

Original Research Article

Beyond the Prompt: A Dual-Layer Framework for AI-Resistant Assessment in English Literature Education

Alwyn Alfred Carvalho,¹ Bhumika Gorakhnath Salvi²

¹Assistant Professor, Department of English, VPM's K. G. Joshi College of Arts and N. G. Bedekar College of Commerce (Autonomous), Thane (W), Maharashtra, India

²Assistant Professor, Visiting Faculty, VPM's K. G. Joshi College of Arts and N. G. Bedekar College of Commerce (Autonomous), Thane (W), Maharashtra, India

Corresponding author E-mail: alwyn.carvalho@vpmtree.org,¹ salvibhumika45@gmail.com²

Received: 10 September, 2026 | Accepted: 23 September, 2026 | Published: 30 September, 2026

Abstract

English Literature departments face a peculiar version of cheating using Artificial Intelligence, particularly Large Language Models (LLMs). Most of the canon these courses teach from is already public domain, already digitized, and already sitting inside the training data of most major language models. A student does not need to ask an AI to think about Hamlet; the model has, in some sense, already read it a thousand times over. This paper looks at what actions departments may take for it, and where those responses tend to fall apart. Timed online examinations have picked up a set of technical countermeasures over the past year or so: hidden prompt injections buried in the question text, watermarked or visually distorted question images, AI-based proctoring. One documented case from July 2026 caught thirty-two of thirty-five students this way. But these tricks have a short shelf life. Students figure them out quickly, model providers patch around them faster, and proctoring software frequently incorrectly flag attempts which may be genuine. Take-home writing is a harder problem, and probably the more important one for a literature department, since AI detectors themselves are unreliable and biased against certain kinds of writers. Our argument, briefly, is that assignments need to stop rewarding a finished essay and start rewarding the visible work

behind it including drafts, in-class discussion tied directly into the prompt, or a short oral defence of the argument. None of this is a permanent fix. It is closer to a set of habits that make AI-generated submissions harder to pass off as one's own, drawn from work in AI security, assessment design, and literary pedagogy that rarely gets read together.

Keywords: Academic integrity, Generative AI in education, Assessment design, Prompt injection, English literature pedagogy

Introduction

Generative artificial intelligence has unsettled an assumption that has long underpinned university assessment: that a submitted piece of writing is reasonably strong evidence of a student's own intellectual work. Large language models (LLMs) can now produce fluent explanations, summaries, comparisons, and literary interpretations within seconds, and the resulting concern about academic integrity has already been documented across higher education generally, well before subject-specific responses were developed (Cotton et al., 2023). This creates a particularly difficult problem for English Literature departments. The objects of study, canonical novels, plays, poems, and widely discussed authors, are frequently available in digital form and surrounded by extensive online criticism. A student can therefore ask an LLM to produce an interpretation of a familiar text without first supplying the text or much contextual information. The problem is not simply that AI can write an essay; it is that conventional literary assessment often evaluates the final essay more heavily than the process through which the interpretation was developed.

This paper argues that the assessment of literary understanding should shift from a predominance of detection toward a dual-layered evaluation system. The first layer creates supervised opportunities for students to demonstrate their ability to engage creatively and independently with literature through close reading, discussion, and brief verbal responses. The second layer focuses on making the process of authorship visible through drafts, annotations, research commentary, revisions, and verbal defence. The goal is not to make assessment “AI proof,” since detection-based approaches are structurally unable to keep pace with more capable models (Sadasivan et al., 2023). A more durable strategy diversifies assessment across several stages, so that no single finished piece of writing can substitute for demonstrated literary thinking.

Objectives

The paper has four objectives. First, it explains why generative AI poses a distinctive threat to conventional assessment practice in English Literature. Second, it evaluates the effectiveness of technological responses to generative AI, including prompt injection, AI proctoring, and automated detection. Third, it identifies pedagogical practices that render students' learning visible. Finally, it proposes a dual-layered approach that combines controlled demonstration with process-based assessment and considers its implications for the teaching of English Literature.

Research Methodology

This study employs a qualitative, exploratory, and literature-based methodology. It does not attempt to quantify student behaviour through surveys or experiments. Instead, it critically reviews literature, institutional policy, and recent studies concerning generative AI, academic integrity, AI detectors, and prompt injection. The study is

guided by one research question: what forms of assessment can provide credible evidence of literary learning when AI-generated answers can be indistinguishable from those produced by students?

The literature and policy documents reviewed were selected for their relevance to this question. Particular attention is paid to guidance issued by UNESCO (Miao & Holmes, 2023) and the Tertiary Education Quality and Standards Agency (Lodge et al., 2023), recent empirical studies evaluating the accuracy of AI detectors (Liang et al., 2023; Sadasivan et al., 2023; Weber-Wulff et al., 2023), and technical work on prompt injection (Li et al., 2026). The discussion also draws on discipline-specific evidence from English Literature classrooms (Bellot et al., 2025) and on broader assessment-redesign scholarship (Awadallah Alkouk & Khlaif, 2024; Lye & Lim, 2024). The findings are related to the challenges and implications this literature poses for academic integrity in university-level English Literature teaching.

Generative AI and the Distinctive Problem of Literary Assessment

The question of how to assess written work in the presence of generative AI is not entirely new. Contract cheating, plagiarism, essay mills, and unauthorised collaboration have long been present in academic life, and concern about their newest form has already reached the wider higher education literature (Cotton et al., 2023). Generative AI intensifies these challenges: an LLM can produce content that does not overlap with any identifiable source even while outsourcing the intellectual work, and it can restructure, lengthen, shorten, or restyle an argument on request. Plagiarism checkers built to scan documents against external sources are therefore poorly suited to text produced by an LLM.

Several characteristics of English Literature make this an acute issue. Literary evaluation typically requires interpreting a text, comparing works, applying a critical theory, or situating a text in its context, and this is precisely the kind of prompt where LLMs can produce responses that sound generic yet plausible. Canonical texts also carry a substantial digital afterlife of editions, summaries, lecture notes, and study guides. An empirical study of an undergraduate literature course found that having students use ChatGPT to generate alternative endings to a canonical novel enriched the literary analysis process and fostered a more interactive classroom, while also giving researchers a way to examine effects on students' analytical skills, engagement, and perceptions of authenticity (Bellot et al., 2025). This suggests that the presence of generative AI does not simply threaten literary assessment; it can also be deliberately built into a task, which is a different problem from a student quietly using it to bypass one. The educational function of literary criticism, however, is not merely to produce a grammatically correct piece of writing. It requires attention to language, form, ambiguity, tone, structure, and the relationship between textual evidence and the claim being built on it, along with a defensible judgement. UNESCO's guidance on generative AI takes a human-centred approach, stressing agency, inclusion, equity, and educational use rather than treating AI purely as a technical concern (Miao & Holmes, 2023). The central assessment question therefore needs to shift from whether an essay was written by a human to whether an evaluation can demonstrate a student's ability to read, analyse, construct, revise, and justify a literary claim.

The Limits of Technological Countermeasures

One response to AI-assisted cheating is to make the question itself difficult for an LLM to process cleanly: prompt injections, unusual formatting, watermarked or distorted question images, and other techniques intended to disrupt a model's comprehension of the question. Prompt injection is a legitimate and well-documented security concern; research on LLM-based automatic grading systems has shown that instructions hidden inside a prompt

can manipulate a model's output in ways the designer never intended (Li et al., 2026). A widely reported case from July 2026 illustrates how this same logic can be repurposed as a detection method rather than an attack: a history instructor at a U.S. university hid an out-of-place instruction in white text within an exam prompt, so that any student who copied the prompt into a chatbot would unknowingly reproduce the planted phrase in their answer. Thirty-two of thirty-five students across two sections were caught this way (TODAY, 2026).

An approach that relies this heavily on confrontation is fragile. It depends on a technological gap between the testing environment and the model, and once either the model provider or the student becomes aware of the trick, it stops working. The hidden device used to catch a submission may also have nothing to do with the outcome the assessment is meant to measure; a literary analysis task should test whether a student understands the literature, not whether they can spot a planted trap. Cases such as the one above catch a cohort once, but they do not scale into a repeatable assessment strategy, and instructors who rely on them risk treating a single viral tactic as a substitute for redesigned assessment.

AI proctoring raises a related set of concerns. Remote surveillance may create an impression of control, but it can also raise problems of privacy, accessibility, and equity. Unusual movement, background noise, reading habits, or device settings may be flagged as suspicious without any malpractice having occurred. UNESCO has argued that AI technologies in education should be examined through a human-rights lens that foregrounds privacy and equity (Miao & Holmes, 2023). Technological measures may still have a role: controlled testing environments, secure browsers, and carefully designed question presentation can reduce opportunities for cheating. The difficulty arises when such measures are treated as the core of an integrity strategy, since they tend to have a short shelf life relative to their cost.

The Problem with AI Detection

An automated AI detector appears attractive because it claims to identify whether a text was produced by a language model. A low detection score, however, does not establish authorship; a detector only estimates whether a text resembles patterns typical of AI-generated writing, and it cannot monitor the process by which a specific student produced their submission. Independent testing of widely used detection tools has found substantial rates of both false positives and false negatives across a range of text types (Weber-Wulff et al., 2023), and separate theoretical and empirical work has shown that simple paraphrasing can defeat most current detectors, including those based on watermarking (Sadasivan et al., 2023).

The problem becomes more serious when detection errors are examined by writer background. Liang et al. (2023) tested several GPT detectors and found that non-native English writing was flagged as AI-generated far more often than native English writing, and that small changes to a prompt could shift a detector's verdict considerably. This raises a serious concern for English Literature classrooms, which often include non-native English speakers, since no student should have to prove their innocence against an automated probability score.

The failure of detection technology does not mean instructors should ignore clear contradictions. An abrupt change in writing style, fabricated references, an inability to explain an argument, or unfamiliarity with claims made in an assignment can be legitimate grounds for a conversation with a student. The line lies between evidence that warrants a conversation and a technology that purports, on its own, to prove academic misconduct.

Layer One: Controlled Demonstration of Literary Thinking

The first component of the proposed framework is controlled demonstration. Its purpose is to create situations in which literary thinking must be shown under conditions where large-scale outsourcing to AI is impractical.

The simplest version of this layer is supervised close reading. Rather than assigning a single lengthy essay on a novel, instructors can periodically ask students to analyse a brief, previously unseen passage from the set text during class. This asks students to identify a significant linguistic detail, explain its significance, connect it to a larger argument, and draw a tentative conclusion, all without external assistance.

Oral defence can serve a similar purpose. Once a research essay has been submitted, a short set of questions about a student's research process, thesis, or revision choices can quickly establish whether the submitted work is connected to what the student actually understands; a few minutes is usually sufficient. Faculty development work on AI-resistant assessment argues that redesigning tasks in this way is not something instructors can be expected to improvise on their own; it recommends that institutions support the shift with structured training, workload adjustments, recognition, and funding for faculty who take it on (Awadallah Alkouk & Khlaif, 2024).

The value of controlled demonstration lies in triangulation. No single exercise can prove authorship on its own; it is the combination of independent close reading, discussion, and the written essay that provides stronger evidence together than any one component could alone.

Layer Two: Making the Writing Process Visible

The second layer addresses a structural flaw in the traditional take-home essay: the final version reveals nothing about how it was produced. Process-visible assessment reverses this by treating the intermediate stages of writing as evidence in their own right.

Students can be asked to submit supporting material alongside a final essay, such as an initial thesis statement, annotated textual excerpts, a bibliography, a paragraph outline, an earlier draft, and a short reflection on the revisions made. Revision logs can supplement this evidence, though they should not be treated as a form of invasive monitoring.

Annotation is particularly useful for literature scholars, since an annotated passage reveals what a student noticed before an interpretation was fully formed. A research journal shows how a student evaluated secondary material, and a reflective comment can explain why one interpretation was kept over another.

This approach supports responsible AI use more effectively than an outright ban. Where institutions permit AI assistance, students can be asked to disclose whether they used AI for brainstorming, language correction, translation, or another purpose, and to reflect on what kind of use was appropriate. This distinguishes assistance from substitution. Both UNESCO and the Tertiary Education Quality and Standards Agency stress the importance of developing critical and ethical judgement in assessment design that anticipates generative AI rather than reacting to it after the fact (Miao & Holmes, 2023; Lodge et al., 2023).

The Dual-Layer Framework

The proposed framework combines the two layers rather than treating them as alternatives.

Layer	Primary purpose	Possible assessment practices
Layer I: Controlled Demonstration	Evidence of independent literary thinking	In-class close reading, supervised writing, discussion, short oral responses
Layer II: Process Visibility	Evidence of development and ownership	Drafts, annotations, research logs, revision notes, reflective commentary, oral defence

Findings and Discussion

This review of the literature leads to three conclusions. First, detection-based responses will always be reactive; as generative AI models and detection models evolve together, neither side gains a lasting technical advantage. Second, automated detection should not be used as the basis for disciplinary action, since it produces false results that fall disproportionately on some groups of students (Liang et al., 2023; Weber-Wulff et al., 2023). Third, assessment redesign offers a more sustainable response than either approach, because it ties academic integrity directly to learning outcomes rather than to detection accuracy. This is consistent with broader calls to redesign assessment around what generative AI makes newly difficult to fake, rather than attempting to exclude the technology altogether (Lye & Lim, 2024).

In English Literature specifically, the clearest evidence of learning is likely to be found in acts of interpretation: identifying textual elements, connecting ideas, engaging with prior criticism, building an argument, and answering follow-up questions about it. This framework also highlights a risk within the broader “AI-resistant” discourse in education: resistance should not translate into increased hostility toward students. The aim is to design assessments in which learning remains visible despite the availability of these tools, since the use of generative AI is unlikely to disappear from students' academic or professional lives.

Implications, Recommendations and Constraints

Departments of English Literature can adopt the proposed model without eliminating essays altogether. First, major assignments should include at least one process-visible element. Second, occasional supervised close-reading exercises should be built into the course to generate independent evidence of literary reasoning. Third, brief oral follow-ups can be used selectively for significant assignments. Fourth, departments should publish guidelines that distinguish permissible AI assistance from outsourcing. Fifth, AI detection tools should not be treated as proof of misconduct.

There are limits to implementation. Process assessment adds workload for faculty, particularly in large classes, and organising oral defences is not realistic for a cohort of several hundred students. Faculty development research on AI-resistant assessment likewise stresses that this kind of redesign only takes hold where institutions back it with training, workload adjustment, recognition, and funding, rather than leaving individual instructors to absorb the extra effort (Awadallah Alkouk & Khlaif, 2024). Digital revision histories may also be unavailable to

some students or may raise concerns about surveillance. Because of these constraints, the framework is intended as an adaptable set of practices rather than a compulsory template.

Conclusion

Generative AI has intensified a problem that predates ChatGPT: the finished essay is often treated as sufficient evidence of learning. In English Literature, which depends on interpreting language, constructing arguments, engaging with scholarship, and working through ambiguity, the essay has become a progressively less secure proxy for that learning.

The solution is unlikely to be a lasting technological trap or a flawless AI detector. Prompt injections can be bypassed once discovered, proctoring systems misclassify ordinary behaviour, and detection scores cannot establish authorship on their own (Sadasivan et al., 2023; Weber-Wulff et al., 2023). A more durable response reforms assessment so that literary reasoning becomes observable in the process, not only in the product.

The framework proposed here combines controlled demonstration with process transparency: it asks students to demonstrate independent interpretation at specific moments while also showing how a piece of academic writing developed over time. This will not fully prevent the misuse of generative AI. What it does is reframe the question that assessment is trying to answer. Instead of asking whether a machine can be detected behind a finished essay, the question becomes whether a student can demonstrate the human labour of interpreting, researching, arguing, and defending a claim. For English Literature, this is not only a defensive response to generative AI but also an opportunity to make the discipline's own values more visible in how it assesses students.

Acknowledgements:

The authors declare no external funding and no conflicts of interest for this study.

References

- Awadallah Alkhouk, W., & Khlaif, Z. N. (2024). AI-resistant assessments in higher education: Practical insights from faculty training workshops. *Frontiers in Education*, 9, Article 1499495.
<https://doi.org/10.3389/educ.2024.1499495>
- Bellot, A. R., Gutiérrez-Colón Plana, M., & Baran, K. A. (2025). Redefining literature education: The role of ChatGPT in undergraduate courses. *International Journal of Artificial Intelligence in Education*, 35(5), 3185–3201. <https://doi.org/10.1007/s40593-025-00497-3>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). 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>
- Li, H., Filippov, F., Lin, Y., He, P., Yang, K., Chu, Y., Cui, Y., Liu, H., & Tang, J. (2026). “Important!” You should give me full credits!: Exploring prompt injection attacks on LLM-based automatic grading systems. *arXiv*.
<https://arxiv.org/abs/2606.03090>
- Liang, W., Yuksekogonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), Article 100779. <https://doi.org/10.1016/j.patter.2023.100779>

- Lodge, J. M., Howard, S., Bearman, M., Dawson, P., & Associates. (2023). *Assessment reform for the age of artificial intelligence*. Tertiary Education Quality and Standards Agency.
<https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf>
- Lye, C. Y., & Lim, L. (2024). Generative artificial intelligence in tertiary education: Assessment redesign principles and considerations. *Education Sciences*, 14(6), Article 569. <https://doi.org/10.3390/educsci14060569>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.
<https://doi.org/10.54675/EWZM9535>
- Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv*. <https://arxiv.org/abs/2303.11156>
- TODAY. (2026, July 30). A professor created a trap in his midterm to catch students using AI. Nearly all made the same mistake. <https://www.today.com/parents/family/professor-catches-ai-cheating-hidden-word-rcna589421>
- 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*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>