



A pragmatic protocol for stress testing assessment formats against GenAI

Philip M. Newton^a and Tomáš Foltýnek^b

^aSwansea University Medical School, Swansea, SA2 8PP United Kingdom

^bFaculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic

*Corresponding author p.newton@swansea.ac.uk

Educators are concerned, interested or sceptical about the potential for students to use new Generative AI (GenAI) tools like ChatGPT to complete their assignments. Some headlines have warned of an assessment 'apocalypse' (1) while others focus, in unflattering terms, on the weaknesses of GenAI (2). The concerns seem legitimate, with GenAI showing excellent performance in many common assessment formats like essays and online exams (3), suggesting these formats are at risk from cheating with GenAI. The tools are improving rapidly, and this is reflected in substantially improved performance of newer models on current assessment formats in higher education (4,5)

A recent report by the UK Higher Education Policy Institute showed that 92% of UK University students had used GenAI in the past year, and recommended that all universities should 'stress-test' their current assessment practices against GenAI (6), in part to identify the risk that students might cheat. There are clear arguments in favour of future assessment reform, to bring GenAI into some assessments, but also to ensure that assessments of foundational knowledge are carried out under secure conditions. This paper is a protocol for how to stress test *current* assessments simply and effectively, with the following aims

- Allowing educators to learn, experientially, what GenAI tools can, and cannot, do with their current assessment practice.
- Allow the reporting of the findings of a stress-test in a way that can be understood and replicated by others.
- Initiate a process of reflection and redesign to focus on secure ways to assess learning in the absence of GenAI, and/or ways to incorporate GenAI use into assessment.

The protocol has been developed using the experience of the authors, based upon many papers which report the performance of GenAI tools, and the challenges associated with replicating the findings of those papers. The protocol is not perfect and will almost certainly go out of date - things are moving very fast. However, if you follow the principles, as closely as you can, the results should still be useful - hence the 'pragmatic' approach.

Part A. Planning the test

- **Compose novel assessment items if possible.** Avoid using sample papers or exam questions that are in the public domain. These will likely have been part of the training materials for the GenAI tools.
- **Assume the content posted into the tool will be 'in the public domain'.** The policies used by GenAI companies for copyright, confidentiality, security etc all vary by tool, by license, and over time. The safest approach is to generate novel content and then assume it is now in the public domain.
- **Clear the memory of the generative AI tool before proceeding.** This is to ensure that the responses of the tool are not influenced by your prior interaction with it. If you don't want to do this, then you might achieve the same result by using 'temporary chat' (see below).
- **Decide how many opportunities the tool will be given to answer correctly.** If cheating is your primary concern, then allowing only one answer is perhaps the most valid test (and the quickest!), but you might learn more by asking the tool to 'try again' more than once. Decide on the extent of clues you want to provide when asking the tool to try again.



European Network
for Academic Integrity





Part B. Performing the test

- **Use temporary chat where possible.** This should mean that content is not saved by the tool or used to inform your future interactions with it.
- **Use a new chat for every question.** To prevent answers from one question influencing the tool's response to the next one.
- **Use only the assessment/exam question and use it verbatim (though see #8).** Copy and paste the assessment item directly into the chat window, with no additional information. This is the simplest and cleanest format and seems likely to best represent the approach used by a student. Avoid putting multiple questions in one prompt.
- **Record any additional prompts that were used to ask the generative AI tool to complete the assessment.** There may be a need or rationale for using additional information, besides the question itself (for instance, the tool might refuse to answer questions on cybersecurity, health or law, unless it is told that this is a (preparation for an) exam and not a real case). Any additional information you add can profoundly influence the outputs from GenAI, so keep a record of any additional information added.
- **If the tool refuses to answer, make it clear in the report.** Despite the effort described in the previous questions, the tools may keep refusing to answer. If this is the case, make a record of it.
- **Save the chat if possible.** Many tools give you the option to email the chats to yourself or save them. When using temporary chats, this may require copying and pasting or screenshot in the chat. It is helpful to do this even if the tool answers incorrectly, as the follow-up conversation can be helpful.
- **Use multiple generative AI tools if possible.** Different tools are good at different things. If your institution has a license for your students to use a particular tool, then maybe start with that. Try to include at least one freely accessible tool in your test.
- **When using multi-modal tools, experiment with more formats.** This may include copy-pasting text, screenshot (i.e., image), or a combination of both. Try to simulate behaviour of a student who wants to put as little effort as possible.

Part C. Analysing the output

- **If the assignment generates free text, use a range of student generated work for comparison.** Try not just to mark the GenAI product by itself.
- **Blind marking.** If the assignment uses free text, like an essay, then it really helps to ask other people to mark it, without telling them whether the text has been generated by GenAI.
- **Compare the performance of the tool vs both the pass-mark and the average score, and distribution, of test-takers.** The performance of a GenAI tool can really only be meaningfully interpreted by comparing to these metrics. As an example, if a GenAI tool scores 70% on an assessment, then the implications of this performance will be very different if the average score achieved by human test-takers is 90%, vs if it is 50%.

Part D. Reporting the output.

- **State which generative AI tools, which models and which versions you are using.** Each tool is, basically, a chatbot that is powered by an underlying large language model (LLM). These models have evolved very, very quickly, and many tools now offer users a choice about which model to use. For example, ChatGPT currently gives users a choice of models: GPT-o1, GPT-4.5 etc. However each of these models may also get updated (i.e. a 'version'), even if the model name itself does not change. If you don't know how to check the version, don't worry, record the model name, and the date you run the test (see below).
- **State the date upon which any tests were undertaken.** This can allow for the reader to gather more information regarding versions and updates.
- **State the interaction mode,** i.e., whether you interact with the tool via a web browser, through the app or via API.





References

1. Mollick E. The Homework Apocalypse [Internet]. 2023 [cited 2023 Aug 7]. Available from: <https://www.oneusefulthing.org/p/the-homework-apocalypse>
2. Hicks MT, Humphries J, Slater J. ChatGPT is bullshit. *Ethics Inf Technol*. 2024 Jun 8;26(2):38.
3. Newton PM, Jones S. Education and Training Assessment and Artificial Intelligence. A Pragmatic Guide for Educators. *Br J Biomed Sci*. 2025 Feb 5;81:14049.
4. Newton P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assess Eval High Educ*. 2024;0(0):1–18.
5. Newton PM, Summers CJ, Zaheer U, Xiromeriti M, Stokes JR, Bhangu JS, et al. Can ChatGPT-4o Really Pass Medical Science Exams? A Pragmatic Analysis Using Novel Questions. *Med Sci Educ* [Internet]. 2025 Feb 4 [cited 2025 Feb 28]; Available from: <https://doi.org/10.1007/s40670-025-02293-z>
6. Freeman J. Student Generative AI Survey 2025 [Internet]. Higher Education Policy Institute; 2025 [cited 2025 Mar 11]. Available from: <https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025/>

Version of the document: 140325

Cite as:

Newton, P. M., & Foltýnek, T. (2024). *A pragmatic protocol for stress testing assessment formats against GenAI* (Version 140325). [Manuscript]. Available at: <https://www.academicintegrity.eu/materials/393>



European Network
for Academic Integrity

