

HOW TO DESIGN ASSIGNMENTS STUDENTS CAN'T OUTSOURCE TO AI

Stephen Murgatroyd, PhD

Contact North | Contact Nord

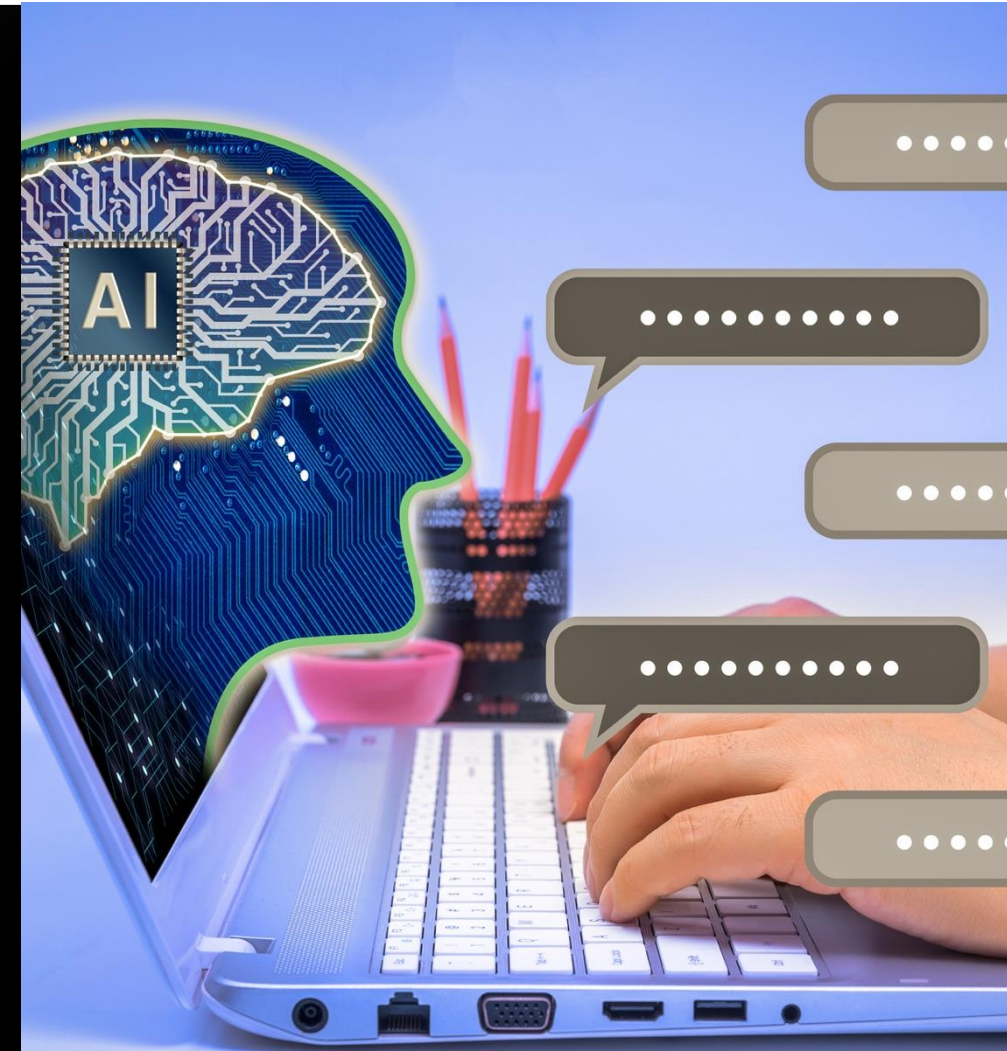
AS FACULTY, WE'RE FACING MORE THAN JUST A CHEATING CRISIS

- We're facing a trust crisis
- We're facing a signalling crisis
- Beneath both, we're facing a formation crisis
- AI didn't create this
- It exposed what was already in plain sight



THE REAL PROBLEM ISN'T CHEATING

- Cheating has existed since the Ming Dynasty
- Surveillance hasn't stopped it
- Detection hasn't stopped it
- AI simply makes cheating faster, easier and harder to detect
- 94% of UK undergraduates now use AI on assessed work (Stephenson & Armstrong, 2026)



THE BETTER QUESTION

Instead of asking:

Did this student use AI?

We should ask:

Can this student think?

Can this student explain?

Can this student defend?

Can this student transfer?

Can this student apply what they know?

WHAT IS THE REAL PROBLEM?

If students can think, explain, defend, transfer, and apply their learning, then AI use is not the central issue.

If they can't, the issue isn't cheating.

The issue is assessment design.

FACULTY ARE UNEASY

- We assign an essay
- We assign a math problem
- We assign a science design challenge
- AI can produce a response in seconds
- It sounds educated
- It reads well
- It can follow a rubric closely
- It may be difficult not to give it a strong grade

WHAT ARE WE ACTUALLY GRADING?

As faculty, we find ourselves asking:

- Are we grading the student?
- Are we grading ChatGPT?
- Are we grading Gemini?
- Are we grading Claude?
- Or are we grading the design of our own assessment?

WHAT MANY ASSIGNMENTS REWARD

Many assignments still reward:

- Regurgitation of knowledge
- Pattern recognition
- Formulaic writing
- Formulaic problem solving
- Surface synthesis

These are exactly the kinds of things large language models are trained to do.

THE CULTURE WE HELPED CREATE

We have also created a culture where grades often count far more than learning.

Failure is too often treated as defeat rather than feedback.

In that culture, cheating can begin to look like a reasonable risk.

THE SURVEILLANCE FANTASY

The fantasy says:

- Policies are clear
- Detection systems are in place
- Consequences are serious
- Students will comply



WHY THE FANTASY FAILS



- No detector is more than 80% accurate – some are at 65% or lower
- False positives create legal and ethical risks
- Detectors are biased against writers whose first language isn't English
- Faculty and detectors often disagree
- Students quickly learn new ways to use AI that detectors can't identify
- Surveillance doesn't rebuild trust

SAME ASSESSMENT, WEAKER INFERENCE

What stayed the same:

Course, content, criteria
and examiner

What changed:

An AI-permitted take-home
exam became a supervised
exam.

Failure rose from 2-6% to
18.4% (Brattli et al., 2026).

THE LOSS WE RARELY NAME

Producing the work was part of how capability was formed.

A student who hands production to AI hasn't mainly misrepresented themselves. They've skipped the process that would have made their knowledge usable.

The formative loss comes before the evidentiary one.

HOW KNOWLEDGE BECOMES CAPABILITY

Knowledge converts to capability through:

- Conceptual organization
- Conditionalization: knowing when it applies
- Enactment
- Feedback
- Metacognitive monitoring
- Evidence generation

Three of these six are things assessment does. Assessment isn't a report on learning. **It's one of the ways learning happens.**

ONE MODEL, TWO DESIGNS, OPPOSITE RESULTS

Unrestricted GPT-4:

48% better in practice

17% worse on the exam

GPT-4 set up as a tutor:

Hints, not answers

127% better in practice

No harm on the exam

Same model. Different design.

Bastani et al. (2025), PNAS

THE FUTURE OF ASSESSMENT

The future of assessment is not more assessment or more supervised assessment.

The future of assessment is assessment redesign.



STUDENTS USE AI PRODUCTIVELY

- Students can use AI as a:
 - Collaborator
 - Co-intelligent partner
 - Brainstorming engine
 - Reflection coach
 - Challenge coach
 - Editor
 - Critical reviewer
 - Thinking partner
 - Resource finder

WE SHOULD ENCOURAGE PRODUCTIVE AI USE

As faculty, we should encourage students to use AI in ways that support learning.

- The goal is **not** unrestricted use
- The goal is wise, transparent, and meaningful use – we call it scaffolding the use of AI



SOME STUDENTS USE AI TO AVOID LEARNING

Some students will:

- Use AI to think and write for them
- Use AI to bypass productive struggle
- Outsource their cognitive processes
- Ask AI to be them instead of becoming themselves

OUR CHALLENGE AS FACULTY

Our challenge is to:

- Stop cognitive surrender and the privatization of cognition
- Stop the outsourcing of learning
- Design assignments that students can't simply hand over to AI

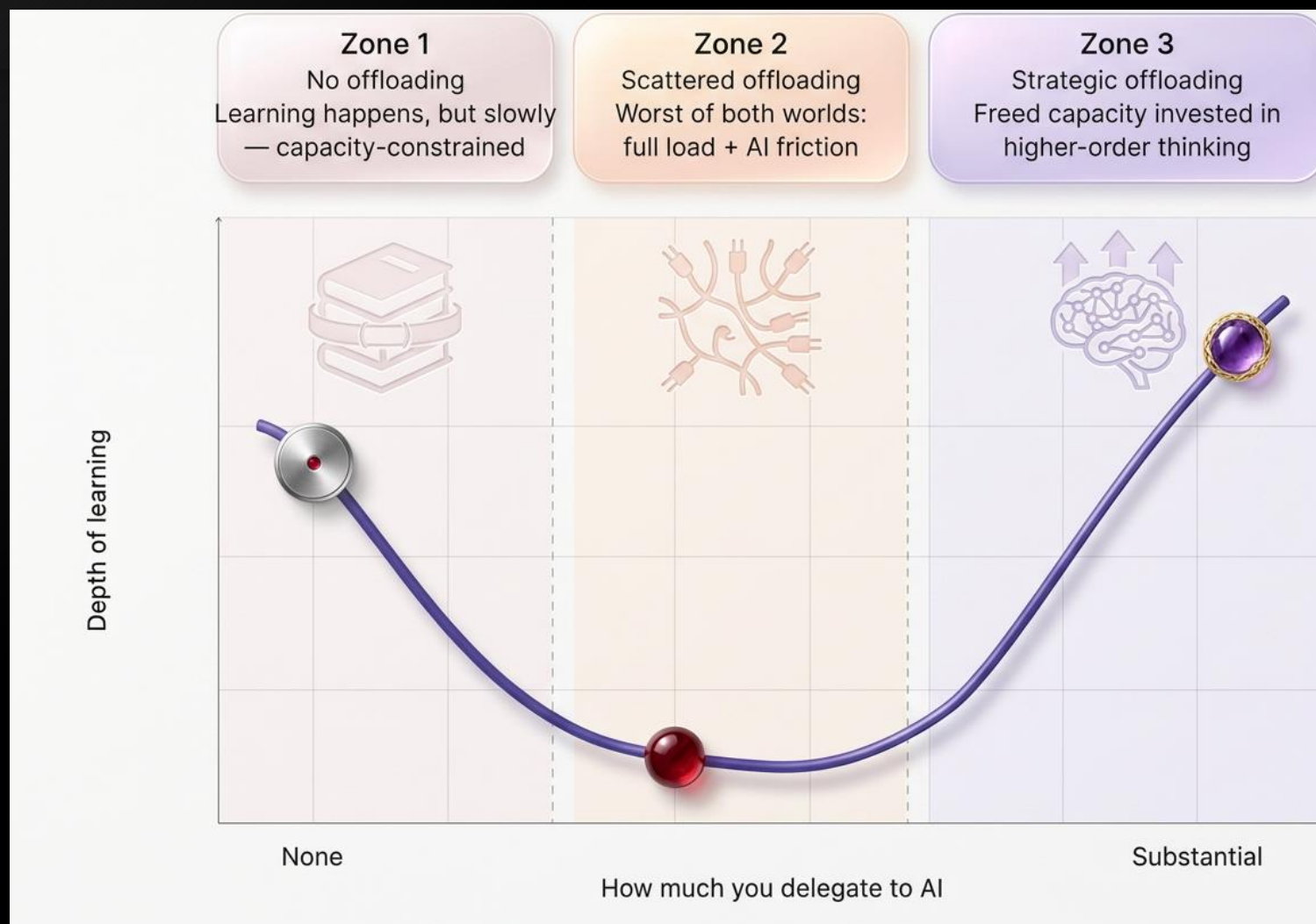
THE COGNITIVE OFFLOADING PARADOX

- Offloading thinking to AI does not automatically reduce learning.
- In some cases, it can improve learning.
- Students who delegate substantive work strategically report deeper learning, not shallower (Wang & Zhang, 2026).

THREE ZONES OF OFFLOADING

- Most students sit in Zone 2: the full load, plus the friction of managing the tool.
- Moving to Zone 3 is a taught capability, not a policy setting.

After Wang & Zhang (2026); Hardman (2026)



THE REAL INTEGRITY QUESTION

Academic integrity is not only
about whether AI is used.

It is about ***how*** AI is used.



THE DESIGN CHALLENGE

Learning outcomes depend on:

- What students delegate
- What students retain
- How students use the capacity AI frees up

THREE CONDITIONS FOR POSITIVE LEARNING

Positive learning requires:

- Enough knowledge to keep hold of the problem
- Meaningful delegation to AI systems
- Deliberate reinvestment of thinking

If we don't design for this, students drift into cognitive surrender.



A TEST FOR YOUR ASSIGNMENTS

- Go to Claude or ChatGPT
- Share your course material
- Ask the system to complete all the assignments for the course
- Grade the work



IF AI PASSES, WE HAVE A DESIGN PROBLEM

- If AI completes the assignments and earns a B or an A, the issue is not student misconduct
- GPT-4 answered up to 85% of assessment questions correctly across 50 university courses (Borges et al., 2024)
- The issue is assessment design

ASSESSMENT NEEDS A RENAISSANCE

We need to move toward:

- Process-based assessment
- Oral defence
- In-class performance
- Multimedia presentations
- Iterative drafts
- AI transparency statements
- More formative assessment
- Fewer high-stakes summative assessments
- Competency-based demonstrations

BUILD, ADAPT, DEFEND

- Build: Students create something meaningful
- Adapt: Students revise, transfer or apply what they've created in a new context
- Defend: Students explain and justify their choices



WHY BUILD, ADAPT, DEFEND WORKS

- Build, Adapt, Defend addresses core integrity risks
- Hidden AI use is made explicit through Build
- Surface learning is countered through Adapt
- AI-generated work is neutralized through Defend

BUILD, ADAPT, DEFEND: A WORKED EXAMPLE

- **Before:** a 2,500-word policy brief, submitted once and graded as an artefact
- **Build:** draft it with AI used openly; hand in the prompt trail and a transparency statement
- **Adapt:** in class, the context changes - new jurisdiction, new constraint, new data. Revise, and annotate what changed and why. Use the paper by Pat Thomson (Calling Out Policy Bullshit – J Ed Policy) as a lens for a review.
- **Defend:** ten minutes, no tool. Justify two choices, name what the AI got wrong, answer one challenge

THE ESSAY, REDESIGNED

- **Before:** a 1,500-word argumentative essay on a set question, submitted once
- **Build:** ask AI for the strongest case on each side, then write your own position; submit the AI exchange and a transparency statement
- **Adapt:** in class, receive a source the AI didn't have - a primary document, local data, a counter-argument - and rewrite one section by hand, annotating what changed
- **Defend:** five minutes. Which AI claim did you reject, and why? What evidence would change your mind?

THE MATH PROBLEM, REDESIGNED

- **Before:** ten textbook problems, answers submitted online
- **Build:** work with AI set up as a tutor, not a solver - hints, not answers; submit worked solutions and mark the one place you got stuck
- **Adapt:** in class, a variant - a changed parameter, a missing value, or an AI solution with a planted error to find and fix
- **Defend:** talk through one step aloud. Why this method and not another? Estimate the answer before calculating it

THE SCIENCE DESIGN CHALLENGE, REDESIGNED

- **Before:** design an experiment to test a hypothesis; submit the written protocol
- **Build:** have AI generate three candidate designs; choose one and record why you rejected the other two
- **Adapt:** in the lab, the equipment, budget or sample changes; revise the protocol and predict what will now go wrong
- **Defend:** questions at the bench. What is your control? What result would prove you wrong? What did the AI's design overlook?

ACROSS APPLIED PROGRAMS

- **Nursing:** AI drafts a care plan; students critique it against a patient case it hasn't seen, then defend their priorities in the simulation debrief
- **Skilled trades:** AI explains the code requirement; students do the job in the shop and talk the assessor through each safety decision
- **Business and accounting:** AI produces the analysis; students find the error planted in the client data and defend their recommendation to a client who pushes back
- **Early childhood and social work:** AI summarizes the theory; students apply it to an observed case and justify their judgement in supervision

DEFEND: TWO TESTS AND AN EVIDENCE RULE

Two tests any instructor can run:

- Can the student perform the delegated task unassisted?
- Can the student say what the AI's answer omits, assumes or gets wrong?

An evidence rule: unsupervised written work is never the sole evidence. Look for more than one instance, in more than one context.

WHAT BUILD, ADAPT, DEFEND DEVELOPS

- Metacognition
- Judgement
- Transfer of learning
- Ethical awareness
- Confidence
- Capability



FIVE QUESTIONS FOR EVERY STUDENT

Ask students to answer, every time:

- Do I understand this?
- Did AI generate this, or did I?
- Can I explain it myself?
- Could I have done it without assistance?
- Was I learning, or merely completing?

This is what an AI transparency statement should ask. Each question is a metacognitive act. The one a student can't answer is the one to teach.

THE COGNITIVE ARCHITECTURE MATTERS

“AI is a design variable - and like every design variable, its impact depends entirely on the decisions made by the person who shapes the interaction.”

Philippa Hardman (2026)

WE NEED TO ASSESS MORE THAN DOMAIN KNOWLEDGE

- We need to think about what we assess, not just how we assess.



EVERGREEN SKILLS MATTER

The skills most valued in the future will increasingly include:

- Emotional intelligence
- Critical thinking
- Judgement
- Creativity
- Originality
- Adaptability
- Teamwork
- Leadership

DOMAIN SKILLS STILL MATTER

- Domain-specific skills still matter
- Foundational knowledge still matters
- But many domain skills change quickly
- Evergreen skills endure longer

EVERGREEN SKILLS ARE MADE FROM KNOWLEDGE

Capability isn't an alternative to knowledge. It's what knowledge becomes when it's converted into action and possibility.

Cut the disciplines and we cut the raw material.

ASSESSMENT MUST PROVIDE FEEDBACK ON BOTH

Assessment needs to provide feedback on:

- Domain knowledge and skills
- Evergreen knowledge and skills

Right now, we too often assess domain knowledge alone.

A FORECLOSURE, NOT A TRADE-OFF

Domain skills depreciate and can be replaced, provided the person can still learn.

Evergreen capability that never formed is hard to acquire later, because acquiring it requires the very capacity it constitutes.

Graduates find the gap where it's most expensive to repair: mid-career, without an institution, an assessor or time.

WHAT WE NEED TO SEE MORE OF

- We need to see more:
- Oral examination
 - Presentations
 - Studio critique
 - Socratic dialogue
 - Simulation-based assessment
 - Collaborative problem-solving
 - Impact-based project work
 - Witnessed performance

THE FUTURE MAY LOOK MEDIEVAL

The future of assessment may look strangely medieval.

- More witnessed performance
- More oral defence
- More demonstration
- More apprenticeship
- More judgement

MEDIEVAL HAS A COST AND A RISK

The cost:

- Judging performance is closer to clinical supervision than to marking
- Few of us were ever trained to do it
- It takes shared descriptors, calibration and time no workload formula counts

The risk: more human judgement means more room for patterned bias. Contain it with several instances, more than one assessor, and descriptors built with the communities being judged.

ASSESSMENT REDESIGN IS PROGRAM WORK

This can't be solved one course at a time.
We also need program-level assessment redesign.

We need to ask:

- What needs to be witnessed?
- What can be documented?
- What can be demonstrated?
- What can be supported by AI?

PROGRAM-LEVEL REDESIGN

We need to incorporate:

- Shared AI literacy
- Competency-based assessment
- Transparent AI policies
- Assessment ecosystems
- Calibration among assessors
- Substantial institutional support

FROM FRAGMENTED ASSESSMENT TO INTEGRATED ARCHITECTURE

Old model:

- Fragmented course by course assessment

New model:

- Integrated program architecture

WHAT PROGRAMS CAN FOCUS ON

- Iterative assessment structures across courses
- Systematic feedback loops
- Reflective process evidence
- AI collaboration models
- Rubrics co-created with faculty and students
- A shift from outcome-focused to process-focused assessment

SWISS CHEESE ASSESSMENT: LAYERS, NOT SINGLE SLICES

- No single assessment proves capability.
- Multiple assessments across time provide overlapping evidence of competence.
- Assessment credibility comes from patterns of evidence.
- This isn't untried: programmatic assessment has worked this way in health professions education for two decades (van der Vleuten et al., 2012).

WHY GOOD REDESIGN STALLS

Four things still count time, not capability:

- Qualifications frameworks: credit hours
- Student aid and funding: periods of study
- Workload formulas: contact hours
- Learning platforms: weekly modules and gradebooks

None of them forbids capability assessment. Together they make it expensive, invisible and non-transferable. Wisconsin's Flexible Option moved three of the four.

THE QUESTION FOR INSTITUTIONS

Will our institutions become places that try to preserve a disappearing world?

Or will they become places that teach students how to live wisely, truthfully and humanely in the age of AI?



HUMAN FORMATION REMAINS THE MISSION

Technology changes quickly.

Human formation and
flourishing remains the
mission.



THE SYSTEMS THAT THRIVE

The systems that thrive in the next decade will not be those that police AI most effectively.

“They will be the systems that most successfully teach young people how to think, discern, create, collaborate and act ethically in partnership with intelligent machines.”

Murgatroyd & Couture, 2026
AI Unplugged – The Hype and Hope for Education Futures.
Routledge.

WHAT WE CAN DO

- Co-create with students, teachers and leaders
- Establish program-level AI norms
- Support assessment redesign conversations
- Protect staff from detector overreach
- Procure AI tools configured to tutor, not to answer
- Encourage experimentation with safeguards
- Build AI literacy for staff and students
- Monitor emerging risks and inequities

WHAT STUDENTS NEED TO HEAR

- AI literacy is now part of responsible citizenship
- We are not surrendering standards
- We are changing how we assess domain and evergreen skills
- Integrity still matters deeply
- Students must still think, reflect, defend, and demonstrate understanding
- Human judgement is central and increasingly valued
- The goal is wise and ethical use, not unrestricted use

WHAT EMPLOYERS NEED TO HEAR AND SEE

- We have taken steps to ensure student achievement reflects students' own knowledge, skills and understanding
- We can show what each student is capable of
- Students leave with a record of capability they own and others can verify
- Students can make effective use of AI-ready tools in support of their work
- Students are AI ready
- Students are effective communicators
- Students have strong evergreen skills

IF WE DON'T, OTHERS WILL

Employers who can't trust credentials have begun building their own assessments.

If we decline to assess what we claim to develop, others will assess it for us.

AI didn't break assessment. It disclosed what assessment had stopped doing.

THANK YOU

For more information

Dr. Stephen Murgatroyd

Contact North | Contact Nord

Research Associate

Stephen@contactnorth.ca and smurgatr@ualberta.ca

SOURCES

- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122.
- Borges, B., Foroutan, N., Bayazit, D., Sotnikova, A., Montariol, S., Nazaretzky, T., ... Bosselut, A. (2024). Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. *Proceedings of the National Academy of Sciences*, 121(49), Article e2414955121.
- Brattli, H., Utne, A., & Lynch, M. (2026). Assessment validity in the age of GenAI: A natural experiment. *Informatics*, 13(4), Article 56.
- Hardman, P. (2026, April 16). The “cognitive offloading” paradox. *Dr Phil’s Newsletter*.
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), Article 100779.
- Lodge, J. M., & Loble, L. (2026). Artificial intelligence, cognitive offloading and implications for education. University of Technology Sydney.
- Murgatroyd, S. (2026). Converting knowledge into capability: Assessment as formation in the age of AI [Manuscript].
- Murgatroyd, S., & Couture, J.-C. (2026). *AI unplugged: The hype and hope in education futures*. Routledge.
- Sigelman, M., Fuller, J., & Martin, A. (2024). Skills-based hiring: The long road from pronouncements to practice. Burning Glass Institute.
- Stephenson, R., & Armstrong, C. (2026). Student generative artificial intelligence survey 2026 (HEPI Report 199). Higher Education Policy Institute.
- van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205-214.
- Wang & Zhang (2026). [TITLE, INITIALS AND ARTICLE NUMBER TO BE VERIFIED]. *International Journal of Educational Technology in Higher Education*.
<https://doi.org/10.1186/s41239-026-00585-x>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26.