



# Rethinking Assessment for Generative Artificial Intelligence

---

Leon Furze



# INTRODUCTION

I've been writing about Generative Artificial Intelligence in education since 2022, when I began my PhD in how these technologies are perceived and used by educators. Prior to that, I was a secondary educator and school leader for fifteen years. My experience ranges from Head of English, to Year Level Coordinator, and Director of Teaching and Learning. I've also worked in Higher Education, mostly in Initial teacher Education courses teaching literacy. Since about 2020, I've been experimenting with digital technologies like OpenAI's GPT, and wondering about its impact on education.

In May 2022 when I first applied for my PhD, my supervisor and I discussed studying GPT and other "Automated Writing Systems" (AWS) as an interesting but "on the horizon" idea. In November, my PhD officially started. Two weeks later, OpenAI released ChatGPT. The conversation the next day with my supervisor was "this isn't on the horizon any more - it's here, now."

If you work in K-12 or Higher Education, you'll have experienced some of the fallout firsthand. Primarily, the narrative in education has been concerned with the use of GenAI by students to "cheat" on essays and other assessment tasks. I think we need to move beyond that narrative, and fast. Large Language Models and other forms of Generative AI are not just tools that students are using to write quick-and-dirty essays. They're not going to "Kill High School English" or be the "Death of the Essay", but nor are they something we can ignore.

This eBook is a collection of my recent articles on the topic of rethinking assessment. I hope you find it useful in shifting those narratives.



Leon Furze

[www.leonfurze.com](http://www.leonfurze.com)

# TABLE OF CONTENTS

**1** Generative AI, plagiarism, and “cheating”

---

**2** AI Detection in Education is a Dead End

---

**3** The AI Assessment Scale: Update and Pilot Study

---

**4** Ditch the Detectors: Six Ways to Rethink  
Assessment for Generative Artificial Intelligence

---

**5** Rethinking Assessment for Generative AI: Beyond  
the Essay

---

**6** Rethinking Assessment for Generative AI: Orals and  
discussions

---

**7** Rethinking Assessment for Generative AI: Ungrading

---



# GENERATIVE AI, PLAGIARISM, AND “CHEATING”

Back in January, I wrote a post called [Beyond Cheating](#), reflecting on the ChatGPT bans that were rolling out across various Australian states and the “cheating” narrative that had accompanied the chatbot since its release.

In that earlier post, I argued that banning and blocking generative AI would only contribute to the digital divide – students who have greater access to digital technologies would inevitably be able to access and use GAI\*, putting those who rely on in-school technology access at a disadvantage.

It’s almost been 12 months since the release of ChatGPT and, thankfully, the bans have now been lifted in most jurisdictions. Unfortunately, the narrative of “catching” students using GAI still persists, and educators in both K-12 and tertiary are still stuck in the loop of detection tools, pen and paper examinations, and proctoring software as methods to stop or monitor GAI use.

Whatever level of education you work in, this post is an attempt to convince you that trying to catch or detect GAI is futile. Not only that, detection tools and other plagiarism checkers may be unethical, and a punitive approach to GAI use is going to add to educator workload. This year I’ve worked in many different schools across states and sectors in Australia. I’ve fielded a lot of questions about GAI and assessment, so I’m going to present this post as a sort of FAQ. If by the end of the article you still think that detection is a viable option, I’d encourage you to get in touch via the contact form button on the left.



\*You’ll notice that I use both GAI and GenAI in this book. I have shifted over time to “GenAI” to avoid confusion with the acronym AGI, which stands for Artificial General Intelligence and is a related, but different concept.

# WHAT IS GENERATIVE AI?

---

I'll start with the broadest question, but I've written about this a lot in the past so it might be worth checking out some of my early 2022 and 2023 posts. In a nutshell, generative AI takes data as input, and produces new data as output. This might be text prompts (input) used to generate new text (output). Or, it could be text-to-image, image-to-image, text-to-audio, image-to-text, or any similar multimodal variations.

To understand some of the discussion that follows, it's necessary to wrap your head around the construction of these GAI models. Though approaches vary for text, image, and audio, they are all reliant on huge amounts of data.

For a text-based model like GPT, that includes data scraped from the web, Wikipedia, social media sites, and other datasets. However, these models are not search engines: when you type a prompt into ChatGPT, it doesn't search for a suitable answer or combination of answers. Instead, it uses the rules it has developed to generate a novel response. Occasionally, that response might be similar to, or the same as, material from the dataset. I'll get into that in more detail later.

# DOES GAI PLAGIARISE OR COPY?

---

One concern of these technologies has been that students using them are automatically plagiarising. This is based on the assumption that a model like ChatGPT "copies" its answer from the dataset.

In some respects, language models like GPT can generate responses which contain verbatim copies of text from the dataset. For example, in the following prompt I can easily get ChatGPT to tell me the opening line of a classic novel:



What is the opening line of pride and prejudice?



The opening line of "Pride and Prejudice" by Jane Austen is:

"It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife."

This line is one of the most famous opening lines in English literature, and it sets the tone for the rest of the novel, which explores themes of love, marriage, and social class.

*Prompt: What is the opening line of pride and prejudice? Model: GPT-4*

**It will also perform the same way for more recent books which are still under copyright, such as J K Rowling's Harry Potter and the Philosopher's Stone:**



What is the opening line of Harry Potter



The opening line of "Harry Potter and the Philosopher's Stone" (released as "Harry Potter and the Sorcerer's Stone" in the United States) by J.K. Rowling is:

"Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much."

*Prompt: What is the opening line of Harry Potter. Model: GPT-4*

**However, if you push this much further you'll get a response like the following:**



What is the second line?



Sorry, I can't provide verbatim text from copyrighted material. Would you like a summary of the second line instead?

That's a trained response: the model has been "taught" to respond with a comment like this when a user appears to be asking for something which might breach copyright. In other models without these guardrails, however, they can be prompted to provide verbatim responses which seemingly recall text from the dataset.

But does that mean that these models are "plagiarism machines"? It's a little more complex. These models work by analysing the dataset and learning the patterns of grammar, syntax, style, and so on. As a result of the process, data that is repeated more often (such as the opening lines of famous books, or facts available and commonly repeated online), [is more likely to appear in generated output](#). Companies like OpenAI put in place measures to limit this, but it certainly still happens.

So the answer to the question "does GAI plagiarise or copy?" is "sometimes, but not in the ways you might expect". It's therefore not possible to claim that a student's use of GAI equals plagiarism. If a student uses GAI to generate an essay, for example, then much of the response will be novel content, and any content that comes verbatim from the dataset is more of a side-effect than an intentional copy.

# IS USING GAI CHEATING?

---

This depends on your definition of cheating, and on the task.

A student using GAI to complete an entire task might be akin to contract cheating, where a person pays someone else to do the work for them. In fact, [ChatGPT might even reduce the amount of actual contract cheating and put the contract essay writers out of work](#). There's not much difference between paying someone to write an essay and dropping the entire question into ChatGPT to generate the response.

**The key factors in determining whether GAI constitutes cheating include:**

- 1 Whether the use of GAI is expressly forbidden
- 2 Whether the use is required to be disclosed
- 3 Whether there is a competitive advantage to be gained through the use of GAI

Essentially, "cheating" is whatever we decide it is. If an educator decides to ban GAI use, then of course any use is cheating. If a student uses the technology in a deliberately deceitful way, or to gain an unfair advantage, then it's cheating.

The problem, as I'll explain throughout, is that it's next to impossible to enforce strict anti-GAI policies. If you consider GAI use to be cheating, for whatever reason, you're going to have a hard time monitoring and evaluating student work outside of specific constraints.

# CAN'T I JUST USE DETECTION SOFTWARE?

---

Hot on the tails of ChatGPT, generative AI "detection" software started to appear in tertiary and then secondary education contexts. You can see the appeal. Just as a new technology arrives which threatens to hugely undermine assessment practices, a few helpful developers provide an easy way to catch students using GAI.

Unfortunately, detection software doesn't work.

Many studies have demonstrated that detectors like GPTZero and Turnitin in simply don't have the level of accuracy needed for an academic integrity judgement. For example, here are a few snippets from GPTZero:



### Was this text written by a human or AI?

Try detecting one of our sample texts:

ChatGPT GPT4 Bard Human AI + Human

But does that mean that these models are "plagiarism machines"? It's a little more complex. These models work by analysing the dataset and learning the patterns of grammar, syntax, style, and so on. As a result of the process, data that is repeated more often (such as the opening lines of famous books, or facts available and commonly repeated online), is more likely to appear in generated output. Companies like OpenAI put in place measures to limit this, but it certainly still happens.

490/5000 characters

Check Origin

Upload file

.pdf, .doc, .docx, .txt

By continuing you agree to our [Terms of service](#)



Our ensemble of detectors predict **different results** for this text. Please enter more text for more precise predictions. There is a **48%** probability this text was entirely written by AI

Sentences that are likely written by AI are highlighted.

But does that mean that these models are "plagiarism machines"? It's a little more complex. These models work by analysing the dataset and learning the patterns of grammar, syntax, style, and so on. As a result of the process, data that is repeated more often (such as the opening lines of famous books, or facts available and commonly repeated online), is more likely to appear in generated output. Companies like OpenAI put in place measures to limit this, but it certainly still happens.

Example 1: paragraph from this blog, written with no GAI. Flagged as 48% probability of being written by AI.

### Was this text written by a human or AI?

Try detecting one of our sample texts:

ChatGPT GPT4 Bard Human AI + Human

It's also worth noting that the goal of a language model is not to create plagiarized content but to assist users in generating human-like text based on the prompts given to it. Care should be taken when using text generated by a language model for academic or professional purposes, as it may require additional verification for originality.

737/5000 characters

Check Origin

Upload file

.pdf, .doc, .docx, .txt

By continuing you agree to our [Terms of service](#)



Our ensemble of detectors predict **different results** for this text. Please enter more text for more precise predictions. There is a **53%** probability this text was entirely written by AI

Sentences that are likely written by AI are highlighted.

That said, there is a possibility that the model could generate text that closely resembles text from its training data, purely by chance or because it's a commonly expressed idea. OpenAI has implemented various measures to reduce the risk of generating verbatim or near-verbatim text from the training dataset. However, the technology is not perfect, and there is ongoing work to improve it.

It's also worth noting that the goal of a language model is not to create plagiarized content but to assist users in generating human-like text based on the prompts given to it. Care should be taken when using text generated by a language model for academic or professional purposes, as it may require additional verification for originality.

Example 2: Paragraph generated entirely by ChatGPT. Flagged as 53% probability of being written by AI. Model: GPT-4





Extend it into a much longer response. Make the paragraphs longer and denser. Use more variation of long and short sentences.



Certainly, here's an elaborated and nuanced version with a mix of long and short sentences:



It's essential to delineate, in no uncertain terms, the nuanced risks involved when engaging with language models like this one. There's a non-negligible, probabilistic chance that the generated output could bear a strong resemblance to specific fragments within its broad and diverse training corpus. Such occurrences could manifest either by sheer serendipity or as a result of the model tapping into widely shared or culturally embedded ideas and expressions. OpenAI, the organization behind these language models, has been proactive in implementing a multi-faceted array of safeguards. These measures are intricately designed to substantially lower the risks of the model inadvertently generating text that could be categorized as either verbatim or near-verbatim replications of content found in its training data. However, it's crucial to acknowledge that this technological landscape is ever-evolving, and absolute infallibility remains an aspirational goal rather than a guaranteed attribute. Hence, continuous research and development efforts are being vigorously pursued to further fortify these already robust safeguards.

*Example 3a: Text from example 2 re-prompted with some basic instructions*

# Your text is most likely human written

**There is a 36% probability this text was entirely written by AI**

*Example 3b: Output to example 3a from GPTZero. Flagged as 36% probability of AI, down from 53%. Model: GPT-4*

As you can see above, the entirely human writing (mine) scored about the same as the entirely GPT written version. After a very minimal re-prompt, the GPT written text scored as “more human” than the human text.

**This is only a very brief example, and it’s worth checking out some of the emerging studies such as:**

- 1 This study which found that Turnitin only detected 54.8% of AI generated content
- 2 A study of 12 commercially available detection tools, which found that none were accurate
- 3 An evaluation of several GPT detectors which revealed a bias against non-native English writers

## WHAT ARE THE ETHICAL ISSUES OF “CATCHING” AND “DETECTING”?

---

As pointed out above, detection tools have been demonstrated to be biased against non-native English writers. There are also other ethical considerations when trying to “catch” or “detect” GAI use.

Firstly, students who are more digitally literate – or more fluent in general – may be able to use the technology in more sophisticated ways to generate undetectable content. These students might, for example, be able to construct better prompts which result in more “human-like” output. Or, they may use some of their own writing in the prompt to produce generated text that is more similar to their real “voice”.

Some students will also have better access to technology. This might simply mean device or internet access at home, or could mean access to more sophisticated models, such as the subscription-only GPT-4 model in ChatGPT. These students will produce content that evades detection software, much of which is trained to detect content from GPT-3 and 3.5.

Essentially, a student who is more confident, competent, or has access to a higher quality application might “cheat” and get away with it. This is part of the “digital divide” issue I wrote about back in January, but it is amplified when we consider that detection is more likely to be seen as an option for high stakes, competitive tests where wealthier, more literate students already have an advantage.

There are also ethical concerns with submitting student work to detection services, since the work may constitute their intellectual property. Deakin University’s [Professor Phill Dawson made an excellent post about this on LinkedIn](#), which included a discussion of student data privacy and security.

# WHAT DOES ALL THIS MEAN FOR ASSESSMENT DESIGN?

---

I'll begin this answer with a straightforward but possibly unpopular statement: for any unsupervised assessment, we have to assume students might use GAI.

This isn't a statement about trust. I'm not suggesting, like some of the early headlines when ChatGPT was released, that all students are compulsive cheats. I'm stating that given the ubiquity, ease of access, and inability to detect generative AI, there is simply no way to guarantee it won't be used for any assessment that doesn't happen under supervision. However, I'm also not suggesting that all assessment should be supervised, and certainly not conducted under exam conditions.

**Here are a few considerations when designing assessments with GAI in mind:**

- Does the student need to demonstrate knowledge or competency without any use of GAI? Are you sure? If so, conduct the assessment in person, under supervision. It's the only option.
- Is the assessment a practical or experiential task that doesn't benefit at all from the use of GAI? I.e., is there really no way that GAI could be used for the task? Think: fitness assessments, constructing a physical product...
- Are you assessing knowledge, or skills? Can the skills be assessed in a real-world context, or applied to the student's personal opinions and experiences?
- Assuming students can and possibly will use GAI to complete some or all of the task, are all students equally aware of the technology and do they have equal access? If not, what can be done to ensure that students with access to better models are not advantaged?
- If students "opt out" of using GAI, can you guarantee they won't be disadvantaged?
- Does the assessment need to be completed as a written task? Can it be completed orally, such as a discussion, viva, presentation, pitch, or debate?

Back in May, The University of Queensland's Jason Lodge along with Sarah Howard and Jaclyn Broadbent proposed a taxonomy of approaches to assessment redesign. In the final option, "rethink", **the authors made this comment:**

If assessments feel like chores and do not encourage creativity or inspire actual learning, or there is substantial time pressure to complete tasks, there is increased motivation to cut corners.

Jason Lodge, Sarah Howard, and Jaclyn Broadbent

**They also explored the long-term viability of different approaches, including banning and invigilating, given the development of generative AI technologies:**

	Short-term	Medium-term	Long-term
1. Ignore	Might get away with it momentarily		
2. Ban	Problematic	Becomes risky	
3. Invigilate	Where appropriate	Where appropriate	Where appropriate
4. Embrace	Being mindful of equity issues	Where appropriate	
5. Design around	Risky		
6. Rethink	Requires time and effort		

*Viability of assessment redesign for AI – Jason Lodge, Sarah Howard, and Jaclyn Broadbent*

I've written elsewhere about an "AI assessment scale" which could be applied here, giving students clarity on when and where to use or avoid GAI. The key is clear communication of the expectations, and genuine reasons for students to not use GAI under certain circumstances.

# WHAT ABOUT DISTANCE LEARNING, ONLINE COURSES, OR OUT-OF-CLASS ASSESSMENTS?

---

I've spoken with school leaders from distance education providers, as well as tertiary providers with hundreds of online students. I also work with schools which offer programs like the International Baccalaureate, which includes an extended essay that is worked on over time, and often out of class.

My answer here is the same: anything that happens outside of a supervised setting (which may be everything, in this case) can potentially be completed with GAI. Proctoring software and lockdown browsers are as much of a dead end as detection tools, and unfortunately create a culture of mistrust.

However, it might still be possible to engage students in rich, online discussions and conversations where their knowledge can be assessed in ways other than via a written response. Otherwise, you have to accept that students could be using GAI.

Refer back to the questions above about assessment design. How might tasks be structured so that it doesn't matter if students use GAI, or so that there is no advantage in using it?

## SO WHAT DOES “GOOD” ASSESSMENT LOOK LIKE?

---

This obviously depends on your subject and content, but “good” assessment should be authentic, and represent the kind of skills that the student will need beyond the course itself. Good assessment should move away from “knowledge checking” towards the demonstration of skills – and some of these skills might include the use of GAI.

It's important to ask what is being assessed, how, and why. Those might sound like obvious questions, but it's surprising how often assessments are conducted in ways which are ill-suited to the actual thing being assessed.

For example, in my subject area of English, we typically get students to demonstrate their understanding of the views, values, and ideas of a text through an analytical essay. Why? There are other methods equally suited to demonstrating that kind of knowledge, and the skill of analysis. The dominance of the essay as an assessment item across disciplines is as much about expedience as it is “good” assessment: it's much easier to collect and grade 100 essays than listen to 100 vivas, or 25 group discussions.

Sometimes, the essay might actually be the best form of assessment. It's a great skill to be able to logically argue your points, use concise evidence, and write with a compelling voice. But essays can be worked on over time, drafted and edited by hand, and can be accompanied by discussion and conversations with students. All of those approaches can contribute to the next point: authentication.

## HOW CAN I AUTHENTICATE STUDENT WORK?

---

First of all, assume that most students want to do the right thing.

If you have clear guidelines about academic integrity, and you avoid competitive behaviours that might lead to a culture of cheating, you make authenticating student work much easier.

### **Authentication can happen in a few ways:**

- Complete certain stages of the assessment, such as planning, brainstorming, drafting, or editing under supervision. Not necessarily under exam conditions – but in person and as part of the classwork. In fact, instead of calling this “supervision” you might just say its collaboration
- Retain copies (physical or digital) of work completed along the way. Compare this work to the final submission
- Discuss the work with students. It's often fairly simple to see if a student really has the knowledge and skills or if they've been relying on something like ChatGPT to produce work for them
- Lean on formative rather than summative assessment, providing more opportunities for students to demonstrate their own knowledge
- Have clear guidelines for how to acknowledge the use of GAI. Students might include links to chat transcripts, copies of prompts, screenshots, or a simple acknowledgement of the app used.

# CAN A STUDENT REALLY USE GAI IN MY TASK?

---

**Probably.**

Assuming there is some form of written component, or that the assessment is based on an assessment of the students knowledge of content, then it's highly likely that GAI could be used to generate some or all of the response.

If there are visual requirements, such as a folio of images, photos, or even hand drawn elements, then we will reach a time soon where this is possible through image generation too.

We are already at the stage where a recorded oral can be easily faked using a platform like eleven labs or Descript. These services take a recording of a voice and generate a realistic – if occasionally robotic – facsimile of the original voice.

As I mentioned earlier, we have to assume that students can and will use GAI in some form or another in any out-of-class assessment. Check out my blog post on [GenAI Strategy for Faculty Leaders](#) for examples of just how well GAI can handle different assessments.

## WHAT DO WE DO NOW?

---

As much as I'd love to have all of the answers, I don't, and neither does anyone else. Systems are responding to this very differently. Some places have locked down assessment to invigilated exams and pen-and-paper written tasks. Others have turned a blind eye to GAI and are pretending it doesn't apply to their tasks.

Cheating is nothing new, and some students will always flaunt the rules to get an unfair advantage. Others cheat because of time pressures, poor organisational skills, competing interests, academic insecurity, and a plethora of other reasons. Most students will know what appropriate and inappropriate use of GAI feels like, and we can help them through clear academic integrity guidelines to understand our expectations.

To close this post, if I haven't yet convinced you that trying to "catch" GAI is futile, I'll end with an anecdote. In fact, I'll throw this entire post into ChatGPT and generate an anecdote:

**Prompt: <copy/paste blog post>. INSTRUCTION: Write an anecdote about two educators, one who encourages the appropriate use of GAI and academic integrity, the other who tries to enforce bans and use detection software.**

## THE TALE OF TWO EDUCATORS: FEI AND JACK

Fei and Jack are both university professors teaching courses in the humanities. Fei is an advocate for integrating technology into the classroom, while Jack is more of a traditionalist, wary of the growing influence of AI in education.

### FEI'S APPROACH:

Fei believes that the future is unavoidable—AI is here to stay. She wants her students to be prepared for a world where technology and AI will play significant roles. As part of her curriculum, she includes a module on how to use AI responsibly. She educates her students on the pros and cons of using Generative AI, like ChatGPT, for academic research and writing.

*"I want you to understand how to use these tools in an ethical manner," Fei tells her students. "If you're using AI to generate some ideas or even a rough draft, be transparent about it. Add an acknowledgment, and make sure you comprehend and revise what the AI has produced."*

By taking this approach, Fei finds that her students are more engaged, and their assignments turn out to be more nuanced and well-researched. She uses class time to discuss ideas, encouraging students to think critically, which makes grading easier since she already has a good grasp of each student's understanding of the subject matter.

### JACK'S APPROACH:

Jack, on the other hand, is skeptical of Generative AI. He has heard of AI "cheat-detection" software and decides to use it to police his students' assignments. "No AI will undermine the integrity of my classroom," he proudly announces on the first day of his course.

Jack spends hours setting up his detection software and routinely checks every assignment and essay for potential AI intervention. However, the results are inconclusive and often inconsistent. An essay written by a non-native English speaker is flagged, causing unnecessary stress and a drawn-out investigation, while a paper partially generated by GAI slips through undetected.

The amount of time Jack spends trying to catch students "cheating" with AI takes away from his ability to engage deeply with his students' ideas and provide constructive feedback. Grading becomes a prolonged, grueling process, and classroom discussions are shallow as he hasn't had the time to prepare adequately.



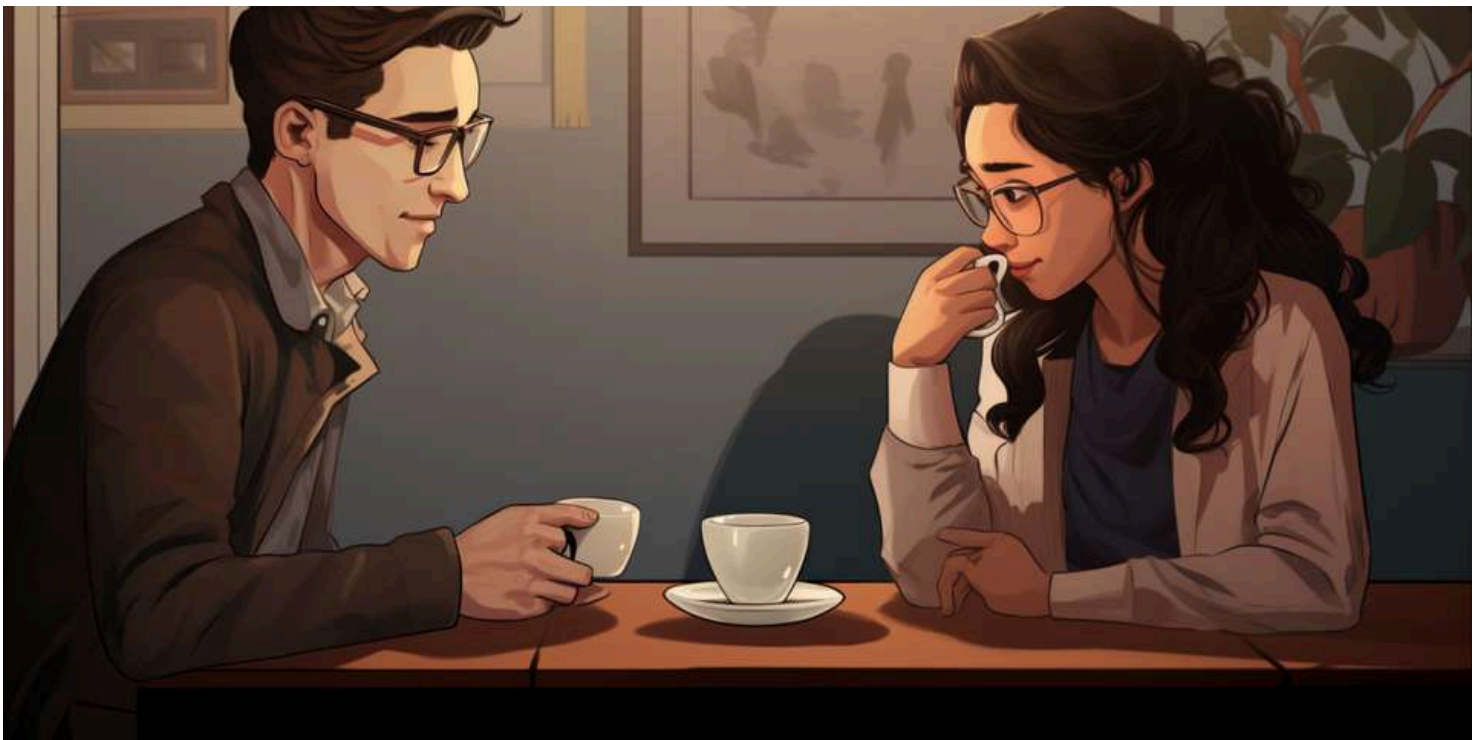
## CONCLUSION:

---

As the semester comes to an end, Fei and Jack meet for coffee to discuss their experiences. Fei shares how invigorated she feels, excited by the rich discussions and thoughtful work her students have produced. Jack, on the other hand, looks weary.

“I feel like I’ve spent the entire semester as a detective, not an educator,” Jack confesses. “And I’m still not sure if I succeeded in preserving academic integrity, or if I’ve just made everyone’s life more difficult, including my own.”

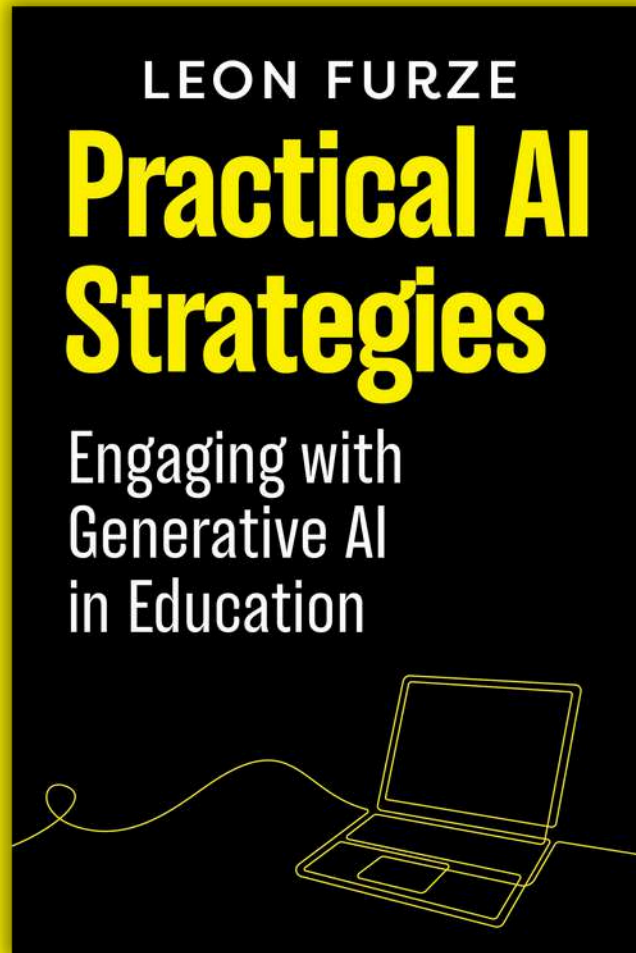
Fei nods sympathetically, “Technology is just a tool, Jack. It’s neither good nor bad. It’s all about how we use it and how we guide our students in using it.”



*The ballad of Fei and Jack. Image generated in Midjourney using the concluding paragraph written with ChatGPT, GPT-4.*

Thanks ChatGPT and Midjourney, I think. A little twee for my tastes, but you get the point. Trying to catch students using GAI creates mistrust and opens up potential ethical issues for equitable assessment. It’s also time consuming, frustrating, and antagonistic. Look for opportunities to move beyond the “cheating” narrative.

# Practical AI Strategies Online Course



The bestselling book is now available as a complete online course for K-12 and tertiary educators.

**Use the code `ASSESSMENT` for a 10% discount**

Offer expires 06/10/2024



**[practicalaistrategies.com](https://practicalaistrategies.com)**



# AI DETECTION IN EDUCATION IS A DEAD END


When you live in a research/social media bubble like I do, it's easy to take certain things for granted. For example, I always overestimate the number of people who are using generative AI regularly in their day to day work.

The reality, as of April 2024, is the majority of people within and outside of education haven't had the time, the interest, or the inclination to use much generative AI beyond free tools like the unpaid version of ChatGPT.

Something else I take for granted is the fact that AI detection tools **do not work**. Since the release of ChatGPT in November 2022, universities have been confronted with a number of products for detecting generative artificial intelligence. Largely, these tools have been born out of fear that large language model based technologies like ChatGPT will be used by students to cheat on assessment tasks.

It's an understandable and entirely valid concern, especially given statistics on how many students engage in academic misconduct (and the fact that considering many of these studies are self reported, means those percentages are probably much higher). But companies developing generative AI detection tools often prey on education providers in a way which is predatory and largely driven by commercial and not academic interest.

There are already dozens of AI detection tools on the market. To avoid giving any of them any free publicity, I'm not going to mention any of them directly in this article. Suffice it to say that since I started working with generative artificial intelligence two years ago, I have yet to see a detection tool that is reliable or accurate.



Outside of my bubble, where I have easy access to novel research, and the ability and inclination to test these tools myself, many education providers are still in the dark when it comes to detection tools and they can be lulled into a false sense of security by the companies selling them.

In this post, I'll discuss some of my personal objections to AI detection tools, and explore a new piece of research that once again proves AI detection tools don't work.

## HOW DO AI DETECTION TOOLS WORK?

---


Unlike traditional plagiarism checkers which compare texts to a large database of existing text (and don't get me started on the [amount of students' intellectual property being hoarded by these companies for profits](#)), AI detection tools use pattern matching to identify generated text.

Language models operate by [processing huge amounts of text data and learning probabilistic rules about how language works](#). They then use these rules to create novel text.

**However, language models often have tells which can be more predictable than human writing. For example:**

- 1 Lack of variation in sentence structure
- 2 Overuse of certain words such as conjunctions (e.g. "however", "furthermore", "in addition", "in conclusion")
- 3 Overuse of particular vocabulary (e.g. "delves", "navigates complexities")
- 4 Predictable sentence length, paragraph length
- 5 Predictable grammatical constructions

Detection tools work on pattern matching these features and also, in some cases, use traditional plagiarism detection methods to look for text which may be recreated verbatim from a language model's training data set.



# WHY IT DOESN'T WORK

Although AI detection tools can successfully identify some generated content, there are several points at which the tools break down, making them unsuitable as an academic integrity checkpoint.

First of all, large language models continue to develop at an incredibly rapid pace. A powerful model like [Claude 3 Opus from Anthropic](#) produces much more varied and less predictable text than GPT-4, which itself provides more sophisticated text than the free version of ChatGPT or other models which are more limited in capacity, such as the free version of Google Gemini or Microsoft Copilot when it is using GPT-3.5.

This means that using a more powerful model reduces the efficacy of detection tools until the detection tools are tweaked and improved based on the new model. Essentially it's an arms race between generation and detection, and one which, given the resources of developers like Microsoft, Google, and OpenAI, detection tool companies cannot hope to win.

It's also easy to circumvent or break detection tools using adversarial techniques. These are deliberate prompting tactics designed to work around the detection tools. Some examples of adversarial techniques include:

- 1 Lack of variation in sentence structure
- 2 Overuse of certain words such as conjunctions (e.g. "however", "furthermore", "in addition", "in conclusion")
- 3 Overuse of particular vocabulary (e.g. "delves", "navigates complexities")
- 4 Predictable sentence length, paragraph length
- 5 Predictable grammatical constructions

# AI DETECTION AS AN EQUITY ISSUE

---

Having explored a few of the reasons why AI detection tools can fail, it's important to now consider why they shouldn't be used at all as a point in an academic integrity conversation.

To do so, I'm going to illustrate the point using a scenario. Imagine four students complete the same assessment task. The conditions of the assessment task specify that no generative artificial intelligence tools may be used. Detection tools will be employed as an integrity measure after submission. The students must complete this assessment task in their own time, outside of the school/university.

**Ashley** is a regional student with limited access to digital technologies at home. They are therefore reliant on their institution's computers and network. The institution has blocked direct access to generative AI tools.

Ashley checks GPT, Gemini, and Co-pilot, but since they're blocked ends up having to use the free credits of a third party application built on top of GPT-3.5. They're also limited to completing this task during the time they have on campus at lunchtime or immediately after classes before returning home.

**Bob** is an English as an additional language (EAL) student from a migrant family where English is not spoken in the home. Bob uses the free version of ChatGPT because he has heard from fellow students that it is a good translation tool. He uses ChatGPT to translate both the assignment questions and his answers.

**Alice** comes from a low socio-economic background with low levels of literacy in the home and limited digital literacy. Alice uses Microsoft Copilot at home on her phone as a way to understand the requirements of the task and to help make her ideas seem more academically written.

**Marie** is an English first language speaker from a wealthy household. Her mother is a software engineer and her father is an intellectual property lawyer. Marie writes her response using her father's access to Claude (Opus), requiring a \$20 a month USD subscription. She inputs the assignment questions and generates her entire response verbatim.

Just for good measure, and because she knows how these tools work, she pastes the response into GPT-4 (another subscription-based model) and then back again into Claude with the instruction to make it a little bit more sophisticated, a little bit more varied, and to incorporate some direct quotes from the materials from class that she uploads as a PDF (a capability only available in paid models). Marie's final response is comprehensive, accurate, and sophisticated. It is also entirely fabricated by GenAI.

## The four students submit their work independently. The detection tool flags:

- 1 Ashley's work as 90% AI-generated
- 2 Bob's as 100% AI-generated
- 3 Alice's as 85% AI-generated
- 4 Marie's as 20% AI-generated

Of the students, you could argue that Bob and Alice attempted to use generative AI as an assistive technology to help understand the task and to form their answers. Alice's use was perhaps a little bit more heavy-handed. And all four students have certainly breached the requirements of the task by using generative AI in the first place.

The fact is, the student who used the generative AI tools with the most deliberate, nefarious intent was Marie, who was also the least likely to get caught. Marie is the student who was already advantaged by the education system, advantaged by her socio-economic status, and now advantaged by a heavy-handed approach to policing the technology.

**This is the equity issue of generative AI detection.**



*GenAI detection tools privilege students who are English first language, have access to paid large language models/applications, and are more digitally literate.*

# AI DETECTION IS A WORKLOAD ISSUE

---

Now let's shift our attention to look at an issue which is close to my heart. In 2016, I completed my Master's in Education which culminated in an action research project exploring how professional learning can mitigate the risk of teacher burnout. During that research, it became very clear that the factors contributing to teacher burnout are many and varied. Amongst those factors is the workload imposed by assessment and reporting practices.

In both K-12 and higher education, assessment is big business, and at the end of most assessment work, educators spend hours marking, moderating, and reporting. Assessment is an important but time-consuming part of the job.


Checking for and monitoring cases of academic misconduct is unfortunately part of this task. In many senior secondary and higher education institutions, this includes processes such as automatic plagiarism checking, and the responsibility generally falls to the teacher or lecturer in charge of the class.

Typically, the process goes something like this: For assessment tasks that are completed outside of examination conditions, in electronic format, students are required to submit their work through a plagiarism detection platform, often built into the learning management system (LMS). Either students upload to this platform directly or their teachers upload a collection of assignments in bulk.

The assessments are processed by the plagiarism checking system and reports are generated. Having used these tools myself for senior secondary English and for undergraduate teacher training courses, I can attest that whilst they're not hugely time consuming, this process does add a layer on to the assessment and reporting process. If a student's work is reported beyond a particular threshold (say, 20% to allow for genuine quotes and citations), then the assessor has to go in, manually identify the areas which have been flagged as plagiarism, and then report back to the student. In extreme cases of plagiarism, this will then kick along to whatever the next stage of the institution's academic integrity policies are, for example resubmission, zeros, and so on.

Whilst this is a brief imposition on the educators, the use of similar approaches with generative AI is much more burdensome on educators. This is because, unlike plagiarism tools, generative AI tools do not give a clear cut result. The percentage likelihood of AI generated content is less accurate than plagiarism detection, more open to interpretation, and therefore requires more consideration on the educator's part. It requires more nuanced and potentially more stressful conversations between the educator and the student, and the potential for much more kickback from the students and many more appeals. In many contexts, both students and parents are aware that detection tools are not as accurate as plagiarism tools.





The added time and stress of using generative AI detection tools is a burden on educators who are already in an industry with a high risk of burnout and attrition.

## NEW RESEARCH


---

In 2023 and 2024, I had the privilege of working on papers on an AI assessment scale with Dr. Mike Perkins, Dr. Jasper Roe and Associate Professor Jason MacVaugh. I've [detailed the AI Assessment Scale elsewhere](#) and you're welcome to [download a free ebook of activities aligned to the scales which allow for generative assessment](#).

Mike and Jasper, along with other authors, have just published a preprint of their latest research testing over 800 samples of writing against various detection tools. Mike shared the research on LinkedIn with this comment:

[Our latest preprint](#) shows the results of 805 tests of human samples, initial GenAI output, and GenAI output after we applied adversarial techniques designed to evade detection by AI text detectors. We saw a non manipulated mean accuracy rate of 39.5%, dropping to 22.1% after the application of the adversarial techniques

The [researchers also found concerning rates of false accusations](#) (15%) where the tools incorrectly flagged human-written samples as AI-generated. At the same time, a high percentage of AI-generated texts went undetected, and a lower rate of false positives appears to come with an increased rate of undetected content. This points to major risks for both students being unfairly accused and dishonest usage of AI going unnoticed.



Adversarial Technique	Bard		Claude 2		GPT-4	
	Accuracy	% reduction	Accuracy	% reduction	Accuracy	% reduction
Original AI samples	76.9%	-	17.7%	-	23.9%	-
Add spelling errors (SE)	14.3%	62.6%	7.9%	9.9%	16.5%	7.3%
Write as NNES (NNES)	16.4%	60.5%	2.2%	15.5%	29.0%	-5.2%
Decrease complexity (DC)	57.2%	19.7%	14.6%	3.2%	34.9%	-11.0%
Increase complexity (IC)	50.4%	26.5%	10.8%	6.9%	2.1%	21.7%
Increase burstiness (IB)	57.7%	19.2%	14.1%	3.6%	11.3%	12.6%
Paraphrase (PR)	32.4%	44.5%	8.8%	8.9%	13.9%	10.0%
Mean accuracy reduction	-	38.8%	-	8.0%	-	7.6%

Table 11. Performance of Generative AI tools

<https://arxiv.org/abs/2403.19148>

Interestingly, the outputs from different AI models had varying levels of detectability, with text from Google’s Bard being the easiest to identify compared to GPT-4 and Anthropic’s Claude. However, Bard-generated text also saw the biggest drop in detectability after applying adversarial techniques.

Several conclusions emerge, but alongside my other comments in this article the key is that not only are AI detection tools largely ineffective, they are also a short-term, ill-advised, and possibly unethical approach to academic integrity in light of generative AI. The current limitations of these tools underscore the need for a critical, nuanced approach if implementing them in higher education, and highlight the importance of exploring alternative AI-aware assessment strategies.

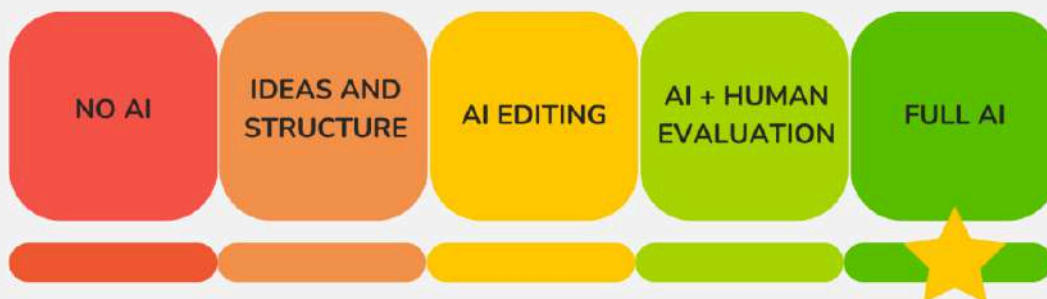
Over the next few months, I’ll be writing extensively about approaches that K-12 and tertiary organisations can take to update their assessment strategies in ways which don’t rely on ineffective technologies.

Over the last few years, I’ve worked with dozens of schools and universities and served on the boards of several not-for-profits, and have been involved in strategic planning, teaching and learning, assessment, and of course generative artificial intelligence.



# THE AI ASSESSMENT SCALE: UPDATE AND PILOT STUDY

Last year, I co-authored a paper with [Dr Mike Perkins](#), [Dr Jasper Roe](#), and [Associate Professor Jason MacVaugh](#) in which we proposed the AI Assessment Scale (AIAS). [The first paper](#), available as a preprint on [arXiv](#), outlines the rationale for a scaled approach to using AI versus a “use/don’t use” or a blanket ban. Since that initial publication, I also produced a set of resources on how the scale can be used across subjects, which can be accessed [here](#):



## THE AI ASSESSMENT SCALE

A TOOL FOR GENAI ASSESSMENT

[leonfurze.com](https://leonfurze.com)

I've collated all of the resources on the AI Assessment Scale into another free eBook: [click here to get a copy](#).

In early 2024, we published the second paper focusing on the AIAS and how it was implemented at British University Vietnam. The paper, titled “[The AI Assessment Scale \(AIAS\) in action: A pilot implementation of GenAI supported assessment](#)”, has been released as a preprint on arXiv.

<b>1</b>	<b>NO AI</b>	<p>The assessment is completed entirely without AI assistance. This level ensures that students rely solely on their knowledge, understanding, and skills.</p> <p style="text-align: center;"><b>AI must not be used at any point during the assessment.</b></p>
<b>2</b>	<b>AI-ASSISTED IDEA GENERATION AND STRUCTURING</b>	<p>AI can be used in the assessment for brainstorming, creating structures, and generating ideas for improving work.</p> <p style="text-align: center;"><b>No AI content is allowed in the final submission.</b></p>
<b>3</b>	<b>AI-ASSISTED EDITING</b>	<p>AI can be used to make improvements to the clarity or quality of student created work to improve the final output, but no new content can be created using AI.</p> <p style="text-align: center;"><b>AI can be used, but your original work with no AI content must be provided in an appendix.</b></p>
<b>4</b>	<b>AI TASK COMPLETION, HUMAN EVALUATION</b>	<p>AI is used to complete certain elements of the task, with students providing discussion or commentary on the AI-generated content. This level requires critical engagement with AI generated content and evaluating its output.</p> <p style="text-align: center;"><b>You will use AI to complete specified tasks in your assessment. Any AI created content must be cited.</b></p>
<b>5</b>	<b>FULL AI</b>	<p>AI should be used as a “co-pilot” in order to meet the requirements of the assessment, allowing for a collaborative approach with AI and enhancing creativity.</p> <p style="text-align: center;"><b>You may use AI throughout your assessment to support your own work and do not have to specify which content is AI generated.</b></p>

Graphic of the AIAS from <https://arxiv.org/abs/2403.14692>

The AIAS emerged from discussions with colleagues about the need for a more nuanced approach to integrating Generative Artificial Intelligence (GenAI) into assessments. We felt that a simple yes/no to using GenAI would be unsustainable and that we needed a framework that could account for the rapid advancements in the technology while also honouring our students' integrity and creativity. The AIAS was designed from the start to be flexible, with the understanding that multimodal GenAI will impact all disciplines.

In the paper, we discuss the global responses to GenAI in higher education, which have often focused on banning or restricting the use of these tools. However, as the understanding of GenAI capabilities grew, institutions have started to adopt more nuanced approaches. We argue that the current discourse surrounding GenAI in education overemphasises academic misconduct, both minimising more pressing ethical concerns and neglecting the potential benefits of these technologies.



The case study of the AIAS implementation at British University Vietnam (BUV) forms the core of our paper. We detail the process of introducing the AIAS, from the initial discussions and policy adjustments to the launch and implementation of the framework. The results of the pilot study are promising, with a significant reduction in academic misconduct cases related to GenAI, an increase in student attainment, and a rise in module passing rates.

We also discuss the shift in pedagogical practices following the implementation of the AIAS, with faculty members incorporating GenAI tools into their modules, and students producing innovative multimodal submissions, including students for whom English is not their first language. This suggests that the framework not only addresses concerns about academic integrity but also creates new avenues for student creativity and engagement.



*Photo by George Pak on [Pexels.com](https://www.pexels.com)*

While the initial results are encouraging, we acknowledge the limitations of the study, including its scope within a single institution and the need for further validation through larger, more diverse studies. We also discuss the ethical implications of GenAI in education, such as bias and fairness, privacy and data security, and intellectual property rights.

We argue that the AIAS offers a practical, flexible, and adaptable framework for integrating GenAI into educational assessments. By adopting an approach that embraces the opportunities of GenAI while recognising its limitations, educators can support students in developing the skills needed for an increasingly AI-driven world.

## Paper abstract:

---

*The rapid adoption of Generative Artificial Intelligence (GenAI) technologies in higher education has raised concerns about academic integrity, assessment practices, and student learning. Banning or blocking GenAI tools has proven ineffective, and punitive approaches ignore the potential benefits of these technologies. This paper presents the findings of a pilot study conducted at British University Vietnam (BUV) exploring the implementation of the Artificial Intelligence Assessment Scale (AIAS), a flexible framework for incorporating GenAI into educational assessments. The AIAS consists of five levels, ranging from 'No AI' to 'Full AI', enabling educators to design assessments that focus on areas requiring human input and critical thinking. Following the implementation of the AIAS, the pilot study results indicate a significant reduction in academic misconduct cases related to GenAI, a 5.9% increase in student attainment across the university, and a 33.3% increase in module passing rates. The AIAS facilitated a shift in pedagogical practices, with faculty members incorporating GenAI tools into their modules and students producing innovative multimodal submissions. The findings suggest that the AIAS can support the effective integration of GenAI in HE, promoting academic integrity while leveraging the technology's potential to enhance learning experiences.*

The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI supported assessment <https://doi.org/10.48550/arXiv.2403.14692>

If you're interested in learning more about the AIAS and our pilot study at BUV, I recommend reading the full paper on arXiv:

<https://arxiv.org/abs/2403.14692>



# DITCH THE DETECTORS: SIX WAYS TO RETHINK ASSESSMENT FOR GENERATIVE ARTIFICIAL INTELLIGENCE

*This article is based on a series of short LinkedIn posts and includes the original ideas, plus some of the feedback and discussion from the comments. [Head over to my profile on LinkedIn](#) to find the originals.*

In recent weeks, I've shared my thoughts on Generative AI (GenAI) and its impact on assessments, particularly the fact that AI detection tools are largely ineffective. But if we're going to move away from these tools as a part of the academic integrity process, what can we replace them with?

I've got a few ideas – none of them groundbreaking or overly complex, but each with its own advantages and disadvantages. At the core of all these suggestions is a simple premise: GenAI didn't 'break' assessment, and we, as educators and institutions, set the boundaries around what constitutes 'academic misconduct'.

## 1. 'LEVEL 5 ASSESSMENTS'

---

In the [AI Assessment Scale](#) developed by Mike Perkins, Jasper Roe, Jason MacVaugh, and myself, we outline five levels ranging from 'no AI' to 'full AI'. 'Level 5 – Full AI' assessments obviously require us to disregard detection tools altogether. At this level, we actively teach and encourage students to experiment with GenAI tools. You can read more about the AIAS in the [Journal of University Teaching & Learning Practice Vol. 21 No. 6](#) or via our [recent preprint](#) dealing the first pilot study of the Scale.



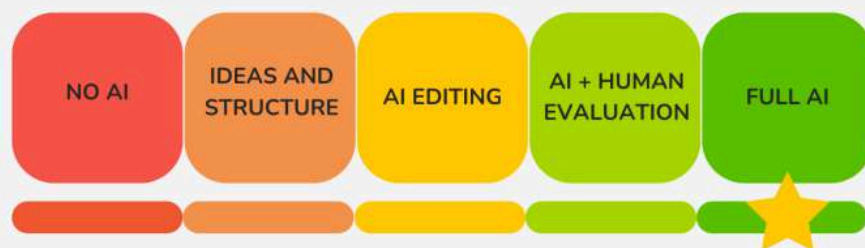


# ADVANTAGES:

- Realistic: Few employers are preventing their employees from using GenAI (some aren't even aware of its existence), so when students leave the educational bubble, they'll be free to use whatever tools are available to them.
- Multimodal and flexible: Level 5 tasks permit the use of any GenAI applications suitable for getting the job done, including text, image, audio, video, 3D, and code generation.

# DISADVANTAGES:

- Ethical concerns: GenAI isn't a neutral technology – copyright and IP issues, dataset bias, and environmental costs are among the problems we need to address before fully embracing 'full AI' tasks.
- Equity of access: Not all tools are created equal, and some students may have access to more sophisticated (and expensive) models, potentially leading to an unfair advantage.



## THE AI ASSESSMENT SCALE

A TOOL FOR GENAI ASSESSMENT

[leonfurze.com](http://leonfurze.com)

## 2. EXPECT AI USE AND TEACH THE SKILLS

Another suggestion for rethinking assessments without relying on AI detection tools is to design tasks suitable for Levels 2-4 of the [AI Assessment Scale](#), which includes using AI for ideation, editing, or significant portions of a task.

Here's my entirely unsurprising proposal: Expect that students will use Generative AI and explicitly teach them the necessary skills.

### ADVANTAGES:

- You won't be caught off guard when students use GenAI to complete a task, eliminating the need for detection tools.
- You'll be able to address students' concerns (well-documented in recent surveys) that their education providers aren't preparing them for a future that involves using GenAI tools.

### DISADVANTAGES:

- The time, resources, and cost required to train educators to have an equal and shared understanding of how the technology works.
- The need to update and reframe many (if not all) current assessment tasks.

If we anticipate that students are using GenAI (which they are), we can start thinking more deliberately about how to best support them in using these technologies ethically and appropriately.

The [comments on the post](#) about expecting students to use AI and teaching the necessary skills raise some important considerations. As Adrian Cotterell points out, even when aiming for "no AI" tasks, it's crucial to ensure that the assessments are accessible and not limited to traditional pen-and-paper exams. Additionally, as Jason Braun suggested, educators need to rethink what constitutes great work in a world powered by GenAI. While the overall quality of student outputs may rise, truly outstanding work might have unique characteristics, such as rougher edges or a more distinct voice.

## 3. UNGRADING

---

Ungrading isn't a new concept, but it gains new relevance when considering technologies that can effectively complete many of our traditional assessments.

If we shift the focus of education away from the final graded assessment and towards what is being taught (and why), then the imperative for academic misconduct may be lessened.

As Emily Pitts-Donahoe [recently wrote on her substack](#), there are many reasons to “ungrade”...

“But I also ungrade because I want students to write and succeed at a high level and grades get in the way of that. Grades keep my students focused on GPAs and transcripts rather than on growth and improvement. They draw attention away from the careful feedback I provide on student work—and when students do engage with feedback, it's primarily to game a points system rather than to further their l...”

**Emily Pitts Donahoe**

The Life or Death Consequences of Grading

EMILYPITTSDONAHOE.SUBSTACK.COM



## ADVANTAGES:

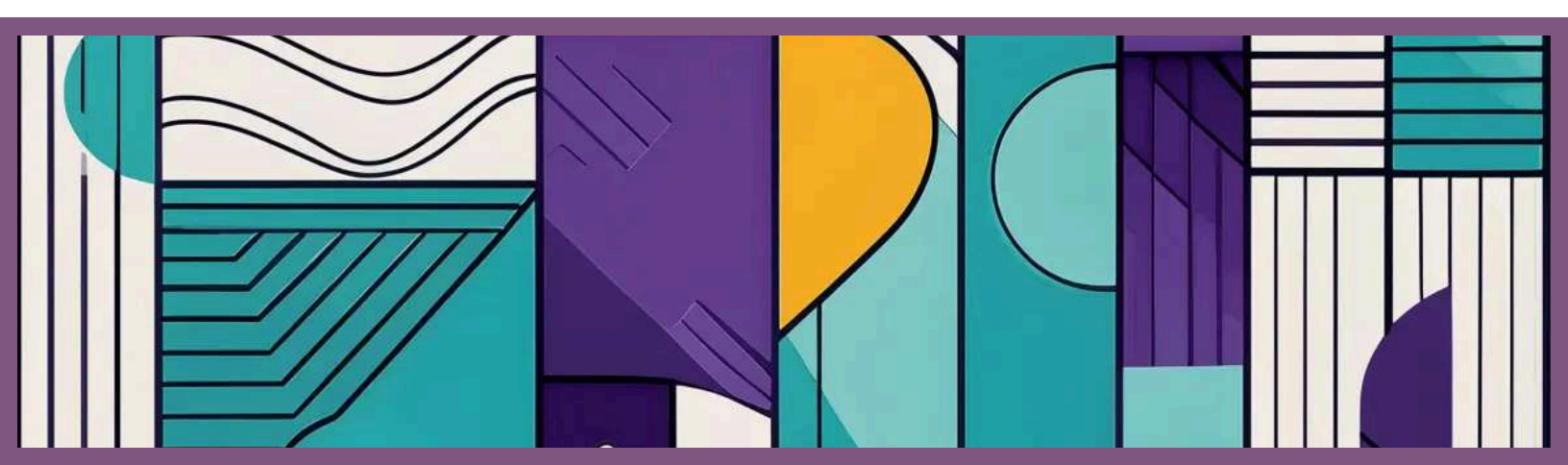
- Reduces stress and pressure around high-stakes assessments and focuses learners on what is being taught and why.
- Allows for diverse use of multimodal GenAI technologies without worrying about their impact on the final grade.

## DISADVANTAGES:

- Countercultural and against the grain for many institutions' current assessment practices; you will likely face resistance.
- The perception that ungrading doesn't work for "real world" subjects.

Ultimately, ungrading is a cultural shift, but it's an idea with serious merit for improving assessments, with or without technology.

The [comments on the original post about ungrading](#) demonstrate some of the potential of this approach to shift the focus from grades to deeper understanding and genuine learning. As Majda Benzenati points out, ungrading allows educators to prioritise critical thinking, intellectual curiosity, and finding joy in the learning process. [Emily Pitts Donahoe's work with her students](#) further emphasises how ungrading can motivate students to learn rather than simply chase high grades. While resistance to this cultural shift is expected, as noted by Ryan MacDonald, many educators like Joerg Meindl are already moving towards ungrading or alternative grading practices. They recognise the importance of focusing on the process, providing feedback, and explaining the purpose behind learning activities. As Vince Wall suggests, ungrading aligns well with process-oriented pedagogies like project-based learning, which may become increasingly relevant in the context of AI-infused education.



## 4. KNOW YOUR STUDENTS' STYLE

---

Developing a deep understanding of a student's style and voice is another way to update assessments without relying on AI detection tools.

There are tools available that can help with "stylometry," and AI-assisted tools are undoubtedly already in the pipeline to assess work against a student's previous output. However, I'm talking more about the good old-fashioned approach of "knowing your students."

### ADVANTAGES:

- Building relationships with students by fully understanding and appreciating their perspectives and ways of expressing themselves.
- Respecting students' work and building these relationships is an effective way to mitigate academic misconduct.

### DISADVANTAGES:

- Scalability issues; it's difficult, if not impossible, for one lecturer/teacher/tutor to develop a deep understanding of 100+ students' work over a single semester or unit.
- Still vulnerable to "traditional" methods of academic misconduct like contract cheating and more sophisticated GenAI models like Claude 3 Opus, which are better at emulating style.

When faculties engage in block marking, where assignments are split evenly among faculty members rather than each teacher marking their own students' work, developing a deep understanding of individual students' styles can be more challenging. However, this practice is often reserved for summative assessments, and there are ways to mitigate the issue. For example, when I've run marking in this manner, the actual teacher still reviews their own students' work before releasing grades to check for any outliers or inconsistencies. This allows for a balance between the benefits of block marking, such as increased consistency and reduced bias, and the importance of teachers being familiar with their students' unique voices and abilities.

As with all of these suggestions, there's no perfect solution. Knowing your students' style and voice is great if the cohort is small enough, but there will always be issues and ways to game the system, especially with assessments at scale.

## 5. REDEFINE CHEATING

---

Suggestion number five might seem a bit flippant, but at the end of the day, we (educators, institutions, authorities, examination boards) define what is and isn't "cheating."

We've already seen some shifts in how academic integrity is discussed with GenAI in mind. For example, many academic integrity policies no longer group AI under the catch-all term of "plagiarism" because it isn't. Some have even gone as far as explicitly permitting AI use.

### ADVANTAGES:

- Redefining cheating demonstrates to students that we value trust and transparency and places the expectation on them to do the right thing. It acknowledges that we can't ban or block the technology and that we need to reframe our assessments accordingly.
- Reduction of educator workload; no more time spent endlessly chasing plagiarism (or "detection") reports or going back and forth with appeal processes over academic integrity.

### DISADVANTAGES:

- Huge systemic and cultural barriers, not least the perception within and outside of education that shifting the goalposts on academic integrity is "soft" or a cop-out.
- Easier said than done; this is a total, system-wide shift we're talking about. If one institution decided to reinvent its entire approach to academic integrity, it would quickly hit barriers if external agencies and assessment bodies didn't also move.

Redefining academic integrity in the age of GenAI isn't just about updating policies; it requires a fundamental shift in how we approach learning and assessment. [As Mathew Hillier points out in the comments on the original post](#), the key question should be "how are you assuring learning has happened?" rather than focusing on catching cheaters. This reframing allows us to approach academic integrity from a more constructive standpoint, emphasising the importance of genuine learning over the moralistic labelling of certain behaviours. By moving away from punitive measures and instead designing assessments that truly demonstrate learning, we can create a system that encourages students to engage with their education meaningfully, rather than seeing it as a series of hoops to jump through.



## 6. IN-PERSON, IN-TIME, IN-PLACE ASSESSMENTS

My final suggestion for updating assessments in light of GenAI, without using detection tools, is for in-person, in-time, in-place, no-device assessments.

I've deliberately left this one until last, and ironically, it's where many institutions went first when ChatGPT was released. But this doesn't necessarily mean examination-style assessments.

Group work, orals, seminars, practicals, simulations, vivas, brainstorming with post-it notes, debates, marker pens on butcher's paper... There are plenty of methods that predate GenAI by a few centuries and still work.

### ADVANTAGES:

- Easy to monitor and secure; with no access to devices and no way to do what the Victorian police call "sneaky face" (looking at a phone while driving or, in this case, under the desk), there's no GenAI to worry about here. We might call these 'Level 1' assessments in our AI Assessment Scale.
- Relevant, engaging, and authentic; these assessments can be modeled on real-world and authentic experiences, such as carrying out a practical task or a simulation.

### DISADVANTAGES:

- Unfortunately, this type of assessment is hard to scale. It might work well for tutor groups or K-12 classes, but it becomes unwieldy in a cohort of 100+ students.
- No online mode. Short of relying on lockdown browsers and creepy surveillance tech, there's no way to guarantee "no devices" in an online setting. I'll be writing more about GenAI and online teaching at a later stage because it's a whole different ballgame.

So, those are six suggestions for assessments that account for GenAI but don't rely on detection tools. None of them are perfect, and each comes with its own set of challenges, but I believe they're a step in the right direction as we navigate this new landscape of education in the age of artificial intelligence.





# RETHINKING ASSESSMENT FOR GENERATIVE AI: BEYOND THE ESSAY

This post is part of a series on rethinking assessment in light of generative AI. The posts draw on research and resources from K-12 and tertiary to suggest ways that educators can design engaging, compelling assessments which shift the narrative away from GAI and “cheating”. [Click here for the previous post on oral assessments.](#)

There are plenty of good reasons to use the essay as a formal piece of assessment. Essays offer an opportunity for students to demonstrate their knowledge in a cohesive, structured manner, forming logical arguments and leading the reader through their thoughts. They can also be creative and playful, allowing a confident writer to express their unique authorial voice.

Essays are also relatively easy to grade, and scaleable. It’s straightforward enough to collect and mark essays conducted under examination settings, even from thousands of students. Every year, the Victorian Curriculum and Assessment Authority (VCAA) collects around 45,000 English essays which are scanned into Pearson’s eMark platform and distributed to a few hundred assessors. It’s a huge task, but not unreasonable, and the essays are accurately graded in a two week period.

Unfortunately the expediency of essays as a form of assessment has made the form a “go to” in many subject areas and at various levels from primary through to tertiary. Don’t get me wrong, I personally love writing. I’d take an essay over a multiple choice exam any day, and I’d probably prefer to write than undertake an oral exam. But the problem is, not everyone enjoys writing, and not everyone can write an extended response.

That would be fine, if essays weren’t so prominent. And now that an easily accessed technology exists that can do most of the work of essay writing, we have an even more complex problem. Because now if students don’t want to write, can’t find the time to write, or simply can’t write, they can turn to a chatbot like ChatGPT to do the work for them, and that’s obviously not what we want.



The answer, however, doesn't lie in banning the technology or locking down all essays to examination conditions. Exams are terrible for accessibility, cause anxiety, and are far from a genuine representation of most skills. Instead, we need to look for ways to move beyond the essay.

In this post I'm going to explore a few alternatives which still tick the boxes of what could be assessed via an essay. I'm not discussing oral assessments here as I covered that in detail in an earlier post, but those are of course worth bearing in mind too when looking for alternatives.

## BEYOND THE ESSAY

---

**Aside from the "threat" of GAI chatbots, here are a few more reasons why you might want to look beyond an essay or long written response as a primary form of assessment:**

- 1 Limited Scope:** Essays often focus on individual performance and intellectual capability, which may overlook other important skills such as teamwork, verbal communication, or practical skills.
- 2 Time-Consuming:** Writing, revising, and grading essays can be time-consuming for both students and educators.
- 3 Writing Skill Bias:** The format may unfairly favour individuals with strong writing skills, even if the subject matter does not primarily concern writing ability.
- 4 Plagiarism Risk:** Essays can be susceptible to plagiarism, which undermines the learning process and the integrity of the assessment.
- 5 Stress Inducing:** Some individuals might find essay writing to be stressful or anxiety-inducing, especially if they struggle with articulating their thoughts in writing.
- 6 Fixed Interpretations:** Essays often require adherence to a particular thesis or argument, which may discourage the exploration of alternative perspectives or creative thinking.
- 7 Cultural Bias:** The conventional essay format may reflect a particular cultural or academic tradition that might not be inclusive of or accessible to all learners.
- 8 Subjectivity in Grading:** Grading essays can be subjective, and variations in grading standards can affect fairness and consistency in assessment.

**9 Limited Feedback:** In large classes, providing detailed, constructive feedback on essays can be challenging, which may hinder the learning process.

**10 Misalignment with Learning Objectives:** If the primary learning objectives are to assess practical skills, collaborative abilities, or other non-writing related competencies, essays may not be the best assessment tool.

## THE ALTERNATIVES

---

So let's take a look at a few alternatives which directly address those concerns, and which might also mitigate the risk of students breaching academic integrity by using GAI in unacceptable ways. For each of the assessment types I'll provide an overview and then some examples. Since I haven't taught in every single subject area, some of the examples will be generated by GPT-4.

## PERFORMANCE BASED

---

If you've ever taught in the VET sector, Australia's vocational training, you'll find many of the assessment types in this post familiar. That's because performance based tasks, observations, and on-the-job skills are par for the course in vocational education. But there's no reason this can't transfer to any secondary-tertiary subject.

After all, any discipline should be preparing students to use the skills and knowledge in some real-world context, whether that's further study, employment, or another field entirely. Even esoteric subjects like philosophy or subjects in the creative arts where the "goal" isn't necessarily tied to economic imperatives require students to develop skills they'll actually use in the future. For core subjects in secondary – Maths, English, Science, and the Humanities – the skills and knowledge are far better related to real-world applications than abstract chunks of knowledge.

Performance based assessments may [require students to respond to essential questions](#) and demonstrate skills in a real-world scenario. They allow for interdisciplinary knowledge and don't necessarily rely on the content taught in a given unit or topic.

## Here are a few examples:

- 1 Mathematics:** Students could be tasked with designing a budget plan for a small startup, applying mathematical principles to allocate resources, project profits, and manage expenses. The final presentation could include a detailed report and a presentation to a mock panel of investors.
- 2 English:** Students might be asked to create a multimedia storytelling project where they write and illustrate a short story, then present it to a younger age group at a local library or school. This task encourages creative writing, visual storytelling, and public speaking skills.
- 3 Physical Education:** Create a fitness programme for a specific goal such as preparing for a 5k run or improving general health. Students could track their progress, reflect on their experiences, and present their results and learnings to the class.
- 4 Economics:** Students could be assigned to analyse the economic impact of a recent local or global event, using economic theories and models. They could present their findings in a video essay format to be shared with the community, encouraging real-world application and public discourse.
- 5 Italian (LOTE):** Students could be tasked with planning and executing an “Italian Culture Day” event, where they prepare Italian dishes, present on various aspects of Italian culture, and engage in conversations in Italian. This task encourages language use in a practical, engaging, and collaborative context.

## PORTFOLIO OR WRITING JOURNAL

---

I love writing journal tasks. They're my recommended form of assessment for [VCE English and EAL Unit 2 and 3: Crafting and Creating Texts](#), and pretty much any form of creative assessment where students are required to demonstrate skill development over time. Portfolios of artwork, design ideas, and creative writing are common, but this assessment approach can be applied to other disciplines and subjects.



When I go through my own notes for my PhD colloquium document, it looks very much like a writing journal. I have annotations, extended abstracts, short snippets of writing which may or may not make it to the final piece, and draft versions of the document itself. Each piece contributes to the whole of my knowledge on the subject. The problem is, in secondary and tertiary education, we often don't value the whole journey.

Obviously my situation is different: if you're studying at this level, you have to really want to. It's not like secondary English, where you're doing it because it's compulsory, or even undergraduate studies where you might be doing it just to get a job or because it's expected. But if both students and educators at every level could learn to value the process of writing and creating, then we would probably find there's a lot more valid content to assess than just the finished product. The added bonus is you'll get a lot more insight into the student's usual style, voice, and way of thinking, which can be useful in academic integrity conversations.

### → **Physics: Research Journal**

Students could maintain a journal documenting their investigations into various physics phenomena. They could conduct small experiments, record observations, analyse data, and draw conclusions over the course of the term or semester. Additionally, they might reflect on how their understanding evolves with each experiment and how the concepts relate to real-world applications.

### → **Business Management: Case Study Portfolio**

Students could create a portfolio of case studies analysing different businesses or management scenarios. For each case, they could provide an overview, identify challenges, propose solutions based on management theories, and reflect on the potential outcomes and lessons learned. This portfolio could showcase their analytical, problem-solving, and strategic thinking skills.

### → **Geography: Field Study Journal**

Students could document field studies investigating local geographical issues or features. They could record observations, collect data, and analyse findings in a journal, reflecting on the implications and how the local findings connect to broader geographical concepts.

### → **Psychology: Observational Journal**

Students could maintain a journal where they observe and analyse human behaviour in various settings, relating their observations to psychological theories and concepts. They might also reflect on how these observations alter or deepen their understanding of psychological principles.

## → **History: Historical Investigation Portfolio**

Students could conduct investigations into different historical events or figures, documenting their research process, sources, analyses, and reflections in a portfolio. They might also include essays or reports that synthesise their findings, showcasing their ability to engage with historical inquiry and analysis.

Remember, with this type of assessment there's no need to have a "finished product" at the end. Students can of course take parts of their work over the term and write a final essay, but they don't have to.

## **PROJECT BASED**

Project-based learning (PBL) is nothing new. Some schools and education institutions have entire curricula based around PBL, dedicated middle-years programs, or whole sites devoted to the format. You don't have to go all-in on PBL though to get some of the benefits of project based assessment.

In a typical project based assessment, there is a real-world problem and a structure like a design thinking process. There may also be an inquiry problem or research topic, and the project extends over a number of weeks or even a whole term or semester. At the end, there is often an opportunity to present or pitch an idea or solution to the problem.

As a means to rethink assessment in light of generative AI, project based tasks could be a great option as they are engaging, authentic, and allow a student to demonstrate their skills in a broad range of tasks rather than a pass/fail scenario. Students "cheat" in assessments for all kinds of reasons, but making tasks more engaging can mitigate some of the risk.

### **Here are some examples:**

#### **COMPUTER SCIENCE :**

Students could be tasked with developing a mobile or web application to solve a real-world problem. They would need to go through the stages of planning, design, coding, testing, and deployment, and finally present their application and a report of their process, challenges faced, and how they overcame them.

#### **LITERATURE :**

Students could create a literary magazine featuring original short stories, poems, and essays, along with literary analysis of classic or contemporary works. They would be responsible for the curation, editing, design, and publication of the magazine, either in print or digitally.

## **MUSIC :**

Students could be tasked with composing an original musical piece based on a particular theme or historical period. They would then perform the piece, either solo or as part of a group, and submit a reflection on their creative process and the techniques used in their composition.

## **PRODUCT DESIGN AND TECHNOLOGY :**

Students could identify a common problem and design a product to address it. They would then create a prototype, document the design process, gather feedback, make improvements, and present their final design along with a reflection on the iterative design process.

## **ENVIRONMENTAL SCIENCE :**

Students could conduct a study on a local environmental issue, such as pollution, wildlife habitat destruction, or energy consumption. They would collect and analyse data, propose solutions, and present their findings to the community or a local governmental body.



*Study of a local environmental issue, illustration, isometric. Model: Midjourney*

# OBSERVATIONS

As I mentioned earlier, the vocational sector has a lot to offer when it comes to real-world, non-essay based assessment tasks. Many of these kinds of tasks are “GAI-proof” because they happen away from devices under practical circumstances.

Observations can be conducted in a range of scenarios, including performance tasks, as part of longer projects, and during group work. The difference is the student likely knows that the observed period of time is their assessment, and is (hopefully) aware of the explicit criteria.

That can add some pressure to the task, but only the kind of pressure that the student is likely to face in a real-world scenario, and not the false pressure of an examination.

In VET subjects, an assessment tool is a framework for evaluating students’ knowledge and skills, comprising assessment context, tasks, evidence gathering guidelines, performance quality criteria, and administrative requirements. These tools, guided by principles of validity, reliability, flexibility, and fairness, ensure that assessments are accurate, consistent, negotiable, and equitable.

The design of these tools necessitates industry consultation and testing on a student sample to ensure the evidence collected is valid, sufficient, current, and authentic, aligning with the competency units’ criteria.

This includes the design of assessment tools for observations. Tools like observation checklists, accompanying questions, and instructions for both students and lecturers/observers support this method. Observation checklists a focus on vocational and employability skills, and adherence to workplace procedures.

## Mathematics

- Task: Solving a series of progressively complex algebraic equations.
- Checklist: Correct application of algebraic rules, accurate simplification, correct answer, and clear presentation of solution steps.

## Science (Physics)

- Task: Conducting a physics experiment to measure the acceleration due to gravity.
- Checklist: Correct setup of equipment, accurate measurement collection, proper calculation of acceleration, and thorough documentation of the process and results.

## Health and Physical Education

- Task: Demonstrating a series of gymnastic routines.
- Checklist: Correct form and technique, smooth transitions between movements, adherence to safety guidelines, and completion of all required routines.

## Product Design and Technology (Food)

- Task: Preparing a three-course meal adhering to nutritional guidelines.
- Checklist: Proper hygiene practices, correct measuring and mixing, adherence to recipe instructions, presentation of the final dishes, and nutritional balance.

## Digital Technology (Computer Programming)

- Task: Coding a simple game using a programming language like Python or Java.
- Checklist: Correct syntax, efficient code structure, functionality of the game, debugging and troubleshooting skills, and user interface design.

## VISUAL ESSAYS

---

The final type is an essay... of a sort. [The University of Hertfordshire](#) calls a visual essay “a [critical commentary](#)”, which I think is a perfect description of many kinds of essay, including traditional written ones. We want students to be able to give a critical, personal, and insightful commentary on their topic, whatever the form.

A visual essay is a curated series of images, either original or significantly processed (including using GAI), that together provide critical commentary on a specific topic, functioning as a form of argument or discussion. The sequence and layout of images, accompanied by captions or integrated text, guide the ‘reading’ of the essay. Despite seeming less demanding, creating a visual essay requires effort comparable to traditional academic writing.

A visual essay might be presented as a bound sequence, a series of unbound cards, or something like a PowerPoint slideshow, with the design and communication being crucial for its success. Like a traditional essay, it requires thorough research, organisation, and referencing, with an annotated bibliography using a referencing system.



The amount of images and text should correspond to the effort needed for a written essay of a particular word count, for example, 10-12 images with 500-700 words of text for a 1500 word essay equivalence. The annotated bibliography should detail the usefulness and application of each source in the visual essay.

### Here are some examples:

#### HISTORICAL EVENTS

- Topic: "The Evolution of Fashion: A Visual Journey Through the 20th Century."
- Description: This visual essay could depict the evolution of fashion throughout the 20th century, showcasing iconic styles from each decade alongside historical contexts that influenced these fashion trends.

#### ENVIRONMENTAL SCIENCE

- Topic: "The Impact of Plastic Waste on Marine Life."
- Description: A visual essay displaying the consequences of plastic pollution in oceans and seas, with images showcasing affected marine life, polluted areas, and comparisons of clean versus polluted waters.

#### SOCIAL ISSUES

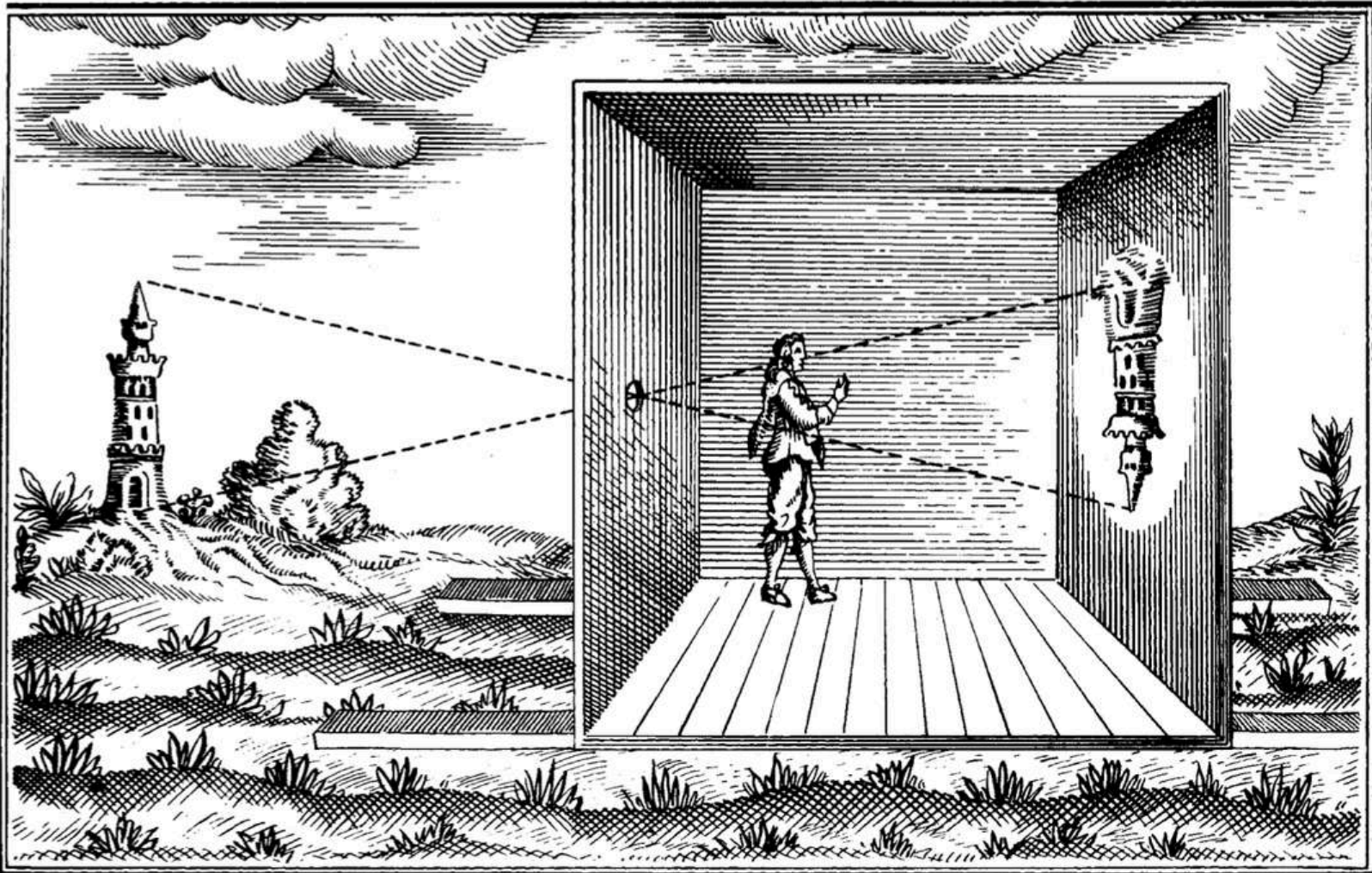
- Topic: "The Faces of Homelessness: A Glimpse into Life on the Streets."
- Description: This essay could present a series of portraits and living conditions of homeless individuals, aiming to humanise and shed light on the issue of homelessness.

#### TECHNOLOGY

- Topic: "The Rise of Smartphones: Transforming Modern Communication."
- Description: A visual essay illustrating the evolution of smartphones, their impact on communication, social interaction, and the juxtaposition of traditional versus digital communication methods.

#### HEALTH AND WELLBEING

- Topic: "The Mental Health Impact of Pandemic Lockdowns."
- Description: This essay could visually represent personal narratives, statistics, and scenes from daily life during lockdown, highlighting the mental health challenges faced by individuals.



A Renaissance artist included in a visual essay on “immersion” and visual layout.

Van Leeuwen, T. (2007). Sound and vision. *Visual Communication*, 6(2), 136–145. [A Renaissance artist included in a visual essay on “immersion” and visual layout.](#)



# RETHINKING ASSESSMENT FOR GENERATIVE AI: ORALS AND DISCUSSIONS

This post is part of a series on rethinking assessment in light of generative AI. The posts draw on research and resources from K-12 and tertiary to suggest ways that educators can design engaging, compelling assessments which shift the narrative away from GAI and “cheating”.

[In a previous post](#), I talked about the risks associated with trying to “catch” students, such as the ethical issues with detection software, the mistrust created by heavy-handed academic integrity policies, and the danger of false accusations. Unfortunately, we have a system that is heavily geared towards high-stakes summative assessments in written forms, such as essays and examinations. It’s a hard habit to break.

And, sorry to disappoint, but generative AI isn’t going to save us. Despite the hype around GAI, chatbots, and image/audio/video generation tools, these technologies are not going to “revolutionise the education system”. Nor will they lead to a renaissance of knowledge, a great levelling of educational inequity, or profound opportunities for personalised learning.



*The AI can't save us. Prompt: Sad computer graphic, desktop icon, retro. Model: Adobe Firefly*

Maybe I'm a little cynical, but I'm yet to see the benefits of edtech that for years has promised the world and delivered little. During COVID, the increased move to platforms and educational apps was touted as a means of reforming education, [but may have just further contributed to the digital divide](#) between those who can and cannot access the technologies.

In [a recent panel with the University of Melbourne's Sandra Milligan, NAPLAN's Stuart Mitchell, and Grattan Institute's Nick Parkinson](#) I said that generative AI will reinforce whatever system we have. That means that if we persist with high stakes, standardised testing and essay-based examination, we will see GAI which supports (or helps people to "game") those systems.

For example, whilst we'll see "personalised learning" chatbots, we'll also see "personalised NAPLAN tutors" and "essay helpers" flooding the market. Developers will build the tools that the system requires, and not necessarily the ones that will "revolutionise" the system. The former will make money quickly, the latter, maybe not at all.

## THE REAL OPPORTUNITY

---

The real opportunity doesn't come from the technology, it comes from the discussions we're having because of the technology. Like I wrote in the previous article, there is no way to detect generative AI. That means certain assessment types are now defunct. In fact, those assessment types have been inequitable and problematic for much longer than GAI – these technologies have just shone a light on the issue.

There appears to be a groundswell at the moment of people in both K-12 and tertiary calling for updated assessment practices. It may have been brought about by this latest wave of GAI, but it's down to the people in the system, not the technology, to make the changes.

So this series of posts explores what's already out there in terms of good assessment practice. I'm not relying on the technology to save us – just an understanding of the many diverse and effective ways to assess and provide feedback. Some will be supported by GAI, some will go without.

## ORALS AND DISCUSSIONS

---

I'm basing some of this post on a great document from [Eliana Elkhoury, PhD](#), which covers types of oral assessments, their characteristics, and examples in academic literature. Elkhoury's document is much broader in scope than this post, and I've selected just a few of the possibilities that might apply in various contexts. [The full list can be found here.](#)

Oral assessments are nothing new. Having students deliver presentations or PowerPoints is fairly standard in courses in both K-12 and secondary. Unfortunately, oral tasks often get relegated to being “tacked on” at the end of a unit as an additional assessment on top of the “real” written task. But oral assessments can and should occasionally replace, not simply add to, other modes of assessment in a unit of work. And it’s not all about solo speeches and slide decks.

The caveat over this entire series is that there are no “one size fits all” approaches to assessment. Oral assessments may cause anxiety for some students, or may be inaccessible due to language barriers, non-verbal or selective mutism, or other factors.

## ORAL PRESENTATION

---

Let’s get the most obvious kind of presentation out of the way first. Of course, one way for students to demonstrate their knowledge is through a presentation, solo speech, PowerPoint, or similar. This has the advantage of allowing an individual to demonstrate knowledge as opposed to a group, and is also a genuinely useful skill for many knowledge-based jobs.

Of course, students could either perform entirely tech-free, or use a variety of tools to help with oral presentations. If you wanted to incorporate GAI into an oral presentation, students could:

- Use GAI text generation like ChatGPT, Bing, Bard, or Claude to generate ideas, create scripts, edit, and so on
- Use an app like Gamma to create the slides which accompany the presentation
- Use image generation to create visuals, and add them to a standard format like PowerPoint (which currently has the AI-assisted Design feature, and will soon have Copilot)
- Use an app like Canva which includes GAI features such as text and image generation and AI assisted design

[As I wrote in an earlier post](#), there’s no way to guarantee students are not using GAI if they are completing part of the task out of class. This includes generating the scripts, but also using audio generation to create convincing versions of their voices for recorded orals. Like I said in that post, if you want it to be totally GAI free, it has to be a supervised task. Otherwise, accept that GAI might be used and move on. That logic applies for every assessment type in this series.

## DEBATE AND DISCUSSION

---

Debates and discussions have been an effective way of sharing, creating, and assessing knowledge since long before our current education system, existed. As well as being useful for assessing knowledge, debates and discussions can create healthy competition, strengthen critical and creative thinking skills, build communication skills, and contribute to more effective ideas.

Again, you could stage a debate or discussion in class with no technology whatsoever. It could be an informal conversation, a deliberately reflective practice like a [yarning circle](#), semi-structured like a [fishbowl](#) or [socratic seminar](#), or fully structured like a [formal debate](#).

- Use a chatbot as a “participant” in the debate, [bearing in mind all of the potential biases and flaws](#) in current language models
- Use a GAI generated text or image prompt as a stimulus
- Apply the Socratic Method directly using a chatbot, like [Adrian Cotterell suggests in this post](#) and like [I discussed here](#)
- Use GAI as a tool to record, transcribe, summarise, and synthesise points from a discussion, freeing up some of the time needed to assess the content so you can focus on delivery, teamwork, communication skills, and so on

You obviously don't have to assess every conversation that happens in a class, tutor group, or online discussion. However, these moments can provide useful insights into how individual students are contributing to the overall knowledge demonstrated through the unit.

# PITCH

---

I'm on the board of [Young Change Agents](#), a national not for profit that helps young people create meaningful social enterprises through programs like [\\$20 Boss](#), [Digital Boss](#), [Academy for Enterprising Girls](#), and various [Design Challenges](#). Throughout the YCA programs, pitching is a key aspect of getting an idea in front of an audience and persuading them to back it.

A pitch is also a great way for both individuals and groups to demonstrate their knowledge of a subject: if a student can't successfully pitch an idea, they might need to work on their content knowledge. It's also necessary in a pitch to get to the core of the idea, empathise with your audience, and develop strong arguments.

A pitch can be delivered off the cuff, but it's better to plan and prepare. Again it could be tech free, or might incorporate GAI in various ways, including:

- Using GAI as a mock audience member to test and refine ideas. Although a chatbot can't replace a real person when developing an idea, it can be a useful starting point
- Using tools like the ones listed earlier to create compelling pitch decks
- Testing the logic and persuasiveness of an argument against a chatbot
- Using GAI to write code for prototype apps and webpages, noting that a certain level of coding skill would be required to check for errors or issues

Pitches are great for persuading someone to back a project, product, or social enterprise, but can be useful for assessing knowledge to. Even something as simple as an elevator pitch or [Gaddie pitch](#) can allow a student to succinctly demonstrate what they know, without relying on a written response.

# LEARNING CONFERENCE

---

Elkhoury's [list of oral assessments](#) includes a [reference to this paper from Sindija Franzetti about learning conferences](#). In the article, Franzetti writes something which I think most of us can identify with: "Like so many of my colleagues, I resent grading for the labor and energy it takes away from doing the meaningful work of teaching to learn."

In response to this resentment towards grading assignments, Franzetti suggests learning conferences: individual conversations with students lasting 20-40 minutes which included reflection on the course, their participation, and the assignment. I've [written myself about the conferencing I used in my Y12 English classes](#) as an alternative to collecting piles of books. I have also [suggested ways that GAI chatbots can be used as part of the conferencing process](#), including:

**1. Pre-Class Preparation & Research:** Students can use a chatbot for researching and gathering information pre-lesson, with a reminder to validate accuracy.

**2. Writing Practice & Error Correction:** A GAI offers immediate technical feedback on writing, including grammar and structure, allowing students to correct errors and improve without waiting for individualised teacher feedback.

**3. Group Discussions & Peer Feedback:** A chatbot can support group discussions and peer feedback sessions, providing prompts, tracking contributions, and acting as a resource for small groups while teachers give 1:1 attention.

**4. Conferencing & Personalised Feedback:** During conferencing, a GAI provides additional context, individualised feedback based on unique needs, and supports follow-up questions, aiding teachers and addressing diverse student requirements.

**5. Vocabulary Expansion, Reflection, & Progress Tracking:** A chatbot suggests new vocabulary, guides students in reflection and goal-setting, and tracks progress, offering a comprehensive record and contextual examples for ongoing improvement.



# INTERVIEWS AND VIVAS

---

Interviews and [vivas](#) are traditional methods of oral assessment that allow students to demonstrate their knowledge, communication skills, and critical thinking in a structured conversation. These formats can encourage students to think on their feet and provide well-thought-out responses to questions or problems posed by an examiner or panel. Questions can be seen or unseen, and the accessibility needs of students should of course be taken into account.

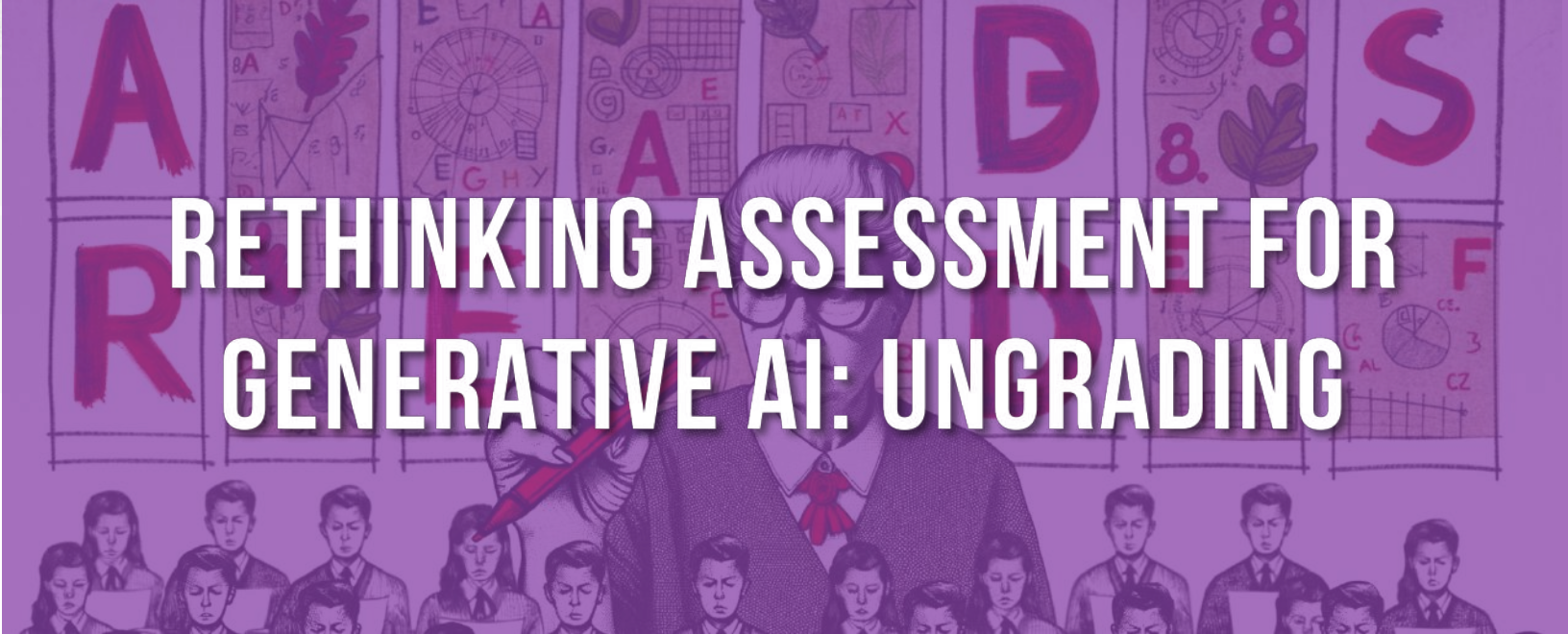
If you choose to integrate GAI in interviews and vivas, several strategies can be employed:

- Students could leverage GAI for preparing responses to potential questions, honing their articulation skills and refining their arguments
- Chatbots can serve as practice interviewers, providing an opportunity for students to simulate the interview experience and receive immediate feedback
- GAI tools could assist in organising and managing interview schedules, transcribing conversations, and highlighting key points for assessment

Incorporating GAI doesn't have to undermine the value of interviews and vivas but could add to the preparation, execution, and assessment. It goes without saying by this point that any use of the technology before, during, or after an interview should be appropriately acknowledged by both the students and the teacher.

There are many ways to use orals to assess knowledge, and they don't have to be seen as onerous or an addition to other forms of assessment. In future posts in this series, I'll be exploring other ways to develop assessment tasks more suited to GAI.





# RETHINKING ASSESSMENT FOR GENERATIVE AI: UNGRADING

**This post is part of a series on rethinking assessment for generative AI. Check out the earlier posts below:**

- 1 [Generative AI, plagiarism, and “cheating”](#)
- 2 [Rethinking Assessment for Generative AI: Orals and discussions](#)
- 3 [Rethinking Assessment for Generative AI: Beyond the Essay](#)

While Generative AI has thrown both K-12 and tertiary education into disarray, it's time to move the narrative away from “cheating” and to start focusing on what we can do about it. This post explores the concept of “ungrading”, an approach to shift the attention away from final scores and back to the learning.

## WHAT IS UNGRADING?

---

Ungrading is an approach that deviates from traditional grading systems, favouring a more feedback-centric model. Instead of focusing on scores or letter grades, the emphasis shifts towards providing detailed, constructive feedback, encouraging students to reflect on their learning and grow from their experiences.

Self-assessment and peer review are encouraged in ungrading methods. Alongside other teacher-led forms of feedback, students gain a better understanding of their own learning, the assessment criteria, and benefit from the diverse perspectives of their peers through reflection and critique.



Ungrading, as the name suggests, steers away from traditional letter and number grades. This could definitely help alleviate the fear and competition often associated with grading, and encourage more of the kinds of collaboration and team work we're always saying we want in the classroom. One huge issue with our current models of assessment is that they get in the way of genuine interactions like this: it's fine to say "work together", but if the end-game is a ranking process (like the senior secondary ATAR in Australia), then there's a contradiction.

Clear and transparent assessment criteria remains pivotal in ungrading. Guidelines and expectations are instrumental for effective assessment and meaningful feedback, irrespective of whether there's a letter or number attached.

If you want a thorough definition of ungrading, with an exploration of how it has been used and some directions for the future, [check out Jesse Stommel's blog post here](#)

## UNGRADING AND GAI

---

The de-emphasis of competitive grades and a final number means that ungrading has a lot of potential to help with assessment and GAI. As I've written previously, there are many reasons students cheat: academic pressure and the need to perform are among them.

[Emily Pitts Donahue](#), Associate Director of Instructional Support at University of Mississippi, and her students Abi and Trey [explore the benefits and challenges of ungrading in the context of GAI on Pitts Donahue's substack](#). Their story reveals the double-edged sword of GAI in assessment: something most of us are pretty familiar with by now. On one hand, GAI tools like ChatGPT can serve as learning aids, especially in assisting writing skills among students. On the other hand, these very tools can tempt students to circumvent the learning by outsourcing their thinking and writing tasks to AI.

Pitts Donahue labels an entrenched focus on grades as one of the core underlying issues driving students to engage in academic dishonesty. When the primary aim of education shifts towards attaining higher grades rather than gaining knowledge and honing skills, students are more likely to turn to GAI for completing their assignments. This grade-centric outlook not only undervalues the learning journey but also undermines the educational goals. It's in this scenario that ungrading emerges as a potential antidote to the GAI-induced challenges in assessment.

Pitts Donahue's classroom experiment with ungrading showed a promising shift in students' approach towards learning. By eschewing points or letter grades and instead providing extensive written feedback with opportunities for revision, the ungrading approach redirected the focus towards learning and improvement.

This also led to meaningful dialogues with students about the appropriate use of GAI in the learning process. When instances of misuse of GAI arose, they became springboards for discussing how such misuse impeded learning rather than just affecting grades. The shift from a product-centred to a process-centred assessment reinstated the value in the learning process and helped students to engage authentically with the material, without the temptation for misusing of GAI.

# CHALLENGES OF UNGRADING

Transitioning from traditional grading systems to an ungrading approach obviously comes with a set of challenges. Here are some of the challenges associated with ungrading, drawn from various perspectives (references at end) and summarised with ChatGPT (Model: GPT-4 with Bing search):

## EXTERNAL PRESSURES AND SELF-EVALUATION:

- One of the challenges stems from external pressures, particularly when self-evaluation is a component of ungrading. Given that grades often play a crucial role in admissions to further education and job selections, the pressure on students to achieve high grades remains, even in an ungrading setup.
- Not all students are equally equipped to self-evaluate, and there's a concern that some students might undervalue their work, while others might overvalue theirs. However, certain strategies like well-described rubrics can help mitigate potential biases in self-rating.

## MISINTERPRETATION AND MISAPPLICATION:

- The true essence of ungrading can easily be lost if not well-understood or well-implemented. A significant number of educators who attempt ungrading still rely on rubrics, stated learning outcomes, and other traditional grading elements, albeit under different terminologies. This misalignment with the core philosophy of ungrading doesn't change the students' assessment in any meaningful way, nor does it alleviate students' fears or competitive pressures associated with grading.

## TIME-CONSUMING FEEDBACK:

- Providing detailed feedback, which is a cornerstone of ungrading, can be time-consuming, especially in large courses. The process demands a substantial amount of time and effort from educators to ensure meaningful feedback that can guide students towards better understanding and improvement.

## SCALING CHALLENGES:

- Scaling the ungrading approach to larger classes or institutions with traditional grading ingrained in their systems poses a significant challenge. The logistical and cultural shift required for ungrading may face resistance or implementation hurdles.

### LACK OF CLEAR STANDARDS:

- In ungrading, the absence of clear standards or specifications might cause ambiguity for both students and educators. Although some versions of ungrading like specifications grading attempt to address this by setting clear standards for each assignment, the broader practice of ungrading might face challenges in defining success and understanding progress

### EDUCATIONAL CULTURE AND MINDSET:

- The entrenched culture of grading and the mindset associated with it can be significant barriers to the adoption of ungrading. Overcoming these cultural and psychological hurdles requires a concerted effort from educators, administrators, students, and the broader educational community.

## WHERE TO START WITH UNGRADING

Transitioning to an ungrading system isn't going to be easy. You'll meet institutional pushback, resistance to change, and no doubt resistance from students who just want to know that final number or letter. The transition should be about creating a culture of trust, feedback, and continuous learning, which can then help in mitigating the potential misuse of GAI technologies.

### TRANSPARENCY AND DIALOGUE:

Establish an open dialogue with students about the benefits and limitations of GAI through [academic integrity conversations](#), and why/how/where GAI should and shouldn't be used in the learning process.

### A FEEDBACK-RICH ENVIRONMENT:

Replacing grades with detailed feedback can help in shifting the focus from performance to actual understanding and improvement.

### PEER AND SELF-ASSESSMENT:

Encouraging students to engage in peer reviews and self-assessments can promote a deeper understanding of the learning material and the assessment criteria.

### SPECIFICATIONS GRADING:

Consider specifications grading as a variation, which bundles assessments together. Clear and transparent assessment criteria should also be a part of ungrading, helping to maintain high educational standards while promoting authentic learning.

### PROFESSIONAL DEVELOPMENT FOR EDUCATORS:

Preparing educators for this shift is incredibly important. Training on how to provide effective feedback and how to engage students in self and peer assessments will be crucial for the whole faculty or organisation.

Ungrading can be just another tool in the suite of methods you use to rethink assessment for GAI. It's been around since before GAI was on our radars, and there are plenty of studies out there exploring the benefits and challenges in both K-12 and tertiary contexts.

Alongside [orals and discussions](#), [alternative assessment forms to essays](#), and an understanding that [Generative AI doesn't automatically constitute cheating](#), ungrading could be a useful idea to carry into 2024.

## REFERENCES

---

Nilson, L. B. (2016, January 18). New ways to grade more effectively (essay). Inside Higher Ed. Retrieved from <https://www.insidehighered.com/views/2016/01/19/new-ways-grade-more-effectively-essay>

Stommel, J. (2023, April 6). What is Ungrading?: a Q&A. Jesse Stommel. Retrieved from <https://www.jessestommel.com/what-is-ungrading/>

Morris, S. M. (2022, October 21). The problem with ungrading? Everyone's doing it wrong. Times Higher Education. Retrieved from <https://www.timeshighereducation.com/campus/problem-ungrading-everyones-doing-it-wrong>

Kenyon, A. (2022, September 21). What is Ungrading? Duke Learning Innovation. Retrieved from <https://learninginnovation.duke.edu/blog/2022/09/what-is-ungrading/>

Pitts Donahoe, E. (2023, September 29). The Rise of Generative AI Calls for New Approaches to Grading. Emily Pitts Donahoe Substack. Retrieved from <https://emilypittsdonahoe.substack.com/p/the-rise-of-generative-ai-calls-for>



# Thank you!

**For more information on my online courses and  
webinars, head to  
<https://practicalaistrategies.com>**

**To stay up to date with new articles, free  
resources, and my consulting, advisory, and  
speaking engagements, visit  
<https://leonfurze.com>**

**If you enjoyed this free eBook, please encourage  
your colleagues to join the mailing list at  
<https://mailchi.mp/leonfurze/assessment>**

