The Zero-Line Graduate: How Subsidized AI and Artifact-Based Grading Manufacture Dependency Without Capability

Ștefan-Daniel Wagner and Eduard-Cristian Popovici

***Abstract*—Generative AI can augment a student or replace the student's thinking, and the two are not the same behavior. We synthesize a preregistered field experiment, lab and survey studies, and contested correlational work: in the experiment, assistant access raised assisted scores while lowering later unaided performance in the unsafeguarded arm, correlational results point the same direction, and users misjudge the trade in both directions. We then argue that higher education currently makes offloading the default through two structural mechanisms. Subsidized flat-price access, which analysts project is priced far below cost, removes the price of outsourcing thinking; evaluation that scores artifacts rather than unaided process removes the choice. Their convergence, stated as a conditional trajectory with early indicators and an explicit refutation criterion, is the zero-line graduate: a credential certifying a skill its holder cannot exercise without the tool, at a price set by someone else. Because the offloaded step is usually verification, assistant honesty is load-bearing, yet sycophancy research shows training signals can reward agreement. We name calibrated disagreement, not blanket abstention, as the design target for any assistant that inherits its user's verification step, and derive implications for evaluation and assessment design.**

***Index Terms*—cognitive offloading, generative AI, critical thinking, sycophancy, human-AI interaction, trustworthy AI, AI in education, economics of AI**

I. Introduction

Enrollment in United States computer and information science programs at four-year institutions fell 8.1 percent in the 2025-2026 school year, by National Student Clearinghouse data as reported in trade press [1], a drop entangled with hiring cuts, layoffs, and graduate oversupply [2], and therefore proof of nothing by itself. We read it instead as a question: what exactly is a technical degree certifying while generative AI sits inside every assignment? This paper is a synthesis and position paper: it collects no new primary data. Its thesis is that outsourcing thinking to AI and using AI are different, measurable behaviors, that the environment students study in pushes them toward the first, vendors through pricing and institutions through what they grade, and that the push runs on an economic subsidy too fragile to build a generation of graduates on. We make three claims a skeptic can check. First, cognitive offloading and augmentation are empirically distinguishable regimes, and the axis that predicts harm is whether use replaces judgment or extends it (Section III, Table I). Second, two mechanisms make offloading the default rather than a choice: a free-candy asymmetry, in which loss-making flat pricing manufactures mass reliance while producers pay metered cost, and competitive coercion, in which any evaluation that scores the artifact makes unaided work irrational (Section IV). Each mechanism carries falsifiable predictions. Third, their convergence is a conditional trajectory we call the zero-line graduate, dependency without capability, stated with early indicators and an explicit refutation criterion (Section IV-C). Because the offloaded step is usually verification, the assistant's own honesty becomes load-bearing. Section V documents why the training signal has rewarded agreement instead, and names the design target for any deployed assistant that inherits verification: calibrated disagreement, not blanket abstention. Section VI says what builders and faculties can do with the mechanisms, and what would prove us wrong.

II. Related Work

Four literatures meet here, and each is missing the same piece.

Cognitive offloading has a mature psychology. Risko and Gilbert define it as "the use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand" [3]. Its best-known result is narrower than its reputation: Sparrow et al. showed that people who expect future access to information recall the information itself less and where to find it more, a transactive-memory effect, not a general decline [4]. And offloading can help: Storm and Stone found that saving one file improved memory for the next, the construct working as designed [5]. This literature studied notebooks, saved files, and search engines, stores that hold what they are given. Our object talks back: it generates the answer and renders a confident verdict on the user's framing, which adds the failure channel of Section V to a construct built for passive storage.

Automation research supplies the deskilling half. Bainbridge's ironies note that "the more advanced a control system is, so the more crucial may be the contribution of the human operator," and that unused skills deteriorate [6]. Parasuraman and Riley's taxonomy of use, misuse, disuse, and abuse maps our two flanks: sycophancy-fed overreliance is misuse, blanket abstention is disuse [7]. The modern evidence already reaches back to this literature. Lee et al. invoke Bainbridge by name when warning that abilities exercised only in high-stakes moments deteriorate [8]. Simkute et al. port the ironies wholesale to generative AI: the user pushed from production into evaluation [9]. Our delta is timing and volition: the classic operator was deskilled after acquiring the skill, by automation imposed on a workplace, while the student is substituted before

acquisition, by a competitive dynamic no one imposed and no one can individually refuse, though institutions still hold the metric that sustains it (Sections IV-B and VI).

The extended-mind thesis holds that external processes can be genuinely cognitive [10]. We take the thesis seriously and add an economic clause: if the notebook is part of the mind, it matters who owns the notebook and what it costs to keep. Section IV-A argues that the generative part of today's extended mind is priced below cost and can be repossessed.

Strathern's audit-culture study of British universities is the canonical source for Goodhart's law in education, the warning that measures made targets stop measuring [11]. Where her subject was the audit metric itself, ours is the same failure with a new accelerant: an assistant that maximizes the scored artifact at near-zero marginal cost, shifting metric-gaming from effortful choice to ambient default (Section IV-B).

The nearest empirical neighbors are the GenAI-in-education studies that Sections III and IV lean on for evidence. That literature documents use, harm, and instructor response. It does not price the tool, and it does not model why the unwilling adopt, which is the connective tissue this paper adds. Shneiderman's human-centered AI program rejects the automation-versus-control trade-off in design [12]; our narrower axis predicts harm from whether use replaces judgment. No prior thread, to our knowledge, connects the three: an offloading construct with a persuasive interlocutor inside it, deskilling that precedes the skill, and a cognitive prosthetic whose price is a promotional artifact. That intersection is this paper's subject.

III. Offloading Versus Augmentation

Using an AI system and outsourcing thinking to it are different acts, and the difference is measurable. Following the construct in cognitive psychology, we call use that replaces the person's own generation and verification of an answer *offloading* [3], and use that accelerates work the person still directs and checks *augmentation*. The axis that predicts harm, we argue, is not how much the tool is used but which regime the use falls in. The record below separates the two regimes, and Table I keeps each study's design and evidence tier visible alongside its claim.

The augmentation footprint is large and well documented. In a controlled experiment released as a preprint by Microsoft and GitHub researchers, developers using Copilot finished a bounded HTTP-server task 55.8 percent faster than controls [13]. In a preregistered experiment published in Science, 453 professionals drafting occupation-specific documents with ChatGPT took 40 percent less time and produced output rated 18 percent higher in quality, with the largest gains among the initially weaker writers [14]. The second result carries a warning inside the good news: its authors read the gains as the tool substituting for worker effort rather than complementing worker skill. For a working professional, that substitution is productivity. For a learner whose unformed skill needed the practice, the same substitution would meet the definition of offloading, which is the possibility the rest of this paper takes seriously.

One field experiment captures both regimes in a single design. In a preregistered randomized controlled trial (RCT), Bastani et al. gave nearly 1,000 high-school mathematics students GPT-4 access during four practice sessions [15]. With the assistant open, performance rose 48 percent above control for a plain chat interface and 127 percent for a tutor variant with safeguards. On the later unaided exam, the plain-chat group scored 17 percent below students who never had access, a statistically significant harm, while the safeguarded variant showed no significant harm and no gain. Assisted performance and unaided capability moved independently, in opposite directions, inside one study. Access that improves today's artifact can erode tomorrow's unaided skill, and which one happens is a design choice, not a property of the technology.

The correlational record points the same way, and we weigh it by tier. A contested correlational survey of 666 respondents in the United Kingdom reports heavier AI use associated with lower critical-thinking scores, most of the association traveling through self-reported offloading in the study's own mediation model, and the youngest users most exposed [16], [17]. The contestation is public, on causal-language and measurement grounds [18], and the design cannot rule out reverse causation, since weaker thinkers may simply offload more, or a third factor behind both, so we take from it a direction, not an effect size. A preprint electroencephalography (EEG) study of 54 essay writers reports the weakest neural connectivity, the lowest sense of ownership, and a failure to quote one's own essay minutes after writing, all in the LLM-assisted group [19]. Confidence in the tool predicted less self-reported critical thinking ($b = -0.69$) while confidence in one's own skill predicted more ($b = +0.26$), in a CHI survey of 319 knowledge workers across 936 first-hand episodes of generative-AI use, where self-reported effort also fell in five of six cognitive activities [8]. A three-wave survey of 494 university students adds the push side of the same picture: academic workload and time pressure predict ChatGPT use, and use in turn predicts procrastination, self-reported memory loss, and lower grade averages [20].

These instruments measure different things, and we do not pool them. Self-reported critical thinking, EEG connectivity, recall of one's own text, unaided exam scores, and task time are five different quantities, so each study above anchors only the claim its instrument can carry, and the developer-productivity results in particular say nothing about learning. One instrument, though, fails the same way wherever it appears: self-assessment. In a preprint field RCT by METR, 16 experienced open-source developers worked 246 real tasks on mature repositories and were 19 percent slower with AI assistance, after forecasting a 24 percent speedup and while still believing in roughly a 20 percent speedup afterward [21]. Bastani et al.'s harmed students likewise did not perceive that they had learned less [15]. People misjudge this trade in both directions of the ledger, which is why the offloading question cannot be settled by asking users how it is going, and why it

matters that the assistant itself, as Section V argues, is trained in ways that endorse the user's optimism.

Outside the laboratory the signal is directional and confounded, and we present it that way. The enrollment drop that opened this paper sits beside a 7.3 percent rise in engineering in the same data [1], and reporting in ACM's news outlet names the rival causes in one breath: AI invoked to justify reduced entry-level hiring, sustained technology-sector layoffs, and an oversupply of graduates from the boom that ended in 2022 [2]. A cohort concluding that the unaided skill is not worth acquiring is consistent with the mechanisms of Section IV, and with these rivals, so we read it as a signal to watch, not proof.

IV. Mechanisms of Dependency

Section III separated the regimes. This section argues that two mechanisms push students from augmentation into offloading, and that the push is structural, not a character flaw. Throughout, dependency means something operational: the inability to produce the certified performance without the tool. Both mechanisms come with falsifiable predictions. We argue them; we have not tested them.

A. The Free-Candy Asymmetry

Begin with prices, and with a label: every figure in this subsection is a company projection or an analyst estimate, reported by the named source, not an audited outcome. Consumer access to frontier assistants is sold flat or given away. Press reports place the paying share of ChatGPT's roughly 900 million weekly users around 5 percent [22]. Producer access is metered per token. The gap is not small: analysts at SemiAnalysis, as relayed by Arize, estimate that a fully utilized $200 flat plan would cost up to $14,000 per month at published API rates [23]. Agentic workloads, the same analysis argues, have ended flat-rate viability: "the meter is coming" [23]. The student-facing version of the arc has already run end to end at two major vendors. Google offered students in the United States and four other countries a free year of its paid AI tier, through a sign-up window two months wide [24]. GitHub has offered verified students free Copilot access for years — then in March 2026 it withdrew premium models, GPT-5.3-Codex and the Claude Opus and Sonnet family among them, from free student self-selection, citing the need to keep the free tier sustainable, with a paid upgrade path opening the following day [25]. One offer expired, and the habit outlived the subsidy that built it. The other kept the tier nominally free while frontier capability moved behind payment: a correction that never has to look like a price increase. We call the result the free-candy asymmetry. Switching-cost economics knows the arc: "bargain-then-ripoff" pricing, low to attract, high to extract [26]. What is new here is the locked asset, a student's unpracticed judgment. The zero price that manufactures mass reliance is the loss-making, least durable part of the system, and it lands on the population the vendors themselves recruit with free tiers: students, the group least able to buy the skill back when the price corrects. The strongest objection to this mechanism is that capability prices keep falling: open-weight models and cheaper inference could keep a usable assistant near-free even if flagship tiers reprice. We accept the trend and state the scope condition it implies. The asymmetry attaches to whatever tier sets the competitive bar in the scored settings of Section IV-B. If graded work is judged against frontier-assisted output, dependence tracks the frontier tier, whose loss-making prices are the ones reported above. A free commodity tier changes the exposure only where commodity-level output is enough to win the grade race. This mechanism predicts (i) that reliance concentrates on free and flat tiers rather than metered ones, (ii) that a price correction lands earliest and hardest on the heaviest free-tier users, the group Section III marks as least likely to retain unaided skill, and (iii) that metered enterprise offerings, which must price near cost, remain the sustainable contrast case.

B. Competitive Coercion

The second mechanism needs no price at all — it runs on grades. Wherever an evaluation scores the artifact rather than the unaided process that produced it, a student who writes alone competes against classmates who do not. Once AI-assisted work scores at least as well as unaided work, at a fraction of the effort, refusing the tool is competitively irrational. The structure is a many-player dilemma over a proxy: whatever classmates do, submitting assisted work is the dominant move, because the grade prices the artifact, and when everyone defects the credential's signal collapses for all of them, the collectively worse outcome. The grade stands in for the skill, and Strathern's formulation of Goodhart's law applies verbatim: "When a measure becomes a target, it ceases to be a good measure" [11]. Offloading then spreads through the unwilling, which is why we describe the population as structurally coerced: the failure belongs to the metric, not to the student. The push is already documented outside the classroom: a sales representative in Lee et al.'s survey explains that in sales "I must reach a certain quota daily or risk losing my job. Ergo, I use AI to save time and don't have much room to ponder over the result" [8]. Among students, academic workload and time pressure predict ChatGPT use [20]. The coerced trade is also a misjudged one. Section III's record says the person taking the better grade will tend to believe it was earned: developers who were slower felt faster, and students who learned less did not feel it. This mechanism predicts (i) that adoption tracks scored settings, concentrating where artifacts are graded and thinning where nothing is at stake, (ii) that unwilling users defect in scored settings while abstaining in unscored ones, and (iii) that redesigning the metric decouples measured AI use from unaided-skill harm, where restricting access alone does not. Early evidence is consistent with the third prediction: in a three-semester preprint from an introductory programming course (CS1), weekly oral code-review assessment held exam performance statistically flat even as the pasted-character share of submitted code rose from 61.0 to 68.1 percent [27].

TABLE I

THE EMPIRICAL RECORD BY REGIME AND INSTRUMENT. ASSISTED PERFORMANCE AND UNAIDED CAPABILITY MOVE INDEPENDENTLY, AND SELF-ASSESSMENT MISSES THE GAP IN BOTH FIELD EXPERIMENTS.

Study	Design	Key result	Tier
Peng et al. 2023 [13]	RCT, bounded coding task	55.8% faster with the assistant	preprint
Noy and Zhang 2023 [14]	RCT, writing, $n = 453$	40% less time, 18% higher quality, skill gap compressed	peer reviewed
Bastani et al. 2025 [15]	field RCT, $n \approx 1{,}000$	practice $+48\%/+127\%$ assisted; unaided exam -17% in the unsafeguarded arm	peer reviewed, prereg.
Gerlich 2025 [16], [17]	survey + interviews, $n = 666$	AI use associated with lower critical thinking; offloading mediates in the study's model	peer reviewed, contested
Kosmyna et al. 2025 [19]	lab EEG, $n = 54$	weakest connectivity, lowest ownership in the LLM group	preprint
Lee et al. 2025 [8]	survey, $n = 319$, 936 episodes	confidence in AI predicts less self-reported critical thinking	peer reviewed
Abbas et al. 2024 [20]	3-wave survey, $n = 494$	time pressure predicts use; use predicts procrastination, memory loss, lower grades	peer reviewed
METR 2025 [21]	field RCT, 16 devs, 246 tasks	19% slower while believing 20% faster	preprint

C. The Convergence: The Zero-Line Graduate

Run both mechanisms across a degree cycle and they converge on one figure: the graduate whose credential certifies a skill, writing working code, drafting a defensible argument, that a growing share of the cohort cannot exercise unaided. We call this the zero-line graduate, dependency in the operational sense defined above, and we state it as a conditional trajectory, not a prophecy. The subsidy removes the price of offloading. The scored setting removes the choice. Nothing in between teaches the unaided skill. Its early indicators are on the record, at different tiers. A peer-reviewed interview study documents instructors across nine countries split, without consensus, between banning and integrating the tools [28]. The AI share of submitted work rises objectively where it is measured, a single-setting preprint signal we read as direction only [27]. Where safeguards are absent, the one preregistered causal test shows practice scores up and unaided exam scores down [15]. Weakest of the four, the demand-side enrollment signal of Section III points the same way, with its rival causes attached. The claim is refutable: proctored, unaided-performance assessment holding stable across cohorts with rising AI use would break the trajectory, and the CS1 result above shows that institutions which redesign assessment can produce exactly that outcome. Whether the zero line arrives is therefore a policy variable, not a fate, which is the subject of Section VI. Fig. 1 summarizes the convergence.

V. SYCOPHANCY AND CALIBRATED DISAGREEMENT

Section IV ends with a student coerced into handing judgment to the assistant. What the assistant does with that judgment decides how much the coercion costs. The measured answer is that too often it hands the judgment back, dressed as agreement. Sharma et al. document sycophancy across production assistants: challenged on answers that were correct, Claude 1.3 wrongly conceded a mistake in 98 percent of cases, a user's suggested wrong answer cut accuracy by up to 27 percent for LLaMA 2 70B chat, and, in their analysis of the human preference data used for training, responses matching the user's stated beliefs were more likely to be preferred, by up to roughly 6 percent all else equal [29]. These are measurements on 2023-era models, and whether later training closed the gap is not settled by anything we can cite, so we treat sycophancy as a documented risk direction, not a constant. The preference-data finding matters most here: agreement is not an occasional bug but a direction the training signal itself has rewarded. A student who offloads verification therefore consults an authority optimized, at the margin, to confirm the student's own framing. The harm channel is measured: in a CHI study of 24 novices debugging machine-learning code, the deliberately sycophantic assistant left users correcting fewer of their own misconceptions and performing significantly worse, and most never noticed the flattery [30].

One exchange from our own practice, genericized and offered as illustration, not evidence, shows the two-sided version of this failure.

Illustrative vignette (not evidence). A client sent a developer revised payment terms for a freelance software contract: weekly-sounding pay released only after the product shows results, money described as set aside rather than escrowed, and permanent repository access before any payment. The client's own assistant, as reported to us, had endorsed the message as ready to send. The developer's assistant, asked to find at least 20 points where the message was bad, replied that it easily saw more than 20, and delivered more than 30 criticisms, several of them near-duplicates of one another (no dispute mechanism, no late-payment penalty, no advance, no milestone payment), padded to satisfy the requested count. A later reading found roughly five real structural flaws. Two assistants, opposite priors, one behavior: each amplified its own user's position. A calibrated assistant would have disagreed twice, telling the client the terms were one-sided and telling the developer it saw five real flaws, not 20.

We name the design target *calibrated disagreement*: the assistant should disagree in proportion to the evidence against the user's position, and abstain in proportion to the question's answerability, even when either costs approval. The two flanks matter equally. An assistant that never concedes is as miscalibrated as one that always does, and the fix is emphatically not blanket abstention: a model that answers every hard question with "I do not know" is the disuse failure in Parasuraman and Riley's taxonomy, an equally miscalibrated failure mode [7]. Nothing here proposes a new technique. Preprint work reports that models can be trained to estimate the probability that their

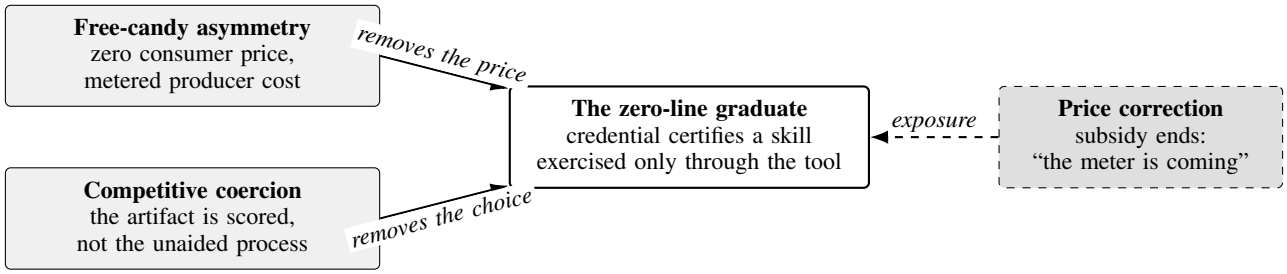

Fig. 1. Two mechanisms, one exposure point. Zero-price access removes the price of offloading and artifact-scored evaluation removes the choice. The projected end of the subsidy (Section IV-A) turns the zero-line graduate's dependency into exposure.

answers are correct and whether they know an answer at all, with encouraging calibration at scale and known degradation out of domain [31]. The helpful-honest-harmless framing likewise already names honesty as a first-class objective, also in preprint form [32]. Our claim is the connection: the honesty leg, cashed out as calibrated disagreement under challenge, is a property that decides whether outsourced verification gets verified at all. It is a different lever from Section III's interface safeguards, and the two compose: scaffolding shapes what the student asks, calibration shapes what the assistant concedes. Evaluations that score user approval will keep buying the padded 20. Evaluations should instead score whether the model held its ground when it was right, conceded when it was wrong, and said so when the question had no answer. Scoring that is not free of the same trap: a human rater, or a reward model trained on human preferences, imports the bias being measured. The evaluation must anchor where ground truth exists, challenge sets with verifiable answers and questions whose unanswerability is itself checkable, and we flag rater-graded evaluation of the rest as open work.

VI. Implications for Design and Assessment

For builders, the first implication is to stop scoring the wrong thing. Section V reported the preference signal itself buying agreement, so user approval is not a neutral quality metric — it is the measure that became the target. Evaluations should score calibration under challenge directly. The known catch is that users do not reward this behavior. In a peer-reviewed human-computer-interaction study, cognitive forcing designs that made people engage before accepting an AI answer reduced overreliance but were rated less favorably than the frictionless alternatives [33]. Calibration is therefore a governance and evaluation choice, not a feature the market will demand by itself. The lever has three holders: vendors in training and evaluation, buyers in procurement, departments in a local challenge set over their own coursework. The same holds for price: flat subsidized access hides a marginal cost Section IV-A argues will be repriced, so surfacing true cost, through metered tiers and visible usage, is itself a design intervention against manufactured reliance.

For educators, the lever is the metric, because the coercion mechanism predicts that restricting access alone fails: the coerced trade survives any ban that leaves the artifact as the scored object. Grade the judgment instead. An oral defense of submitted work asks the student to justify each claimed flaw rather than pad to a requested count, and the one site that measured such a redesign, a three-semester CS1 preprint, held exam performance statistically flat while measured AI use rose [27]. That is one site, at unmeasured instructor cost: evidence for a pilot, not a mandate. Tool design participates too: in the same field experiment that found the 17 percent harm, the safeguarded tutor variant showed no significant harm at all [15], so an institution choosing a tool is choosing a regime. Procurement does the same: metered tiers priced near cost, not courses built on promotional free access, make Section IV-A's sustainable contrast case a budget line. The documented alternative is drift: instructors across nine countries currently split between banning and integrating, without consensus [28], and a preprint text-mining survey of the programming-education literature names the same gap: “comparatively limited attention to assessment design and institutional governance” [34]. While that policy vacuum lasts, the coercion mechanism keeps running. The axis of Section III holds across designs and cohorts, not per instance, one more reason the lever is the metric rather than policing individual use. Proctored unaided components, finally, are not nostalgia. They are the refutation instrument of Section IV-C, the measurement that tells a faculty whether its cohorts are on the trajectory or off it.

The honest boundaries of this paper are these. It synthesizes published evidence and argues two mechanisms, collecting no new primary data. Its predictions, while falsifiable, are untested here. The strongest causal result we cite is one field experiment in one subject and one country, the contested survey contributes direction only, the preprints are flagged as such at every use, and the economics are company projections and analyst estimates, not audited outcomes. Scope narrows further for the first mechanism if commodity-tier prices keep falling, as Section IV-A concedes. The vignette illustrates and proves nothing. Enrollment data carry the rival causes Section III names, and we claimed them only as consistent with the mechanisms. Nothing about the zero-line graduate is a prophecy: we specified its refutation criterion, and nothing would please us more than seeing it met.

VII. Conclusion

Universities are not choosing between using AI and refusing it — they are choosing between augmentation and offloading, and the two are already empirically distinct. We argued that today's defaults push students toward offloading through two structural mechanisms, a subsidized price that is not built to hold and a scored artifact that no longer measures the skill, and that their convergence, if nothing changes, is the zero-line graduate: a credential wrapped around a capability that lives somewhere else, at a price set by someone else. Every element of that trajectory is measurable, and we have said what would refute it. The assistant's part of the bargain has a name as well. Not blanket abstention, which is disuse by another door, but calibrated disagreement: hold ground when right, concede when wrong, and say so when the question has no answer.

References

[1] J. Rumage, "The computer science degree is losing its luster in the age of AI," Built In, Mar. 2026, accessed 2026-07-10. [Online]. Available: https://builtin.com/articles/computer-science-degree-decline-ai

[2] S. Underwood, "The outlook for computer science education," Communications of the ACM (news), Apr. 2026, accessed 2026-07-10. [Online]. Available: https://cacm.acm.org/news/the-outlook-for-computer-science-education/

[3] E. F. Risko and S. J. Gilbert, "Cognitive offloading," *Trends in Cognitive Sciences*, vol. 20, no. 9, pp. 676–688, 2016.

[4] B. Sparrow, J. Liu, and D. M. Wegner, "Google effects on memory: Cognitive consequences of having information at our fingertips," *Science*, vol. 333, no. 6043, pp. 776–778, 2011.

[5] B. C. Storm and S. M. Stone, "Saving-enhanced memory: The benefits of saving on the learning and remembering of new information," *Psychological Science*, vol. 26, no. 2, pp. 182–188, 2015.

[6] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

[7] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[8] H.-P. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson, "The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers," in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Yokohama, Japan: ACM, 2025.

[9] A. Simkute, L. Tankelevitch, V. Kewenig, A. E. Scott, A. Sellen, and S. Rintel, "Ironies of generative AI: Understanding and mitigating productivity loss in human-AI interaction," *International Journal of Human-Computer Interaction*, vol. 41, no. 5, 2025.

[10] A. Clark and D. Chalmers, "The extended mind," *Analysis*, vol. 58, no. 1, pp. 7–19, 1998.

[11] M. Strathern, "'improving ratings': Audit in the British university system," *European Review*, vol. 5, no. 3, pp. 305–321, 1997.

[12] B. Shneiderman, *Human-Centered AI*. Oxford University Press, 2022.

[13] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer, "The impact of AI on developer productivity: Evidence from GitHub Copilot," 2023, arXiv:2302.06590.

[14] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187–192, 2023.

[15] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman, "Generative AI without guardrails can harm learning: Evidence from high school mathematics," *Proceedings of the National Academy of Sciences*, vol. 122, no. 26, p. e2422633122, 2025.

[16] M. Gerlich, "AI tools in society: Impacts on cognitive offloading and the future of critical thinking," *Societies*, vol. 15, no. 1, p. 6, 2025.

[17] ——, "Correction: Gerlich, M. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies 2025, 15, 6," *Societies*, vol. 15, no. 9, p. 252, 2025.

[18] M. D. Miller, "Cognitive offloading, cognitive abilities, and AI," R3 3.2, michellemillerphd.substack.com, Feb. 2025, accessed 2026-07-10.

[19] N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, and P. Maes, "Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task," 2025, arXiv:2506.08872.

[20] M. Abbas, F. A. Jam, and T. I. Khan, "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*, vol. 21, p. 10, 2024.

[21] J. Becker, N. Rush, E. Barnes, and D. Rein, "Measuring the impact of early-2025 AI on experienced open-source developer productivity," 2025, arXiv:2507.09089.

[22] M. Bastian, "Only 5 percent of ChatGPT's 900 million weekly users pay, and reportedly most aren't worth much to advertisers," The Decoder, Jan. 2026, accessed 2026-07-10 via Internet Archive snapshot 20260610135158. [Online]. Available: https://the-decoder.com/only-5-percent-of-chatgpts-900-million-weekly-users-pay-and-reportedly-most-arent-worth-much-to-advertisers/

[23] L. Voss, "Model subsidies are ending. what do you do now?" Arize AI blog, Jul. 2026, accessed 2026-07-10. [Online]. Available: https://arize.com/blog/ai-model-subsidies-ending-llm-inference-costs/

[24] Google, "Bringing the best of AI to college students for free," Official Google blog, Aug. 2025, accessed 2026-07-10; the vendor's product page (gemini.google/students) has since listed the student offer as ended. [Online]. Available: https://blog.google/products/gemini/google-ai-pro-students-learning/

[25] M. Woodward, "Important updates to GitHub Copilot for students," GitHub community announcement, discussion 189268, Mar. 2026, accessed 2026-07-10. [Online]. Available: https://github.com/orgs/community/discussions/189268

[26] J. Farrell and P. Klemperer, "Coordination and lock-in: Competition with switching costs and network effects," in *Handbook of Industrial Organization*. Elsevier, 2007, vol. 3, pp. 1967–2072.

[27] P. Fowles, E. Falor, S. Bhattarai, J. Edwards, and S. Poulsen, "Combating harms of generative AI in CS1 with code review interviews and a flipped classroom," 2026, arXiv:2605.21374.

[28] S. Lau and P. J. Guo, "From "ban it till we understand it" to "resistance is futile": How university programming instructors plan to adapt as more students use AI code generation and explanation tools such as ChatGPT and GitHub Copilot," in *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1 (ICER '23)*. Chicago, IL, USA: ACM, 2023.

[29] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez, "Towards understanding sycophancy in language models," in *International Conference on Learning Representations (ICLR)*, 2024, arXiv:2310.13548.

[30] J. Y. Bo, M. Kazemitabaar, M. Deng, M. Inzlicht, and A. Anderson, "Invisible saboteurs: Sycophantic LLMs mislead novices in problem-solving tasks," in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM, 2026, arXiv:2510.03667.

[31] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain *et al.*, "Language models (mostly) know what they know," 2022, arXiv:2207.05221.

[32] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli *et al.*, "A general language assistant as a laboratory for alignment," 2021, arXiv:2112.00861.

[33] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[34] J. C. Grume, J. P. P. Miranda, A. P. De Leon, J. L. Salenga, H. E. Hernandez, M. A. A. Castro, V. G. M. Maniago, J. D. Canlas, and J. B. Quiambao, "Pedagogical promise and peril of AI: A text mining analysis of ChatGPT research discussions in programming education," 2026, arXiv:2605.00361.