Date of creation: February 2026
Last update: May 2026

Assessment methods & AI

Universiteit Leiden

Special Interest Group SIG

Created by: LLInC - special thanks to Ella Akin, Zohre Mohammadi, Maisy Charles & Anthea Aerts
Graphic design: Bing Tang

Contact: AIED@leidenuniv.nl

LLInC

Why is this toolkit created?

The rapid spread of generative AI has challenged traditional ideas about what summative assessments measure and how confidently teachers can interpret student performance. The toolkit is not meant to help choose the "best" assessment method for a course. That decision should already be made through constructive alignment of learning objectives, teaching activities, and assessment. Instead, the toolkit helps educators reflect on whether an existing assessment can be made AI-resilient, and under which conditions this is possible.

What is needed besides this toolkit?

This toolkit does not work on its own. To use it well, educators need professional judgement, and check for justifiability and feasibility, including workload, scalability, and practical constraints. Legal, ethical, logistical, and institutional factors. These remain the responsibility of programmes and teaching teams. Users are encouraged use the accompanying e-learning module, which offers didactical guidance on the pedagogical implications of AI use.

How is this toolkit organized?

This toolkit was developed by reviewing commonly used summative assessment methods in higher education and examining how these methods interact with generative AI in current teaching. Each assessment method was analysed in terms of its usual strengths and weaknesses for validity and reliability, and the ways in which AI can influence assessment outcomes depending on design, context, and implementation called AI positioning. It also provides optional solutions to help reduce assessment method factors, AI positioning and related risks. This toolkit makes the role of AI in assessment design explicit by linking AI use directly to assessment quality, specifically validity (are we measuring the intended learning?) and reliability (are assessment conditions and judgements consistent across students and contexts?). The central idea behind the toolkit is that AI risk is not a fixed property of an assessment method. Instead, the impact of AI on validity and reliability depends largely on assessment design, the assessment context, and teachers' professional knowledge and judgement. At the end of this document an explanation on how to interpret the risk level and AI positioning related to each assessment method is included.

The assessment methods included in this toolkit are:

- Oral exams
- Oral defence
- Presentations
- Capstone projects
- Peer review
- E-portfolios
- Written assignments

Next step?

We recognise that no toolkit can address every situation, and that meaningful assessment design depends on local contexts, disciplinary practices, and professional judgement. We therefore welcome feedback, questions, and shared experiences from users of the toolkit, as these will help us improve and refine it over time. This document is under active maintenance and will be updated regularly in response to developments in AI, assessment practice, and educational policy.

Contact email: AIED@leidenuniv.nl

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
Oral exams  	• Excellent at assessing application of knowledge to authentic scenarios or problems. • Suitable for assessing professional competencies, language skills, and situational judgement. • Allows examiners to redirect or vary difficulty in real time. • Efficient for probing content knowledge and practical skill at the same time. • Strong face validity in professional and vocational contexts.	• Labor and time intensive: difficult to scale for large cohorts. • Risk of outside factors influencing the assessment, such as student anxiety or language barriers. • Subjectivity and examiner bias • Not suited for assessing knowledge across many topics.	Inherently AI-resilient (design dependent) AI exposure: Low – AI cannot be accessed during live interaction; preparation may draw on AI, but the application in the moment, especially to new scenarios, requires students demonstrating their knowledge and skills. AI impact: Low to moderate – AI-assisted preparation has little effect when understanding needs to be demonstrated in real time. Impact rises if the scenario type is predictable and preparation materials closely match the assessment task.	**Assessment** • Record the oral exam with permission of the student, so that a second assessor doesn't need to be present in the moment. Hold calibration sessions before and after to reduce inter-rater variation. • Combine oral exams with another type of assessment to reduce student anxiety and increase validity (less dependent on 'the' moment). • Develop a large and varied scenario bank and rotate scenarios between cohorts to reduce leakage. **AI exposure** • Accept AI-assisted preparation for general case types as non-problematic, but ensure the exam demands situational judgement that preparation alone cannot cover. • Design scenarios that require integration of course-specific content, local context, or the student's own prior work (e.g. linking to their own contribution in a workgroup or lecture). **AI impact** • Students are encouraged to create and asked to bring notes, diagrams, or their own prior work to show their process. • Require students to justify their choices from multiple disciplinary angles rather than arriving at a single correct answer. • Incorporate follow-up questions that require students to adapt their reasoning to a changed condition mid-scenario.

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies

Oral defence
 
 	
	• Excellent at assessing how students think on their feet, adapt solutions, and defend reasoning.
	• Allows examiners to probe depth, ask for clarification, test alternative approaches.
	• High construct validity when targeting argumentation and critical thinking.
	• Immediate follow-up prevents rehearsed or superficial answers.
		• Labor and time intensive: difficult to scale for large cohorts.
		• Risk of outside factors influencing the assessment, such as student anxiety or language barriers.
		• Subjectivity and examiner bias
		• Students who have their defence at a later stage may benefit from other students' experience for their preparation.
			Inherently AI-resilient (design dependent)
			
			AI exposure: Low – AI cannot be accessed during live interaction; preparation with AI is limited by the need to respond spontaneously.
			
			AI impact: Low to moderate– AI-assisted preparation can help students prepare questions and answers based on the product they are defending. However, this has little effect when questions are unpredictable and are focused on reasoning and argumentation. Impact rises when the same type of questions is asked to each student.
				Assessment
				• Record the oral exam with permission of the student, so that a second assessor doesn't need to be present in the moment. Hold calibration sessions before and after to reduce inter-rater variation.
				• Develop structured rubrics that define what good reasoning looks like at each performance level.
				• Offer students a (student-led) practice session or mock defence to reduce anxiety-related distortion.
				
				AI exposure
				• Students can be asked to submit a recorded mini talk on a project or solution explaining:
				  • decisions made.
				  • trade-offs.
				  • what they would change.
				This may also lower the stress level and allows the examiner to inquire about discrepancies.
				• Ask students to trace how their argument developed over time, anchoring responses to personal learning history that AI cannot replicate.
				
				AI impact
				• Require students to connect every answer explicitly to their own work, process, or stated reasoning.
				• Avoid predictable question banks by rotating and personalising questions per student(group).
				• Prepare counterarguments to the students' reasoning in their initial work and ask students to respond to them.

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
Presentations  	• Strong in communicating solutions and arguments clearly. • Good for assessing how well students convey ideas / justify decisions. • Offers insight into digital artefact use (slides, media). • A perfectly aligned assessment method for project-based learning • AI resilient when it requires real time reasoning.	• Less strong in assessing long-term planning, deep iteration of solutions. Final performance may mask weaker underlying work unless process is also assessed. • May privilege students with higher verbal skills rather than those strong at underlying solution work. • Students may over-rehearse AI-generated scripts. • Charisma may be rewarded over accuracy (examiner bias may interfere) • Group presentations often mask unequal contribution and risk individual accountability	Conditionally AI-resilient (depends on probing and individual accountability) AI exposure: Medium–High: AI can easily support scripting, slide creation, visuals, and argument framing. AI impact: Moderate–High: AI can substantially improve surface performance without guaranteeing underlying understanding.	**Assessment** • Consider recording the presentations with permission of the student, so that a second assessor can be included to reduce bias. Hold calibration sessions before and after to reduce inter-rater variation. • Reduce the influence of public-speaking anxiety on content marks by making delivery and content separate, independently weighted criteria in a rubric. • Combine presentations with another type of assessment to reduce student anxiety and increase validity (less dependent on 'the' moment). **AI exposure** • Normalise AI transparency by requiring students to document their use and model this practice yourself. Frame and treat this as a professional habit of transparency, not a surveillance or punitive measure. • Require iterative drafts or a planning document as part of the submission to create a visible process record. • Require slides to be submitted alongside a short-written rationale explaining key design, content choices and a brief reflection on how the students' thinking developed.

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
				AI impact • Adding mandatory questions and answer section at the end and make this part as a component to be graded too. • Prepare a variety of questions that challenges the students' presented arguments, question something on a specific slide, or relates to their personal learning experience. • Include criteria that reward genuine personal perspective, such as connecting the topic to course-specific discussions or the student's own position.
Capstone projects  	• Combines generating solutions (problem definition, multiple solutions, iteration) with full project work (time/resources/teams) • Realistic, integrative task and mirrors professional work. Rich evidence for many skills including digital and communication.	• Resource and labour-intensive (supervision, assessment). • Individual contributions can be hidden in teamwork unless monitored. • Risk that the "final product" overshadows process or reflection; must ensure rubric covers all steps. • Labor intensive for supervisors. • Difficult to distinguish student capability from tool capability?	AI-sensitive / Design-dependent (AI resilience depends on assessing process, decisions, and justification) AI exposure: High – AI can be used throughout the project lifecycle (idea generation, drafting, coding, analysis, writing, visualisation) AI impact: High – Extensive AI use can significantly influence both process and final output, potentially obscuring students' independent contribution.	**Assessment** • Ensure a regular check-in structure throughout the project to reduce the risk of students going off-track undetected. • Use milestone assessments (research proposal, progress presentation, draft chapter) so the final grade reflects a sustained process rather than a single product. • Develop clear assessment criteria that distinguish between process quality (research design, iteration, supervision engagement) and product quality (argumentation, conclusions). • Provide explicit guidance on research ethics, referencing, and academic integrity expectations early in the process. **AI exposure** • Permit responsible AI use and require an AI transparency statement from students. Model this practice yourself. Frame and treat this as a

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
				professional habit of transparency, not a surveillance or punitive measure. • Require regular supervised writing moments or in-person progress meetings where students articulate their developing argument in their own words. • Use process portfolios that document iterations, dead ends, and supervisor feedback, making the intellectual journey visible alongside the final product. **AI impact** • Combine the capstone project with an oral defence assessment of ten minutes where students demonstrate their process and understanding in an unscripted context. • Only allow capstone projects that are sufficiently specific, related to the students' context, authentic, locally situated, and preferably tied to a unique dataset or setting.
Peer review  	• Promotes collaboration, peer-learning, and feedback-literacy. • Encourages metacognitive activities like reflecting on others' work preparing for the work-environments.	• Labor intensive • Quality depends heavily on peers' competence and rubric clarity. • Process may not be well documented for audit or grading clarity. • Leniency or friendship bias	AI-vulnerable unless designed carefully. AI exposure: High – AI can easily generate feedback, summaries, or critiques with little effort and limited detectability. AI impact: Moderate–High: AI-generated feedback may appear plausible but may not reflect genuine	**Assessment** • Provide structured review templates with specific criteria and guiding questions to reduce the risk of superficial or unhelpful feedback. • Train students explicitly in what constructive peer feedback looks like before the task begins, using examples of strong and weak reviews. • Make the quality of the review itself a graded component, separate from the grade given, to encourage real engagement.

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
			engagement, understanding, or independent judgement.	**AI exposure** • Provide explicit guidance on AI, ethics, referencing, and academic integrity expectations early in the process, specifically related to (uploading) someone else's work. • Time peer review activities within supervised sessions where possible or use in-class peer review as a low-stakes guided precursor to the graded version. • Require reviewers to identify one question they would ask the author in a follow-up conversation, creating a personalised trace of reading and writing. • Combine written peer review with a short paired discussion where reviewer(s) and author(s) talk through the feedback. **AI impact** • Require reviewers to refer to specific passages or decisions in the work they are reviewing. • Ask students to justify their evaluative judgements in their own words, linking feedback to the assessment criteria and to concrete evidence in the reviewed work. • Include a short oral or written reflection in which the reviewer explains what they learned from reading a peer's work in relation to the course. • Design review tasks around works-in-progress or drafts rather than final products. Require reviewers to reflect on the changes between products.
E-Portfolios  	• Strong for project work: shows planning, task breakdown, progress, resource/time management.	• Requires good scaffolding or students may just submit final versions with minimal process evidence.	AI-sensitive / Conditionally AI-resilient (depends on how reflection and process are assessed)	**Assessment** • Provide a clear portfolio framework early, specifying what evidence is required, what reflection depth is expected, and how both will be assessed.

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
	• Good for generating solutions when portfolio includes multiple versions, reflections on solution paths. • Encourages self-reflection and digital artefact management.	• Examiners may struggle to standardize across different types of artefacts (e-portfolio). • Labor intensive (unless e-portfolio)	AI exposure: AI can support reflection writing, artefact descriptions, synthesis, and narrative coherence across entries. AI impact: can substantially improve reflective language and narrative quality, potentially masking the depth of authentic learning.	• Ensure a regular check-in structure throughout the course to reduce the risk of students assembling their portfolio last minute and retroactively at the end of the course. • Develop rubrics that assess the quality of reflective thinking explicitly, distinguishing between descriptive accounts and actual critical self-assessment. • Offer worked examples of strong and weak portfolio entries so students understand what meaningful reflection looks like in practice. • Provide explicit guidance on ethics, referencing, and academic integrity expectations early in the process. **AI exposure** • Normalise AI transparency by requiring students to document their use and model this practice yourself. Frame and treat this as a professional habit of transparency, not a surveillance or punitive measure. • Require some portfolio entries to be completed during in-class moments, creating entries that are clearly student-generated. • Combine the portfolio with another type of assessment to increase validity and reliability. For example, with an added oral defence where a selection of the students' entries is discussed. • Build in a mandatory peer exchange where students share and respond to each other's portfolio entries, creating a social and relational dimension. Making the exchange also part of their portfolio increases personalisation.

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
				AI impact • Design reflection prompts that require students to respond to specific, dated events (e.g., a specific lecture, a piece of feedback) that AI cannot fabricate convincingly. • Ask students to connect each entry to a personal learning goal or a named course experience, grounding the portfolio in their individual trajectory. • Include evidence types that are inherently personal and time-stamped: photographs of fieldwork, annotated drafts with track changes, screenshots of iterative design work, or audio recordings of practice. • Assess growth and change over time explicitly, so a portfolio that presents polished final outputs without traces of development cannot score highly
Written assignments  	• Allows students to demonstrate sustained, structured reasoning. • The students' products are permanent and reviewable, which supports transparency. • Excellent for assessing research skills, argumentation, and synthesis. • Gives students autonomy in planning when and how they are working on the assignment.	• No direct verification of authorship during completion unless created in a controlled environment. • Labor intensive • Subjectivity and examiner bias • Feedback loop is delayed: feedback follows on a later date than other 'live' formats. • Risk of surface level completion. Structure and language use can mimic understanding without demonstrating it.	AI-sensitive. AI exposure: High – AI can be used throughout the process without any live oversight for generating arguments, synthesising sources, and editing texts. AI impact: High - AI can produce content that closely resembles student-authored work across typical assessment criteria: argumentation, evidence use, language quality, and structure. When writing	**Assessment** • If possible, organise a calibration session with colleagues also involved in the course before and after the assignment. Create a shared document of 'anchor' examples that can be used during grading. • Rotate topics, articles and/or scenarios between cohorts to limit circulation of previous student work and encourage authentic work.

Method	Strengths	Weaknesses	AI positioning	Suggested mitigation strategies
			quality and knowledge demonstration are central criteria, AI use can substantially influence outcomes without reflecting the student's own learning, directly threatening construct validity.	**AI exposure** • Normalise AI transparency by requiring students to document their use and model this practice yourself. Frame and treat this as a professional habit of transparency, not a surveillance or punitive measure. • Ask students to submit a draft mid-course and require a reflection paragraph with the end product that explains what has changed since the draft and why. Make this part of the grading criteria. • Combine written assignments with another type of assessment to increase validity and reliability. For example, an added oral defence of ten minutes. • When having multiple workgroup sessions, make talking about the assignment part of the learning activities to inquire about the students' reasoning so far, their opinion on articles, and their process. **AI impact** • Move away from generic essay prompts toward tasks that require integration of personally situated knowledge. • Accept AI as a legitimate thinking and drafting tool, but design tasks where the distinctly human contribution such as judgement, personal position, and course-specific insight is what is actually being graded. • Consider breaking a bigger assignment up into smaller assignments that can be completed in a controlled environment.

How to interpret the risk level and AI positioning?

This toolkit makes the role of AI in assessment design explicit by linking AI use directly to assessment quality, specifically validity (are we measuring the intended learning?) and reliability (are assessment conditions and judgements consistent across students and contexts?).

The toolkit uses a simple three-part framework to clarify the role of AI and its effect on assessment quality. AI Exposure refers to how easily AI can be used during assessment preparation or performance. This is important for reliability, as uneven access to AI can lead to unequal assessment conditions. AI Impact describes how much AI use can influence the assessment outcome. This affects construct validity, because when AI can significantly improve performance without demonstrating the intended learning (for example, better writing without deeper understanding), the assessment may measure the wrong things. The Overall AI Position brings these two aspects together and indicates whether AI resilience is built into the assessment design itself or depends on specific design decisions, such as questioning strategies, process evidence, or individual accountability. This supports both validity and reliability by keeping the focus on intended learning outcomes and consistent judgement across students. Importantly, the AI position does not label assessment methods as universally "AI-safe" or "AI-vulnerable," but highlights that AI resilience is a feature of assessment design, not of the method alone.

To support consistent and defensible judgement, assessment methods in this toolkit are grouped using a three level AI risk framework. The risk levels are not fixed labels, but indicators to support reflection, discussion, and justification.

- **Level 1 – Minimal risk (AI- resilient by design)**
  At this level, key risks are addressed directly through the assessment design itself. Even when AI tools are available, they cannot meaningfully replace the intended learning or strongly influence student performance. As a result, validity and reliability are largely protected without additional safeguards.

- **Level 2 – Moderate risk (design-dependent)**
  At this level, AI exposure is realistic and AI impact can be meaningful, but the level of risk depends strongly on how the assessment is designed and used in context. With clear guidance, expectations for transparency, and criteria that prioritise reasoning, process, or judgement, risks to validity and reliability can be reduced. AI resilience is achieved through intentional design and careful implementation, not through the assessment method alone.

- **Level 3 – High risk (design-sensitive and safeguard critical)**
  At this level, AI can meaningfully shape cognition and directly overlap with the core learning outcomes being assessed. Without strong and non-negotiable safeguards, there is a high risk that the assessment measures effective AI use rather than student learning.

References

Anthropic. (2026). *Claude* (Claude Sonnet 4.6) [Large Language Model]. https://claude.ai
Berkel, H. van, Bax, A., & Joosten-ten Brinke, D. (2014). *Toetsen in het hoger onderwijs*. Bohn Stafleu van Loghum.
Joughin, G. (2010). *A short guide to oral assessment*. Leeds Met Press in association with University of Wollongong.
Microsoft. (2026). *Microsoft Copilot* (Copilot with GPT-4o) [Large Language Model]. https://copilot.microsoft.com

Disclaimer

For the final editing of this text, Claude (Anthropic, 2026) and Microsoft Copilot (Microsoft, 2026) were used for AI-assisted text editing. The content and argumentation are the authors' own.