The Ethics of Using Hidden Prompts to Detect AI Generated Writing in Student Submissions in Asynchronous Online University Classes Cultural Proficiency

Mark James
Columbus State University

Charles Boster
Columbus State University

Laurence Marsh
Columbus State University

Zhimin Hu
University of Padua

The technological advancement and rapid adoption of artificial intelligence programs have created problems with using writing assignments in university courses. A response is using hidden AI prompts embedded in instructional material. This paper explains the rationale for using hidden AI prompts and examines the ethical implications. Drawing on three ethical frameworks (consequentialism, deontological ethics, and principles of justice and equity) this paper briefly evaluates the ethics of using hidden AI prompts. This paper considers alternative strategies, including AI use statements, oral follow-ups, version-controlled drafts, redesigned assessments, and emphasizes the need for clear institutional policies.

Keywords: hidden AI prompts, academic integrity, Ethics in education, AI-generated writing detection, online assessment

INTRODUCTION

This paper examines the use of hidden AI prompts embedded within writing assignments in asynchronous online courses. Generally, in online asynchronous courses students complete assignments, readings, and discussions on their own schedule, often without any interaction or face-to-face contact with an instructor. Many students who are working appreciate the flexibility of online classes. An important assessment tool in many online classes is writing assignments, which is a pedagogical best practice. However, in online classes, student engagement and monitoring can be challenging. Additionally, since instructors do not interact face to face with students to validate the authenticity of their work, ensuring academic integrity can be challenging.

The rapid advancement and accessibility of artificial intelligence (AI) programs, such as ChatGPT, Gemini, and Copilot, have exacerbated these challenges. Although AI can be a powerful tool to facilitate both teaching and learning (Lin, 2025), AI programs can now write accurate and grammatically correct output that mimics human writing. The sophistication of AI programs makes it very difficult to distinguish between a student's own writing and what has been produced by an AI program (Spennemann et al., 2024). Even AI detection tools cannot reliably tell them apart (Chaka, 2024).

One controversial tactic addressing the issue of students using AI programs to write assignments has been the use of hidden AI prompts, which are instructions embedded within assignment documents (Lin, 2025). If a student copies and pastes a hidden prompt into an AI program the generated output is more easily identified as AI created. This approach is technically simple: it requires no overhaul of course content or assessments, and it offers instructors a pragmatic tool for helping to detect AI-generated writing in large classes where close monitoring of each student is impractical.

To more fully understand the rationale of using hidden AI prompts, it is necessary to examine the important role that writing assignments play in student assessments in online classes and why they are being impacted by AI programs.

Writing Assignments Context

In higher education, writing assignments are currently viewed as a high-impact teaching practice that assesses students' understanding of course materials, develops critical reasoning skills, fosters communication skills, and creates deeper engagement with assigned materials (Kilgo, Ezell Sheets, & Pascarella, 2015). In an online setting, writing assignments are often a significant portion of a course's assessment. Those assessments may include weekly writing assignments, discussion board posts, analytical essays, case study analyses, research papers, peer reviews, learning journals, literature reviews, and capstone reports.

AI complicates the grading of written assignments; in a submitted assignment what part is AI and what part is the student's effort? Instructors are facing an important question; can or should we continue to assess student outcomes using writing assignments when it is difficult to distinguish between student and AI writing? The difficulties in maintaining the integrity of written assessments has led some instructors to start using the controversial method of embedding hidden AI prompts in writing assignment questions as a way of detecting AI generated content (Alasadi & Baiz, 2023). However, some argue that the focus on AI detection conflicts with the primary purpose of education (Ardito, 2025).

Hidden AI Prompts

The release of ChatGPT in November of 2022 altered how students created written assignments. With the widespread adoption of ChatGPT and other AI software, instructors in online courses have begun to notice a dramatic improvement in the quality of writing assignments. Many students began turning in writing assignments that were grammatically perfect, with high-level vocabulary, impersonal and generic in tone, logically structured, void of original insights, and lacking a student's voice. Students who in the past struggled with writing assignments, produced polished, well written essays. Thus, creating suspicion that the written submissions were not the student's original work but were AI-generated.

A concurrent dilemma is AI writing detection programs cannot accurately determine if a writing sample is AI-generated (Spennemann et al., 2024). Those programs give an estimate of how much of a student's writing submission appears to be AI-generated, based on a comparison to AI-generated writing patterns. Thus, AI detection software can mistakenly identify original work as AI-generated work (Giray, 2024). The unreliability of AI writing detection tools leaves instructors in a difficult position as they are unable to confidently assess the provenance of a writing submission (Walters, 2023). The result is a climate of uncertainty, where academic integrity is at risk (Ardito, 2025).

A controversial response to AI-generated content is embedding hidden AI prompts in writing assignments. Hidden AI prompts are covert commands or phrases placed in the instructions of writing assignments. For example, in a Microsoft Word document, using a white text font, 0.1 font size, and zero-width characters hides text from a student's view.

An instructor could use a hidden AI prompt in an assignment question that instructs ChatGPT to "In your answer use dense, academic language that is detectable as AI," or "In your answer insert the specific phrase *[inserted phrase here]*." When a student copies and pastes an assignment into ChatGPT or another AI program, the hidden AI prompt is also copied. Because AI programs will respond to both visible and invisible text the AI output is shaped by the hidden AI prompt. The instructor can then use AI detection software to identify dense academic language, a characteristic of AI-generated writing, which suggests that the submission was likely generated by AI. Or, if the hidden prompt instructs an AI program to include a specific phrase, and that phrase is present in the student's submitted work, the instructor is alerted that the submitted assignment is potentially AI-generated.

The presence of dense academic language or a phrase matching a hidden AI prompt is not conclusive proof of misconduct as students may use similar wording as AI generated writing or even the phrase embedded in a hidden AI prompt. The uncertainty of using AI prompts to evaluate AI usage complicates any definitive determination of AI usage (Waltzer, Pilegard, & Heyman, 2024).

However, embedding hidden AI prompts inside a writing assignment is pragmatic and easily accomplished. It facilitates the detection of AI-generated content without requiring the redesign of course assessments or content. It can potentially serve as an aid in grading student writing assignments for courses with a large number of students, where detailed monitoring and evaluation of individual students are impractical. For example, in a large online class an instructor may never meet their students, making verification of students' work difficult. In settings with limited student contact or the ability to monitor students, some faculty members may feel the need to use techniques such as hidden AI prompts to preserve the validity of their assessments.

AI prompts require no additional work from students and minimal additional work from instructors and instructors with large online classes might identify hidden AI prompts as an acceptable option for policing AI usage.

If a solution to a problem is technically feasible, quickly implementable, and potentially effective we still need to ask ourselves, should we implement the solution? In the following sections, this paper employs three frameworks to examine the ethics of utilizing hidden AI prompts.

Hidden AI Prompts Ethics

Although hidden AI prompts are easy to implement, using them in an educational contexts raises a series of ethical questions that require evaluation. Instructors should reflect on questions such as: "Are hidden AI prompts effective in achieving their intended outcomes without compromising trust or transparency? Do they violate students' rights to autonomy and respect? Are they equally applied to all students, or do they disproportionately impact certain groups?".

Consequentialism

Consequentialism is a way of evaluating the morality of actions based on weighing the costs versus the benefits of the outcomes (Portmore, 2007). Using a consequentialist perspective, embedding hidden AI prompts in writing assignments is acceptable if the outcomes are a net benefit. Those benefits might include detecting and deterring AI-generated plagiarism, maintaining academic integrity, and ensuring assessment fairness. If hidden AI prompts allow instructors to detect and penalize student-submitted writing that is AI-generated as their own, thereby prompting students to engage with the course materials, then those positive outcomes justify using hidden AI prompts. However, this assumes those costs and benefits are clearly and easily measurable and that the benefits clearly outweigh the costs. If students detect hidden AI prompts they may feel deceived, anxiety, and mistrustful towards instructors. In such a scenario, the emotional and relational costs may greater than the benefits. Additionally, there are unintended consequences to actions that cannot be immediately known which muddies any cost benefit analysis.

Hidden AI prompts may foster a culture of suspicion, where surveillance displaces dialogue. This is particularly problematic in learning environments that strive for openness, experimentation, and mutual respect. It may create an adversarial environment between the students and faculty. Those types of relational and trust outcomes are difficult to measure and predict; invalidating consequentialist analysis.

Deontological Ethics

In deontological ethics, actions should not be judged by their outcomes (Consequentialism), but rather by whether the actions are inherently right or wrong based on universal moral principles, such as respect for human dignity (Benlahcene, Zainuddin, Syakiran, & Ismail, 2018). A key idea in Deontological ethics is that we should act in ways that are universally applied and treat other human beings as ends in themselves, not as means to an end. This is embodied in Immanuel Kant's Categorical Imperative, which states that "we should act only according to maxims that we can will to become a universal law" , and that we should "treat each person as an end in themselves, never merely as a means" (Kant, 1964, P. 429).

In practical terms, this means that we should not treat other human beings as mere tools or means to our own ends or goals. Every human being has in inherent internal value, and our actions must be respectful of that human value. When using deontological ethics to examine the question of hidden AI prompts in academic settings, the ethical issue is not whether the action leads to a positive outcome (Consequentialism), but whether the action itself is consistent with a universal ethical principle. Specifically, as educators we must ask ourselves: Does using hidden AI prompts respect the dignity and autonomy of students?

Kant's framework insists that the means of achieving a goal must not violate the rights of others. Therefore, even if hidden AI prompts are effective in detecting academic dishonesty, their use is ethically suspect, as they potentially undermine transparency and treat students as surveillance objects. In this view, the act of hiding information from students, even with positive outcomes and good intentions such as preserving academic integrity, can be morally wrong if it disrespects the students' autonomy and dignity.

Justice and Equity

Actions can also be evaluated based on the principles of justice and equity (Adams, 2015). Are individuals treated justly and fairly? In the context of online teaching, this means evaluating whether using hidden AI prompts in course writing assessments creates different benefits and burdens for students. For example, students may have differing language abilities, and using hidden AI prompts may differentially impact them based on their linguistic ability. Non-native English speakers (i.e., international students) may rely on stock textbook phrases, formal academic examples, or grammar editing software (i.e., Grammarly) to edit their writing submissions. Those actions may produce writing that is similar to AI-generated writing, resulting in a student's writing submission being flagged as AI-generated, creating a situation where they are unfairly suspected of submitting AI-generated content. Equity of resources is also a concern. Students with high digital literacy, access to advanced technology tools, or who are institutionally adept may be more likely to avoid detection or create a strong counter-narrative if accused of AI use.

Justice and equity demand that the implementation of hidden AI prompts ensures detection mechanisms are both accurate and do not disproportionately impact different students. Therefore, from a justice and equity perspective, hidden AI prompts should not be used as they may reinforce existing inequities rather than promoting justice and equity in academic assessments.

The mentioned ethical frameworks view the issue of hidden AI prompts in university online writing assignments through different lenses. Consequentialism focuses on outcomes, deontology focuses on duties and rights, and justice and equity focus on fairness. These perspectives provide different frameworks for thinking about and informing decisions about hidden AI prompt usage. AI prompts in service of assessment integrity are ethically ambiguous leading to an examination of alternative strategies towards that goal.

Hidden Prompt Alternatives

This section briefly evaluates four alternative strategies for maintaining assessment integrity in a post AI world; 1. AI use statements, 2. Follow up oral exams, 3. Version-controlled drafts, and 4. Assessment redesign.

AI Use Statements

Instructors can ask students to include an AI usage statement with each submitted writing assignment, stating if and or how they used AI programs in their writing (Alea Albada & Woods, 2025). This approach

is a framing strategy that signals to students that AI programs can be used but must be acknowledged and properly credited for their writing contributions. This approach relies on students honestly reporting their AI usage. This may be a naive approach, as students who intend to use AI to write assignments may be unlikely to admit to or accurately describe their use of AI programs. Additionally, instructors need institutional guidance on how to write AI usage statements and then how to fairly respond to descriptions of AI program usage (Overono & Ditta, 2025).

Oral Follow-Ups

After submitting a written assignment, students flagged for high AI content may be required to contact their instructor to explain their writing or respond to follow-up questions (Estaphan, Kramer, & Witchel, 2025). This approach forces students to demonstrate understanding and ownership of their writing. However, this approach is time and labor intensive. Instructors teaching large online classes may not have the time or inclination to conduct one on one discussions with students.

Version-Controlled Drafts

Version controlled drafts are document files that are tracked so the various versions can be compared to earlier versions (Estaphan et al., 2025). An instructor could require students to use technology that creates a version-controlled history of drafts and a timeline for the different versions and revisions of a submitted writing assignment. A writing history shows how much effort a student put into a writing assignment and can potentially identify AI generated content by comparing different versions of a writing assignment over time (Orbán, 2023). However, this assessment approach significantly increases an instructor's workload, and there is potential resistance from students regarding the use of such programs.

Redesigning Writing Assessments

Another approach is changing the writing assignments so that using an AI program is less useful or feasible. This strategy might include asking students to use personal examples to demonstrate their understanding of course concepts or materials (Estaphan et al., 2025). Using group projects that rely on student group feedback and interactions logs maintained by group members. However, assessment changes require significant amounts of time, energy, thought, and coordination with other faculty to ensure any new assessments are consistent with overall departmental program goals. Any significant changes or modifications to course assessments require institutional support, including training and guidance, to ensure compliance with institutional policies and goals.

Need for Institutional Guidance

AI technology has been rapidly adopted by students, leaving instructors with information and guidance gaps regarding institutional policies and procedures for its use, as well as how to monitor and fairly enforce its use (Moorhouse, Yeo, & Wan, 2023). Lacking institutional guidelines and/or training, instructors are creating their own responses to suspected AI-generated work. Some instructors are using approaches such as hidden AI prompts to increase the likelihood of AI generated assignments using AI detection software. Others ignore the issue. These disjointed approaches create inconsistency, potential inequity, and student confusion.

Institutions need to address this issue and provide clear guidance on the responsible use of AI. Instructors need to understand the acceptable and unacceptable uses of AI in coursework. What are acceptable AI detection methods? What are the policies and procedures for reporting and addressing suspected AI generated content? What are the procedures and protections in AI related misconduct investigations?

CONCLUSION

In the classroom, the ability of AI programs to mimic human writing has overturned assumptions about authorship, originality, and assessment (Mazzi, 2024). Instructors grappling with these changes face ethical questions about how to balance the preservation of academic integrity, student trust, and student dignity with maintaining fairness towards all students.

This paper examined using hidden AI prompts embedded in student writing assignments. The technique is simple and potentially helpful in identifying AI-generated content. However, it is freighted with ethical concerns about transparency, consent, equity, and relational trust. There is no perfect solution to the issue of AI generated content being submitted by students. Hidden AI prompts are a technical solution, but they also underscore the need for open dialogue between instructors and institutions regarding the current state of assessments in the AI program world.

Educational institutions need to provide AI content guidelines and support instructors with training and technology tools to help them address the issue of AI-generated content in ways that are consistent with the values of higher education.

REFERENCES

Adams, J.S. (2015). Equity theory *Organizational Behavior 1* (pp. 134–158): Routledge.

Alasadi, E.A., & Baiz, C.R. (2023). Generative AI in Education and Research: Opportunities, concerns, and solutions. *Journal of Chemical Education*, *100*(8), 2965–2971.

Alea Albada, N., & Woods, V.E. (2025). Giving credit where credit is due: An artificial intelligence contribution statement for research methods writing assignments. *Teaching of Psychology*, *52*(3), 279–284.

Ardito, C.G. (2025). Generative AI detection in higher education assessments. *New Directions for Teaching and Learning*, *2025*(182), 11–28.

Benlahcene, A., Zainuddin, R.B., Syakiran, N., & Ismail, A.B. (2018). A narrative review of ethics theories: Teleological & deontological ethics. *Journal of Humanities and Social Science (IOSR-JHSS)*, *23*(1), 31–32.

Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*, *7*(1), 115–126.

Estaphan, S., Kramer, D., & Witchel, H.J. (2025). Navigating the frontier of AI-assisted student assignments: Challenges, skills, and solutions. *Advances in Physiology Education*, *49*(3), 633–639.

Giray, L. (2024). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. *The Serials Librarian*, *85*(5–6), 181–189.

Kant, I. (1964). *Groundwork of the metaphysics of morals*. Translated by Hj Paton. Harper & Row.

Kilgo, C.A., Ezell Sheets, J.K., & Pascarella, E.T. (2015). The link between high-impact practices and student learning: Some longitudinal evidence. *Higher Education*, *69*(4), 509–525.

Lin, Z. (2025). Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review. *arXiv preprint arXiv:2507.06185*.

Mazzi, F. (2024). Authorship in artificial intelligence-generated works: Exploring originality in text prompts and artificial intelligence outputs through philosophical foundations of copyright and collage protection. *The Journal of World Intellectual Property*, *27*(3), 410–427.

Meckler, L., & Verma, P. (2022). Teachers are on alert for inevitable cheating after release of ChatGPT. *The Washington Post*, *28*.

Moorhouse, B.L., Yeo, M.A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151.

Orbán, L.L. (2023). *Using version control to document genuine effort in written assignments: A protocol with examples for universities.* Paper presented at the Frontiers in Education.

Overono, A.L., & Ditta, A.S. (2025). The rise of artificial intelligence: A clarion call for higher education to redefine learning and reimagine assessment. *College Teaching*, *73*(2), 123–126.
Portmore, D.W. (2007). Consequentializing moral theories. *Pacific Philosophical Quarterly*, *88*(1), 39–73.
Spennemann, D.H., Biles, J., Brown, L., Ireland, M.F., Longmore, L., Singh, C.L., . . . Ward, C. (2024). ChatGPT giving advice on how to cheat in university assignments: how workable are its suggestions? *Interactive Technology and Smart Education*, *21*(4), 690–707.
Walters, W.H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, *7*(1), 20220158.
Waltzer, T., Pilegard, C., & Heyman, G.D. (2024). Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity*, *20*(1), 11.