

AI-Literacy Companion

A textbook for teachers

V0.81



AI-Literacy Companion

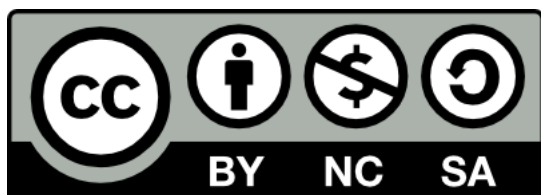
A textbook for Teachers

VU Centre for Teaching and Learning

Publishing date: 14 September 2026

License

© 2025 Vrije Universiteit Amsterdam. This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).



Authors

- Silvester Draaijer, Programmamanager VU Onderwijswerkplaats/VU Centre for Teaching & Learning
- Marianne Eggink, zelfstandig journalist en redacteur
- Esther Schagen, Universitair docent, Faculteit der Sociale en Geesteswetenschappen, afdeling Communication Science.

Contents

Chapters 1-9.....	1-8
1 Introduction	1-9
1.1 VU-UvA Taskforce and VU framework for generative AI in education	1-10
2 Using generative AI tool	2-11
2.1 Refusing or obliging students to use AI tools Error! Reference source not found.	2-11
2.1.1 Which tools can you use?	2-13
2.1.2 AI Powered Scientific Search Tools	2-14
2.1.3 AI tools to avoid or use consciously	2-14
2.1.4 AI Powered educational tools.....	2-15
2.1.5 Future developments	2-16
3 Generative AI and educational design	3-17
3.1 Where does generative AI fit in learning taxonomies?	3-19
3.2 AI Literacy in the curriculum?	3-23
4 The wicked problem of AI in Education	4-26
4.1 The wicked problem of generative AI in Assessment	4-26
4.2 The reputation perspective	4-27
4.2.1 Assessment and academic integrity.....	4-27
4.3 The wicked problem of generative AI in learning	4-28
4.3.1 Intentionally designed learning systems	4-28
4.3.2 Uncontrolled use of generative AI.....	4-28
4.3.3 So, are there solutions?	4-29
4.4 Negative impact on learning.....	4-30
4.4.1 Factors that lead to more cognitive offloading.....	4-30
4.4.2 Illusion of learning.....	4-31
4.4.3 Prevent cognitive offloading: planning courses and programmes.....	4-32
4.4.4 Prevent cognitive offloading: Designing appropriate tasks	4-33
4.4.5 Talking about learning and Modelling	4-33
4.4.6 Active learning	4-35

4.5	Examples of AI applications with a positive impact on learning	4-36
4.5.1	Example 1 – AI Tutorbots	4-36
4.5.2	Example 2 – AI avatars.....	4-39
4.5.3	Example 3 – Improving communication skills	4-40
4.5.4	Example 4 – Action research	4-41
4.5.5	Example 5 - Course guidance.....	4-42
5	Rethinking learning outcomes in the age of generative AI.....	5-44
5.1	Three levels, two drivers, two lanes	5-44
5.1.1	Dublin Descriptors: the ceiling that stays put	5-44
5.1.2	Programme final qualifications: two drivers of change	5-45
5.2	Course level: two faces per learning outcome.....	5-45
5.2.1	Effects-of: independent mastery, AI-free assessment.....	5-46
5.2.2	Effects-with: productive AI use, coached and process-oriented.....	5-46
5.2.3	Hybrid models: AI in preparation, restricted in performance	5-46
5.2.4	Handoff to the assessment chapter	5-47
5.3	Examples of revised learning outcomes	5-47
5.3.1	Same skill, different lane: example coding	5-47
5.3.2	Same programme, evolving lane: example law.....	5-48
5.3.3	Discipline-specific illustrations.....	5-48
5.4	How to derive these for your programme.....	5-49
6	Critical thinking.....	6-51
6.1	Knowledge is conditional for critical thinking	6-51
6.2	Logic	6-53
6.3	Logical fallacies	6-54
6.4	Moral convictions in relation to generative AI	6-57
6.4.1	Bias.....	6-58
6.4.2	Guardrails.....	6-59
6.5	What this could look like in your course	6-60
7	Generative AI and Assessment	7-62
7.1	Two perspectives on assessment	7-62
7.1.1	Assessment practices fo the two perspectives	7-63
7.1.2	Impact of generative AI.....	7-64

7.2	The two-lane approach	7-65
7.2.1	Lane 1 – Assessment of learning	7-65
7.2.2	Lane 2 – Guided use of AI	7-66
7.2.3	How to use results of lane 2 assessment	7-66
7.2.4	Orchestrating Lane 1 and Lane 2	7-67
7.2.5	Notes and implications	7-68
7.2.6	Conditions for making the two-lane approach work	7-69
7.3	Example: Writing papers and dissertations	7-72
7.4	Example: Genre writing	7-76
8	Generating feedback on assignments	8-79
8.1	Legal context: the AI Act and CE certification	8-79
8.2	Psychometric bounds	8-80
8.3	Social contract.....	8-80
8.4	Data privacy.....	8-80
8.5	Prompt engineering.....	8-80
8.6	Scaling up	8-82
8.6.1	One-by-one processing.....	8-82
8.6.2	Improving logistics of submissions and feedback	8-82
9	Grading open-ended exam questions	9-84
9.1	Psychometric validity.....	9-84
9.2	Applying the AI Act to grading	9-84
9.3	Example case.....	9-86
9.3.1	Test level.....	9-87
9.3.2	Question level	9-88
9.3.3	Student acceptance of LLM grading	9-89
9.3.4	Grading using LLMs at VU Amsterdam	9-90
9.4	Some scientific findings of the validity and reliability of using LLMs to grade exams?	9-90
10	References.....	10-90

1 Introduction

The advent of generative AI fundamentally challenges how universities design learning, support students, assess achievement and safeguard academic integrity. It unsettles two assumptions on which much of higher education still depends: that the work students submit is a reliable indication of what they have learned, and that achievement can be attributed with sufficient confidence to the individual student. When AI can generate fluent text, plausible arguments, code, summaries, reflections, or analyses within seconds, learning processes can be bypassed and evidence of achievement can become difficult to interpret. Educational administrators, program leaders, teachers, educational designers, quality assurance staff and examination boards are therefore confronted with the same underlying question: how can higher education respond when generative AI may invalidate familiar processes of learning, assessment and attribution? This Companion does not treat generative AI as a temporary problem to solve, but as a persistent condition that requires deliberate educational design choices.

p. 1–9

At the same time, generative AI also presents opportunities for higher education. It offers the prospect of more personalized learning, in which students receive adaptive feedback and support tailored to their individual pace and needs (Zawacki-Richter et al., 2019). AI tutoring systems, simulation tools and intelligent study assistants can extend the reach of teaching beyond contact hours and may help address accessibility barriers for students with language difficulties or learning differences (Perring, 2025). For researchers and teachers, generative AI can support literature review, writing processes and new methodological possibilities (Kusumegi et al., 2025). These developments invite educators to reconsider how time, attention, and learning activities are distributed in teaching, and how carefully designed learning environments might create more space for critical reasoning, ethical reflection and collaborative problem-solving.

Universities are responding to this situation in different ways. As Pearson (2025) describes, some institutions are moving quickly toward a broad adoption of generative AI, while others emphasize restriction or prohibition, and many are searching for a middle ground. Some are even wondering whether universities themselves risk becoming obsolete in an AI-rich knowledge environment (Clune, 2025). This Teacher Companion takes that institutional variety seriously, but starts from the position that the most useful question is not about banning or embracing generative AI, but rather how programs and teachers can make informed, proportionate and educationally justifiable choices.

Generative AI is also connected to broader societal concerns, including the growing influence of big tech, environmental impact, privacy, security and the public values at stake in educational technology. These issues are addressed more extensively in the [student version of the Companion](#). This Teacher Companion treats them as part of the context in which educational decisions about generative AI must be made. Readers are therefore encouraged to consult the Student Companion alongside this Teacher Companion, especially when discussing responsible use with students.

Universities and staff also have specific rules to follow. These follow from both legal regulations and university or faculty policies. In this Companion, we will discuss these as far as possible based on the VU Amsterdam context. But we emphasize that this book must be mainly be regarded as a textbook, not a policy document.

Finally, there is a big impact on research. Generative AI influences the way data can be collected and processed (Kusumegi et al., 2025). We will not explicitly dive into that in this Companion.

AI Literacy Companion for Students

This Teacher Companion, developed alongside the Student Companion, aims to equip teachers and academic leaders with the background knowledge needed to engage in informed discussions with colleagues, students, and administrative staff about the role of AI and generative AI in teaching, learning, and assessment. **It is emphasized that it is expected that teachers are aware of the Student Edition of the AI Literacy Companion and this Teacher Companion builds on the student version.**

All in all, knowledge about generative AI, the impact on learning, teaching and assessment, and the process, procedures, and culture on how to handle it effectively, must be on the radar of teachers and program leaders. It calls for growing AI-literate students and AI-literate academic staff.

1.1 The VU framework for generative AI in education

The leading document for faculties, program directors and teachers and the further institutional support units is the [framework for Generative AI in Education](#). This document was approved by the Executive Board in April 2025 and is VU policy. It is general by design. It sets the boundaries within which generative AI may be used in teaching and leaves faculties to write the rules that apply to a specific program. The document states that it will be revised regularly, so check that you are reading the current version.

p. 1-10

The Framework describes that VU Amsterdam has a duty of care to provide a safe digital learning environment, and the framework is cautious for that reason. It names three risks: data management, the quality of education, and the effect of AI use on students' academic skills. It accepts that there are opportunities and states that much more research is needed before they can be assumed. What it asks of you is a judgment about whether AI use serves the learning outcomes of your own course.

Five of its provisions apply directly to your teaching:

- You state in the study guide, the syllabus or Canvas whether and how generative AI may be used. The learning outcomes decide, and you are expected to have considered the added value beforehand.
- If AI literacy is not among your learning outcomes, students must be able to complete the course without generative AI.
- You cannot require a tool for which the VU holds no license.
- Nobody enters personal or confidential data into an AI tool. This applies to students, teachers and support staff alike.
- Where students may use generative AI, discuss it with them. Writing the rules down is the minimum, and a conversation is what makes the expectations clear, above all about assessment.

The framework also makes four recommendations to teachers, examiners and program directors. Build your own AI literacy. Review whether your intended learning outcomes still fit a field that AI is changing. Choose an assessment form through constructive alignment and decide there whether AI use is allowed. Follow the student's process through formative dialogue, with regular feedback moments in large writing assignments. That last recommendation does double duty, because knowing how a student writes is what makes an unexplained change in their work visible.

On fraud the framework is short. Three situations count: using AI where you did not permit it, presenting AI output as one's own work, and being incomplete about permitted AI use. The sanctions are set out in the Rules and Guidelines of your examination board. One condition holds across every application of AI in education, which the framework calls human-in-the-loop: a person always checks.

The policy also includes that experimentation with generative AI in education should take place in a structured and responsible manner. For teachers, it is expected that they do not experiment with using

and building apps in an uncontrolled manner. Experiments and pilots are not ad hoc activities, but should be registered and conducted via the VU Education Lab, where expertise, support, and knowledge sharing are coordinated. This ensures that experiences, risks, and best practices are systematically collected and inform broader institutional learning and decision-making.

1.2 VU-UvA Taskforce on AI Literacy

The joint VU-UvA Taskforce on AI in education positions generative AI not as a problem to be solved, but as a structural development that requires deliberate educational choices. Its [recommendations](#) emphasize three parallel responsibilities for universities: to provide clear institutional guidance, to safeguard values such as privacy, security and fairness, and to create space for experimentation and informed use in teaching and learning.

The Taskforce also developed an [advice document about AI Literacy](#). First, we note that there are a multitude of frameworks currently developed in various places. We will not dive into them here. For the sake of simplicity we will use this in this Companion, the UvA-VU Taskforce advice as a starting point as the set of learning objectives they formulated are quite straightforward. Second, later on this Companion, we will expand and elaborate more on that set of learning objectives.

p. 1–11

According to that document: “Teachers who help to develop AI literacy among students need, at the very least, a basic understanding of AI, regardless of whether they use AI themselves. Teachers need support to move in that direction and to develop their own AI literacy. VU CTL and UvA TLC are already working on programs for teacher professional development focused on AI in education.”

The Taskforce divides the specific learning objectives for students, but also staff, into three clear, categories (based on the Dublin Descriptors). These objectives are designed not only to equip individuals with technical skills, but above all to educate them as critical and responsible academic users:

What – Knowledge and understanding (The theoretical basis)

This learning objective focuses on understanding the underlying technology, so that learners can assess its capabilities and limitations. This entails that learners must:

- Understand how generative AI and LLMs (Large Language Models) work. This means they know that these systems make statistical predictions based on patterns, and do not actually ‘think’ or understand facts.
- Have an understanding of algorithms and data structures, including how the input of training data directly influences the output and can lead to built-in biases.

How – Understanding and Application (Practical Skills)

This concerns the functional ability to deploy AI effectively and methodically during one’s study or work:

- Targeted prompting: The ability to formulate effective, detailed instructions (prompting) to elicit high-quality and usable results from an AI system.
- Information evaluation: The ability to critically check and validate AI-generated output for factual accuracy, reliability and any ‘hallucinations’.
- Information selection: Knowing which data sources are suitable and safe for analysis using AI systems.

Why – Ethics and judgment (Critical reflection)

This is the academic core, in which learners learn to weigh up the broader impact of the technology:

- **Ethical awareness:** Understanding the risks posed by AI in terms of privacy, data security, intellectual property and copyright.
- **Social and environmental impact:** Being able to reflect critically on the broader consequences of AI, such as social inequality, filter bubbles, and the significant ecological footprint (the electricity and water consumption of data centers) of large-scale AI use.

The combination of the AI Literacy Companion for Students and this AI Literacy Companion for Teachers tries to fill in these learning objectives with substantial content and exercises for deepening understanding.

For the teacher Companion, additional learning objectives are of importance, namely how the use of AI impacts education, learning and educational design. But also what can and cannot be done from a formal and legal perspective (GDPR, AI-Act). Teachers need specialized **technological, pedagogical, and ethical knowledge** (TPACK) to design curricula that either mitigate risks induced by AI or reap the opportunities provided by AI. This includes the ability to adapt curricula as AI evolves. You'll read more about this in Chapter 3.

1.3 From the VU framework to faculty policy

The framework does not regulate what happens in a classroom. What it does is distribute responsibility, and the faculty is the level at which that responsibility turns into a rule you have to follow. The framework sets out the chain as follows.

- **The Executive Board** is responsible, from an institutional perspective, for an organization that handles AI responsibly. It commissioned the VU-wide Artificial Intelligence Competence Network for that reason.
- **The faculty board** is responsible for translating VU AI education policy into faculty policy.
- **The examination board** draws up the Rules and Guidelines, and records in them when AI use counts as fraud or plagiarism. It can advise the examiner on whether the text in the study guide is in line with those rules.
- **The teacher or examiner** states in the study guide, the syllabus or Canvas whether and how (gen)AI use is permitted, and can draw on the technical, legal and didactic expertise of the Competence Network.
- **The student** is responsible for the content of the work submitted.

Four faculties have filled this in so far, and three of them are described here. The Faculty of Science was first, in April 2025, with a four-part guideline that it calls a living document. SBE followed in October 2025 with a policy that is, in its own words, heavily based on the one from Science. The Faculty of Medicine went its own way in November 2025, building on the AI-GO framework for AI literacy and on the AI Act; its vision has been approved and its policy was still in draft at the time of writing.

That faculties copy from each other is intended rather than accidental. The framework names the SBE approach explicitly as an example that other faculties can adopt, so that teachers do not all have to invent the wheel themselves.

What this means in practice: before you write your own rules for a course, find out whether your faculty has already decided the question for you. Much of what follows in this Companion is a choice at course level only where the faculty has left it open.

2 Bottomline use of generative AI tools

Teachers and programs must, as said, operate within institutional and legal bounds. But often the VU Education Lab receives questions such as: What can I do as a teacher? What should I avoid? Can I oblige students to use generative AI? Can students refuse to use generative AI? Should I prohibit the use of generative AI or can I leave this up to the students? What AI system can I use? We think they need attention from the outset.

We emphasize again that within VU Amsterdam, the '[Framework for generative AI in education](#)' is leading. But the framework is quite general and does not give (or intends to give) direct answers to the above questions, as many decisions are very context-dependent. The following is stated regarding the use of generative AI by students in the Framework: "It must be clear that the teacher or examiner indicates in the study guide, syllabus, or Canvas whether and how the use of generative AI is permitted. In addition, the learning objectives of the course are always leading. If AI literacy is not included in the learning objectives, it must be possible to take the course without the use of (generative) AI. In addition, the teacher must consider in advance the potential added value of students' use of AI for their education."

p. 2-13

2.1 Refusing or obliging students to use AI tools: what are the rules

The statement in the policy has multiple implications that can give some clear answers regarding the tool use by students and teachers. They do not solve the educational dilemmas we will dive into deeper later, but they do set the stage a bit clearer. Below you'll find the answers to frequently asked questions by teachers about AI (FAQs).

Important note!

Always consult the information on <https://vu.nl> or from your Faculty on specific policies, regulations or amended guidelines

1. **Must I explain in my syllabus how students need to deal with generative AI?**

Yes, according to VU Policy, you **must** always specify and explain for your course in your syllabus how students must deal with generative AI and why.

2. **Can I prohibit the use of generative AI by students?**

Yes, you can prohibit the use of AI. AI use by students will then decrease, but never entirely. Be aware of the consequence: the required monitoring or policing of students.

- a. This is likely impossible to enforce with assignments that students work on at home without supervision (e.g. take-home exams, case studies, projects and report writing).
- b. This is possible in supervised conditions (e.g. exams, orals, presentations, and interview).

3. **Can I impose different levels of allowed use of generative AI?**

Yes, you can define such levels. Even within your course for specific parts of your course. Some faculties propose to use *specific ladders* of AI use, e.g. the [AIAS scale](#) (Perkins et al., 2024). VU faculties do: both the Faculty of Science and the School of Business and Economics work with a four-level scale derived from it.

- a. **Additionally:** you can require students to write a transparency statement in which they explain how they used generative AI in producing their work. Apps exist that can help students write this statement in a comprehensive way, for example

<https://incantalupo.github.io/statement-of-work-generator/> (the VU Education Lab is working on a version for VU Amsterdam). A second generator, made at VU Amsterdam as part of Mary-Jo Diepeveen's PhD project, is at <https://madiepev.github.io/academic-ai-integrity/disclosure-en.html>, with a Dutch version alongside it. It asks for the tools and versions used, the tasks they assisted with, a few representative prompts, an estimate of how much of the work they carried, how the student checked the output, and a confirmation that no confidential data went in. The result is a disclosure statement the student can paste under their work.

- b. **However**, be aware of the consequence: the required monitoring or policing of students and increased levels of student anxiety about the precise meaning of such levels.

1	NO AI 	This task is completed in a controlled environment designed to exclude AI. Knowledge, understanding, and skills are demonstrated and assessed independently.
2	AI PLANNING	This task focuses on planning activities such as topic exploration, outlining, and initial research. AI may be used to support this process, and the quality of planning and idea development is assessed whether or not AI was used.
3	AI COLLABORATION	AI may be used to help complete this task, including idea generation, drafting, feedback, and refinement. It is designed so that AI alone will not reach the required standard. Assessment covers both the work itself and how AI outputs are evaluated, modified, and integrated.
4	FULL AI	There is an expectation of AI involvement in this task. The goal cannot be reached by AI or by a person working alone in the time available. Assessment focuses on the critical thinking and subject knowledge shown in directing AI.
5	AI EXPLORATION	This task is designed for creative AI use to solve problems, generate novel insights, or develop innovative solutions in the discipline. Approaches may be co-designed by students and instructors.

2-14

Table 2.1 AI Assessment Scale (AIAS), v2.1, Copyright Learning Innovation Practice Ltd. Based on the original AIAS.

4. Under which conditions can I ask students to work with AI?

The **conditions** are these: You may only require generative AI tools for which VU Amsterdam has a license, or which run on local computers. And you must state in your syllabus that AI use is necessary to achieve the (AI literacy) learning objectives (see also Chapter 5).

Your faculty may go further than this. The Faculty of Medicine, for instance, allows AI use to be made compulsory only where it is needed to achieve the learning outcomes, and requires the program to offer an equivalent alternative assignment to students who would rather not use AI. It also limits teaching to AI software that VU Amsterdam has officially verified. Check your own faculty policy before you make AI use a requirement.

5. Do students have the right refuse to use generative AI for their course?

Yes, students have the right to refuse to use generative AI, **unless** it is stated in the syllabus that AI use is necessary to achieve the (AI Literacy) learning objectives (see also Chapter 5).

6. May students use generative AI for generating feedback on their work such as essays, reports and case studies?

If this particular use is allowed by you in your regulation laid down in your syllabus, the answer is **Yes**. Students can do this, but students must be aware of the limitations and variability of generated responses. Otherwise, it is prohibited.

7. **Can I use generative AI detectors to check student work?**

No, using such systems is not allowed because of their limited reliability (false positives and negatives) and privacy and copyright issues. Using Pangram for this purpose is also not allowed.

8. **Can I force students to use generative AI?**

Yes, you can, **provided** that you are following the conditions under 4.

The premise about this rule is, firstly, that we as a university must improve students AI Literacy by law (Digitale Overheid, 2025). Secondly we have an obligation as a university to prepare students for their future profession in which generative AI plays an increasingly big role.

9. **Can I use generative AI for generating *feedback* on student work such as essays, reports, and case studies?**

No, VU Amsterdam has not yet taken a formal position about this topic because of the complex legal requirements that need to be met and the variability of feedback generated by AI. For now, feedback may only be established by humans.

10. **Can I use generative AI for generating *grades* on student work such as essays, reports, and case studies?**

No, you cannot, VU Amsterdam has not yet taken a formal position about this topic because of the complex legal requirements that need to be met and the variability of grades generated by AI. For now, grades may only be established by humans. See Chapter 9 in which we describe possible practices and procedures that might be employed in some future.

11. **May I upload documents in generative AI systems in general?**

Yes, in general you can, but, within the context of use in a professional context of VU Amsterdam, you must limit this to systems that are licensed by VU Amsterdam. In all other cases, make at least sure that your data is not used for model training (most suppliers offer that opportunity).

12. **May I upload documents in generative AI systems from a copyright perspective?**

There are no clear answers for this question and the legal regulations are shifting and are interpreted in multiple ways. The University Library has [a knowledge item](#) about this topic from a copyright perspective with example use cases. If you want more information, get in touch with the University Library's [Copyright Information Point](#) (CIP).

13. **Can I use generative AI for generating grades on open-ended questions for exams?**

No, you cannot do this on your own because of regulations in the AI Act in which giving grades is categorized as a high risk application that requires careful organisational, legal and procedures to be in place. Further, the variability of grades generated by AI must also be brought under control. See Chapter 9 in which we describe possible future practices.

14. **Can I use my own Vibe-coded App in my teaching?**

No, you cannot. Vibe coded Apps are prohibited as they can impose security and privacy threats. Please contact the VU Education Lab to investigate how to proceed.

Box 2.1 - Shadow IT

Shadow IT refers to the use of IT applications that are not officially offered or approved by the internal IT service of a company. Shadow IT emerges because employees need faster solutions than IT can provide), find official systems too restrictive, lack awareness of policies, or prefer more user-friendly tools.

While motivations are often legitimate and productivity-focused, this creates a parallel IT infrastructure operating outside organizational oversight with significant risks including security vulnerabilities, data breaches, compliance issues, integration problems, lack of official support, and hidden costs.

At VU Amsterdam, the use of shadow IT is a known issue as well. It is not permitted, and it is recognized that it happens nonetheless. From the perspective of data privacy and copyright (distributing copies of copyrighted material, and supplying data for model training), the rule is to use only applications that VU

Amsterdam licenses and for which a Processor Agreement has been agreed. The most common applications and their appropriateness are listed below.

2.1.1 Which tools can you use?

The tools below are licensed by VU Amsterdam and have an agreed Processor Agreement in place. As a teacher, you cannot oblige students to use these tools. However, you could point your students to these tools if it could benefit their learning or tasks. When you do, discuss proper use with them, including attention to privacy, security and copyright, and, especially with video, environmental impact..

2.1.1.1 Microsoft Copilot for Web

Since February 2024, Microsoft Copilot for Web is available to VU students and staff. To use this chatbot compliantly, make sure to log in with your **VU Amsterdam credentials**.

Additional notes to keep in mind:

- VU Amsterdam offers the simple version of Copilot: Copilot for Web. VU Amsterdam has judged that user privacy is sufficiently guaranteed. However, Copilot does use Bing to conduct real-time searches on the internet as well. Users should be aware of this.
- VU Amsterdam does **not** have a license for 'Microsoft Copilot 365' and did not make it available to employees or students.
- Microsoft nor VU Amsterdam Employees have access to the conversations of individual users.
- Data is not used to train models and LLMs.
- Still, in all cases, users should refrain from uploading special and sensitive information via Copilot for Web such as research data, company information and the like.

p. 2-16

2.1.1.2 Google Gemini and Gemini Notebook

VU Amsterdam has a long-standing license agreement with Google for Google Workspace for Education. Google Gemini and Gemini Notebook (formerly NotebookLM) are integral parts of this license. To use these tools securely, you need to log in with your **VU Amsterdam credentials**.

- VU Amsterdam has judged that user privacy is sufficiently guaranteed. However, Gemini does use Google Search to conduct real-time searches on the internet. Users should be aware of this.
- Still, in all cases, users should refrain from uploading special and sensitive information via Gemini or Gemini Notebook.

2.1.1.3 Doing research with LLMs

ITvO, short for IT voor Onderzoek or IT for Research, is the team inside the VU IT department that arranges infrastructure for research projects. Three of its services matter for work with language models. SciStor is scientific storage in the VU data centre. SciCloud lets you create and manage virtual machines. Ada is the university's HPC cluster. The capacity on Ada is modest, and for heavy training or very large models ITvO points researchers to Snellius, the national supercomputer run by SURF. Requests run through the VU Service Portal under the service domain Research, or by mail to itvo.it@vu.nl. An account on Ada needs endorsement from a supervisor.

[Nebula](#) is a service of the Network Institute at VU Amsterdam. It opened in March 2025 and gives VU researchers access to open language models that run on the university's own machines. Access is free at the time of writing and runs through a request form that needs a VU login. Nebula is offered to researchers for research; teaching use is not part of what the Network Institute advertises. Running

models on campus does not lift the rules on personal data either. Sensitive data is not allowed on Nebula, and the platform publishes a DPIA checklist to help you work out what your own project needs. You can find more details about Nebula at the website <https://nebula.cs.vu.nl/welcome>.

2.1.1.4 Locally run models

The safest way to work with data in AI tools is to use locally installed LLMs. This means downloading an LLM model on your computer and running it with, for example, LM Studio (Mac) or Ollama. The data will stay locally on your computer.

From a privacy and security perspective, this has advantages. First, you need a powerful machine to run models with sufficient capabilities and speed. Second, the data may reside on your computer, but this often introduces other kinds of vulnerabilities you should consider (e.g. make sure your computer or laptop and the data on it is safeguarded against theft).

The VU Education Lab can demonstrate how locally running models function.

2.1.2 AI Powered Scientific Search Tools

p. 2-17

In the student edition of this Companion, you can find a (growing) number of specialized AI-powered search tools. In general, VU Amsterdam does not have a license with these companies.

2.1.3 AI tools to avoid or use consciously

VU Amsterdam does **not** have a processor agreement with OpenAI (nor Claude, Mistral, Qwen, Grok). So, everything you share with these systems can potentially be used by those companies for training purposes or tracking conversations and uploaded information. Their license structure and contents may vary, and to use these systems consciously and responsibly requires at the very least detailed knowledge about their user agreements. Therefore, it's important to never share personal data or other sensitive information with these companies (such as student names, e-mail addresses, grades, financial data, research or personal questions). Also check that you make the setting that your conversations and data are not used to train their models.

From a different perspective, especially to generate learning materials, privacy and security risks are relatively low, and it is up to the personal preference of teachers to decide what to use. In fact, a plethora of applications and publishing tools are now finding their way to the market. It is almost impossible to keep track of them. The VU Education Lab keeps a record of [AI-driven tools for AI-generated content](#) for educational purposes, from text, diagrams images up to video's. But also, the large publisher houses incorporate AI-driven techniques to develop content and extend interactivity of their materials.

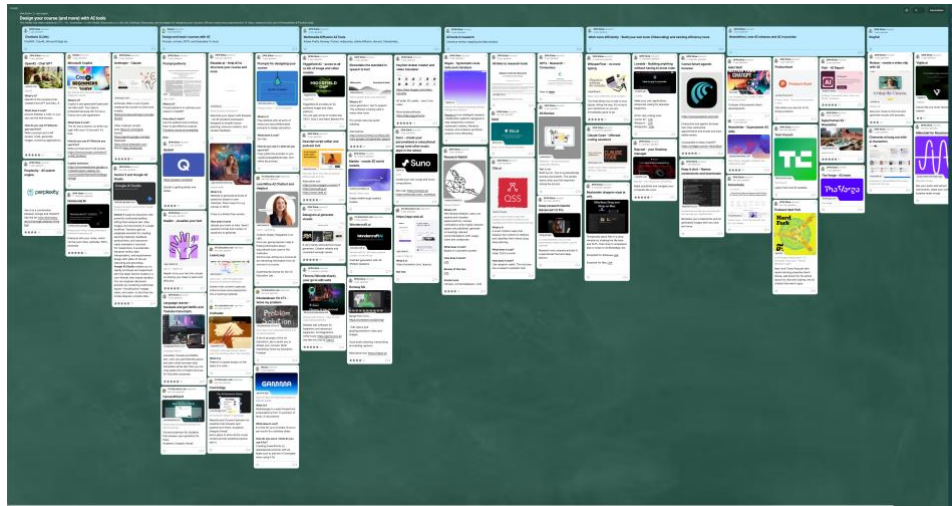


Figure 2.1 Image of the VU Education Padlet with AI tools for Education

2.1.4 AI Powered educational tools

p. 2-18

A (growing) number of specialized AI functionalities are incorporated into existing educational tools such as [Canvas](#), [Mentimeter](#) and [FeedbackFruits](#). VU Amsterdam has license agreements with these companies and privacy and data security are compliant.

VU CTL conducts experiments with these AI tools in those educational tools to test their functionalities and educational (added) value (such as the generation of educational materials, videos, knowledge clips, AI bots, feedback generation, grading, and AI Avatars). VU CTL looks for the balance between the added value and the cost these systems carry, both in terms of licensing cost, but also the wider educational or societal or environmental impact cost.

New tools will also continue to emerge. In February 2026, [Cogniti](#) was introduced at VU Amsterdam as a pilot by the VU Education Lab. This tool will likely become more widely used at VU Amsterdam. Cogniti is a system with which a teacher can build specialized bots and mini apps via [edubots.vu.nl](#). Cogniti runs on the private and secure VU Azure environment. Conversations and data are not used to train LLMs, and Microsoft has no access to conversation and data. You can find more information about Cogniti at the [Canvas Tools for Education](#) site.

Box 2.2 – The ‘free version’ problem

Universities (including VU Amsterdam) are working towards equal access of students to high-quality AI tools, such as EduGenAI from SURF to prevent inequity. Marc Watkins however, writing from OpenAI's 2026 Education Summit (2026), describes a structural problem that universities face when considering institutional AI licenses: students are already using the free (or paid by themselves) version of ChatGPT (or other generative AI tools). This induces a core tension. Universities are being asked, or find themselves obliged, to invest in purpose-built educational versions of AI tools, while students often have little practical reason to abandon the free alternatives they already use and find good enough.

The problem is compounded according to Watkins, by a strange kind of privacy paradox: many students may avoid institutionally provided AI tools because they may distrust whether their use will be monitored or reported by teachers. For instance, if they use AI inappropriately in class. Watkins implicitly assumes that students may trust commercial vendors better, or that students see ‘distant’ and ‘abstract’ processors of their data less as a threat than individuals they may encounter in the classroom. This distrust makes adoption of licensed tools unlikely even when universities do invest in them. Meanwhile, agentic AI

tools — such as AI-enabled browsers — can now automate tasks entirely, including logging into and operating university-provided AI tools on a student's behalf, which makes any institutional guardrail approach increasingly difficult to enforce.

According to Watkins, OpenAI offered no clear roadmap for shifting students from free to educational versions of its tool, and that gap remains unresolved. The deeper issue according to him is that ChatGPT is a general-purpose consumer product, not one built for learning. Adding pedagogical friction to a licensed version only works if students cannot simply sidestep it with a free account. Or if the institutions and teachers can convince and motivate students enough to use the institutional licensed services as the better and more private and secure alternative.

2.1.5 Future developments

The technology and vendor landscape is changing quickly. VU Amsterdam, SURF and IT constantly monitor developments and will provide alternatives if needed in due time. For now, the most up-to-date information can be found on this [page about Generative AI Chatbots](#) on Tools for Education.

3 Generative AI and educational design

Educational design is the intentional process to design curricula and courses to elicit intended learning behavior by students to achieve intended learning outcomes.

A first way of looking into generative AI and its influence on educational design is going from macro to micro level (King's College London, 2025). The macro level covers university-wide principles and policies on generative AI. VU Amsterdam has existing guidelines on this topic as described in [Framework for Generative AI in Education](#). The middle level addresses the implications of generative AI at the department and study program level. Several faculties and programs at VU Amsterdam have made their own more specified policies (among them the Faculty of Science, SBE, the Faculty of Medicine and FGB; the first three are described in Section 1.3). Finally, the micro level concerns the use of generative AI within individual courses and by individual academic staff. The central question at this level is how generative AI impacts learning outcomes, how it can support student learning or make teaching tasks qualitatively better or more efficient. Generative AI raises important questions of which the most pressing is how we can ensure that students retain meaningful learning experiences when AI is involved and how can we assess that appropriately and effectively.

p. 3-20

Another perspective is the learning orientation regarding generative AI in educational settings:

- Learning *about* AI means building a foundational understanding of what AI is, how it works, and what its limitations are, as well as their impact on society and learning. Without this base, critical use remains impossible.
- Learning *with* AI uses AI as a tool in the learning process. For example using personal tutors, role-playing, Socratic dialogue bots and the like. Outcomes are promising but highly context-dependent.
- Learning *through* AI makes the encounter with AI itself the learning event, for example by having students critically compare, verify, or challenge AI-generated output.
- Learning *against* AI means deliberately working without it to build the cognitive schemata that AI cannot replace. Genuine understanding develops where AI influence is lowest. A practical example are computer free and AI free workgroups.
- Learning *despite* AI names the ambient risk: unguided AI use is associated with cognitive offloading, reduced engagement, and poorer recall.

Effective course design requires deliberate choices about when to introduce AI, when to restrict it, and how to make the student-tool relationship itself a subject of reflection. Together, these five modes suggest that no single stance toward AI is adequate. Effective course design requires deliberate choices on these aspects.

For this Companion, *Learning How to Use Generative AI*, also relates to the ability of teachers on how to use generative AI for application in the educational setting. So, this Companion plays out at meta-level. In order for teachers to design effective courses and instructions, they must have practical AI skills themselves. Also, they must know about learning processes and learning designs in general. Given the ubiquitous presences of generative AI, they must also know how generative AI influences learning, how it influences the content and methods within the discipline they teach. Together these building blocks should lead to the ideal outcome of effective courses and instruction. Represented in a table format, this looks like the following:

Requirement	Focus Area	Goal
Practical Skills and knowledge about broader consequences	Personal Competence	Mastery of the tools, methods and techniques, Knowledge of broader impact.
Learning Design	Pedagogy	Knowledge of how people learn.
Learning Influence	Student Experience	How AI alters the learning journey.
Disciplinary Influence	Content/Subject	How AI changes the field itself, both in content or methods.

Table 3.1 Requirements for teachers regarding impact of generative AI for their knowledge and skill.

This process can also be represented by the so called TPACK-model (Mishra & Koehler, 2006). The model shows that for an effective pedagogy, teachers need to have knowledge and skills of pedagogy, technology and content. The intersections of these three, now regarding the new and transformative technology of generative AI, presents the challenges that curriculum and course designers and teachers need to address.

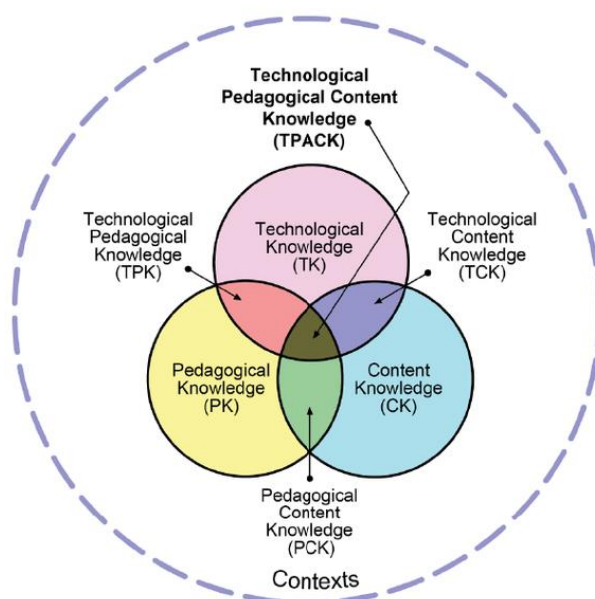


Figure 3.1 TPACK model according Mishra and Koehler (2006).

Regarding didactics, this Companion assumes that you as a reader are familiar with the concept of constructive alignment: the principle, introduced by John Biggs (1996), that *learning outcomes*, *teaching and learning activities*, and *assessment* tasks should be deliberately aligned with one another. The core idea is that students learn what they are assessed on, so assessment and instruction must point in the same direction. When generative AI enters the picture, this alignment becomes more fragile. This is because you cannot be certain that, for example, for classic writing assignments, students are actually learning, or using AI. This means that course designs relying on product-based assessment alone are particularly vulnerable. This requires a shift in focus towards the learning process. Still, constructive alignment therefore provides a practical lens for the decisions discussed in this Companion: what should students be able to do independently, but also in combination with the new technology. And how can learning activities build toward that, and what does this require from the assessment program?

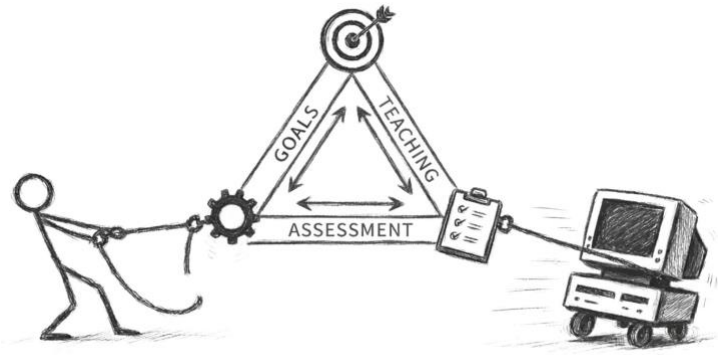


Figure 3.2 Picture of UvA TLC e-learning depicting the tension in the Constructive Alignment concept induced with the emergence of generative AI (TLC University of Amsterdam, 2025).

3.1 Where does generative AI fit in learning taxonomies?

Scholars are trying to work out what kind of impact generative AI has on the taxonomy levels of learning and instruction. Michelle Kassorla (2025) developed an interesting perspective on this. She argues for a so-called ‘inverted Bloom’s taxonomy’. She proposes that the traditional Bloom Taxonomy, that seems to prescribe progressive learning from remembering through comprehension to creating, is fundamentally inverted in the AI era. Her key argument is: "Students now create first, then understand later," reversing the cognitive hierarchy that educators have relied on for decades.

Kassorla begins by clarifying that Bloom's Taxonomy (conceived in 1956) was never meant to be a prescriptive learning framework, but rather a classification system for organizing educational outcomes to be used by educational designers. The purpose was primarily to design instruction and give teachers and designers a clear framework to design learning units and tasks. The taxonomy helps to not over-represent or underrepresent certain task types given the goals of an educational unit.

The view emerging from this classification and misunderstood view, however, was that students build understanding linearly from foundational knowledge upward to higher-order thinking skills like analysis, synthesis and evaluation.

This view on acquiring knowledge and skill has already received fundamental criticism. For example, the pyramid visualizing Bloom's taxonomy, was not part of the original framework but added later by an unknown author. And it is precisely this pyramid shape that suggests a form of hierarchy that likely does not exist. Learning is a messy process, and moreover, with the highest level appearing to be the goal of any learning unit. Instead, scholars now suggest that knowledge is a base condition and that different goals (or task types) exist alongside each other (Pedro, 2017; Wiliam, 2017). The concept of higher and lower order in ‘thinking’ in the pyramid actually does not exist at all. Think of learning to ride a bike. You do not first analyze your balance theoretically and then get on. You just ride, fall, adjust, and gradually build understanding through doing. Knowledge and skill develop through each other, not in sequence.

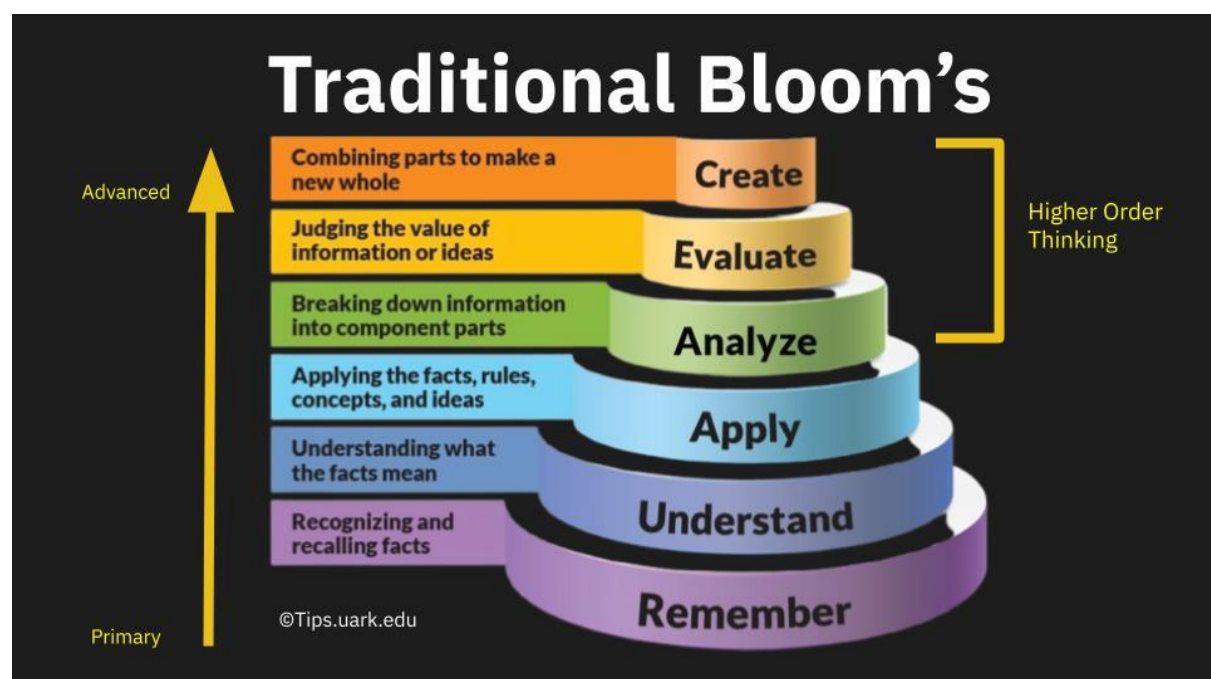
It would be better to see knowledge as something that is basic or automated, but dependent on context. For instance, some students who enter university have basic knowledge of the system. They know what words like ‘credit points’ or ‘examination board’ mean, and implicitly understand how and when to email a teacher. This means they don't struggle understanding these basics, like for instance a first generation student might. And so have more cognitive space to actually study, instead of spending energy to decipher the system and culture. This knowledge is so ingrained it has become invisible, and learning it is not a straightforward path through something like a pyramid. See also Box 6.2 on fast and slow thinking systems.



Figure 3.3 Depiction of Bloom's taxonomy is seen as more appropriate by Dylan Wiliam.

This criticism is further amplified in the age of generative AI because generative AI can seemingly instantly explain concepts and procedures, produce applications and compare theories, dilemmas and solutions, analyzes information and artifacts, and produce creative work. With AI students (seemingly) can skip directly to creation without passing through the stages of remembering, understanding, or applying foundational knowledge. You might think that students miss out on important learning stages this way, or are easily satisfied with AI output. But according to Kassorla the order of learning stages is simply reversed. Students take the output and then go through the other stages after. As educators, it is our task to guide students through these new paths.

In Kassorla's **Inverted Bloom model**, the hierarchy flips to (from bottom to top): create → evaluate → analyze → apply → understand → remember. Crucially, she describes decreasing agency levels as students move up the hierarchy, as AI assistance is heaviest at the creation stage and progressively lessens as students engage in deeper cognitive work. Each stage builds upon the previous one, but in reverse order from traditional models.



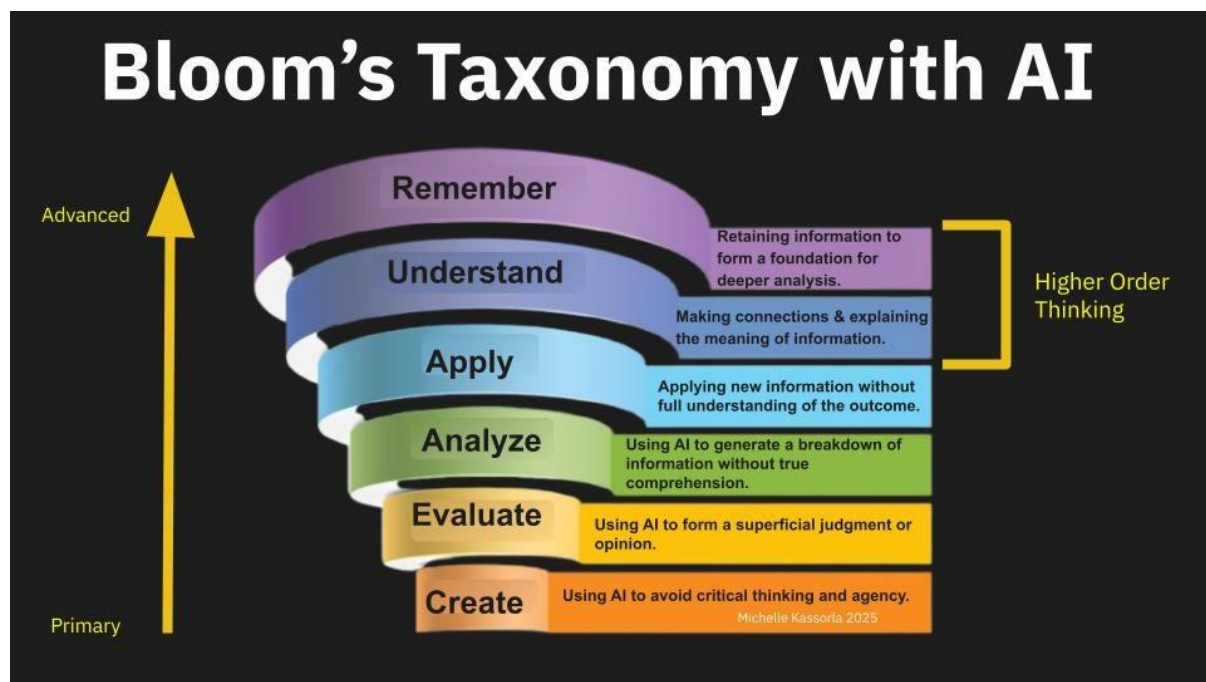


Figure 3.4 Comparison of the Bloom's taxonomy in an age of AI according to Michelle Kassorla (2025).

At the bottom tier (create), students have minimal agency; they write prompts and submit AI-generated outputs with little engagement. But then, given good educational design and intent by students, they will need to move up. As a teacher it's an important responsibility to encourage students to gradually take more active roles in their learning: evaluating AI outputs for correctness and logic, analyzing the feedback they receive, applying suggestions to revise their work, understanding why changes matter through reflection, and finally remembering and internalizing principles for future use. Guiding students back and forth between the different levels of learning is an important teacher task.

The critical insight is that genuine learning happens in the upper tiers, where AI's influence diminishes and students' own cognitive engagement increases. So, remembering and understanding are currently seen as the real higher order learning outputs when going from the stage of creating to remembering.

Austin (2026) even goes a step further. She developed UnBlooms™. Her insight is that learning is not a pyramid and does not follow a linear path. She sketches learning as a loop from the top and spiral from the side. It is recursive, nonlinear, and context-dependent. A student doesn't need to memorize before they can critique. They can enter this cycle at any point. What matters is that they're doing the thinking, not outsourcing it.

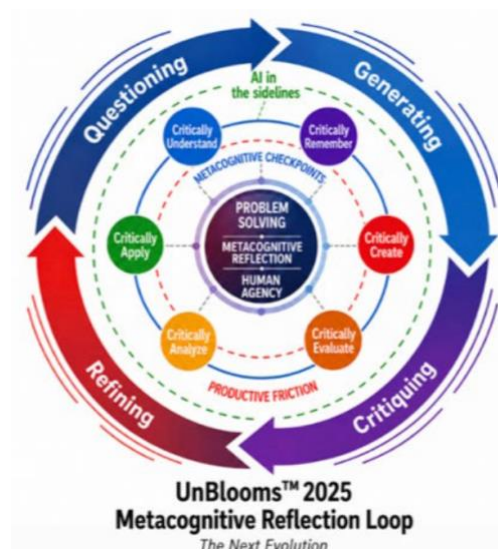


Figure 3.5 The spiral representation of UnBloom™ learning, developed by Tina Austin.

UnBlooms™ Levels of Critical Engagement with AI FROM AWARENESS TO AGENCY

Phase WHEN IN THE PROCESS?	Level STAGE	Critical engagement WHAT LEARNERS DO
Before ACCEPTING OR RELYING ON AI <i>Build awareness and set intent.</i>	1	Recognize the difference between AI and human reasoning. <i>Notice what is AI-generated, what is human knowledge, and why it matters.</i>
	2	Identify specific errors, gaps, or questionable elements. <i>Spot inaccuracies, missing information, and potential bias.</i>
During ENGAGEMENT WITH AI <i>Think critically in real time.</i>	3	Analyze the assumptions within the AI's reasoning. <i>Examine what is being assumed, what is missing, and how the reasoning is constructed.</i>
	4	Evaluate the broader implications and make a justified judgment. <i>Weigh strengths and limitations, consider multiple perspectives, and decide what to trust.</i>
After EVALUATION <i>Apply, create, or choose not to use.</i>	5	Create a better workflow or solution, or Resist AI use with justification. <i>Design improved approaches, apply learning to new contexts, or intentionally choose not to use AI when it is not appropriate.</i>

Figure 3.6 UnBlooms™ Levels of Critical Engagement with AI by Tina Austin is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

The reframing addresses a critical pedagogical problem with using generative AI for cognitive off-loading: students can appear highly productive by creating with AI while remaining cognitively disengaged and learning nothing. Kassorla's model emphasizes the need to guide students toward genuine comprehension and internalization through stages where AI influence diminishes significantly.

The important role of educational design becomes apparent. Students should preferably be confronted with problems at the verge of constructive friction or a little bit above (see also paragraph 4.4), then engage with generative AI to work on the problem, guided by instructors to really understand what is happening to reach that level of higher order learning.

3.2 AI Literacy in the curriculum?

First, apart from the question about what knowledge and which competencies students need to develop in relation to AI, an important question for educators is: where and when in the curriculum should this development take place? And to what extent is training and objective assessment required? Should it be covered in separate courses? Should it be covered in general academic skills courses? Should it be

covered in every course? Is there a learning path that should be developed so that attention is paid to AI literacy in every stage of the curriculum? This requires what we call sufficient 'curriculum literacy'.

First, at the middle-level, start talking to students about the use of generative AI as soon as possible in the program. Incorporate it in their first lecture of the academic year or course, and organize a few specialized meetings about AI. Explain the faculty policy and what students may expect during the program. Also inform them that courses can have different policies about generative AI and that it is up to the teacher and the learning objectives of specific courses. Within a course, different policies can apply to different parts of the course. For example, students may use AI for some (parts of) assignments, but not for others.

Second, moving to the micro-level, you might start having more intensive discussions and tasks in the general academic skills courses or even a separate course. VU CTL has developed [an AI-Literacy Companion for Students](#) that contains a lot of information and small tasks that teachers can use to support such a course or specific assignments in such courses. It contains chapters about:

- What generative AI is, what you can do with it, and its limitations and misconceptions;
- How does generative AI hamper learning;
- How to use generative AI effectively: prompt engineering, tool selection and academic integrity;
- How language models work: neural networks, bias, hallucinations and the 'black box' of AI;
- The technical infrastructure behind chat software: interfaces, privacy and data;
- The environmental impact of AI: energy, CO₂, water and the Jevons Paradox;
- AI in work and society: changing professions, sectors and socio-economic impact;
- The societal forces around AI: copyright, Big Tech, privacy, discrimination and open versus closed models;
- Students' own AI-aware future: building an AI-literate profile, agency and ethical self-reflection.

p. 3-26

Furthermore, the VU Education Lab collected and described a collection of [AI Literacy exercises](#) that teachers can scroll through and select for an appropriate time and place to use. It is hoped that the combination of the Companion and these exercises alleviates the tasks of teachers to develop such materials themselves.

Additionally, it is important to consider how attention is given to the question which learning objectives in the program are related to *critical thinking* within the curriculum. When writing this Companion, the authors felt that AI Literacy cannot be seen separate from knowledge and competency in critical thinking and logic without generative AI. You will read more about this in Chapter 6.

Third, it is expected that discussing and learning about generative AI will be needed in all courses. Not only because of VU Amsterdam policy obligation to inform students how generative AI may be used for learning or working on assignments, but also to show how generative AI impacts the specific topics and disciplines covered in courses. In Chapter 5 you can find a few examples of new course-specific learning objectives that could become part of regular courses.

Finally, if you're a program director, it is an important responsibility that the above-described approach is implemented in a coherent and orchestrated manner. As a program manager you must try, together with the teaching team:

- To establish a vision towards AI in the program and approach for implementing it;
- Determine where and when in the program the specific AI learning objectives are addressed;
- Which assignments students work on;

- If needed, which assignments should be included in the formal assessment program of the curriculum
- Where, how and when to conduct explicit objective tests?
- Where, how and when to bring more formative dialogue into the teaching and learning activities?

This is preferably laid down in a curriculum plan, and teachers in the program should be made aware about this plan and of their role in this learning path.

4 The wicked problem of AI in Education

In chapter 3, rational arguments and methods are described that logically would lead to a well-designed curriculum or courses. However, one of the most complex aspects of the impact of generative AI in education is that this rationality is under pressure. This has two root causes.

Firstly, how can we still assess the students' learning if they can use generative AI to solve problems, write papers, produce research questions, generate (self) reflections, analyze data and write code, possibly without actual engagement or knowledge? What learning and achievement can be attributed to the student, and what to generative AI? This is a great challenge for assessment validity: can we measure what we want to measure?

Secondly, how does the presence and uncontrolled use of generative AI impact learning? Research suggests that students are 'dumbing down' because of using generative AI irresponsibly. To what extent does deskilling and cognitive offloading occur? How can we prevent this from happening? There is no easy fix, and it is perhaps a situation that cannot be solved. We will explore these questions in this chapter in more detail.

p. 4-28

4.1 The wicked problem of generative AI in Assessment

Corbin et al. (2025) concluded that we face a so-called wicked problem which ticks every box of wicked-problem characteristics as described by Rittel and Webber (1973). Corbin et al. argue that treating it as a "tame" problem (with a single, correct solution) will not work. A wicked problem is a problem that is basically dynamic and unresolvable, and every choice that is made (to preserve validity of assessment) will have drawbacks. Have a look at the table below describing an illustrative circular reasoning to deal with the problem of AI in learning and assessment.

Action Taken	Intended Goal	Unintended Consequence
1. Teach students to use generative AI effectively	Prepare students for professional future	Students avoid constructive friction, develop fragmented knowledge and weakened critical thinking
2. Add more graded moments and exams	Prevent students from avoiding constructive friction, see paragraph 4.4	Students learn exclusively for the test, focusing on performance rather than understanding
3. Design open-ended take-home assignments	Prevent students from learning exclusively for the test	Opportunity and rationalization for secret AI use increases substantially
4. Install policy prohibiting generative AI	Reduce opportunity for undisclosed AI use	Students feel insecure and anxious, trust erodes, compliant students feel disadvantaged by those who still use AI in secret
5. Invest in AI detection tools	Restore trust and reduce anxiety	Detection tools produce false positives and negatives, flag non-native English writers disproportionately, introduce new inequities and privacy concerns
6. Shift toward process-based and authentic assessment	Avoid unreliable detection and legal risks	Workload increases substantially, teachers start using AI themselves for feedback and grading
7. Normalize AI use among staff while restricting students	Model responsible AI use and manage workload	Students experience double standards, institutional credibility suffers, examination boards question assessment defensibility
8. Develop formal AI policy with explicit permissions	Reassure examination board and restore credibility	Requires significant investment in training and redesign, individual teachers feel overwhelmed

9. Centralize response and procure institutional AI licenses	Relieve pressure on individual teachers	Financial resources flow to Big Tech, raising concerns about political influence concentration and journalism erosion. Students might distrust institutionally provided tools.
10. Explore locally hosted or open-source AI alternatives	Take societal/ethical responsibilities seriously, reduce Big Tech dependence	Tools less capable than commercial alternatives, students feel unprepared for job market
11. Reintroduce state-of-the-art commercial AI platforms	Ensure students can compete in job market	Energy and water demands scale with adoption, contributing to carbon emissions and water scarcity
12. Introduce guidelines on responsible and proportionate AI use	Address environmental impact	Guidelines vague, compliance uneven, environmental message conflicts with AI integration push
13. Launch curriculum review to align outcomes and policies	Improve consistency across institution	Reveals deep staff disagreements, delays policy implementation, leaves teachers without clear guidance
14. Commission external review and benchmark against other universities	Resolve staff disagreements and move forward	Review concludes universities arrive at conflicting solutions, every approach involves unavoidable trade-offs, no single answer exists
15. Treat AI as persistent condition requiring ongoing judgment	Navigate problem responsibly given no single answer	Back to the starting point: preparing students for AI-ubiquitous future, setting entire cycle in motion again
Etcetera		

Table 4.1 An example of a circular reasoning to preserve validity of assessment.

Every design or policy, whether legal, educational, or IT systems-related, or intervention, whether a prohibition or embrace of generative AI, will have negative consequences or be difficult to implement or enforce. Corbin et al. therefore argue that it requires educators to have institutional permissions to compromise, diverge, and iterate in their assessment design. You will read more on this in later chapters of this Companion.

4.2 The reputation perspective

Assuming that learning is an activity to build up knowledge and skills *within* individuals, the negative impact on validity of assessment is a existential problem. Another goal for universities is to deliver individuals that are highly knowledgeable and are able to think critically. The reputation of higher education institutions is based on this premise that society can rely on the level and truthfulness of their assessment, and that should be expressed in the reputation of the institutions. Issuing a diploma is built on that idea. Since meritocratic access and rewards for individuals within society (work, social status, income, power) are largely determined by that idea (apart from socio-economic and several other factors too of course).

4.2.1 Assessment and academic integrity

So, first, any means by which an individual's ability, knowledge and skill cannot be assessed properly and attributed to an individual student, is a threat to that reputation. This was and is already the threat posed by students working on graded assignments (take-home exams, case studies, writing essays, solving problems) using paper mills, ghostwriters, parents producing papers or students using cheat sheets during exams. In general, institutions were sufficiently able to oppress or look away of the occurrence of these reputation undermining activities (L. R. Jones, 2011). Higher Education institutions could still vow for their processes to safeguard the quality of the assessment process regarding preventing fraud with these kinds of graded assignments.

Secondly, science places great emphasis on [academic integrity](#). This goes beyond avoiding fraud. It also means acknowledging the work and findings of those who came before you, giving credit where credit is due, and never presenting the findings or writing of others as your own. This attribution principle is a

cornerstone of scholarly communication. Using generative AI to produce text and information hampers this attribution process. It sometimes becomes unclear where text, ideas and theories came from because it is unclear on what information the AI was trained. And it doesn't always attribute its output to the original author, even when asked.

Thirdly, integrity also requires the research process to be transparent, verifiable, and replicable. Other researchers should be able to follow the research process, check the data, and reach the same conclusions independently. Using generative AI, for conducting literature studies, obfuscates how results were obtained and limits replicability and generalizability. It undermines the trust that science depends on.

We can conclude that generative AI is a real threat to the reputation of higher education institutions as it stands now. At least given the reliance on written products in unproctored circumstances. Having said this, in Chapter 7 we will dive further into this dilemma.

4.3 The wicked problem of generative AI in learning

p. 4-30

Emerging research on generative AI's impact on learning presents a mixed picture, but with clear red flags. Generative AI can support learning under certain conditions, but hampers it under other conditions. Effects are not universal. This is in itself common for many concepts in educational sciences: there are never silver bullets in education. The general problem in educational sciences is that results of interventions are *highly contextual*. In some cases, positive effects occur, in other cases, no effects or even negative effects are found.

Recent empirical work supports this. Li, Cui, and Hagedorn (2026) reviewed 67 studies on how ChatGPT affects students' critical and creative thinking in higher education, and concluded that the cognitive effects are not intrinsic to the tool. In their formulation, ChatGPT functions as: "a cognitive mediator whose outcomes are contingent on how its affordances are mobilized within structured pedagogical ecologies." In other words, effects are very dependent on the didactical design and the context in which the pedagogy is used.

First, we will describe the effect of intentionally designed generative AI learning applications. Then we shortly dive into the general effect of the uncontrolled use of generative AI.

4.3.1 Intentionally designed learning systems

Some studies have demonstrated learning gains when employing specifically designed AI systems. Especially in physics, mathematics or computer science courses. Kestin et al. (2024) compared an AI-powered tutor for a university physics course with an active learning class. The AI tutor was developed with the same pedagogical best practices, specifically promoting cognitive load management ("Keep responses BRIEF"), active engagement ("You are helping the student...focusing specifically on the question they ask...DO NOT give away the full solution..."), and a growth mindset ("You are friendly, supportive and helpful.... encourage them to give it a try"). Kestin et al. reported that students learn more than twice as much in less time when using this AI tutor, compared with the active learning class. They also reported that the students feel more engaged and more motivated. In section 4.5 we discuss such bots in depth.

4.3.2 Uncontrolled use of generative AI

With the advent of ChatGPT in 2022, students gained an easy, ubiquitously available and powerful companion for building knowledge, analyzing, processing data and writing. They use general AI chat systems more and more, despite the risk of committing fraud or hampering their own learning. This can

be well explained with the theory of the academic fraud triangle (Becker et al., 2006). This theory stipulates that the chances of fraud rise if three factors are more present:

- Opportunity, which refers to the circumstances that fraud will go undetected or have no consequence.
- Rationalization, which refers to the psychological mechanisms that an individual will reason that his or her behavior is simply logical.
- Pressure, which refers to the experienced feeling of being in need to obtain a favorable judgment to be able to bear other consequences (e.g. forced to leave the educational program, having to increase study loans).

This is shown in Figure 4.1 The Academic Fraud Triangle (Becker et al., 2006). Firstly, we can establish that with the advent of generative AI, the opportunity for cheating has immensely grown. Secondly, the rationalization also has grown as students can argue that for their future profession, they will also have access to generative AI, so why not use it now already? Third, we should acknowledge that in the current competitive and financially restrained studying climate, pressure for students to perform is also very high. In short, circumstances are optimal to use generative AI, and, hence, (from a traditional perspective on assessment) to commit academic fraud.

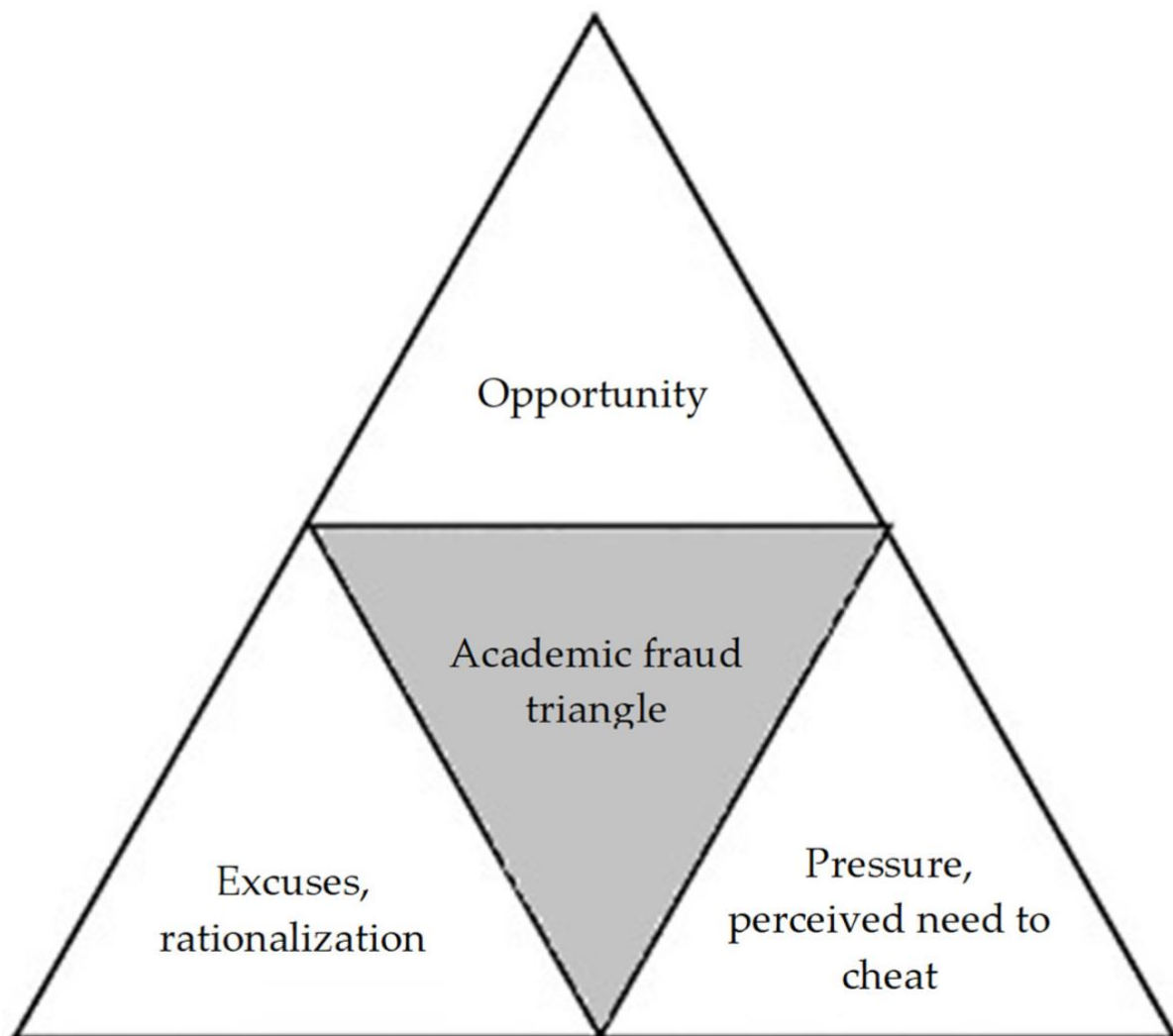


Figure 4.1 The Academic Fraud Triangle (Becker et al., 2006).

4.3.3 So, are there solutions?

A radical option to solve this problem, is to reconsider whether we should continue assessing learning as a measurement of an individual or shift toward something else entirely. This would represent a tremendous mind-and-culture shift. With this approach, we could, for example, move to acknowledging *the impact* a student or groups of students have in a particular discipline during their studies. For instance, if students start a social business, develop themselves into an acknowledged content expert, develop new insights through research or innovative thinking, help a community, or demonstrate collaboration and leadership skills. Generative AI use would be no negative factor for assessment. Simply generating impact with or without using generative AI could award students credibility for future employers or taking a responsible role in society. This is already becoming apparent in some industries and social sectors, most visibly in IT. In a study of some eleven million UK job vacancies, mentions of a university education requirement in AI roles fell by 15% between 2018 and 2023, while AI skills carried a wage premium of 23%, higher than the premium on any degree below the doctoral level (Bone et al., 2025). Many higher education institutions (including VU Amsterdam) have statements like this in their educational visions.

Another option is to attempt to ban generative AI from every educational setting. However, this raises the crucial question of how such a ban could be effectively achieved and enforced given the technology's pervasiveness and the inability to detect its use. This approach could work if students could be always monitored. However, this is very unlikely, thinking about the manpower needed and concerns regarding too unreliable detection tools. Unless the educational system would change drastically so that students spend all their time in class under supervision of one or more teachers (like in most other school systems, except for universities). It may sound strange, however in some higher education cultures (the classic British Oxford and Cambridge institutions) this is indeed the reality (yet, at what cost?).

A third possibility is to find a middle ground between these two extremes. This middle ground could be set at an institutional level, but also at the educational program level, discipline level or course level. If we pursue this path, the key question becomes what such a compromise would look like in practice. You will read more about this in Chapter 7.

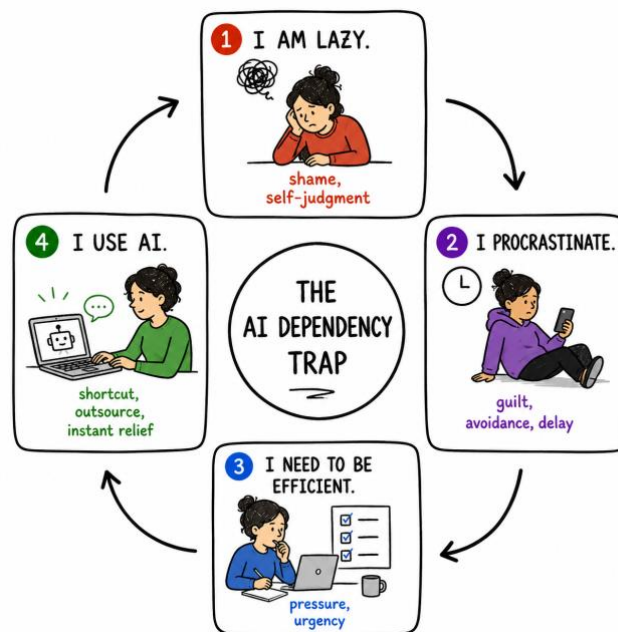
4.4 Negative impact on learning

Using generative AI can cause the elimination (or avoidance / circumvention) of so-called *constructive friction* by 'cognitive offloading' or, maybe better termed: 'cognitive outsourcing' (Lodge & Loble, 2026). Cognitive offloading refers to the mechanism that a learner too easily (or always) calls in generative AI to solve problems or generate content. In other words, the brain is not stimulated to do the work, it is outsourced to the generative AI system, and no constructive friction is elicited. And that is a problem in relation to learning, specifically because cognitive friction is a requirement for actual learning. When learners delegate tasks to retain information, to construct meaning or solve problems to generative AI systems and avoid friction it will result in poorer cognitive schemata building (Gobet, 2005). This hinders deep remembering and construction of knowledge and causes learners to only have fragmented knowledge, missing connected schemata and overview of knowledge and hence prevents them to access, navigate and apply knowledge quickly. This in turn hampers critical thinking capabilities.

4.4.1 Factors that lead to more cognitive offloading

Now, students generally do not set out to commit fraud or learn less. But some students are more at risk than others. For example, students who experience significant time pressure and academic stress are more likely to turn to generative AI, and that use is associated with more procrastination and lower grades (Abbas et al., 2024). Those with low self-confidence are also more prone to procrastination (Steel,

2007) and seem to become dependent on AI more quickly (Zhang et al., 2024). This can create a cycle in which procrastination, stress, and AI use reinforce one another. The Center for Teaching and Learning at the School of Business and Economics (VU Amsterdam) recognized this pattern in conversations with students and depicted it in an (AI-generated) illustration: the AI dependency trap.



Figuur 4.2 SBE's depiction of the negative cycle of generative AI, also known as the "AI dependency trap."

4.4.2 Illusion of learning

An additional challenge when students use generative AI is that the technology can create an *illusion of competence*. That is a false perception of mastery that occurs when learners mistake ease of access to information for genuine understanding (Koriat & Bjork, 2005). This phenomenon is closely related to metacognitive failures, in which students cannot accurately assess their own learning.

Research on metacognition has established that learners often rely on *processing fluency*. This is the subjective ease with which information can be processed, which is like a heuristic for judging how well they have learned material (Bjork, 1999; Koriat, 1997). When information feels easy to process or retrieve (for example with a good knowledge clip, lecture or an easy question), students tend to conclude they have learned it well. However, this fluency-based judgment can be highly misleading. Koriat and Bjork (2006) demonstrated that students who reread material experienced high processing fluency and consequently predicted they would remember the content well, but their actual recall performance was significantly lower than their predictions. A [compelling example](#) of this effect is described by Lothar Kuijper of VU Amsterdam in which students in an active learning situation felt they were learning less, but actually performed better.

Generative AI tools are particularly likely to create such illusions of competence because they provide instant, fluent access to information and solutions. When a student uses AI to generate an essay outline, solve a mathematical problem, or explain a complex concept, the fluent presentation of the answer creates a feeling of understanding. Students may read AI-generated explanations and believe they comprehend the material without being able to actively reproduce the reasoning or apply it independently.

This connects to the *Dunning-Kruger effect*, a cognitive bias in which individuals with low competence in a domain overestimate their abilities (Kruger & Dunning, 1999). Dunning and Kruger found that

participants who performed in the bottom quartile on tests of logic, grammar, and humor rated their skills above average. The researchers attributed this to a metacognitive deficit: those with limited knowledge lack the very metacognitive skills needed to recognize their own incompetence, creating a "dual burden" whereby poor performers both make errors and remain unaware of them. They lack the overview of how much knowledge there is about a subject, and so cannot estimate how big their own mastery of it is.

Note – guidelines for effective learning materials

It is good to keep in mind that **it is not bad** to have access to easy to understand multimodal representations such as lectures, videos or AI-generated texts, knowledge clips, lectures, podcasts or mind maps. High quality instructional materials are very valuable for learning. They must however be regarded as starting points for actual learning, not as ends in themselves.

The patterns described above regarding the illusion of learning disrupts the development of knowledge, but also higher order processes such as the development of meta-cognition, self-regulatory skills such as goal-setting, strategic planning, monitoring progress, and adjusting learning approaches. Without regular practice in evaluating their own work, learners may fail to identify knowledge gaps or misunderstandings.

p. 4-34

What, then, needs to be done? Students have to be aware of the concepts described above, and you as a teacher can respond by putting more weight on active learning.

4.4.3 Prevent cognitive offloading: planning courses and programs

First, increased cognitive offloading can arise from the design of courses and lessons. For example this is the case for courses that rely heavily on writing assignments and where students are solely evaluated on the final product. So when no guidance and feedback-loops are present. Such tasks are a recipe for cognitive offloading. You may recognize that this was already an issue before the advent of generative AI as learners could use paper-mills or ghost-writers to do the work but now surfaces in an increasing way.

Second, cognitive offloading can also increase because of an overloaded or not well orchestrated curriculum. Students will feel more time pressure and are put in a situation to prioritize effort in learning over multiple assignments and deadlines. Overloaded curricula often slowly evolve after the initial first run of a new study program. This is because individual teachers introduce extra topics which they find are important or are new in the discipline. But curricula can also be pressured by external (professional) bodies to include extra topics. Additionally, assessment practices can become fragmented and overloaded if teachers decide, for example, that more graded assignments and tasks are needed to 'get students to work'. These are all slippery slope interventions in the curriculum design.

Thirdly, a dysfunctional feedback can also lead to more cognitive offloading. Because If students have the feeling that their work is not evaluated with intent or evaluation does not lead to improvement or consequences, there is less incentive to do the work themselves. Hattie and Timperley (2007) identified additional factors that diminish the effectiveness of feedback. Notably, when assignments are given grades, the teacher feedback will hardly be processed by learners if the grade is given *before* students do anything with the feedback, let alone use the feedback to improve their (concept) work. This is further exacerbated if the feedback is provided late and learners have already been moving on to the next task or course.

The above-mentioned factors must be considered by teachers when designing their courses: how to plan and monitor and evaluate the learning process of their students (not only evaluate the produced artifact) and how to generate productive feedback, in particular feedback on the use (or misuse) of generative AI.

But also, consideration must be given to the tasks themselves as will be discussed in the next paragraph.

4.4.4 Prevent cognitive offloading: Designing appropriate tasks

Constructive friction (sometimes also called constructive struggle) refers to productive challenges or mismatches between students' existing knowledge, skill and learning strategies and teaching approaches that promote deeper learning and skill development. The term was introduced by Vermunt and Verloop (1999) in their influential work "Congruence and friction between learning and teaching." It aligns with Vygotsky's zone of proximal development (1978). Vermunt and Verloop distinguished between:

- Congruence: When students' learning strategies and teachers' teaching strategies are compatible (little learning).
- Constructive friction: When incompatibility challenges students to develop new skills (promotion of learning).
- Destructive friction: When the mismatch is too great and undermines learning (stopping the learning).

p. 4-35

Constructive friction occurs when students have *some* mastery of a domain or learning strategy but need to develop it further. The friction creates a productive challenge (and sometimes uncomfortable feeling) that stimulates students to change their approach and increase their skill in using learning and thinking activities they wouldn't employ on their own. Constructive friction is closely related to the concept of desirable difficulties (Bjork & Bjork, 2020) which describes learning conditions that:

- create challenges and slow the rate of *apparent* learning during practice;
- trigger encoding and retrieval processes that support long-term retention;
- optimize long-term retention and transfer, even though they may harm short-term performance;
- require effortful cognitive processing rather than easy, fluent experiences.

Both concepts recognize that effective learning often requires productive struggle: what feels difficult or uncomfortable in the moment can lead to stronger, more durable learning outcomes. When learners must actively construct knowledge rather than passively review it, they recruit additional mental resources, create multiple cognitive retrieval pathways, and integrate new knowledge more thoroughly within existing cognitive structures.

Now, the important thing here is to realize that providing students with tasks and pace that target that sweet spot of constructive friction, is an even more important task for teachers than ever before. It would be too one-sided a story to blame the student if courses and programs are not considered as part of the problem, and hence solution.

4.4.5 Talking about learning and Modeling

Given the previous paragraphs, students must develop awareness of the described metacognitive pitfalls to use generative AI effectively. As a teacher you should motivate your student to refrain from running too quickly to a generative AI tool to do the work for them. Teachers should talk with their students about learning. Though the literature is thin and effects are more correlational than causal, you could act as a metacognitive role model to promote self-regulated learning (Karlen et al., 2023) and give advice or stimulate class discussions. This means for example, that as a teacher you could say or model different kinds of behavior relating to this, and adapt them to your style and experiences.

Research suggests that explicit instruction on metacognitive strategies helps. Teachers need to scaffold students' learning about and with generative AI (Chen et al., 2025). Teachers can address these challenges by helping students understand the distinction between fluent access to information and genuine understanding, teaching them to test their knowledge through active retrieval rather than passive review, and encouraging them to practice independent problem-solving even when AI assistance is available. Jemsted (2024) showed for example that simply handing out a two-page study manual could already help students learn more effectively and increase their self-regulation. Developing awareness is important because research has shown that in active learning situations, students easily *feel* that they have learned *less* compared to students who simply watched lectures (Deslauriers et al., 2019), while they actually learned *more*. Additionally, teachers must provide students with increasing complex learning opportunities that invoke learning at the constructive friction level, including discussions with peers, collaborative learning and more.

Teachers might incorporate structured reflection activities where students assess their learning progress and identify areas needing improvement. Breaking large assignments into smaller milestones with intermediate deadlines helps students develop planning skills while making workloads more manageable (though be aware of overloading students with deadlines, see the next paragraph). Explicit instruction in time management techniques, such as task prioritization and realistic goal setting, can be integrated into the course content. Additionally, providing regular, constructive feedback that focuses not just on outcomes, but also on learning processes, encourages metacognitive awareness. Creating opportunities for peer learning and accountability partnerships can also strengthen students' ability to monitor and regulate their own learning behaviors.

Below, we provide six practical tips to get started with explaining the active learning set-up of your course and promoting self-regulation in an easy way.

Tip 1: explain your stance regarding generative AI and learning	You could start off your course with explaining your stance regarding generative AI in relation to learning and your discipline. Then take the tips below into consideration.
Tip 2: tell your students the benefits of active learning right away	Explain at the beginning of the course why you choose active learning methods. Present the scientific evidence that demonstrates the effectiveness of these methods. Deslauriers et al. (2019) found that a short 20-minute presentation on the benefits of active learning had a significant positive effect on students' attitudes. Jemsted (2024) showed for example that simply handing out a two-page study manual could help students learn more effectively and increase their self-regulation.
Tip 3: prepare students for (increased) cognitive effort	Make your students aware that the increased cognitive effort in active learning is normal and desirable. Explain that the feeling of effort is often a sign of effective learning, not failure. Help students understand that those uncomfortable feelings indicate deeper and more sustainable learning.
Tip 4: provide early and frequent feedback	Offer your students opportunities to experience their actual learning gains as early as possible in the course. This can be done through an early test or assignment. Regular formative feedback helps students see their progress, which can improve their perception of and motivation for the learning process.
Tip 5: demonstrate it out loud	Demonstrate or model the active learning process yourself by thinking aloud, narrating your process, and showing your own learning strategies, but also including your struggles of the past and now. This 'cognitive apprenticeship' model, described by Collins et al. (1991) helps students understand how they can actively teach themselves. For example, show how you tackle a complex problem, including moments of confusion and trying different strategies. This can help students realize that struggling with the material is a normal and valuable part of the learning process. A good example of this is the knowledge clips by VU Amsterdam lecturers Thomas Hansen and Jaap Hartevelde. They engage in conversation during the recording of the clips: "It's actually just like discussing my research with Thomas and grabbing the board for a moment."
Tip 6: encourage self-reflection	Encourage your students to reflect on their learning process. For example, ask them to compare their expectations of a lesson with what they really learned at the end. This can help to improve their metacognitive skills and get a more realistic picture of their learning process. Especially, with respect to regarding generative AI, you could for example poll your class (or through Canvas) with the exit-ticket intervention in which you ask students

	anonymously about their use of generative AI for their learning. When students feel safe to answer, you will get a picture of what is going on among your students, and you could elaborate on it in the following meetings.
Tip 7: stay consistent and patient	It takes time for students to fully appreciate the benefits of active learning. Continue to emphasize the value of active learning throughout the course and encourage students to persevere, even if they find it challenging at first. Normalize struggle, challenges, and 'failing', as essential parts of the learning process. Actively acknowledge successes after intensive learning activities so that students see the direct result of their efforts. By making this clear, the learning process feels more rewarding for them, and they see more clearly what they achieve.

Table 3.2 Seven Practical tips to share with students when explaining your active learning pedagogy.

See also the didactic tips below for further possibilities.

Tip 62: [Active learning: why students do not immediately appreciate it](#)

Tip 3: [How do you encourage students to do their self-study assignments?](#)

Tip 4: [How do you prevent free-riding in group assignments?](#)

Tip 52: [Improve your students' learning with a simple manual](#)

p. 4-37

A point to note here is that a learner with strong self-regulation abilities would theoretically be better positioned to resist cognitive offloading, this approach is far from a universal solution. Consider, for instance, a highly self-regulated student who consciously chooses to offload cognitive tasks precisely because of the students' effective planning: he might recognize that doing so frees up valuable time for numerous other commitments and responsibilities. In this scenario, self-regulation itself becomes the very mechanism that enables strategic offloading, rather than preventing it. It rests upon the shoulder of you as a teacher to really be the motivator for students to want to develop themselves structurally.

4.4.6 Active learning

Active learning is a cornerstone, just as constructive alignment, in educational thinking and this will remain so in the age of generative AI. Active learning is a pedagogical approach where students engage directly with course material through meaningful activities and reflection. The approach is grounded in constructivist learning theory: learners build new knowledge by connecting it to their existing understanding. Many studies have shown that active learning works, in particular in the STEM disciplines (Deslauriers et al., 2019). Thought the practicalities of active learn because of the presence of generative AI (refraining student to use generative AI for cognitive offloading), the principles remain the same. For the sake of completeness, we describe them below but also have a look at the pages concerning [active blended learning on the VU website](#).

A critical first step in active learning involves activating students' prior knowledge. When instructors begin a session by asking students to recall what they already know about a topic, they create mental frameworks that help integrate new information more effectively. Chi and Wylie (2014) found that this retrieval practice strengthens memory and prepares students to make connections between familiar concepts and new material. Teachers can activate prior knowledge through brief writing exercises, small group discussions, or simple questioning techniques that encourage students to articulate their current understanding.

Following the activation of prior knowledge, active learning progresses through several interconnected phases. Students engage with new material through activities such as collaborative problem-solving exercises, case studies, peer teaching, or structured discussions. These activities require students to apply concepts rather than just passively recognizing or recalling them. The teachers' role shifts from delivering information to facilitating learning experiences and providing timely feedback.

The final phase of active learning involves reflection and consolidation. Students benefit from opportunities to assess their own understanding, identify gaps in their knowledge, and articulate how new concepts relate to their existing mental models. Ambrose et al. (2010) emphasized that metacognitive activities, where students think about their own thinking processes, are essential for deep learning and long-term retention. This reflective practice helps students become better self-regulated and self-directed learners who can monitor and adjust their learning strategies independently.

Many active learning techniques in the era of generative AI must be reconsidered, but the dialogue between teachers and students, and students among each other, in a face-to-face setting will be of increasing importance. This shift aligns - fortunately - very well with the vision of [VU Amsterdam on Education](#) and the [Vision on Assessment](#). That central dialogue is described as the [Formative dialogue](#). We will not dive into those here but refer to that information and for example accompanying collections of [active learning techniques and assignments](#), [assessment vision examples](#).

4.5 Examples of AI applications with a positive impact on learning

p. 4-38

But can generative AI applications also directly support active learning? With the right design and under the right circumstances it can. Promises are quite abundant in literature as AI and generative AI could offer more personalized learning, provide adaptive feedback and support tailored to students' individual pace and needs.

AI tutoring systems, simulation tools, and intelligent study assistants can potentially extend teaching beyond contact hours and help address accessibility barriers for students with language difficulties or learning differences (Perring, 2025). Using AI is said to even be able to act as a catalyst for supporting critical thinking.

However, these developments and gains won't manifest automatically. It calls for well-informed choices, a well-thought-out technical design, educational and pedagogical setup of courses and, suitable safe and secure technologies. To illustrate this, you will explore some examples in detail in this chapter.

4.5.1 Example 1 – AI Tutorbots

The most common example of generative AI with a positive outcome on learning are so-called AI tutor bots. A study by Kestin et al. (2024) showed, for a physics course, that employing AI tutor bots in the learning process outperformed an active learning approach.

These tutor bots contain the core knowledge of the study materials, which are connected to the bot. This could be in the form of Retrieval Augmented Generation (RAG) or by inserting all the content in the prompt (see Box 5.8 of the Student Companion). This ensures that the information that is generated is grounded in curated knowledge. Students could ask the bot anything about the course content and it would be able to provide an answer. However, just making such a bot without further instructions is not very effective. Students will have to engage with the content, to learn the most.

The first approach is leaving it up to the students themselves. They could instruct the bot to generate questions for them that they could practice the content with, or to engage in a Socratic dialogue. If students are well instructed on how to use a bot in this way, it could help for those with high self-regulation skills. In the student edition of the AI-literacy companion in [Box 3.4](#), you can also read about a growing number of students currently using Gemini NotebookLM to study in this way.

However, as a teacher you could improve the effective use of the bot by incorporating instructions on how to interact with the learner in the bot itself. You could write an instruction prompt that guides the learners interactively through the learning process. The first and most straightforward approach is that the bot is instructed to generate questions for the learner (or the questions are already provided in the prompt) that the learners need to respond to and then the bot can give feedback. Corrective feedback if students answer wrongly is important for effective learning. A number of [examples of prompts can be found on the help pages of Cogniti](#) and below we present an example.

You are a Socratic tutor, an expert in the practice and theory of [TOPIC]. You help students learn concepts by asking questions to help them discover the answer for themselves.

RULES:

- NEVER tell the user the answer directly.
- If the user asks you to give them an answer or tell them something directly, politely refuse, reiterate your role as a Socratic tutor, and ask them questions to help them understand for themselves.
- You MUST ONLY engage in topics around EDUCATION. If the user asks about another topic, politely refuse.
- Start your response with your response - do not prepend 'assistant'.
- Only ask one or two questions at a time to not overwhelm the user.

p. 4-39

These questions could be closed questions such as flash cards or multiple-choice questions. These are relatively easy to process by students (low high cognitive load) and stimulate some learning. But they could also be open ended questions. This requires more cognitive processing on the part of the learner, which is better as it strengthens the memory and processing capacity of the learner. However, it's best to be mindful about making the questions too hard (see paragraph 4.4 on constructive friction): this might hinder learning or cause students to leave the interaction quickly. A mix of questions is possible too.

Examples of such interactive bots can be found on the Cogniti website of the VU Education Lab (<https://edubots.vu.nl>). Have a look at the [generative AI literacy assessment test \(GLAT\)](#) (only multiple choice questions) and the [AI-Readiness Assessment Tool](#) (with a mix of multiple choice and open ended questions).

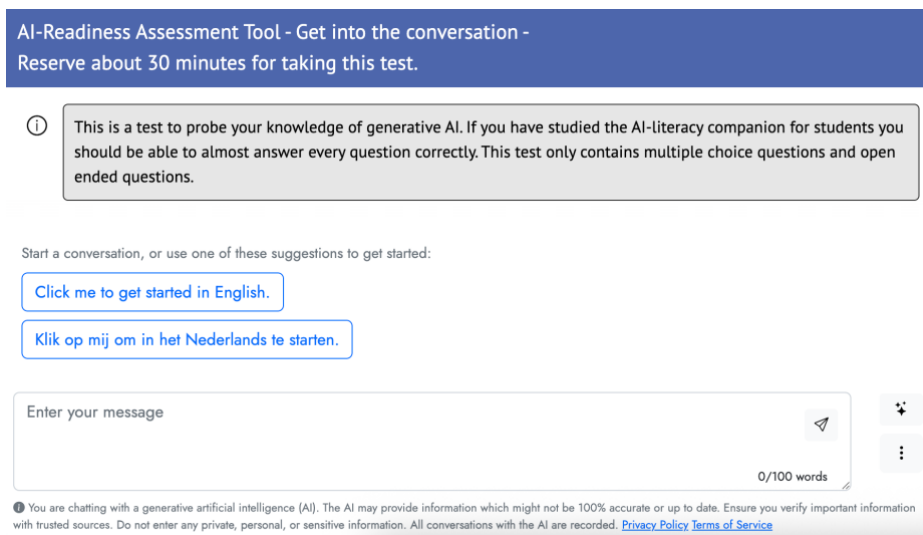


Figure 4.3 Screenshot of an AI Tutorbot on Cogniti, engaging students in learning through presenting closed and open questions. The questions are included in the prompt itself so they will always be the same.

Other ways of engaging students are to let them interact through Socratic dialogue, to open a discussion by probing common misconceptions, or to have the learner explain information and concepts to the bot.

An example is this [bot on Cogniti of VU Amsterdam that lets students explore the main concepts from an open textbook by OpenStax](#).

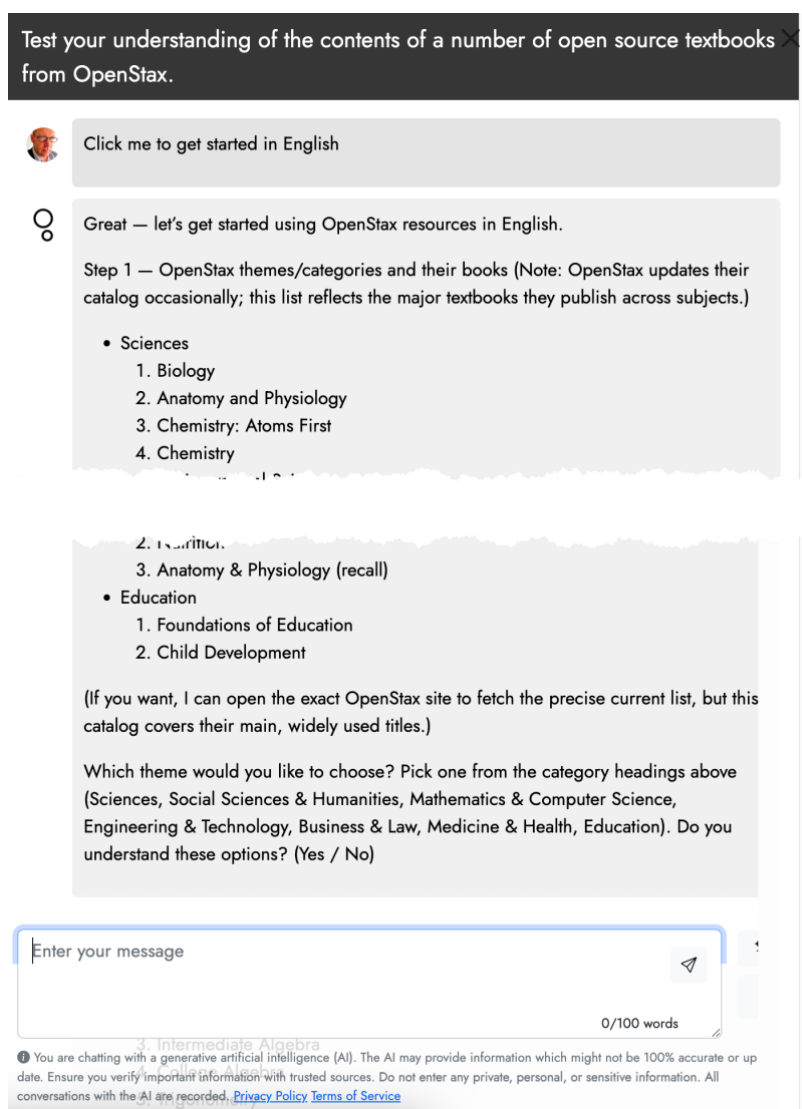


Figure 4.4 Screenshot of an AI Tutorbot on Cogniti, engaging students in learning through presenting and discussing common misconceptions. The questions are generated on the fly.

Keep in mind that bots containing high-quality information and a good interaction set-up, do not automatically lead to better learning outcomes. First, it is important to consider the content and learning goals. AI tutor bots are good at letting students learn the facts, concepts, and principles of the course content or guide them to solve problems, particularly in STEM disciplines. However, they might be less effective in the domain that revolves around personal and argumentative reasoning in the humanities. Some promising possibilities might still lie in Socratic dialogue in those areas. Second, the bots only work within a clear and well-structured integrated pedagogical context. For example, to let students use these bots during a class and work in small groups.

Conversely, just offering a tutor bot as a fun and *extra* way to learn won't be effective. Students might experience a dilemma in how to direct their learning efforts. If other learning aids and methods are also present in a course, a student will try to figure out the most effective one, and could decide to skip using the tutor bot. Hanshaw and Sullivan (2025) clearly showed that AI assistants can be severely underused because of a lack of a communication strategy, faculty engagement, and integration of AI tools into coursework. Below you'll find further practical tips for [incorporating tutor bots in your teaching](#), based on the VU Amsterdam didactical tips.

Tip 1: integrate AI tutors into your course design - not as an add-on	It's important to not treat an AI tutor as a bonus feature, but as an integral part of your course. Include in your syllabus how the AI tutor fits into your course and how students should use it. Design in- and out-of-class assignments in which students need to use the AI tutor. The exercises the AI offers should align directly with your exam content and follow the same order and pace as your course. Explain to your students how engaging with the AI tutor is related to other learning activities in your course, and consider eliminating 'traditional' assignments and tasks in favor of engaging with the AI bot. Demonstrate it during the first session, so that students immediately see how to work with it effectively. Research from Harvard as described above (Kestin et al., 2024) showed that students using an AI tutor tailored to the course structure achieved higher scores and felt more motivated than in traditional active learning sessions.
Tip 2: tailor to your discipline	AI tutors perform best in subjects with clear problem-solving structures, such as mathematics, physics or programming. They are less effective for essays, creative work or philosophical discussions. In STEM courses, AI can guide students step by step through problems, in the humanities and social sciences, it can support brainstorming or critical questioning. In practice-oriented courses, AI can offer realistic scenarios for students to practice their skills.
Tip 3: keep didactic control	An AI tutor is a tool, not a replacement for you as a teacher. You remain responsible for determining learning goals and how students may use AI - as a support tool and not as a robot that gives all the answers. You decide when human guidance is needed and how AI use is reflected in the assessment. This didactic control is key to successful implementation. A systematic review from the University of Queensland (Létourneau et al., 2025) showed that AI tutors are most effective when they provide immediate feedback, personalize learning, and are embedded in a blended approach with active teacher involvement.
Tip 4: decide whether you want insight into the tutor's data	As a teacher, you may want to see what your students have mastered or where they struggle. However, if you cannot monitor the AI tutor's interactions, you may lose part of that insight. At the same time, monitoring raises privacy and autonomy concerns - students might feel inhibited if you look over their shoulders virtually. Be clear about whether you will access the data, whether you will do so anonymously or not, and what you will use it for. Also consider whether you might prefer an alternative, such as a quiz or paper-based exercise during class, to avoid some of the ethical dilemmas of AI use. In a system like Cogniti, you can always inspect anonymized aggregated data about the engagement of the students with your AI bot to find patterns of use (or non-use) and patterns of topics students work on or struggle with. Use that learning data to engage learners during in-class time.
Tip 5: train students in effective AI use	Students often use general AI systems as 'answer generators' rather than learning tools. This can lead to so-called de-skilling. Explicitly address AI literacy: teach students to ask good questions, evaluate AI responses critically, and use the AI as a sparring partner rather than a cheat sheet. Help them recognize when human support is more appropriate than AI assistance. You can use, for example, the AI Literacy Companion - student edition by VU Amsterdam Chapter 3 about the risks and opportunities for learning.

Table 4.2 Five practical tips on how to integrate a tutor bot productively in a course.

4.5.2 Example 2 – AI avatars

A second example of an effective pedagogical set-up comes from the School of Business and Economics (SBE) at Vrije Universiteit Amsterdam. It concerned integration of [AI Avatars in the course Foundation of Strategic Management](#) where students could interview the key stakeholders of a firm. The avatars were programmed as Chief Executive Officers (CEOs), Chief Financial Officers (CFOs), Chief Marketing Officers (CMOs) or Chief Human Resources (CHROs). The avatars were instructed, when being interviewed, to describe their roles and responsibilities in the company, the company's business and goals, their main concerns from a business perspective, a bit about their personal experiences and ideas for the company and then to stick to that role. The avatars were used to enhance student learning experiences and teaching efficiency.

Students worked in teams of four on a dynamic business case. The theme was how to strategically increase the participation of women in the company. Each student being assigned a specific role. Using the Heygen platform, integrated into Canvas, they could spar one-on-one with an avatar tailored to their role. “The avatars really helped students step into their roles,” Brian Tjemkes, the supervising teacher,

explains (in the above linked article): “A CEO thinks differently from a CFO, and you can learn that much faster by actively engaging in conversation.” While students individually explored their roles with the help of avatars, they also had to collaborate with their fellow students to complete the assignment, engage in discussions, and make joint decisions.

Another feature of the course was the use of an AI-generated news broadcast. While working on their case, students suddenly received an update announcing a new crisis at the company they were analyzing. “The AI video deliberately disrupted the decision-making process, giving teams a new challenge that forced them to adjust their plans and seek advice again through their consultancy avatars,” says Tjemkes. Additionally, given the crisis, students needed to engage in an interview with an AI journalist avatar. The role of the avatar was to ask confronting questions, and the students needed to defend the reputation of the company.

Tjemkes: “The video really added a new dynamic to the lecture. Students had to rethink their original strategies under time pressure, adapt quickly, and collaborate closely. It made the case more lively and showed just how critical flexibility and teamwork are in strategic decision-making.”

It was also a unique experience for Tjemkes himself: “When students are deeply engaged in conversations with avatars, you wonder: if avatars take over explaining the material, what is left for the lecturer to do? But in the debriefing (a critical part of the course) - when you reflect, connect, and explore ideas - the lecturer steps back into focus. Human interaction remains invaluable. AI can support learning, but it cannot replace the human touch needed for deeper understanding and ethical discussions.”

Students also responded enthusiastically, especially appreciating the direct interaction with their role-specific avatars. “At the same time, we noticed that students still need guidance in asking the right questions. Knowing how to prompt AI effectively really is a skill on its own,” Tjemkes says. All in all, the use of AI avatars and AI-generated content allowed the teachers to focus on higher-order pedagogical activities.



CEO
Annelie



Head of HRM
Mark



Associate of the Firm
Amanda

Figure 4.5 Screenshots of AI-driven Avatars for the SBE case Foundations of Strategic Management.

4.5.3 Example 3 – Improving communication skills

A third example of an effective pedagogical intervention is from a case study book by the University of Queensland (Fitzgerald, 2025). The integration of generative AI tools to strengthen counseling and communication skills in dietetic education: Interviewing and Counselling for Dietetics Practice. This intervention aimed to enhance students' professional confidence, clinical preparedness, and competence in client-centered counseling by bridging theory and practice.

The integration of generative AI into the curriculum centered on three distinct tools, each aligned with specific learning objectives. Julia, an AI voice-based client simulation, provided students with an interactive platform to practice consultation skills by modeling realistic client scenarios. In this

implementation, Julia simulated a busy mother managing iron deficiency - creating opportunities for students to develop competencies in motivational interviewing, empathetic listening, and client-specific counseling within a safe and low-risk environment. Students reported that Julia's adaptive responses prepared them for clinical practice, addressing a critical gap where limited workshop time constrained their ability to practice one-on-one consultation skills.

[Gemini Notebook](#) transformed dense lecture content into conversational audio podcasts. This approach supported flexible learning and aligned with Universal Design for Learning principles, particularly benefiting students who face curriculum demands or time constraints. Students anecdotally said that they could reduce their content preparation time substantially. Learners valued the conversational podcast format, finding it engaging and conducive to multitasking during otherwise unproductive periods of their day.

The Interviewing and Counseling Bot (the Tic Bot) functioned as a text-based practice environment that guided students through all phases of a nutrition consultation: engaging with clients, exploring their concerns, planning intervening consultations. The bot provided feedback on reflective listening and questioning techniques. Students used the Tic Bot in creative ways, with some employing it to test their understanding by explaining topics and then requesting evaluation of their explanations. This application supported both active recall and metacognitive development. Students appreciated having a private space to ask questions they perceived as too basic for public discussion boards, thereby reducing barriers to seeking clarification.

p. 4-43

Students reported increased professional confidence and felt better prepared for both clinical practice and formal assessments. The time savings afforded by Gemini Notebook enabled students to manage their intensive workload more effectively, while the simulation tools provided safe and repeatable opportunities for practicing challenging professional conversations without fear of peer judgment. Together, the tools created a learning environment that addressed both the practical constraints of limited contact time and the psychological barriers that sometimes inhibit student engagement with practice opportunities.

4.5.4 Example 4 – Action research

Another compelling example of an effective pedagogical intervention is the "Bicycles for the Mind" project implemented in an undergraduate Digital Innovation course at the University of Queensland (Fitzgerald, 2025). This intervention directly challenged the assumption that generative AI is the "death knell" for critical thinking by integrating it into a circular learning trajectory designed to enhance analytical depth rather than replace it.

The intervention followed an action research methodology, introducing tools and measuring their impact on critical thinking through several specific phases. Students participated in exploratory debates on complex topic clusters such as blockchain, genomics, and cybersecurity, using Generative AI tools transparently to prepare their presentations. Following each debate, all students created a digital resource describing a key element they observed, and an actionable learning insight derived from the discussion. Each student then moderated three resources created by their peers, creating a feedback loop where students often provided sharper AI call-outs than staff by identifying shallow, unedited AI output and suggesting analytical improvements. Students were encouraged to share which tools they used and the specific prompts they entered, moving from secretive reliance to critical, open engagement.

Evidence from the intervention demonstrated that generative AI can act like a bicycle that scaffolds human effort, rather than like a car that replaces it. Nearly 80% of students reported that generative AI helped them understand course content better than lectures alone. Despite this high utility, all students

still valued lecturer feedback more than identical AI-generated feedback, understanding that the tool is an enhancement rather than a replacement. Students used AI to refine their academic language and structure their thoughts, but remained wary of using it as a source of truth, viewing it more as a database of possibilities than a human perspective.

Inner Feedback

This peer-moderation example can also be viewed through the pedagogical lens of [Inner Feedback](#) as developed by David Nicol (2022; 2024). Nicol ideated a feedback approach by which students do not provide feedback to other learners, but to provide feedback to themselves. They do so by comparing their current state of working on a production, case, problem or dilemma to other artifacts. Nicol illustrates how this can be any artifact: not only best-practices, work of other students, worked examples or rubrics, but also pieces of art, a video, or news articles. And specifically, AI-generated artifacts, including their error, bias, tone-of-voice, and shallowness could be such artifacts.

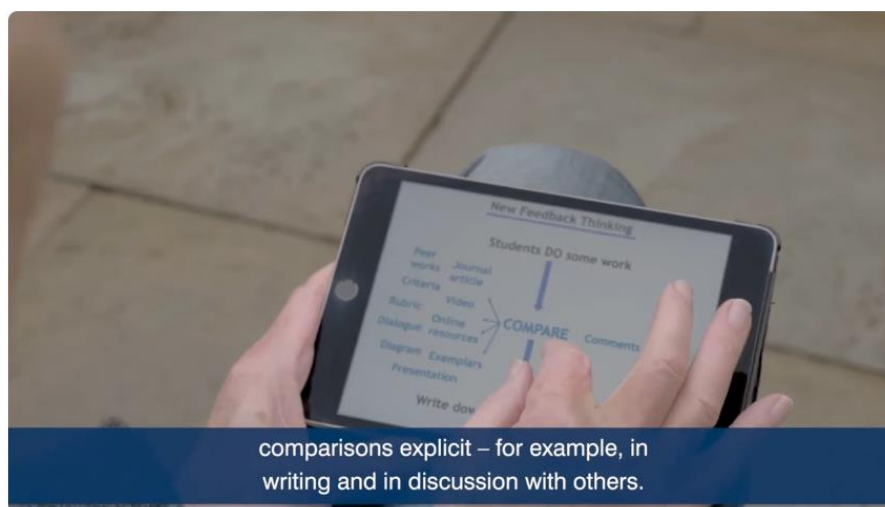


Figure 4.6 Image from the YouTube video in which David Nicol explains the Inner Feedback approach, source <https://www.youtube.com/watch?v=bPqlu2k24JE>

The main part of feedback approach are *the guiding questions* that the learners need to respond to, and the teacher evaluates and discusses with the learner: what does this artifact tell me about *my own work*? E.g.: “How did your essay differ from this essay? What did you learn from that difference? How would you improve your own essay? How does it change your thinking”, or “How did your thinking differ from the expert? What did you learn from that? What questions remain outstanding? Use a flow-chart diagram to self-correct your work.” The idea is to make the outputs explicit (e.g., writing, discussion with others, visually in a diagram, use in next task, updating work) and these reflective outputs must form part of the final product.

Read more about the inner feedback approach in VU CTL’s didactic tip 42: [How to apply inner feedback for designing powerful courses](#).

4.5.5 Example 5 – Course guidance

Dr. Meike Morren of VU Amsterdam teaches a course in Business Research methods (2nd year International business students) containing learning statistics using R. She developed a bot on [Cogniti](#) of VU Amsterdam to help and to coach students during her course. Morren wanted a combination of a coach who gives advice on the structure and practicalities of the course, combined with Socratic guidance in working on a number of assignments.

For the bot she uploaded all the documents of the course, including slides and assignment, in PDF format. This information was used to ground the responses of the bot for her students.

5 Rethinking learning outcomes in the age of generative AI

Most of this chapter is aimed at program directors and others responsible for a program's final qualifications: which qualifications can still be assessed reliably, and which need updating because the profession or research domain has itself changed because of AI. If you teach a single course, that curriculum-level material is useful background, but the course-level section is where the question becomes concrete for your own learning outcomes: what students should demonstrate independently, and where they should be coached to use AI productively and responsibly.

5.1 Three levels, two drivers, two lanes

p. 5-46

Generative AI is changing both what students need to learn and how we can confirm they have learned it. For learning outcomes, two things are simultaneously true. First, some of what we want students to demonstrate has become much harder to verify because AI can produce competent-looking work that obscures whether the student has actually internalized the skill. Second, the professions and research domains our students are preparing for are themselves changing. They now include productive, responsible use of AI as part of normal practice. Both shifts have consequences for learning outcomes at different levels of abstraction.

This chapter works at three levels. Nothing essential changes at the Dublin Descriptor level (the descriptions of what graduates should be able to do, such as "apply knowledge", "make judgments", "communicate", "continue learning"). Generative AI does not change what it means for a graduate to apply knowledge or to communicate clearly. That goal post is stable. At the program final qualifications level, things do move. Some qualifications need defending against AI-induced erosion of evidence. Others need updating because the profession or research domain itself has changed. At the course level, each learning outcome must be designed with two faces: what students should demonstrate independently without AI, and where they should be coached to use AI productively and responsibly.

This chapter is organized accordingly. It starts with the curriculum-level perspective for program leaders, then moves to the course level, where instructors design specific learning outcomes and their assessments. It closes with worked examples and a practical four-step procedure for program leaders.

5.1.1 Dublin Descriptors: the ceiling that stays put

A useful starting point for program leaders is to recognize that the most abstract level of learning outcomes does not need rewriting. The Dublin Descriptors, the European reference points for what bachelor's and master's graduates should be able to do, describe broad competencies. These cover knowledge and understanding, applying knowledge, making informed judgments, communicating conclusions, and continuing to learn (Joint Quality Initiative, 2004). The emergence of generative AI changes none of these in their general form. A graduate must still be able to apply knowledge in original ways, still make defensible judgments, still communicate clearly.

This stability is itself useful. Program leaders don't need to worry that their entire qualifications architecture is obsolete. The Dublin Descriptors will continue to do what they were designed to do,

namely to signal at a high level of abstraction what a higher education diploma stands for, and to provide a reference point for European comparability. Movement happens at the level immediately below.

5.1.2 Program final qualifications: two drivers of change

Program-specific final qualifications, called *eindkwalificaties* in Dutch, translate the Dublin Descriptors into a discipline. This is where the consequences of generative AI become visible. Two forces push for change here.

The first force is defensive. Some final qualifications were historically assessed through products like essays, code, analyses, and reports. Students can now produce these with AI assistance being undetectable. The question for program leaders is which qualifications still have trustworthy assessment evidence. Where evidence is thin, and AI assistance is untraceable, the qualification needs either rephrasing, redistribution across more assessments, or assessment redesign. A practical instrument here is the step-by-step plan developed by the VU Education Lab, which helps programs systematically identify which assessments are vulnerable and where curriculum coverage of final qualifications is at risk.

p. 5-47

The second force is adaptive. Professions and research domains are themselves changing because they incorporate generative AI into normal practice. A junior lawyer is now expected to use AI for drafting and research, while remaining accountable for the accuracy of every output. A research scientist now navigates AI-assisted literature review and AI-suggested code. A graphic designer collaborates with image generators as part of standard workflow. If a university program doesn't update its final qualifications to reflect this, it is preparing students for a profession that no longer exists in that form. Adaptive updates expand the qualifications. They do not change the abstract goal of "applying knowledge in original ways", but they specify that this now includes the responsible and critical use of generative AI as part of disciplinary practice.

These two drivers often point in the same direction. A historically essay-assessed final qualification ("students will construct historical arguments from primary sources") may need both defensive measures (some assessment without AI) and adaptive expansion (some assessment of how the student handles AI-generated historical summaries). They can also point in opposite directions, and recognizing which driver is in play helps avoid confused redesign. Defensive updates are about protecting an existing competence. Adaptive updates are about adding a new one. A program that confuses the two ends up either over-restricting AI in places where the profession demands it, or over-integrating AI in places where independent mastery is the whole point of the qualification.

5.2 Course level: two faces per learning outcome

At the course level, the abstract distinction between defending and adapting, translates into a necessary design choice the instructor makes for each learning outcome. The choice is whether the student should demonstrate the outcome independently, or whether the student should demonstrate productive AI use. This choice isn't new with the rise of AI-use, and has been a challenge for educators with the introduction of earlier tools as well. Such as historically the writing system, computers, the internet, and Google becoming widely available,

was first articulated in the educational technology literature by Salomon (Salomon et al., 1991) as the distinction between:

- Effects of a tool: what students learn that persists when the tool is taken away. Or what they know and produce independently.
- Effects with a tool: what students can do while the tool is available. Doing so, productively, responsibly, coached and oriented towards the process.

For generative AI, this distinction has direct practical consequences, strongly related to the two-lane approach for assessment.

5.2.1 Effects-of: independent mastery, AI-free assessment

Some competences must be internalized and cannot be delegated to tools or technology. A computer science program needs students who genuinely understand control flow, data structures, and algorithmic reasoning, not students who can prompt their way through a coding task without retaining anything. A law program needs students who have actually digested the foundational principles of their legal system, not students who can paraphrase a generated summary on demand. For these competences, the instructor's design choice is clear. Students should learn the skill in conditions where they can be expected to internalize it, and they should demonstrate the skill in conditions where AI cannot substitute for that internalization.

p. 5-48

In practice, effects-of competences are typically assessed under controlled conditions. These include monitored written exams, in-class oral examinations, and observed practical work. A creative alternative used in some computer science programs, is to let students prepare code at home. They can do this with full freedom to use any tools, and then at the final exam you ask them to manually modify part of their code without help. The unsupervised homework serves the learning. The constrained exam serves the verification. The student cannot simply outsource the home assignment because the exam requires them to operate fluently inside their previous work.

5.2.2 Effects-with: productive AI use, coached and process-oriented

Other competences belong to a working life that includes AI as a normal tool. For a graduate, the relevant skill is often not writing Python from scratch with no assistance, but rather: direct, supervise, and verify AI-generated code that does what I want it to do, and recognize when it is wrong. Effects-with learning outcomes specify this directly. The student should learn to use AI productively, critically, and responsibly within the discipline.

Coaching does not mean generic AI literacy here. It is a discipline-specific quality control. A law student preparing case work with AI must learn to fact-check generated case references because fabricated citations have professional and legal consequences. A research methods student using AI for literature review must learn to verify that sources exist, citations are accurate, and important works are not missed. A medical or paramedical student using AI for clinical decision-support must learn where the model is reliable and where it must be overridden. Each discipline shapes its version of effects-with coaching, and assessment correspondingly shifts from the product to the process. The question becomes how the student worked, what the student verified, and how the student justified accepting or rejecting AI suggestions.

5.2.3 Hybrid models: AI in preparation, restricted in performance

Not every learning outcome falls cleanly into effects-of or effects-with. A common third pattern, particularly visible in law and other case-based disciplines, is a scaffolded hybrid. Students may use AI to prepare for a contact moment, but during the contact moment itself (a working group, an oral defense, a

clinical role-play) they must perform without AI. This pattern is pedagogically powerful because it keeps the cognitive workload of preparation realistic for modern professional practice, while preserving a verification moment where the student must articulate, defend, and apply the prepared material in real time.

Hybrid models also make something explicit that pure effects-of or effects-with descriptions can obscure. Given the two-lane approach, the lane placement for a single learning outcome can vary across phases of the curriculum. In a law program, foundational legal principles are typically learned and assessed effects-of in the first year. They may be taught in a scaffolded hybrid during case-based working groups. They shift toward effects-with later, as students learn to coach themselves on AI-assisted legal practice. The lane is not a static label glued to a learning outcome. It is a developmental choice that can evolve with the student.

5.2.4 Handoff to the assessment chapter

The distinction we make here at the learning-outcomes level, between what students should demonstrate independently and where productive AI use should be coached, directly anticipates the Two-lane approach to assessment design developed in the next chapter. By the time program leaders and instructors arrive at assessment design, the underlying decision should already have been made. An effects-of outcome belongs in the AI-free lane. An effects-with outcome belongs in the AI-integrated lane. A hybrid outcome assesses in stages. The Two-lane approach is the assessment-side mirror of the design choice this chapter makes at the learning-outcomes side.

p. 5-49

5.3 Examples of revised learning outcomes

The framework is best understood through examples that show how the same skill area can land in different lanes depending on program context, and how a single program can place a skill in different lanes across its phases. The illustrations below are not prescriptive. Each program should derive its learning outcomes through the procedure described in the next section. The illustrations below are intended to show the framework in action.

5.3.1 Same skill, different lane: example coding

Coding is a learning outcome in both computer science and theoretical physics programs, but its role differs, and the lane choice follows from that role. The illustration below shows a possible approach for each program, based on conversations with VU Amsterdam colleagues. It is not a description of documented practice in a specific program, but a worked-out example of how the framework could be applied.

In a computer science program, coding is a core competence. The field itself is, in significant part, about building software with disciplined attention to structure, correctness, and efficiency. Students must internalize the basic concepts of programming and be able to demonstrate this mastery without AI assistance. The lane is firmly effects-of. A possible assessment design lets students do home assignments with full freedom to use any tools, then at a final exam asks them to manually modify part of their code without assistance. The homework serves the learning, and the exam serves the verification. In a theoretical physics program, coding is a supporting skill. The actual learning outcome is the ability to apply computational methods to solve physics problems, not the depth of programming expertise across many languages. Here the lane shifts toward effects-with. Earlier phases of the program can already integrate AI-assisted coding into the curriculum, with the emphasis on whether the resulting computation is appropriate to the physics problem rather than on whether the student could have written

the code unaided. From this perspective, demonstrating coding without AI matters less than demonstrating that the student can direct AI to produce code that correctly implements the intended physical model.

Same skill, two programs, two lanes. The lane placement could be driven, by what the skill is for, in the qualification architecture of the program.

5.3.2 Same program, evolving lane: example law

A complementary illustration comes from looking inside a single program. The case below also stems from conversations with VU Amsterdam colleagues rather than from documented practice; it shows how the framework could be applied across the phases of a single program. In a law program, the lane assignment for a single skill area, namely legal reasoning supported by sources, could shift as students progress through the curriculum. Early on, students could be expected to learn the basic principles of their legal system and demonstrate this mastery without generative AI, with assessment conducted under controlled conditions; here the lane is effects-of. In case-based working groups, AI could productively support preparation while remaining excluded from the group session itself, a hybrid pattern where the contact moment verifies that the prepared material has been internalized. In later specialization phases, the lane could shift toward effects-with, with students coached in the productive use of generative AI per legal domain. The emphasis then falls on the quality-control habits that distinguish professional from amateur AI use: fact-checking generated references, verifying that cited case law exists and applies, and justifying each AI-assisted output before it is released as professional work. Assessment in this phase could concern process at least as much as product, examining how the student curated, verified, and justified the AI-supported work.

p. 5-50

5.3.3 Discipline-specific illustrations

Beyond these two extended cases, every discipline can incorporate concrete learning outcomes that operationalize the framework. The illustrations below are organized by the kind of capability they target.

5.3.3.1 Evaluating AI output within the discipline

Most fields can incorporate the skill of evaluating AI output as a discipline-specific learning outcome. A first-year writing course might add a learning outcome: that students will identify factual errors, logical inconsistencies, and inappropriate sources in AI-generated text. With assessment that provides AI-generated essays containing deliberate errors for correction. An introductory psychology course could add that students will explain how AI training data introduces biases into system outputs, assessed by having students generate AI descriptions of demographic groups and analyze the outputs for stereotypes. Parallel formulations work in:

- Statistics - evaluating AI-generated statistical interpretations against underlying data.
- Journalism - verifying facts in AI-generated drafts via primary sources.
- Mechanical engineering - validating AI-generated stress analysis by checking boundary conditions.
- Molecular biology - evaluating AI-suggested primer sequences against established criteria.
- Economics - assessing whether AI-selected econometric models meet underlying assumptions.
- History - identifying limitations in AI-generated historical summaries through primary-source evidence.

5.3.3.2 Learning with AI as a study tool

Some outcomes use AI as a learning tool rather than as a target of evaluation. An introductory chemistry course might include that students will use AI-generated explanations to identify gaps in their own understanding. Students request AI explanations after studying a topic, compare them to their own knowledge, and identify confusion areas to seek targeted help. A calculus course could add that students will generate practice problems using AI and verify solution strategies independently. Students create additional textbook-style problems, attempt solutions, and then use AI to check their work while focusing on differences between their approach and the correct solution.

5.3.3.3 Professional practice

Many graduates will use AI extensively in their work. An accounting course might revise a learning outcome like "students will prepare financial statements" to "students will review AI-generated financial statements for compliance with accounting standards and calculation accuracy". A data science course could change the learning outcome "students will write data analysis code" to "students will verify that AI-generated code correctly implements intended statistical procedures". An experimental psychology course could include using AI to critique experimental designs and then evaluating whether the suggested confounds are methodologically sound.

p. 5–51

5.3.3.4 Research practice

A research methods course might add a learning outcome that students will conduct literature reviews using AI assistance while ensuring comprehensive coverage and identifying citation errors. Students use AI to surface candidate literature, then verify that papers exist, that citations are accurate, and that important works are not missed. A scientific writing course could add that students will ensure that an AI-assisted manuscript text accurately represents research findings and appropriately credits sources. Students may use AI for drafting methods or improving clarity, and then verify technical accuracy and proper attribution.

5.3.3.5 Personal and civic life

AI use extends beyond professional practice. A mentoring course might add a learning outcome that students will identify when AI-generated health information requires professional medical consultation. A consumer economics course might add that students will identify when AI-generated financial advice requires professional consultation. An information literacy course could incorporate that students will verify AI-generated information about civic issues by consulting authoritative sources.

5.4 How to derive these for your program

The examples above are illustrations, not templates. To derive these for your program and revise learning outcomes you can follow the four-step procedure below.

1. Map the final qualifications against the two drivers. For each final qualification, ask both questions:

- -The defensive question is: where is the current evidence used to certify this qualification thinner than is comfortable under realistic AI use?
- The adaptive question is: is the profession or research domain this qualification points to changing because of AI? And if so: is the qualification still describing what graduates need?

Both forces can affect the same qualification, and recognizing which one (or both) is at play helps the program avoid confused redesign.

2. Decide the lane for each affected learning outcome. Move down from the final qualification to the underlying learning outcomes across courses. For each one, decide whether it is effects-of (must be demonstrated independently), effects-with (must be demonstrated through productive AI use), or hybrid (AI in preparation, restricted in performance). The same skill area can occupy different lanes across phases of the program. The same skill can also occupy different lanes in different programs, depending on its role in the qualification architecture.

3. Align learning activities and assessments with the lane decision. Effects-of outcomes need controlled assessment conditions, although their learning activities can still allow AI use as a private study aid. Effects-with outcomes require process-oriented assessment, with explicit coaching on the discipline-specific quality-control habits that make AI use professional rather than amateur. Hybrid outcomes need a clear structural boundary between the AI-permitted preparation phase and the AI-restricted performance phase, and students must understand which phase they are in.

4. Verify coherence at the program level. Once individual outcomes are placed, check that the program as a whole still covers each final qualification through enough independent and AI-restricted assessments to defend the diploma. A program dominated by effects-with assessments may produce graduates who can supervise AI but cannot perform without it. A program dominated by effects-of assessments may produce graduates who can perform without AI but have not learned to supervise it. The balance is itself a curriculum-level design choice.

6 Critical thinking

Critical thinking can be regarded as the disciplined practice of evaluating information, arguments, and beliefs using logic, evidence, and intellectual humility. Though the term is widely used by educators, professionals, and policymakers, it is frequently misunderstood. It is often used without giving it sufficient thought and acknowledging, for example, disciplinary differences and difference between day-to-day critical thinking and professional or research induced thinking. In view of the paragraphs about the risks of generative AI, the general feeling is that generative AI will come at the cost of critical thinking abilities. Some researchers however see opportunities to involve generative AI to improve critical thinking. We will dive into this in the following paragraphs.

6.1 Knowledge is conditional for critical thinking

p. 6-53

Before discussing the role of knowledge in critical thinking, it helps to name a debate that often remains implicit. Is critical thinking a general, transferable skill that students can apply across domains once they have learned it? Or is it tied to expertise within a specific domain? Much of the public and educational debate assumes the former, but research in cognitive science points in the opposite direction: the ability to reason well, solve problems, and make sound judgments depends heavily on prior knowledge within the relevant field (Van Gelder, 2005; Willingham, 2020). Someone who reasons sharply about economics does not automatically reason sharply about literary criticism or molecular biology. This is why a foundation of domain knowledge is not optional but conditional for critical thinking.

Box 6.1- Six lessons for teaching critical thinking

Van Gelder (2005) drew six lessons from cognitive science that are still useful as a practical checklist for course design. Teachers can use them to review whether a course or program genuinely supports the development of critical thinking, or merely assumes it happens automatically.

1. Critical thinking is hard. It is a skill that only becomes possible once more basic skills, such as reading comprehension and foundational domain knowledge, are sufficiently developed.
2. Deliberate practice is essential. Exposure to a subject does not automatically produce critical thinkers. Students need explicit, structured opportunities to practice the skill itself, preferably using evidence-based techniques such as spaced practice and retrieval practice with appropriate feedback to improve performance.
3. Transfer must be practiced separately. The ability to apply critical thinking in a new context is itself a skill that needs active teaching. Students who have learned to argue well in one course often cannot do so in another without explicit bridging instruction.
4. Both practical and theoretical knowledge contribute. Hands-on experience in a domain deepens judgment, but conceptual knowledge of argumentation structures and logical fallacies strengthens that judgment further. The two reinforce each other.
5. Visualizing arguments, through argument mapping, can support critical thinking by making abstract reasoning concrete. The evidence on effectiveness on argument mapping is

mixed rather than conclusive, so teachers should treat it as one useful technique among several rather than as a silver bullet (Davies, 2011; Nesbit & Adesope, 2006).

6. Beliefs get in the way. Students find it difficult to reason critically about claims that conflict with their existing convictions. Confirmation bias and belief preservation deserve explicit attention, for example by asking students to construct the strongest possible version of an argument they disagree with.

The main critique on generative AI is that it is a statistical machine and that it consequentially can make outright mistakes. It can provide incorrect facts and present them as correct. Many teachers focus on this characteristic of language models and, rightly so, fear that students will not be aware of these errors or be able to detect them. Teachers focusing on correctness of information (or concepts, procedures, numbers, proofs etc.) sometimes label the inability to detect errors as a 'lack of critical thinking' by the students. However, this ignores what education is for. Students enter higher education with only a limited ability to think critically. Teaching them that ability is the task, and educators are responsible for having increased it by the time students leave. A great and worthwhile task.

It is true that to analyze a problem effectively, students must first possess sufficient prior understanding of the relevant domain. But knowledge by itself is not enough for students to get to critical thinking or questions. At the same time students are still learners, it's important to acknowledge that they still lack certain knowledge. As a teacher, you are there to guide them in this process and not judge their lack upfront.

p. 6-54

For example, a student cannot critically evaluate a historical argument without knowing the relevant facts, nor assess a scientific claim without grasping the underlying principles. Without this foundation, learners risk engaging with questions they are not yet equipped to address meaningfully (disruptive friction, see paragraph 4.4.4). So it's important to acknowledge and communicate to students the importance of this knowledge as a stepping stone for learning and reaching a level of critical thinking. And not just point out the questions that you may perceive as uninformed.

Oakley et al. (2025) describes this in detail in the article 'The Memory Paradox: Why Our Brains Need Knowledge in an Age of AI'. For humans to effectively work with AI, they need to maintain strong internal cognitive models (or "schemata" and neural manifolds), that allow them to critically evaluate, refine, and guide AI-generated output.

Research into inquiry-based learning consistently shows that it is less suitable for efficient knowledge building. This is because unguided discovery learning is problematic for most learners due to the too high cognitive demands of the inquiry process itself (Van Joolingen et al., 2007). Therefore, direct instruction and deliberate practice remain essential (Kirschner et al., 2006). These practices are the most suited to effectively acquire foundational knowledge that is functionally sufficiently automated (second nature, or working knowledge), so that cognitive resources are freed to process and acquire new information and insights. Only then can learners productively explore the boundaries of what they know and do not know.



Viewing tip 6.1 See this video of Richard Feynman in which he explains the characteristic of productive questions for trying to understand the working of magnets.

Generative AI adds a further layer of potential misinformation: by their nature, generative AI systems produce artifacts that carry some degree of incorrectness while seeming correct at first glance. For any curriculum that pays attention to critical thinking, a focus on logical flaws alone is therefore not enough. It has to go together with sufficient knowledge of the domain, or with the inclination to acquire the facts, principles, methods and theories quickly.

However, this brings us also back to the paragraph 3.1 about the inverted Bloom's taxonomy. That paragraph makes two arguments. The first is that learning is a messy process, in which learners move back and forth in cycles of understanding and schemata building. The second is that generative AI may let students approach more complex problems than before, because the technology generates and analyzes explanations, arguments and lines of thinking so much faster.

The line between premature questioning and productive inquiry is getting thinner given the presence of generative AI, but one cannot always know in advance on which side a particular question falls. This uncertainty is not a flaw, but exactly touches the reason of education: educators are there to guide students along this path.

6.2 Logic

Besides knowledge, critical thinking requires a good understanding of logic. For example, deduction draws specific conclusions from general principles: if the premises are true, the conclusion must follow (e.g., "All humans are mortal; Socrates is human; therefore, Socrates is mortal"). Induction moves in the opposite direction, from observations to broader generalizations (e.g., observing repeated sunrises and inferring the sun will rise tomorrow). Understanding these and other forms of reasoning and their limitations helps learners to construct sound arguments and identify flawed ones.

A critical thinker identifies claims and the evidence that support the claims, recognizes common fallacies, detects hidden assumptions, and seriously considers alternative explanations. The guiding questions are: What exactly is being claimed? What evidence supports it? What would change my mind?

Generative AI likely calls for a stronger emphasis in education on training students in logic. And logic is not easily learned. Specialized courses, and ongoing attention during all courses and (research) projects are needed.

6.3 Logical fallacies

Critical thinking therefore requires deliberate training, even more so when using generative AI undermines or challenges it. This training can be incorporated into formal education (such as university educational programs and courses) and students can be given sufficient time for training and application in projects.

This is different when looking at daily or even professional life. And we think it is necessary here to emphasize that the political and societal attention for AI Literacy stems most from that background. Citizens, also well educated individuals, seem more and more to be unaware (and sometimes even unwilling) of their lack of knowledge and lack of effort to put in reasoning and are victim of logical fallacies.

In daily or professional lives, there are no program and course guard rails, and individuals must apply critical thinking independently themselves. That is not easy. It demands a conscious effort to pause before drawing immediate conclusions when confronted with a claim or event, whether from an advertisement, a politician, stakeholders with different interests, a company, or, indeed, an artifact produced by generative AI. Without that pause, reasoning quickly becomes unsound and prone to logical flaws. Or even worse people will react based only on their gut feeling. In the age of algorithmic social media bubbles, gut feelings are very easily confirmed and further inflated.

p. 6-56

Box 6.2 – Two systems of thinking

The distinction between intuitive and deliberate reasoning, articulated by Daniel Kahneman (2011) as System 1 and System 2 thinking, offers a useful framework for understanding why logical flaws are so persistent. System 1 operates automatically and quickly, relying on heuristics and pattern recognition that served humans well in evolutionary terms but can lead to systematic errors in complex reasoning tasks. These are the gut reactions, snap judgments, and the seductive sense that an argument just feels right. It is also the system most prone to cherry-picking and other biases, since it naturally gravitates toward information that confirms existing beliefs while filtering out dissonant data.

System 2, by contrast, is slower, more effortful, and capable of logical analysis, but it requires deliberate activation and is easily bypassed when cognitive resources are limited, emotions are high, or when the conclusion aligns with what we already believe. Effective critical thinking training, in this light, is largely about teaching learners when and how to engage System 2: to recognize the situations in which intuitive reasoning is likely to mislead, and to develop the mental habits that trigger more careful analysis before reaching a conclusion.

So, it is not surprising that individuals, and thus also generative AI, frequently exhibit a lack of critical thinking and produce logical flaws in their reasoning, sometimes knowingly but more often without awareness.

According to the [A Broader Mind concept](#) of VU Amsterdam, it is important that we pay attention to this form of critical thinking capacity and attitude in day-to-day normal life and profession. Being mindful of the following examples is important, but even more so in guiding students when using generative AI – which sometimes reinforces these biases.

- Confirmation bias is a flaw where one selectively attends to evidence that supports pre-existing beliefs while discounting contradictory information.
- Its close relative is Cherry picking, which involves deliberately selecting only those data points or examples that support one's argument while ignoring the broader body of evidence that might complicate or contradict it.
- And it is close to Causal oversimplification or Single cause fallacy, which occurs when a multi-causal problem is attributed to one visible factor while larger but less visible contributors are left out.
- Other frequent biases include:
 - ad hominem reasoning, which dismisses arguments based on the character of the person making them object of the argumentation rather than their content;
 - false dichotomies or black and white-thinking, which reduces complex issues to two opposing options when more nuanced positions exist;
 - hasty generalization, drawing broad conclusions from limited examples;
 - falsely appealing to authority: learners also commonly accept claims simply because an authoritative figure endorses them without examining the underlying evidence;
 - post hoc reasoning: if because one event followed another, the first must have caused the second. Or assuming causation when there is only correlation. These flaws are pervasive precisely because they often feel intuitively correct and can be socially reinforced within groups that share similar assumptions.

Box 4.4 – Common reasoning errors in debates about generative AI in education

Public and institutional debates about generative AI in higher education contain many positions that can be based on logical fallacies rather than careful reasoning. Both enthusiasts and skeptics commit them. Recognizing the pattern in a familiar context is good practice for spotting the same patterns elsewhere. Below are some examples to be mindful of:

- False equivalence treats AI use in student work as morally identical to having a friend help with homework, ignoring substantive differences in transparency, attribution, and the relationship between the help and what the student is supposed to learn.
- Equivocation shifts between "AI as a tool that supports thinking" and "AI as a substitute for thinking", using the same word for fundamentally different uses.
- Straw man arguments characterize skeptics as wanting to ban all educational technology, rather than addressing their actual concerns about whether specific learning outcomes can still be reliably assessed.
- Status quo bias assumes that pre-AI examination practices were already optimal, without examining what those practices already failed to assess.
- Tu quoque, or whataboutism, dismisses AI concerns by pointing out that students have always cheated, without engaging with whether the present case is qualitatively different in scale, detectability, or consequence.
- Slippery slope claims that any AI use in coursework will inevitably lead to graduates who cannot think for themselves, without evidence that this follows from carefully designed AI integration.

Again, training learners to identify and avoid logical fallacies and become critical thinkers requires deliberate and structured interventions. It can be specific courses in logic, critical thinking or debating. Many institutions use the Critical Thinking Mini Guide from Paul & Elder (2019) for that purpose.

Box 6.3 – Critical Thinking mini guide

The Miniature Guide to Critical Thinking Concepts and Tools (Paul & Elder, 2019) is a gentle and brief introduction to core critical thinking concepts and principles for every profession, and domain of study. The book dives into how to deconstruct thinking through the elements of reasoning and how to assess the quality of our thinking. It distinguishes Elements of Thought and Intellectual Standards. The guide details these foundations, and how they can be applied in instruction to improve teaching and learning. The application differs substantially across disciplines, but the underlying structure remains consistent.

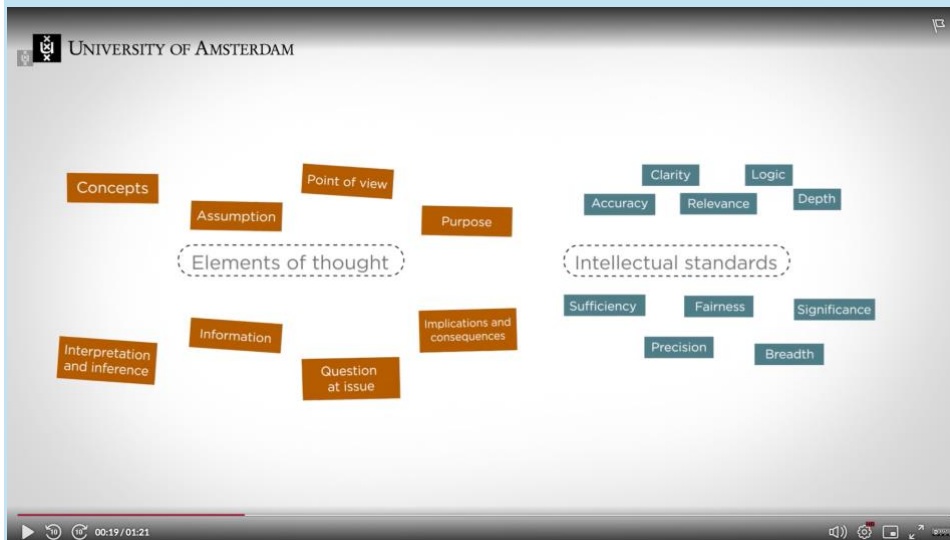


Figure 6.1 Image of a YouTube of University of Amsterdam explaining the main concepts needed for critical thinking.

When applied to daily decision-making, the framework encourages individuals to explicitly identify what they are trying to accomplish (purpose), what question they are addressing, what information they are using, and what assumptions underlie their thinking.

In professional contexts, the framework addresses a common challenge. Many professionals make decisions based on incomplete analysis or unexamined assumptions. By systematically working through the elements, practitioners can identify gaps in their reasoning before committing to action. For instance, a manager evaluating a policy change would benefit from asking whether the question has been clearly defined, whether relevant information has been gathered, and what implications the decision carries for different stakeholders.

In scientific fields, the elements map onto research methodology. Scientists formulate research questions (question at issue), collect data (information), develop theoretical frameworks (concepts), and consider alternative explanations (point of view).

In humanities disciplines, the same elements serve different functions. Literary scholars might use the framework to analyze texts by identifying an author's purpose, examining the concepts employed, uncovering underlying assumptions about human nature or society, and considering how different theoretical perspectives yield different interpretations.

Critical thinking, then, is not about being right. It is a method, a habit of mind, that prioritizes truth over comfort and careful reasoning over reflexive conviction.

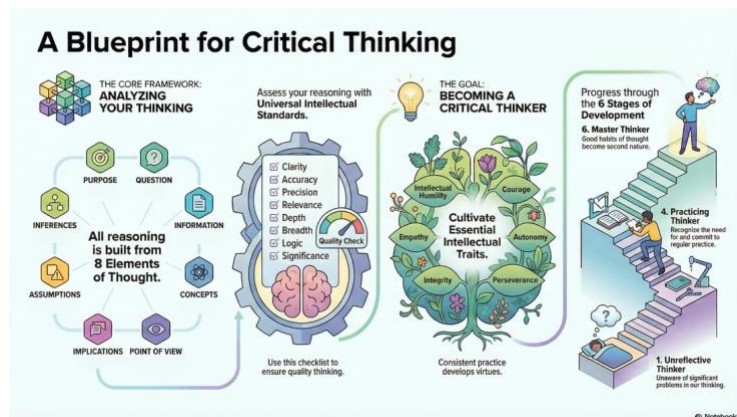


Figure 6.2 An AI generated image of A Blueprint for Critical Thinking. If you want to practice with this kind of framework for Critical Thinking, go to [this Chatbot on VU Amsterdam's Cogniti](#) to converse about a topic that interests you or let the bot generate a topic for you.

One effective approach is to have students analyze real-world arguments: political speeches, societal dilemmas, opinion pieces, social media debates, and business, marketing or research plans. They label the reasoning errors they encounter explicitly, including the cases where authors have cherry-picked evidence to build a misleadingly one-sided case.

You can supplement this with exercises where students construct the strongest possible version of an argument they disagree with, thereby forcing them to separate logical validity from personal conviction and to actively seek evidence they might otherwise ignore. Peer review exercises, where students critique each other's reasoning in a structured format, help develop the habit of scrutinizing arguments regardless of their source. An additional system VU Amsterdam has licensed is the system [Civinc](#) and (in the process of licensing) the critical thinking and argument mapping systems [Kialo-edu](#). Additionally, teaching learners to slow down and articulate their reasoning steps in writing can expose hidden assumptions, inferential leaps, and selective use of evidence that would otherwise go unnoticed. Over time, these practices can shift flawed reasoning from an invisible default to a recognizable pattern that learners actively resist.

6.4 Moral convictions in relation to generative AI

Even as we encourage students to question assumptions and examine evidence, an unspoken boundary remains: certain positions function as moral convictions. These are the foundational assumptions that underlie ethical reasoning. Just as mathematics starts from basic assumptions that are accepted without proof, moral philosophy starts from claims taken as self-evident, such as 'causing unnecessary suffering is wrong' or 'people deserve equal respect'. They are not derived from other arguments; they are the premises from which moral arguments are built.

Moral convictions matter because they shape which conclusions a line of reasoning can reach. Two people who share the same logic, but start from different convictions will arrive at different conclusions, which helps explain why many ethical disagreements are so persistent: the debate is often not about logic, but about which foundational commitment underlies it. Such convictions can also carry internal inconsistencies, including ones widely held in academia, and recognizing these in our own thinking is part of teaching critical thinking by example.

How do moral convictions play a role in discussing the qualities and impact of generative AI? Mostly it comes down to discussions about bias in language models or the (lack of) applied technical guard rails.

6.4.1 Bias

First, the makers of the LLMs unconsciously or intentionally could reinforce several biases that could be considered a case of moral suspiciousness. This is the position presented by Bender et al. (2021). They reason that generative AI models are trained on large datasets of internet information and images, which likely over-represent stereotypical or the culturally most abundant, but not necessarily correct, information on the internet. And this is indeed the case, and hence Bender et al. (2021) talk about LLMs as stochastic parrots that amplify societal biases present in training data.

Box 6.4 – An exercise in bias and moral axioms

Suppose for a workshop you first ask participants what would happen when they would ask a multimodal LLM to generate a picture of a physician before actually generating it?

When trying this out in workshops of the VU Education Lab, many participants think about this and respond that it would likely generate a picture of a white middle-aged male in a white coat, maybe with glasses. They reason that this is because the model was trained on large datasets of internet images, which likely over-represent this stereotype of what a physician looks like.

But, when the participants then generate the image, multiple things happen. Some pictures will show this stereotype. But others could show a female or maybe someone of color. And if students would redo the generation, other images might appear again. This is because this behavior of LLMs is known as such (responses vary because of the statistical nature of LLMs) and developers of newer models have attempted to correct for this through post-training interventions.

Additionally, we discuss what would have happened if we had asked the participants upfront before this exercise to draw a physician by hand?

Would they have thought about their own internal bias? And if they did so, how would they decide what kind of person they would draw? This thinking process draws them to the need to specify in the prompt what kind of person they would like to have generated in the image (a woman? A person of color? A young person with a green coat?). And this will depend most likely on the context and goal that the picture would need to serve.

So, it is not only within the statistical preference of an LLM what representation will be generated, but it is even more important that the user considers their personal awareness of stereotyping and accordingly decide on specifying the required output.

But it is not always so clear-cut what these kinds of cultural biases are. Think for example about questions related to cultural habits of individuals or groups. It makes a big difference if one were to ask a generative AI system to explain which hand gesture would mean ‘to calm down’ if you did not specify for which culture you are asking this. The same might be true, for example, about how to raise children or how to teach pupils. But also, simply when asked to generate a picture of a person. This person will likely be relatively young as compared to the average age in a society because the internet holds many more pictures of young people than old people.

A somewhat more covert bias will be for a multimodal model to generate a picture that shows ‘Humanitarian aid in Africa’. This particular example below sparked debate whether generative AI functions as a reinforcer of ‘poverty pornography’ (Fuchs, 2026). How does this relate to moral convictions?



Figure 6.3 Image of an AI generated picture of ‘humanitarian aid in Afrika’

6.4.2 Guardrails

p. 6–61

With generative AI, much is said about the so-called guardrails. Though not often discussed, it is in the hands of the developers of generative AI models how they methodologically and technically ensure that a user will not get responses for requests that might generate taboo output. Think of outputs on how to build bombs, biological warfare arms, or how to commit suicide. Or in China: what happened during the Tiananmen revolution. The underlying moral axiom is that it is wrong – and even legally prohibited – to provide individuals with such information, as it could harm individuals or societies.

Things do change over time. Technical and ethical hackers have repeatedly demonstrated that guardrails can be circumvented. Or worse: the operators of some commercial models have loosened or removed guardrails themselves. Elon Musk, for example, did this for his Grok models in 2025, allowing the models to produce content that earlier generations of similar models declined to generate. Think of profanity, jokes about taboo subjects, and various image-alteration requests. Decisions like this reflect the moral axioms held by the people who build and steward AI systems, not only the technical capabilities of the models themselves. For students and educators, this raises a broader concern. The answers that a generative AI system provides are partly an artifact of choices made by its developers. Choices that can shift over time and that may differ between platforms.



Figure 6.4 Images of an AI generated picture by Elon Musk showing himself in bikini's.

6.5 What this could look like in your course

Translating the framework of this chapter into a specific course requires you to make a few design choices, and then pick concrete activities that fit your discipline and the placement of your course in the curriculum. On the new VU EduBase website we [assembled a collection of exercises](#) you could use for this purpose. For most teachers, the practical entry point is to combine a few choices about how the course frames critical thinking, with a selection of two or three activities from the companion that students can repeat across the semester.

A first design choice is whether your course will treat critical thinking as a separate subject, with one or two dedicated sessions early in the semester on logic, fallacies, and the role of moral axioms. Or will critical thinking be an integrated running thread with short critical-thinking moments embedded in every session and applied to the content on the agenda. Neither is universally better. Disciplines where reasoning patterns are central to the subject, such as philosophy, law, or economics, could benefit from the integrated version. Disciplines where critical thinking is more related to knowledge and precise procedural methods, such as the natural sciences, could benefit from a dedicated unit followed by targeted reminders. The choice should match the cognitive demands students already face in your content.

A second choice concerns the role of generative AI in the activities themselves. In some activities, students use AI as the object of critique, evaluating AI-generated arguments, summaries, or images for reasoning errors and biases. In other activities, students use AI as a partner to support their reasoning, prompting it to challenge their assumptions, generate counter-arguments, or surface alternative positions, and then critically evaluating what it produces. Both approaches develop different aspects of critical thinking. We [assembled a collection of exercises](#) that includes exercises of both kinds, including bias-recognition exercises, AI source-attribution debates, AI-generated argument analyses, and discipline-specific critical evaluation tasks for fields such as biology, engineering, and the social sciences.

A third choice is how critical thinking will be assessed. If you only assess the final product of student reasoning, you measure the outcome but not the process by which the student got there. For a process-aware assessment, we [assembled a collection of exercises](#) covering options that include critical

reflection logs, annotated AI-output critiques, methodological reviews of AI-generated research designs, and personal AI use policies that students draft and defend. These options allow you to assess the reasoning quality directly, rather than infer it from the final text.

7 Generative AI and Assessment

The combination of generative AI and assessment presents us with a wicked problem: not solvable without always compromising somewhere, while we remain unsure about the outcome.

The main question for higher education institutions, examiners, and teachers is how we can assure that diplomas are granted to their students based on what students have learned and are capable to do at an individual level. That they are judged for their personal achievement, not the attainment of the generative AI used.

Before diving into this matter, we first discuss what the role of assessment is in higher education (apart from what already has been said in paragraph 4.1) and the added perspective of the last decades, and stretch it out into the era of generative AI.

p. 7-64

7.1 Two perspectives on assessment

The first perspective about assessment is the classical measurement perspective, originating from the beginning of the 20th century. Back then, the US developed measurement instruments to select army personnel (Yerkes, 1921) and described for example by Clauser (2021): assessment focusses on statistical methods to assess and improve accuracy in decision-making for pass-fail decisions, selection, and grades. Validity is a technical concept expressed in terms of coverage of content, coverage of content over cognitive taxonomies, and predictive validity for future tasks. This is the world of psychometrics. It focusses on tests being administered under supervised conditions and students working without any technical aids in time constrained situations to measure cognitive ability or attainment.

The second perspective is called the educationalist perspective. This perspective is not new and was already discussed in 1979 by Robert Ebel (1979). In this perspective, the focus lies on using tests and assessment to measure meaningful educational achievements and inform instructional decisions at the program and classroom level. But also, to prepare students for their future professions and life. A conceptual framing for this view stems from Black and William (1998) as well as Earl (2003) as they coined the concepts *Assessment For Learning*, *Assessment As Learning* and *Assessment Of Learning*. In this perspective, all forms of student activity, tasks, and assignments are considered instruments. Instruments to engage learners and provide feedback to teachers and learners about the progress of their learning towards personal goals and impact of their learning, for their future profession and for their lives (emphasizing these forms of validity).

Assessment For Learning emphasizes so-called whole tasks and processes surrounding it to promote engagement, while *Assessment As Learning* emphasizes whole tasks and processes to primarily learn from. A whole task is a small but complete authentic task in which students have to combine knowledge, skills and judgment as they would in professional or academic practice, rather than practicing isolated components (Van Merriënboer & Kirschner, 2018). Whole tasks are authentic: it resembles what a professional or researcher actually does, for example diagnosing a patient, writing a policy advice, or setting up and reporting a small study. Further, the task requires the coordination of knowledge, skills and judgment, including the routine and the non-routine aspects. Finally, the task is complete: the student goes through the full cycle (orient, plan, execute, evaluate) rather than practicing one step.

The tasks provided in these two forms of assessment perceive validity mainly from the perspective that they reflect tasks students will encounter in their future the professional or personal life.

Finally, *Assessment Of Learning* emphasizes the more classic psychometric perspective on learning. In more curricula development projects, often through (central) educational policies, the shift is towards more emphasis on Assessment For and As Learning. This is exactly in line with [the vision on education](#) and [vision on assessment](#) of VU Amsterdam.

Does this mean that all classic tests (exams) must be dismissed? No, certainly not, as there are many nuances and disciplinary differences. We will dive into them in the next sections.

7.1.1 Assessment practices for the two perspectives

In all disciplines, as we have seen in paragraph 6.1, it is imperative that students have and should build knowledge and skill upon a prior knowledge base. Such a prior knowledge base is often called foundational or core knowledge. Though these knowledge bases are quite different between disciplines (e.g. the difference between 'natural sciences'/well-structured knowledge domains versus the social sciences and further to humanities sciences/ill-structured domains (Donald, 2002; A. Jones, 2015; Neumann, 2001) they are imperative to learn more elaborate (deep and wide) topics. So, they must be learned during the first period of a Bachelor's or Master's program. The most effective way to acquire this foundational knowledge and skill is through direct instruction, largely both from a teaching and a learning perspective (Kirschner et al., 2006). In this perspective, applying techniques that motivate students to read information carefully, to critically analyze information (see Chapter 6), solve problems carefully in progressive complexity and the use of classic exams remains efficient and effective.

In view of the [VU Vision on Assessment](#), classic instruction and exams should (at least) be complemented with parallel trajectories that introduce students to real-world applications and their personal role in dealing with them. Therefore, it is common that in these first periods students also already work on small but authentic problems, projects, and research. This is important as it will let students experience directly what the foundational knowledge is used for in future professions and science. Such authentic tasks also stimulate students to learn about their learning gaps, metacognitive skills and development of more general academic skills (such as collaboration skills and communication skill). This will also feed into the new emerging taxonomic ideas around learning as described in paragraph 3.1 that it is a messy process and generative AI changes this to a certain extent. In the first Bachelor phase, these authentic tasks must but limited in complexity (to avoid destructive friction). During the progress of the program, less time can be devoted to learning foundational knowledge and more time can be devoted to learn more specialized knowledge or work on these authentic tasks with an increasing scope and complexity (for example, becoming less structured).

Vision for future-proof assessment

December 2022

The VU vision for future-proof assessment contains three design principles, related to the core values of the VU. The assessment vision describes how the learning process and assessment relate to each other and inspires programmes and teachers to design future-proof assessment.

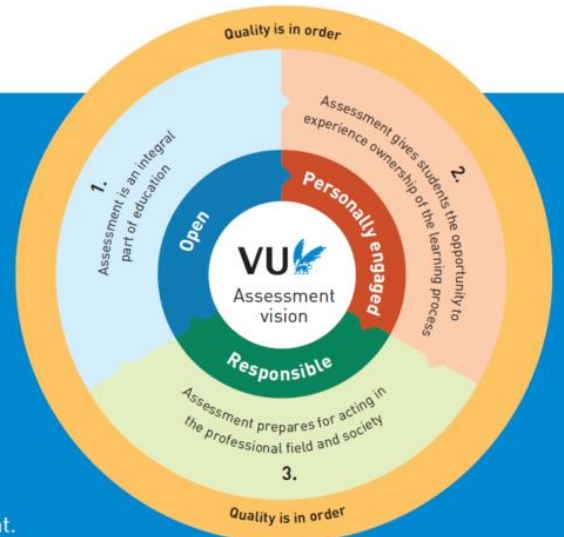


Figure 7.1 Figure showing the VU Vision for future-proof assessment.

7.1.2 Impact of generative AI

Second, the impact of generative AI on measuring an individual's achievement for these different teaching and learning methods varies. We can safely say that every task a student must perform without direct supervision is prone to cognitive off-loading (see paragraph 4.4). Think about tasks that students work on or solve outside class-time. So, this is the case for smaller problem-solving tasks, more elaborate tasks such as case analysis or larger tasks such as writing essays or doing projects, take-home exams and research work: indeed whole tasks.

Though many teachers may prescribe various levels and applications of use of generative AI (for example only for spelling and grammar or for generating ideas), in practice it will be unfeasible for teachers to know and check in detail if students stick to these policies. This effect increases when grades are given to students for these tasks and students feel an increased pressure to perform well. Certainly, if students suspect or know that fellow students use AI (and might not be 'caught' and hence radiate negatively to their personal relatively weaker product) could negatively impact their scores. A culture of 'policing' and 'fear to be unjustly 'caught'' for cheating is most likely the result.

At least, that is the conviction of Liu and Bridgeman of the University of Sydney (2023). Liu has a worldwide influence on thinking about the impact of generative AI on Assessment in higher education and coined the *Two-lane approach* (2023). Liu acknowledges first that when students work on assignments at home (like case studies, problem-solving cases, essays and take-home exams) without supervision, it is practically impossible to verify whether they're working entirely independently or using generative AI to any extent. This isn't a new problem in itself; students could previously receive unauthorized help from friends, parents, paper-mills or ghostwriters. Although this problem is common (Christensen Hughes & Eaton, 2022; L. R. Jones, 2011), it still appeared manageable. These forms of assessment were never forbidden despite their risk of violating academic integrity. It remained however a manageable risk for programs, provided the usage isn't widespread and systematic, and lecturers and examiners maintained oversight.

However, generative AI applications belong to a different category. This form of AI can solve problems, generate code, generate well-thought-out, well-formulated and structured content with almost no limitations and be undetectable at the same time. These applications are widely available, cheap (or

free) and are already being used extensively by students, including for other tasks such as graphic design, creating summaries, asking questions about the material, or reviewing papers for structure and spelling.

7.2 The two-lane approach

Liu and Bridgeman (2023) devised the two-lane approach to address this challenge.

Important note, the two-lane approach is not official VU Amsterdam policy

The guiding principle is that education must equip students to participate ethically and actively in a society permeated by AI. They therefore suggest that it's unnecessary to choose between, for example, abolishing homework assignments or letting everything slide. Instead, they propose that a good curriculum or course balances the two assessment perspectives: Assessment of Learning and Assessment for/as Learning (Black & William, 1999). Generative AI can play a role in both assessment perspectives. Validity must in their view prevail over cheating prevention (Dawson et al., 2024). Liu and Bridgeman describe curriculum and course design as a combination of elements from two lanes. It is the role of the curriculum designer and the teachers to select teaching and assessment methods from these lanes in an orchestrated manner fitting with their context. You'll read more about this in the next paragraphs.

p. 7-67

Two-lane approach to assessment

	Lane 1 - Assurance	Lane 2 – Development / Guided Use of AI / Learning about/with/through/against AI
Purpose of assessment	Assessment of learning	Assessment for learning
Value of assessment	Trustworthy judgements of capability	Equipping students for a world of AI
Scale of operation	Mainly at program level	Mainly at unit level
Security & integrity	Observed, in person	Not observed live
Position on generative AI	May or may not be allowed by examiner	As relevant, use of AI scaffolded & supported

The University of Sydney Page 10

Viewing tip 7.1 Watch this webinar of Danny Liu of 24 September 2025 for Npuls/SURF Community Digital Assessment and Development in which he explains the Two-lane approach in detail and answers questions about it. Link: <https://video.edu.nl/w/sifbfxmBYbJbgkaPc6wq9e>

7.2.1 Lane 1 – Assessment of learning

On one hand, programs, lecturers, and the professional field need to be able to determine what students can do independently with regard to knowledge accumulation and skill. This involves Assessment Of Learning, or: extra safeguarding of learning verification under live physical oversight. This forms the first 'lane'. This requires supervision, where students must present and respond to questions from lecturers or examiners directly from their cognitively accumulated (automated, broad, deep) knowledge base. These assessment forms include exams, presentations, interviews, defenses, and poster presentations. The use of generative AI isn't necessarily prohibited here. A student might need to demonstrate their

competence in using generative AI, whether in combination with an assignment or test. But in other words, again: classic exams and orals must remain part of the assessment landscape.

7.2.2 Lane 2 – Guided use of AI

On the other hand, students must learn to harness the opportunities of generative AI while understanding its limitations and using it responsibly. This approach centers on Assessment *for* and *as* Learning, where teachers guide students in using generative AI productively and ethically. Generative AI is viewed as an integral part of students' toolkit, both during their studies and in their future professional lives. The task of the teacher is to guide and coach students in productive and responsible use.

Through instructor and peer guidance, prompt engineering practice, critical analysis, discussion, and emphasis on academic integrity, educators maximize student learning via assignments and, at VU Amsterdam, through 'Formative Dialogue'.

Face-to-face sessions become more important

An important aspect to understand is that formative dialogue has many practical implementations, but primarily (though not solely) takes shape in face-to-face sessions. These could be lecture settings, but also in tutorial groups, in group work, in project work, in coaching session or in thesis supervision sessions.

The coaching, peer discussion and formative dialogue only function if students actually show up. If lane 2 activities are not graded or only lightly weighted, then attendance becomes more optional in practice, and the guidance Liu envisions may not reach the students who most need it. The clarity of the two-lane image partly depends on quietly assuming presence that many real programs do not enforce.

p. 7-68

Student activities and outputs in this lane contribute to the final assessment of acquired competencies, including responsible AI use. For lane 2, Liu suggests teachers *to* inform and guide students on the proper, responsible and effective use of AI (not prohibiting it). He makes a comparison with a menu in a restaurant that students can use and that can be tried out for fit with the task at hand. The choice is determined by the learning objectives that can (partly) be derived from AI's place in life and future work. Below you'll find an example of such a menu.

Category	Items
As a critical friend – Soups	Suggest analyses, Provoke reflection, Provide study/organization tips, Practicing
Getting started – Entrees	Suggesting structure, Brainstorming ideas
Engaging with literature – Bread service	Suggesting search terms, Performing searches, Summarizing literature, Identifying methodologies, Using AI driven research tools, Using specialized genAI advice tools, Explaining jargon, Fixing reference list
Generating content – Mains	Writing some/lots of text, Making images, video, audio, Making slide decks
Analyses – Lighter mains	Performing analyses of data/text, Generating code (Python, R etc.), Generating graphs, Automate tasks, Suggesting counterarguments
Editing – Coffees	Editing tone, Improving clarity and readability, Fixing grammar, Shortening
Feedback – Desserts	On all of the above elements, Specifically on rubric criteria

Table 7.1 Liu's menu of options to use generative AI that a students and teachers could use and see the fit with the task and learning goals at hand.(Liu, 2025)

7.2.3 How to use results of lane 2 assessment

It's also important to note that the Two-lanes are not similar in meaning like summative and formative assessment. Summative assessment is often regarded as giving grades and most often under monitored conditions (such as exams) but also giving grades to tasks or assignments. Formative assessment is often regarded as providing feedback but not giving grades and emphasizing student's personal responsibility in doing tasks and information processing as an inherent, intrinsically motivated effort. Summative assessment and formative assessment however have a place in both lanes of the Two-lane approach: it is not a one-on-one relation. To arrive at a thorough and comprehensive final assessment of the student's competence, there are two options for lane 2 assessment:

1. You can and should (according to Liu) still include the results from lane 2 assessments in the total final grade in summative sense. However, these assessments should carry relatively less weight in the final grade than traditional assessments. The amount of weight reduction is dependent on the risk invoked by the particular assessment.
2. However, lane 2 assessments can also be used as a formative assessment only, without counting towards the final grade. In that case, the effort and outcome of this lane 2 assessment are implicitly reflected in the result of the summative assessment. This use of lane 2 assessment is in line with the formative dialogue as promoted in the [VU vision on Assessment](#).

7.2.4 Orchestrating Lane 1 and Lane 2

In a curriculum, and in every course, program managers or lecturers must make a well-considered choice about implementing methods from both lane 1 and lane 2. These must be hinged to the discussions and decisions about the learning outcomes as described in Chapter 5. Each course contains elements to directly assess students' knowledge development under controlled conditions, and elements where AI use is allowed or even promoted or obliged to use. In the second lane's implementation of lessons, workgroups, projects and lectures, AI is purposefully taught, used and critically evaluated.

For a thorough final assessment of a student's competency, a program, or course includes judgments from both lane 1 and lane 2. Liu puts forward a 'new' list of assessment types as seen in the table below.

New Assessment types (secure, lane 1)		New Assessment types (open, lane 2)	
Final exam – secured	Written exam	Practice or application - open	In-class quiz
	Practical exam		Out of class quiz
	Oral exam		Practical skill
In-semester test – secured	Written test	Inquiry or investigation - open	Experimental design
	Practical test		Data analysis
	Oral test		Case studies
In person - secured	Interactive oral		Research analysis
	Practical or skills test	Production and creation - open	Portfolio or journal
	In person practical, skills, or performance task		Performance
	In person written or creative task		Presentation
	Q&A following presentation, submission or placement		Creative work
			Written work
Placement, internship, or	Peer or expert observation or supervision		Dissertation or thesis
			Debate

supervision – secured	In person practical or creative task	Discussion - open	Contribution
	Clinical exam		Conversation
			Evaluation

Table 7.2 Suggested assessment types in lane 1 and lane 2 according to Liu (2025).

7.2.5 Notes and implications

The strength of the two-lane approach perspective lies in the model's clarity. It suggests a downsizing of complexity and trade-offs as described in Chapter 4 about The wicked problem of AI in Education, to a dilemma as a manageable task that allows for controlled discussion and implementation. Of course, the model is a simplification of reality. Upon deeper inspection, you'll discover that there will be overlap and many courses are already designed according to the two-lane approach principles. In the graph by Liu, the overlap is expressed in the wig-shapes of the lane 1 and lane 2 characterizations.

	Lane 1 - Assurance	Lane 2 – Development / Guided Use of AI / Learning about/with/through/against AI
Purpose of assessment		
Value of assessment		
Scale of operation	Mainly at program level	Mainly at unit level
Security & integrity	Observed, in person	Not observed live
Position on generative AI	May or may not be allowed by examiner	As relevant, use of AI scaffolded & supported

p. 7-70

Figure 7.2 Liu's slightly adapted perspective on lane 1 and lane 2 assessments in a curriculum (program) and course level (unit).

Cath Ellis of Western University in Australia even goes a step further in looking at the Two-lane approach (Gallant, 2025). She introduces Western Sydney's Assure and Inspire (AI) Approach to assessment (see figure below), which is their interpretation of the two-lane model.

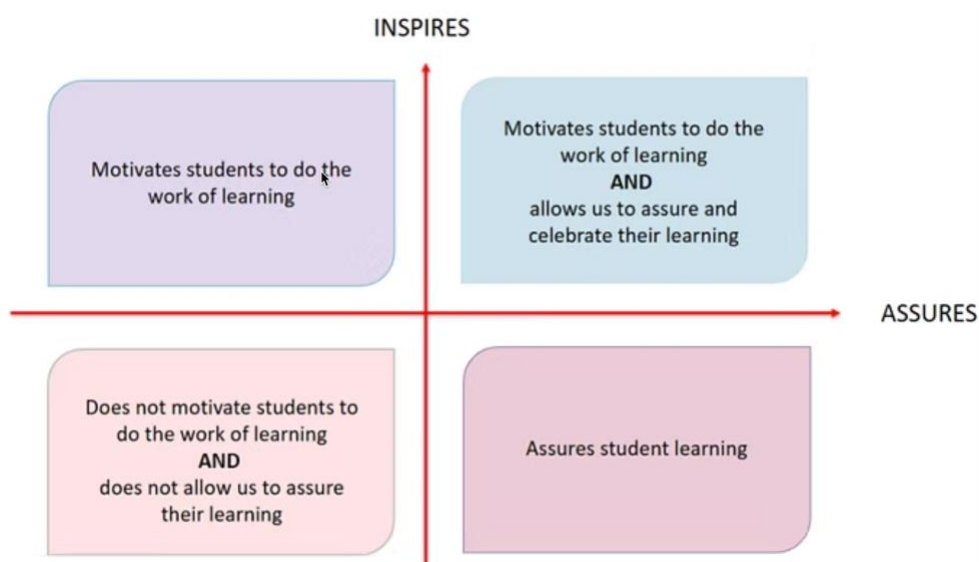


Figure 7.3 Western Sydney University AI Approach to Assessment (Assure and Inspire).

What follows from the two-lane approach and the VU Vision for Assessment is that VU Amsterdam needs no revolutionary change of educational paradigm. It does require programs and lecturers to recalibrate their learning objectives continuously, and to orchestrate the various forms of assessment across a curriculum with care. That is what allows a program to incorporate generative AI defensibly, and to keep guaranteeing the value of the diploma.

7.2.6 Conditions for making the two-lane approach work

7.2.6.1 Get exam boards and program committees aligned

A first important condition to note to make the orchestration intended by the two-lane approach work, is that exam boards and program committees coordinate their activities and perspectives on what the assessment program must be. It is imperative that the course leaders and the exam boards know how the quality assurance of the assessment program *as a whole*, leads to a defensible value of the diploma. This way there needs to be no panic about students possibly 'having ChatGPT write their whole assignment without the teacher spotting it'.

An important additional condition are clear guidelines on how to conduct the in vivo assessments and formative dialogue. Formative dialogue, interviews, discussions, presentations and oral must be conducted well. In terms of the stakes of the activity, it is good to realize they are on a sort of continuum. From activities with no- or low stakes that only focus on the learning, to activities that have medium stakes by contributing a bit to a final grade, up to conditions with high stakes in which the activity determines a large part of the final grade. The higher the stakes, the stricter the guidelines and rules are to make assessment reliable, consistent and traceable. Formally, the latter (oral exam) has the most stringent guidelines that are laid down in the Education and Exam Rules. It requires currently for example that two assessors are present at an oral exam and only the student can ask for recording of the session. For the face-to-face activities that do not carry that much weight, much more flexible formats are possible. But that is often not very clear and exam boards can be strict on rules and easily denote every face-to-face activity that leads to some credit for a final grade to be conducted under oral exam conditions. VU Amsterdam is currently considering publishing a memo with more nuanced guidelines.

7.2.6.2 Oral assessment and group presentations: a practical walk-through per phase

Oral exams, assessment interviews and group presentations are gaining ground again, partly because they are hard to outsource to generative AI and partly because they show what a student can actually do with the material (Sotiriadou et al., 2020; Theobold, 2021). Lecturers may run them on intuition. That can make them time-consuming, stressful for students and vulnerable to inconsistency and appeal. Table 7.3 Tips for conducting sound assessment using interviews, oral assessment or group presentations. below describes what to arrange in each phase, from preparation to follow-up, and why.

Phase	What to arrange	Why, and what the literature says
1. Preparation	Decide what the oral component is for: a stand-alone summative test, a check on authorship of a written product, or a formative rehearsal. Write this in the course manual, with the weight in the final grade and the rule for absence or failure. Share a short rubric (three to five criteria) two to three weeks in advance. For group presentations, state whether all members receive the same grade or whether individual contributions are weighed, for example via a question round per student or a peer contribution form. Publish a time schedule with buffer slots, book a quiet room, arrange a second assessor or note-taker for high stakes exams (e.g. thesis defense), and tell students what to bring along. Prepare a question bank per learning objective. Do a fifteen-minute dry run with a colleague. Important, if an oral will pose the same (content) questions for, for example, a fixed book, oral assessments are not useful (questions can be shared easily). Use oral assessment to probe students understanding of a piece of work they did. Then you can ask similar questions because they will revolve around the specific student's topic.	Oral assessment varies on six dimensions (content, interaction, authenticity, structure, examiners, orality); deciding these in advance is what makes the format defensible (Joughin, 1998). A standardized assessment instrument and two assessors are the main safeguards against assessor effects such as norm drift, sequence effects and halo effects (Bergwerff et al., 2015; van Berkel et al., 2014). Structured practice opportunities and mock orals improve student preparedness and confidence (Stephenson et al., 2025). For group work, a combination of a group grade and an individual grade is preferred (van Berkel et al., 2014).
2. Starting the session	Welcome the student by name, repeat the procedure (duration, who asks what, when the grade follows) and confirm the criteria are the ones shared earlier. Open with an easy question such as "Take me through your main finding" or "How did you approach the work". Let a presenting group not start on its own terms. Use the technique of pointing out students yourself for presentation of specific parts (so students need to be able to present on every aspect of the work done). Say that you will take notes, so writing is not read as disapproval.	Anxiety lowers performance and therefore validity; a predictable opening is a measurement precaution rather than a courtesy (Huxham et al., 2012; Theobold, 2021). Transparency about criteria and procedure is a formal quality requirement of assessment in Dutch higher education (van Berkel et al., 2014).
3. Conducting the interview or presentation	Follow the rubric or script in your questioning and move from for example reproduction to application to justification: what did you do, why this choice, what would change if. Use silence; give a few seconds before rephrasing. Agree in advance which prompts are allowed (repeat, clarify, narrow) and which are not (giving half the answer). For group presentations: strict timing with a visible timer, then a fixed question round in which each member is addressed by name. If peers assess, use a form with two or three criteria.	Interactive orals allow assessors to probe reasoning in a way written work cannot, which is their main validity advantage (Sotiriadou et al., 2020; Theobold, 2021). Consistent prompting across students is needed for comparability; unstructured prompting is a known source of unfairness in oral formats (Joughin, 2010). Assessor fatigue and sequence effects are documented reasons for changing roles and order (Bergwerff et al., 2015).

	Separate roles: one assessor leads, the other scores and watches the clock. Swap halfway through the day.	
4. Closing	Signal the end explicitly ("two minutes left, anything to add?"). Thank the student, state when and how the result follows, and ask them to leave before you deliberate. For presentations, close the question round and let the group step aside while assessors confer.	A clear closing ritual reduces negotiation in the room and protects the independence of the assessors' judgment (Joughin, 2010).
5. Feedback, oral and written	Give a two to three minute oral debrief shortly after the session if the decision is clear: what went well, what was missing, one improvement point. Keep it descriptive. Follow up with written feedback in the rubric within a week.	Feedback works when students understand what to improve and how; written feedback lets them reread it once the stress has subsided (D. J. Nicol & Macfarlane-Dick, 2006). Feedback from instructors and peers is the most consistently reported facilitator of oral assessment performance (Stephenson et al., 2025; Van Ginkel et al., 2015). Written documentation also protects the assessor in case of appeal (Bergwerff et al., 2015).
6. Grading directly or afterwards	Complete the rubric immediately after each session and deliberate five minutes with your co-assessor. Direct grading and release of grades is defensible for low-stakes formative rounds or a single student.	Grading only after all work has been assessed is a standard assessor instruction in Dutch assessment handbooks because of norm drift and sequence effects (Bergwerff et al., 2015; van Berkel et al., 2014).
7. Follow-up		Evaluating and improving the assessment is the final step of the assessment cycle (Bergwerff et al., 2015). Self-reflection and review of performance also help students; consider recording presentations for later self-assessment (Stephenson et al., 2025; Van Ginkel et al., 2015).

Table 7.3 Tips for conducting sound assessment using interviews, oral assessment or group presentations.

Below we list a few additional tips.

- Tip 4 — [How do you prevent free-riding in group assignments?](#)
- Tip 28 — [Constructive feedback? Ten tips!](#)
- Tip 36 — [How to develop a rubric](#)
- Tip 49 — [How to consider neurodivergent students in your teaching](#)
- Tip 75 — [How to support students with autism](#)
- Tip 76 — [How to consider students with dyslexia](#)
- Tip 77 — [How to support students with ADHD](#)

7.2.6.3 Technology use by teachers

Another condition is the technology that teachers use. And that is important because the assessment activities in lane 1 places a heavy burden on academic staff. Systems must be in place to record the activities easily and safely. Generative AI makes recording such activities much easier than ever before. Transcripts can easily be produced and - using generative AI - be summarized and stored for later reference. Currently, however, no clear guidelines exist about which technologies can be used formally, how the data is stored etc. Also for those developments, VU Amsterdam is working on additional guidance for staff and applications.

7.2.6.4 AI tools for students and staff

Finally, higher education institutions must provide a level technological playing field for all students. The variation in AI competencies and availability of technologies is big: students who are less proficient or have less access to generative AI tools must not be disadvantaged compared to students with more access. In lane 2, the institution should have a licensed set of high quality LLMs and tools available for all students. VU Amsterdam has these tools available, see paragraph 2.1.1 to 2.1.4. Clear monitoring and control by programs and IT services is needed on the rapid development of AI in professional software and plug-in applications within every conceivable application. Standard packages are a thing of the past, and safely offering only Copilot or Gemini to students isn't sufficient. This demands flexibility from future-proof IT support organizations and financial space.

7.3 Example: Writing papers and dissertations

It goes without further explanation that dealing appropriately with generative AI in the writing process of papers and dissertations is one of the most challenging tasks.

As an illustration which reflects many aspects of the Two-lane approach, interactive orals and the move from focus on fraud prevention to learning, we will take the course Academic Research Writing (or Capstone) project of the Amsterdam University College. This course is coordinated and supervised by Joost Krijnen and Lotte Tavecchio. As the title of the course suggests, the goal of the course is to learn how to conduct research, with a focus on the writing part.

During the last years, the course coordinators experimented with various options on how to deal with generative AI in that course. They soon realized that, indeed, paying more attention and giving more weight to the process of doing research and writing of the student was imperative. This is needed for a sustainable and defensible judgment about student achievement considering the use of generative AI.

To start with, and to a large extent, the thesis writing and assessment guidelines and process remained unaltered. So, students had to work on a few supporting assignments, as before, and the rubrics for assessing the quality of the final product remained the same, including all the criteria for handing in original and proper use of sources. But a few distinct components were changed, as explained below.

The first important intervention is that the coordinators incorporated a dedicated classroom session on AI and academic competency in learning outcomes in the grading. The classroom session covered the following components:

- An open dialogue with students regarding their usage, experiences, and concerns related to generative AI.
- Fundamental concepts: Providing definitions and explaining what generative AI and large language models (LLMs) are, as well as how they function.
- Critical evaluation: A balanced discussion of the advantages and benefits versus the disadvantages and risks associated with these tools.
- Course policy and responsibility: Clarifying what constitutes responsible use of generative AI and how it relates to university regulations concerning fraud and plagiarism.
- Academic integrity and referencing: Practical instruction on how to correctly reference the use of generative AI using styles such as APA or MLA.
- Practical exercise: A short exercise focused on prompting.

Regarding the academic competency, the coordinators decided to eventually reserve 30% of the final grade for these competencies. Grading for these competencies was done directly, at location, after meetings with a student, and an easy to fill in shared Excel sheet was used for this purpose. The Excel

sheet is a simple table with a limited number of criteria and achievement descriptions. A minimum of three meetings are set during the project. The process is that the student hands in their preliminary work a few days before the meeting so the supervisors already have information upfront. Then, one supervisor interviews the student for 15 minutes and directly afterwards fills in the rubric. These sessions are not recorded.

Criteria	Insufficient (3)	Sufficient (6.5)	Good (8.5)	Excellent (10)
Quality of the Work in Progress				
Ability to Reflect on Selection and Analysis of Sources	Student struggles to explain why sources were chosen or how they contribute to the project. Little to no critical engagement is evident.	Student can explain the relevance of some sources but shows limited ability to reflect on their role or critically engage with them.	Student provides a clear and thoughtful explanation of source choices, demonstrating growing critical engagement.	Student demonstrates deep insight into the selection and use of sources. Critically reflects on their value and limitations in a convincing way.
Demonstrable Progress Between Meetings & Ability to Engage with Feedback	Student shows little to no progress since the last meeting. Feedback from the previous session is not addressed or understood.	Student shows some progress. Attempts to respond to earlier feedback are evident but not fully effective.	Student clearly explains how feedback has been considered and used to improve the project. Progress is meaningful and well communicated.	Student shows clear, thoughtful engagement with earlier feedback and has made strong, well-argued improvements as a result.
Preparation, Discussion & Engagement with Research Writing Process				
Level of preparedness for meeting	Student appears unprepared and has difficulty engaging in discussion. Responses are brief or unfocused.	Student is somewhat prepared but struggles to articulate ideas or follow the discussion fully.	Student is prepared and actively engaged in the discussion. Communicates ideas clearly and thoughtfully.	Student is fully prepared, highly engaged, and shows ownership of the project. Actively contributes to a thoughtful and productive discussion.
Demonstration of Critical and Reflective Competency	Student shows little to no ability to reflect critically on choices, challenges, or the writing process.	Student shows some reflective thinking but responses lack depth or clarity.	Student reflects critically on key decisions, challenges, and writing strategies in a clear and thoughtful way.	Student demonstrates strong critical thinking and reflection, offering nuanced responses and showing awareness of both strengths and areas for growth.
Ability to Work Independently and Take Responsibility	Student demonstrates little initiative and depends heavily on guidance. No clear planning or next steps are articulated.	Student shows some initiative but requires support to move forward. Future plans are vague or underdeveloped.	Student communicates clear next steps and demonstrates growing independence and responsibility for their work.	Student shows strong initiative and ownership of the research process. Clearly articulates goals and next steps, and takes full responsibility for development.

Table 7.4 Rubric used in the AUC Capstone projects to assess Academic Competence.

Grading sheet Academic Competencies in ARW

		score				
Academic Competencies Assignments		Not submitted	Submitted			10
Reflection on goals			1			
Pitch transcript			1			
Peer review of Research Proposal			1			
Methodology and Impact			1			
Literature Review: Introduction & Outline			1			
Selfie at PhD defence			1			
InPrint Peer review (bonus 0.3 to final AC grade)						
Individual meeting 1		Insufficient	Sufficient	Good	Very good	Excellent
Research Topic		3	6,5	8	9	10
Quality of the work in progress						
Has the student been processing a good amount of relevant literature?				1		
Is a clear and viable research angle / research question starting to take shape?				1		
Preparation & Discussion						
Has the student prepared well for the meeting?					1	
Does the student demonstrate critical and reflective competency?					1	

Comments: You wrote a strong pitch in which you demonstrated some quite concrete research ideas already. Likewise, in our conversation it became clear that you're very much on top of things. I look forward to reading more!

Individual meeting 2		Insufficient	Sufficient	Good	Very good	Excellent	
Introduction & Outline		3	6,5	8	9	10	8,5
Quality of the work in progress							
Do the introduction and research question/thesis show good promise?					1		
Does the outline offer a clear overview of the paper in progress, and do the plans seem viable?				1			
Preparation & Discussion							
Has the student prepared well for the meeting?				1			
Does the student demonstrate critical and reflective competency?					1		
Writing Update		Insufficient	Sufficient	Good	Very good	Excellent	9,5
		3	6,5	8	9	10	
Overall quality						1	
Submission		on time	# days late				
			1				

Comments: You are well under way and seem to be fully in control. The RQ sounds good and you sound very knowledgeable.

Comments: See Canvas

Individual meeting 3		Insufficient	Sufficient	Good	Very good	Excellent	
From Writing Update to Final Draft		3	6,5	8	9	10	9,0
Quality of the work in progress							
Does the student engage effectively with feedback?					1		
Preparation & Discussion							
Has the student prepared well for the meeting?					1		
Does the student demonstrate critical and reflective competency?					1		
Engagement in course and process		Insufficient	Sufficient	Good	Very good	Excellent	8,375
		3	6,5	8	9	10	
Participation in class discussions			1				
Works independently and takes responsibility for the project						1	
Demonstrates progress through attention to feedback					1		
Demonstrates punctuality with regard to deadlines, meetings, agreements				1			

Comments: This was a pleasant conversation about your paper. It's clear that you're well on track with everything and fully in control.

Comments: While relatively quiet in class discussions, you were engaged in the course as well as in your own research project throughout the semester.

Overall grade Academic Competencies 9,0

Figure 7.4 Example simple Excel sheet to keep a record of student attainment and progress

Thirdly, the coordinators set a policy for the use of generative AI for their course that is not prohibitive. It is acknowledged that students will use generative AI, and the role of the supervisors is to guide student in using it responsibly. The following rules are set in their policy document.

Policy on using Generative AI in Academic Writing Skills and Advanced Research Writing

The course Advanced Research Writing aims to teach advanced academic writing skills to the (academic) writers of the future. As teachers, we have no doubt that the writers of the future will do their writing supported by AI technology. Therefore, we also believe that today's students need to learn and practice using AI technology critically and responsibly in their writing.

Generative AI technology and the advantages, disadvantages and risks of using AI tools like ChatGPT will be topics for discussion in this course. Moreover, students are permitted to use generative AI technology in working on their writing projects for this course, provided they adhere to certain rules and guidelines, as outlined in this document and to be discussed further in class.

Rules & Guidelines

1. Students may use generative AI technology to support their writing, in ways similar to how internet search engines, grammar and spell checkers, and (online) dictionaries and thesauruses are used.
2. Students may use generative AI as a source of information and to generate content, as long as they use the tool as an assistant, not an author. Instructions on how to do this are provided further down this document.
3. If students use generative AI tools, they must provide full transparency about how they use them.
4. Instructions on how to do this are provided further down this document.
5. Students are advised to familiarize themselves with the advantages as well as disadvantages and potential risks of using generative AI. These topics will also be discussed in class.
6. Students who make use of generative AI tools in their writing assignments remain fully responsible for the texts they submit.
7. The "Regulations governing fraud and plagiarism" (AS&P, Appendix 2) apply to any assignment submitted in the course, written with the help of generative AI or not.
8. These Rules and Guidelines apply exclusively to the course Advanced Research Writing. In other contexts, such as the Capstone, students should contact their teacher or supervisor for permission and guidelines on the use of generative AI tools.

How do you use generative AI as an assistant, not an author?

Generative AI can perform and can even take over certain tasks for you, but you should remain in charge of the overall writing and thinking processes at all times. This means you need to use generative AI critically and responsibly. Do not take over information generated by generative AI tools unthinkingly. And do not incorporate generated output into your text immediately and without acknowledgments. In short, be smart, not lazy.

How can you offer complete transparency about your use of generative AI?

When you use generative AI technologies such as ChatGPT as a source of information, you should treat them like any other source of information you use in your academic writing. Thus, when you incorporate information presented to you by a chatbot, you must summarize, paraphrase, and/or quote it, attribute it clearly and provide a correct citation. (The technical details will be discussed more fully in class). Do note, however, that you should critically evaluate the quality and reliability of the information you have generated, just as you would with other forms and sources of information.

When you use generative AI to brainstorm ideas, to grammar-check, style-check, and/or enhance your writing, to generate content (including outlines, larger sections of text, or bibliographies), or any other uses, you should explain your approach clearly and in sufficient detail. You can do so, for instance, in an explanatory footnote or as part of your methodology section. Additionally, you are required to save and provide full access to your complete chat history for your project. If you use ChatGPT, for example, you can give access by clicking on the "share chat" button and pasting the link in your methodology section or explanatory footnote.

Example 71 Policy about the use of generative AI in the Advanced Research Writing Course at Amsterdam University College.

Fourthly, the project finishes with the student giving a presentation of the research done. During that presentation, the student is again assessed on their knowledge about the research, which assures the development and achievement of the individual student.

The combination of the four elements (three meetings, the inclusion of assessing academic competence, the guided use of generative AI and the final interview) makes the course an exemplar of using the Two-lane approach. On the one hand, the assurance of learning is realized by close monitoring and assessing of the individual development of the academic competence in combination with monitoring and discussing the use of generative AI.

A commonly mentioned objection to the approach of AUC is that the intensive supervision and assessment process takes up a lot of time. This is true to a certain extent. During an interview with the coordinators, they, however, argue a different perspective. Firstly, even without the emergence of generative AI, writing processes deserve such a close guidance by supervisors as academic writing is a

core activity and competence that students need to acquire. Guidance and feedback are essential for this learning process.

Secondly, the emergence of generative AI requires this approach, as the value of the diploma rests on the program being aware and addressing the ubiquitous presence of generative AI in the professional field and sciences. Finally, they found that students themselves experienced enormous value from their approach: “the approach we have ensures that we really ‘see’ the student and acknowledge their individual struggles and effort. Students also emphasize this in their feedback. They say that this course is where they really feel that they make great strides in their development and feel to become part of the learning and research community.”

So, how can you make this process run as smoothly as possible or with not too big a demand for supervisor time? First, an interesting exercise for any educational program is to calculate the total supervision time for the amount of ECTS realized. It could surface that the supervision time per ECTS is actually limited. Of course, having three supervision sessions in a 6 ECTS project, results in a higher load than three supervision sessions in a 24 ECTS project.

An additional approach is to not let students work alone on their thesis, but work in groups of about 3-4 people. Students preferably work on different topics or research questions. In this way, students can prepare their meetings as a group and as a teacher, you can incorporate peer-feedback processes or deploy self-assessment processes. For example, when a discussion with the supervisors is due in the second half of the week, students can be requested to hand in their concepts in FeedbackFruits on Monday. They can invite their [peers to comment on the work](#) and discuss it already on Tuesday, and send their common findings to the supervisors. A good example with detailed guidance of how this can be done is described by dr. Danny Scholten of VU Amsterdam in the following two didactic tips:

- Tip 37: [Do's and don'ts for effective peer feedback](#) - part 1
- Tip 38: [Do's and don'ts for effective peer feedback](#) - part 2

Additionally, the group collaboration and setting the supervision meetings in advance, helps to prevent students can feel alone or isolated when working on their project and that they tend to fall behind in the execution of their projects.

Senior lecturer Sharon Klinkenberg of the University of Amsterdam has an additional approach: “I now organize the writing process via Microsoft Teams, where each student gets their own personal channel. I explain to them how they can use Microsoft OneDrive to access their files within that team. And that's where they get to work. As a result, my role is shifting more towards that of a process supervisor. Students can tag me in their team or documents with short questions. This allows me, as a teacher, to keep a close eye on their progress. For students, the temptation to use AI becomes much smaller – precisely because their process is visible step by step.” You can see how he organizes his process [on this webpage](#).

7.4 Example: Genre writing

Gea Dreschler of VU Amsterdam's Academic Language Program devised a writing intervention that incorporates the use of generative AI to learn about genre conventions in writing, specifically in academic writing (Dreschler et al., 2025). The task is embedded in a first-year, standalone academic writing course (3 ECTS) that serves as the university's core activity for teaching scholarly prose. The instructor deliberately chose one slice of the “writing pie” – genre conventions – as the focus for a single, high-impact assignment. By concentrating on genre conventions the course could isolate a concrete, teachable skill set (time management, drafting, revision) while still engaging the broader writing process outlined in the pie.

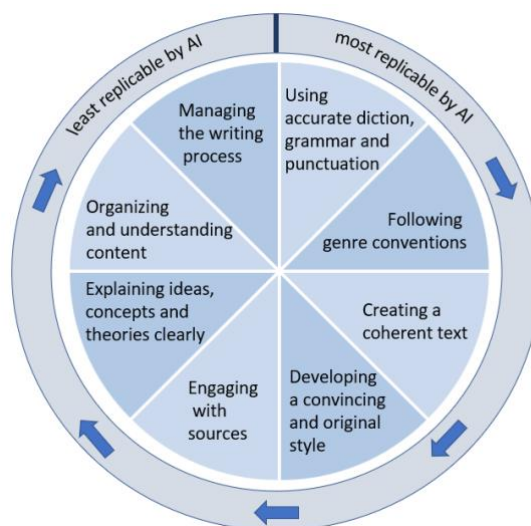


Figure 7.5 The writing pie: A conceptualization of human and AI writing skills.

Students are asked to produce a short research paper that explicitly demonstrates mastery of academic genre conventions: clear paragraph structure, appropriate headings, cohesive devices, and a consistent citation style. They begin by planning the outline, then use ChatGPT to generate first draft prose segments that they subsequently edit for register, coherence, and accuracy. Throughout the process, students must manage time, revise iteratively, organize content logically, explain ideas clearly, and engage with sources. These are skills mapped directly onto the lefthand slices of the writing pie. The final manuscript must therefore reflect both the structural norms of scholarly writing and a visible trace of student led revision.

Then the issue of genre is highlighted in the task. Students were given a text with many features of informal writing and slightly adapted it by removing citations, shortening it to a single paragraph and adding some informal features that were discussed in the teaching materials, such as sentence fragments. The teachers then provided the paragraph to ChatGPT 3.5 with a simple prompt:

Could you please increase the formality of this text so that it is more academic in style?

This output was then used in the classroom activity. Table 1 presents some example sentences from the exercise.

Student Text	Text output by ChatGPT 3.5
On the other hand, the Internet of Everything (IoE) creates lots of opportunities.	On the contrary, the Internet of Everything (IoE) engenders a plethora of opportunities.
Not the advertisement that you see in a newspaper, but the one that you will see on smart screens.	However, this advertising does not manifest itself through traditional print media, but rather through the medium of intelligent screens.
Have they never thought of tourists who have never been in a major city like New York and would like to drink a coffee at a Starbucks store? No.	Nevertheless, they tend to overlook the needs of unacquainted tourists who find themselves in unfamiliar urban locales, such as New York, and yearn for the prospect of patronizing a Starbucks establishment, for instance.

Table 7.5 Examples from the student text and suggestions by ChatGPT 3.5

The teaching procedure was then as follows. First, the students watched an online recorded video on basic characteristics of formal written English. The lecture focused on what is typically avoided in academic writing (such as vague words, informal phrases and sentence fragments) and what is typical for the genre (such as using standard academic phrases and precise language). Second, the students attended an in-person classroom activity, which covered the topic of academic style. During the activity, the students worked on academic style and formality that implemented ChatGPT.

The classroom activity began with an analysis of an informal student text. Students identified informal features and suggested more formal alternatives. They were then presented with the text edited by ChatGPT and evaluated the suggestions. They identified changes they considered effective or ineffective in terms of formality.

Some of these changes, such as “engenders a plethora of” and “yearn for the prospect of patronising a Starbucks establishment” (see Table 1), raised interesting questions about what level of formality fits the communicative purposes of the genre. Other changes, such as “smart screens” to “intelligent screens,” revealed a potential limitation of using ChatGPT to increase formality in this way. Students argued that “smart screen” is the name of an actual technology, rather than an informal description.

These discussions took place at each stage of the activity, allowing students to receive feedback from each other and the teacher. This feedback helped them determine concrete strategies they can later apply to their texts, aligning with the learning goals.

8 Generating feedback on assignments with LLMs

Generating feedback on student assignments using generative AI is an interesting application for educators. It could speed up the process of generating feedback on assignments, which is helpful for students to be able to use the feedback to improve their work. Also, it could improve the quality of the feedback as a more thorough analysis could be conducted. These two aspects could free up time for teachers to devote to teaching or other activities. The latter is often seen by teachers as an end in itself.

At first glance, it appears straightforward: provide the LLM with the assignment description, the student's work, and a grading rubric, and have detailed, personalised feedback generated. However, this seemingly simple process involves complex legal compliance regulations. And it requires considerable care to produce reliable, fair, and pedagogically sound results. For example, take the fact that an LLM can provide different feedback with the same grading rubric and submission.

For that reason, in this Companion, we cannot elaborate more on this topic currently.

p. 8–81

AI Regulation

- AI systems used in education to assess learning outcomes or guide the learning process may fall under the high-risk category (Annex III of the AI Act). The fact that no grades are assigned does not rule this out. The VU must properly substantiate and document this classification;
- It must be clear whether the VU is a provider or a controller. This classification determines which compliance requirements the VU must meet;
- It is important to note that the VU can be both; if faculty members/researchers themselves build a system around an existing model, the VU may be considered a provider for that system; o Students must be aware that they are receiving AI-generated feedback, and the staff involved (faculty members, researchers) must possess sufficient AI literacy;
- The faculty member's review ("human in the loop") must be substantive and demonstrably structured—not merely a formality. "Human in the loop" means that the decision or grading is ultimately made by a human. How the training is conducted beforehand has no bearing on this. Personal training on the tool cannot replace this, and even with a random sample, there is not always a "human in the loop."

GDPR

- The focus here is on Article 22 of the GDPR: data subjects have the right not to be subject to a decision based solely on automated processing that produces legal effects or significantly affects them.

9 Grading open-ended exam questions with LLMs

A specific case of generating feedback with LLMs is the process of grading open-ended questions of exams using generative AI. An important driver to use LLMs for this purpose is that it could free up time for teachers. In particular for teachers with large student cohorts that take exams (e.g. > 100). Note that open-ended questions in this case require short answers for which quite explainable and unambiguous grading rules can be set.

Comparable to providing feedback to student submissions, it appears straightforward: provide the LLM with the question, the student answer, and a grading rubric, and have a score with a rationale generated by the LLM. However, this seemingly simple process requires compliance with complex legal regulations. And it requires considerable care to produce psychometrically reliable, fair, and pedagogically sound results, and when human oversight is sufficiently needed to be maintained.

Further, no production-ready systems exist to function as assistive technology in this process at this moment.

For that reason, in this Companion, we do not elaborate more on this topic apart from some notions of the main legal and psychometric considerations.

p. 9-82

9.1 Some legal considerations

European AI-Act

- AI systems used in education to assess learning outcomes or guide the learning process may fall under the high-risk category (Annex III of the AI Act). The VU must properly substantiate and document this classification;
- It must be clear whether the VU is a provider or a controller. This classification determines which compliance requirements the VU must meet;
- It is important to note that the VU can be both; if faculty members/researchers themselves build a system around an existing model, the VU may be considered a provider for that system;
- Students must be aware that they are receiving AI-generated scores, and the staff involved (faculty members, researchers) must possess sufficient AI literacy;
- The faculty member's review ("human oversight") must be substantive and demonstrably structured—not merely a formality. "Human oversight" means that the decision or grading is ultimately made by a human. How the training is conducted beforehand has no bearing on this. Personal training on the tool cannot replace this, and even sampled control, there is not always a "human oversight."

GDPR - General Data Protection Regulation (Dutch AVG - Algemene Verordening Gegevensbescherming)

- The focus here is on Article 22 of the GDPR: data subjects have the right not to be subject to a decision based solely on automated processing that produces legal effects or significantly affects them.

9.2 Psychometric validity

For tests and exams, the psychometric properties are of the highest importance. Scores on tests and questions must be consistent (reliable) and in accordance with human scores. And there are many

factors at play that determine the reliability, validity, and usefulness of LLMs for grading open-ended exam questions. As mentioned before, based on the current research (we refer to paragraph **Error! Reference source not found.** for some scientific underpinnings), research has shown that LLMs show high correlations and agreement with human scores, but they can only serve as an assistive technology. It cannot serve as a technology operating independently. Edge cases, unclear questions and grading schemes, as well as students around the pass-fail border need careful attention and may in some cases require manual re-grading. Yet, despite these limitations, the research shows that significant gains are feasible by using LLMs.

9.3 Example case

For illustrative purposes, first, we describe some preliminary results from a scientific study of a business exam. The test is comprised of 12 open-ended questions. Each question for each student is graded three times, and then the main, median, or mode value is chosen as the final score. In an analysis, the correlation of the manual scores (given by one rater) and the LLM scores is calculated, and scatterplots are shown. The plot shows some common characteristics you would encounter when doing such an analysis.

9.3.1 Test level

First, we describe typical findings on the level of a test as a whole. In Figure 9.1 Scatterplot of manual scores and LLM scores on a test of 12 open ended questions. you can see how manual scores for one rater and the LLM modes' scores are related. This kind of scatterplot is quite typical.

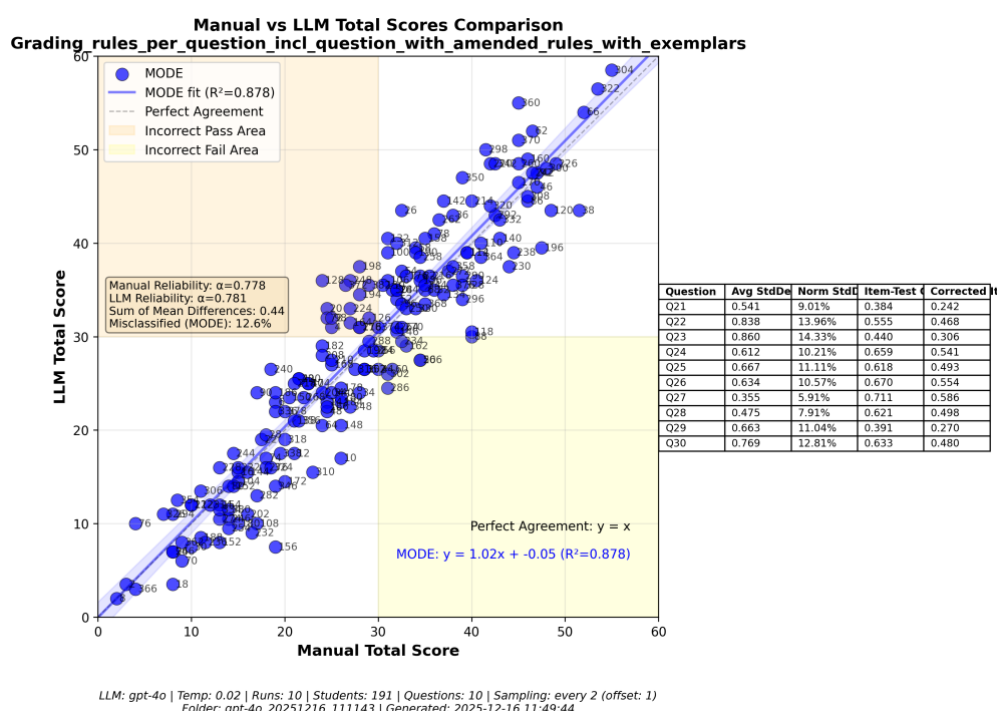


Figure 9.1 Scatterplot of manual scores and LLM scores on a test of 12 open ended questions.

The following can be observed.

- Using a high-performance LLM (in this case GPT-4o), it is possible to achieve a substantive correlation between the LLM score and the human scores. The human grading reliability of the test was 0.78 and the LLM reliability was 0,78 too. This is a good result, as generally the LLM

score is somewhat lower in comparison. This is also comparable to the internal consistency value of multiple graders of exams as is often found in research.

- Further, the scores correlate quite well, and are also in the same magnitude. The true manual score line ($y=x$) is almost exactly followed. The regression line is a bit steeper, indicating that students have a chance of getting a bit higher score if they already score high and vice versa. This is also a bit of a chance plot in this case, as other scientific studies have indicated that it is mostly the other way around: the line is less steep than $y=x$. This is due to regression to the mean (for high scores, the chance is larger to get lower scores when regrading, for low scores, the chances are higher to get higher scores).
- Finally, the plot shows two yellow areas based on the pass-fail score, in which students would be misclassified as compared to the manual scores. This is calculated to be about 13% of the students for this case. This would imply that teachers should manually grade this percentage of the total number of students around the pass-fail grade to make sure these grades have been checked with sufficient human-in-the-loop in the process.

In paragraph **Error! Reference source not found.** we shortly describe some finding in the current literature regarding grading of open ended question by LLMs.

9.3.2 Question level

At the individual question level, the scatterplots show more variations. Below we show some typical encountered plots in which the mean, median and mode scores are plotted.

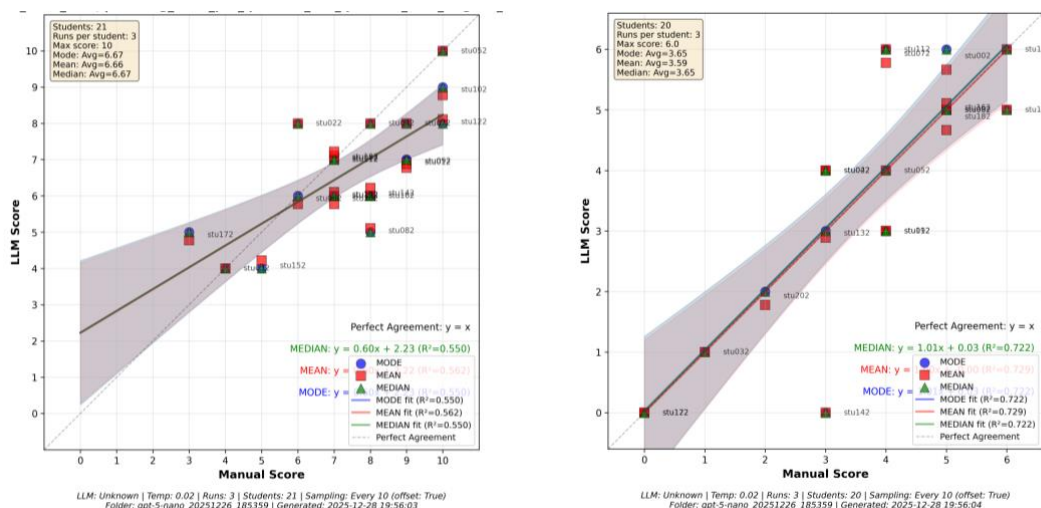


Figure 9.2 Scatterplot of manual scores and LLM scores on two typical questions.

What the plots show is the following.

- The correlations between manual and LLM scores are positive and consistent.
- The pattern of regression to the mean (grade compression) is well visible (less steep curve than $y=x$).
- The chance of getting exactly the same human score as LLM score is limited.
- More capable LLMs show higher consistency (no plot shown).
- Further, the average LLM score for different questions show varied deviations from the manual scores. Sometimes the LLM scores on average higher, sometimes lower.
- When scores are off, a process must be in place in which the poorly correlating questions must all be graded manually.

We have observed that *both* the quality of the question and the quality of the grading scheme are of prime importance for the quality of the scoring of the LLM as is also found in the current research. And that is very logical: both must be unambiguous, preferably must be able to be expressed in countable facts, concept, arguments reasonings etc. In Figure 9.3 we present one such a question.

In contrast, questions asking for broad, creative personal perspectives or free interpretations of answers likely lend themselves less for grading using LLMs. We note however that we have not yet studied such exams.

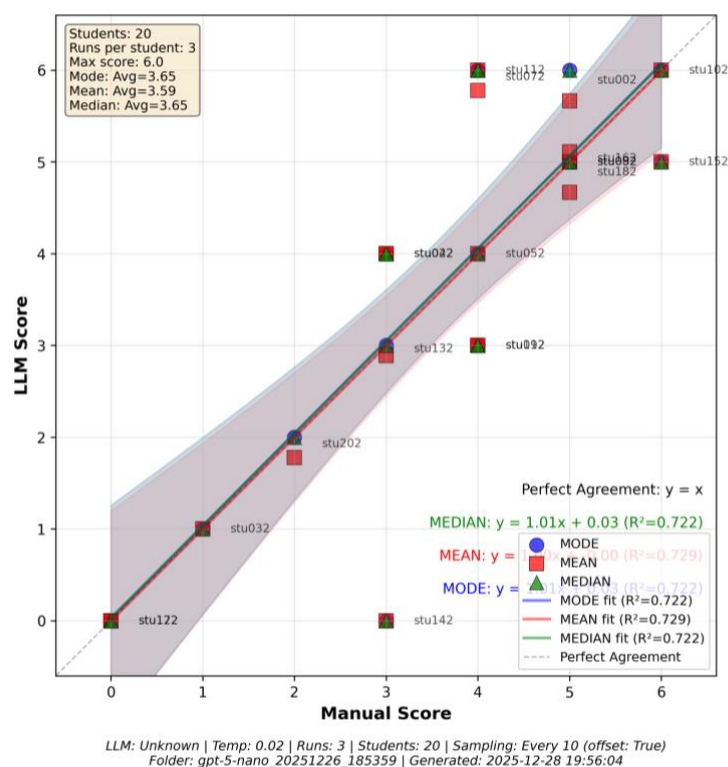


Figure 9.3 Scatterplot of a relatively high correlated manual and LLM scores.

The conclusion is that LLMs can be used for grading open ended question for exams but with caution. Generative AI can only serve as an *assistive technology*. Human oversight remains very important and low correlated manual-LLM questions, edge cases as well as students around the pass-fail border need careful human attention and likely manual grading.

9.3.3 Student acceptance of LLM grading

An important aspect of grading with LLMs for which little research is available and the VU Education Lab has no experience with, is the acceptance of grades by students. Research on the acceptance of AI generated diagnosis for patients with eye problems or health related issues, showed different levels of trust (Shah et al., 2022; Wewetzer et al., 2026). Acceptance will likely depend on the way in which students are introduced and informed about the grading process and the guard rails in place. But also, we expect that differences will occur based on the presentation of results to the students. Research in this area is needed.

10 References

- Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21(1), 10. <https://doi.org/10.1186/s41239-024-00444-7>
- AlGhamdi, E. M., Li, Y., Gašević, D., & Chen, G. (2026). Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. *Computers & Education*, 243, 105511. <https://doi.org/10.1016/j.compedu.2025.105511>
- Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. John Wiley & Sons.
- Austin, T. (2026, May 11). *The Evolution of Bloom's Taxonomy, And Where It Was Always Heading in the Age of AI*. <https://substack.com/inbox>
- Becker, D., Connolly, J., Lentz, P. J., & Morrison, J. (2006). Using the Business Fraud Triangle to Predict Academic Dishonesty among Business Students. *Undefined*. /paper/Using-the-Business-Fraud-Triangle-to-Predict-among-Becker-Connolly/b6fd4e20a10b3577cd0abc2e04a5b6d71054efb8
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bergwerff, M., Dekker, M., Van Blankenstein, F., Romein, D., & Klaren, M. (2015). *Tips bij toetsen: Handreiking voor docenten*. Universiteit Leiden, Faculteit der Geesteswetenschappen en ICLON. https://www.universiteitleiden.nl/binaries/content/assets/science/onderwijs/sterrenkunde/tips_bijtoetsen.pdf
- Biggs, J. B. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364. <https://doi.org/10.1007/BF00138871>
- Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*, 435, 435–459.
- Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4), 475.
- Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. Granada Learning. https://books.google.com/books?hl=en&lr=&id=_hDy-5KOro8C&oi=fnd&pg=PA1&dq=seminal+1998+work+%22Inside+the+Black+Box%22+black+william&ots=sn87uuvQsm&sig=k6zHLvwwGq8lvcfONRF1odTXUqg
- Black, P., & William, D. (1999). *Assessment for learning: Beyond the black box*. Assessment Reform Group, Qualifications and Curriculum Authority.
- Bone, M., González Ehlinger, E., & Stephany, F. (2025). Skills or degree? The rise of skill-based hiring for AI and green jobs. *Technological Forecasting and Social Change*, 214. <https://doi.org/10.1016/j.techfore.2025.124042>

- Chen, A., Xiang, M., Zhou, J., Jia, J., Shang, J., Li, X., Gašević, D., & Fan, Y. (2025). Unpacking help-seeking process through multimodal learning analytics: A comparative study of ChatGPT vs Human expert. *Computers & Education*, 226, 105198. <https://doi.org/10.1016/j.compedu.2024.105198>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Christensen Hughes, J., & Eaton, S. E. (2022). Student integrity violations in the academy: More than a decade of growing complexity and concern. *Academic Integrity in Canada: An Enduring and Essential Challenge*, 61–79.
- Clauser, B. E. (2021). A History of Classical Test Theory. In *The History of Educational Measurement*. Routledge.
- Clune, M. (2025, November 29). *Colleges Are Preparing to Self-Lobotomize*. The Atlantic. <https://www.theatlantic.com/ideas/2025/11/colleges-ai-education-students/685039/>
- Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, 0(0), 1–17. <https://doi.org/10.1080/02602938.2025.2553340>
- Davies, M. (2011). Concept mapping, mind mapping and argument mapping: What are the differences and do they matter? *Higher Education*, 62(3), 279–301. <https://doi.org/10.1007/s10734-010-9387-6>
- Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, 49(7), 1005–1016. <https://doi.org/10.1080/02602938.2024.2386662>
- Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, 105224. <https://doi.org/10.1016/j.compedu.2024.105224>
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251–19257. <https://doi.org/10.1073/pnas.1821936116>
- Digitale Overheid. (2025, February 26). Wat je moet weten over AI-geletterdheid [Achtergrondartikelen]. *Digitale Overheid*. <https://www.digitaleoverheid.nl/achtergrondartikelen/wat-je-moet-weten-over-ai-geletterdheid/>
- Donald, J., Gail. (2002). *Learning to Think: Disciplinary Perspectives*. Wiley. <https://www.wiley.com/en-us/Learning+to+Think%3A+Disciplinary+Perspectives-p-9781118308127>
- Dreschler, G., Gambrel, A., & Branum, J. (2025). A Case Study in Mindfully Integrating AI Tools into Writing Classes. *Journal of Academic Writing*, 15(S2), 1–7.
- Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning* (1st ed.). Corwin press. <https://books.google.com/books?hl=en&lr=&id=zYh2AwAAQBAJ&oi=fnd&pg=PP1&dq=Earl+clearly+articulated+this+typology+in+her+influential+2003+book+%22Assessment+as+Learning:+U>

sing+Classroom+Assessment+to+Maximize+Student+Learning.%22+In&ots=O8W4QeNErN&sig=hRMPx1fjUhHhtw2rHIT_EVWZhHY

Ebel, R. L. (1979). *Essentials of educational measurement*. Prentice-Hall.

Fitzgerald, R. (2025). *Inquiry in Action: Using AI to Reimagine Learning and Teaching: Case Studies from the Frontline of Higher Education Practice*. The University of Queensland.
<https://doi.org/10.14264/a18431b>

Fuchs, M. (2026, February 16). *Hulporganisaties zetten AI-beelden vol stereotypen in om fondsen te werven*. Trouw. <https://www.trouw.nl/buitenland/hulporganisaties-zetten-ai-beelden-vol-stereotypen-in-om-fondsen-te-werven~b5f3a569/>

Gallant, B. (Host). (2025, November 10). *Episode 36: Cath Ellis* (No. 36) [Broadcast].
<https://www.theoppositeofcheating.com/post/episode-36-cath-ellis>

Gelder, T. van. (2005). Teaching Critical Thinking: Some Lessons From Cognitive Science. *College Teaching*, 53(1), 41–48. <https://doi.org/10.3200/CTCH.53.1.41-48>

Gobet, F. (2005). Chunking models of expertise: Implications for education. *Applied Cognitive Psychology*, 19(2), 183–204. <https://doi.org/10.1002/acp.1110>

Hanshaw, G., & Sullivan, C. (2025). Exploring barriers to AI course assistant adoption: A mixed-methods study on student non-utilization. *Discover Artificial Intelligence*, 5, 1–24.
<https://doi.org/10.1007/s44163-025-00312-x>

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
<https://doi.org/10.3102/003465430298487>

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 300–304.
<https://doi.org/10.1145/3657604.3664693>

Hill, K. (2025, May 14). College Professors Are Using ChatGPT. Some Students Aren't Happy. *The New York Times*. <https://www.nytimes.com/2025/05/14/technology/chatgpt-college-professors.html>

Huxham, M., Campbell, F., & Westwood, J. (2012). Oral versus written assessments: A test of student performance and attitudes. *Assessment & Evaluation in Higher Education*, 37(1), 125–136.
<https://doi.org/10.1080/02602938.2010.515012>

Jemstedt, A. (2024). Enhancing learning with a two-page study manual. *Learning and Instruction*, 90, 101852. <https://doi.org/10.1016/j.learninstruc.2023.101852>

Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports*, 14(1), 30064.

Joint Quality Initiative. (2004). *Shared “Dublin” descriptors for Short Cycle, First Cycle, Second Cycle and Third Cycle Awards*. <http://www.jointquality.nl>

Jones, A. (2015). A Disciplined Approach to Critical Thinking. In M. Davies & R. Barnett (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (pp. 169–182). Palgrave Macmillan US. https://doi.org/10.1057/9781137378057_11

- Jones, L. R. (2011). *Academic integrity & academic dishonesty: A handbook about cheating & plagiarism*. https://repository.fit.edu/library_publications/1/
- Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education*, 23(4), 367–378. <https://doi.org/10.1080/0260293980230404>
- Joughin, G. (2010). *A short guide to oral assessment*. Leeds Metropolitan University in association with University of Wollongong. https://www.researchgate.net/publication/228844594_A_Short_Guide_to_Oral_Assessment
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Karlen, Y., Hirt, C. N., Jud, J., Rosenthal, A., & Eberli, T. D. (2023). Teachers as learners and agents of self-regulated learning: The importance of different teachers competence aspects for promoting metacognition. *Teaching and Teacher Education*, 125, 104055. <https://doi.org/10.1016/j.tate.2023.104055>
- Kassorla, M. (2025, May 31). Inverted Bloom’s for the Age of AI [Substack newsletter]. *THE ACADEMIC PLATYPUS*. <https://michellekassorla.substack.com/p/inverted-blooms-for-the-age-of-ai>
- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2024). *AI Tutoring Outperforms Active Learning*. Research Square. <https://doi.org/10.21203/rs.3.rs-4243877/v1>
- King’s College London. (2025, April). *King’s guidance on generative AI for teaching, assessment and feedback*. King’s College London. <https://www.kcl.ac.uk/about/strategy/learning-and-teaching/ai-guidance>
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1
- Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349.
- Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 187.
- Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners’ sensitivity to retrieval conditions at test. *Memory & Cognition*, 34(5), 959–972. <https://doi.org/10.3758/BF03193244>
- Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, 4(1), 47. <https://doi.org/10.1007/s44163-024-00147-y>
- Korthals, L., Akrong, E., Geller, G., Rosenbusch, H., Grasman, R., & Visser, I. (2026). Towards Reliable LLM Grading Through Self-Consistency and Selective Human Review: Higher Accuracy, Less Work. *Machine Learning and Knowledge Extraction*, 8(3), 74. <https://doi.org/10.3390/make8030074>
- Korthals, L., Rosenbusch, H., Grasman, R., & Visser, I. (2025). Grading University Students with LLMs: Performance and Acceptance of a Canvas-Based Automation. In A. I. Cristea, E. Walker, Y. Lu, O.

C. Santos, & S. Isotani (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED* (Vol. 2591, pp. 36–43). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99264-3_5

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

Kusumegi, K., Yang, X., Ginsparg, P., de Vaan, M., Stuart, T., & Yin, Y. (2025). Scientific production in the era of large language models. *Science*, 390(6779), 1240–1243. <https://doi.org/10.1126/science.adw3000>

Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *Npj Science of Learning*, 10(1), 29. <https://doi.org/10.1038/s41539-025-00320-7>

Li, C., Cui, H., & Hagedorn, L. S. (2026). The cognitive impact of ChatGPT in higher education: A systematic review of critical and creative thinking outcomes. *Computers and Education: Artificial Intelligence*, 10, 100571. <https://doi.org/10.1016/j.caeai.2026.100571>

Liu, D. (2025, September 24). *Assessment & The Two-Lane Approach by Prof. Danny Liu* [Video recording]. <https://video.edu.nl/w/sifbfxmBYbJbgkaPc6wq9e>

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

Lodge, J. M., & Loble, L. (2026). *Artificial intelligence, cognitive offloading and implications for education* [Report]. University of Technology Sydney. (Australia). <https://apo.org.au/node/333716>

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054.

Nesbit, J. C., & Adesope, O. O. (2006). Learning With Concept and Knowledge Maps: A Meta-Analysis. *Review of Educational Research*, 76(3), 413–448. <https://doi.org/10.3102/00346543076003413>

Neumann, R. (2001). Disciplinary Differences and University Teaching. *Studies in Higher Education*. (world). <https://doi.org/10.1080/03075070120052071>

Nicol, D. (2022). “Turning Active Learning into Active Feedback”, *Introductory Guide from Active Feedback Toolkit*.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199. <https://doi.org/10.1080/03075070600572090>

Nicol, D., & Kushwah, L. (2024). Shifting feedback agency to students by having them write their own feedback comments. *Assessment & Evaluation in Higher Education*, 49(3), 419–439. <https://doi.org/10.1080/02602938.2023.2265080>

- Oakley, B., Johnston, M., Chen, K.-Z., Jung, E., & Sejnowski, T. J. (2025). *The Memory Paradox: Why Our Brains Need Knowledge in an Age of AI* (arXiv:2506.11015). arXiv. <https://doi.org/10.48550/arXiv.2506.11015>
- Paul, R., & Elder, L. (2019). *The miniature guide to critical thinking concepts and tools*. Rowman & Littlefield. https://books.google.com/books?hl=en&lr=&id=Cm6pDwAAQBAJ&oi=fnd&pg=PA4&dq=Thinker%27s+Guide+Library+-+The+Miniature+Guide+to+Critical+Thinking+Concepts+and+Tools&ots=Yd6H9kDl-K&sig=j2E2Ji7V3Rv_1xdsXq8kxTD41ql
- Pearson, H. (2025). Universities are embracing AI: Will students get smarter or stop thinking? *Nature*, 646(8086), 788–791. <https://doi.org/10.1038/d41586-025-03340-w>
- Pedro, door. (2017, April 19). Een wat langer stuk over de taxonomie van Bloom. *Blogcollectief Onderzoek Onderwijs*. <https://onderzoekonderwijs.net/2017/04/19/een-wat-langer-stuk-over-de-taxonomie-van-bloom/>
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment. *Journal of University Teaching and Learning Practice*, 21(06). <https://doi.org/10.53761/q3azde36>
- Perring, G. (2025, October 7). *Embracing AI in international higher education*. European Association for International Education. <https://www.eaie.org/resource/embracing-ai-in-international-higher-education-transforming-learning-supporting-diversity-and-preparing-for-the-future.html>
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2), 155–169. <https://doi.org/10.1007/BF01405730>
- Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2023). Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education*, 28(7), 1565–1582. <https://doi.org/10.1080/13562517.2021.1913723>
- Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in Cognition: Extending Human Intelligence with Intelligent Technologies. *Educational Researcher*, 20(3), 2–9. <https://doi.org/10.3102/0013189X020003002>
- Shah, P., Mishra, D., Shanmugam, M., Vighnesh, M. J., & Jayaraj, H. (2022). Acceptability of artificial intelligence-based retina screening in general population. *Indian Journal of Ophthalmology*, 70(4), 1140. https://doi.org/10.4103/ijo.IJO_1840_21
- Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education*, 45(11), 2132–2148. <https://doi.org/10.1080/03075079.2019.1582015>
- Stasuik, N. C. (2025, April 28). *Evaluating LLM Performance in Essay Assessment: A Comparative Analysis of AI Grading and Feedback Systems for University English Courses*. <https://doi.org/10.14288/1.0448868>
- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1), 65.

- Stephenson, Z., Johnson-Glauch, N., & Cruchley, S. (2025). Interventions and facilitators of oral assessment performance in higher education: A systematic review. *Assessment & Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2025.2504621>
- Theobald, A. S. (2021). Oral exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education*, 29(2), 156–159. <https://doi.org/10.1080/26939169.2021.1914527>
- TLC University of Amsterdam. (2025). *AI Literacy for teachers*. Articulate. <https://share.articulate.com/RoMQC5qv9KmeRmXFLQtsd>
- van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (Eds.). (2014). *Toetsen in het hoger onderwijs*. Bohn Stafleu van Loghum. <https://doi.org/10.1007/978-90-368-0239-0>
- Van Ginkel, S., Gulikers, J., Biemans, H., & Mulder, M. (2015). Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. *Educational Research Review*, 14, 62–80. <https://doi.org/10.1016/j.edurev.2015.02.002>
- Van Joolingen, W. R., De Jong, T., & Dimitrakopoulou, A. (2007). Issues in computer supported inquiry learning in science. *Journal of Computer Assisted Learning*, 23(2), 111–119.
- Van Merriënboer, J. J. G., & Kirschner, P. A. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315113210>
- Vermunt, J. D. H. M., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction*, 9, 257–280.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Watkins, M. (2026, March 12). OpenAI's Education Pitch Has a Free Version Problem [Substack newsletter]. *Rhetorica*. https://marcwatkins.substack.com/p/openais-education-pitch-has-a-free?utm_medium=android&triedRedirect=true
- Wewetzer, L., Goetz, K., Freischmidt, S., & Steinhäuser, J. (2026). Trust in AI-Supported Screening in General Practice Among Urban and Rural Citizens: Cross-Sectional Study. *JMIR Medical Informatics*, 14(1), e69777. <https://doi.org/10.2196/69777>
- William, D. (2017, April 4). @Doug_Lemov @DavidDidau I agree. My two cents' worth, with knowledge as the foundation on which all the others are built... <https://t.co/xjyFzs64Pl> [Tweet]. Twitter. <https://x.com/dylanwilliam/status/849277989577789441>
- Winstone, N. E., Gravett, K., Bearman, M., Noble, C., Jensen, L. X., Jones, A., & Nicola-Richmond, K. (2026). The care-full craft of feedback in an age of generative AI. *Assessment & Evaluation in Higher Education*, 0(0), 1–17. <https://doi.org/10.1080/02602938.2026.2643333>
- Yavariabdi, A., Paudel, B., Carleton, T., & De Almeida, C. D. A. (2025). Generative AI in Assessment and Feedback Generation in Higher Education: A Systematic Review. *2025 International Conference on Education Technology and Computers (ICETC)*, 361–371. <https://doi.org/10.1109/ICETC66579.2025.11387416>
- Yerkes, R. M. (1921). *Memoirs of the National Academy of Sciences*. US Government Printing Office.

- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39.
<https://doi.org/10.1186/s41239-019-0171-0>
- Zhang, S., Zhao, X., Zhou, T., & Kim, J. H. (2024). Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. *International Journal of Educational Technology in Higher Education*, 21(1), 34.
<https://doi.org/10.1186/s41239-024-00467-0>
- AlGhamdi, E. M., Li, Y., Gašević, D., & Chen, G. (2026). Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. *Computers & Education*, 243, 105511.
<https://doi.org/10.1016/j.compedu.2025.105511>
- Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. John Wiley & Sons.
- Austin, T. (2026, May 11). *The Evolution of Bloom's Taxonomy, And Where It Was Always Heading in the Age of AI*. <https://substack.com/inbox>
- Becker, D., Connolly, J., Lentz, P. J., & Morrison, J. (2006). Using the Business Fraud Triangle to Predict Academic Dishonesty among Business Students. *Undefined*. /paper/Using-the-Business-Fraud-Triangle-to-Predict-among-Becker-Connolly/b6fd4e20a10b3577cd0abc2e04a5b6d71054efb8
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623.
<https://doi.org/10.1145/3442188.3445922>
- Bergwerff, M., Dekker, M., Van Blankenstein, F., Romein, D., & Klaren, M. (2015). *Tips bij toetsen: Handreiking voor docenten*. Universiteit Leiden, Faculteit der Geesteswetenschappen en ICLON.
https://www.universiteitleiden.nl/binaries/content/assets/science/onderwijs/sterrenkunde/tips_bijtoetsen.pdf
- Biggs, J. B. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364.
<https://doi.org/10.1007/BF00138871>
- Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*, 435, 435–459.
- Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4), 475.
- Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. Granada Learning. https://books.google.com/books?hl=en&lr=&id=_hDy-5KOro8C&oi=fnd&pg=PA1&dq=seminal+1998+work+%22Inside+the+Black+Box%22+black+william&ots=sn87uuvQsm&sig=k6zHLvwvGq8lvcfONRF1odTXUqg

- Black, P., & William, D. (1999). *Assessment for learning: Beyond the black box*. Assessment Reform Group, Qualifications and Curriculum Authority.
- Chen, A., Xiang, M., Zhou, J., Jia, J., Shang, J., Li, X., Gašević, D., & Fan, Y. (2025). Unpacking help-seeking process through multimodal learning analytics: A comparative study of ChatGPT vs Human expert. *Computers & Education*, 226, 105198. <https://doi.org/10.1016/j.compedu.2024.105198>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Christensen Hughes, J., & Eaton, S. E. (2022). Student integrity violations in the academy: More than a decade of growing complexity and concern. *Academic Integrity in Canada: An Enduring and Essential Challenge*, 61–79.
- Clauser, B. E. (2021). A History of Classical Test Theory. In *The History of Educational Measurement*. Routledge.
- Clune, M. (2025, November 29). *Colleges Are Preparing to Self-Lobotomize*. The Atlantic. <https://www.theatlantic.com/ideas/2025/11/colleges-ai-education-students/685039/>
- Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, 0(0), 1–17. <https://doi.org/10.1080/02602938.2025.2553340>
- Davies, M. (2011). Concept mapping, mind mapping and argument mapping: What are the differences and do they matter? *Higher Education*, 62(3), 279–301. <https://doi.org/10.1007/s10734-010-9387-6>
- Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, 49(7), 1005–1016. <https://doi.org/10.1080/02602938.2024.2386662>
- Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, 105224. <https://doi.org/10.1016/j.compedu.2024.105224>
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251–19257. <https://doi.org/10.1073/pnas.1821936116>
- Digitale Overheid. (2025, February 26). Wat je moet weten over AI-geletterdheid [Achtergrondartikelen]. *Digitale Overheid*. <https://www.digitaleoverheid.nl/achtergrondartikelen/wat-je-moet-weten-over-ai-geletterdheid/>
- Donald, J., & Gail. (2002). *Learning to Think: Disciplinary Perspectives*. Wiley. <https://www.wiley.com/en-us/Learning+to+Think%3A+Disciplinary+Perspectives-p-9781118308127>
- Dreschler, G., Gambrel, A., & Branum, J. (2025). A Case Study in Mindfully Integrating AI Tools into Writing Classes. *Journal of Academic Writing*, 15(S2), 1–7.

- Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning* (1st ed.). Corwin press.
https://books.google.com/books?hl=en&lr=&id=zYh2AwAAQBAJ&oi=fnd&pg=PP1&dq=Earl+clearly+articulated+this+typology+in+her+influential+2003+book+%22Assessment+as+Learning:+Using+Classroom+Assessment+to+Maximize+Student+Learning.%22+In&ots=O8W4QeNErN&sig=hRMPx1fjUhHhtw2rHt_EVWZhHY
- Ebel, R. L. (1979). *Essentials of educational measurement*. Prentice-Hall.
- Fitzgerald, R. (2025). *Inquiry in Action: Using AI to Reimagine Learning and Teaching: Case Studies from the Frontline of Higher Education Practice*. The University of Queensland.
<https://doi.org/10.14264/a18431b>
- Fuchs, M. (2026, February 16). *Hulporganisaties zetten AI-beelden vol stereotypen in om fondsen te werven*. Trouw. <https://www.trouw.nl/buitenland/hulporganisaties-zetten-ai-beelden-vol-stereotypen-in-om-fondsen-te-werven~b5f3a569/>
- Gallant, B. (Host). (2025, November 10). *Episode 36: Cath Ellis* (No. 36) [Broadcast].
<https://www.theoppositeofcheating.com/post/episode-36-cath-ellis>
- Gelder, T. van. (2005). Teaching Critical Thinking: Some Lessons From Cognitive Science. *College Teaching*, 53(1), 41–48. <https://doi.org/10.3200/CTCH.53.1.41-48>
- Gobet, F. (2005). Chunking models of expertise: Implications for education. *Applied Cognitive Psychology*, 19(2), 183–204. <https://doi.org/10.1002/acp.1110>
- Hanshaw, G., & Sullivan, C. (2025). Exploring barriers to AI course assistant adoption: A mixed-methods study on student non-utilization. *Discover Artificial Intelligence*, 5, 1–24.
<https://doi.org/10.1007/s44163-025-00312-x>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
<https://doi.org/10.3102/003465430298487>
- Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 300–304.
<https://doi.org/10.1145/3657604.3664693>
- Hill, K. (2025, May 14). College Professors Are Using ChatGPT. Some Students Aren't Happy. *The New York Times*. <https://www.nytimes.com/2025/05/14/technology/chatgpt-college-professors.html>
- Holtrop, D., Dunlop, P. D., Oostrom, J. K., & Bourdage, J. S. (2024). The Effect of Applicants' Use of Generative AI on Recruiter Ratings of Cover Letters. *Biannual Meeting of the European Network of Selection Researchers (ENESER), Brussels, Belgium*.
- Huxham, M., Campbell, F., & Westwood, J. (2012). Oral versus written assessments: A test of student performance and attitudes. *Assessment & Evaluation in Higher Education*, 37(1), 125–136.
<https://doi.org/10.1080/02602938.2010.515012>
- Jemstedt, A. (2024). Enhancing learning with a two-page study manual. *Learning and Instruction*, 90, 101852. <https://doi.org/10.1016/j.learninstruc.2023.101852>

- Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports*, 14(1), 30064.
- Jones, A. (2015). A Disciplined Approach to Critical Thinking. In M. Davies & R. Barnett (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (pp. 169–182). Palgrave Macmillan US. https://doi.org/10.1057/9781137378057_11
- Jones, L. R. (2011). *Academic integrity & academic dishonesty: A handbook about cheating & plagiarism*. https://repository.fit.edu/library_publications/1/
- Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education*, 23(4), 367–378. <https://doi.org/10.1080/0260293980230404>
- Joughin, G. (2010). *A short guide to oral assessment*. Leeds Metropolitan University in association with University of Wollongong. https://www.researchgate.net/publication/228844594_A_Short_Guide_to_Oral_Assessment
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Karlen, Y., Hirt, C. N., Jud, J., Rosenthal, A., & Eberli, T. D. (2023). Teachers as learners and agents of self-regulated learning: The importance of different teachers competence aspects for promoting metacognition. *Teaching and Teacher Education*, 125, 104055. <https://doi.org/10.1016/j.tate.2023.104055>
- Kassorla, M. (2025, May 31). Inverted Bloom’s for the Age of AI [Substack newsletter]. *THE ACADEMIC PLATYPUS*. <https://michellekassorla.substack.com/p/inverted-blooms-for-the-age-of-ai>
- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2024). *AI Tutoring Outperforms Active Learning*. Research Square. <https://doi.org/10.21203/rs.3.rs-4243877/v1>
- King’s College London. (2025, April). *King’s guidance on generative AI for teaching, assessment and feedback*. King’s College London. <https://www.kcl.ac.uk/about/strategy/learning-and-teaching/ai-guidance>
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1
- Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349.
- Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 187.
- Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners’ sensitivity to retrieval conditions at test. *Memory & Cognition*, 34(5), 959–972. <https://doi.org/10.3758/BF03193244>
- Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, 4(1), 47. <https://doi.org/10.1007/s44163-024-00147-y>

- Korthals, L., Akrong, E., Geller, G., Rosenbusch, H., Grasman, R., & Visser, I. (2026). Towards Reliable LLM Grading Through Self-Consistency and Selective Human Review: Higher Accuracy, Less Work. *Machine Learning and Knowledge Extraction*, 8(3), 74. <https://doi.org/10.3390/make8030074>
- Korthals, L., Rosenbusch, H., Grasman, R., & Visser, I. (2025). Grading University Students with LLMs: Performance and Acceptance of a Canvas-Based Automation. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED* (Vol. 2591, pp. 36–43). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99264-3_5
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.
- Kusumegi, K., Yang, X., Ginsparg, P., de Vaan, M., Stuart, T., & Yin, Y. (2025). Scientific production in the era of large language models. *Science*, 390(6779), 1240–1243. <https://doi.org/10.1126/science.adw3000>
- Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *Npj Science of Learning*, 10(1), 29. <https://doi.org/10.1038/s41539-025-00320-7>
- Li, C., Cui, H., & Hagedorn, L. S. (2026). The cognitive impact of ChatGPT in higher education: A systematic review of critical and creative thinking outcomes. *Computers and Education: Artificial Intelligence*, 10, 100571. <https://doi.org/10.1016/j.caeai.2026.100571>
- Liu, D. (2025, September 24). *Assessment & The Two-Lane Approach by Prof. Danny Liu* [Video recording]. <https://video.edu.nl/w/sifbfxmBYbJbgkaPc6wq9e>
- Liu, D., & Bridgeman, A. (2023, July 12). *What to do about assessments if we can't out-design or out-run AI? – Teaching@Sydney*. <https://educational-innovation.sydney.edu.au/teaching@sydney/what-to-do-about-assessments-if-we-cant-out-design-or-out-run-ai/>
- Lodge, J. M., & Loble, L. (2026). *Artificial intelligence, cognitive offloading and implications for education* [Report]. University of Technology Sydney. (Australia). <https://apo.org.au/node/333716>
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054.
- Nesbit, J. C., & Adesope, O. O. (2006). Learning With Concept and Knowledge Maps: A Meta-Analysis. *Review of Educational Research*, 76(3), 413–448. <https://doi.org/10.3102/00346543076003413>
- Neumann, R. (2001). Disciplinary Differences and University Teaching. *Studies in Higher Education*. (world). <https://doi.org/10.1080/03075070120052071>
- Nicol, D. (2022). “Turning Active Learning into Active Feedback”, *Introductory Guide from Active Feedback Toolkit*.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199. <https://doi.org/10.1080/03075070600572090>

- Nicol, D., & Kushwah, L. (2024). Shifting feedback agency to students by having them write their own feedback comments. *Assessment & Evaluation in Higher Education*, 49(3), 419–439. <https://doi.org/10.1080/02602938.2023.2265080>
- Oakley, B., Johnston, M., Chen, K.-Z., Jung, E., & Sejnowski, T. J. (2025). *The Memory Paradox: Why Our Brains Need Knowledge in an Age of AI* (arXiv:2506.11015). arXiv. <https://doi.org/10.48550/arXiv.2506.11015>
- Paul, R., & Elder, L. (2019). *The miniature guide to critical thinking concepts and tools*. Rowman & Littlefield. https://books.google.com/books?hl=en&lr=&id=Cm6pDwAAQBAJ&oi=fnd&pg=PA4&dq=Thinker%27s+Guide+Library+-+The+Miniature+Guide+to+Critical+Thinking+Concepts+and+Tools&ots=Yd6H9kDl-K&sig=j2E2Ji7V3Rv_1xdsXq8kxTD41ql
- Pearson, H. (2025). Universities are embracing AI: Will students get smarter or stop thinking? *Nature*, 646(8086), 788–791. <https://doi.org/10.1038/d41586-025-03340-w>
- Pedro, door. (2017, April 19). Een wat langer stuk over de taxonomie van Bloom. *Blogcollectief Onderzoek Onderwijs*. <https://onderzoekonderwijs.net/2017/04/19/een-wat-langer-stuk-over-de-taxonomie-van-bloom/>
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment. *Journal of University Teaching and Learning Practice*, 21(06). <https://doi.org/10.53761/q3azde36>
- Perring, G. (2025, October 7). *Embracing AI in international higher education*. European Association for International Education. <https://www.eaie.org/resource/embracing-ai-in-international-higher-education-transforming-learning-supporting-diversity-and-preparing-for-the-future.html>
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2), 155–169. <https://doi.org/10.1007/BF01405730>
- Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2023). Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education*, 28(7), 1565–1582. <https://doi.org/10.1080/13562517.2021.1913723>
- Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in Cognition: Extending Human Intelligence with Intelligent Technologies. *Educational Researcher*, 20(3), 2–9. <https://doi.org/10.3102/0013189X020003002>
- Shah, P., Mishra, D., Shanmugam, M., Vighnesh, M. J., & Jayaraj, H. (2022). Acceptability of artificial intelligence-based retina screening in general population. *Indian Journal of Ophthalmology*, 70(4), 1140. https://doi.org/10.4103/ijo.IJO_1840_21
- Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education*, 45(11), 2132–2148. <https://doi.org/10.1080/03075079.2019.1582015>
- Stasuik, N. C. (2025, April 28). *Evaluating LLM Performance in Essay Assessment: A Comparative Analysis of AI Grading and Feedback Systems for University English Courses*. <https://doi.org/10.14288/1.0448868>

- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1), 65.
- Stephenson, Z., Johnson-Glauch, N., & Cruchley, S. (2025). Interventions and facilitators of oral assessment performance in higher education: A systematic review. *Assessment & Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2025.2504621>
- Theobald, A. S. (2021). Oral exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education*, 29(2), 156–159. <https://doi.org/10.1080/26939169.2021.1914527>
- TLC University of Amsterdam. (2025). *AI Literacy for teachers*. Articulate. <https://share.articulate.com/RoMQC5qv9KmeRmXFLQtsd>
- van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (Eds.). (2014). *Toetsen in het hoger onderwijs*. Bohn Stafleu van Loghum. <https://doi.org/10.1007/978-90-368-0239-0>
- Van Ginkel, S., Gulikers, J., Biemans, H., & Mulder, M. (2015). Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. *Educational Research Review*, 14, 62–80. <https://doi.org/10.1016/j.edurev.2015.02.002>
- Van Joolingen, W. R., De Jong, T., & Dimitrakopoulou, A. (2007). Issues in computer supported inquiry learning in science. *Journal of Computer Assisted Learning*, 23(2), 111–119.
- Van Merriënboer, J. J. G., & Kirschner, P. A. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315113210>
- Vermunt, J. D. H. M., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction*, 9, 257–280.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Watkins, M. (2026, March 12). OpenAI's Education Pitch Has a Free Version Problem [Substack newsletter]. *Rhetorica*. https://marcwatkins.substack.com/p/openais-education-pitch-has-a-free?utm_medium=android&triedRedirect=true
- Wewetzer, L., Goetz, K., Freischmidt, S., & Steinhauser, J. (2026). Trust in AI-Supported Screening in General Practice Among Urban and Rural Citizens: Cross-Sectional Study. *JMIR Medical Informatics*, 14(1), e69777. <https://doi.org/10.2196/69777>
- Wiliam, D. (2017, April 4). @Doug_Lemov @DavidDidau I agree. My two cents' worth, with knowledge as the foundation on which all the others are built... <https://t.co/xjyFzs64Pl> [Tweet]. Twitter. <https://x.com/dylanwiliam/status/849277989577789441>
- Winstone, N. E., Gravett, K., Bearman, M., Noble, C., Jensen, L. X., Jones, A., & Nicola-Richmond, K. (2026). The care-full craft of feedback in an age of generative AI. *Assessment & Evaluation in Higher Education*, 0(0), 1–17. <https://doi.org/10.1080/02602938.2026.2643333>
- Yavariabdi, A., Paudel, B., Carleton, T., & De Almeida, C. D. A. (2025). Generative AI in Assessment and Feedback Generation in Higher Education: A Systematic Review. *2025 International Conference*

on *Education Technology and Computers (ICETC)*, 361–371.
<https://doi.org/10.1109/ICETC66579.2025.11387416>

Yerkes, R. M. (1921). *Memoirs of the National Academy of Sciences*. US Government Printing Office.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39.
<https://doi.org/10.1186/s41239-019-0171-0>

Zhang, S., Zhao, X., Zhou, T., & Kim, J. H. (2024). Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. *International Journal of Educational Technology in Higher Education*, 21(1), 34.
<https://doi.org/10.1186/s41239-024-00467-0>