Learning With AI, and Teaching With Integrity

An evidence-based guide: how to learn well with AI, and how to protect academic integrity in the age of generative AI.

John Hay Biomedical · Cleveland Metropolitan School District

July 2026

Every claim below is cited to peer-reviewed research or primary documentation. Notably, when popular explainer videos were adversarially fact-checked, no video-only claim survived; the durable evidence is all research and first-party sources.

Part 1 · Learning Well With AI

One finding runs through everything the evidence says: AI reliably makes you faster and more productive, but hands-off use erodes or prevents the very skills that make the output good. The skill is staying in the loop on purpose. Two throughlines carry the whole guide: **ground the AI in real sources, and keep yourself doing the thinking that matters.**

1. NotebookLM (now Gemini Notebooks): a safer tool by design

Why it is safer. It answers only from sources you upload, using retrieve-then-generate. By default it will not pull from the open web or its own memory. That structure is the reason it hallucinates less.

How much less. A peer-reviewed 2025 study (Tozuka et al., Japanese Journal of Radiology) handed NotebookLM and GPT-4o the same references on 100 lung-cancer staging cases. NotebookLM scored **86 percent versus 39 percent.** Same sources, far less invention.

- **But not zero.** A separate 2025 measure put its residual hallucination near 13 percent, versus roughly 40 percent for open chatbots. Grounding reduces, it does not eliminate. You still verify.
- **Citation caveat.** Its inline citations are clickable and jump to the passage, but they lack page or paragraph precision and the highlighted span can be loose. Traceable, not court-grade.
- **The workflow.** Add your sources, ask grounded questions with citations, then generate study aids: Audio and Video Overviews, Flashcards, Quizzes, Mind Maps, Infographics, and Slide Decks.
- **Know your mode.** Opt-in features (Discover Sources, web search, agentic Deep Research) can reach beyond your uploads. They add sources rather than change the grounded default, but they change the trust profile.

2. Minimizing hallucinations (works everywhere, not just NotebookLM)

- **Ground it.** Adding retrieval cut hallucinated content from about 21 percent to about 2 percent in a measured NAACL 2024 case study. Grounding is the single biggest lever.
- **Demand citations, then check them.** A strict-citations prompt gave the highest verifiable grounding in a 2026 medical-QA benchmark, but up to 57 percent of citations can be post-rationalized. The citation is half the job; opening the source is the other half.
- **Reasoning prompts are not automatically safer for facts.** An expert chain-of-thought prompt pushed models to trust internal logic over the retrieved text (unsupported-sentence ratio 95 to 100 percent). Reasoning is great for problems, risky for stick-to-the-source.
- **Self-verification.** Chain-of-Verification (Meta AI, ACL 2024): have the model draft, write its own check questions, answer them independently, then correct. It measurably lowers hallucination, and you can just prompt for it.

- **Two-model cross-check.** Teaming architecturally different models beats one model checking itself (NeurIPS 2025 workshop). A model that hallucinates confidently wins its own vote; a different model can out-vote it.

3. Keeping AI on task

Honest flag: this area is less study-backed than the rest, so treat it as practitioner guidance that is consistent with the evidence.

- **Scope tight, decompose, one job at a time.** AI degrades outside its zone, so shrink the task to where it is strong.
- **Constrain the output:** format, length, only from these sources, hints not answers.
- **Iterate in small steps and re-anchor when it drifts.** Do not let a single thread sprawl.

4. Leveraging different models

Match the model to the task: grounded tools (NotebookLM) for source-faithful study, strong reasoning models for problem-solving, fast and cheap models for bulk or deterministic work.

Capability is jagged. The Harvard and BCG study (758 consultants, Organization Science 2025): inside the frontier, GPT-4 users did about 12 percent more tasks, roughly 25 percent faster, at more than 40 percent higher quality. On a task engineered to fall outside the frontier, AI assistance cut correctness by about 19 points. Different models have different frontiers, so a second model is not just a checker, it covers ground the first one cannot.

5. Skill, and the reallocation question

The core danger, the learning paradox. In a PNAS randomized trial of about 1,000 students, an unguarded ChatGPT tutor raised practice scores 48 percent, yet those students scored **17 percent worse on a later unaided exam** than peers who never used AI. A hints-not-answers tutor with teacher guardrails erased the harm. Same tool, opposite outcomes, purely down to how it was used. And in-the-moment fluency is a false signal: Bjork's desirable-difficulties work shows current performance is not a reliable index of learning, which only shows on delayed recall.

Deskilling of trained experts is real. The Lancet Gastroenterology and Hepatology ACCEPT trial (2025): after AI-assisted colonoscopy became routine, the same endoscopists' unaided detection rate fell from **28.4 percent to 22.4 percent** (about 20 percent relative). Precedents run the same way for pilots' manual flying and heavy GPS users' spatial memory. Fair caveat: the colonoscopy study is observational and case volume nearly doubled, so the cause is argued even though the drop is real.

The reallocation model, graded honestly

- **Right (well-supported).** Skills go rusty, not destroyed. Procedural skill decays roughly 0.06 to 0.08 SD per month, half the gains gone in about 6.5 to 13 months (2026 Psychological Bulletin meta-analysis),

and they come back faster than first learned (relearning savings, replicated since Ebbinghaus). First-language attrition validates the Spanish-and-English intuition: the native language is not lost, it suffers production interference, and re-exposed speakers become statistically indistinguishable from monolinguals after roughly a week.

- **Oversimplified.** The tidy inverse-proportional, rusty-in-proportion-to-disuse framing. Attrition research finds no reliable link between raw frequency of use and how much you lose, and the relationship is bidirectional, not a clean one-way seesaw. The direction is right; the neat proportionality is not how it behaves.
- **Where it breaks: novices.** The model assumes the skill was built first, then rusts. For someone who offloads before ever building it, there is no schema to recover. An RCT (N=78) on novice programmers with unrestricted AI: they matched or beat manual peers on immediate output (92.4 versus 65.2 percent), but their ability to repair their own work collapsed once AI was removed. Fragile experts.
- **Where it breaks: the reallocation itself.** Freed capacity is not automatically redirected to higher-value work. And the key question, did overall quality rise, is largely unanswered and what exists is cautionary: in METR's developer RCT, experienced developers were 19 percent slower with AI while believing they were 20 percent faster. Professionals cannot reliably self-assess AI's effect on themselves, which is exactly why quality has to be measured, not felt.

AI is genuinely good at	Humans must own
Breadth, speed, tireless first drafts	Context and real-world stakes
Pattern recall and summarizing	Judgment and verification
Inside-the-frontier tasks	Knowing where the frontier is
Volume and iteration	Values, accountability, the final call

The winning pattern is neither hands-off (AI does it all, which loses quality to missing context) nor hands-clean (I do it all, which loses the volume). It is the split: divide and interleave the work, and the human verifies.

6. Best practices for learning with AI

1. **Ground it.** Prefer source-grounded tools so answers come from your materials.
2. **Demand citations, then open them.** Half the value is in the click.
3. **Make AI a coach, not a crutch.** Hints and questions, not the answer. That single guardrail was the difference between harm and no-harm in the PNAS trial.
4. **Attempt first, then check.** Produce from your own head before asking; use AI to critique, not to originate.

5. **Cross-check what matters with a second, different model.**
6. **Build the skill before you offload it,** especially for beginners. No AI on the fundamentals until the schema exists.
7. **Measure real learning on a delay, unaided.** Fluency now lies; recall later tells the truth.
8. **Judge the portfolio, not the task,** and verify it, because you cannot feel it accurately.

Part 2 · Academic Integrity in the AI Era

Start with the finding that should reframe the whole problem: generative AI did not make students cheat more. Stanford and Challenge Success surveys of about 40 U.S. high schools (Lee, Pope et al., peer-reviewed 2024) show that **60 to 70 percent** of students self-reported at least one cheating behavior per month for 15-plus years before ChatGPT, and that rate stayed flat or dipped slightly afterward, matching a roughly 64 percent cheated-on-a-test baseline from ICAI surveys of 70,000-plus students.

> **The reframe**
>
> The tool changed; the rate did not. Cheating is driven by grades, workload, and stakes, not by AI availability. So the winning response is integrity by design, not detection or heavier penalties. Honest caveats on the data: it is self-reported, it is an overall rate (AI-specific incidents rose from a low base), and it is time-bounded. The researchers themselves say this may change as teen use of ChatGPT for schoolwork has roughly doubled.

1. What changed, and what did not

The impulse is old; the delivery is new. Before generative AI, cheating meant copy-paste plagiarism, essay mills, and copying peers, all catchable by similarity matching. After it, a language model produces original-looking text instantly, cheaply, and invisibly to similarity tools. The ease and undetectability changed, not the motivation. This is exactly the gap AI opened in your toolkit: a Turnitin similarity score still catches copy-paste, but it cannot catch AI-ghostwritten original prose.

2. The detector problem (which validates your stance)

AI text detectors are unreliable and biased. Across seven widely used tools, more than half of human-written non-native-English essays were misclassified as AI-generated (**61 percent** average false-positive rate); one flagged 89 of 91; native-writer essays were near-perfect (Liang et al., 2023, Patterns / Cell Press, replicated since).

> **What this means for your policy**
>
> The false accusations fall hardest on English-language learners and some neurodivergent writers, which makes detector-alone enforcement pedagogically and legally risky. Your policy already says detectors are unreliable and never penalizes on the score alone. The one refinement the evidence adds: even your highest AI-probability tier should trigger the oral exam and the request for drafts, never a penalty on the detector number by itself. You are already doing this.

3. What works: integrity by design

Ranked by strength of evidence:

1. **Assessment redesign plus process artifacts plus oral defense.** The strongest, and it is what you already do. Bertram Gallant (Cambridge, 2026): multi-layered policy, assessment redesign, and cultivating ethical behavior, explicitly beyond technology-only fixes. The College Board AP model is the clearest working example: a purpose principle (support learning, do not bypass it), a plain allowed / limited / forbidden framework, and interim checkpoints where students make their thinking visible, similar to an oral defense, with a score of 0 for failing to complete them, and no mention of detectors. Your 70-percent-triggers-an-oral-exam-plus-drafts rule is this exact model.
2. **Transparent, tiered AI-use permissions.** The AI Assessment Scale (Perkins et al., 2024): five clearly named levels from No AI to AI Exploration, used as a communication tool and an assessment-redesign framework, not a policing mechanism. State the permitted level on every assignment so students always know the rule for this task.
3. **Communicate the policy and involve students in its norms.** In a 2024 study, severity of punishment had no measurable deterrent effect, while communicating the policy and involving students did (honor codes had only a mild effect). The lever is not the size of the penalty.
4. **Detectors used alone.** Not recommended, see section 2.

4. Device-free assessment: how much is reasonable?

Honest flag: no verified source settled a specific percentage. What follows is reasoned design judgment, not a cited number.

The case for device-free work is that it is the one mode AI cannot touch, and it doubles as the true measure of learning (delayed, unaided recall). It gives you a secure anchor for grades that cannot be ghostwritten. **The case against an all-device-free course** is that your students are headed into biomedical and health careers where fluent, ethical use of devices, records systems, lab software, and AI is the professional skill. A blanket ban fails the real-world-preparation goal.

> ### Recommendation: a two-lane course
>
> A secure lane (device-free or in-person-verified): in-class handwritten writing, problem-solving, oral defenses, lab practicals, closed-device quizzes. This is your verification layer. An open lane (device-rich, AI allowed and taught): authentic projects, lab reports, and research where you teach professional AI use and grade the process. Anchor roughly one-third to one-half of the grade to the secure lane. Skew toward one-half for foundational skills you must certify individually; toward one-third for capstone or authentic-work courses. The secure minority must be large enough that a student cannot pass on AI-generated work alone.

The elegant part: the secure lane makes the open lane self-policing. If a student must also demonstrate the skill unaided, then having AI do the take-home work becomes pointless, it only sets them up to fail the

verification. That is your 70-percent oral-exam mechanism working as a system, not as a punishment. So the answer to what percentage is not a fixed number, it is enough secure, unaided assessment that grades certify individual skill, roughly a third to a half of the weight, with the rest device-rich and AI-literate.

5. Your policy: aligned versus worth adjusting

Part of your policy	Verdict
No detector-alone penalties; holistic review with appeal	Strongly aligned. This is the scholarly consensus.
70 percent similarity triggers an oral exam plus drafts as evidence of idea development	Strongly aligned. This is the College Board and Bertram Gallant gold standard.
A single, clear similarity threshold (70 percent)	Good. One number is clearer and fairer than competing cutoffs.
Confirmed dishonesty results in a non-recoverable zero	Worth reconsidering. The evidence says penalty severity does not deter.

On the non-recoverable zero. The 2024 evidence predicts that severity alone will not increase deterrence and may cost trust, while a path to demonstrate authentic learning plus clear communication does deter. Your earlier rebuild-through-authentic-work instinct was actually closer to the evidence. A defensible middle: keep the zero on the invalidated work, but pair it with a restorative path (redo it authentically, or pass the oral defense), and above all communicate the why and involve students in the norms.

> ### The frame that makes it land
>
> Sarah Elaine Eaton: integrity is ethical decision-making, not plagiarism-policing. Sell the policy to students as protecting the value of what they earn, and connect it to the learning research: cheating with AI does not just break a rule, it robs them of the skill. The honest pitch is simple. The AI can pass the assignment. It cannot pass the oral exam, the board exam, or the clinical floor for you. Go Hornets.

Sources that held up

- **Learning tools and hallucination:** Google Gemini Notebook documentation; Tozuka et al. 2025 (Japanese Journal of Radiology); Chain-of-Verification (Meta AI, ACL 2024); the Harvard and BCG Jagged Frontier study (Organization Science 2025).
- **Learning effects:** Bastani et al. 2025 (PNAS); Bjork and Bjork on desirable difficulties; Fan et al. 2025 (BJET) on metacognitive laziness; Mollick and Mollick, Seven Approaches.
- **Skill and deskilling:** Budzyn et al. 2025 (Lancet Gastroenterology and Hepatology); METR 2025; the 2026 Psychological Bulletin skill-decay meta-analysis; first-language attrition (Schmid; Chamorro, Sorace

and Sturt).

- **Integrity:** Lee, Pope et al. 2024 (Computers and Education: AI); Liang et al. 2023 (Patterns); Bertram Gallant, Academic Integrity in the Age of AI (Cambridge 2026); Eaton (University of Calgary); the AI Assessment Scale (Perkins et al. 2024); College Board AP AI policy.
- **A practitioner worth following (written, not video):** Ethan Mollick, One Useful Thing.