International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 23, No. 2, 2026

AI-Supported Feedback for Second Language Mandarin Learners: Affordances and Limitations Across YCT, HSK, and HSKK in Multilingual International School Settings

Ru Bai

University of Education Freiburg, Freiburg, Germany

Abstract: The rapid advancement of artificial intelligence (AI) has opened new possibilities for second language assessment, particularly in the context of Chinese learning and Chinese proficiency certification. This paper examines the potential of AI-supported feedback tools in addressing the persistent challenge of timely and personalized feedback for second language Mandarin learners across the full spectrum of Chinese proficiency examinations-the Youth Chinese Test (YCT), the Hanyu Shuiping Kaoshi (HSK, Levels 1–5), and the HSK Speaking Test (HSKK)-in multilingual international school settings. Drawing on second language acquisition (SLA) theory, feedback models, and classroom observations from international school environments in Singapore and the Philippines, this conceptual review analyses the affordances and limitations of freely available AI tools, including ChatGPT, HelloChinese, Duolingo Chinese, Google AI Studio, and Speechling, which support vocabulary acquisition, tonal accuracy, character recognition, oral production, and communicative competence. The paper further contextualizes these tools within multilingual policy frameworks, with particular reference to Singapore's bilingual education model and IBDP international school settings in the Philippines. Findings suggest that while AI tools offer significant potential for real-time, adaptive feedback, their alignment with YCT, HSK, and HSKK assessment criteria and the cultural-pragmatic dimensions of Mandarin remains limited. Pedagogical implications for practitioner-researchers and teacher educators in multilingual international school settings are discussed.

Keywords: YCT, HSK, HSKK, AI Feedback, Second Language Acquisition, Mandarin Language, Multilingual Education, IBDP, International School, Formative Assessment.

1. Introduction

The globalization of Mandarin as a language of commerce, diplomacy, and cultural exchange has driven a significant rise in the number of non-native learners worldwide. The HSK (Hanyu Shuiping Kaoshi), China's official standardized proficiency examination, has become the primary benchmark for Mandarin learners, with over one million candidates sitting the examination annually across more than 60 countries [4]. Despite its widespread adoption, the pedagogical challenge of preparing learners, particularly those from diverse linguistic backgrounds for HSK remain considerable. A central difficulty lies in the provision of timely, individualized, and linguistically informed feedback during the preparation process.

This challenge is especially pronounced in multilingual and vocational education contexts. Based on over eight years of classroom observation in international school settings across Singapore and the Philippines, the author has consistently encountered learners from diverse language backgrounds, including speakers of English, Tagalog, Malay, Tamil, and various European languages, pursuing HSK certification at Levels 1 to 5. In these classrooms, teacher feedback is inherently constrained by class size, scheduling, and the sheer linguistic diversity of learners. A student whose first language is Tagalog, for example, faces fundamentally different phonological and orthographic challenges when acquiring Mandarin than a learner from a heritage Chinese-speaking background. Providing differentiated feedback to each learner in real time is, in practice, beyond the capacity of any individual teacher.

The emergence of AI-powered language learning tools offers a potential solution to this persistent problem. Platforms such as ChatGPT, HelloChinese, Duolingo Chinese, Google AI Studio, and Speechling now provide learners with immediate, algorithmically generated responses to language input, feedback that is available at any hour, in any location, and at no or minimal cost. However, the pedagogical value of these tools, particularly in relation to HSK-specific assessment criteria and the needs of multilingual learners, has not yet been systematically examined.

This paper addresses that gap through a conceptual review of the relevant literature and practitioner observation. It asks: (1) What types of feedback do freely available AI tools provide, and how do these align with YCT/HSK assessment? (2) What are the affordances and limitations of these tools for learners in multilingual and vocational education contexts? And (3) what implications do these findings hold for pedagogical practice and future research? The paper contributes to an emerging body of scholarship on AI in language education while offering practical guidance for teachers and researchers working in similar contexts.

2. Literature Review

2.1. Feedback in Second Language Acquisition

The role of feedback in second language acquisition has been extensively theorized and empirically investigated. Krashen [9] Input Hypothesis positioned comprehensible input as the primary driver of language acquisition, with explicit feedback playing a limited role. However, subsequent work by Swain [13] on the Output Hypothesis and Long [10] Interaction Hypothesis foregrounded the importance of feedback in negotiating meaning and drawing learners'

attention to gaps between their interlanguage and the target language. Corrective feedback-whether explicit or implicit, immediate or delayed, has since been recognized as a significant factor in L2 development [2].

Hattie and Timperley [5] influential feedback model provides a particularly useful framework for evaluating the quality of AI-generated responses. Their model distinguishes four levels of feedback: task level (correctness of a specific answer), process level (strategies for completing a task), self-regulation level (learner monitoring and management), and self level (personal attributes). Effective feedback, they argue, operates primarily at the process and self-regulation levels, supporting learners in developing independent strategies rather than merely correcting surface errors. As this paper will demonstrate, most current AI feedback tools for Mandarin learning operate predominantly at the task level-a significant pedagogical limitation.

2.2. HSK: Framework and Assessment Criteria

The Chinese proficiency examination system administered by Hanban encompasses three distinct but interrelated assessments, each targeting different learner profiles and competency domains. The HSK is structured around a six-level framework (with a revised seven-level structure introduced in HSK 3.0, [4]), corresponding broadly to the Common European Framework of Reference for Languages (CEFR) descriptors from A1 to C2. At Levels 1 and 2, the examination assesses recognition of high-frequency vocabulary (150 and 300 words respectively), basic listening comprehension, and simple reading tasks. Levels 3 and 4 extend this to 600 and 1,200 vocabulary items, introducing more complex grammatical structures, extended listening passages, and short written compositions [4]. The Youth Chinese Test (YCT), by contrast, is specifically designed for primary and secondary school students who are learning Mandarin as a foreign language. Comprising four levels, the YCT employs age-appropriate topics, simplified vocabulary, and visually supported tasks to assess communicative competence in younger learners, which makes it the most relevant benchmark for early years and primary-level instruction. The HSK Speaking Test (HSKK), meanwhile, is an independent oral examination that assesses spontaneous spoken production across three levels (Elementary, Intermediate, and Advanced), targeting the dimensions of pronunciation accuracy, fluency, and discourse coherence that the written HSK does not assess. Crucially, the HSK assesses communicative competence across integrated skills, not merely discrete linguistic knowledge-a distinction that has important implications for the design and evaluation of AI feedback tools.

A notable feature of the revised HSK 3.0 framework is its increased emphasis on pragmatic and intercultural competence, particularly at higher levels. This reflects a broader shift in language assessment theory towards construct validity models that foreground authentic, contextualized language use [11]. For AI tools to be genuinely useful in HSK preparation, they must therefore address not only lexical and grammatical accuracy but also the pragmatic dimensions of Mandarin communication.

2.3. AI Tools in Language Learning: Current Research

Research on AI-assisted language learning has grown substantially over the past decade. Early studies focused primarily on automated writing evaluation (AWE) systems such as Turnitin and e-rater, which provide feedback on grammatical accuracy and textual organization [14]. More recent work has examined the application of large language models (LLMs), including GPT-based systems, in providing open-ended feedback on learner writing and speech [8]. These studies generally report positive learner attitudes towards AI feedback, particularly with regard to immediacy and availability, while noting concerns about accuracy, depth, and the absence of human interactional qualities.

Research specifically addressing AI feedback in Chinese as a second language (CSL) contexts remains limited. Existing studies have examined automatic speech recognition (ASR) systems for Mandarin tonal feedback [7] and the use of gamified applications in vocabulary acquisition at lower proficiency levels [1]. However, few studies have systematically evaluated the alignment of AI feedback with HSK assessment criteria, or examined the particular needs of learners in multilingual and vocational education environments. This paper therefore fills a gap in the literature while drawing on practitioner knowledge to ground its analysis.

3. AI Tools for HSK Feedback: Affordances and Limitations

This section analyses five freely available AI tools that have demonstrable relevance to YCT, HSK, and HSKK preparation across different learner age groups and proficiency levels. The analysis draws on published research, developer documentation, and the author's direct classroom use of these tools in Singapore and the Philippines. Table 1 provides an overview before each tool is discussed in turn.

Table 1. Overview of AI Tools for YCT, HSK, and HSKK Preparation

Tool	Feedback Type	HSK Skills Covered	Cost	Key Limitation
ChatGPT	Writing correction, grammar explanation, mock dialogue	Writing, speaking (text), vocabulary; HSK 3–5, HSKK (Intermediate/Advanced)	Free (GPT-3.5)	Not aligned to HSK rubric
HelloChinese	Immediate right/wrong + pronunciation scoring	Vocabulary, grammar, tones; YCT 1–4, HSK 1–2	Free (basic)	Limited to lower levels
Duolingo Chinese	Gamified immediate feedback on vocabulary/grammar	Vocabulary, reading; YCT 1–4, HSK 1–2	Free	Shallow grammatical depth
Google AI Studio	Spoken input analysis, conversational response	Speaking, listening; HSKK (all levels), HSK 2–4	Free	Tonal accuracy inconsistent
Speechling	Pronunciation scoring with native-speaker comparison	Speaking, tones; HSKK (Elementary/Intermediate), HSK 1–4	Free (limited)	No writing or reading support

3.1. ChatGPT

Of the tools examined, ChatGPT [12] offers the broadest range of feedback capabilities for HSK preparation. As observed in international school classrooms in Singapore, the tool was used by students at HSK Levels 3 and 4 to receive corrections on short compositions, to practise HSK-style sentence completion tasks, and to engage in simulated dialogues on HSK topic domains such as travel, shopping, and daily routines. ChatGPT's ability to explain grammatical errors in the learner's preferred language, English, Tagalog, or otherwise, represents a significant advantage in multilingual classrooms where metalinguistic explanation is otherwise constrained by teacher capacity.

However, a critical limitation is ChatGPT's lack of alignment with the HSK assessment rubric. When asked to evaluate a Level 4 composition, the system tends to apply general writing quality criteria rather than the vocabulary range, grammatical complexity, and communicative function descriptors specified by Hanban. This gap between AI feedback and HSK construct validity represents a significant pedagogical risk: learners may be guided towards linguistic patterns that are fluent but not optimal for examination performance. Moreover, ChatGPT does not address tonal accuracy, which is assessed in the spoken component of HSKK (the oral extension examination) and is a persistent difficulty for learners from non-tonal language backgrounds.

3.2. HelloChinese and Duolingo Chinese

HelloChinese and Duolingo Chinese both offer gamified, immediate feedback systems that are particularly well suited to the vocabulary range and task formats of the Youth Chinese Test (YCT) and the lower levels of the HSK (Levels 1–2). The YCT, designed specifically for primary and secondary school learners of Mandarin as a foreign language, emphasises age-appropriate vocabulary, visually supported tasks, and communicative scenarios that align closely with the scaffolded, game-like structure of these applications. HelloChinese, in particular, includes a speech recognition component that provides a numeric score for tonal production, alongside visual waveform comparisons. In classroom use in the Philippines, where the author taught early years learners (ages 3–6) with no prior Mandarin exposure alongside older students, HelloChinese proved effective as a supplementary tool for building initial lexical recognition and character writing practice at HSK Level 1 and YCT Level 1. The immediate corrective feedback and low-stakes gamified environment appeared to reduce learner anxiety-a factor consistently identified in the literature as inhibiting L2 production (Horwitz, [6]).

Nevertheless, both applications share a significant limitation: their feedback operates exclusively at Hattie and Timperley's [5] task level. Correct responses are rewarded; incorrect responses are flagged and the correct form displayed. No process-level explanation of why an error occurred, or how a learner might self-monitor for similar errors in future, is provided. This limits the depth of learning that these tools can support and suggests that they are best understood as practice tools rather than assessment feedback mechanisms.

3.3. Google AI Studio and Speechling

Google AI Studio provides a conversational AI interface with speech recognition capabilities that allow learners to practise spoken Mandarin in real time, capabilities that align closely with the demands of the HSKK, which assesses spontaneous spoken production through tasks including sentence repetition, open-ended question responses, and extended monologue production across Elementary, Intermediate, and Advanced levels. The author used this tool with IBDP students in Singapore to simulate HSKK-style spoken response tasks and HSK listening comprehension scenarios. The tool's multimodal capabilities, processing both text and spoken input make it potentially valuable for integrated skill development across HSKK levels. However, its tonal accuracy detection is inconsistent, particularly for non-native speakers producing second and third tones in connected speech. This limitation is significant given that tonal accuracy is a core assessment criterion in the HSKK and is assessed from the earliest HSK levels.

Speechling offers a more focused approach to pronunciation feedback, providing learners with a score based on comparison with a native-speaker model. Its clear pronunciation scoring makes it a valuable tool specifically for the tonal and segmental accuracy demands of HSKK preparation at Elementary and Intermediate levels. Its limitation, however, is scope: it does not address reading, writing, or grammatical complexity, and its free tier restricts the number of daily submissions. For comprehensive preparation across HSK written and HSKK oral components, it would therefore need to be combined with other tools.

4. Multilingual and Vocational Education Contexts

4.1. Singapore: Bilingual Policy and Mandarin Learning

Singapore's constitutionally mandated bilingual education policy, which requires all students to study English alongside their designated mother tongue language (Mandarin, Malay, or Tamil), creates a distinctive learning environment for Mandarin as a second language. For ethnically Chinese students, Mandarin is officially designated as their mother tongue regardless of whether it is used at home-a policy that has generated significant debate about the relationship between linguistic heritage and classroom language instruction. For non-Chinese students pursuing HSK certification, Mandarin functions as a fully foreign language, acquired exclusively through formal instruction.

This policy context has direct implications for AI feedback tool selection. Students whose home language is Mandarin or a Chinese variety may require AI tools that address pragmatic and register differences between spoken and written Mandarin, as these are the areas where heritage speakers typically show the greatest variability. By contrast, learners with no prior exposure to Mandarin, common in international schools, require tools that provide systematic feedback on tonal production and character recognition from the earliest stages. A one-size-fits-all AI tool is thus unlikely to serve the full range of learners in Singapore's multilingual schools, and teacher mediation in tool selection and use remains essential.

4.2. The Philippines: IBDP International School Contexts

In the Philippines, Mandarin learning has expanded significantly in recent years, driven partly by growing

economic ties with mainland China and partly by the aspirations of internationally minded learners seeking university places in Singapore, Taiwan, and mainland China. The author's teaching experience at Singapore School Cebu, an English-medium international school in the Philippines, provides a contrasting but complementary context to Singapore. Here, the student population ranges from early childhood learners to International Baccalaureate Diploma Programme (IBDP) students aged 16 to 18. Among older learners pursuing the IBDP Mandarin ab initio programme, Mandarin proficiency typically corresponds to HSK Levels 3 to 5, a range that reflects sustained formal instruction across the secondary years. A significant proportion of these students aspire to pursue higher education in Singapore, Taiwan, or mainland China, where Mandarin proficiency is either a formal entry requirement or a significant competitive advantage. This aspirational dimension gives HSK certification a direct and meaningful purpose, shaping learners' motivation and engagement with assessment preparation in ways that differ markedly from learners pursuing Mandarin as a purely elective subject.

This context introduces specific demands on AI feedback tools that differ from those observed at lower proficiency levels. IBDP Mandarin ab initio learners require feedback that extends beyond vocabulary and tonal accuracy to encompass extended written production, integrated listening-reading tasks, and oral interaction skills-competencies assessed in both the IB external examinations and at HSK Levels 4 and 5. Classroom observation in this context suggests that students made regular use of ChatGPT to draft and revise written assignments, seeking feedback on cohesion, argument structure, and lexical variety, as well as to practise spoken dialogue scenarios aligned with IB and HSK topic domains. However, the absence of HSK-specific rubric alignment in these AI tools meant that students received feedback calibrated to general communicative adequacy rather than the specific constructs assessed at HSK Level 4 or 5. Furthermore, for students preparing to study or work in diverse Sinophone environments, mainland China, Taiwan, or Singapore, where the pragmatic and cultural dimensions of Mandarin communication are particularly salient, yet remain largely unaddressed by current AI feedback tools. This points to the need for AI tools that engage with HSK's communicative competence constructs more deeply at intermediate and upper-intermediate levels, and for teacher guidance in contextualizing AI feedback within learners' specific academic and migration goals.

4.3. Broader Implications for Multilingual Contexts

The patterns observed in Singapore and the Philippines have resonance across a range of other multilingual contexts where Mandarin is taught as a second or foreign language. In East and Southeast Asian educational systems where learners may speak a Chinese variety at home, such as Cantonese, Hokkien, or Hakka, the transition to Standard Mandarin for HSK purposes does not involve a fully foreign language but a related variety with significant phonological, lexical, and grammatical differences. AI tools calibrated primarily for learners from non-Sinitic language backgrounds may thus be poorly suited to the specific interlanguage patterns of heritage or dialect-background learners. Similarly, in European and North American contexts where Mandarin is taught as an elective foreign language in multilingual schools, the learner population may span a wide range of first languages, requiring AI feedback tools that are adaptable to diverse phonological starting points.

Across these varied contexts, a common theme emerges: no single AI feedback tool is universally suited to all multilingual learner profiles. The tool selection process must therefore be understood as a pedagogical decision requiring teacher expertise, contextual knowledge, and ongoing evaluation. Future research comparing AI feedback tool effectiveness across different multilingual settings would make a valuable contribution to the field, enabling more evidence-based recommendations for practitioners working in diverse educational environments worldwide.

5. Challenges and Limitations of AI Feedback for HSK

Notwithstanding their considerable affordances, AI feedback tools for HSK preparation face several significant challenges that warrant careful consideration by practitioners and researchers alike.

First, the alignment problem: as noted throughout this review, none of the freely available AI tools examined is explicitly calibrated to the HSK assessment rubric. This means that feedback may be accurate from a general Mandarin linguistic standpoint while being misleading from an examination preparation perspective. Teachers using these tools should supplement AI feedback with explicit instruction on HSK task types, marking criteria, and examiners' expectations.

Second, the cultural-pragmatic gap: Mandarin communication norms, including forms of address, directness conventions, and politeness strategies are deeply culturally embedded and cannot be reduced to linguistic rules that current AI systems can reliably detect or teach. For learners in multilingual contexts who are simultaneously navigating their own cultural communication norms, the absence of culturally sensitive feedback in AI tools represents a significant limitation. This is particularly relevant in the vocational education contexts described above, where professional Mandarin communication requires cultural as well as linguistic competence.

Third, the teacher role: a consistent finding across AI language learning research is that tools are most effective when integrated into a structured pedagogical framework mediated by a skilled teacher [3]. Learners who use AI tools in isolation, without teacher guidance on how to interpret and act on feedback, may develop inaccurate metalinguistic representations or fail to transfer AI-mediated gains to examination performance. This finding has particular implications for vocational education contexts, where learners may be largely self-directed and have limited access to specialist Mandarin teachers.

Fourth, data privacy and institutional considerations: the use of third-party AI tools in educational settings raises questions about learner data protection, particularly in contexts involving minors or sensitive personal information. Schools and vocational institutions adopting these tools should ensure compliance with applicable data protection regulations and develop clear policies on AI use in assessment preparation.

6. Conclusion

This paper has examined the potential of AI-supported

feedback tools for second language Mandarin learners preparing for YCT, HSK, and HSKK examinations in multilingual international school contexts. Drawing on SLA theory, the Chinese proficiency assessment framework, and practitioner observation from international school settings in Singapore and the Philippines, the review has identified significant affordances, particularly in the areas of immediate task-level feedback, tonal pronunciation scoring for HSKK preparation, vocabulary reinforcement for YCT and lower HSK levels, and written feedback for HSK Levels 3–5, alongside important limitations relating to assessment alignment, cultural-pragmatic depth, and the risk of displacing teacher-mediated feedback.

The analysis suggests that freely available AI tools are best understood as pedagogical supplements rather than replacements for expert teacher feedback. Their value lies in extending the feedback available to learners beyond the temporal and personalisation constraints of classroom instruction, but this value is maximised only when tools are selected thoughtfully with reference to learner needs, proficiency level, and HSK assessment objectives, and when teacher guidance is provided on how to interpret and apply AI-generated feedback.

For future research, several directions are indicated. First, empirical studies examining the actual impact of AI feedback tool use on YCT, HSK, and HSKK examination outcomes across different multilingual learner populations are needed. Second, research into the development of assessment-aligned AI feedback systems that provide feedback explicitly calibrated to YCT, HSK, and HSKK rubrics and communicative competence constructs would represent a significant advance. Third, practitioner-researcher studies conducted in Singapore, the Philippines, and comparable multilingual international school contexts worldwide would enrich the evidence base and inform context-sensitive pedagogical recommendations. As AI technology continues to develop, the intersection of standardized Chinese language assessment and artificial intelligence feedback represents a fertile and practically important area for applied linguistics research.

Acknowledgments

The author thanks the students and colleagues at international schools in Singapore and the Philippines whose classroom experiences informed the observations presented in this paper.

References

[1] Chen, X., & Li, H. (2020). Gamification and vocabulary acquisition in Chinese as a second language: A review of app-based approaches. Language Learning & Technology, 24(2), 45–63.

[2] Ellis, R. (2009). Corrective feedback and teacher development. L2 Journal, 1(1), 3–18.

[3] Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. Computer Assisted Language Learning, 27(1), 70–105.

[4] Hanban. (2021). HSK Standard Course: Level descriptions and vocabulary lists (3rd ed.). Beijing Language and Culture University Press.

[5] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112.

[6] Horwitz, E. K. (2001). Language anxiety and achievement. Annual Review of Applied Linguistics, 21, 112–126.

[7] Huo, Z., Zhang, M., & Waibel, A. (2019). Automatic speech recognition for Mandarin tonal feedback in L2 learning. Speech Communication, 108, 31–44.

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

[9] Krashen, S. D. (1985). The input hypothesis: Issues and implications. Longman.

[10] 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.

[11] McNamara, T. (2000). Language testing. Oxford University Press.

[12] OpenAI. (2023). ChatGPT (GPT-3.5). https://openai.com

[13] Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), Input in second language acquisition (pp. 235–253). Newbury House.

[14] Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies: An International Journal, 3(1), 22–36.