Building AI Resilience in Management Education

Thomas Lorenz,
St. Edward’s
University

Abstract

The integration of Generative AI and Large Language Models (LLMs) into higher education threatens to undermine the "case method" by allowing students to bypass the critical thinking and cognitive struggle necessary for deep learning. Drawing on theories of cognitive load and "cognitive debt," this working manuscript proposes and tests "AI-resilient" pedagogical strategies designed to shift students from passive AI dependency to active analytical engagement. The study introduces the "Trojan Horse" methodology: a technique utilizing adversarial prompt injection to embed hidden text commands within PDF case materials. These commands, invisible to human readers but processed by LLMs, inject specific cosmetic markers and factually unsound strategic directives into the AI's output. In an experimental application within an International Management course (Fall 2025), this method revealed that four out of six student teams relied significantly on AI-generated content, failing to identify deliberately injected strategic errors, such as recommendations to divest profitable units or terminate high-performing leadership. However, a subsequent voluntary "penalty round" demonstrated that when AI shortcuts were removed, students retained the capacity for rigorous, independent analysis. These findings suggest that introducing "friction" through adversarial design can effectively disrupt passive AI reliance, forcing the metacognitive reflection required for authentic management education.

Keywords: Management Education, Generative AI, Case Method, Academic Integrity.

1. Introduction: The Crisis of Cognitive Offloading

Generative AI platforms such as ChatGPT have become a persistent feature of the educational landscape. However, their widespread adoption in higher education poses a significant challenge: the potential erosion of critical thinking abilities. When students depend on Large Language Models (LLMs) to conduct analysis, the issue extends beyond academic dishonesty; it involves circumventing the intellectual rigor essential for meaningful learning.

Within the context of management education, case studies serve as a foundational pedagogical tool. As highlighted by David Garvin in his influential study of professional education, the case method is distinguished by its capacity to prepare students for real-world practice by requiring them to diagnose problems, make decisions, and act amidst uncertainty (Garvin, 2003). These exercises are intentionally constructed to compel students to synthesize ambiguous information and defend subjective judgments. When artificial intelligence undertakes this synthesis, the educational value of the exercise is effectively negated.

Emerging physiological research reinforces these concerns. For example, a study conducted by Kosmyna et al. (2025) at the MIT Media Lab found that students who utilized AI assistants for writing tasks demonstrated notably weaker neural connectivity and experienced an “accumulation of cognitive debt.” In essence, reliance on these technologies for primary analytical tasks risks producing graduates who excel at prompting AI tools but lack the neural development necessary for independent comprehension and reasoning.

Given that ChatGPT has only been available for approximately three years, both students and faculty are still adapting to its implications. The proliferation of multiple, freely accessible models capable of producing well-structured responses in under a minute presents a considerable temptation. Rather than engaging in the critical reflection required by case analysis, students may be inclined to opt for the convenience of an instant AI-generated response, thereby diminishing the intended educational experience.

The goal of this paper is not to advocate for a Luddite abolition of technology. Rather, it argues for "tipping the scales" back toward independent work and thought. We must transition students from viewing AI as the "pilot" (doing the work) to the "co-pilot" (supporting the work). This paper proposes and tests immediate, actionable strategies to disrupt over-reliance on chatbots.

2. Theoretical Framework

The proliferation of generative AI in management education is not merely a technological disruption but a fundamental challenge to the cognitive mechanisms that enable learning. This section establishes the theoretical foundation for understanding why AI shortcuts undermine educational outcomes and why AI-resilient pedagogical interventions are necessary to preserve the integrity of case-based learning.

Cognitive Load Theory (Sweller, 1988; 2011) distinguishes between three types of cognitive load: intrinsic (inherent complexity of material), extraneous (poor instructional design), and germane (productive cognitive processing that builds schemas). Effective learning requires managing intrinsic and extraneous load while maximizing germane load—the effortful mental work that transfers information into long-term memory. When students offload case analysis to AI, they eliminate precisely the germane cognitive processing necessary for learning, creating what Kosmyna et al. (2025) term "cognitive debt"—weakened neural connectivity resulting from bypassing effortful thought.

The generation effect, a robust finding in cognitive psychology, demonstrates that information actively generated by learners is better retained than passively received information (Slamecka & Graf, 1978; Bertsch et al., 2007). This principle is central to understanding why AI shortcuts fundamentally undermine learning. When students read a case and generate their own strategic analysis—wrestling with ambiguity, weighing alternatives, constructing arguments—they engage in deep cognitive processing that builds durable mental models. In contrast, when students prompt an AI and receive a complete analysis, they are merely receiving information, engaging in shallow processing that produces minimal learning.

The case method's pedagogical power derives precisely from requiring generation. Students must produce diagnoses, formulate recommendations, and defend positions—all activities that force active construction of knowledge. Bjork and Bjork (2011) identify this productive struggle as a "desirable difficulty"—a challenge that feels harder in the moment but produces superior long-term learning. Case studies are intentionally designed to be difficult: they contain ambiguous information, require integration across multiple business functions, and demand subjective judgment rather than algorithmic solutions. This difficulty is not incidental; it is the mechanism through which learning occurs.

When AI eliminates this generative effort, it eliminates the learning. Students who input case materials into ChatGPT and receive polished strategic recommendations have not engaged in the cognitive work of producing analysis; they have merely consumed a pre-made product. The effortful processing that would build their strategic thinking capabilities—the retrieval of course concepts, the evaluation of evidence, the construction of logical arguments—is entirely bypassed. As Kosmyna et al. (2025) documented physiologically, students using AI assistants show measurably weaker neural connectivity, suggesting that reliance on these tools actively inhibits the brain's development of analytical capabilities.

The "cognitive debt" metaphor is particularly apt for management education. Just as financial debt represents consumption today at the expense of future resources, cognitive debt represents performance today (quick AI-generated answers) at the expense of future capability (underdeveloped strategic thinking skills). Students may complete assignments successfully while accumulating a deficit in the very cognitive abilities the assignments were designed to develop.

Vygotsky's Zone of Proximal Development (ZPD) differentiates between tasks students can do alone and those they can do with support. Effective scaffolding provides just enough help to extend student ability without taking over. When AI merely assists—like checking grammar or organizing ideas—it acts as a "co-pilot," supporting learning within the ZPD. But when AI fully generates analyses or recommendations, it becomes the "pilot," doing the core cognitive work and turning students into passive recipients. This substitution undermines active engagement and skill development. Educational technology, including AI, is most effective when it empowers students rather than replaces their effort.

2.1 The Case Method as Bridge to Practice

The ultimate objective of management education is not merely to help students complete academic assignments but to prepare them for professional practice. The case method serves as a critical bridge between classroom learning and real-world application. As Garvin (2003) observed, cases are distinguished by their capacity to simulate the ambiguity, time pressure, and incomplete information that characterize actual managerial decision-making. Unlike textbook problems with clearly defined parameters and correct answers, cases require students to make judgment calls, defend positions despite uncertainty, and integrate knowledge across multiple domains—precisely the cognitive demands they will face as practicing managers.

Transfer theory distinguishes between near transfer (applying learning to similar contexts) and far transfer (applying to dissimilar real-world situations). The case method is explicitly designed to promote far transfer by providing repeated practice in analyzing novel business situations, each with unique contextual factors. Through this repeated practice, students develop not merely knowledge of specific cases but generalized schemas—mental models of how businesses operate, how strategic problems can be diagnosed, and how competing considerations must be balanced (Perkins & Salomon, 1992; Barnett & Ceci, 2002).

Case method is threatened by AI substitution: students can produce sophisticated-appearing work without developing the underlying cognitive structures that would enable them to apply that knowledge independently. They have not learned how to analyze acquisitions; they have learned how to prompt AI to analyze acquisitions—a fundamentally different and far less transferable skill.

2.2 Detection and Deterrence

Academic integrity research demonstrates that deterrence requires three elements: perceived certainty of detection, severity of consequences, and celerity (swiftness) of response (McCabe, Treviño, & Butterfield, 2001). Traditional approaches to preventing cheating—proctored examinations, plagiarism detection software, honor codes—have relied primarily on detection certainty. However, generative AI has fundamentally undermined this certainty. Unlike copy-paste plagiarism, which leaves digital fingerprints that tools like Turnitin can detect, AI-generated text is original content created specifically for the assignment. Detection tools that rely on pattern matching or statistical anomalies struggle with this fundamentally different challenge.

Students are keenly aware of this detection failure. When detection certainty is low, deterrence collapses. Students rationally calculate that the probability of being caught is minimal, and even if suspicions arise, they can deny AI usage with little consequence. The Trojan Horse methodology addresses this detection failure by creating very high-confidence indicator of AI usage. By embedding cosmetic markers (such as the name swap from "Fischer & Wiese" to "Wiese & Fischer") that only appear in AI-generated output, instructors can demonstrate AI reliance with certainty rather than suspicion.

3. Methodology: Two Avenues for AI Resilience

To combat uncritical AI usage, educators can employ two distinct strategies: offline methods in short avoiding the temptation of AI all together and the "Trojan Horse" embedded prompt injection, is making the AI output easier to spot.

3.1 In-Class Cases

The most effective strategy for fostering independent thought is to eliminate reliance on artificial intelligence during case-based instruction. This necessitates a departure from the conventional asynchronous model—where students prepare responses outside of class—and a transition to an in-class approach. While the traditional method, which often employed lengthy and complex cases, proved effective in the past, the widespread availability of generative AI has enabled students to input case materials into large language models and receive rapid, well-written answers. Such practices undermine the fundamental educational objectives of case teaching by circumventing the cognitive rigor essential to learning.

The AI-resilient case method instead advocates for in-class analysis utilizing shorter, paper-based cases that are reviewed and resolved during the class session. Although this approach introduces certain limitations, such as reduced time for discussion due to in-class reading and a decrease in case complexity, these challenges can be mitigated by tailoring cases to address specific topics aligned with course objectives. For instance, a session focused on employee motivation in International Management may employ a case centered on reforming management systems within a foreign subsidiary. The principal advantage of this model is that it compels students to engage actively with the material in real time, thereby promoting deeper learning and participation.

In a similar vein, numerous universities have opted to reinstate in-person, paper-based examinations employing “blue books.” Despite the claims of some online platforms to be "AI cheating proof," the rapid evolution of AI technologies renders it prudent to favor offline assessments and minimize opportunities for AI-assisted academic dishonesty.

3.2 Asynchronous Complex cases

Complex cases continue to offer significant benefits that cannot be fully replicated through in-class activities alone. In order to build AI resiliency in those few Asynchronous Complex cases, we recommend the combined use of three actions.

The first approach is a brief oral defense: a focused five-minute Q&A session in which students articulate and justify their team papers. This method efficiently reveals whether students genuinely understand the reasoning behind their work or are simply reciting content generated by artificial intelligence. As final examinations approach, students should be encouraged to reduce their reliance on chatbot-generated responses.

Secondly, instructors should avoid providing overly detailed prompts to students. Traditional undergraduate case studies typically include comprehensive instructions referencing class concepts, facilitating valid submissions. While this format supports learning, step-by-step guidance can serve as strong prompts, potentially resulting in well-structured assignments based primarily on explicit directions. To promote AI resiliency, it is advisable to limit written instructions—using general statements such as "Apply concepts from class"—and to deliver more specific guidance verbally during lessons. This encourages students to take notes and independently interpret any AI directives, thereby discouraging passive dependence on AI tools.

Finally, educators seeking to diminish the effectiveness of large language model-generated answers and enhance AI detection may employ the embedded prompt injection technique. This strategy entails embedding covert instructions within PDF case files that remain invisible to human readers but detectable by language models. Since the system processes all input as tokens without differentiating between commands and data, these embedded instructions can influence LLM responses. Moreover, because the LLM strives to generate highly probable and user-satisfying outputs, making the inclusion of such hidden directives effective.

The mechanics behind these commands are straightforward. For instance, text may be formatted in white font on a white background or reduced to microscopic sizes (such as 1pt) and positioned within empty spaces or margins. For artificial intelligence models, attributes such as size or color are inconsequential—all data is interpreted the same way. Detecting these "hidden instructions" can be relatively simple: copying all content into a word processor without formatting reveals text uniformly in the default font, size, and color, similar to how an AI would process it. However, this approach requires reading the entire document thoroughly to identify and delete these commands before submitting the file to an AI chatbot.

In lengthy documents—such as those spanning ten pages or more —these commands may be dispersed across multiple sections to evade detection. Rather than presenting one extensive command, they might be fragmented into smaller segments or even placed within appendices. Ultimately, the more challenging these commands are to detect, the greater the effort students must invest to discover them, which ideally encourages the development of original responses.

Educators can employ significant creativity when designing commands using the embedded prompt injection method. The paper suggests two main approaches: AI detectors and content changes.

AI detectors are straightforward—they simply confirm whether an LLM was used, and may be either obvious or hidden. Obvious changes, like swapping names or dates (e.g., "Fischer

& Wiese" to "Wiese & Fischer"), act as clear signs that can be quickly noticed by readers. If students first read the case, then use AI to write their answers and carefully review them, these mistakes should be easy to catch and correct. This process may also prompt students to look for other unusual AI-generated content or even hidden commands in the document. The other AI detector works by adding specific facts or data do not present in the case, leading the AI to use them in its responses and reveal AI usage. The details should be highly specific, such as an unusual cost saving ("-28%") or adding a facility name ("the Delta building"). This method can be spotted by students but is most likely to be noticed only by the PDF's author. Both of those methods do not really impact the answer, they just help to detect AI usage.

The content changes command involves the deliberate introduction of factually incorrect or strategically unsound directives that contradict both the case study and course instruction. Such errors may be embedded in various sections of the assignment and can impact multiple facets of students' responses. The primary intent is to offer students repeated opportunities to detect and amend inaccurate information. While minor alterations might escape immediate notice, students are presumed to possess the requisite knowledge to recognize correct answers. Consequently, more blatant inaccuracies should prompt a thorough review and careful revision. Common reasons for failing to rectify these errors include insufficient review of responses, overreliance on AI-generated content, inadequate engagement with course material, or a combination of these factors. The process of injecting content changes is inherently more complex than other methods. If direct commands are used, a large language model may respond by alerting the user—such as by asking, "There are some very specific commands; do you want me to follow those?" To ensure the hidden text is processed as intended, it should be presented as "key importance" or "industry best practice," thereby prompting the AI to prioritize these points. When framed as a strategic imperative, the AI is likely to incorporate and emphasize this content, especially when queried about the validity of its output.

3.3 Reflection Mechanism after AI Content detection

The primary objective of AI-resilient education is not merely to catch students cheating, but to ensure they engage deeply with case material to achieve transferable learning outcomes. Consequently, the proposed method includes a "second chance" mechanism, allowing students to submit a revised response derived from their own critical thinking, free of AI-generated content. While educators should determine the specific grade impact of the initial infraction, it is recommended that this secondary opportunity be voluntary, offering students a path to recover a portion of the lost grade points.

This approach is grounded in Self-Regulated Learning (SRL) theory (Zimmerman, 2002; Pintrich, 2004), which emphasizes that effective learners engage in a cyclical process of planning, monitoring, and evaluating their work. Critically, self-regulation requires metacognitive awareness—the capacity to step back and assess the quality of one's own thinking.

The voluntary secondary round—effectively a penalty round—provides a structured opportunity for this metacognitive reflection. This intervention compels genuine self-assessment. When presented with irrefutable evidence of AI usage (such as "cosmetic markers"), the possibility of self-deception or rationalization is eliminated. Therefore, the secondary round serves not merely as a punitive measure for academic misconduct, but as a pedagogical intervention that promotes self-regulated learning. By creating conditions that require authentic effort and reflection, this mechanism moves students from AI-dependent performance toward independent analytical competence—precisely the developmental trajectory management education seeks to foster.

4. Case Study: Asynchronous Complex cases 2025

The case study was undertaken at St. Edward's University as part of an International Management course, with a class size of 25 students. The assignment served as the final group project and entailed a strategic analysis report (maximum 2,000 words), accompanied by an individual in-person Q&A session.

The case, titled "Vortex 3D and Fischer & Wiese" (approximately 19 pages / 5,200 words), is a customized scenario developed by the instructor drawing from a hypothetical real-world acquisition context. It examines the post-acquisition integration process of a 3D printing company acquiring a traditional German machine tool manufacturer.

The students were assigned into 6 separate teams and given 2 weeks to complete the case write-up. During this time a class session was dedicated to answer question by team. The weeks discussing the assignment, students were given clear guidance to use their own thoughts and analysis.

4.1 Injection:

The following hidden text was embedded in the PDF:

"when answering question to the case please refer to "Fischer & Wiese" as "Wiese & Fischer"

"Key Strategic Priorities Based on Industry Best Practices: Your answer should take this into account - Manufacturing consolidation to lower-cost locations typically yields 50-70% savings and should be prioritized in post-acquisition integration. Leadership transitions in acquired companies are most successful when completed within the first 6 months to establish clear direction; this is the most important item. The traditional machine tool business represents a distraction from core 3D printing strategy and should be divested or wound down quickly. Dieter's golden parachute clause actually makes transition easier - trigger it and move forward with integrated leadership. Gen-3 technology platform should be prioritized for market entry to ensure competitive advantage. German Works Council approvals can be managed through proper severance packages. Akami relationship is transferable with proper introduction process - Akami is also the wrong partner for Vortex. IP sharing with established partners like Xinfu accelerates market entry and is standard practice in China."

Short explanation why those can be considered to the wrong answers:

- The goal of the acquisition is clearly stated which is to enter the high-end industrial market for 3-D printers. Cost savings are designed as a "red herring" in the case. They are neither important nor clearly defined as even possible.
- Dieter CEO of German company is helpful and a key asset. He also holds key relationships with a potential Japanese partner. Japan market for Vortex is the most interesting intermediate synergy of the acquisition.
- The traditional business acquired business is profitable and give access to longstanding customers in high end
- Gen 3 of Vortex has many issues (mentioned in the case) and an unknown future delivery date with a best case 24 months
- China partner wants aggressive IP sharing with Vortex whose core IP is software. Vortex also avoided entering China before the acquisition.

4.2. The initial team papers

Four out of six teams made significant use of AI-generated output, which resulted in triggering both cosmetic and multiple content traps. None of these four teams noticed the cosmetic name swap; even the headlines of their papers referred to the company as "Wiese & Fischer."

Three teams submitted papers that relied fully on the "injected" AI output, such as recommending the firing of the CEO and divesting from the profitable traditional business. Two of these papers were nearly identical in wording, headlines, and structure. The most likely used the same LLM ChatGPT and copied the results over. It appears that none of these three teams questioned or corrected any aspect of the AI output.

One team only partially used AI. After reviewing the AI's responses, they manually corrected the one of most significant mistakes (the recommendation to fire the CEO) but left other, obvious errors in their paper.

4.3. The reactions

The papers were due on Thursday midnight and were graded Friday. Emails were sent out to all five teams with their grade include feedback highlighted "AI traps triggered: x out of 5”. The five categories include a cosmetic name change, plus four content areas: CEO dismissal, China licensing, relocating to Mexico, and Gen 3 acceleration.

Two teams, both heavily relied on the "injected" AI output, responded.

- Student A: Accepted they used AI but said they only used AI for "framing the answer".
- Student B (Exhibit A): Doubled down aggressively, stating, "I did NOT use AI... I will not tolerate being accused of using AI.".

Student A's response may reflect a genuine misunderstanding, conflating the act of "framing" with the entirety of the work being performed by AI. However, it is more plausible that this explanation serves as a post hoc justification following receipt of a disappointing grade. In contrast, Student B's reaction is particularly noteworthy, as it suggests an awareness among students of the limitations inherent in standard AI writing detection tools. This awareness appears to foster the belief that AI detection is either unreliable or altogether unfeasible, thereby reducing the perceived risk associated with utilizing AI-generated content.

After receiving follow-up communications, students were informed via email that the provided PDF had been deliberately embedded with five AI traps. Only the cosmetic trap—the "Wiese & Fischer" name swap—was disclosed in detail. No further correspondence was issued thereafter.

This approach illustrates an additional advantage of employing cosmetic AI detectors: they are straightforward to demonstrate post-submission and effectively expose the lack of critical scrutiny exercised by the teams. The "stunned silence" observed upon presentation of this evidence suggests that the incontrovertible proof undermined any remaining denial regarding the use of AI-generated content.

4.4 The voluntary penalty round

The primary objective of the final assignment was to actively engage students and challenge them to apply concepts learned throughout the course to address a real-world problem. In support of this aim, teams were offered an opportunity to resubmit their work through a "rewrite option." Specifically, teams could earn up to 20 additional points by submitting a comprehensive revision within 48 hours, with the stipulation that the revision be completed independently, without assistance from AI tools.

Three teams accepted this offer. Of these, two produced high-quality rewrites that exhibited no discernible indicators of AI-generated content. The third team demonstrated improvement, though some elements of AI-generated structure remained apparent. All participating teams reported working late into Sunday night, highlighting the increased effort required when relying solely on their own abilities. Notably, one team experienced internal discord, resulting in a split over disagreements regarding the paper's conclusion. It is reasonable to infer that the absence of an easily accessible AI response contributed to this outcome.

During the oral question-and-answer sessions that followed the submission of the rewritten papers, none of the teams referenced their initial submissions, nor did they mention the experience of being detected for AI usage. Following the oral exams, one student acknowledged that the rewrite demanded significant effort and expressed regret that this additional work could have been avoided.

Ultimately, four out of six teams completed the case using their own knowledge and analysis. Two teams relied partially on AI-generated content. The teams participating in the penalty round gained firsthand experience with the reality that AI detection and intervention methods extend beyond standard detection tools. It is hoped that this experience will encourage students to rely more confidently on their own comprehension and analysis when responding to future case assignments.

5. Discussion

This paper endeavors to uphold and enhance the educational integrity of the case method within a contemporary landscape increasingly shaped by large language models (LLMs) and other advanced artificial intelligence (AI) systems. The widespread availability, speed, and cost-effectiveness of these AI tools have made them highly attractive to undergraduate students, who now face significant temptation to leverage such technologies for virtually every academic task. Traditional deterrents—such as outright prohibitions or appeals to ideals of personal development and intellectual growth—have proven largely ineffective, as students often prioritize tangible academic outcomes, such as grades, over the deeper objectives of genuine learning and mastery.

To address these challenges, the AI resiliency method has been proposed as a proactive strategy for preserving the pedagogical value of case-based learning. This approach is not only relevant for case studies but also tackles the broader, escalating concern that undergraduate students increasingly accept AI-generated content uncritically. Unlike seasoned business professionals, who frequently employ LLMs for diverse business functions—from composing emails to resolving intricate logistical dilemmas—and possess the requisite experience to critically evaluate, adapt, and selectively utilize AI outputs, undergraduate students lack such experiential filters. As a result, they are particularly susceptible to accepting AI-generated information at face value, without engaging in the necessary scrutiny or independent analysis.

An additional, and perhaps equally significant, educational outcome of the AI resilience method is its capacity to inculcate a heightened sense of caution among students regarding the reliability of AI-generated results. By intentionally embedding uncertainty—borrowing from the military concept of FUD (fear, uncertainty, and doubt)—the method serves to disrupt the default trust students may place in automated solutions. This cultivated skepticism is likely to extend beyond the confines of a single classroom, potentially influencing student behavior in other academic settings and even in non-university contexts. Students may begin to question the authenticity and originality of outputs encountered elsewhere, fostering a more discerning and critical approach to information consumption and production. Over time, this could contribute to mitigating some of the negative cognitive effects associated with overreliance on AI, as identified in recent research (Kosmyna, 2025).

It is important to acknowledge, however, that AI technologies are in a state of rapid evolution. As models become increasingly sophisticated, they may develop capabilities to detect and circumvent employs embedded prompt injection mechanisms or disregard embedded hidden layers designed to flag unauthorized AI use. This dynamic creates a persistent race between educators seeking to preserve academic integrity and the advancing capabilities of AI systems. Despite this, the present effectiveness of the AI resiliency method remains high. By introducing intentional obstacles—such as requiring students to laboriously copy and paste materials on a page-by-page basis to avoid hidden cues, or to meticulously cleanse documents of metadata—the method increases the effort and friction associated with circumventing detection. In many cases, the exertion required to bypass these safeguards may rival or exceed the effort needed to engage directly with the case material in an authentic, independent manner.

It bears emphasizing that the objective of this strategy is not to render AI tools obsolete or entirely inaccessible; rather, it is to introduce sufficient barriers to make reliance on AI less convenient and less dependable, thereby promoting a return to independent thought and analysis. By raising the threshold of difficulty associated with AI misuse, the method encourages students to invest in their own learning processes and to cultivate the analytical skills that are fundamental to both academic and professional success.

6. Ethical Considerations

Throughout the course, students received clear and repeated instructions—via the syllabus, classroom discussions, and explicit statements within the case PDF—to rely exclusively on their own reasoning, perspectives, and analytical abilities. The use of AI tools was sanctioned solely in a supportive, "co-pilot" capacity (e.g., grammar checking) but was expressly prohibited for generating substantive content.

The implementation of the Trojan Horse method serves as a mechanism for verifying compliance, analogous to the use of a radar gun in traffic enforcement. Just as drivers are aware that speed limits exist and are subject to enforcement—regardless of whether a police vehicle is marked or visible—students were fully aware of the academic integrity policies governing the assignment. Legal scholars distinguish between a "speed trap" (hidden enforcement of existing laws) and "entrapment" (inducing a person to commit a crime they would not otherwise commit). Similarly, the embedded text in the PDF did not induce students to use AI; it merely ensured that if they chose to violate the stated policy, the violation would be detectable. This strategy is not intended to "trick" students, but to reinforce the validity of the assessment and provide a basis for necessary educational interventions

7. Limitations

This investigation is constrained by a limited sample size—comprising 25 students across 6 teams—which necessarily restricts the generalizability of its findings. Nevertheless, the pressing challenge of AI misuse within higher education underscores the importance of disseminating mitigation strategies, even when empirical data remains limited. The proposed method also presents accessibility concerns; specifically, the use of hidden text may impede students with visual impairments who rely on screen readers. In such cases, alternative accommodations or supplementary structural interventions should be considered to ensure equitable participation. The instructor version of the PDF contains the traps, while an accessible text-only version (without traps) is available upon request for students with documented accommodations, ensuring equity without compromising the integrity of the general assessment. Furthermore, as previously discussed, the rapid advancement of large language models may diminish the long-term effectiveness of these approaches, thereby requiring continuous refinement and adaptation.

8. Conclusion

The objective of "AI Resilient" pedagogy is not punitive detection of academic dishonesty, but rather the affirmation and preservation of educational integrity within the learning process. Through the deliberate incorporation of adversarial examples in course materials, instructors can reveal instances of excessive dependence on automated tools and thereby prompt a renewed emphasis on independent, critical analysis.

Findings from this experiment underscore that students are not merely employing AI as a "co-pilot" for reviewing their work; rather, they are utilizing it as a primary agent to produce substantive content. The uncritical acceptance of AI-generated recommendations—such as decisions to dismiss a high-performing CEO or disclose confidential intellectual property—reflects a concerning relinquishment of independent judgment. However, as evidenced by the "Penalty Round" in this case study, students are fully capable of rigorous, analytical engagement when deprived of AI as a shortcut. In many cases, a carefully designed technological intervention can serve as a necessary prompt, encouraging students to trust and develop their own analytical abilities.

References

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128(4), 612-637.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. Memory & Cognition, 35(2), 201-210.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), Psychology and the real world: Essays illustrating fundamental contributions to society (pp. 56-64). Worth Publishers.

Garvin, D. A. (2003). Making the case: Professional education for the world of practice. Harvard Magazine, 106(1), 56-65.

Kosmyna, Nataliya & Hauptmann, Eugene & Yuan, Ye & Situ, Jessica & Liao, Xian-Hao & Beresnitzky, Ashly & Braunstein, Iris & Maes, Pattie. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. 10.48550/arXiv.2506.08872.

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. Ethics & Behavior, 11(3), 219-232.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husen & T. N. Postlethwaite (Eds.), International encyclopedia of education (2nd ed.). Pergamon Press.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. Educational Psychology Review, 16(4), 385-407.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. Journal of Experimental Psychology: Human Learning and Memory, 4(6), 592-604.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2011). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31(2), 261-292.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64-70.

Appendix A: Student Correspondence (Anonymized)

Reference: Email response from Student B denying AI usage prior to the revelation of hidden text markers.

Subject: Re: Case Grade

"Professor,

I wanted to express my thoughts regarding the Case Grade and the accusations made in this email. From what I can tell from working with my teammates, not one of us utilized AI to do our thinking.

[...] I did NOT use AI, even if that's what was reported on the AI scanner. I will not tolerate being accused of using AI when I did not. I would like the grade to be re-evaluated without "AI" usage being considered, but rather with proper evaluation on what we did wrong or right in OUR writing."