香 港 大 學
THE UNIVERSITY OF HONG KONG

LAIDLAW FOUNDATION

Laidlaw Undergraduate Research and Leadership Scholarship

Research Project Essay

Investigating Assessment Design in the Context of Generative AI

Galiya Akhmetova

Bachelor of Arts, Year 2

Supervisor: Professor David Carless

The University of Hong Kong

2026

Table of Contents

1. Introduction

Generative AI has been available to university students for nearly four years and has become routine for most undergraduates. Stephenson and Armstrong (2026) report that 95% of undergraduate students in the UK use AI at least in one way. Since the public release of ChatGPT in 2022, universities across the globe have responded largely by trying to detect and restrict AI use. Software systems like Turnitin now supplement traditional similarity scores with an estimated AI-use indicator (Turnitin, n.d.). Some instructors have gone further, embedding hidden prompt injections in assignment guidelines, so that any answer produced by an AI agent carries a detectable tell. Two students in this study reported encountering such "traps", in unrelated interviews.

Detection measures assume the issue is with the students breaking the rules. A student I interviewed recalled writing an essay himself but scoring lower than a classmate who completely AI-generated his and ran it through a "humanizer." He said he felt "betrayed," as no detector caught this. Detectors themselves can be unreliable, but with new "humanizer" tools their task became even more complicated. Perkins et al. (2024) showed that detectors' accuracy dropped from 39.5% to 17.4% after an AI-generated text was modified to avoid being caught.

This episode can be easily interpreted as a failure of policy enforcement; however, I argue it depicts a failure of assessment design. In the context of these powerful tools, many forms of assignments make outsourcing to AI the easier and sometimes even more rational choice. Redesign is the logical response, but there is little consensus on what it should look like among scholars. One discussed proposal is the "two-lane" approach, which sorts out assignments into two groups where AI is prohibited or allowed (Liu & Bridgeman, 2023). Curtis (2025), however, argues that this all-or-nothing approach leaves out the more useful options that fall in between. What tends to be missing, though, is the point of view of the people actually taking these assessments. This study draws on twelve interviews with undergraduate students across ten disciplines to investigate effective assessment redesign in higher education.

2. Methods

This study is based on qualitative analysis of individual interviews with 12 HKU undergraduate students from across ten disciplines, such as engineering, history, psychology, journalism, chemistry, speech and language pathology, computer science, business administration, philosophy, and biomedical science. Participants were recruited from the author's personal network and were mostly in their first or second year of study during the 2025-26 academic year.

The semi-structured interview format was deliberately chosen, as it allowed for deeper exploration of certain points students wanted to elaborate on and for skipping other questions less relevant to them. Topics included current AI use, the assessments participants had experienced, their perceptions of easiest and hardest ones to outsource to AI, which taught them most and least, and what they would change if they were dean.

The interviews were recorded, transcribed, coded by the author, and anonymized. Recurring themes were grouped, each theme marked with the number of interviews supporting it. Participants were over 18 years of age, and their informed consent was taken through Google Forms.

3. Findings and discussion

3.1 “AI-proof” assessment types

The assessment types that the interview participants indicated as most AI-proof fall into three categories. I have grouped them by the mechanism through which they resist AI: removing access, requiring live performance, and drawing on students’ own experience.

3.1.1 Removing the access

In-person examinations

Rather expectedly, in-person closed-book testing was deemed one of the most AI-resistant formats (5/12). Interviewee 10 went as far as to claim that only paper-based, no-tech assessment fully prevents AI use. The assignment itself does not require change, as removing access to the tool is enough in this case. However, several students reported that this single high-stakes assessment format creates excessive stress and leads to procrastination. Interviewee 2 described his experience with a midterm as “demotivating” and pushing him to “self-sabotage” to the extent he wanted to give up on it altogether. Interviewee 3 reported that in-person examination does not resemble the real-world work environment and, thus, does not appropriately prepare him for his future career. So, the format considered most resistant to AI is also the one students feel is most damaging. Several participants pointed to an alternative that retains the advantages but eliminates high stress.

Continuous assessment

Continuous assessment distributes the stakes throughout the semester and reduces last-minute cramming (4/12). Usually in the form of weekly in-class quizzes, these assignments are still proctored and, in most cases, remove access to AI. Though it requires more short-term effort, Interviewee 2 argues that “every week, you have an opportunity to push yourself,” leading to better long-term retention. Interviewee 11 explicitly compared high-stakes exams and low-stakes frequent assessment. He recalled being more prepared for the final test in a course with weekly quizzes, compared to another class, for which he crammed 800 pages before a single final exam.

3.1.2 Requiring live performance

This group includes live presentations (5/12), podcasts and short films (4/12), and roleplay assignments (1/12). The main AI resistance mechanism here comes from the requirement to perform live. This forces students to process information, memorize it, and deliver it orally, which no AI could do on their behalf.

Podcast

Interviewee 8 described a podcast assignment in an introductory politics course as fundamentally different from traditional formats, because she had to “strike the balance” between eye-catching visuals, engaging script, and a compelling live performance to “sound like a true politician.” The requirement to deliver her segment personally made outsourcing to AI simply insufficient, because no model could perform in her place.

Live-performance-based assessment does not exclude AI, since students can still use it to draft their speech and do background research. Interviewee 9 described using AI to locate and translate sources in Chinese - a language the student does not speak. This suggests the format can prompt students to use artificial intelligence in ethical ways that support their learning, rather than over relying on it.

Roleplay

The effectiveness of live-delivery assessment, however, depends on implementation. This is best illustrated by Interviewees 8 and 10’s opposite experiences with similar roleplay assignments. While the former student reported personally enjoying the unusually creative task and performed it with relative ease, Interviewee 10 expressed frustration about the uncertainty that stemmed from that same creative openness. She described the assignment as unstructured and unaligned with exam content, and even said she felt it was a waste of time. This contrast raises the question of the quality of implementation of these new “creative” assignments.

One reason may be that instructors leave such tasks deliberately open-ended so that students have to work out their approach themselves (Daly et al., 2016). This ambiguity, however, may confuse students, which in turn pushes them to procrastinate and rush the assignment by handing it to AI. This exact scenario happened to Interviewee 10, as the lack of communication on *why* this assignment lacks usual point-to-point instructions led her and her teammates to delay the work and resort to AI’s help at the last minute.

3.1.3 Hands-on learning

Similar to the live-performance format, hands-on assignments by their nature limit AI use to scaffolding function. This category includes personal self-experiments (3/12) and internships (3/12).

Internship

In Interviewee 6’s internship-based course, she had to propose a solution to a real-world problem using the knowledge acquired in classes. Generative AI was allowed for routine tasks, so she had more time to focus on interviewing stakeholders and understanding the problem on a deeper level. More importantly, the participant evaluated her experience as far more valuable and applicable in real life than a standard research paper. However, it should be acknowledged that this assessment would potentially be difficult to implement, as large class sizes are a known barrier to authentic assessment (Islam & Ahmed, 2018; Vlachopoulos & Makri, 2024).

Personal experiment

Three participants (Interviewees 4, 7, 10) in separate interviews reported engaging in a behavioral self-experiment as part of the introductory psychology course. The course enrols over 500 students in a semester, so it can be a useful example of "AI-proof" assessment implemented at scale. In this class, participants ran a two-week behavior-change experiment of their choice, with Generative AI use restricted to the outlining stage.

Interviewee 7 recalled working on phone-free study sessions, leaving her mobile phone in the dorm and going to campus to study for several hours. The student reported that the assignment had a lasting impact on her behavior and helped her understand the underlying psychological theories, as the task required her to explain self-observed behavior using concepts taught in class. She highlighted that this assessment specifically was how she managed to remember theories that would then be tested during the midterm and final examinations. Her account is consistent with the self-reference effect, where information processed in relation to oneself is recalled better (Symons & Johnson, 1997). The fact that students chose their own target behavior may further reinforce that effect.

Interviewee 10, who earlier expressed frustration at the lack of clear guidelines in creative assignments, actually found this experiment-based one "most useful" precisely *because* the instructions were detailed. An entire grading rubric, an experiment framework, and three months to complete it allowed her to put in the necessary effort and "actually learn from it."

Referring back to her roleplay experience, this suggests that what frustrated her was not really the creative or unfamiliar format but the absence of clear expectations from instructors.

3.2 Why students outsource to AI

Not all AI use is cheating under HKU's policy, and most of what interview participants described was study support. However, when students do outsource to models, the reasons tend to fall into two larger categories - the ones rooted in task design and the ones stemming from course organization.

3.2.1 Reasons in the task

Student's input

Essays were named simplest to hand to AI most consistently (4/12). Interviewee 1 specifically pointed to written reflections on readings as the clearest example of tasks that could be uploaded and entirely generated by AI. Everything the task requires - the readings, instructions, word count - can be handed to a model in one upload. What makes this format vulnerable is that a reflection does not require a fact check or any "hard" inputs from the student. The ratio of output to effort is unusually high, which can make outsourcing look like the rational choice to students who are short on time. In contrast, one of the most AI-proof assessments discussed earlier - personal experiment - depends on the behavior change data that only the student can provide. Both tasks end in a piece of writing, but the difference is whether the assignment asks for material a model has no access to.

AI permission without redesign

One response to essay format's vulnerability is to permit Generative AI use openly rather than fight it. Interviewee 3 described a case where the instructor introduced a task to test AIassisted essay writing, allowing students to use it freely. This seemingly liberating strategy, however, backfired. Almost everyone used AI and scored close to 100%, so the intended grade curve became simply meaningless. This suggests that the limiting factor is not the AI policy - rather, the task design. The internship-based assessment described earlier also provided freedom to use tools but produced the opposite result. The student there used the tools for routine operations and spent more time on higher-level tasks that a model could not do. By the same logic as the personal experiment, the internship assignment required something the student had to gather herself, while a routine essay usually does not.

No oversight

Another assessment type students found particularly easy to outsource is non-proctored online quizzes. Interviewee 10 highlighted that most of her peers approach these online tests with the goal of just “getting done with them.” Students can paste every question into AI and submit the answers, often without even checking them first. These quizzes usually count for only a few percentage points, so the effort of completing them properly may be difficult to justify. Another explanation behind this choice lies in the absence of oversight. The same participant testified that her friend, who used AI on every Moodle quiz, never cheated during any invigilated in-person exams. Though the grade credit was the same, the online setting removed the sense of responsibility that a physical, proctored room creates. In a way, non-proctored online tests almost invite students to cheat with AI.

3.2.2 Reasons in course organization

Excessive workload

Students also reported outsourcing work in well-designed assessments. Interviewee 10 illustrated this in her experience with a Common Core course on a topic she was genuinely interested in and was looking forward to. She reported that the course, however, had “ten times as much work” as her major courses. Apart from other assignments, students had to read a rather long book and write a review on it. “It was humorous, sarcastic, there was satire,” she said, as she recalled enjoying the first few chapters of that book. But the unusually high amount of assessment tasks in the course and the length of that book left her unable to finish on her own and she ran it through AI for a summary. The intrinsic interest was present in this case, so the usual explanation that students outsource because they find course material boring does not apply here. The student resorted to AI’s help because of the number of competing demands within the course and not because the task itself was poorly designed.

Poor grade distribution

Within the same category lies another possible driver of outsourcing - grade weighting. Interviewee 11 described a course with three essays, each worth 15 to 20%, and said this made him feel there was little point in putting real effort into all three. Each essay demanded a full cycle of reading, planning, and writing, but the effort required probably did not correspond to its weight. Splitting a grade this way instead of concentrating it in one larger assessment may assign less importance to the individual smaller tasks, prompting students to save time and effort on them.

3.3 Students’ assessment redesign wishes

Apart from discussing their experiences with AI-resistant assessment, participants shared their views on broader structural change. Four participants indicated teaching legitimate AI use as the most desired reform. Interviewee 7 explicitly called for assignments that help students learn to use these tools correctly, rather than restricting them altogether. She described one such task in her major course, where students had to find and correct errors in AI-generated answers. Catching those errors required a high enough understanding of the material to know when the model is wrong. Interviewee 11 suggested adding one smaller-weight assessment in his major courses that would allow unrestricted use of artificial intelligence. Arguing that "you can't just ignore AI and pretend it's not there," he framed this as a better preparation for a future career in the workplace, where such tools are already deployed routinely.

The two participants who identified course structure as a driver of outsourcing also proposed changes to it. Interviewee 10 wanted academic requirements to be "a lot more reasonable" in terms of workload, so that students do not have to decide which assignments are important to do themselves and which ones could be AI-d. A related reform was proposed by Interviewee 11, who would prefer fewer assignments but with more weight, so that the course work builds towards something substantial, instead of repeating the same tasks multiple times. These are mainly structural rather than pedagogical fixes that do not change the tasks themselves but *how* they are organized. So, according to students, if an assignment is meant to be taken seriously, it should not be competing with five others.

4. Limitations

This study recruited twelve participants from the author's personal network, which limits the generalizability of these findings. It must also be acknowledged that most interviewees were in their first and second years of undergraduate at the same institution, so the perspectives of upperclassmen and of other university students are absent. Due to the study's qualitative design, students self-reported their own and their peers' behavior on a sensitive topic of AI use in academic work. This may have led some to understate their own reliance or overstate others' use.

Some findings are based on a single detailed account and are identified in the text. The counts (e.g. 5/12) provided in the text indicate only how many of the twelve interviews supported a theme, and not any statistical claim.

5. Conclusion

This research offers insight into assessment redesign in the context of Generative AI, drawing on interviews with twelve undergraduate students. The three categories of assessment that appear to resist AI most effectively are in-person examinations that remove access to models, live-performance-based assignments, and assessments that draw on a student's own experience. What unites the second and third groups is that a student’s work depends on something AI cannot supply - the student’s presence or personal data.

The reasons students gave for outsourcing coursework fell into two groups. Some stem from the task design, where an assignment did not require any information only a student can provide, or where there was no human oversight. Other motives are grounded in the course structure, rather than assessment design. Students tend to hand homework to Gen AI even in courses they genuinely enjoyed, as too many assessment components competed for their attention. Mismatch between required effort and grade weight distribution also compelled undergraduates to outsource.

The reforms students proposed followed from the rationales of their AI misuse. Participants asked for explicit teaching of legitimate AI use, more reasonable workload, and a better-distributed grade weighting. Notably, these suggestions have little to do with better detection. What students asked for instead were changes to how much a course demands and how it distributes credit.

References

Curtis, G. J. (2025). The two-lane road to hell is paved with good intentions: Why an all-or-none approach to generative AI, integrity, and assessment is insupportable. *Higher Education Research & Development*, *44*(8), 2151–2158. https://doi.org/10.1080/07294360.2025.2476516

Daly, S.R., Mosyjowski, E.A. and Seifert, C.M. (2016), Teaching Creative Process across Disciplines. *The Journal of Creative Behavior*, 53: 5-17. https://doi.org/10.1002/jocb.158

Islam, N. N., & Ahmed, I. (2018). Thinking unconventional: Alternative assessment at tertiary level education in Bangladesh. *International Journal of New Technology and Research, 4*(1), 18–23. https://www.ijntr.org/download_data/IJNTR04010040.pdf

Liu, D., & Bridgeman, A. (2023, July 12). *What to do about assessments if we can't out-design or out-run AI?* Teaching@Sydney, University of Sydney. https://educational-innovation.sydney.edu.au/teaching@sydney/what-to-do-about-assessments-if-we-cant-out-design-or-out-run-ai/

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education, 21*, 53. https://doi.org/10.1186/s41239-024-00487-w

Stephenson, R., & Armstrong, C. (2026). *Student generative AI survey 2026* (HEPI Report 199). Higher Education Policy Institute. https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*(3), 371–394. https://doi.org/10.1037/0033-2909.121.3.371

Turnitin. (n.d.). *Using the AI writing report*. Turnitin Guides. https://guides.turnitin.com/hc/en-us/articles/22774058814093-Using-the-AI-Writing-Report

Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation, 83*, 101425. https://doi.org/10.1016/j.stueduc.2024.101425