Educational and Humanities
ISSN Print: 3104-4352
ISSN Online: 3104-4360
CD ISSN: 3104-4379
https://doi.org/10.63313/EH.9129

Practice and Challenges: The Application of Large Language Models in English Education and Development Research in the Greater Bay Area

Wenyu Guan
School of Foreign Studies, Zhaoqing University, Guangdong Province, 526020, China

How to cite this paper: Guan, W. (2026). Practice and challenges: The application of large language models in English education and development research in the Greater Bay Area. Educational and Humanities, 5(2), 52–61. ISSN Print: 3104-4352, ISSN Online: 3104-4360.
https://doi.org/10.63313/EH.9129
Published: 2026-11-26

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
CC BY 4.0

Abstract

Large language models (LLMs) are now embedded in English teaching, learning, and research across the Guangdong–Hong Kong–Macao Greater Bay Area (GBA). Drawing on a focused critical synthesis of ten recent studies and reviews, this article maps five sites of practice—writing feedback, oral interaction, personalized reading, teacher professional learning, and research support—and examines six challenges: assessment integrity, output reliability, cognitive over-reliance, linguistic equity, data governance, and teacher readiness. The evidence indicates that LLMs widen access to timely feedback and reduce the transactional costs of scholarship, but that benefits depend on task design, metacognitive engagement, and institutional support. Because the GBA's biliterate and trilingual ecology makes generic guidance insufficient, a context-sensitive framework is proposed around pedagogical orchestration, transpar-ency, human accountability, and local adaptation. Implications for policymakers, teacher educators, and researchers are discussed.

Keywords

Large Language Models; Generative Artificial Intelligence; English Language Education; Greater Bay Area; Academic Integrity; Teacher Development

1. Introduction

The release of conversational generative artificial intelligence in late 2022 marked a discontinuity in educational technology. Unlike scripted tutors, large language models (LLMs) generate contextually sensitive text across a wide range of tasks. English is especially exposed to this shift: it is the object of instruction, the medium of academic communication, and the working language of the tools themselves. A learner can obtain a grammar explanation, a model paragraph, a simulated conversation partner, or a critique of an argument within one interface (Lee et al., 2025).

A systematic review of 49 classroom studies found higher education dominant and

English the primary target language, with acquisition research concentrated in writing (Lee et al., 2025). A meta-analysis of 57 studies and 5,389 university students reported positive aggregate effects on language skills, achievement, and motivation, but no significant effect on metacognition (Chen & Cheung, 2025).

The pattern carries a warning: the dimensions most readily improved are not necessarily those central to independent language development. Where a tool supplies answers, it may improve a product while leaving the processes that produce competence untouched. The tension is acute in the Guangdong–Hong Kong–Macao Greater Bay Area (GBA), where 11 cities operate under one regional framework but retain three administrative, legal, and educational systems.

Hong Kong's "two scripts and three languages" policy positions English as a global resource alongside Chinese cultural and civic functions; mainland Guangdong institutions follow national curriculum standards and, in many programmes, English-medium instruction; Macao adds a Portuguese-influenced dimension.

An LLM entering this ecology is not neutral: it carries an English-dominant training distribution and is interpreted through institutional cultures that differ sharply across jurisdictions. This article addresses the fragmentation in research connecting LLM use to GBA conditions. It maps principal applications, examines the associated challenges, and proposes a context-sensitive framework. As a critical synthesis rather than new fieldwork, its contribution lies in organizing recent evidence and specifying a regionally grounded research agenda.

2. Conceptual and Contextual Foundations

2.1. From Tool to Participant

Earlier AI in language learning operated behind the interface: automated writing evaluation, intelligent tutoring, and speech recognition classified and scored on behalf of teachers or learners. LLMs alter this relationship.

By generating rather than selecting language, one system can act as feedback provider, tutor, cognitive stimulator, and conversation partner—simplifying a text, producing a parallel conversation, and explaining why one revision works better than another (Lee et al., 2025).Such mediation is not automatically beneficial. A tool may scaffold performance without fostering internalisation, or displace the struggle through which competence develops. Performance gains observed while a tool is available say little about what learners can do once it is withdrawn.

This distinction is central to the evidence: much classroom research relies on self-reported perceptions rather than measured acquisition (Lee et al., 2025), and the meta-analysis by Chen and Cheung (2025) found no significant metacognitive effect. Integration therefore depends less on access to a tool than on whether it is embedded in tasks requiring learners to plan, monitor, and evaluate their own performance.

2.2. The Greater Bay Area as a Language Ecology

The GBA comprises systems with different curricula, examination regimes, qualification requirements, and data rules, all of which shape the demand for LLM support and the acceptability of particular uses.

Hong Kong's bilingual and trilingual policy, Guangdong's national curriculum standards, and Macao's Portuguese-influenced multilingual environment create overlapping but non-identical expectations of what English competence means and how it should be demonstrated. These differences are compounded by migrant populations, international campuses, and vocational institutions, whose learners bring varied literacies and purposes to the same tools.

A model trained on standard, high-resource English may reproduce assumptions uneasy with local priorities; for Cantonese-speaking learners its handling of translanguaging and code-switching may be unreliable; and an explanation aligned with one assessment regime may mislead in another.

"AI literacy" in the region must therefore include judging whether an output suits a specific linguistic, curricular, and institutional setting, rather than treating fluency and plausibility as evidence of appropriateness.

3. Analytical Approach

This study adopts an integrative critical review designed to map practice rather than estimate a pooled effect. Peer-reviewed articles and policy documents published mainly between 2023 and 2025 were identified through searches combining large language model, generative AI, ChatGPT, English language teaching, teacher education, academic integrity, and Greater Bay Area / Hong Kong / Macao.

Sources were selected for a substantive focus on LLM-enabled practice, a language- or higher-education setting, and relevance to the region. They were organised by site of application (classroom, teacher development, research) and nature of claim (measured effect, perception, or conceptual argument).

The organisation aims at transparency about the strength of evidence. Experimental work supports some applications, while others rest mainly on interviews and self-report—valuable for understanding experience but unable to establish effectiveness. This distinction is maintained throughout the discussion: where this article calls an application "promising" rather than "effective", it is because the literature has that character (Lee et al., 2025).Evidence of learner satisfaction, for example, is reported as such and is not treated as proof of acquisition.

The review is interpretive rather than exhaustive, and the framework developed in Section 6 is a reasoned proposal for adaptation and testing, not a validated model. Its purpose is to make explicit the pedagogical and institutional conditions that recent studies suggest are decisive.

4. Practices in English Education

4.1. Writing Instruction and Feedback

Writing has the densest evidence base. LLMs generate model texts, gloss vocabulary, identify grammatical and discourse-level problems, and suggest revisions within seconds; reviews report substantial feedback opportunities alongside inconsistent quality (Lee et al., 2025). What matters most, however, is how feedback is arranged and whose judgement prevails when sources disagree.

Comparing teacher and GPT-assisted written corrective feedback, Lin and Crosthwaite (2024) found that neither source is uniformly superior: teachers are better placed to diagnose learner-specific problems and to interpret them in the light of a course trajectory, whereas GPT offers breadth, speed, and availability outside class hours. Learners consequently tend to favour mixed feedback ecologies rather than a single provider.

Teng (2025a) found that learners valued ChatGPT feedback for its practicality and interactivity but wanted teacher confirmation; behaviours ranged from copying strings of text to strategic, critical use, with metacognitive awareness the strongest differentiator.

The question thus shifts from whether AI feedback works to which learners are equipped to use it well—a distinction consistent with the absence of metacognitive gains in the wider evidence (Chen & Cheung, 2025).

In GBA universities, students write for both local assessment and international publication, using LLMs to navigate genre conventions across the two. A model trained on dominant academic English conventions may flatten the rhetorical choices local scholars make when writing back to international journals. Feedback is most useful when teachers help learners judge which suggestions serve their purpose, rather than treating output as an authoritative verdict.

4.2. Speaking, Listening, and Interaction

Speaking is the most intuitive use case: an LLM can simulate conversation without the time pressure and judgement anxiety of a human interlocutor. Interviewing 21 first-year students and three instructors in Hong Kong, Kohnke and Moorhouse (2025) found that GenAI enhanced motivation, reduced anxiety, and fostered a supportive affective environment, although cultural context, technical problems, and the instructor's role moderated these effects.

Affective benefit should not be confused with speaking gains. Reviews report improved willingness to communicate, but longitudinal evidence for oral development remains limited (Lee et al., 2025). An endlessly accommodating partner may reinforce errors rather than correct them, and a default to standard English may reduce opportunities to use Cantonese or Putonghua as resources.

Speaking applications are best framed as rehearsal within human-led instruction, not substitutes for it.

4.3. Reading, Vocabulary, and Personalisation

LLMs can adapt text difficulty, generate glosses, and explain lexical items in context. Personalization—responding to a learner's proficiency and interests—is attractive in large classes, and reviews identify tailored input as a frequently reported benefit (Lee et al., 2025). Yet simplification may strip a text of the features worth reading, and generated questions may test recall rather than inference.

The same caution applies to vocabulary work: a gloss delivered on demand supports immediate comprehension, but learners still need repeated, varied encounters with a word before it becomes part of their productive repertoire. Adaptation should therefore preserve a manageable level of difficulty rather than remove difficulty altogether.

4.4. Teacher Professional Learning

Teachers are the bottleneck in implementation. Intervention research in initial language teacher education shows that professional generative AI competence can be developed through designed coursework covering ethical judgement, prompt design, and output evaluation (Moorhouse et al., 2024). Familiarity with a tool is not the same as knowing when and why to deploy it.

In the GBA, provision differs across the three systems and cross-border learning is constrained by certification and language. A practical starting point is to treat AI competence as situated professional judgement, so teachers ask not "Can this tool do the task?" but "Is this an appropriate use for these learners, in this curriculum, under these assessment rules?"

4.5. Applications in Development Research

LLMs are changing how research is conducted. Examining graduate students' use of ChatGPT for academic revision, Koltovskaia et al. (2024) found distinct behavioral, cognitive, and affective engagement patterns, suggesting the tool alters how students relate to their own writing.

Broader uses include literature mapping, transcription cleanup, translation of interview data, initial coding, and academic editing for scholars writing in English as an additional language.

These uses bring ethical obligations. A model may omit relevant studies, misattribute claims, fabricate citations, or flatten dialectal nuance in translation; AI-assisted editing may improve clarity while shifting an argument toward dominant conventions. Responsible practice requires verification against sources, documentation of tool use, and—in many journals—explicit disclosure. These requirements concern the integrity of the knowledge produced, not merely technical procedure.

Table 1. Principal applications of LLMs in English education and development research

Domain	Representative uses	Main benefit	Principal caveat
Writing	Feedback, revision, modelling	Rapid, individualised, iterative support	Quality varies; may encourage copying
Speaking	Rehearsal, role-play, conversation	Lower anxiety, greater practice volume	Limited evidence of transferable oral gains
Reading and vocabulary	Simplification, glossing, questioning	Tailored input and explanation	Risk of removing productive difficulty
Teacher development	Lesson planning, evaluation, critical AI literacy	Builds context-sensitive competence	Requires structured programmes, not tool demos
Research	Literature mapping, coding support, academic editing	Reduces transactional costs	Verification, disclosure, and bias risks

5. Challenges and Tensions

5.1. Academic Integrity and Assessment Validity

If an essay can be produced in minutes, an assessment rewarding only the final product no longer measures what it claims. A study of guidelines across teaching, learning, research, and administration found that many remain general and leave frontline decisions unresolved (An et al., 2025). Declaration mechanisms are fragile for the same reason: integrity cannot be secured through detection and disclosure alone.

Assessment must instead make the process visible—through drafts, oral defenses, in-class writing, or tasks tied to local data a model cannot fabricate. Tiered vocabularies such as the AI Assessment Scale specify the permitted degree of AI involvement for a given task, converting an abstract rule into a concrete agreement between teacher and learner (Furze et al., 2024).

Such shared vocabularies are especially valuable in the GBA, where students may submit work across institutions operating under different policies and where a single, regionally intelligible language of acceptable use can prevent misunderstanding. Their value lies less in enforcement than in making expectations explicit before the work begins.

5.2. Reliability and Feedback Quality

LLMs generate fluent but not always accurate text. Errors may be subtle: a fabricated citation, an inappropriate register, a false rule explanation, or a "correction" that changes meaning. Reviews stress that AI output should supplement rather than replace other sources (Lee et al., 2025), and feedback quality varies across prompts and text types (Teng, 2025a).

Reliability is distributed: a teacher unable to evaluate a suggestion and a learner assuming fluency equals correctness are both exposed, particularly in specialized English for academic purposes where disciplinary conventions matter. Reliability is a property of the human–tool system, not of the tool in isolation.

5.3. Over-Reliance and Metacognition

The most serious pedagogical risk is cognitive offloading. Chen and Cheung (2025) found positive effects on language skills and motivation but none on metacognition—a coherent result, since metacognition develops through planning, monitoring, and self-evaluation that a responsive tool can perform on the learner's behalf.

Teng (2025a) likewise found that metacognitive awareness distinguished effective from ineffective use: the tool reveals self-regulation rather than creating it. Tasks should therefore require learners to predict feedback, justify accepting or rejecting it, and reflect on differences from the machine's judgement. Without such scaffolding, repeated assistance may produce hard-to-detect dependence because products continue to look good.

5.4. Equity, Language Ideology, and Cultural Bias

LLMs are not linguistically neutral. Their training distributions privilege high-resource languages and standardized varieties. For GBA learners this means English-dominant models may devalue translanguaging, automated feedback may enforce a narrow correctness, and uneven access to tools and AI-literate teachers may widen inequality.

Francis et al. (2025) identify the digital divide and algorithmic bias among the central ethical concerns of GenAI in higher education. Where English teaching is linked to values education in Hong Kong or national standards in Guangdong, imported tools may introduce content requiring critical filtering. Such concerns require shared regional principles on transparency and the protection of local linguistic practices.

5.5. Data Governance and Policy Fragmentation

Student writing, interview transcripts, and assessment records may be uploaded to systems hosted in other jurisdictions. Storage, consent, and transfer rules differ across mainland China, Hong Kong, and Macao, while institutional guidelines differ further and often fail to resolve operational questions (An et al., 2025).

Francis et al. (2025) argue for transparent policies ensuring fairness and accountability; in the GBA these must also specify where data go and whose law applies. Fragmentation affects collaboration, since a tool approved at one university may be restricted at another. Regional, non-binding principles on minimisation, disclosure, retention, and sensitive material would reduce duplication.

5.6. Teacher Readiness and Institutional Capacity

Many teachers encounter LLMs through personal experimentation, so judgement may be shaped by novelty rather than evidence. Intervention research shows competence can be built, but requires time and designed experience rather than brief demonstrations (Moorhouse et al., 2024). Institutional capacity matters equally: without leadership, updated regulations, and technical support, teachers manage ethical dilemmas alone and practice becomes inconsistent.

The challenge is systemic, requiring alignment across curriculum, assessment, development, and governance.

6. Discussion: Toward a Context-Sensitive Framework

The evidence supports a qualified conclusion: LLMs offer genuine value in expanding access to feedback and reducing research costs, but benefits are conditional on pedagogical and institutional arrangements rather than technical capability. Three principles follow.

Pedagogical orchestration over tool adoption. Begin with a learning goal, not a technology. Tasks should specify what learners do before, during, and after AI involvement—an initial draft, a rationale for accepted suggestions, a reflection, or AI rehearsal paired with human interaction. The teacher's role shifts to designing sequences and adjudicating outcomes.

Transparency and human accountability. Institutions should adopt a shared vocabulary for acceptable use and require disclosure proportionate to involvement (An et al., 2025). Transparency protects students from ambiguity and teachers from unenforceable policies. A named human remains responsible for submitted work and published claims, including citations and translations; verification should be a documented research step.

Regional collaboration with local adaptation. Diversity becomes a liability when each institution solves the same problems alone. A regional forum for sharing policies and bilingual AI-literacy materials, adapted to Cantonese–English, Putonghua–English, and Portuguese-influenced contexts, would reduce duplication. Generic guidance should be a starting point, not a solution (Francis et al., 2025).

Table 2. A context-sensitive integration framework for the GBA

Principle	Classroom application	Research application	Governance requirement
Pedagogical orchestration	AI used within designed task sequences; human follow-up	Documented purpose for each AI-assisted step	Curriculum and assessment alignment
Transparency	Shared acceptable-use vocabulary; learner disclosure	Disclosure of tool use and verification steps	Regional policy principles
Human accountability	Teacher adjudicates feedback; learner retains authorship	Author verifies citations, translations, claims	Clear responsibility rules
Local adaptation	Materials reflect local language ecology	Attention to dialect and register in data	Cross-jurisdictional data protocols

7. Conclusion

Large language models are already part of the GBA's educational reality. Their applications in writing, speaking, reading, teacher development, and research are now well documented, but their value depends less on technical capability than on the pedagogical and institutional arrangements surrounding them. The principal challenges—assessment validity, unreliable output, cognitive over-reliance, inequity, fragmented governance, and uneven readiness—are equally clear (An et al., 2025; Chen & Cheung, 2025; Francis et al., 2025). Since the GBA combines multiple educational systems within one multilingual ecology, solutions imported from monolingual settings are unlikely to travel well.

The framework proposed here rests on pedagogical orchestration, transparency, human accountability, and local adaptation, and is offered as a basis for dialogue rather than a finished prescription. Its four principles are deliberately stated at a level of generality that institutions can translate into their own curricula, assessment rules, and data protocols. What is needed now is longitudinal, comparative evidence attentive to the region's multilingual realities and to the uneven distribution of AI literacy across its cities and institutions. Such evidence would allow the GBA to contribute not merely as a site of adoption but as a source of context-sensitive theory and practice for language education in an AI-mediated world.

Acknowledgements

We thank the following projects for the fund support:

Research Project of Zhaoqing University: The Application of Large Language Models in the Research on English Teacher Education and Development in the Greater Bay Area (fw202509)";

Foreign Language Teaching and Research Project: A Study on Strategy for Enhancing University English Teachers' Digital Competence in the Context of Digital Intelligence

(2025070302);
2026 Zhaoqing University Development and Planning Project: Reform of English Reading Course Evaluation Mechanism for English Majors Empowered by Digital Intelligence (JYPJ202604);
2026 Zhaoqing Philosophy and Social Sciences Planning Project: Enhancing Digital Literacy of County-Level Secondary School Teachers in Zhaoqing Amidst Digital Transformation: A Strategy Research Study(26GJ-09)

References

[1] An, Y., Yu, J. H., & James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. International Journal of Educational Technology in Higher Education, 22(1), 1–23. https://doi.org/10.1186/s41239-025-00507-3

[2] Chen, S., & Cheung, A. C. K. (2025). Effect of generative artificial intelligence on university students' learning outcomes: A systematic review and meta-analysis. Educational Research Review, 49, Article 100737. https://doi.org/10.1016/j.edurev.2025.100737

[3] Francis, N. J., Jones, S., & Smith, D. P. (2025). Generative AI in higher education: Balancing innovation and integrity. British Journal of Biomedical Science, 81, Article 14048. https://doi.org/10.3389/bjbs.2024.14048

[4] Kohnke, L., & Moorhouse, B. L. (2025). Enhancing the emotional aspects of language education through generative artificial intelligence (GenAI): A qualitative investigation. Computers in Human Behavior, 167, Article 108600. https://doi.org/10.1016/j.chb.2025.108600

[5] Koltovskaia, S., Rahmati, P., & Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. Journal of Second Language Writing, 65, Article 101130. https://doi.org/10.1016/j.jslw.2024.101130

[6] Lee, S., Choe, H., Zou, D., & Jeon, J. (2025). Generative AI (GenAI) in the language classroom: A systematic review. Interactive Learning Environments, 34(1), 335–359. https://doi.org/10.1080/10494820.2025.2498537

[7] Moorhouse, B. L., Wan, Y., Wu, C., Kohnke, L., Ho, T. Y., & Kwong, T. (2024). Developing language teachers' professional generative AI competence: An intervention study in an initial language teacher education course. System, 125, Article 103399. https://doi.org/10.1016/j.system.2024.103399

[8] Teng, M. F. (2025a). Metacognitive awareness and EFL learners' perceptions and experiences in utilizing ChatGPT for writing feedback. European Journal of Education, 60(1), Article e12811. https://doi.org/10.1111/ejed.12811

[9] Moorhouse, B. L., Wan, Y., Wu, C., Kohnke, L., Ho, T. Y., & Kwong, T. (2024). Developing language teachers' professional generative AI competence: An intervention study in an initial language teacher education course. System, 125, Article 103399. https://doi.org/10.1016/j.system.2024.103399

[10] Teng, M. F. (2025a). Metacognitive awareness and EFL learners' perceptions and experiences in utilizing ChatGPT for writing feedback. European Journal of Education, 60(1), Article e12811. https://doi.org/10.1111/ejed.12811