Teaching Responsible AI Use as a First-Year University Skill: A Structured-Engagement Action Research Proposal

Benjamin Farenhorst, B.Sc., MBA

Doctoral Student in Artificial Intelligence, Golden Gate University

Abstract

Generative artificial intelligence (AI) has become routine in how first-year university students complete academic work, yet most institutional responses focus on detection rather than skill development. This research-in-progress report proposes an action research study that reframes responsible AI use as a teachable academic skill. The argument draws on Haidt (2024), who contends that Generation Z grew up overprotected in the physical world and underprotected in the digital world, with reported consequences for sustained attention, self-regulation, and mental health. While Haidt's broader claims remain contested in the developmental psychology literature (Odgers, 2024; Orben & Przybylski, 2019), the structural element of his argument—that powerful new technologies require shared norms and bounded contexts—is consistent with established findings in self-regulated learning. The article extends this argument to higher education: a cohort entering university amid uneven digital norms now encounters generative AI under similarly uneven conditions, and structured course-level conditions are needed if the first year of university is to develop the cognitive capacities it claims to develop. The proposal also draws on cognitive offloading, self-regulated learning, and retrieval practice. The proposed intervention introduces three mechanisms in first-year courses: optional source restrictions, in-progress comprehension check-ins embedded in AI-assisted assignments, and brief personalised verification activities completed in class. Responsible AI use is operationally defined and measured through three observable indicators. No outcome data are reported; the article presents the study as planned work intended to contribute a practice-oriented model for teaching responsible AI use in the first year of university.

Introduction and Context

Generative AI tools entered undergraduate classrooms faster than universities produced guidance for how students should use them (Kasneci et al., 2023; Ng et al., 2024). First-year students now arrive on campus having already used AI throughout secondary school, and they encounter a university system in which policies vary by course, by instructor, and sometimes by assignment. Students are uncertain about what is permitted, instructors are uncertain about what is happening, and responses remain weighted toward detection rather than toward the cognitive and academic skills university is meant to build (Cotton et al., 2024).

This landscape does not arrive on campus as a neutral starting point. Haidt (2024) argues that Generation Z entered higher education following a period of rapid change in adolescent digital environments, and that this shift coincided with declines in sustained attention and self-regulation. Haidt's interpretation of these trends is debated: Odgers (2024) and Orben and Przybylski (2019) note that the empirical association between adolescent screen use and mental health is small and inconsistently replicated, and caution against attributing population-level changes to a single cause. The present study does not adopt Haidt's full causal account. It draws on the narrower, less contested element of his argument: that powerful new technologies tend to be used more constructively when shared norms and structured contexts are in place. That structural claim is consistent with long-standing findings in self-regulated learning research (Panadero, 2017; Zimmerman, 2002) and provides the rationale for a course-level rather than purely individual intervention.

The problem this study addresses is the absence of practice-ready interventions that treat responsible AI use as a teachable skill in the first year of university. First-year undergraduates are old enough to be treated as autonomous learners, yet their self-regulatory and metacognitive capacities are still consolidating (Panadero, 2017; Zimmerman, 2002). They face a step-change in academic expectations: longer reading loads, longer written assignments, and greater independence. The same AI tools that can scaffold this transition can also substitute for the cognitive processes the first year is designed to develop (Bastani et al., 2024).

This report outlines a proposed action research study in first-year undergraduate courses. It follows the journal's Research-in-Progress format: it describes a real-world problem, proposes an intervention, and reflects on early implementation challenges.

Theoretical Background

The Case for Structured Course-Level Conditions

Any intervention designed for first-year students must take seriously the developmental context they arrive in. As noted above, Haidt (2024) characterises recent cohorts as having entered higher education during a period of rapid change in adolescent attentional environments. The empirical strength of his population-level claims is contested (Odgers, 2024; Orben & Przybylski, 2019), and this study does not depend on them. What it does adopt is the narrower structural argument: that the constructive use of a powerful new tool is shaped by the norms and contexts surrounding it, not only by individual willpower. This is a familiar idea in education research, where self-regulation is understood to develop within environments that scaffold it rather than as a fixed individual trait (Panadero, 2017; Zimmerman, 2002).

Applied to the present case, the implication is straightforward. First-year students encountering generative AI without shared course-level expectations face a coordination problem: an individual student is unlikely to refrain from unstructured AI use if they expect their peers to use it freely. Responsible AI use is therefore unlikely to emerge from honour codes alone. It is more plausibly developed within courses that make their expectations and supports explicit.

Cognitive Offloading and the Cognitive Debt Hypothesis

Cognitive offloading is the use of external resources to reduce internal cognitive demand (Risko & Gilbert, 2016). Offloading routine operations can be adaptive; offloading the developmental target of an assignment is not. When a first-year student uses AI to generate an essay or a problem solution, the offloaded task is the skill the assignment was designed to build.

Recent work has begun to examine whether AI assistance produces learning costs that persist beyond the assisted task. Kosmyna et al. (2025), in a preprint that has not yet been peer-reviewed, reported reduced frontal engagement during AI-assisted essay writing and described this pattern as cognitive debt. The study's small sample size and preprint status mean its findings should be treated as preliminary and as a hypothesis to be tested rather than as established evidence. Two peer-reviewed studies provide more cautious points of reference: Abbas et al. (2024) reported a negative association between higher AI tool use and critical thinking performance among university students, and Bastani et al. (2024), in a randomised field experiment, found weaker independent performance on transfer tasks among students who had

received AI assistance during initial problem solving. Together, these studies motivate—but do not yet confirm—the concern that unstructured AI use during skill-acquisition phases may interfere with the development of the underlying capacity. The present study treats this as an open empirical question rather than a settled finding.

Self-Regulated Learning and Retrieval Practice

Self-regulated learning (SRL) describes the cyclical process through which learners plan, monitor, and evaluate their own cognitive activity (Zimmerman, 2002). Panadero (2017) identified metacognition as the central mechanism linking SRL to learning outcomes. Unstructured AI use can short-circuit the SRL cycle: the student may neither plan nor monitor their own reasoning, and the submitted product may not reflect their thinking at all. Bjork et al. (2013) documented an illusion of knowing that follows exposure without retrieval, a pattern AI-generated summaries appear well placed to amplify.

Retrieval practice produces stronger and more durable learning than passive review (Karpicke & Blunt, 2011; Roediger & Butler, 2011). A course that couples AI assistance with brief retrieval activities provides the structural conditions under which self-regulated learning can actually be practised.

From Detection to Skill Development

Institutional responses to AI in higher education have largely focused on detecting AI-generated text rather than on designing instruction that treats AI use as a skill (Cotton et al., 2024; Holmes et al., 2019). Detection framings position AI as an adversary and place students and instructors on opposite sides of a compliance relationship. Skill-development framings treat AI literacy as part of the curriculum and assume students can learn to use AI in ways that support their own cognitive engagement (Luckin et al., 2016). This study adopts the skill-development framing explicitly.

Research Question and Aim

The study is guided by the question: How does a structured AI engagement protocol, delivered in first-year undergraduate courses, shape students' responsible AI use, self-regulated learning behaviours, and academic performance relative to unstructured AI access?

The aim is to develop and evaluate a classroom-ready protocol for teaching responsible AI use as an academic skill in the first year of university, and to generate practice-focused recommendations that instructors can implement without specialised infrastructure.

Operational Definition of Responsible AI Use

For the purposes of this study, responsible AI use is defined as the use of generative AI tools in academic work in a manner that supports, rather than substitutes for, the cognitive and academic skills the assignment is designed to develop. This definition is intentionally task-anchored: what counts as responsible use depends on what the assignment is trying to build.

Three observable indicators are used to operationalise the construct:

1. **Disclosure and traceability.** The student declares which AI tools were used, for what purpose, and at what stage of the assignment, in a brief structured statement attached to the submission.
2. **Engagement with one's own work.** The student is able to articulate, in their own words and without AI assistance, the central claims, evidence, and reasoning of the submitted work during a short verification activity.
3. **Alignment with task purpose.** AI use is directed at activities that support the developmental target of the assignment (e.g., clarifying terminology, checking grammar, exploring counter-arguments) rather than at producing the assessed deliverable itself.

Each indicator is rated on a three-point scale (not demonstrated, partially demonstrated, fully demonstrated) using a rubric applied to the student's disclosure statement, check-in responses, and verification performance. The composite score forms the primary measure of responsible AI use; the individual indicators are also retained to allow more granular analysis.

Proposed Intervention

The intervention introduces three mechanisms in first-year undergraduate courses, each grounded in the theoretical foundations above.

Mechanism 1: Optional Source Restrictions

Instructors may specify which readings, datasets, or sources students are permitted to draw on for a given AI-assisted assignment. This anchors student work to course materials and reduces the scope for AI to generate off-syllabus content. For first-year students, the anchoring is a scaffold rather than a constraint: it communicates what counts as legitimate source material and models the disciplinary reading practices the course is trying to build.

Mechanism 2: In-Progress Comprehension Check-Ins

As students work on AI-assisted assignments, they respond to short comprehension questions at defined intervals, articulating in their own words what they have understood so far. Check-ins function as retrieval practice embedded in the assignment (Karpicke & Blunt, 2011) and as process-level evidence of engagement instructors can review alongside the final product (Holmes et al., 2019). Each check-in consists of two to three short-answer prompts, takes approximately five minutes to complete, and is delivered through the existing learning management system (LMS) at three points across the assignment timeline (after planning, after a draft, and before submission).

Mechanism 3: Personalised In-Class Verification

After submission, students complete a brief in-class verification activity derived from the content of their own submission. The activity is individualised, closed-book, and approximately ten minutes long, fitting within the opening of a regular class period. It gives instructors a direct comparison between the understanding shown in the submitted work and what the student can reproduce independently. Because verification is derived from each student's own submission, it is resistant to AI capability improvements: a student cannot prepare for it without genuinely engaging with their assignment.

AI Literacy Component

The intervention also includes a short sequence of AI literacy activities (approximately two 30-minute sessions) in the first three weeks of the course. These introduce students to the cognitive considerations associated with AI use, the conditions under which AI assistance supports learning, and the expectations of the course. The framing is explicit: responsible AI use is a skill the course is teaching, not a rule it is enforcing.

Proposed Methodology

Design

The study will use a quasi-experimental mixed-methods design. The quantitative strand will compare outcomes across two conditions, structured engagement and unstructured access, using pre- and post-measures. The qualitative strand will use instructor interviews and student focus groups. Data from both strands will be collected concurrently and integrated at the interpretation stage.

Setting and Participants

The proposed setting is introductory undergraduate courses at a Canadian university, in disciplines with substantial writing or problem-solving components. The pilot will recruit four to six first-year courses (two to three in the structured-engagement condition, two to three in the comparison condition), targeting an analytic sample of approximately 300–450 students (assuming 75–100 enrolled per course and roughly two-thirds providing consent and complete data). This sample size provides approximately 80 percent power to detect a small-to-medium between-condition effect (Cohen's $d \approx 0.30$) on the primary outcomes at $\alpha = .05$, accounting for clustering of students within courses (intraclass correlation assumed at 0.05). Courses will be matched on discipline area, enrolment size, and assessment structure where possible. Recruitment will run through course coordinators and the university's teaching and learning centre over a one-semester preparation window. Inclusion criteria require instructor agreement, LMS access sufficient to deliver the protocol, and the ability to run a brief in-class verification activity within a regular class period.

Implementation Timeline

The pilot is planned across two academic terms. Term one is used for instructor recruitment, protocol customisation to each course, LMS setup of the check-in templates, Research Ethics Board approval, and a small dry run with a single course section. Term two is the full implementation: AI literacy sessions in weeks one to three, structured-engagement mechanisms applied to two AI-assisted assignments per course, and pre- and post-measures administered in week one and week twelve respectively. Instructor interviews and student focus groups will be scheduled in the final two weeks of term.

Outcome Measures

Three primary outcome domains will be assessed. Critical thinking will be measured using the Critical Thinking Assessment Test (CAT) or an equivalent validated instrument appropriate for university populations, administered pre and post. Self-regulated learning behaviours will be measured using an adapted Motivated Strategies for Learning Questionnaire. Responsible AI use will be measured through the three-indicator rubric described above, applied to disclosure statements, instructor-rated check-in artefacts, and verification performance, supplemented by a brief student self-report instrument. Course performance will serve as a secondary outcome, drawn from instructor-assigned grades on comparable assessments across conditions.

Analysis

Quantitative analysis will use multilevel models with pre-intervention scores as covariates and students nested within courses, with course-level random intercepts to account for clustering. Qualitative analysis will use thematic analysis (Braun & Clarke, 2006) of instructor interviews and student focus groups. Integration will occur at the interpretation stage, using quantitative patterns to identify contexts for deeper qualitative exploration and qualitative findings to explain unexpected quantitative results.

Ethical Considerations

The study will require Research Ethics Board approval prior to data collection. Informed consent will be sought from all student participants, and participation will be decoupled from course grades: no student will receive a lower grade for declining to contribute research data, and instructors will not be told which students declined. Instructors will be offered the protocol as a teaching tool regardless of whether their students consent to the research component.

Early Reflections and Implementation Challenges

Several challenges have emerged during the design phase of this study.

The first challenge is framing. Colleagues asked about this project frequently assume it is an academic integrity study. The skill-development framing has to be stated repeatedly, and the intervention must visibly avoid surveillance. The in-class verification, in particular, can be misread as a trap. Making it brief, ungraded for research purposes, and derived from the student's

own work rather than an external answer key is an attempt to communicate the framing through the mechanism itself.

The second challenge is instructor readiness. First-year courses at scale are often taught by a mix of tenure-track and sessional instructors, and capacity to implement the protocol varies across these roles. The check-in mechanism can be delivered through existing LMS quiz functionality without new infrastructure, but the verification activity requires a short, protected window of class time. Securing this window depends on the instructor of record.

The third challenge is avoiding a compliance reading of the protocol by students. The AI literacy component is intended to make the reasoning behind the protocol visible: check-ins are framed as retrieval practice that benefits students' own learning, and verification as an opportunity to see, privately and without grade consequence during the pilot, what students can reproduce independently. Whether students read the protocol this way, or as a disguised academic integrity regime, is a question the qualitative strand will address.

The fourth challenge is the pace of change in the AI tools themselves. The verification mechanism is designed to be resistant to AI improvements because it draws on each student's own submission. Even so, the literacy component will need updating as the tool landscape shifts.

Anticipated Contribution

If the study proceeds as proposed, its anticipated contribution is a practice-ready protocol for teaching responsible AI use as a first-year university skill, together with an evidence base about its effects on responsible AI use, self-regulated learning, and course performance. The protocol is designed to run in standard first-year courses using common LMS tools. The framing is action-oriented: the study is organised around a real instructional problem, documents an intervention, and generates reflection and implications for practice across disciplines.

Acknowledgement of AI Tool Use

In preparing this report, the author used a generative AI assistant for language editing, structural organisation, and clarity of expression. The research question, theoretical framing, intervention design, methodological decisions, and reflections on implementation are the author's own and draw on the author's ongoing doctoral study in artificial intelligence. No AI tool was

used to generate the research content, interpret the literature, or formulate the scholarly argument. This use is disclosed in line with the journal's AI-use policy.

Conflict of Interest Disclosure

The author declares no financial conflicts of interest. The author is a doctoral student in artificial intelligence at Golden Gate University, and this interest informs the motivation for the study but does not affect its design, conduct, or reporting.

References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Generative artificial intelligence and critical thinking: Double-edged effects of information literacy and cognitive fatigue on university students. *Thinking Skills and Creativity, 53*, Article 101579. https://doi.org/10.1016/j.tsc.2024.101579

Bastani, H., Jones, M., Lifshitz-Assaf, H., & Webb, T. (2024). Generative AI, productivity, and learning: Evidence from a randomized field experiment. *Science, 384*(6693), 742–747. https://doi.org/10.1126/science.adk7899

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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

Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Penguin Press.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., & Kasneci, G. (2023). ChatGPT for good? Opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* [Preprint]. arXiv. https://arxiv.org/abs/2506.08872

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2024). Generative artificial intelligence in K–12 education: A systematic review. *Frontiers in Education, 9*, Article 1298457. https://doi.org/10.3389/feduc.2024.1298457

Odgers, C. L. (2024). The great rewiring: Is social media really behind an epidemic of teenage mental illness? *Nature, 628*(8006), 29–30. https://doi.org/10.1038/d41586-024-00902-2

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour, 3*(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2