Imperial College Business School | IDEA Lab

Imperial means Intelligent Business

GenAI stress testing

Dr Cloda Jenkins
Associate Dean of Education Quality

Sean O'Grady
Lead Learning Designer

Peter Atkinson
Senior Learning Designer

Lok Yee Liu & Sophie Talbot
Editorial Team

Last updated: 7 August 2024

Lessons learned for assessment design

The **GenAI stress testing**, led by the IDEA Lab and Education Quality, stress-tested Autumn term 2023 module assessments, including assignments, quizzes and programming tasks, across Imperial College Business School.

Using a range of AI tools, we tested what output students could obtain if they asked AI to complete module assessment tasks. Testers evaluated each task for accuracy, clarity, relevance, compliance, referencing and ease of use, documenting all AI interactions. Results were reviewed by the IDEA Lab's learning designers, and Dr Cloda Jenkins, who calculated average scores and assigned risk levels (low, moderate, high). This process offered insights into AI-tool usage and strategies to mitigate AI interference.

In this document, we share our findings about:

- What AI can do well
- What AI struggles with
- How assessments can be redesigned to mitigate against misuse of AI.

A key discovery has been that, outside of an in-person, invigilated exam, all assessment tasks can be at least partially completed by AI with varying levels of accuracy and quality. This guide offers suggestions for redesigning assessments, ranging from minor changes that can be adopted immediately to more significant changes which may require approval.

For this iteration of testing, the following terminology and AI models were used:

- **AI tools** – the primary tools used were ChatGPT-4, Bard, and Claude 2, though testers could use their preferred AI tools for additional insights.
- **More advanced models** – referring to ChatGPT-4 and ChatGPT-4o.
- **Hallucinations** – fabricated or incorrect outputs generated by AI.
- **Writing tasks** – essays, strategies, plans, solutions, evaluations, case study-analyses, research proposals, research projects, peer reviews, project write-ups, reflections, and literature reviews.

If you would like to submit a module to undergo GenAI stress-testing or discuss assessment redesign with a member of our team, scan the QR codes below.

Request a stress test

Assessment design consultation

For further details or collaboration requests please contact: **s.ogrady@imperial.ac.uk** and **p.atkinson@imperial.ac.uk**.

Contents

1. What AI can do well

1.1 Across assessment types

Generate ideas
Students can prompt AI tools to generate initial ideas for their assignment. For example, they might ask AI to suggest companies, problems, solutions, etc. related to the topic of their assessment.

Transform notes into prose
AI can take students' notes about an experience or a task/project and write them out in full sentences and paragraphs. More advanced models will allow students to upload a document(s) that contains content such as project notes, meeting minutes and other information related to an experience or task/project and will produce text based on this content.

Conduct research
Some AI models, such as GPT4/4o, can access the internet and can be prompted to find relevant and timely sources about a topic. AI is much less likely to generate fake sources than it was in the past. Students can also use dedicated research AI tools, such as Scite, to find links to real, relevant sources. Such tools will usually default to more generic online texts or articles but can be prompted to find peer-reviewed literature with some success. However, AI-generated sources may include made-up citations, so students must be vigilant and check all sources for authenticity.

Structure assignments
AI can create an outline or structure for student assignments, with headings, sub-headings and indicative content, following a given reflective framework if required.

Analyse data and other sources
Students can upload data or other information required in their assignment, and more advanced AI models will perform a range of analyses and even produce graphs, charts and other visuals. AI performs better with smaller data sets (e.g. <100 rows of data) and discrete tasks.

Create images and visualisations
Some AI models can create basic images, and visuals such as tables, flowcharts and diagrams, which can be incorporated into presentations, reports, etc. More advanced AI models can perform a wide range of data analysis techniques to produce visualisations such as tables, graphs and charts. The accuracy of visualisations tends to be high with small to medium-sized data sets and where data is not particularly complex.

Suggest companies, products or other entities
AI can suggest companies, products or other entities to base the assignment on. It will usually default to very well-known examples but can be prompted to suggest lesser-known alternatives.

Suggest strategies and solutions
AI can propose specific strategies or solutions for an assignment provided it is given enough information, such as a description of the company or problem provided in the assessment brief.

Summarise/analyse the contents of a document
Many tools can analyse a document (PDF, Word, PPT, CSV) uploaded by the user, and summarise its contents. More advanced AI models can generate a text that reviews a document's main themes and arguments, and which critiques the quality of the work against a given framework.

Suggest improvements to a document
Based on its own analysis, or analysis provided by a student, AI can suggest ways to improve or further develop a piece of work that has been uploaded as a document.

Supplement information provided
AI can take limited information provided by a student (such as notes or summaries) and invent further details (for example 'lessons learned') that appear genuine. More advanced models can convincingly imitate a student by generating reflections that are highly relevant to the actual project or experience being discussed.

1.2 Presentations (live and video)

Write scripts
AI can produce scripts for each section of a presentation based on indicative content and can refer to specific sources if prompted to do so. AI tools require effective prompting to produce high-quality scripts.

Create slides
AI tools such as Gamma can take students' ideas, or those generated by another AI tool, and create professional-looking slides with text and images — for example, a student could create a script using ChatGPT and then ask Gamma to create the corresponding slides. AI-generated slides tend to be simple, with bullet points and generic images, and will need to be refined by the student to feature more complex information and visuals.

Edit videos
Students may use AI tools to edit their videos, for example to increase or decrease the speed of a video so that it fits within or reaches the time limits of the presentation assignment.

Prepare for Q&A
Students can ask AI tools to suggest potential questions that may come up in the Q&A following their presentation, and subsequently to evaluate their responses and provide feedback to help them prepare for spontaneous questions.

1.3 Quizzes ('take-home' and in-class)

Help students prepare for quizzes
AI can create sample questions to help students prepare for quizzes — although, there may be significant variation in the level of difficulty between the questions. If a quiz assessment is closed-book, AI can help students to prepare ahead of the quiz by creating sample questions, marking their answers and providing feedback.

Produce/provide correct answers
Most AI tools are able to provide correct answers to multiple-choice questions that are based on established concepts and theories, and discrete calculations. If the quiz assessment is open-book, students may use their device to copy and paste the questions into an AI tool, or they may upload a screenshot of them, and the AI tool will work its way through each question in turn.

More advanced AI models will also produce correct answers to other question types, such as open-text questions and 'fill-in-the-blanks' questions, where there is some flexibility in terms of acceptable answers (i.e. where correct answers can comprise different words or phrases).

1.4 Data analysis tasks

Read and analyse data sets
AI models (including some freely available ones) allow students to upload a data set (depending on the file format), and the AI models will read the data, provide an accurate summary of the type of data available and perform basic data analysis. This analysis is generally accurate for small- to medium-sized, discrete data sets where the data is not particularly complex, but not so much for larger, more complex data.

Perform specific data analysis techniques
More advanced AI models can perform a range of specific data analysis techniques as requested by the student, with the aforementioned caveat that data size and complexity affect accuracy.

Produce written analysis and evaluation of data
AI can generate written analysis and evaluation of data following its analysis, but hallucinations can occur in its discussions of large and complex data sets.

Produce data visualisations
More advanced AI models can take a data set and produce a range of data visualisations, such as charts and graphs, with the aforementioned caveat that data size and complexity affect accuracy.

1.5 Maths tasks

Perform discrete calculations
AI can generate solutions for one-off maths problems and calculations, with more advanced AI models offering increasing accuracy.

Work with maths images
Some advanced AI models perform well when provided with an image (e.g. a screenshot) of a maths problem or set of problems, making the process of using AI to solve calculations even easier for students.

Explain how to approach a maths problem
AI can accurately describe the steps needed to find a solution to a maths problem, although it will not always follow the same approach that has been taught in the module.

1.6 Coding tasks

Break down a coding problem
AI can break down a programming problem into its constituent parts and explain the necessary requirements and steps to solve it.

Write code
AI can write code in various languages. More advanced models can produce accurate code and identify bugs if the code does not work in the first instance. Accurate AI-generated code tends to be limited to smaller, discrete coding tasks. The number of errors in the code increases with more complex tasks and projects.

Execute code
More advanced AI models can execute Python code, for example, for some limited tasks, or produce the necessary files to run and execute the code in another environment.

1.7 Writing tasks

Write individual sections
AI can generate text for individual sections of a written assignment. With effective prompting, it can write to an approximate word count and include specified ideas, data and references.

The best results are produced when a student asks the AI tool to first plan and create an outline for a section of the assignment, and the student then provides feedback on this outline before asking the AI tool to produce a first draft of the section. The student can then ask the AI tool to refine what it has written through further prompting (e.g. “mention this source in paragraph two”, “add a specific example to support the argument in paragraph three”).

Analyse literature
Most AI tools can analyse an individual text and summarise the main arguments. More advanced models can analyse a limited number of texts (two or three) simultaneously, and compare, contrast and synthesise information. AI can also create a limited literature review that focuses on each text provided, or a limited number of texts, in turn.

Suggest research topics or methods
AI can generate research topics for students to explore, ranging from broad themes to more specific research questions — although, it is unlikely to suggest anything completely original. AI can also make suggestions for research methodology suitable to specific research questions, aims and objectives.

Write aims and outcomes
If the student provides sufficient context about their research topic and methodology, AI can generate or refine research aims and objectives.

Create a project plan
AI can create a plan for a research project with realistic deadlines for each stage of the project. More advanced models can produce visuals such as Gantt charts to visualise the project plan.

2. What AI struggles with

2.1 Across assessment types

Generating specific ideas
Generally, AI can produce high-level analysis of advantages, disadvantages, problems, solutions and other aspects of the company/product/case in question, but it struggles to provide more detailed analysis unless the company/product/case is well known or where a lot of information is publicly available.

Suggesting original topics and research questions
AI is likely to suggest topics and research questions that have already been addressed elsewhere. It can produce more specific topic suggestions with effective prompting but struggles to generate anything completely original.

Providing sufficient detail
Current AI tools tend to produce limited word counts and deal with topics at surface level, without moving into more detailed analysis. Unless students provide substantial notes, AI will fail to discuss specificities and will instead focus on a higher-level overview. As a result, AI outputs are usually sufficient for brief, low-stakes assessments, but for more in-depth assignments that require considerable detail and analysis, a student would have to put in a lot of effort to bring their AI-generated response up to an acceptable standard.

Maintaining cohesion across multiple sections
Maintaining cohesion across multiple related sections of an assignment is also an issue for current AI tools, as they tend to forget information and instructions from earlier on in the task. This exacerbates the issue of AI tending to deal with topics in a superficial manner that lacks critical analysis.

Incorporating module content and experiences
When asked to refer to specific module materials and experiences (such as lectures, texts, simulations), AI often provides generic outputs that lack sufficient detail or analysis — it struggles to synthesise multiple sources in a meaningful way. AI may incorporate a small number of specific module sources in its response, such as an article or lecture transcript, but it fails to capture the richness of learning that results from live sessions and peer-to-peer interactions, such as group tasks or forum discussions. AI's responses could be improved, however, if students collate and upload all the relevant module materials, including their personal notes and reflections.

Note that while AI tools can offer generic feedback about a document that a student has uploaded, the tools will not suggest ways to better incorporate specific module content, such as concepts, readings or other material.

Incorporating personal experience and reflection
AI is able to generate pieces that make reference to students' personal experiences, if students provide these, but reflections produced by AI tend to be generic, as the AI does not have access to students' backgrounds, beliefs and goals, which would lead to richer reflections.

Currently, AI tends to focus on a small number of directly related experiences provided by students but is less able to draw on more numerous, complex and less obviously connected events. However, more advanced models are getting better at faking meaningful reflections and finding themes across multiple unrelated experiences provided by students.

Citing high-quality academic literature
While AI can be prompted to find relevant and timely sources, including peer-reviewed academic literature — using dedicated tools such as Scite, for example — it often misses key sources, which can easily be found in a library search, and tends to revert to using generic, sometimes outdated, online sources. Even when AI tools are prompted to use only real texts, hallucinations can still occur, requiring students to be diligent. Students who rely on AI tools for research purposes are likely to miss out on some of the most relevant and high-quality sources.

In addition, while AI tools can do a good job of discussing one specific text in a particular section of an assignment, they struggle to synthesise multiple sources and use them across the piece.

Demonstrating advanced critical analysis
AI will often provide generic, surface-level analysis and struggle to engage in deeper, critical analysis. This is especially true when it is faced with novel or lesser-known cases or examples where it cannot access pre-existing analysis e.g. of famous case studies and large, multinational companies.

Unless AI is provided with detailed information about the project, such as minutes from meetings, it will default to generating generic ideas that display a lack of familiarity with the actual student experience.

Performing advanced data analysis
When AI is provided with large amounts of complex data (for example, a large number of interviews or survey responses), its data analysis and resulting visualisations tend to be less accurate compared to its analysis of smaller, less complex data sets.

Using data consistently
AI may correctly incorporate data provided by the student (or retrieved online) in its response, but it begins to make errors when it is tasked with using the same data multiple times. For example, its interpretation or analysis of the same data may vary from instance to instance in the same piece of work.

2.2 Presentations (live and video)

Deep-faking students' voices and images
Various AI tools, such as HeyGen and Synthesia, allow users to clone their voice and image to create a 'digital twin', which can then deliver a script. Using the tools currently available to most students, the resulting videos would clearly be AI-generated, due to unnatural gestures and lip movement, and accent inconsistencies. However, the performance of these tools is improving rapidly, and they may soon produce passable results, i.e. where a video appears to show a student, but it is actually their digital twin or avatar.

2.3 Quizzes ('take-home' and live)

Multi-part and maths questions
Presently, most AI tools will produce some incorrect answers to complex, multi-part questions, particularly those that involve maths. This is especially true where the answer to one question forms part of the next question to create a series of related calculations. In this situation, the AI may start by producing correct answers but lose track along the way.

It is worth noting that newer and more advanced AI models claim to be much better at correctly answering maths questions than earlier models.

Answering more complex question types
AI tools may also perform less well in response to open-text and certain other question formats (as opposed to multiple-choice questions), and they perform unreliably when asked about specific texts or sources.

Answering questions on novel or lesser-known content
The ability of AI tools to correctly answer questions deteriorates when they are faced with content they have not seen before, such as questions on specific module content or bespoke problems/cases.

2.4 Data analysis tasks

Working with large and complex data sets
AI will often fail to engage with particularly large and complex data sets (e.g. >300 rows). Even if it is 'willing' to do so, the accuracy of its analysis and resulting visualisations will be unreliable.

Performing advanced data analysis techniques
There are some data analysis techniques that AI tools cannot perform. In these instances, the AI will offer explanations and suggestions for how to perform the analysis with other tools.

Using taught approaches and techniques
AI may perform a data analysis task but in a way that differs from how students have been taught in their module.

Producing detailed analysis and evaluation of data
AI tends to provide reasonable high-level overviews of data, but when asked to go into more detail it struggles to produce accurate and reliable analyses and evaluations. In such cases, hallucinations and contradictions can occur, especially with large and complex data sets.

2.5 Maths tasks

Dealing with novel problems that resemble well-known problems
Where a mathematical problem resembles one that is well known, that the AI model is very familiar with, the AI model will often produce an answer to the widely known problem, not the specific one it has been given.

Performing complex, multi-part calculations
AI can easily make mistakes when asked to perform complex, multi-part calculations where there are multiple dependent variables. Often, the AI will use the correct approach but make a mistake at some point in the process, which then impacts all subsequent calculations.

2.6 Coding tasks

Solving problems using taught methods
AI may produce a working solution to a coding problem but will often use unconventional methods which differ from those that students have been taught.

Writing code for larger, more complex tasks
More advanced AI models will attempt to create code for larger tasks and projects, but accuracy varies across different programming languages. In general, performance deteriorates as task complexity increases.

2.7 Writing tasks

Maintaining detail and cohesion across multiple, related sections
As the word count and complexity of a written assignment increase, AI struggles to provide sufficient detail and maintain cohesion across different sections in its response. The AI-generated text may deviate from the agreed outline in later sections and begin to introduce ideas that do not fit with what came before. For example:

- AI's proposed solutions to a problem may not align with or address certain aspects of the problem, or the AI tool may introduce new aspects of the problem that were not previously mentioned.
- Evaluations that feature later in the text may not align with earlier discussion and analysis.
- AI's discussions of a research methodology may contradict the research aims and objectives, and its analysis of findings may not fully align with the actual results.
- An AI-generated literature review may present contradictory ideas without acknowledgement or explanation.

Synthesising information from multiple sources
AI can analyse and synthesise information from a limited number of sources (two or three), but when it is asked to analyse a larger number of sources (i.e. perform an entire literature review) its analysis tends to be very superficial, and the number of hallucinations increases.

3. Ideas for preventing AI misuse

3.1 Across assessment types

Change criteria/weighting
If an assignment task is easily outsourced to AI, consider reducing the weighting of the task. Adjust the assessment or marking criteria to place more emphasis on elements/tasks in the assignment that can be less easily produced by AI, such as synthesising information, critical analysis, originality of research topic and question, and analysis and synthesis of peer-reviewed academic literature.

Modify task
Modify assessment tasks so that they cannot be easily outsourced to AI. For example:

- Require students to explain their process and/or submit a personal reflection on any challenges they faced and how they overcame them.
- Require students to refer to specific module content and/or aspects of the learning experience (such as lectures, readings, people, tools, problems and solutions).
- Design a task based on novel or lesser-known companies, products or other entities.
- Design a task that requires consistent and cohesive linking across multiple, related sections.
- Require students to carefully log assignment progress, including notes and minutes/transcripts from meetings, and submit details of this alongside their assignment.

Generally, assessments where there are multiple sections that require detailed critical analysis and information synthesis are less easily outsourced to AI.

Include a live/spoken element
To reduce the risk of AI misuse in assessments, incorporate a live, spoken element, such as a presentation or discussion, where students discuss their assignment and respond to questions from faculty and/or peers.

Change assessment format
Consider switching to an assessment format less susceptible to AI misuse, such as an in-class quiz, exam, in-class participation or live discussion (providing the learning outcomes can still be achieved).

Change assessment topic/focus
Replace generic topics, or topics on which information is readily available, with more specialised ones, or require students to incorporate their own experience and reflections as well as specific module content in their answers.

3.2 Presentations (live and video)

Focus on Q&A
Depending on the module's learning outcomes, the assessment criteria for live presentations should focus more on aspects of the presentation less easily outsourced to AI, such as performance in the Q&A section.

If it is practical, consider making a video presentation a live presentation with a Q&A section. Alternatively, consider adding a live Q&A component to a video presentation — i.e. students first submit their recorded video presentation and then meet online or in-person to answer questions from faculty and/or peers.

3.3 Quizzes ('take-home' and live)

Redesign questions
If your quiz primarily consists of multiple-choice questions and tests knowledge of established concepts, theories or other publicly available information, consider alternative question types (e.g. open-text) and include questions that refer to specific sources.

Include multi-part questions, where the answer(s) to one question forms part of another. Ask students to explain their thinking/workings in addition to providing the correct answer.

3.4 Data analysis tasks

Increase complexity of data analysis task
Require students to use large and complex data sets.

Focus on process
Require students to explain and demonstrate the process they underwent to perform the data analysis.

Include an evaluation
Require students to incorporate a detailed written evaluation of the data and their resulting analyses.

3.5 Maths tasks

Increase interrelatedness of maths task
Include tasks/questions that comprise multi-part, related calculations, where the answer(s) to one question forms part of another.

3.6 Coding tasks

Increase complexity/interrelatedness of coding task
Replace individual, discrete coding tasks with a series of more complex, related tasks.

3.7 Writing tasks

Include a reflection task
Require students to complete a personal reflection about the process of writing their research proposal, project, literature review, etc.

Authors

Author	About
Name: Dr Cloda Jenkins **E:** c.jenkins@imperial.ac.uk	Dr Cloda Jenkins is Associate Dean, Education Quality and a Professorial Teaching Fellow in the Department of Economics and Public Policy in the Imperial Business School. She is Head of Year for BSc Economics, Finance and Data Science. Prior to joining Imperial, Cloda was a Professor (Teaching) in the Department of Economics at University College London where she taught students from 1st year BSc through to MSc. Cloda is an expert in regulatory economics and mechanism design. She has applied this expertise to practical policy making, having advised organisations in a range of regulated sectors since 1997, including heading Ofgem's Review of Energy Network Regulation (2008-2010), sitting on the Expert Panel of the UK Regulatory Network and providing expert advice to the water regulator for England and Wales (Ofwat). She is also an expert in economics education, with particular interest in developing research-based education, developing employability skills in economics degrees, designing authentic assessments, adapting economics curriculum to make education resilient to shocks and improving student learning and well-being by working in partnership with students. As Associate Director of the Centre for Teaching and Learning Economics (CTaLE), member of the RES Education and Training Committee and member of the EEA Education Committee she is passionate about sharing lessons with peers globally to improve the evidence base for enhancements in economics education.
Name: Sean O'Grady **E:** s.ogrady@imperial.ac.uk	As Lead Learning Designer in the IDEA Lab, I collaborate with faculty, academic and programme management and media to design and evaluate online programmes within the business school, primarily the Global Online MBA (GMBA). I explore how technology can be used to create impactful learning experiences while maintaining a pedagogy first approach to my work. My background is primarily in teaching as well as academic management and staff development. I hold a masters degree in Education and Technology from the Institute of Education (IOE) at UCL.
Name: Dr Peter Atkinson **E:** p.atkinson@imperial.ac.uk	I have been working as a Senior Learning Designer in the IDEA Lab within Imperial's Business School since September 2023. I work primarily on the Business School's online programmes, including Strategic Marketing Online and the Global MBA. I am interested in using emerging technologies and evidence-based pedagogy to create effective and engaging online learning experiences for students around the world. Prior to working at Imperial, I worked in online learning at The Open University and also worked for a number of years in higher education publishing. I have a Postgraduate Certificate in Online and Distance Education from The Open University, and a PhD in Musicology from the University of Birmingham. I also have experience teaching in HE.