How to Build AI Resilience in Assessment

This resource was produced by the School's working group on AI and Academic Integrity (a sub-group of the LSE Departmental AI leads group, convened March–November 2025). It will be reviewed on a regular basis.

What is AI Resilience?

AI resilience is the ability to maintain meaningful assessment and other educational practices given the risk of inappropriate use of AI technology. An AI-resilient assessment is one that either makes it difficult for students to misuse AI or enables educators to verify the authenticity of student work by making AI use evident. Importantly, AI resilience is a matter of degree rather than an all-or-nothing property. No assessment is perfectly immune to AI misuse, just as no assessment has ever been entirely immune to other forms of academic misconduct. The goal is to reduce AI misuse through means that are pedagogically appropriate and feasible within workload constraints.

Since the emergence and rapid development of generative AI in November 2022, LSE has sought to balance two strategic priorities: maintaining the academic integrity of its assessments and ensuring that students develop the skills to work effectively in an AI-enabled world. In advance of each academic year, all academic departments must agree department-wide or module-level policies on the authorised use of generative AI in assessment within one of the following positions:

- Position 1: No authorised use of GenAI in assessments
- Position 2: Limited authorised use of GenAI in assessment, and
- Position 3: Full authorised use of GenAI in assessment.

In June 2025, a School position on Observed Assessment was introduced to strengthen these positions. It includes ideas on how staff can better protect academic integrity given concerns about AI misuse. This document offers further guidance with practical steps for building AI resilience into assessment design.

When is AI Resilience Needed?

Across different disciplines and modules at LSE, there is a diversity of policies regarding AI use. This diversity reflects legitimate differences in pedagogical aims, disciplinary norms, and the specific learning outcomes that assessments are designed to measure. Despite this diversity, some form of AI resilience will often be valuable, including in contexts where AI use is not prohibited. For instance, when AI is permitted for some tasks but not others, AI resilience can encourage students to use AI only for permitted tasks. If AI is permitted under a "use and declare" policy, AI resilience can discourage students from using AI *without* declaration. In a large range of cases then – from prohibition to regulated use – AI resilience serves to maintain the trustworthiness of assessment and ensure that students develop the skills the module is designed to cultivate.

Building effective AI resilience requires careful thought and planning. There are no quick fixes. Some may assume that AI-generated work can be reliably identified using the AI detection tools available online. However, the accuracy of these detectors remains a matter of ongoing debate and research. AI detectors produce both false positives (flagging human-written work as AI-generated) and false negatives (failing to identify work that was in fact produced by AI). For these reasons, it is important not to rely solely on AI detection software to defend academic integrity against AI misuse. Effective AI resilience will likely require a range of methods.

AI resilience measures should be implemented alongside clear rules and guidance for students. Departments have a responsibility to help students understand why achieving learning outcomes through their own work matters. Programme directors, course convenors, and department AI leads should ensure students receive consistent messaging about appropriate AI use and understand relevant policies. Developing students' AI literacy and critical understanding of these tools is essential to reducing unauthorised use.

Prevention is Better than Detection (but the Two are Related). Early Detection is Better than Later Detection.

Prevention is preferable to detection when it comes to AI resilience because there are significant costs, both to students and to staff, of pursuing AI academic misconduct investigations. (See new staff guidance on academic misconduct in the case of suspected unauthorised use of AI). That said, prevention and detection are interrelated. Preventive measures often make genuine learning achievement visible and verifiable, rendering AI misuse easier to detect. Conversely, the knowledge that effective detection mechanisms are in place can itself serve as a deterrent. A robust approach to AI resilience will typically combine elements of both.

Just as prevention is preferable to detection, early detection is preferable to later detection. When AI misuse can be identified early, from drafts and other formative work, educators can intervene before summative assessment and avoid the costs and stress of formal misconduct investigations for both students and staff. Formative work and drafts are typically not subject to academic misconduct penalties. The evidence threshold for having a supportive conversation with a student during the work process, to clarify the rules around AI use and guide them toward appropriate practices, is lower than the evidence required to pursue allegations of academic misconduct after summative work been submitted. These considerations provide educators with additional reason, beyond the pedagogical benefits, to include opportunities for formative work, draft submissions, and feedback ahead of summative assessment.

How to Choose a Method of AI Resilience

When deciding which method of AI resilience to adopt, educators must balance three distinct but interconnected goals:

1. **Pedagogical suitability**: The assessment should be the right kind for the module, aligning with the learning outcomes of the course and designed inclusively to ensure equitable access for all students.
2. **Feasibility**: The method of AI resilience should be manageable in terms of staff workload, avoiding excessive demands on time in design, grading, oversight, or enforcement.

3. **Protection of academic integrity**: The approach should safeguard against inappropriate AI use, ensuring that students' work genuinely reflects their own learning and capabilities.

It is worth noting that all three of these goals are, in an important sense, student-centred. Academic integrity matters profoundly for students because failure to uphold it undermines the credibility of their qualifications. Similarly, pressures on staff time and workload have direct effects on students. The more time educators spend on policing and enforcing academic integrity, the less time is available for other aspects of teaching and support of student learning. **Pursuing AI resilience, therefore, is not about trading off academic integrity against what is best for students, but rather about achieving the best for students by pursuing the right balance of these three goals.**

Ideally, introducing greater AI resilience into assessment would be win-win-win: the new method of assessment chosen would be more pedagogically appropriate, less demanding of staff time and more effective in protecting academic integrity than the previous form of assessment. Often, however, difficult choices will have to be made. In making these choices, staff will need to exercise their best judgment about how to balance these competing considerations in light of the specific context of their module.

Five Methods of AI Resilience

There are many approaches to AI-resilient assessment, but these tend to fall within five overall categories. Here, we outline the five categories and note their possible advantages and disadvantages. These methods of AI resilience are not mutually exclusive. Educators may wish to combine features from various methods to suit their module learning objectives.

	Details	Advantages	Disadvantages
1. In-Person Assessment	Written exams, oral exams, presentations and in-class quizzes.	• Probably the *most* effective method of protecting academic integrity.	• Exams tend to be unpopular with students. • May increase the need for reasonable adjustments. • Depending on the discipline, may not be the best method to test student learning outcomes.
2. Edit Tracking	Students work within an LSE-supported platform, such as Cadmus, that is purpose-built for assessment integrity, tracking student edits as they type.	• Copying and pasting of AI-generated work is likely to be detected. Typing up AI-generated work might also	• Requires specialised software and training. • With sufficient time and effort, students could potentially

	Version history in Google Docs, OneDrive/SharePoint, or Git/GitHub (for coding-based assessments) may offer a partial alternative, though these tools capture less detail and are more cumbersome to review.	be detected and the costs and risks of doing so may deter AI misuse. • Compatible with 'take-home' assessment methods such as essays. • In student surveys, edit tracking has proved more popular than exams. • Provides students with security against false accusations of AI misuse. • The specified platform might have its own advantages e.g. Cadmus integrates additional tools and resources for students to support them in the development of their assessments.	type up AI-generated work to make it appear genuine. • If using Cadmus, or similar software, students cannot work in familiar platforms such as MS Word. • While edit tracking does not track student activity outside of the platform, it does record how students work within the platform. Some students may object to being monitored in this way.
3. Vivas	After students have completed their assessment, students are invited to discuss their work. The	• Compatible with 'take-home' assessment formats such as essays.	• Students may be able to effectively fake authorship of AI written work during a viva. Gaps in

	interviews themselves could be graded or ungraded. Educators could potentially spot AI misuse if students seem unable to explain their written work.	• Does not require significant changes to assessment, nor specialised training. • May have a deterrence effect even if detection is difficult.	understanding could result from students over-reaching themselves in written work rather than AI misuse. Even when student responses raise red flags, vivas may fail to provide sufficient proof of AI use in the absence of other evidence. • Significantly time consuming if every student is interviewed. Possible variant: only a selection of students receive vivas. See School guidance on conducting selected interviews.
4. Supervised Assessment	Students meet regularly with an educator to discuss their work in progress. While this model is typically reserved for dissertations, it could be extended to other assessments. Students could be graded on final work or interim submissions throughout the course. In supervisions, educators could potentially spot AI misuse if students seem unable to explain their written work.	The same advantages as vivas plus: • Potential opportunities to spot AI misuse early. • Significant pedagogical benefits: students receive feedback during the working process. • Students value contact time with educators.	The same disadvantages as vivas: • Significantly time consuming. • May fail to provide sufficient evidence of AI misuse.

5. Task Requires Human	Three kinds: 1. Secret Knowledge. Students are asked to refer to knowledge that LLMs do not have access to and cannot be easily uploaded, such as an in-class conversation. 2. AI-Resilient Medium. Assessment involves a medium that AI cannot currently complete e.g. a 'vodcast' episode in which students interview each other about a topic. 3. Participatory Projects. To complete assessment, students are required to engage in some form of face-to-face interaction, such as consultancy work, placements, or client-facing projects.	• Effective at *limiting* AI use. (Students might still be able to use AI for some tasks). • Compatible with 'take home' assessment. • Likely to be more popular than exams.	• Not appropriate for every course. • Limited shelf-life as technology advances.

The Importance of Referencing Standards

In addition to the above five methods, it is worth noting that insistence on rigorous referencing, including page references for quotations and specific claims, can provide an additional form of AI resilience. The strongest evidence of AI academic misconduct often comes from errors in referencing: confabulated sources, incorrect page numbers, quotations that do not appear on the cited pages, and so on. When students are not required to provide page numbers or when referencing errors carry no penalty, it becomes easier for students misusing AI to avoid detection. For this reason, departments and module leaders might consider imposing a modest penalty for referencing errors on all assessed work. Such a penalty not only incentivizes good scholarship but also creates a practical deterrent to AI misuse. The additional work required to verify references should reduce the temptation to cheat.

What next?

For further guidance on assessment design, see the LSE Assessment and Feedback Toolkit.

Information on School policy and practice relating to generative AI in education is available on the Eden Centre website.

For further consultation, contact your departmental AI lead or your Eden Centre departmental adviser.