found in this study hold across different classes, teachers, and learner profiles. Longer intervention periods would make it possible to examine whether learners show greater speaking development over a longer period of practice and whether clearer patterns emerge across the two feedback conditions.

Future research could also test hybrid models directly. The present findings suggest that AI-supported and teacher-led feedback may offer complementary strengths, but this was inferred from two separate conditions. A future study could compare teacher-led practice, AI-supported practice, and a hybrid model in which AI is used for rehearsal, expression-building, and post-class consolidation, while teachers focus on pronunciation, fluency, interaction, and exam-relevant judgement. This would make it possible to examine whether combining AI and teacher feedback produces stronger outcomes than using either feedback mode alone.

Finally, future studies could examine different learner groups, proficiency levels, AI tools, and task types. The present study focused on L1 Mandarin learners preparing for IELTS Speaking and used Doubao within a structured online programme. Other learner populations and other AI systems may produce different patterns. Further research is needed before broader claims can be made about the role of AI feedback in EFL speaking instruction.

5.6 Chapter Summary

This chapter discussed the findings in relation to the research questions and the literature reviewed in Chapter 2. The main quantitative finding was that both groups showed significant within-group gains in IELTS Speaking, while no significant between-group difference in speaking gain was detected. This suggests that both AI-supported and teacher-led practice were associated with speaking improvement in this small-scale study, although the findings should be interpreted carefully because of the sample size and study design.

The mixed methods findings showed both convergence and divergence. The convergence was that learners in both groups reported greater confidence, clearer expression, more developed responses, and a desire for more speaking practice. These findings suggest that participants experienced the structured practice, feedback, and opportunities for revision positively across both conditions.

The divergence was that the two feedback modes appeared to support different aspects of speaking development. Participants perceived AI-supported feedback as particularly useful for answer expansion, alternative expressions, repeated practice, and a more manageable feedback experience, while teacher-led participants particularly valued pronunciation correction, live explanation, peer interaction, and real-time diagnosis of spoken performance. These differences suggest that AI feedback and teacher feedback cannot be considered interchangeable. Instead, they seem to offer complementary forms of support.

The chapter supports a cautious complementary interpretation. Guided AI-supported practice may extend opportunities for IELTS Speaking rehearsal and response development, especially when teacher time is limited. However, teacher guidance remains important for feedback quality, pronunciation, fluency, interaction, and the critical evaluation of AI-generated suggestions. The next chapter reflects on the research process, the researcher's role, and the learning gained from conducting this study.

Chapter Six: Reflections on the Use of AI for Language Learning and Research

This chapter reflects on my use of artificial intelligence (AI) throughout the whole research project. Unlike the previous chapters, which focused on the study itself, this chapter considers my position as a teacher, researcher, and AI user. It reflects on how AI shaped both the project itself and the wider research process.

A central feature of this project is that AI was both the object of research and a tool that supported this research. Thus, the study examined how AI-supported IELTS speaking practice compared with teacher-led practice and used AI to support the research work itself. Specifically, I used AI to support literature searching, the initial screening of journal articles, summarising abstracts and methods, brainstorming, receiving feedback on my academic writing, and organising the project. I also used it to learn unfamiliar research tasks, such as organising data files, choosing suitable descriptive statistics, and understanding how to report statistical results. The use of AI in these two capacities gave me an opportunity to reflect on how AI can be used more generally in learning and teaching environments.

The chapter is organised into three main parts. First, I reflect on my use of AI as a teacher, including the motivation for the study, the design of the AI-supported speaking practice, and my observations of participants' interaction with AI. Second, I reflect on my use of AI as a researcher, focusing on how AI supported my reading, thinking, writing, and project management, as well as the risks I encountered. Finally, I identify practical takeaways for language learners, teachers, researchers, and AI prompt design.

The purpose of this chapter is not to present AI as a perfect solution for language learning or research. Instead, it shows how AI can be useful when it is used critically, transparently, and within clearly guided educational and research processes. The central reflection is that AI's value in this project came from the researcher's teaching experience, research responsibility, and ongoing critical reflection.

6.1 My Use of AI as a Teacher

The use of AI in language teaching depends on how it is designed, guided, and used in practice. This section reflects on my position as a teacher-researcher, the design of the AI-supported speaking practice, participants' interactions with AI, and what I learned from the teaching experience.

6.1.1 Positioning Myself as a Teacher-Researcher

This study emerged from my dual position as an IELTS speaking teacher and a researcher with a long-standing interest in AI technologies. My interest in this topic was shaped by both my teaching experience and my personal experience of using AI tools over several years. As someone who has used AI regularly, I have observed how quickly the technology has developed. In some cases, the quality of AI responses seemed to improve within months. This made me wonder how AI could be used in language learning, especially in speaking practice.

My own experience of learning English also influenced the direction of this study. When I was learning English as a child, learners had far fewer digital resources. Opportunities for speaking practice were usually limited to what the classroom could provide. Whether speaking was practised efficiently depended on the teacher's lesson design and on classmates' willingness to participate. Outside class, more regular individual speaking practice was often only available through private tutoring, which could be expensive. Feedback on speaking was also limited and often delayed. In contrast, current learners have access to a much wider range of digital tools for language learning, including AI chatbots, voice interaction tools, translation tools, online courses, and automated feedback systems. These tools are not limited to speaking practice, but they create new possibilities for practising and receiving language support outside the classroom. This creates new opportunities for independent learning, especially for students who may not be able to afford expensive English classes or who have limited access to educational resources.

At the same time, I was aware that access to AI does not automatically lead to effective learning. Learners may use AI passively, ask vague questions, accept inaccurate feedback, or rely too heavily on AI-generated answers. From a teaching perspective, this raised an important question for me: how can teachers guide learners to use AI in a more purposeful and critical way, rather than simply allowing unrestricted or uncritical use?

Before starting this study, I had discussed the possibility of using AI in IELTS Speaking classes with my previous teaching institution. However, there was some hesitation. One concern was that fee-paying students and parents might see AI as less valuable than teacher-led instruction. This response reflected a wider uncertainty around AI in education. While AI has the potential to save time and reduce costs, many teachers are worried AI may be used to replace them, and as a result, do not take full advantage of what AI has to offer.

This hesitance and mistrust of AI motivated me to conduct the project outside my regular teaching role. I wanted to move beyond assumptions about AI and examine its use in a structured way. The study allowed me to design an AI-supported speaking practice model, compare it with teacher-led practice, and reflect on what AI can and cannot do in an IELTS Speaking learning context without being restricted by the goals and regulations of a teaching business.

6.1.2 Designing AI-Supported IELTS Speaking Practice

The AI-supported practice was designed as a structured learning activity rather than free AI use. In other words, learners were not simply told to use AI for speaking practice. The intervention focused on IELTS Speaking Part 2, and the AI-supported group followed the same general practice structure as the teacher-led group. The main difference was the source of feedback.

My prompt design was informed by my teaching experience and by insights gained from completing introductory AI and prompting training modules, including DeepLearningAI's *AI for Everyone* and *Google Prompting Essentials* (Google Career Certificates, 2024; Ng, 2019). They helped me understand that useful AI output depends on the clarity of goals, context, and constraints contained in the prompt and that prompts require iterative refinement. I tested early prompt versions using researcher-created sample IELTS Speaking Part 2 answers based on common learner difficulties I had observed in my teaching, rather than using participant data. I then revised the prompt when the AI feedback was too general, too long, too formal, or not clearly linked to IELTS Speaking criteria. Figure 7 summarises the stages through which the AI feedback prompt was developed.

Figure 7

Prompt Development Process



As shown in Figure 7, the prompt developed gradually. The first version was simple: I gave AI an IELTS Speaking Part 2 cue card, provided a learner's answer, and asked for feedback. That broad instruction often produced general comments. The feedback was correct, but it was too broad and did not always show learners exactly what to improve or how to improve it. It also did not fully match the targeted feedback I, as a teacher, would give in an IELTS speaking class.

I then modified the prompt to specify the role and tone of the feedback. The new prompt asked AI to act as a supportive IELTS speaking coach, which positioned it as a feedback provider rather than a general chatbot. The prompt specified the use of a supportive tone as students may otherwise feel personally criticised by the feedback, because research has shown that speaking feedback in particular can feel like a personal criticism (Ergül, 2021). Therefore, feedback needs to help learners notice problems without making the practice feel discouraging.

The next design decision was to align the feedback with IELTS Speaking criteria. The feedback needed to connect with fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. At the same time, I did not want the AI to act as an official examiner. Therefore,

the final prompt did not ask for an overall band score. I wanted learners to focus on constructive feedback and improvement rather than focus on their score.

Answer expansion became another key focus. In my teaching experience, many Chinese EFL learners need support not only with language accuracy, but also with developing ideas. IELTS Speaking Part 2 requires candidates to speak for an extended period and make their ideas explicit to an examiner who does not share their personal context. The prompt, therefore, was adjusted to encourage examples, explanations, details, and follow-up thinking. This design choice was later reflected in the findings, where participants in the AI-supported group particularly valued answer expansion, alternative expressions, and support for developing more detailed responses.

I also limited the amount of feedback. The final prompt asked for two strengths, two grammar-focused issues, two targeted fixes, and two follow-up questions. The limit reflected normal teaching practice. In a speaking lesson, a teacher usually prioritises the most important issues because learners can only absorb a limited amount of feedback at one time (Shute, 2008). Focused feedback is more useful than a long list of corrections.

One important instruction was to quote the learner's exact words. Without exact quotes, AI feedback can become vague. It may say "improve your grammar" or "add more detail" without showing where the problem appears. Exact phrases made the feedback more concrete and evidence based. Learners could see what they had said, why it worked or did not work, and how it could be improved.

The prompt also required natural spoken English. In early testing, AI sometimes gave rewrites that sounded too formal or academic. That style was not suitable for IELTS Speaking, where topics are usually everyday topics such as people, places, hobbies, memories, and experiences. The final prompt therefore asked AI to keep the language spoken, natural, and not essay-like.

Follow-up questions were included to support answer expansion. They were not only questions for learners to answer immediately. They also helped learners see other directions in which an answer could develop, such as reasons, examples, comparisons, feelings, or personal details. That feature directly addressed a problem I often saw in class: students did not always know what else to say.

The instruction "wait for my reply" was also included to keep the interaction active. Without this instruction, AI might provide feedback and then immediately generate a fully revised answer. By asking AI to wait, the learner was given time to develop the answer. This helped create a coaching cycle in which AI supported revision.

As described in the methodology chapter, the prompt was reviewed with input from two experienced EFL teachers. I shared AI feedback examples with them and asked what was useful, what was missing, and what needed to be clearer or more concise. Their comments informed later revisions. The process showed that prompt design was not only a technical skill. In this project,

prompt design became a form of pedagogical design. Figure 8 summarises the iterative cycle of testing the prompt, reviewing the AI output, obtaining feedback from experienced teachers, and revising the instructions.

Figure 8

AI Prompt Revision Cycle

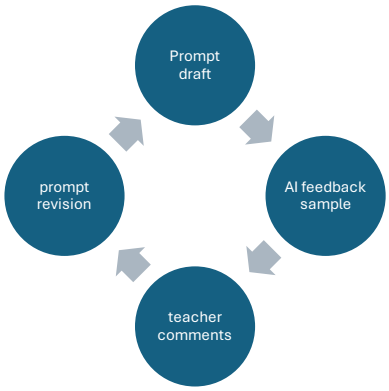


Table 11 summarises the main features of the final AI feedback prompt and the pedagogical purpose of each feature.

Table 11

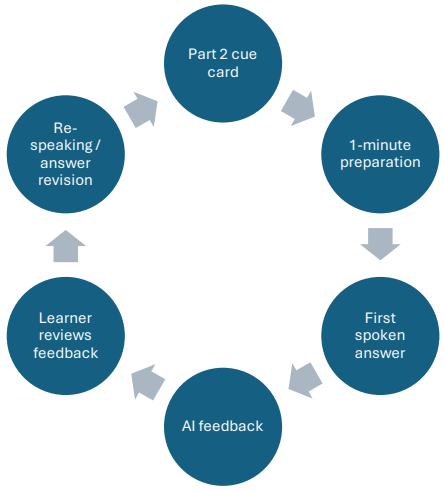
Anatomy of the Final AI Feedback Prompt

Prompt feature	Why it was added	Pedagogical purpose
IELTS speaking coach persona	To give AI a clear role	Positions AI as a feedback provider
Supportive tone	To reduce pressure	Keeps feedback constructive
IELTS Speaking criteria	To connect feedback with the task	Keeps feedback relevant to IELTS Speaking
No overall band score	To avoid repeated score anxiety	Focuses learners on improvement rather than judgement
Exact quotes from learner answer	To make feedback evidence-based and specific	Shows learners where the issue appears
Limited feedback quantity	To avoid overwhelming learners	Makes feedback highlight main issues only
Grammar-focused issues	To provide targeted correction	Makes feedback more actionable
Spoken rewrites	To avoid essay-like language	Keeps the language suitable for speaking
Follow-up questions	To support answer expansion	Gives learners clear directions for development
"Wait for my reply"	To continue interaction	Creates a coaching cycle

The prompt was embedded in a practice cycle. As illustrated in Figure 9, learners received a Part 2 cue card, prepared for one minute, gave a spoken answer, received AI feedback, reviewed the feedback, and then re-spoke or revised the answer. The cycle encouraged learners to use feedback immediately. It also ensured that both AI-supported and teacher-led conditions were comparable, as both groups followed the same cycle and used the same materials in the same order.

Figure 9

AI-Supported IELTS Speaking Part 2 Practice Cycle



The design of the AI-supported practice cycle reinforced an important point: AI generated the feedback, but the learning process was still shaped by teacher judgement. The value of the tool depended on how it was placed inside a structured learning activity.

6.1.3 Observing Participants' Interaction with AI

During the intervention, I observed several patterns in how participants interacted with AI. Some participants in the AI-supported group appeared curious and engaged, and AI seemed novel to them. The use of AI may also have changed how some learners experienced the feedback process, making it feel less directly evaluative than receiving all feedback from the teacher.

Some participants also asked whether they could continue using the prompt after the study. This refers to the final researcher-designed prompt described in the methodology chapter, which was used to structure feedback in the AI-supported practice sessions. Their interest suggested that they saw the prompt as useful beyond the research sessions. From a teaching perspective, this would be a useful strategy as the use of AI-supported practice may extend learning beyond class time if students are given a clear structure.

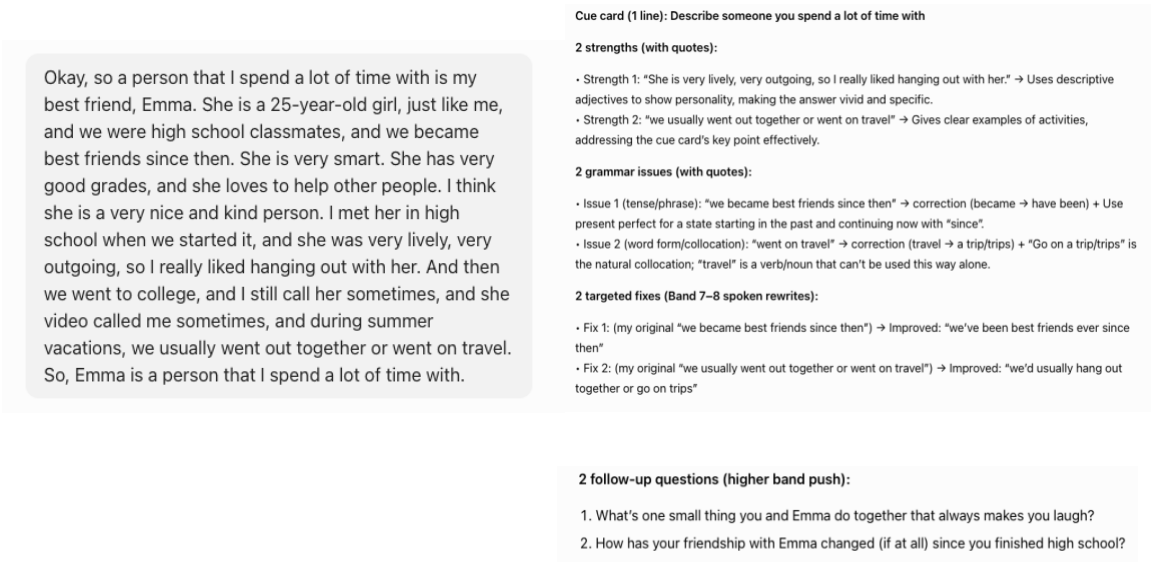
One of the strongest advantages of AI was the chat history. In traditional face-to-face speaking classes, much feedback is given orally and does not leave a lasting record. The same issue appeared in the teacher-led online sessions in this study. Feedback was mainly given orally in real time, supported by brief notes I made while listening to students' answers. However, because the sessions were not recorded or transcribed, the oral feedback was not captured in a complete written record. I could note the main ideas, useful phrases, errors, and points for improvement, but I could not record every word while also listening, diagnosing, and preparing feedback.

AI was stronger in this area because it preserved the full interaction. When developing the AI prompt, I initially focused mainly on whether the feedback was accurate, useful, and relevant to IELTS Speaking criteria. During the AI-supported sessions, however, I realised that the chat history itself was also valuable because it gave learners a written record they could review after class. The learner's answer, AI feedback, suggested rewrites, and follow-up questions remained visible in the chat history, as shown in Figure 10. Each learner could keep a written record of the AI interaction involving their own answer and review it after class. This turned speaking practice from an in-the-moment activity into a more tangible learning process.

Figure 10 shows an anonymised example of how learner output and AI feedback remained visible in the interaction history.

Figure 10

Anonymised AI Interaction Record



Note. Anonymised screenshot of an interaction with Doubao during an AI-supported speaking session. All identifying information has been removed.

I also noticed a difference in attention patterns between the two groups. In the teacher-led group, when I gave personalised feedback to one student, other students did not always appear equally engaged. The feedback was directed at that learner's specific answer, so other students may have felt it was less relevant to them.

In the AI-supported group, students seemed more likely to pay attention when AI feedback was shown. Several factors may have contributed to this. AI felt new and interesting. The feedback was also visible and structured. Because the prompt required exact quotes, learners could clearly see the link between the answer and the feedback. AI feedback was also slightly more general than teacher feedback, which may have made it easier for other learners to relate the points to their own speaking.

This observation should not be taken to mean that AI feedback was better than teacher feedback. The two forms of feedback created different kinds of learning attention. Teacher feedback was more personal, responsive, and nuanced. AI feedback was more visible, structured, and reusable. For me, this observation highlighted the potential of AI to make feedback more visible while also confirming the need for teacher guidance in helping students interpret and use it critically.

6.1.4 What I Learned as a Teacher

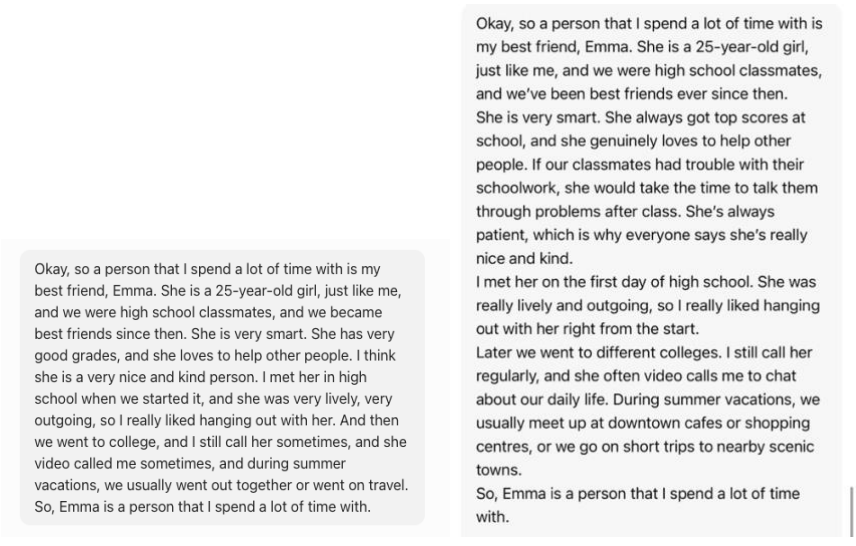
The project helped me develop a more balanced view of AI in language teaching. Before the study, I was interested in AI because it seemed efficient, accessible, and increasingly capable. After the study, I still see these strengths, but I also understand more clearly its limitations and am more aware of the areas where teacher judgement is needed.

One of the areas where the use of AI appeared especially helpful was answer development. AI feedback featured alternative expressions, expanded ideas, and immediate feedback. These features are useful for IELTS Speaking Part 2 because learners often need to produce more specific and developed answers. A short answer can be extended through examples, reasons, feelings, or comparisons, and AI was able to illustrate these approaches quickly.

Figure 11 provides an example of how AI feedback supported answer expansion by suggesting additional details and possible directions for development.

Figure 11

Example of AI-Supported Answer Expansion



Note. Screenshot of AI-generated feedback used to support answer expansion during speaking practice. All identifying information has been removed. The left screenshot shows the original answer, and the right screenshot shows the AI-expanded version.

The feedback provided by the AI programme used in this study was more limited in areas such as pronunciation and spoken delivery. Speaking involves not only wording and content. It also

includes pronunciation, rhythm, intonation, fluency, and communicative effect. General voice-input or automatic speech recognition systems often appear to prioritise recognising the intended message. This differs from pronunciation learning, which requires more diagnostic feedback on how the speech was produced. Recent work on Automatic Speech Recognition (ASR) and pronunciation learning suggests that ASR can support intelligibility-oriented practice, but such pronunciation-focused feedback requires careful design (McCrocklin & Levis, 2025).

The distinction mattered in my own study. During the AI-supported sessions, AI sometimes appeared to understand learners' spoken answers even when aspects of pronunciation or delivery were not fully clear. From a communication perspective, the message had been recognised by AI. From a teaching perspective, however, some pronunciation or delivery issues could remain less visible to the learner. In these instances, teacher feedback would have been useful because a teacher can notice whether pronunciation affects understanding, whether delivery sounds natural, and whether the answer works communicatively. These moments made me reflect on the limits of the particular AI tool used in this project and on the continuing need for teacher judgement when interpreting spoken performance.

The project also changed how I think about the relationship between AI and teachers. The most useful question is not whether AI replaces teachers, but how teachers' roles change as AI is increasingly used in language classrooms. Teachers are still needed as task designers, prompt designers, feedback evaluators, pronunciation coaches, emotional supporters, and guides for critical AI use.

Table 12 summarises the different affordances of AI feedback and teacher feedback identified through the study.

Table 12

Different Affordances of AI Feedback and Teacher Feedback

AI feedback useful for	Teacher feedback important for
Immediate feedback	Pronunciation diagnosis
Repeated practice	Real-time judgement
Answer expansion	Emotional support
Alternative expressions	Classroom interaction
Visible chat history	Interpreting learner needs
Low-pressure practice	Adapting feedback to individuals
Structured written feedback	Helping learners evaluate feedback critically

The complementary affordances observed in this study also suggest a possible role for hybrid models in future teaching design. AI could support repeated practice and first-stage feedback, while teachers could focus on deeper diagnosis, pronunciation, interaction, and critical guidance.

6.2 My Use of AI as a Researcher

AI is also becoming increasingly common in the research and writing process. This section reflects on how I used AI for literature searching and reading, learning research procedures, and supporting academic writing, as well as the risks and limitations I encountered.

6.2.1 Literature Searching and Reading

In my research workflow, I mainly used ChatGPT and Claude as AI support tools, while Doubao was the AI system used in the AI-supported speaking practice sessions described in the methodology chapter.

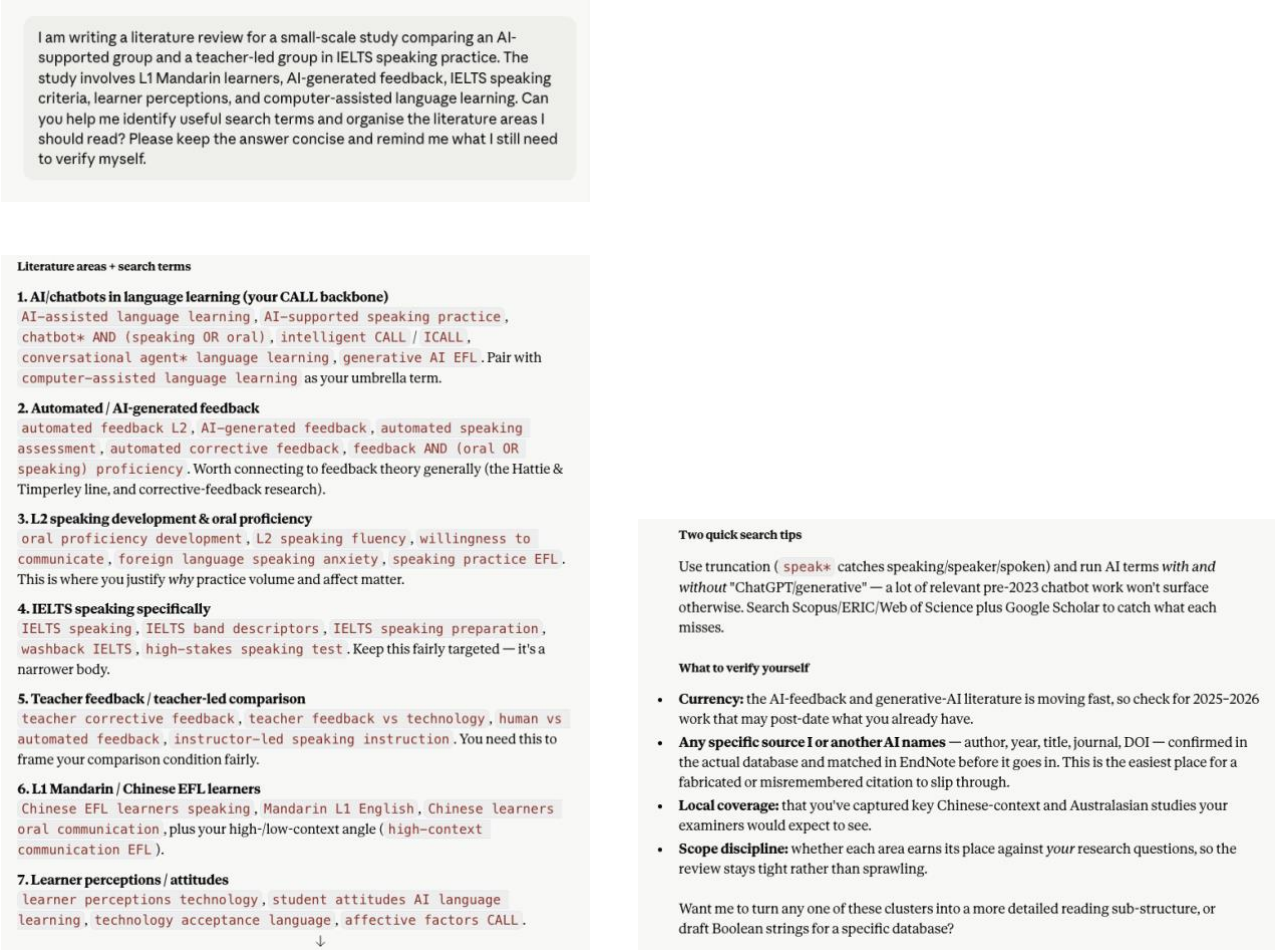
AI also supported my work as a researcher, especially in literature searching and source screening. This was useful because my research topic crossed several areas, including speaking practice, AI feedback, computer-assisted language learning, generative AI, and speaking assessment. AI helped me identify related search terms, locate possible reading directions, and organise the literature across these overlapping fields.

AI proved helpful during the early stages of my research. Thus, I used it to identify search terms, related concepts, and possible reading directions. For example, AI identified relevant key words to explore, such as AI feedback, automated feedback, dialogue-based CALL, generative AI in language learning, IELTS speaking practice, Automatic Speech Recognition (ASR), and learner perceptions. These suggestions helped me approach the literature more systematically.

Figure ***12*** provides an example of how I used AI during the literature-searching process. The first interaction shows how AI helped me generate and organise relevant search terms, while the second shows how it suggested ways to refine searches and check the relevance and currency of the literature.

Figure 12

Example of AI Support for Identifying Literature Search Terms



Note. Screenshots of the researcher's interactions with Claude during the literature-searching process. The left screenshot shows AI-generated search terms related to the study topic, while the right screenshot shows suggestions for refining search strategies and identifying relevant literature

Figure 13 shows an example of how AI was used to locate and check potentially relevant literature.

Figure 13

Example of AI Support for Locating and Checking Relevant Literature



Note. Screenshots of the researcher's interaction with Claude during the literature-searching process. The screenshots show the search criteria provided to the AI, the candidate studies it identified, its explanations of their relevance to the research topic, and its suggestions about which studies were the closest matches. The AI also prompted the researcher to verify the publication details and sources independently.

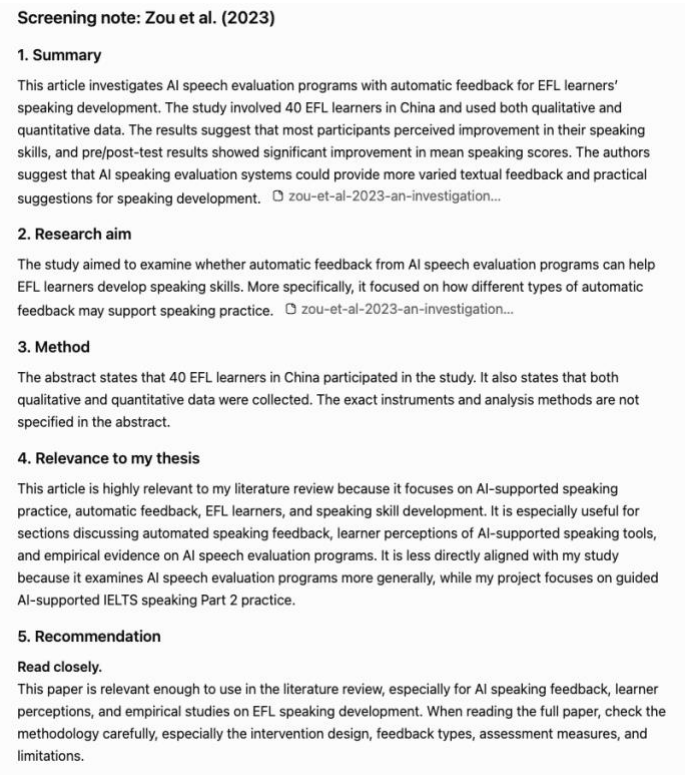
AI also offered useful support in the identification of relevant texts. When reading a new paper, I could ask AI to summarise the abstract, identify the research aim, explain the method, and suggest

possible relevance to my study. These summaries offered helpful syntheses of texts, which both helped identify the texts I needed to read in greater depth and established an overview of the scope of the topic.

Figure 14 shows an example of AI-supported initial screening of a journal article.

Figure 14

Example of AI-Supported Initial Screening of a Journal Article



Screening note: Zou et al. (2023)

1. Summary

This article investigates AI speech evaluation programs with automatic feedback for EFL learners' speaking development. The study involved 40 EFL learners in China and used both qualitative and quantitative data. The results suggest that most participants perceived improvement in their speaking skills, and pre/post-test results showed significant improvement in mean speaking scores. The authors suggest that AI speaking evaluation systems could provide more varied textual feedback and practical suggestions for speaking development. zou-et-al-2023-an-investigation...

2. Research aim

The study aimed to examine whether automatic feedback from AI speech evaluation programs can help EFL learners develop speaking skills. More specifically, it focused on how different types of automatic feedback may support speaking practice. zou-et-al-2023-an-investigation...

3. Method

The abstract states that 40 EFL learners in China participated in the study. It also states that both qualitative and quantitative data were collected. The exact instruments and analysis methods are not specified in the abstract.

4. Relevance to my thesis

This article is highly relevant to my literature review because it focuses on AI-supported speaking practice, automatic feedback, EFL learners, and speaking skill development. It is especially useful for sections discussing automated speaking feedback, learner perceptions of AI-supported speaking tools, and empirical evidence on AI speech evaluation programs. It is less directly aligned with my study because it examines AI speech evaluation programs more generally, while my project focuses on guided AI-supported IELTS speaking Part 2 practice.

5. Recommendation

Read closely.
This paper is relevant enough to use in the literature review, especially for AI speaking feedback, learner perceptions, and empirical studies on EFL speaking development. When reading the full paper, check the methodology carefully, especially the intervention design, feedback types, assessment measures, and limitations.

Note. Screenshot of an AI-generated preliminary summary used to support initial article screening. The original article was subsequently read, checked by the researcher and included in the literature review.

Importantly, I used AI summaries only as a first point of orientation. The summaries generated by AI helped me decide which papers might be worth reading more closely, but they did not replace my own reading of the original sources. In some cases, AI made a paper sound more directly relevant to my study than it actually was or simplified the method and findings too much. For this reason, I read each source carefully before deciding whether it was relevant for my study and should be included in my literature review.

I also used AI to check practical publication details, such as DOI information, journal names, and links to possible sources. However, these details were not accepted automatically. I checked them manually through Google Scholar, the AUT library, journal websites, or the DOI link before using the source. Similarly, when AI suggested that a paper was relevant to my study, I treated this only

as a suggestion. A source was included in the thesis only after I had read the original article and determined that it was relevant to my research focus.

Table 13 summarises the literature search and reading workflow and identifies which tasks at each stage were performed by AI and which remained the researcher's responsibility.

Table 13

AI-Supported Literature Searching and Reading Workflow

Step	AI support	My responsibility
Identify search directions	Suggest keywords and concepts	Decide which concepts fit the study
Search literature	Generate searchable terms	Search databases, Google Scholar, and library resources
Screen papers	Summarise abstracts or sections	Read the original source
Check credibility	Provide DOI, links, and publication details	Verify the source manually
Use in thesis	Suggest possible relevance	Cite only checked sources

I gradually developed a more careful workflow for using AI in my research process. When asking AI for literature support, I asked it to provide links, DOI information, publication details, and, where possible, recent studies from higher-quality journals. I also sometimes gave the same prompt to both ChatGPT and Claude, then compared their responses to see which suggestions were more relevant, specific, or useful. Comparing outputs across these two AI systems became another safeguard because it encouraged me to evaluate the quality of the suggestions rather than simply accepting the first response.

Before using any source in the thesis, I treated AI-generated outputs as leads rather than evidence, read the original article myself, and assessed its credibility and relevance to the research topic before deciding whether to include it.

6.2.2 AI as a Process Tutor

AI also supported me by helping me learn unfamiliar research processes. This was especially important during data analysis. Before this project, I had limited experience with statistical analysis software and was not confident about how to analyse the quantitative data. I had collected several types of data, including pre- and post-test IELTS scores, Likert-scale survey responses, and open-ended comments. I knew what data I had, but I was less clear about how to organise the analysis process.

AI helped me turn this uncertainty into more concrete learning steps. I could explain my dataset to AI, including the two groups, the pre- and post-test structure, the small sample size, the IELTS score data, and the survey responses. Instead of simply giving me answers to copy, AI helped me identify what I needed to learn next. For example, it suggested possible analysis tools, including JASP, and pointed me towards the kinds of tutorials, examples, and explanations that would help

me understand the process. I then watched online tutorial videos, tried the steps outlined in the software, and returned to AI when I needed help understanding what a test, table, or output meant.

In this sense, AI worked more like a process tutor than a source of final answers. It helped me break the research process into smaller steps, such as preparing the dataset, organising variables, choosing suitable tests, running the analysis, reading the output, and reporting the results. However, AI did not do the analysis for me. I still had to learn how to complete these tasks myself by checking the data, operating the software, comparing possible analysis options, and deciding whether the results made sense in relation to my research questions.

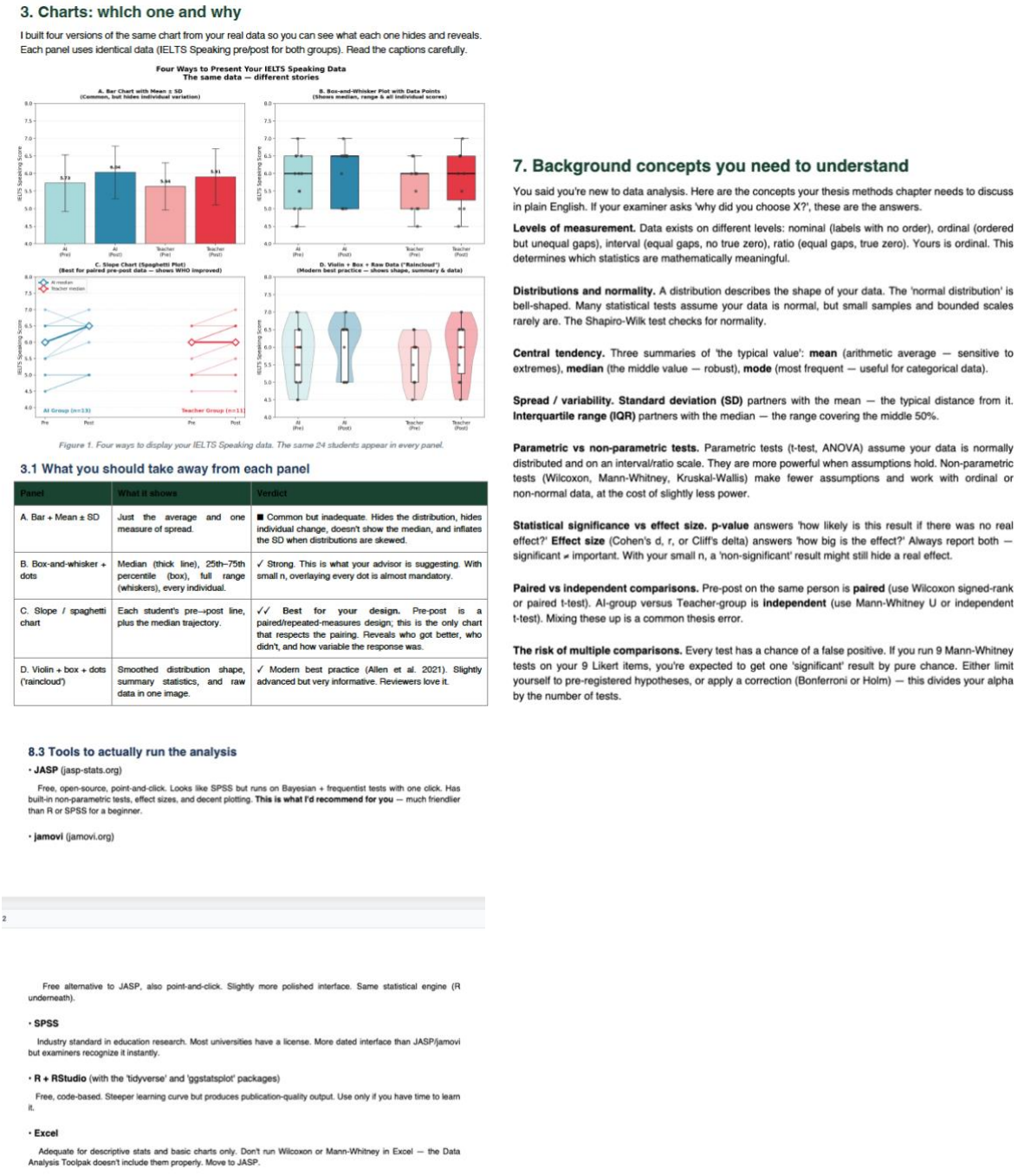
AI also helped me understand the reasoning behind reporting choices. For example, I asked whether I should report means or medians, how to summarise Likert-scale responses, and what types of charts might be suitable for presenting the findings. These questions helped me understand how data type, sample size, and research purpose affect the way results should be analysed and presented. For example, AI helped me think through why medians and interquartile ranges may be useful for small samples or ordinal-style responses, while means and standard deviations may still be reported descriptively where appropriate.

The same kind of support extended to the presentation of my findings. I asked AI for advice on how to visualise IELTS score changes, group comparisons, Likert responses, and integrated findings across quantitative and qualitative data. Some suggestions were too complicated or not suitable for the thesis. However, comparing these options helped me understand why some figures or tables were clearer than others. Later, when revising the findings tables, I found that the more specific I phrased the prompt, the more useful AI's advice became. General requests often produced long or unclear outputs, while precise instructions, such as asking AI to shorten a table while keeping key data, produced better results.

Figure 15 and Figure 16 provide examples of how AI helped me learn the steps involved in data analysis and the interpretation of statistical output.

Figure 15

Example of AI Support for Learning Data-Analysis Procedures: Part 1



3. Charts: which one and why

I built four versions of the same chart from your real data so you can see what each one hides and reveals. Each panel uses identical data (IELTS Speaking pre/post for both groups). Read the captions carefully.

Figure 1. Four ways to display your IELTS Speaking data. The same 24 students appear in every panel.

3.1 What you should take away from each panel

Panel	What it shows	Verdict
A. Bar + Mean ± SD	Just the average and one measure of spread.	■ Common but inadequate. Hides the distribution, hides individual change, doesn't show the median, and inflates the SD when distributions are skewed.
B. Box-and-whisker + dots	Median (thick line), 25th–75th percentile (box), full range (whiskers), every individual.	✓ Strong. This is what your advisor is suggesting. With small n, overlaying every dot is almost mandatory.
C. Slope / spaghetti chart	Each student's pre→post line, plus the median trajectory.	✓✓ Best for your design. Pre-post is a paired/repeated-measures design; this is the only chart that respects the pairing. Reveals who got better, who didn't, and how variable the response was.
D. Violin + box + dots ('raincloud')	Smoothed distribution shape, summary statistics, and raw data in one image.	✓ Modern best practice (Allen et al. 2021). Slightly advanced but very informative. Reviewers love it.

7. Background concepts you need to understand

You said you're new to data analysis. Here are the concepts your thesis methods chapter needs to discuss in plain English. If your examiner asks 'why did you choose X?', these are the answers.

Levels of measurement. Data exists on different levels: nominal (labels with no order), ordinal (ordered but unequal gaps), interval (equal gaps, no true zero), ratio (equal gaps, true zero). Yours is ordinal. This determines which statistics are mathematically meaningful.

Distributions and normality. A distribution describes the shape of your data. The 'normal distribution' is bell-shaped. Many statistical tests assume your data is normal, but small samples and bounded scales rarely are. The Shapiro-Wilk test checks for normality.

Central tendency. Three summaries of 'the typical value': **mean** (arithmetic average — sensitive to extremes), **median** (the middle value — robust), **mode** (most frequent — useful for categorical data).

Spread / variability. Standard deviation (SD) partners with the mean — the typical distance from it. **Interquartile range (IQR)** partners with the median — the range covering the middle 50%.

Parametric vs non-parametric tests. Parametric tests (t-test, ANOVA) assume your data is normally distributed and on an interval/ratio scale. They are more powerful when assumptions hold. Non-parametric tests (Wilcoxon, Mann-Whitney, Kruskal-Wallis) make fewer assumptions and work with ordinal or non-normal data, at the cost of slightly less power.

Statistical significance vs effect size. p-value answers 'how likely is this result if there was no real effect?' **Effect size** (Cohen's d, r, or Cliff's delta) answers 'how big is the effect?' Always report both — significant ≠ important. With your small n, a 'non-significant' result might still hide a real effect.

Paired vs independent comparisons. Pre-post on the same person is **paired** (use Wilcoxon signed-rank or paired t-test). AI-group versus Teacher-group is **independent** (use Mann-Whitney U or independent t-test). Mixing these up is a common thesis error.

The risk of multiple comparisons. Every test has a chance of a false positive. If you run 9 Mann-Whitney tests on your 9 Likert items, you're expected to get one 'significant' result by pure chance. Either limit yourself to pre-registered hypotheses, or apply a correction (Bonferroni or Holm) — this divides your alpha by the number of tests.

8.3 Tools to actually run the analysis

• **JASP** (jasp-stats.org)

Free, open-source, point-and-click. Looks like SPSS but runs on Bayesian + frequentist tests with one click. Has built-in non-parametric tests, effect sizes, and decent plotting. **This is what I'd recommend for you** — much friendlier than R or SPSS for a beginner.

• **jamovi** (jamovi.org)

2

Free alternative to JASP, also point-and-click. Slightly more polished interface. Same statistical engine (R underneath).

• **SPSS**

Industry standard in education research. Most universities have a license. More dated interface than JASP/jamovi but examiners recognize it instantly.

• **R + RStudio** (with the 'tidyverse' and 'ggstatsplot' packages)

Free, code-based. Steeper learning curve but produces publication-quality output. Use only if you have time to learn it.

• **Excel**

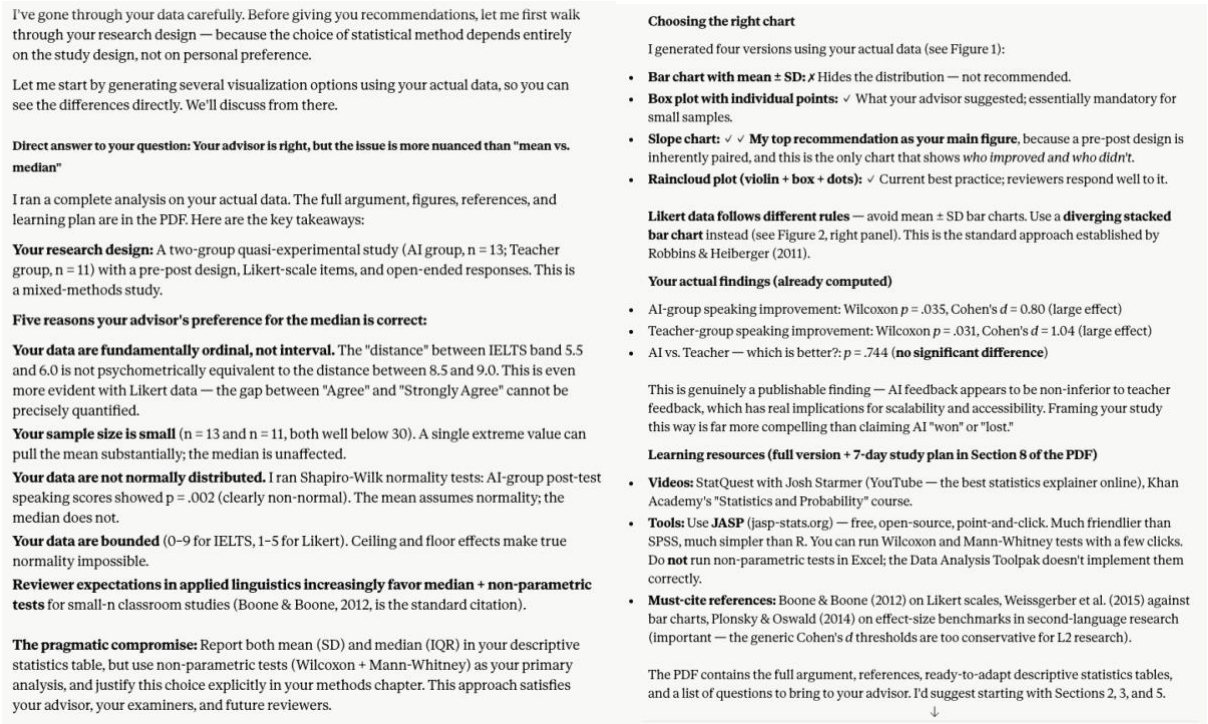
Adequate for descriptive stats and basic charts only. Don't run Wilcoxon or Mann-Whitney in Excel — the Data Analysis Toolpak doesn't include them properly. Move to JASP.

Note. Screenshots of the researcher's interaction with ChatGPT showing guidance on comparing data visualisation options, understanding key statistical concepts, and selecting software tools for conducting the analysis.

Overall, AI did not replace my learning of the research process or the guidance I received from my supervisors. Rather, it worked as an additional support tool between supervision meetings. It helped me break unfamiliar tasks into manageable steps and locate useful resources. The main academic direction and final decisions still depended on supervision, research training, and my own judgement as a researcher.

Figure 16

Example of AI Support for Learning Data-Analysis Procedures: Part 2



I've gone through your data carefully. Before giving you recommendations, let me first walk through your research design — because the choice of statistical method depends entirely on the study design, not on personal preference.

Let me start by generating several visualization options using your actual data, so you can see the differences directly. We'll discuss from there.

Direct answer to your question: Your advisor is right, but the issue is more nuanced than "mean vs. median"

I ran a complete analysis on your actual data. The full argument, figures, references, and learning plan are in the PDF. Here are the key takeaways:

Your research design: A two-group quasi-experimental study (AI group, n = 13; Teacher group, n = 11) with a pre-post design, Likert-scale items, and open-ended responses. This is a mixed-methods study.

Five reasons your advisor's preference for the median is correct:

Your data are fundamentally ordinal, not interval. The "distance" between IELTS band 5.5 and 6.0 is not psychometrically equivalent to the distance between 8.5 and 9.0. This is even more evident with Likert data — the gap between "Agree" and "Strongly Agree" cannot be precisely quantified.

Your sample size is small (n = 13 and n = 11, both well below 30). A single extreme value can pull the mean substantially; the median is unaffected.

Your data are not normally distributed. I ran Shapiro-Wilk normality tests: AI-group post-test speaking scores showed p = .002 (clearly non-normal). The mean assumes normality; the median does not.

Your data are bounded (0–9 for IELTS, 1–5 for Likert). Ceiling and floor effects make true normality impossible.

Reviewer expectations in applied linguistics increasingly favor median + non-parametric tests for small-n classroom studies (Boone & Boone, 2012, is the standard citation).

The pragmatic compromise: Report both mean (SD) and median (IQR) in your descriptive statistics table, but use non-parametric tests (Wilcoxon + Mann-Whitney) as your primary analysis, and justify this choice explicitly in your methods chapter. This approach satisfies your advisor, your examiners, and future reviewers.

Choosing the right chart

I generated four versions using your actual data (see Figure 1):

- **Bar chart with mean ± SD: ✗** Hides the distribution — not recommended.
- **Box plot with individual points:** ✓ What your advisor suggested; essentially mandatory for small samples.
- **Slope chart:** ✓ ✓ **My top recommendation as your main figure**, because a pre-post design is inherently paired, and this is the only chart that shows *who improved and who didn't*.
- **Raincloud plot (violin + box + dots):** ✓ Current best practice; reviewers respond well to it.

Likert data follows different rules — avoid mean ± SD bar charts. Use a **diverging stacked bar chart** instead (see Figure 2, right panel). This is the standard approach established by Robbins & Heiberger (2011).

Your actual findings (already computed)

- AI-group speaking improvement: Wilcoxon $p = .035$, Cohen's $d = 0.80$ (large effect)
- Teacher-group speaking improvement: Wilcoxon $p = .031$, Cohen's $d = 1.04$ (large effect)
- AI vs. Teacher — which is better?: $p = .744$ (**no significant difference**)

This is genuinely a publishable finding — AI feedback appears to be non-inferior to teacher feedback, which has real implications for scalability and accessibility. Framing your study this way is far more compelling than claiming AI "won" or "lost."

Learning resources (full version + 7-day study plan in Section 8 of the PDF)

- **Videos:** StatQuest with Josh Starmer (YouTube — the best statistics explainer online), Khan Academy's "Statistics and Probability" course.
- **Tools:** Use **JASP** (jasp-stats.org) — free, open-source, point-and-click. Much friendlier than SPSS, much simpler than R. You can run Wilcoxon and Mann-Whitney tests with a few clicks. Do **not** run non-parametric tests in Excel; the Data Analysis Toolpak doesn't implement them correctly.
- **Must-cite references:** Boone & Boone (2012) on Likert scales, Weissgerber et al. (2015) against bar charts, Plonsky & Oswald (2014) on effect-size benchmarks in second-language research (important — the generic Cohen's *d* thresholds are too conservative for L2 research).

The PDF contains the full argument, references, ready-to-adapt descriptive statistics tables, and a list of questions to bring to your advisor. I'd suggest starting with Sections 2, 3, and 5.

Note. Screenshots of the researcher's interaction with Claude showing guidance on selecting appropriate descriptive and inferential statistics for the dataset, interpreting the reasons for these choices, and comparing suitable visualisations for presenting the results.

6.2.3 AI as a Writing Support Tool

AI also supported my writing process. In this project, I used AI as a writing support tool to improve structure, sentence flow, clarity, and coherence. It was especially helpful when I already had an idea but struggled to express it clearly in academic English. However, I did not treat AI-generated text as final thesis writing. The final writing still required me to check the meaning, add study-specific details, connect the text to my evidence, and decide whether the wording accurately represented my argument.

One important function was cautious academic wording. The study had a small sample size and a short intervention, so I needed to ensure that my conclusions were appropriately hedged and that I avoided broad generalisations. For example, I could not write that the AI-supported group and teacher-led group were equivalent. A more appropriate wording was that the study did not detect a significant between-group difference in speaking gain. AI helped me explore this kind of careful phrasing, but I still had to check that the final wording matched the data, the research design, and the methodological limitations of the study.

AI also supported practical writing tasks. It helped me plan tables, simplify figures, organise chapters, prepare supervisor meeting questions, and break large tasks into smaller steps. This

support helped me meet my deadlines and ensured I could complete the study within the given time frame. Instead of facing broad tasks, such as “finish the chapter”, I could focus on smaller steps, such as drafting one section, adding one example, or revising one table. In this way, AI supported the writing process, but it did not replace the work of building the thesis argument.

Writer’s block was another area where AI helped. At several points, I found it difficult to start because the thesis felt too large. AI helped me begin with headings, bullet points, rough paragraphs, or a clearer plan. This support was not only practical. It also helped me maintain momentum during the long and sometimes lonely writing process. However, these AI-supported drafts still needed substantial revision from me. I had to decide what was relevant, remove general statements, add details from my own study, and make sure the writing sounded like my own work.

As a researcher writing in my second language, I found AI especially useful. Sometimes I could express complex thoughts in Chinese first and then ask AI to help organise or translate them into English. This allowed me to focus first on content and logic before refining the academic expression. For multilingual researchers, this can reduce the burden of language conversion during early thinking and drafting. At the same time, translation or rewriting support did not remove my responsibility as the writer. I still needed to check whether the English version carried my intended meaning and whether it was suitable for the thesis context.

AI also helped with grammar, sentence structure, paraphrasing, and awkward expressions. However, AI’s output was not always accurate or suitable. Sometimes AI writing became too general, too formal, or too distant from my study. When this happened, I had to add back specific details and revise the sentences so that they reflected my evidence and argument. This experience reminded me that fluency is not the same as quality. A sentence can sound academic but still be too vague, unsupported, or disconnected from the research.

Table 14 summarises several writing-related experiences and what they taught me about AI-supported research writing.

Table 14

Researcher Learning From AI-Supported Writing

Experience	What I learned
Using AI for cautious academic wording	AI can suggest careful phrasing, but the final wording must match the data and research design.
Using AI to break writing into smaller tasks	AI can support writing management, but it does not build the thesis argument by itself.
Comparing AI systems	Different AI tools produce different answers, so evaluation remains essential.
Revising tables with my supervisor	The quality of AI output depends on human judgement, not AI alone.

Throughout this project, I gradually realised that effective AI use depended on professional judgement as much as prompt design. Although AI could produce well-written answers within seconds, it could not determine whether those answers were appropriate for my research. The

responsibility for identifying problems, evaluating outputs, and making final decisions remained mine.

One strategy that helped me was forcing myself to think before prompting. Rather than immediately asking AI for solutions, I first clarified the current situation, the goal, what I had already considered, the constraints of the task, and the type of help I needed. This habit encouraged independent thinking before involving AI and reduced the risk of accepting generated answers too quickly. Over time, I found that the quality of AI responses often reflected the clarity of my own thinking.

I also learned not to rely on a single AI system. On several occasions, I entered the same prompt into both ChatGPT and Claude and compared their responses. The two systems often differed in organisation, level of detail, tone, and practical usefulness. Neither answer was consistently better than the other. Comparing multiple responses encouraged me to evaluate the strengths and weaknesses of each response rather than accepting the first plausible answer. The final decision was always based on my own judgement.

One experience I had during the thesis writing process clearly illustrated what AI could and could not do in this research. While revising one of the findings tables, I first asked AI to improve the presentation of the table. The revised version was still too long and did not communicate the results clearly. During a supervision meeting, my supervisor immediately recognised the underlying problem. Rather than asking AI simply to improve the wording, she suggested a more specific goal: present the numerical findings more directly and remove unnecessary text. Using that revised prompt, AI produced a substantially clearer table.

What struck me was not simply that my supervisor wrote a better prompt. More importantly, she recognised the problem almost immediately, whereas I did not. The limiting factor was not the AI tool itself, but my own ability to identify weaknesses in the work and communicate a clear objective. AI could generate many alternative versions, but it could not decide which version best served the academic purpose of the table. That judgement still required disciplinary experience and critical evaluation.

That experience reshaped how I understood AI-supported writing and research. At the beginning of the project, I saw AI mainly as a tool for generating ideas, improving expression, and increasing productivity. By the end of the thesis, I understood that AI’s usefulness depended heavily on the researcher’s guidance and direction. In this sense, AI is only as useful as the instructions it receives, which again requires a good understanding of what needs to be accomplished.

6.2.4 Risks and Limitations of Using AI in Research

Despite its benefits, AI also introduced important risks throughout this project. Some related to the reliability of AI-generated information, while others concerned responsible research practice. Table

15 summarises the main risks identified during the project and the safeguards used to manage them.

Table 15

Risks and Safeguards in AI-Supported Research

Risk	What could go wrong	Safeguard
Inaccurate information	False references or unsupported claims	Verify original sources
Cognitive offloading	AI replaces independent thinking	Think before prompting and evaluate outputs critically
Privacy risk	Participant data exposure	Anonymise examples and follow ethics approval
Attention fragmentation	Waiting for AI responses or switching between tools leads to distraction	Keep phone away, use a timer, and prepare a parallel academic task

The most obvious risk was inaccurate information. AI occasionally generated incorrect references, DOI details, summaries, or unsupported claims that appeared convincing. This was particularly relevant during the literature review, where inaccurate information could easily be incorporated into academic writing if left unchecked. I therefore treated AI-generated references and summaries only as starting points and verified publication details, original sources, and the relevance of each paper before it was included in the thesis.

A second risk was becoming over-reliant on AI, a process often referred to as cognitive offloading. Cognitive offloading means using an external tool to reduce the cognitive demands of a task (Risko & Gilbert, 2016). In this project, this risk appeared when AI quickly produced coherent and persuasive text, making it tempting to accept its responses without sufficient independent analysis. Although AI was valuable for exploring ideas and improving drafts, it could not replace the researcher's responsibility for interpreting evidence, constructing arguments, or making methodological decisions. Throughout the project, I regarded AI as a support tool only and made sure that all major academic decisions remained my own.

A third risk was attention fragmentation, which refers here to the way attention can become divided or disrupted when moving between tasks, tools, or sources of information. This is related to research on interruptions and task switching, which suggests that interruptions can increase stress and that attention may not fully move from one task to another immediately after a switch (Leroy, 2009; Mark et al., 2008). In this project, short AI responses usually appeared quickly, but more complex tasks, such as reviewing a long section, comparing drafts, or producing detailed feedback, sometimes took a few minutes. I noticed that even this short waiting time could interrupt my focus, because it was tempting for me to check my phone, do another quick online search, or do other tasks not related to the thesis task at hand. To reduce this, I tried to keep my phone away, use a 5-minute timer, and prepare a small parallel academic task, such as checking a reference, formatting a table, or reading a short section while waiting for the AI output. This helped me use AI as part of the research process without letting the process become another source of distraction.

Ethical considerations were equally important. The study involved student learning data, and participant privacy needed to be protected throughout the research process. Examples and screenshots included in the thesis were anonymised, and identifiable participant information was never entered into AI systems. AI use therefore remained consistent with the ethical approval granted for this study.

Overall, this project reinforced that AI should be viewed as a supportive research tool rather than an autonomous research partner. While AI can substantially improve efficiency, responsibility for evidence, ethics, interpretation, and the quality of academic work continue to rest with the researcher.

6.3 Key Takeaways

My experiences of using AI in both teaching and research led to several practical lessons. This section brings together the main takeaways for language learners, teachers, and researchers and presents the prompt design guidelines developed through the project.

6.3.1 For Language Learners

For language learners, the main value of AI is that it can create more opportunities to practise, revise, and reflect. However, AI is most useful when learners use it as part of a practice cycle rather than as a source of ready-made answers. Simply asking AI to “improve my answer” is less useful than asking for specific feedback on grammar, answer expansion, natural spoken expressions, or coherence.

Learners also need to avoid turning AI-generated answers into another form of template dependence. AI can suggest useful expressions and possible directions for development, but learners still need to adapt these suggestions into their own language. Follow-up questions are particularly useful for IELTS Speaking Part 2 because they help learners think of more details, examples, reasons, and personal experiences. The chat history can also become a learning record, allowing learners to review their original answers, feedback, corrections, and revised expressions after class.

6.3.2 For Teachers

For teachers, the main takeaway is that AI does not need to be seen as a threat to teaching. Instead, it can be used as a pedagogical resource when it is carefully designed and guided. Simply telling students to use AI is unlikely to be enough. Without structure, learners may ask vague questions, accept poor feedback, or copy AI-generated answers. Teachers therefore need to design the task, prompt, feedback routine, and follow-up activity.

Prompt design can be understood as part of teaching design. In this study, the prompt shaped what feedback learners received, what they noticed, and what they did next. Teachers are also

needed to help students develop the ability to evaluate AI feedback, including whether suggestions are accurate, natural, and suitable for IELTS Speaking. AI can be used to extend teaching practice, but it still needs to be guided by professional judgement.

6.3.3 For Researchers

For researchers, AI can support early-stage orientation, organisation, and writing clarity, especially when a project crosses several fields or when the researcher is writing in a second language. It can help generate search terms, summarise papers, organise notes, refine wording, and improve sentence flow.

However, AI output should not be treated as evidence. References, DOI information, summaries, and claims still need to be checked against original sources. AI can suggest possible directions, but the final argument must come from the data, methodology, literature, and the researcher's interpretation of the findings. Researchers also need to follow institutional rules, protect participant information, and acknowledge AI support where required. Researchers may also benefit from comparing outputs from different AI systems, as this can help them evaluate suggestions more critically rather than relying on one response.

A further practical lesson is the importance of organising AI use. During the later revision stage, I found it difficult to locate some earlier AI conversations because different tasks, chapters, and research processes were often mixed within the same chat. This made it harder to retrieve useful prompts, screenshots, and process records. For researchers, it may be useful to separate AI conversations by chapter or task, such as literature review, methodology, data analysis, findings visualisation, and reflection. If organised clearly, AI chat history can become a useful record of the research process, rather than only a temporary writing aid.

6.3.4 AI Prompt Design Guidelines

A practical takeaway from this reflection is that AI prompt design matters. For researchers, teachers, and learners, the quality of an AI response depends not only on the AI tool itself but also on how clearly the task is framed. General prompting guidance emphasises the importance of giving clear instructions, providing context, specifying the task, using references or examples where relevant, evaluating the output, and revising the prompt through iteration (Google Cloud, 2026; Grow with Google Editorial Team, 2026; McHugh-Johnson, 2024). In educational contexts, these principles are especially important because a weak prompt may produce vague, generic, or unsuitable feedback, while a stronger prompt can make AI support more specific, relevant, and actionable.

Table 16 presents practical AI prompt-design guidelines for researchers, teachers, and learners.

Table 16

AI Prompt Design Guidelines for Researchers, Teachers, and Learners

Guideline	Explanation
State the situation	Give the AI enough context to understand the learning, teaching, or research situation.
Define the goal	Explain what the prompt is expected to achieve, such as feedback, revision, explanation, planning, or idea generation.
Assign a role	Give the AI a clear role, such as IELTS speaking teacher, academic writing tutor, research assistant, or critical reader.
Specify criteria	Tell the AI what standard to use, such as IELTS Speaking criteria, assignment requirements, marking rubrics, or research questions.
Require evidence	Ask the AI to refer to specific words, examples, data, or parts of the learner's answer instead of giving general comments.
Limit the scope	Avoid asking for too much at once. A focused prompt usually produces more useful output than a broad prompt.
Specify style	Control the tone, register, language level, and format of the response so that it suits the intended user.
Make feedback actionable	Ask for concrete next steps, examples, corrections, or practice tasks rather than only general evaluation.
Build in interaction	Design the prompt so that the user continues thinking, responding, revising, or practising instead of passively receiving an answer.
Evaluate and revise	Test the prompt, check whether the output is useful and accurate, and revise the prompt when the response is unclear or unsuitable.

Note. The guidelines were developed from the researcher's AI-supported workflow and informed by general prompting principles, including context, task, persona, format, specificity, and iterative refinement (Davis-Muffett, 2024; Google Cloud, 2026; Grow with Google Editorial Team, 2026).

These guidelines are not a fixed formula. Different users and tasks require different prompt designs. However, they show that effective prompting is a deliberate process. A useful prompt gives the AI context, purpose, standards, boundaries, and an expected output format. It also leaves space for critical judgement, because the user still needs to check whether the response is accurate, ethical, and appropriate for the intended purpose.

6.4 Chapter Summary

This chapter reflected on my use of AI as both a teacher and a researcher. As a teacher, I used AI to support IELTS Speaking Part 2 practice through a structured prompt and guided practice cycle. The project suggested that AI can be used to provide immediate feedback, repeated practice, answer-expansion support, and visible learning records within a structured, teacher-guided setting. It also showed that teachers remain important for task design, prompt design, pronunciation feedback, and professional judgement. This reflection therefore reinforces the empirical finding that AI-supported and teacher-led feedback offered different but potentially complementary forms of support: AI-supported practice was perceived as particularly useful for expansion and repeated revision, while teacher-led feedback was valued for pronunciation, live explanation, and real-time judgement.

As a researcher, I used AI to support reading, thinking, writing, and organisation. AI was useful for initial orientation, organising messy ideas, improving writing clarity, and supporting multilingual

academic writing. At the same time, the chapter identified several risks of AI use for academic writing, including inaccurate information, cognitive offloading, and privacy concerns. These risks need safeguards such as checking original sources, thinking before prompting, managing attention, evaluating AI output critically, and protecting participant information.

Overall, this reflection suggests that AI's value in education and research does not come from automation alone. It depends on how human users define the purpose, design the process, evaluate the output, and take responsibility for final decisions. For me, this project shifted the question from whether AI can replace teachers or researchers to how teachers and researchers can use AI critically, ethically, and effectively.

Chapter Seven: Conclusion

This thesis examined how AI-supported and teacher-led feedback supported IELTS speaking practice for L1 Mandarin EFL learners. The study was motivated by my practical interest, as an IELTS speaking teacher and researcher, in how generative AI could be used in EFL speaking practice and how AI-supported feedback might compare with teacher-led feedback in a structured IELTS speaking course. The study used a mixed methods methodology incorporating a standard comparative experiment. Twenty-four L1 Mandarin EFL learners took part in a four-week online IELTS Speaking programme, which included eight structured practice sessions. Participants were allocated to either an AI-supported feedback group or a teacher-led feedback group. In the AI-supported sessions, learners received feedback through a teacher-designed AI prompt, while in the teacher-led sessions, feedback was provided directly by the researcher-teacher. The study compared IELTS Speaking score changes across the two conditions and examined participants' experiences through Likert-scale survey responses and open-ended comments.

The quantitative findings showed that both groups made significant within-group gains in IELTS Speaking; however, the between-group comparison did not detect a significant difference in speaking gain. The qualitative findings added further detail by showing how learners responded to the two modes of feedback in class. Participants in both groups reported that after the course they felt more confident, were able to express ideas more clearly, and were better at developing their spoken responses. However, each group also identified different gains that were associated with its specific feedback mode. Because the study was small in scale and short in duration, these findings should be interpreted cautiously. Overall, the findings suggest that the structured speaking programme was useful in both conditions, rather than showing that one feedback mode was clearly more effective than the other.

The study contributes to current discussions about generative AI in EFL speaking education by suggesting that AI-supported and teacher-led feedback may have different but complementary functions. AI-supported feedback appeared especially useful for answer expansion, alternative expressions, and repeated practice. AI-supported practice also had the advantage of leaving a traceable record of learner output, feedback, and class interaction, which could be reviewed later. Teacher-led feedback was more strongly associated with pronunciation correction, a stronger sense of human interaction, and a more specific and flexible diagnosis of speaking performance. These findings suggest that AI-supported feedback is not the same as teacher-led feedback, but that this difference can be used productively in the speaking classroom. More needs to be understood about the kinds of learning support AI can provide, what it does and does not do well, and how it can be integrated into carefully designed speaking practice.

The findings also offer a tentative link back to the high-context and low-context communication framework discussed earlier in the thesis. Participants' responses indicate that both feedback

modes may have supported more explicit spoken communication, although in different ways: AI-supported feedback through answer expansion and elaboration, and teacher-led feedback through direct clarification and real-time interaction.

A further contribution of this thesis is its evidence-informed consideration of how AI can be used for teaching, learning, and in academia. The reflection chapter showed how AI was used both in the speaking classes and across the wider research process. This experience led to practical guidelines for teachers, researchers, and learners. For teachers, the findings suggest that AI feedback needs clear task design, prompt design, and classroom guidance. For researchers, AI can improve research efficiency, but its output still needs checking, source verification, and professional judgement. For learners, AI can be a convenient and accessible tool for additional speaking practice beyond scheduled class time, but they still need to evaluate suggestions critically and use them selectively.

Future research could build on this study by using larger samples, longer programmes, and different learner groups. It would also be useful to compare guided and unguided AI use, examine hybrid models that combine AI-supported rehearsal with teacher-led feedback, and test how different AI platforms, models, and prompt designs shape the quality of feedback learners receive. Future studies could also collect audio data to analyse pronunciation, fluency, interaction, and learner uptake more closely, and examine how AI-supported feedback may work in other language skills such as writing, listening, or reading. Overall, this thesis suggests that AI-supported speaking practice can be a useful part of EFL speaking education when it is carefully designed, guided, and evaluated. Its value lies not only in the technology itself, but in the teaching and learning conditions created around it.

References

Abbas, T., Gadiraju, U., Khan, V.-J., & Markopoulos, P. (2021). Making time fly: Using fillers to improve perceived latency in crowd-powered conversational systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *9*, 2–14. https://doi.org/10.1609/hcomp.v9i1.18935

Bai, S. (2023). The predictive effects of foreign language anxiety and boredom on willingness to communicate among Chinese struggling EFL learners. *Heliyon*, *9*(9), Article e19610. https://doi.org/10.1016/j.heliyon.2023.e19610

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, *32*(8), 827–877. https://doi.org/10.1080/09588221.2018.1535508

Bibauw, S., Van den Noortgate, W., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, *26*(1), 1–24. https://doi.org/10.64152/10125/73488

Boone, H., & Boone, D. (2012). Analyzing Likert data. *Journal of Extension*, *50*(2), Article 48. https://doi.org/10.34068/joe.50.02.48

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

British Council. (n.d.). *IELTS speaking practice test: Part 1*. https://takeielts.britishcouncil.org/take-ielts/prepare/free-ielts-english-practice-tests/speaking/part-1

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). Longman.

ByteDance Seed. (n.d.). *Seed2.0*. https://seed.bytedance.com/en/seed2

Cha, J., Han, J., Yoo, H., & Oh, A. (2024). CHOP: Integrating ChatGPT into EFL oral presentation practice. *CEUR Workshop Proceedings*, *3840*.

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge University Press.

Chen, Z. (2024). A study of Chinese undergraduate students' English language speaking anxiety, expectancy-value beliefs and spoken English proficiency. *Sage Open*, *14*(1), Article 21582440231219312. https://doi.org/10.1177/21582440231219312

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Davis-Muffett, T. (2024, July 17). Beyond the prompt round-up: Our series on getting the best out of generative AI. *Google Workspace*. https://workspace.google.com/blog/ai-and-machine-learning/beyond-prompt-round-our-series-getting-best-out-generative-ai

Dewaele, J.-M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign language enjoyment and anxiety: The effect of teacher and learner variables. *Language Teaching Research*, *22*(6), 676–697. https://doi.org/10.1177/1362168817692161

Ding, D., & Yusof, A. M. B. (2025). Investigating the role of AI-powered conversation bots in enhancing L2 speaking skills and reducing speaking anxiety: A mixed methods study. *Humanities and Social Sciences Communications*, *12*(1), Article 1223. https://doi.org/10.1057/s41599-025-05550-z

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Doubao. (2023). *豆包：字节跳动旗下 AI 智能助手*. https://www.doubao.com/

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, *6*, Article 100230. https://doi.org/10.1016/j.caeai.2024.100230

Ergül, H. (2021). Mitigating oral corrective feedback through linguistic strategies and smiling. *Journal of Pragmatics*, *183*, 142–153. https://doi.org/10.1016/j.pragma.2021.07.018

Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, *121*. https://doi.org/10.1016/j.system.2024.103254
Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—Principles and practices. *Health Services Research*, *48*(6pt2), 2134–2156. https://doi.org/10.1111/1475-6773.12117
Foster, P. (2020). Oral fluency in a second language: A research agenda for the next ten years. *Language Teaching*, *53*(4), 446–461. https://doi.org/10.1017/S026144482000018X
Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2022). Opposing effects of response time in human–chatbot interaction: The moderating role of prior experience. *Business & Information Systems Engineering*, *64*(6), 773–791. https://doi.org/10.1007/s12599-022-00755-x
Goh, C. C. M., & Burns, A. (2012). *Teaching speaking: A holistic approach*. Cambridge University Press.
Google. (n.d.). *Google Translate* [Computer software]. https://translate.google.com/
Google Career Certificates. (2024). *Google Prompting Essentials [MOOC]*. https://www.coursera.org/specializations/prompting-essentials-google
Google Cloud. (2026). *Overview of prompting strategies*. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/prompts/prompt-design-strategies
Grow with Google Editorial Team. (2026). *Prompting best practices*. https://grow.google/grow-your-career/articles/prompting-best-practices
Hall, E. T. (1976). *Beyond culture*. Anchor Press/Doubleday.
Hall, E. T., & Hall, M. R. (1990). *Understanding cultural differences: Germans, French, and Americans*. Intercultural Press.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487
House, J., & Kasper, G. (1981). Politeness markers in English and German. In F. Coulmas (Ed.), *Conversational routine* (pp. 157–186). De Gruyter Mouton. https://doi.org/10.1515/9783110809145.157
Hubbard, P. (2011). Evaluation of courseware and websites. In L. Ducate & N. Arnold (Eds.), *Present and future promises of CALL: From theory and research to new directions in foreign language teaching* (pp. 407–440). CALICO.
IELTS. (2023). *IELTS speaking sample tasks*. https://ielts.org/cdn/ielts-sample-tests/ielts-speaking-sample-tasks-2023.pdf
IELTS. (2025a). *IELTS speaking key assessment criteria* [PDF]. https://ielts.org/cdn/ielts-guides/ielts-speaking-key-assessment-criteria.pdf
IELTS. (2025b). *IELTS test taker performance 2024–2025*. https://ielts.org/en-us/for-researchers/test-statistics
IELTS. (n.d.–a). *IELTS academic format: Speaking*. https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-speaking
IELTS. (n.d.–b). *Test statistics*. https://ielts.org/researchers/our-research/test-statistics
IELTS. (n.d.–c). *Welcome to IELTS*. https://ielts.org/
JASP Team. (2026). *JASP* (Version 0.96.0) [Computer software]. https://jasp-stats.org/
Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T.,…Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274
Kim, D., Pan, Y., & Park, H. S. (1998). High- versus low-context culture: A comparison of Chinese, Korean, and American cultures. *Psychology & Marketing*, *15*(6), 507–521. https://doi.org/10.1002/(SICI)1520-6793(199809)15:6<507::AID-MAR2>3.0.CO;2-A
Kittler, M. G., Rygl, D., & Mackinnon, A. (2011). Special review article: Beyond culture or beyond control? Reviewing the use of Hall's high-/low-context concept. *International Journal of Cross Cultural Management*, *11*(1), 63–82. https://doi.org/10.1177/1470595811398797
Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates.
Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). Routledge. https://doi.org/10.4324/9781315775661
Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*, Article 100174. https://doi.org/10.1016/j.caeo.2024.100174
Leroy, S. (2009). Why is it so hard to do my work? The challenge of attention residue when switching between work tasks. *Organizational Behavior and Human Decision Processes*, *109*(2), 168–181. https://doi.org/10.1016/j.obhdp.2009.04.002
Li, S., & Vuono, A. (2019). Twenty-five years of research on oral and written corrective feedback in System. *System*, *84*, 93–109. https://doi.org/10.1016/j.system.2019.05.006
Lin, Y., Wang, B., Zhang, G., & Su, Y. (2025). The relations among foreign language anxiety, academic buoyancy and willingness to communicate in EFL classroom. *Frontiers in Psychology*, *16*, Article 1634054. https://doi.org/10.3389/fpsyg.2025.1634054
Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815
Loewen, S., & Sato, M. (2018). Interaction and instructed second language acquisition. *Language Teaching*, *51*(3), 285–329. https://doi.org/10.1017/S0261444818000125
Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.
Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40. https://doi.org/10.1017/s0261444812000365
Mackey, A., & Gass, S. M. (2022). *Second language research: Methodology and design* (3rd ed.). Routledge. https://doi.org/10.4324/9781003188414
Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: More speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 107–110). Association for Computing Machinery. https://doi.org/10.1145/1357054.1357072
McCrocklin, S., & Levis, J. (2025). Automatic speech recognition and pronunciation learning. *Language Teaching*, *59*(3), 293–309. https://doi.org/10.1017/s0261444825100852
McHugh-Johnson, M. (2024, July 29). 5 tips for writing great prompts for Gemini in the Workspace side panel. *Google*. https://blog.google/products-and-platforms/products/workspace/google-gemini-workspace-ai-prompt-tips/
Ng, A. (2019). *AI for everyone [MOOC]*. https://www.coursera.org/learn/ai-for-everyone
Park, H. S., Levine, T. R., Weber, R., Lee, H. E., Terra, L. I., Botero, I. C., Bessarabova, E., Guan, X., Shearman, S. M., & Wilson, M. S. (2012). Individual and cultural variations in direct communication style. *International Journal of Intercultural Relations*, *36*(2), 179–187. https://doi.org/10.1016/j.ijintrel.2011.12.010
Peng, J.-E. (2012). Towards an ecological understanding of willingness to communicate in EFL classrooms in China. *System*, *40*(2), 203–213. https://doi.org/10.1016/j.system.2012.02.002
Phakiti, A. (2014). *Experimental research methods in language learning*. Bloomsbury Academic.
Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687. https://doi.org/10.1017/s0272263113000399
Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079
Qualtrics. (n.d.). *Randomizer*. https://www.qualtrics.com/support/survey-platform/survey-module/survey-flow/standard-elements/randomizer/
Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002
Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shao, Q., & Gao, X. (2016). Reticence and willingness to communicate (WTC) of East Asian language learners. *System*, *63*, 115–120. https://doi.org/10.1016/j.system.2016.10.001
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795
Tencent. (n.d.). *WeChat*. https://www.wechat.com/
Tencent Meeting. (n.d.). *腾讯会议官方 [Tencent Meeting official website]*. https://meeting.tencent.com/
Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. *System*, *127*, Article 103533. https://doi.org/10.1016/j.system.2024.103533
White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*, 1–31. https://doi.org/10.48550/arXiv.2302.11382
Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087
Xiaohongshu. (n.d.). *RedNote*. https://apps.apple.com/us/app/rednote/id741292507
Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370
Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System*, *124*, Article 103351. https://doi.org/10.1016/j.system.2024.103351
Yang, Y., Cui, Y., & Yao, S. (2024). Teacher support, grit and L2 willingness to communicate: The mediating effect of foreign language enjoyment. *BMC Psychology*, *12*(1), Article 383. https://doi.org/10.1186/s40359-024-01877-5
Zhang, D., Wu, J. G., & Fu, Z. (2024). From shy to fly: Facilitating EFL learners' willingness to communicate with an AI chatbot and an intelligent tutoring system. *System*, *127*, Article 103501. https://doi.org/10.1016/j.system.2024.103501
Zhang, M., Yi, N., & Zhou, D. (2023). The effects of task repetition schedules on L2 fluency enhancement. *Languages*, *8*(4), Article 252. https://doi.org/10.3390/languages8040252

Appendix A

Participant Information, Eligibility Screening, and Consent Materials

Participant Information Sheet（信息说明）



Participant Information Sheet / 参与者信息说明书
Project title / 项目名称
A Comparative Study of Two Teaching Approaches- Evaluating Two Speaking-Practice Approaches for IELTS Learners
面向雅思学习者的两种口语练习方式评估
Date / 日期： 21 January 2026（2026 年 1 月 21 日）
Student researcher / 学生研究者： Stella Li（李烨）, Master's student, Auckland University of Technology (AUT), New Zealand
Email / 邮箱： ftx0939@autuni.ac.nz
Project supervisor / 项目导师： Associate Professor Ann Morrison, Auckland University of Technology (AUT), New Zealand
Email / 邮箱： ann.morrison@aut.ac.nz

---

1) Invitation / 邀请说明
EN: You are invited to take part in a Master's student research project at Auckland University of Technology (AUT), New Zealand. Please read this information sheet before deciding whether to participate. You may ask questions at any time.
中： 诚邀你参加一项由新西兰奥克兰理工大学（AUT）开展的硕士学生研究。在决定是否参与之前，请仔细阅读本信息说明书。你可在任何时候向研究者提问。

---

2) Purpose of the study / 研究目的
EN: This study compares two speaking-practice approaches for adult IELTS learners. We focus on (1) clarity of spoken communication in English (e.g., being more direct and specific), and (2) changes in official IELTS Speaking results before and after the programme.
中： 本研究比较两种面向成人雅思学习者的口语练习方式，重点关注：（1）英语口语表达的清晰度（例如更直接、更具体的表达），以及（2）课程前后官方雅思口语成绩的变化。

---

3) Why you have been invited (Inclusion criteria) / 你为何被邀请（纳入标准）
EN: You may participate if you meet all of the following criteria:

- You are 18 years or older;
- Your first language (L1) is Chinese (Mandarin);

- You have an official IELTS Test Report Form (TRF/eTRF) from the past 6 months;
- You agree to sit another official IELTS test within 1 month after the final session;
- You agree to provide verification of both IELTS results (pre- and post-study);
- You can attend the fixed class schedule (Tuesdays 5:00–6:00 pm and Thursdays 5:00–6:00 pm or Tuesday 6:30-7:30pm and Thursday 6:30-7:30pm, Beijing time) for 4 weeks via Tencent Meeting.

中： 若你满足以下全部条件，即可参与：
1）年满 18 周岁；
2）母语为中文（普通话）；
3）过去 6 个月内有一次官方 IELTS 成绩单（TRF/eTRF）；
4）同意在最后一次课程结束后的 1 个月内再参加一次官方 IELTS 考试；
5）同意提供研究前与研究后两次 IELTS 成绩的证明材料；
6）可以通过腾讯会议参加为期 4 周的固定上课时间（北京时间每周二 17:00–18:00、每周四 17:00–18:00或者每周二 18:30–19:30、每周四18:30–19:30）。

Important / 重要说明
EN: Because official IELTS results are an inclusion criterion, providing them is required (not optional) for this study.
中： 因为官方 IELTS 成绩属于纳入标准之一，因此本研究中提供成绩是必须项（不是可选项）。

---

4) What participation involves / 参与内容（你需要做什么）
EN: If you take part, you will:
• Complete a short online form (eligibility screening + consent).
• Attend an online speaking programme totaling 8 hours over 4 weeks (2 hours per week):
– Time: Tuesdays 5:00–6:00 pm and Thursdays 5:00–6:00 pm (Beijing time, China Standard Time) or Tuesdays 6:30–7:30 pm and Thursdays 6:30–7:30 pm (Beijing time, China Standard Time).
– Platform: Tencent Meeting.
• Be randomly allocated to one of two feedback approaches for Sessions 2–8 (same total hours and similar IELTS-style prompts).
• Complete a short exit survey after the programme.
• Provide your official IELTS results (pre- and post-study) as described below.
• After you complete the online form, your Study ID and Tencent Meeting access details will be shown on screen. Please save them (e.g., screenshot), as we do not collect contact details and will not send reminders.
中： 如果你参与本研究，你将：
• 填写一个简短的线上表格（筛选问卷 + 知情同意）。
• 在 4 周内完成总计 8 小时的线上口语学习（每周 2 小时）：
– 时间：北京时间每周二 17:00–18:00、每周四 17:00–18:00 或者北京时间每周二 18:30–19:30、每周四 18:30–19:30。
– 平台：腾讯会议 Tencent Meeting。
• 在第 2–8 次课程中被随机分配到两种反馈方式之一（总课时一致、题型与提示相近）。
• 课程结束后完成一份简短的结课问卷。
• 按下文说明提供研究前与研究后的官方雅思成绩信息。
• 完成线上表格后，系统会在页面上显示你的 Study ID 以及腾讯会议上课信息。请自行保存（例如截图），因为本研究不收集联系方式，也不会发送提醒。

---

5) Group allocation / 分组方式
EN: Participants will be randomly allocated to one of two feedback approaches for Sessions 2–8. Both groups receive the same total practice time and similar IELTS-style prompts.
中： 在第 2–8 次课程中，参与者将被随机分配到两种反馈方式之一。两组练习总时长一致，使用相似的雅思口语题型与提示。

6) Study ID / 研究编号（Study ID）
EN: You will be assigned a Study ID. We use this Study ID (not your name) to label your research data wherever possible.
中： 你会获得一个研究编号（**Study ID**）。研究数据将尽可能使用 **Study ID** 标识，而非你的真实姓名。

7) Online class setting and privacy / 线上课堂形式与隐私说明
EN: Sessions are delivered online in a small group. Other participants may be able to see your display name and hear you during class (similar to a normal online class). You may choose to use a nickname and keep your camera off. We will ask everyone to respect privacy and not share class content or personal information outside the group, but we cannot fully control what other participants do.
中： 课程以线上小组形式进行。同班同学可能会看到你的显示名称并听到你的发言（与普通线上课类似）。你可以使用昵称并选择不开摄像头。我们会提醒所有参与者尊重隐私，不在群体外传播课堂内容或他人信息，但无法完全控制其他参与者的行为。

---

8) IELTS results: what we collect and how / 雅思成绩：收集内容与方式
EN: We use official IELTS results as part of the research. The research team does not obtain IELTS results directly from the test provider on participants' behalf.
You will enter your IELTS test date and official IELTS Speaking score in the online form for:
• your IELTS result from the past 6 months (pre-study), and
• your IELTS result taken within 1 month after the final session (post-study).
We do not discuss individual IELTS scores in class. Any IELTS information you provide is stored under Study ID only.
中： 本研究将使用官方 IELTS 成绩作为研究数据的一部分。研究团队不会代表参与者向考试机构直接调取成绩。
你需要在**线上表格**中填写以下两次考试的**考试日期与官方 IELTS 口语分数**：
• 过去 6 个月内的一次官方成绩（研究前），以及
• 最后一次课程结束后 1 个月内参加考试的官方成绩（研究后）。
课堂中不会讨论个人分数。你提供的 IELTS 信息将仅以 Study ID 保存与使用。

---

9) Contact details and reminders / 联系方式与提醒说明（不收集联系方式）
EN: This study does **not** collect participants' email addresses, phone numbers, or other direct contact details. Because we do not collect contact details, we will not send scheduling messages or reminders. The class schedule is fixed, and access details will be shown on screen after you complete the online form. Please save your Study ID and class access details for your own reference.
中： 本研究**不收集**参与者的邮箱、电话或其他直接联系方式。由于不收集联系方式，我们不会发送排课信息或上课提醒。课程时间固定，你在完成线上表格后会在页面上看到上课信息。请自行保存你的 Study ID 和上课信息，便于后续使用。

---

10) Risks and discomforts / 风险与不适
EN: Speaking tasks may cause mild nervousness or performance anxiety. You may pause, skip an activity where possible, or withdraw at any time.
中： 口语任务可能带来轻度紧张或表现焦虑。你可以随时暂停、在允许情况下跳过部分活动，或退出研究。

---

11) Benefits / 收益
EN: You may benefit from structured IELTS speaking practice and feedback on clearer communication. However, improvement in IELTS scores cannot be guaranteed.

The findings may provide useful insights for IELTS learners, teachers, and education providers about ways to support clearer spoken communication.The results may contribute to wider understanding of effective IELTS speaking practice approaches (including technology-supported options).

中： 你可能从结构化雅思口语练习与表达清晰度反馈中获益，但无法保证成绩一定提升。

研究结果可能为雅思学习者、教师与教育机构提供参考，帮助支持更清晰的口语表达。研究也可能为更广泛的雅思口语练习方式（包括技术辅助的练习）提供经验与启示。

---

12) Recording / 录音说明（明确不录音）
EN: Sessions will not be audio-recorded or video-recorded for research purposes.
中： 本研究的课程不会出于研究目的进行音频或视频录制。

---

13) Confidentiality, data storage, and access / 保密、数据保存与访问权限
EN: Research data will be stored securely in password-protected systems under AUT accounts. Only the student researcher and the supervisor will have access to the study data. Findings will be reported in summary form; any examples (if used) will be de-identified. Data will be retained for 6 years after the study and then securely destroyed.
中： 研究数据将存放在 AUT 账号体系下、受密码保护的安全系统中。仅学生研究者与导师可访问研究数据。研究结果将以汇总形式呈现；如使用示例内容（如有），也会进行去标识化处理。数据将在研究结束后保存 6 年，随后安全销毁。

---

14) Voluntary participation and withdrawal / 自愿参与与退出
EN: Participation is voluntary. You may withdraw at any time without penalty. If you withdraw, you may request that your identifiable information and data be removed where possible before analysis and de-identification. Once data have been de-identified and combined for analysis, it may not be possible to remove individual contributions.
中： 参与完全自愿。你可在任何时候退出且不会产生不利影响。若你退出，可在分析与去标识化之前请求删除你的可识别信息与相关数据；一旦数据已去标识化并进入汇总分析阶段，可能无法再单独剔除个人数据。

---

15) Questions, concerns, and complaints / 疑问、关注与投诉
EN: If you have questions about this study, you may contact the student researcher: **ftx0939@autuni.ac.nz**.
Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor: **Associate Professor Ann Morrison (ann.morrison@aut.ac.nz)**.
Concerns regarding the conduct of the research should be notified to the Executive Secretary of AUTEC: **ethics@aut.ac.nz**, phone **+64 9 921 9999 ext 6038**.
中： 如对本研究有疑问，可联系学生研究者邮箱：**ftx0939@autuni.ac.nz**。
如对本项目的内容/性质有疑问或关注，请首先联系项目导师：**Ann Morrison 副教授（ann.morrison@aut.ac.nz）**。
如对研究执行/研究行为有担忧或投诉，可联系 AUTEC 执行秘书：**ethics@aut.ac.nz**，电话 **+64 9 921 9999 分机 6038**。

---

16) Ethics approval / 伦理审批信息
EN: This study was reviewed by the Auckland University of Technology Ethics Committee (AUTEC) and approved on **[05/02/2026]**. AUTEC Reference Number: **25-429**.
中： 本研究已提交奥克兰理工大学伦理委员会（AUTEC）审查，并于 2026年2月5日获批。AUTEC 参考编号：**25-429**。
EN: Please keep a copy of this Information Sheet for your records.
中： 请保留本信息说明书以备后续查阅。

Eligibility screening（筛选问卷）

EN: Are you 18 years old or above?

中： 你是否已满 18 周岁?

- ○ Yes 是
- ○ No 否

EN: Is your first language (L1) Chinese (Mandarin)?

中： 你的母语（第一语言，L1）是否为中文（普通话）？

- ○ Yes 是
- ○ No 否

EN: Do you have an official IELTS Test Report Form (TRF/eTRF) issued within the past 6 months?

中： 你是否有过去 6 个月内出具的官方 IELTS 成绩单（TRF/eTRF）？

- ○ Yes 是
- ○ No 否

EN: Do you plan and agree to sit another official IELTS test within 1 month after the final study session?

中： 你是否计划并同意在最后一次课程结束后的 1 个月内再参加一次官方 IELTS 考试?

- ○ Yes 是
- ○ No 否

EN: Do you agree to provide verification (a photo/screenshot) of your official IELTS TRF/eTRF for both the pre-study and post-study tests?

中： 你是否同意提供研究前与研究后两次官方 IELTS 成绩单（TRF/eTRF）的证明材料（照片/截图）？

- ○ Yes 是

○ No 否

EN: Can you attend the online sessions at the following times (Beijing time): Tuesdays 5:00–6:00 pm and Thursdays 5:00–6:00 pm, or Tuesday 6:30 pm-7:30 pm and Thursday 6:30 pm - 7:30 pm for 4 weeks (total 8 hours)? 中： 你是否能参加以下时间的线上课程（北京时间）：每周二 17:00–18:00 与每周四 17:00–18:00 或者每周二 18:30–19:30、每周四 18:30–19:30，连续 4 周（总计 8 小时）？

○ Yes 是
○ No 否

Consent Form 知情同意书

Consent Form 知情同意书



Project title / 项目名称：

A Comparative Study of Two Teaching Approaches: Evaluating Two Speaking-Practice Approaches for IELTS Learners

面向雅思学习者的两种口语练习方式评估

Project Supervisor / 项目导师：

Associate Professor Ann Morrison, Auckland University of Technology (AUT), New Zealand

Researcher / 研究者:

Stella Li（李烨）, Master's student / 硕士研究生, Auckland University of Technology (AUT), New Zealand

- **EN:** I have read and understood the information provided about this research project in the Information Sheet dated **05/02/2026**. **中：** 我已阅读并理解 **2026 年 2 月 5 日** 的《参与者信息说明书》中关于本研究项目的相关信息。

- **EN:** I have had an opportunity to ask questions and to have them answered. **中：** 我有机会就本研究提出问题，并已得到解答。

- **EN:** I understand that taking part in this study is voluntary (my choice) and that I may withdraw from the study at any time without being disadvantaged in any way. **中：** 我理解参与本研究是自愿的（由我自行决定），并且我可以在任何时间退出研究，且不会因此受到任何不利影响。

- **EN:** I agree to take part in this research. **中：** 我同意参加本研究。

EN: I agree to all of the above statements and voluntarily agree to participate in this study.

中： 我同意以上所有声明，并自愿参加本研究。

- ○ Yes 是
- ○ No 否

EN: I wish to receive a summary of the research findings (please tick one):

中： 我希望在研究结束后收到研究结果摘要（请选择一项）:

○ Yes 是

○ No 否

Signature 签名

Participant's signature: 参与者签名

Please sign below to indicate your consent.

请在下方签名以表示您同意参加本研究。



Participant's name: 参与者姓名

Please enter your full name.

请输入您的姓名。

Date: 06/02/2026

This study has been approved by the Auckland University of Technology Ethics Committee on 06/02/2026.

AUTEC Reference number: 25/429

Note: Participants should retain a copy of this form.

日期：2026年02月06日

本研究已于 2026 年 02 月 06 日获得奥克兰理工大学伦理委员会（AUTEC）批准。

AUTEC 参考编号：25/429

注：建议参与者截图或保存本页面，以便日后查阅。

Study ID 生成与保存

Your Study ID is: ${e://Field/study_id}. Please take a screenshot or copy and save it—you will need to enter it multiple times later.

你的 Study ID 是：${e://Field/study_id}，**请截图/复制保存，之后会多次填写。**

Please confirm and enter your Study ID.

请确认并输入你的Study ID.

Course Link 课程信息

**此页十分重要，建议截图保存哦！

**This page is very important—it's recommended that you take a screenshot and save it!

Platform: Tencent Meeting (online).

Schedule: Tuesdays 5:00–6:00 pm Beijing Time and Thursdays 5:00–6:00 pm Beijing Time.

平台：腾讯会议（线上）。 时间安排：每周二 17:00–18:00（北京时间）和每周四 17:00–18:00（北京时间）或者北京时间每周二 18:30–19:30、每周四 18:30–19:30。

The first class is an individual session. The date and time are: 6:30pm-7:30pm, 10/02/2026

Class link:

第一节课是区别于后面课程的，日期和时间是：2026/02/10 18:30-19:30

课程链接：**（请复制保存）**

Stella Li 邀请您参加腾讯会议 会议主题：Stella Li预定的会议

会议时间：2026/02/10 18:30-19:30（GMT+08:00）中国标准时间 - 北京

该会议开启了报名功能，请点击链接报名后入会：

Powered by Qualtrics

Appendix B

Group Allocation Instructions

Group Allocation (Post-Session 1) 课后分组



Please enter your Study ID (e.g., S000111). 请输入你的Study ID（例如 S000111）。

Thanks for attending Session 1. Please continue to receive instructions for your practice sessions (Sessions 2–8). Further information about the study design will be

provided at the end of the study.
恭喜你完成第一节课！接下来是第2-8节练习课的注意事项。关于研究设计的更多信息将在研究结束时提供。

Group A – Teacher-led

Group A – Teacher-led
Practice Session Instructions (Sessions 2–8)
EN:
During Sessions 2–8, you will attend online practice sessions led by the instructor. You will receive feedback aligned with the IELTS Speaking criteria. To ensure fairness, please do not use any external AI tools (e.g., chatbots) during these sessions.
Schedule (Beijing Time): Tuesdays 17:00–18:00 and Thursdays 17:00–18:00 (4 weeks).
Tencent Meeting link: [] (Meeting ID/Passcode, if applicable: []).
中：
在第 2–8 次课程中，你将参加由授课老师带领的线上练习课，并获得基于雅思口语评分标准的反馈。为保证公平，请在课程进行期间不要使用任何外部 AI 工具（例如聊天机器人）。
上课时间（北京时间）： 周二 17:00–18:00、周四 17:00–18:00（共 4 周）。
腾讯会议链接：【待提供】（会议号/密码如适用：**【待提供】**）。

**This page is very important—it's recommended that you take a screenshot and save it!
**此页十分重要，建议截图保存哦！

Group B – AI-led

Group B – AI-led
Practice Session Instructions (Sessions 2–8)
EN:
During Sessions 2–8, you will use Doubao (豆包) to receive feedback, following the researcher's instructions. Please do not enter any personal identifiers (e.g., your name, email, phone number) into the AI tool.
Schedule (Beijing Time): Tuesdays 18:30–19:30 and Thursdays 18:30–19:30 (4 weeks).
Tencent Meeting link: [] (Meeting ID/Passcode, if applicable: []).
中：
在第 2–8 次课程中，你将按照研究者的指引使用豆包（Doubao）获取反馈。请不要在 AI 工具中输入任何可识别个人身份的信息（例如姓名、邮箱、手机号等）。
上课时间（北京时间）： 周二 18:30–19:30、周四 18:30–19:30（共 4 周）。

腾讯会议链接：【待提供】（会议号/密码如适用：【待提供】）。

**This page is very important—it's recommended that you take a screenshot and save it!
**此页十分重要，建议截图保存哦！

Powered by Qualtrics

Note. Tencent Meeting links and access details have been removed from the reproduced materials.

Appendix C

Exit Survey

Exit Survey 研究结束问卷

Exit Survey 研究结束问卷



Project title / 项目名称
Evaluating Two Speaking-Practice Approaches for IELTS Learners
面向雅思学习者的两种口语练习方式评估
Date / 日期： 21 January 2026（2026 年 1 月 21 日）
Student researcher / 学生研究者： Stella Li（李烨）, Master's student, Auckland University of Technology (AUT), New Zealand
Email / 邮箱： ftx0939@autuni.ac.nz
Project supervisor / 项目导师： Associate Professor Ann Morrison, Auckland University of Technology (AUT), New Zealand
Email / 邮箱： ann.morrison@aut.ac.nz

Questions, concerns, and complaints / 疑问、关注与投诉
EN: If you have questions about this study, you may contact the student researcher: **ftx0939@autuni.ac.nz**.
Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor: **Associate Professor Ann Morrison (ann.morrison@aut.ac.nz)**.
Concerns regarding the conduct of the research should be notified to the Executive Secretary of AUTEC: **ethics@aut.ac.nz**, phone **+64 9 921 9999 ext 6038**.
中： 如对本研究有疑问，可联系学生研究者邮箱：**ftx0939@autuni.ac.nz**。
如对本项目的内容/性质有疑问或关注，请首先联系项目导师：**Ann Morrison 副教授（ann.morrison@aut.ac.nz）**。
如对研究执行/研究行为有担忧或投诉，可联系 AUTEC 执行秘书：**ethics@aut.ac.nz**，电话 **+64 9 921 9999 分机 6038**。

Ethics approval / 伦理审批信息
EN: This study was reviewed by the Auckland University of Technology Ethics Committee (AUTEC) and approved on **[5 February 2026]**. AUTEC Reference Number: **25-429**.
中： 本研究已提交奥克兰理工大学伦理委员会（AUTEC）审查，并于 **[2026年2月5日]** 获批。AUTEC 参考编号：**25-429**。

Your previous overall IELTS band score（e.g., 6.5）
你上一次考试的雅思总分（比如 6.5）

previous IELTS Speaking band score（e.g., 6.5）
上一次雅思口语分数（比如 6.5）

previous IELTS Listening band score（e.g., 6.5）
上一次雅思听力分数（比如 6.5）

previous IELTS Reading band score（e.g., 6.5）
上一次雅思阅读分数（比如 6.5）

prvious IELTS Writing band score（e.g., 6.5）
上一次雅思写作分数（比如 6.5）

Your most recent overall IELTS band score（e.g., 6.5）
你最近一次的雅思总分（比如 6.5）

most recent IELTS Speaking band score（e.g., 6.5）
最近一次雅思口语分数（比如 6.5）

most recent IELTS Listening band score（e.g., 6.5）
最近一次雅思听力分数（比如 6.5）

most recent IELTS Reading band score（e.g., 6.5）
最近一次雅思阅读分数（比如 6.5）

most recent IELTS Writing band score（e.g., 6.5）
最近一次雅思写作分数（比如 6.5）

Course experience

EN: Please indicate how much you agree with the statements below about your experience in this study.

中： 请根据你在本研究中的体验，选择你同意以下陈述的程度。

	Strongly agree非常同意	Agree同意	Neutral一般	Disagree不同意	Strong disagr非常不 意
EN: The sessions helped me express my ideas more clearly in English. 中： 这些课程帮助我用英语更清晰地表达自己的想法。	○	○	○	○	○
EN: I became more direct and specific when expressing opinions or making requests in English. 中： 我在用英语表达观点或提出请求时变得更直接、更具体。	○	○	○	○	○

	Strongly agree非常同意	Agree同意	Neutral一般	Disagree不同意	Strong disagr 非常不 意
EN: I feel more confident speaking in IELTS-style tasks. 中： 我在进行雅思口语题型练习时更有信心。	○	○	○	○	○
EN: The feedback I received was useful for improving my speaking. 中： 我收到的反馈对提升口语很有帮助。	○	○	○	○	○
EN: After each session, I understood what I needed to improve. 中： 每次课后我都更清楚自己需要改进的点。	○	○	○	○	○
EN: The practice activities were well structured and easy to follow. 中： 练习活动结构清晰、容易跟上。	○	○	○	○	○
EN: The pace and workload were reasonable for me. 中： 课程节奏与任务量对我来说是合理的。	○	○	○	○	○
EN: I felt comfortable participating in the online sessions. 中： 我在参加线上课程时总体感到舒适。	○	○	○	○	○
EN: I would recommend this type of speaking practice to other IELTS learners. 中： 我愿意向其他雅思学习者推荐这种口语练习方式。	○	○	○	○	○

Open-Ended Questions

What was the most helpful part of the study for your speaking?
本研究中对你口语最有帮助的部分是什么?

What could be improved? （你觉得哪些地方可以改进？）

EN: Any other comments?
中： 还有其他想补充的吗?

最后一页 Debrief（结尾说明）

EN: Thank you so much for completing this survey and our study. If you have any questions or would like to withdraw, please contact the researcher.
中： 十分感谢你完成本问卷和我们的研究。如有任何疑问，或希望撤回参与，请联系研究者。

Powered by Qualtrics

Appendix D

AI Feedback Prompt Used in the Study

You are my IELTS Speaking coach (Band 6–9 standard). Act like a supportive feedback provider.

How we will do this (interactive coaching)

• I will paste a Part 2 cue card and answer for up to 2 minutes.

You will

• Give 2 strengths (must quote my exact phrases),

• Give 2 grammar-focused issues (must quote my exact phrases),

• Give 2 targeted fixes (rewrite my sentences in natural spoken English, Band 7–8),

• Then ask 2 follow-up questions to expand my answer.

Rules

• No overall band score. Evidence-based coaching only.

• Always quote my exact words as evidence (short quotes).

• Consider all four IELTS Speaking criteria, while prioritising the most important issues in the learner's response.

(Flexible input example: Pay particular attention to grammatical range and accuracy, including tense, articles, agreement, sentence structure, and word form.)

• Keep your response spoken, natural, not essay-like.

• Keep explanations short (1–2 lines each).

Output format

Cue card: …

2 strengths (with quotes):

• Strength 1: "…" → (why it works, 1 line)

• Strength 2: "…" → (why it works, 1 line)

2 grammar issues (with quotes):

• Issue 1 (category): "…" → correction (before → after) + 1-line rule

• Issue 2 (category): "…" → correction (before → after) + 1-line rule

2 targeted fixes (Band 7–8 spoken rewrites):

• Fix 1: (my original "…") → Improved: "…"

• Fix 2: (my original "…") → Improved: "…"

2 follow-up questions (higher band push):

1. …

2. …

Then wait for my reply.