Chapter 4

Assessing EFL Students' Learning in the AI Era

Arida Susyetina
Universitas Kristen Duta Wacana
arida@staff.ukdw.ac.id

A. THE CONTEMPORARY CHALLENGE

Since late 2022, generative artificial intelligence (AI) has moved from a novelty to a routine feature of how many students prepare written and spoken assignments. A British Council global survey of 1,348 English language teachers across 118 countries and territories found that teachers were already using AI for a range of teaching tasks, while explicitly identifying assessment as an area requiring further research (Edmett et al., 2023). For English Language Teaching (ELT), this shift touches something foundational: much of what teachers use as evidence of learning, essays, reports, presentation scripts, reflective journals, can now be partly or wholly generated or restructured by a chatbot before a student submits it. The deeper issue is one of assessment validity: whether a submitted product still constitutes credible evidence for the decisions teachers make about a student's knowledge and communicative competence. The central assessment question is therefore not whether AI was involved, but whether the task still elicits the knowledge, skills, and performances it was designed to measure.

This problem is especially consequential in ELT because language proficiency is demonstrated through performance, not

merely reported: a polished written text may display sophisticated vocabulary without revealing whether a learner can formulate ideas independently, respond spontaneously, or negotiate meaning in real time. Nikolic et al. (2023) tested ChatGPT against real assessment prompts drawn from ten university subjects and found that, with only minor adjustments to task wording, the tool could generate responses good enough to pass many existing assessments. Applied to English language education, this finding is unsettling: a written task designed to show that a student can construct an argument, organise ideas coherently, or use a target grammatical structure accurately may no longer reliably show whether the student, or the machine assisting them, is the one demonstrating that competence.

Two further complications sharpen the problem, and both carry particular weight for language learners. First, AI-detection software, a common institutional response to this uncertainty, has shown important limitations in accuracy and robustness, and simple evasive edits can reduce that accuracy further still (Perkins et al., 2024; Weber-Wulff et al., 2023). More troubling for ELT, these detectors systematically misclassify writing by non-native English speakers as AI-generated far more often than writing by native speakers, since simpler, more constrained vocabulary choices are read as signs of machine authorship (Liang et al., 2023). For a discipline whose students are, by definition, developing writers of English, leaning on detection risks penalising the very learners it is meant to serve. Second, UNESCO's guidance on generative AI in education warns that unregulated use can erode human agency in learning unless institutions deliberately redesign how learning is validated (Miao & Holmes, 2023), and scholarship agrees that punitive, detection-based responses treat a pedagogical problem as a policing problem (Cotton et al., 2024; Evangelista, 2025). For English teachers, the unresolved question is therefore not whether to permit AI, but how to keep collecting credible evidence of learning in a landscape where submitted work can no longer be assumed to be a student's alone, and where the tools meant to verify authorship cannot be trusted to do so fairly.

B. THE PRACTICAL SOLUTIONS

Responding to this challenge does not mean banning AI from the English classroom, nor does it mean surrendering assessment entirely to whatever a chatbot can produce. It means deliberately redesigning assessment so that the evidence teachers collect is harder to substitute and easier to trust. The four practical solutions below offer a starting point for English teachers who want to preserve authentic evidence of learning while still allowing students to benefit from AI where it genuinely supports language development.

Redesign Assessment Around Authentic Language Use

Generic, decontextualised writing prompts are exactly the kind of task generative AI handles most comfortably, because they require little situated knowledge. English teachers can make tasks harder to outsource by anchoring them in authentic communicative purposes and audiences: a persuasive letter to a real local authority about a genuine community issue, a recorded interview with a classmate, a role play negotiating a real-world scenario, or a project requiring students to gather information unique to their own context. In a public speaking course, this might mean a pitch followed by unscripted questions from classmates; in an ESP class, a response to a workplace scenario requiring an appropriate register and professional strategy (Astawa et al., 2026). Speaking tasks and presentations with unpredictable follow-up questions resist AI substitution because they require language produced in real time rather than submitted as a polished artefact. Authenticity is not, however, a complete safeguard: Kofinas et al. (2025) found that a task's authenticity had limited effect on markers' ability to detect GenAI involvement, suggesting authentic design must be combined with the strategies below rather than used alone. The trade-off is practical: authentic tasks take longer to design and mark, so teachers with large classes may need to phase this redesign in gradually.

Assess the Learning Process, Not Only the Final Product

If a finished essay can no longer be trusted as the sole evidence of learning, teachers can widen what counts as evidence to include the process behind it: drafts, revision histories, brief written reflections on why a sentence was changed, short one-to-one consultations, or simple learning logs kept over a unit of work. The purpose of collecting this evidence is not to police students but to make the ordinarily invisible process of language learning visible again, so a teacher can see how a piece of writing developed rather than judging only its final polish. In practice, this can be as simple as asking students to submit an early draft alongside the final version, or holding brief two- or three-minute conferences in which a student talks through a choice they made in their own writing. The main limitation is time: process-based assessment demands more regular contact between teacher and student than a single end-of-unit submission, which can be difficult in under-resourced or examination-driven systems.

Make AI Use Transparent and Purposeful

Rather than pretending AI does not exist, teachers can define, task by task, what level of AI assistance is acceptable and require students to disclose what they used (Fajaria et al., 2026). Furze et al. (2024) propose the AI Assessment Scale (AIAS), a five-level framework running from "no AI" to "full AI" that lets teachers specify exactly which parts of a task may involve AI support and which must be produced independently; a pilot implementation at British University Vietnam using this scale was associated with a reduction in AI-related misconduct cases alongside improved student attainment. Roe et al. (2026) have since adapted this scale specifically for English for Academic Purposes contexts, recognising that language learners already routinely use tools such as grammar checkers and translators, so the boundary between acceptable linguistic support and inappropriate AI substitution needs to be drawn more carefully than in frameworks built for other disciplines. English teachers can adopt a simplified version of this idea: stating on an assignment sheet whether AI may be used for brainstorming but not for drafting, for example, and asking students to attach a short statement of how, if at all, they

used AI. This will not eliminate misuse, but it shifts the relationship with AI from a hidden, guessed-at risk to an openly negotiated classroom practice.

Strengthen Human Evidence Through Reflection and Oral or Performance Assessment

Where a written product remains uncertain evidence, teachers can pair it with a short spoken component that only the student who did the work can plausibly complete. Nikolic et al. (2023) found that ChatGPT could not directly substitute for a student in oral assessment, although they cautioned that AI-generated text could still be memorised and delivered as a rehearsed script, and recommended greater use of unscripted, interactive questioning to capture genuine understanding. This suggests oral interaction provides a meaningfully different layer of evidence only when students must respond spontaneously rather than reproduce prepared text. This logic is now being applied at a national scale: in August 2026, Denmark's education ministry began requiring around 9,000 upper-secondary students to orally defend a major annual written assignment, explicitly framing the oral defence as a way to confirm understanding rather than to detect machine authorship (The Guardian, 2026). A student who can explain and extend their own argument passes regardless of what tools touched an earlier draft; one who cannot, does not. In an English classroom, a two-minute follow-up question after a written submission, a short viva on a group project, or simply asking a student to paraphrase their own conclusion aloud can serve the same purpose, with the added benefit, specific to language teaching, of directly assessing the speaking competence the course is meant to develop. Oral assessment should nonetheless be implemented carefully: a systematic review of the factors that support strong oral-assessment performance in higher education highlights the importance of scaffolded practice, clear criteria, and attention to student anxiety and fairness, rather than treating a spoken component as a simple add-on (Stephenson et al., 2025). The limitation is logistical: oral components take class time and are harder to schedule fairly in very large classes,

so teachers may need to sample rather than examine every student on every task.

C. THE FUTURE DIRECTION

The direction of travel in assessment research and policy is moving away from the question of how to catch AI, toward the more productive question of how to design assessment so that it still tells us something true about learning. UNESCO's AI Competency Framework for Teachers reflects this shift at a systemic level, arguing that teachers need dedicated professional competencies, not just technical familiarity with AI tools, but the pedagogical judgement to decide when AI use supports learning and when it substitutes for it (Miao & Cukurova, 2024). For English teachers, this points toward AI literacy becoming as basic a professional skill as assessment literacy itself (Hasbi, 2026), which is understanding what generative tools can and cannot do with language, so that task design can anticipate rather than merely react to their capabilities, and so that judgements about validity rest on evidence rather than suspicion. A second, related shift is treating AI itself as an object of learning and assessment, not only a tool that threatens it. Students who are taught to critically evaluate AI-generated language, spotting where a chatbot's phrasing is generic, checking its claims, or comparing its output with their own drafting choices, are simultaneously practising the close reading and metalinguistic awareness that language teaching has always valued. The British Council's human-centred framework makes a related point: AI should support rather than replace the teacher's judgement, and equitable access to both the technology and the training to use it critically needs to be built into how institutions plan for AI, rather than left for individual teachers to work out alone (Felice et al., 2025).

None of this suggests that the future of English language assessment is, or should be, AI-free. Banning a technology already embedded in students' daily digital lives is unlikely to be enforceable and would cut students off from tools that, used well, can support drafting, feedback, and revision. What the evidence so far suggests instead is that the future of trustworthy assessment lies in tasks and processes that keep human reasoning, spoken performance, and

genuine communicative purpose visible alongside whatever AI assistance a student may have used, so that, in an AI-enabled classroom, teachers can still say with confidence that a piece of evidence reflects what a particular student has actually learned.

Author



Arida Susyetina is a lecturer in the English Language Education Department at Universitas Kristen Duta Wacana (UKDW), Yogyakarta, Indonesia. Her teaching expertise includes English for Specific Purposes (ESP), public speaking, and cross-cultural communication. Her research interests encompass teacher professional development, intercultural competence in EFL contexts, educational assessment, and technology integration in language education. With international academic experience, she actively engages in research and international collaboration, with a particular commitment to advancing language education and intercultural studies. She can be reached at arida@staff.ukdw.ac.id.

References

Astawa, N. L. P. N. S. P., Hasbi, M., Islamiah, N., Hamid, S. F., Male, H., Kurniati, Khairatunnisa. (2026). *Teaching English for Specific Purposes*. Rizquna.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Edmett, A., Ichaporia, N., Crompton, H., & Crichton, R. (2023). *Artificial intelligence and English language teaching: Preparing for the future* (2nd ed.). British Council. https://doi.org/10.57884/78ea-3c69

Evangelista, E. D. L. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. *Contemporary*

Educational Technology, *17*(1), ep559. https://doi.org/10.30935/CEDTECH/15775

Fajaria, N. H., Maknun, L., Hasbi, M., Islamiah, N., Male, H., Harida, E. S., Kusmaningrum, M. (2026). *AI in English Language Teaching and Research*. Rizquna.

Felice, M., Spiby, R., O'Sullivan, B., & Edmett, A. (2025). *Human-centred AI: Lessons for English learning and assessment*. British Council.

Furze, L., Perkins, M., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. *Australasian Journal of Educational Technology*, *40*(4), 38–55. https://doi.org/10.14742/AJET.9434

Hasbi, M. (2026). AI-based educational video production projects for pre-service EFL teachers. In *AI in English Language Teaching and Research* (pp. 25-37). Rizquna.

Kofinas, A. K., Tsay, C. H. H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*, *56*(6), 2522–2549. https://doi.org/10.1111/bjet.13585

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7). https://doi.org/10.1016/j.patter.2023.100779

Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. In *AI competency framework for teachers*. UNESCO. https://doi.org/10.54675/zjte2084

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. In *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/ewzm9535

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to

investigate assessment integrity. *European Journal of Engineering Education*, *48*(4), 559–614. https://doi.org/10.1080/03043797.2023.2213169

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education 2024 21:1*, *21*(1), 53-. https://doi.org/10.1186/S41239-024-00487-W

Roe, J., Perkins, M., & Tregubova, Y. (2026). The EAP-AIAS: Adapting the AI Assessment Scale for English for Academic Purposes. *TESOL Journal*, *17*(2), e70122. https://doi.org/10.1002/TESJ.70122

Stephenson, Z., Johnson-Glauch, N., & Cruchley, S. (2025). Interventions and facilitators of oral assessment performance in higher education: a systematic review. *Assessment & Evaluation in Higher Education*, *50*(7), 1140–1153. https://doi.org/10.1080/02602938.2025.2504621

The Guardian. (2026, August 6). *Danish pupils will have to orally defend essays in attempt to combat AI cheating | AI (artificial intelligence).* *https://www.theguardian.com/technology/2026/aug/06/students-ai-cheating-schools-denmark*

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity 2023 19:1*, *19*(1), 26-. https://doi.org/10.1007/S40979-023-00146-Z