Why Johnny Can't Think: GenAI's Impacts on Cognitive Engagement

RUDRAJIT CHOUDHURI, Oregon State University, USA
CHRISTOPHER A. SANCHEZ, Oregon State University, USA
MARGARET BURNETT, Oregon State University, USA
ANITA SARMA, Oregon State University, USA

Context: Many students now use generative AI (genAI) in their coursework, yet its effects on their intellectual development remain poorly understood. While prior work has investigated students' cognitive offloading during episodic interactions, it remains unclear whether using genAI *routinely* is tied to more fundamental shifts in students' thinking habits.
Objective: To explore this possibility, we investigate (RQ1-How): how students' trust in and routine use of genAI affect their cognitive engagement—specifically, reflection, the need for understanding, and critical thinking in STEM coursework. Further, we investigate (RQ2-Who): which students are particularly vulnerable to these cognitive disengagement effects.
Method: We drew on dual-process theory, cognitive offloading, and the automation bias literature to develop a statistical model explaining how and to what extent students' trust-driven routine use of genAI affected their cognitive engagement habits in STEM coursework, and how these effects differed across students' diverse cognitive styles. We empirically evaluated this model using Partial Least Squares Structural Equation Modeling on survey data from **299** STEM students across five North American universities.
Results: Students who trusted and routinely used genAI reported significantly lower cognitive engagement. Unexpectedly, students with higher technophilic motivations, risk tolerance, and computer self-efficacy—traits often celebrated in STEM—were more prone to these effects. Interestingly, students' prior experience with genAI or academia did not protect them from cognitively disengaging.
Implications: Our findings suggest a potential *cognitive debt cycle* in which routine genAI use progressively weakens students' intellectual habits, potentially driving over-reliance and escalating usage. This poses critical challenges for curricula and genAI system design, requiring interventions that actively support cognitive engagement.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Generative AI, Overreliance, Critical Thinking, Reflective Thinking, Need For Understanding, Cognitive Styles, STEM Education

1 Introduction

> *"People are neither fully rational nor completely selfish — and they are often lazy thinkers." — Daniel Kahneman*

Human cognition evolved to be lazy and for good reason. Our ancestors who conserved mental energy for genuine threats survived; those who exhaustively analyzed every decision did not. This "law of least effort" pervades human behavior [66, 109, 126]. We outsource memory to shopping lists, arithmetic to calculators, and navigation to GPS, trading internal cognitive work for external support. Generative AI (genAI) tools (e.g., ChatGPT, Claude, Copilot) tend to amplify this tendency by delivering fluent, confident support on demand. As a result, interacting with genAI involves less "thinking by doing" and more "choosing from outputs" [77]. In many domains, this reallocation saves time, improves productivity, and creates a compelling sense of progress [19].

But what works for productivity can be problematic for learning. In STEM education, for example, genAI may remove the desirable difficulties students need [9]: slow, effortful, and often uncomfortable cognitive work that builds intuition and transferable skills [23]. Students may still produce equal (or even better) quality work yet learn less [137]. Recent

Authors' Contact Information: Rudrajit Choudhuri, Oregon State University, Corvallis, OR, USA, choudhru@oregonstate.edu; Christopher A. Sanchez, Oregon State University, Corvallis, OR, USA, christopher.sanchez@oregonstate.edu; Margaret Burnett, Oregon State University, Corvallis, OR, USA, burnett@eecs.oregonstate.edu; Anita Sarma, Oregon State University, Corvallis, OR, USA, anita.sarma@oregonstate.edu.

studies indeed show that genAI assistance reduced depth of inquiry, evidence use, and self-regulatory behaviors in students [9, 48, 73, 74, 127, 141].

Moreover, automation bias research shows that when systems appear competent, users trust outputs with limited scrutiny—even when those outputs conflict with their own knowledge or judgment [53, 87]. In human-AI interaction (HAI) contexts, this tendency can manifest as over-reliance [26, 61], with users incorporating AI suggestions without much critical evaluation [16, 77]. These effects are amplified in settings that reward immediate efficiency, including professional work [19, 77], and increasingly academic contexts [137], where genAI use is becoming routine.

Consider these two scenarios:

- **Scenario A (episodic)**: A student uses genAI to bypass a single difficult problem, showing momentary *overreliance* and *reduced cognitive effort*.
- **Scenario B (routine)**: Through trust-driven routine genAI use, that student develops a general unwillingness to *engage* in reflection, understanding, and critical thinking *across coursework*—fundamentally recalibrating what cognitive effort investment feels "worth it".

Prior work has largely focused on Scenario A [9, 21, 48, 73, 74, 127, 141]. In contrast, this paper focuses on Scenario B—how and to what extent *routine* use of genAI affects students' cognitive engagement habits. If such effects exist, does it mean that STEM education is cultivating a generation adept at querying intelligence but unwilling to exercise their own? Will they enter professional work having never developed the cognitive habits that prior generations built through practice? As a first step towards answering these long-term questions, we ask: **RQ1-How: *To what extent does students' trust in and routine use of genAI affect their cognitive engagement habits in STEM coursework?***

Drawing on decades of research in educational psychology and learning sciences [12, 15, 36, 54, 67, 81, 98], we conceptualize cognitive engagement as three intellectual habits: (1) *Reflection* (the metacognitive monitoring and evaluation of one's understanding) [39, 67, 120], (2) *Need for understanding* (the intrinsic motivation to seek conceptual coherence rather than surface-level correctness) [12, 20, 22], and (3) *Critical thinking* (the analytic evaluation of evidence, assumptions, and alternatives) [44, 45, 124] (detailed in Sec. 3.1). Foundational education research shows that these habits foster durable learning and skill development [14, 43]; RQ1-How investigates how/how much all three are simultaneously affected when students trust and routinely offload cognitive work to genAI.

However, these effects are unlikely to be the same for all individuals. Prior work shows that differences in individuals' cognitive styles [18, 128] substantively shape the ways they use technology [5, 17, 90, 136]. (We expand upon the specific cognitive styles in Sec. 3.2). Further, recent HAI studies suggest that these differences, alongside trust, influence how individuals adopt genAI [25, 29]. Thus, we ask: **RQ2-Who: Who (i.e., students with which cognitive styles) are particularly prone to cognitive disengagement?**

Our study makes three contributions: (1) a theoretically grounded statistical model explaining (for RQ1-How) how students' trust-driven routine use of genAI affects their cognitive engagement habits in coursework, and (for RQ2-Who) who are particularly affected. We then empirically evaluate this model using Partial Least Squares–Structural Equation Modeling (PLS-SEM) on survey data from 299 STEM students across five North American universities. This model and its empirical evaluation also provide (2) a psychometrically validated survey instrument for measuring these constructs in student–genAI interaction contexts. Finally, we derive (3) implications and open challenges for educators and researchers seeking to support diverse individuals' cognitive engagement in genAI-augmented STEM education.

2 Related Work

The past few years have seen a rapid adoption of genAI in educational settings, prompting growing interest in how these systems shape students' learning processes and outcomes. In response, a substantial body of work examined the effects of genAI use on cognition, learning, and task performance across educational levels and domains [9, 21, 51, 73, 100, 104, 137].

Much of this work reports a consistent pattern: genAI supported short-term efficiency gains (e.g., faster ideation, improved writing fluency, note-taking, and problem-solving) while coinciding with weaker higher-order learning outcomes (e.g., creativity, metacognition, self-regulation) [21, 74, 137, 141]. For instance, in secondary education, Kreijkes et al. [74] found that students who relied on AI-generated notes perceived comprehension tasks as easier and expended less effort, yet demonstrated weaker retention than peers who generated their own notes.

Related findings were reported in higher education contexts. Several studies observed that while genAI use reduced perceived task difficulty and mental effort, it was also associated with diminished depth of inquiry, evidence use, reasoning quality, and self-regulatory behaviors [9, 48, 127]. For example, Barcaui [9] found that undergraduates given unrestricted ChatGPT access for learning performed substantially worse on assessments administered weeks later, attributing this decline to the bypassing of "desirable difficulties" (e.g., self-quizzing, note-taking) essential for consolidating knowledge. Similarly, Fan et al. [48] reported that students revising essays with ChatGPT showed reduced motivation to rely on source material, exhibiting "metacognitive laziness" when progress felt effortless.

Other work examined how these effects varied with how genAI was incorporated into learning activities [27, 103]. In one such study, Choudhuri et al. [27] found that software engineering students benefited when genAI was used to clarify concepts or resolve impasses following independent effort, but experienced breakdowns when it was used prematurely—during initial learning or for complex implementations without sufficient foundational understanding.

Beyond immediate task performance, several studies suggested that genAI use could create an "illusion of competence" and a sense of progress that outpaced actual understanding [143]. For example, Prather and colleagues [101, 102] found that students using genAI tools reported higher confidence and completed tasks more quickly (sometimes producing higher-quality artifacts), yet faced metacognitive difficulties (e.g., challenges in debugging, explaining their own solutions, or adapting when AI-generated suggestions failed) [100, 102].

Complementing these behavioral findings, Kosmyna et al. [73] examined neurophysiological correlates of cognitive offloading. In a controlled lab study, college students writing with genAI assistance showed reduced activation in brain networks associated with creativity and cognitive control, alongside weaker subsequent memory recall and a lower sense of ownership over their written contributions.

Notably, these dynamics extend beyond classrooms. Studies of genAI use among knowledge workers reported similar patterns of cognitive offloading [19, 77]. Trust in AI appeared to further amplify these effects: when systems were perceived as competent, users were more likely to accept outputs with limited scrutiny [16, 79]. For example, Lee et al. [77] found that workers with higher confidence in genAI engaged less in creativity and critical thinking during tasks, shifting from deliberate reasoning toward passive evaluation of AI-generated content. Broader population studies likewise linked frequent reliance on genAI with increased offloading, work-slop, and overconfidence [51, 91].

Overall, prior work has largely focused on episodic cognitive offloading and its immediate effects on learning and task performance. What remains underexplored is how routine genAI use during formative education reshapes students' enduring habits of cognitive engagement. Our study addresses this gap by examining how students' trust in and routine use of genAI affect students' reflection, need for understanding, and critical thinking in STEM coursework (*RQ1-How*). Further, while prior work has shown that variations in individuals' cognitive styles affect technology

adoption [5, 6, 25, 29], our work is the first, to our knowledge, to investigate how these individual differences condition students' vulnerability to genAI-related cognitive disengagement (*RQ2-Who*).

3 Foundations and Model Development

We draw on foundational work in HCI and cognitive psychology to develop a theoretical model explaining two pathways of influence (see Fig. 1): (1) how students' trust in and routine genAI usage affects their cognitive engagement habits (RQ1-How), while (2) accounting for variations in their cognitive styles (RQ2-Who).

3.1 Cognitive engagement

Consider what happens when you encounter a confusing passage in a textbook. One response is to highlight it and move on, trusting that re-reading will somehow clarify things later. Another response is to stop, notice the confusion, and deliberately work to resolve it: *"Wait—how does this claim follow from the previous one? What am I missing?"* The second response reflects *cognitive engagement*. It involves making sense of ideas, recognizing gaps in your own understanding, and investing effort to close them, rather than simply collecting information or reproducing facts [14, 50, 75].

Educational psychology has long characterized cognitive engagement habits across three core intellectual dimensions: (a) *Reflection*, (b) *Need for understanding*, and (c) *Critical thinking* [12, 15, 36, 54, 67, 81, 98]. Learners who cultivate these habits are more likely to develop durable understanding and transferable skills than those who rely on more superficial strategies [12, 54, 67, 98]. In this study, we therefore conceptualize cognitive engagement along these dimensions. Below, we define each dimension and describe how it is operationalized in our model.

Reflection is the metacognitive habit of monitoring one's own understanding, evaluating assumptions and strategies, and recognizing gaps in thinking [39]. For example, when you read that ocean temperatures are rising by 0.13°C per decade, do you catch yourself wondering whether that's fast or slow, and what it implies? Education reformer John Dewey described reflection as "overcoming the inertia that inclines one to accept suggestions at their face value" [36]. Research shows that reflective learners are better positioned to catch their own errors, recognize when they're stuck, and adjust learning strategies adaptively [39, 67, 120].

Reflection is a latent psychological construct (subjective, intangible, and not directly observable) and is therefore captured through validated self-report questionnaires [12, 67, 106]. In our study, we adopt Biggs' validated framework [12], which conceptualizes *Reflection* as a combination of *Reflective Motives* (why learners reflect) and *Reflective Strategies* (how they do so). Consistent with this framework, we model Reflection as a higher-order construct (HOC) composed of these two lower-order constructs (LOCs) [57], represented as circles in Fig. 1. Each LOC is itself latent and measured through multiple self-report items (questions); in our model, Reflective Motives and Reflective Strategies are measured using three and five items respectively (denoted as squares in Fig. 1).

Need for Understanding captures a learner's intrinsic drive to make sense of things [20, 22]. Some learners are satisfied once they can execute a procedure correctly. Others experience cognitive discomfort until they grasp *why* it works. This is not stubbornness or perfectionism—it is a stable difference in what feels intellectually satisfying [20]. Learners high in this trait habitually press for explanations and expend cognitive effort to construct meaning, not because they're told to, but because surface-level answers feel incomplete. Unsurprisingly, research shows they learn more effectively: their constructive sensemaking processes foster effective self-regulation that produces durable learning [23, 43]. Following prior work [12], we model Need for Understanding as a unidimensional construct, measured using items adapted from Kember's validated instrument [67] (Fig. 1).

Critical Thinking is the analytic work learners do in inspecting assumptions, weighing evidence, and considering

alternatives [44, 45, 124]. For example, when a solution fails, some may fixate on the most obvious cause. A critical thinker, on the other hand, considers multiple hypotheses: *Is the error due to my inputs, or an unstated assumption in the problem? What evidence would distinguish between these possibilities?*

Critical thinking has been assessed in multiple ways spanning self-report scales, performance tasks, observations, and expert evaluations [41, 44, 67, 70, 96, 144]. The APA Delphi consensus [44] characterizes critical thinking through two complementary dimensions: (1) *Critical Openness*, the willingness to entertain alternative perspectives and/or competing ideas, and interpret information from multiple angles; and (2) *Reflective Skepticism*, the discipline of questioning evidence and scrutinizing claims. Effective critical thinking requires both [42]: openness without skepticism produces gullibility; skepticism without openness produces rigidity. Accordingly, we model Critical Thinking as an HOC composed of these two LOCs (Fig. 1) [44–46], using a validated questionnaire adapted from [124].

We adopt this definition over alternatives such as Bloom's taxonomy [13] or the Paul-Elder framework [95] for three reasons: First, the Delphi-report formulation is grounded in empirical consensus among multiple experts in psychology and education. Second, its parsimonious two-factor structure (unlike the Paul-Elder framework [95], which spans eight "elements of thought", ten "intellectual standards", and eight "intellectual virtues") aligns with our survey design and structural modeling goals. Third, it lowers measurement complexity and multicollinearity (highly fragmented constructs become unwieldy [57]), yielding a model that is more straightforward to estimate and interpret.

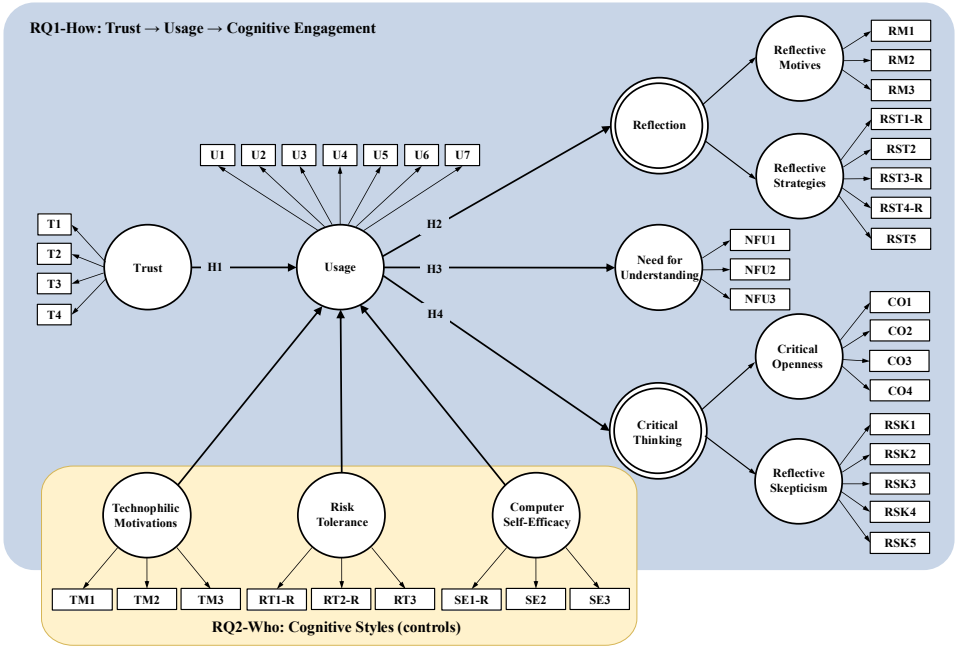

Fig. 1. Proposed theoretical model and measurement specification. The blue region (RQ1-How) models how *Trust* and routine *Usage* affect cognitive engagement: (a) *Reflection*, (b) *Need For Understanding*, and (c) *Critical Thinking*. The yellow region (RQ2-Who) controls for cognitive styles (*Technophilic Motivations*, *Risk Tolerance*, *Computer Self-Efficacy*). Latent constructs (circles) are reflectively measured by survey items (squares) [112]; higher-order constructs (double circles) are reflectively measured via their lower-order constructs (e.g., *Reflection* is measured via *Reflective Motives* & *Strategies*). Reverse-coded items carry the "-R" suffix (e.g., RT1-R). Directed arrows represent hypothesized relationships (H1–H4) and cognitive-style controls. Item wordings (questionnaire) are in [1].

3.2 Model development: How trust, usage, and cognitive styles affect cognitive engagement

To understand how genAI tools affect cognitive engagement, we must first understand something fundamental about how human cognition works—or more precisely, how it avoids working. These theories form the basis of our hypotheses:

3.2.1 ***The laziness of thought*** *(dual process and cognitive offloading theory).* Our minds did not evolve to think carefully about everything. That would be metabolically ruinous and strategically foolish [66, 109]. According to Kahneman's dual-process theory [66], human cognition operates through two systems: System 1 (*"fast thinking"*) is intuitive, automatic, and low-effort, drawing rapidly on associations and patterns. For example, it is what delivers the answer *"2 + 2 = ?"* without deliberation, or completes the phrase *"bread and..."*. System 2 (*"slow thinking"*), on the other hand, is deliberate, controlled, and effortful. It is invoked when, for example, you calculate 17 × 24, evaluate a complex argument, or consciously monitor your behavior in a social situation.

System 2 is lazy. It avoids effort whenever System 1 can provide a plausible solution. This is not a bug; it is an evolutionary feature. Our ancestors who conserved cognitive resources for genuine threats survived; those who exhaustively analyzed every decision were outcompeted. The result is what psychologists call the *"law of least effort"* [66]—a pervasive tendency to accept the path of least cognitive resistance.

This explains why cognitive offloading is so natural [109]. When external aids can shoulder cognitive demand, we instinctively let them—*why waste cognitive resources solving problems that have already been solved?* In learning contexts, however, it can cause students to skip over mental processes that are integral to the formation of knowledge. In the past, we have seen the *"Google effect"* [126]: people became less likely to remember information they believed would be available online. Recent HAI work shows similar offloading patterns with genAI, where users lean on AI-generated suggestions at the expense of their own thinking [73, 77].

3.2.2 ***Automation bias.*** The tendency of cognitive offloading is compounded by trust [53, 87]. When automated systems appear competent, users may accept outputs with limited scrutiny—even when those outputs conflict with their own knowledge or judgment [53, 87]. In HAI contexts, this manifests as over-reliance: users defer judgment to AI, rapidly integrating its suggestions while failing to consider the likelihood or occurrence of errors/problems [16, 61, 78].

3.2.3 ***From laziness of thought and automation bias to RQ1-How's hypotheses.*** We posit, drawing on dual-process, cognitive offloading, and automation bias theories, a set of cascading relationships through which students' trust in and routine use of genAI substantively affect their cognitive engagement habits in coursework.

First, *trust drives routine usage.* We adopt a widely used definition of trust in AI: *"the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"* [78, 82, 135, 138]. Across automation and HAI research, trust has been shown to predict adoption [28, 78, 140]. Students who view genAI as reliable are likely to consult (and defer to) it more frequently for their work. Accordingly,

- **H1:** *Trust positively affects routine genAI usage.*

Second, we hypothesize that *routine usage affects cognitive engagement habits.* Repeated offloading to genAI reinforces a new System-1 habit: *"problem → offload to genAI"*, a fast response that bypasses deliberate cognitive engagement:

- **H2a.** *Routine genAI usage negatively affects Reflection.*
- **H3a.** *Routine genAI usage negatively affects the Need for Understanding.*
- **H4a.** *Routine genAI usage negatively affects Critical Thinking.*

Third, we posit that *trust acts as an accelerant*, increasing the readiness with which students defer to genAI, which in turn trains the System 1 response described above and erodes cognitive engagement habits. Accordingly:

- **H2b.** *Trust transitively reduces Reflection through increased routine usage.*
- **H3b.** *Trust transitively reduces the Need for Understanding through increased routine usage.*
- **H4b.** *Trust transitively reduces Critical Thinking through increased routine usage.*

Following prior work, we adapt validated instruments to measure students' trust (4-item TXAI instrument [28, 97]) and routine use of genAI (academic genAI use instrument [2]).

3.2.4 ***Beyond how's to who's (RQ2-Who).*** Individuals are not identical in their propensity to use tools. Instead, they differ in their *cognitive styles*, i.e., *the way they perceive and process information, problem solve, and work with technology* [18, 90, 128], which affects their technology adoption [5, 6, 17, 136]. Recent HAI work shows that certain cognitive styles: *technophilic motivations* (intrinsic interest in engaging with technology), *risk tolerance* (comfort with uncertainty), and *computer self-efficacy* (confidence in using technology) significantly predict genAI adoption [25, 29].

Thus, to avoid conflating these stable traits with the effects of trust modeled in RQ1-How, we include technophilic motivations, computer self-efficacy, and risk tolerance as **control variables** on genAI usage: they help explain baseline differences in *who* is most likely to heavily adopt genAI [28], and in turn, may be prone to cognitive disengagement. We capture these cognitive styles using the validated GenderMag facet survey [60], following prior work [28].

Finally, given evidence that academic level and prior genAI experience can influence usage [27, 121, 133], we performed **group analyses** to test whether our model relationships varied across these groups.

4 Method

4.1 Survey design

To address our RQs, we conducted a large-scale survey with STEM students across five North American institutions. The study was approved by the University IRB. We followed Kitchenham's guidelines for designing empirical surveys [68], and drew on established frameworks from educational psychology and Human–AI Interaction studies to measure our constructs (Tab. 1). Each construct in our model (circles in Fig. 1) represents a latent variable, i.e., a variable not directly observable but inferred through validated indicators (e.g., survey questions, squares in Fig. 1). We adapted existing questions to the genAI context, where applicable. For example, a risk-tolerance item reading *"I avoid advanced [technology] features or options"* was revised to *"I avoid advanced genAI features or options"* (see supplemental [1]).

Our survey comprised three sections:

(1) **Cognitive engagement.** After providing informed consent, participants reported their cognitive engagement habits in coursework, including their reflection (reflective motives and strategies), need for understanding, and critical thinking (critical openness and reflective skepticism), captured using validated instruments listed in Tab. 1.

(2) **Trust, usage, and cognitive styles.** Participants reported their trust in and usage of genAI tools in coursework, alongside their cognitive styles, including risk tolerance, technophilic motivations, and computer self-efficacy (Tab. 1).

(3) **Background and genAI experience.** Finally, participants rated their prior experience with genAI, academic level, and (optionally) their major, age, and gender (see supplemental material [1]). During the pilot phase, participants were also invited to share open-ended feedback on the survey.

We administered the survey in Qualtrics [105]. All closed-ended questions used a five-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*) with a neutral option and a sixth "*I'm not sure*" option to distinguish ignorance from

Table 1. Constructs in the theoretical model and validated instruments used to measure them

Construct	Instrument
Trust	TXAI instrument [28, 97]
Usage	Academic genAI usage instrument [2]
Reflection	Bigg's instrument [12]
Need for Understanding	Kember's instrument [67]
Critical Thinking	CTDS instrument [124]
Cognitive Styles	GenderMag facet survey [28, 60]

indifference [55]. The survey took 7-10 minutes to complete. We included attention-check questions and randomized item order within blocks to reduce response bias. The survey questionnaire is available in the supplemental material [1].

Sandbox and pilot. We sandboxed the survey one-on-one with psychology/HCI/CS-education researchers ($n = 9$) and STEM instructors ($n = 4$) to assess interpretability, clarity, and realism; revising ambiguous question phrasing based on their feedback. We then piloted it with 67 STEM undergraduates, confirming item clarity and construct validity. We made minor wording edits and excluded pilot responses from the final analysis.

Validation. We psychometrically validated our survey instrument following [57, 112] (see details in Sec. 4.3).

4.2 Data collection

Distribution. The survey was available for three months (April-June, 2025). We recruited STEM instructors at our institution and via education conferences and networks, who then distributed the survey to students in their classes. We asked instructors to encourage participation and, where appropriate, offer it as an optional, non-graded exercise. Contacting instructors directly, rather than students, broadened our reach and aligned with our IRB guidelines.

Sample size determination. We determined the minimum sample size for validating the model via power analysis using the G*Power tool [49]. We specified an F-test with multiple linear regression to detect even small effect sizes ($f^2 = 0.10$) at a significance level of ($\alpha = 0.05$) with power = 0.95, and assuming at most four predictors per construct.[1] The analysis suggested a minimum sample size of 191.

Responses. We received a total of 392 responses. We removed patterned responses ($n = 20$), partial submissions ($n = 39$) and those that failed attention checks ($n = 34$). We treated "*I'm not sure*" selections as missing data, and did not impute missing values given the unproven efficacy of imputation in SEM group contexts [116]. We retained ***n*=299 valid responses** after filtration, which exceeded the *apriori* sample size requirement of 191. Respondents reported a wide distribution of genAI experience, usage frequencies, academic level, age, and gender, consistent with prior studies [77, 111]. A summary of respondent demographics is available in the supplemental material [1].

4.3 Data analysis & survey validation

We analyzed the survey data using *Partial Least Squares–Structural Equation Modeling* (PLS-SEM) to empirically evaluate our theoretical model. PLS-SEM suits exploratory studies like ours because it estimates multiple relationships among constructs while accounting for measurement errors in one comprehensive analysis [28, 111, 112, 134]. Importantly, it relaxes multivariate distributional assumptions. Instead, it calculates the statistical significance of path coefficients (i.e.,

[1] In Fig. 1, the construct with the most incoming arrows is Usage, predicted by (1) Trust, (2) Technophilic Motivations, (3) Risk Tolerance, and (4) Computer Self-Efficacy.

relationships between constructs) through *bootstrapping*: resampling thousands of subsamples (5,000 in our case) to derive inferences [58].

We conducted the analyses using *SmartPLS 4* [122] in two phases. First, we evaluated the *measurement model to validate our survey*: i.e., verify that the items (questions) reliably and validly capture the intended constructs. Second, we evaluated the *structural model* to test the hypothesized relationships among these constructs. Below, we detail the *measurement model* evaluation (**Steps 1-5**). The *structural model* evaluation and results are detailed in Sec. 5.

Survey validation *(measurement model evaluation)*: We used a two-stage procedure to validate our survey instrument, i.e., assess both lower-order and higher-order reflective constructs (circles and double circles in Fig. 1) following recommended PLS-SEM practices [10, 58, 115]. The data met standard assumptions for measurement-model analysis: Bartlett's test of sphericity was significant for all constructs ($\chi^2(946) = 7167.278$, $p < .001$), and the KMO measure of sampling adequacy was 0.901, exceeding the 0.60 threshold [65].

Stage 1: Lower-order reflective construct evaluation: We first assessed the lower-order reflective constructs (Trust, Usage, Reflective Motives, Reflective Strategies, Need for Understanding, Critical Openness, and Reflective Skepticism, circles in Fig. 1) using tests for **(1)** convergent validity, **(2)** internal consistency reliability, **(3)** discriminant validity, and **(4)** indicator collinearity [58, 112].

(1) Convergent validity examines whether indicators (questions) intended to measure the same construct share sufficient variance (for reflective constructs, changes in the construct should be reflected in its indicators) [71]. We assessed convergent validity using (a) factor loadings and (b) Average Variance Extracted (AVE) [58].

- Factor loadings capture the association between indicators and their constructs (e.g., T1–T4 for *Trust*); values $\geq .60$ suffice for exploratory studies [58]. We removed two weak indicators—U6 of *Usage* and CO1 of *Critical Openness*—that did not adequately reflect their constructs.[2] All retained indicators exceeded the threshold, with loadings ranging from .60 to .925.
- AVE captures the shared variance across a construct's indicators and should exceed .50, indicating the construct explains at least 50% of shared variance among its indicators [58]. All constructs met this criterion (Tab. 2).

(2) Internal consistency reliability evaluates whether indicators are internally consistent and reliably measure the same construct. We evaluated internal consistency reliability using Cronbach's α and Composite Reliability (ρ_a, ρ_c) [112]. All constructs met established reliability thresholds (.70–.95) [58] (see Tab. 2).

(3) Discriminant validity assesses the distinctness of constructs. We used the Heterotrait–Monotrait (HTMT) ratio [63], with adequacy indicated by values $< .90$. HTMT ratios among all constructs met this criterion, supporting discriminant validity. We report HTMT ratios for higher-order constructs and other model constructs in Tab. 3; ratios for the LOCs of Reflection and Critical Thinking, as well as the Fornell–Larcker criterion and cross-loadings, alongside confirmatory factor analysis, are provided in the supplemental for completeness [1].

(4) Indicator collinearity assessment checks for multicollinearity among reflective indicators. We used the Variance Inflation Factor (VIF) to check for multicollinearity. All indicator VIFs were < 3, well below the cutoff of 5 [58].

Common method bias. Since we collected data via a single survey, we tested for Common Method Bias (CMB) [112]. To do so, we first applied Harman's single-factor test on the latent constructs [99]. No single factor explained more than 21.9% variance. An unrotated exploratory factor analysis with a forced single-factor solution explained 26.6% of variance, well below the 50% threshold. Additionally, we used Kock's collinearity check [72]: VIFs for all latent constructs were < 2, well under the cut-off. These results indicate CMB was not a concern in our study.

[2]Removing CO1 and U6 also increased AVE from .48 to .52 for *Critical Openness* and from .64 to .67 for *Usage*.

Table 2. Internal consistency reliability, and convergent validity for reflective lower-order and higher-order constructs in the theoretical model. These metrics reflect whether each construct is measured reliably and captures sufficient shared variance among its indicators. Cronbach's α, Composite Reliability (CR(ρ_a), CR(ρ_c)), and Average Variance Extracted (AVE) values indicate internal consistency and convergent validity respectively, with acceptable thresholds of ≥.7 (for reliability) and ≥.5 (for AVE). Higher values reflect stronger reliability and validity of the questionnaire.

	Cronbach's α	*CR(ρ_a)*	*CR(ρ_c)*	*AVE*
Lower-order constructs				
Critical Openness	.746	.802	.735	.524
Need for Understanding	.902	.905	.938	.835
Reflection Motives	.896	.899	.937	.833
Reflective Skepticism	.828	.903	.845	.511
Reflective Strategies	.789	.797	.855	.542
Trust	.818	.831	.883	.655
Usage	.896	.897	.923	.665
Technophilic Motivations	.731	.772	.805	.685
Risk Tolerance	.721	.752	.773	.672
Computer Self-Efficacy	.711	.796	.821	.645
Higher-order constructs				
Critical Thinking	.700	.710	.803	.671
Reflection	.801	.809	.909	.834

Cronbach's α tends to underestimate reliability, whereas composite reliability (CR: ρ_c) tends to overestimate it. The true reliability typically lies between these two estimates and is effectively captured by CR(ρ_a) [112].

Table 3. Heterotrait–Monotrait (HTMT) ratios among model constructs (ranges between 0 and 1). Values $< .90$ indicate that constructs are sufficiently distinct from each other. HTMT ratios for the LOCs of Reflection (Reflective Motives, Reflective Strategies) and Critical Thinking (Critical Openness, Reflective Skepticism) are reported in [1].

	Critical Thinking	**Tech. Motivations**	**NFU**	**Reflection**	**Risk Tolerance**	**Computer S.E.**	**Trust**
Tech. Motivations	.597						
NFU	.07	.179					
Reflection	.66	.823	.322				
Risk Tolerance	.552	.735	.217	.669			
Computer S.E.	.587	.603	.063	.414	.635		
Trust	.529	.711	.266	.843	.608	.489	
Usage	.587	.661	.228	.806	.786	.526	.725

Tech. Motivations: Technophilic Motivations; NFU: Need for Understanding; Computer S.E.: Computer Self-Efficacy

Stage 2: Higher-order reflective construct evaluation: Our model includes two reflective higher-order constructs (HOCs): (1) Reflection and (2) Critical Thinking. Recall, reflection is a motive-strategy pair that acts together (see Sec. 3). Following Biggs et al. [12]'s work, we modeled *Reflection* as an HOC with *Reflective Motives* and *Reflective Strategies* as its lower-order constructs (LOCs). Similarly, following the APA Delphi consensus [44], we modeled *Critical Thinking* as an HOC comprising *Critical Openness* and *Reflective Skepticism*, as in prior work [124]. Collapsing the LOCs into HOCs reduces measurement complexity, aligns the model with theory, and lowers collinearity, yielding a more parsimonious, easier-to-test structure.

(5) HOC assessment. We validated HOCs using the disjoint two-stage approach recommended by Sarstedt and Cheah [114]. First, we estimated latent scores for the lower-order constructs in SmartPLS as linear combinations of indicator scores and outer weights. We then used these scores as indicators for the HOCs: (i) Reflective Motives and

Reflective Strategies for *Reflection*, and (ii) Critical Openness and Reflective Skepticism for *Critical Thinking*. Finally, we evaluated the HOCs using the same assessment criteria applied to the LOCs:

- *Convergent validity.* All factor loadings exceeded .70 (Fig. 3), and AVE values were above .50 for both HOCs (Tab. 2), indicating sufficient explained variance of its indicators (the LOCs).
- *Internal consistency reliability.* Cronbach's α and Composite Reliability values exceeded .70 for both HOCs, indicating sufficient reliability and internal consistency of the constructs (Tab. 2).
- *Discriminant validity*. HTMT ratios were below .90 for all construct pairs (Tab. 3), indicating empirical distinctness.
- *Collinearity.* All VIF values were below 2 (see supplemental material [1]), indicating no multicollinearity concerns.

Together, these stages (1&2) *validated our survey instrument*: questions reliably capture their intended constructs, and these constructs satisfy convergent and discriminant validity, enabling us to test structural relationships next (Sec. 5).

4.4 Limitations

As with any empirical study, our work has limitations [69]. All constructs were measured using validated self-report instruments grounded in established theory. Even then, surveys can introduce response biases or misinterpretations. We guarded against these threats by involving instructors in survey design, sandboxing and piloting, randomizing question blocks to reduce order effects, adding attention checks, and screening patterned responses. All constructs met standard thresholds for reliability and validity.

As a cross-sectional study [129], we report associations rather than causation. Self-selection bias remains possible, as students with stronger views may have been more likely to participate. As with all survey-based research, reported experiences may not align perfectly with in-situ behavior. Nonetheless, given the internal consistency of the results and their triangulation with established theory, we consider this approach sufficient to support the reliability of our conclusions. Further, a theoretical model cannot capture an exhaustive set of factors, i.e., unmeasured confounds may remain. Accordingly, our findings should be interpreted as a theoretically grounded starting point; future work should consider longitudinal designs and replication in varied contexts to strengthen causal claims and generalize broadly.

Finally, while the study was designed to be region-agnostic and included students with varied demographics and levels of genAI experience, the sample reflects a primary US-based educational context. Its size and composition are comparable to prior empirical studies in STEM education [4, 89], providing a reasonable foundation for theory development. Consistent with this goal, we emphasize theoretical rather than statistical generalizability [119]. Future work should examine transferability across broader populations and contexts.

5 Results

After validating our survey (measurement model), we test our hypotheses by evaluating the structural model (recall Fig. 1). From here on, when we say 'significant', we mean 'statistically significant'.

5.1 RQ1-How/How much: Trust → Usage → Cognitive Engagement?

RQ1-How examines whether and to what extent students' *trust in genAI predicts their routine usage of these tools*, and, in turn, *how that trust and usage affect their cognitive engagement habits* in STEM coursework—specifically reflection, need for understanding, and critical thinking (recall Sec. 3 for how we model these constructs). We report these findings in a deliberate sequence: we first describe how routine genAI usage affects cognitive engagement, and then turn to how trust affects these outcomes.

5.1.1 ***Usage and cognitive engagement***. Fig. 2 shows distributions of cognitive engagement by usage quartile (Q1 = lowest 25%; Q4 = highest 25%).[3] Across all three dimensions, median cognitive engagement factor scores declined monotonically from Q1 to Q4. Participants reporting the highest levels of genAI use (Q4) consistently reported the lowest levels of reflection, need for understanding, and critical thinking. Tab. 4 reports the statistical significance of these relationships (visualized as directed paths among constructs in Fig. 3. E.g., Usage→Reflection is H2).

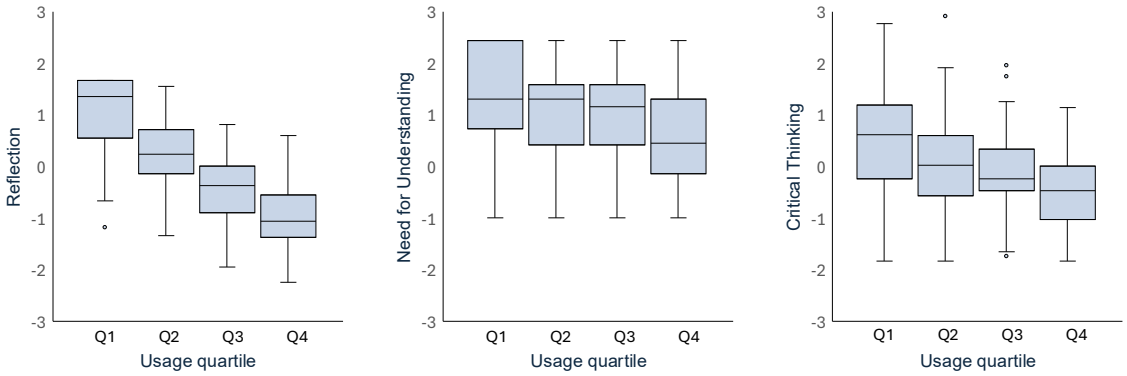

Fig. 2. Cognitive engagement scores by genAI usage. Boxplots show standardized latent scores for Reflection, Need for Understanding, and Critical Thinking across quartiles of reported genAI usage (Q1 = lowest 25%; Q4 = highest 25%). This figure provides a close-up view of data distributions; subsequent figures examine patterns using median-split groups.

Table 4. Effects of students' routine genAI usage on their cognitive engagement habits (H2a-H4a): Standardized path coefficients (β), standard deviations (SD), 95% confidence intervals (CI), p-values, and effect sizes (f^2). We consider $f^2 < 0.02$ no effect, $f^2 \in [0.02, 0.15)$ small, $f^2 \in [0.15, 0.35)$ medium, and $f^2 > 0.35$ large [31].

	B	SD	95% CI	p	f^2
H2a Usage→Reflection	−0.66	0.05	(-0.76, -0.55)	**0.000**	0.51
H3a Usage→Need for Understanding	−0.21	0.06	(0.09, 0.34)	**0.000**	0.12
H4a Usage →Critical Thinking	−0.41	0.02	(-0.46, -0.38)	**0.001**	0.20

Note: PLS-SEM estimates all structural paths jointly within a single model. Significance tests (and their Type-I error rates) are defined within this joint estimation; separate family-wise error corrections are not required [57, 58].

(H2a) Usage and reflection: GenAI usage had a significant negative association with reflection ($\beta = -0.66$, $p < .001$), with a large effect size ($f^2 = 0.51$), supporting **H2a**. Specifically, a unit (variance) increase in routine genAI usage predicted a 0.66-unit reduction in reflection, as indicated by the path coefficient (Tab. 4). Reflection requires monitoring oneself—noticing when something does not make sense or when understanding feels superficial [39, 120]. The magnitude of this association suggests that routine genAI usage substantively affected such process—the pausing, puzzling, and noticing that something requires deeper thought.

(H3a) Usage and need for understanding. GenAI usage had a significant negative association with need for understanding ($\beta = -0.21$, $p < .001$), supporting **H3a**. The effect size ($f^2 = 0.12$) suggests that while routine genAI use

[3]Each score is a model-estimated latent factor score: the measurement model computes how strongly each question signals the underlying construct, then uses these item weights to aggregate the factor score [57].

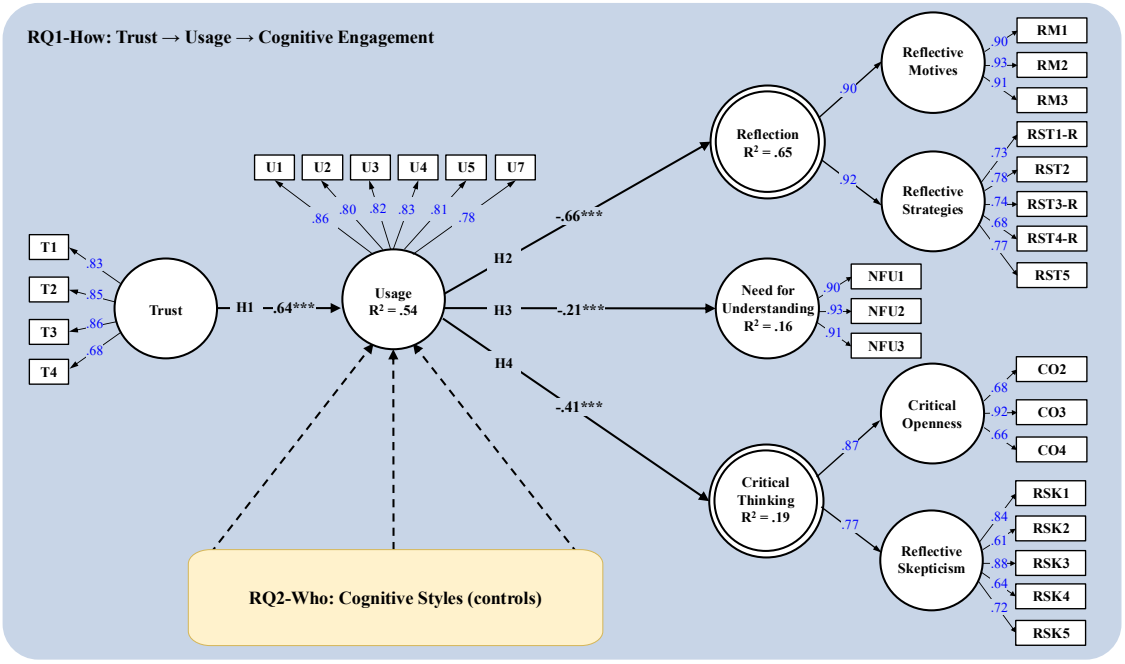

Fig. 3. Structural model results (RQ1-How): Standardized path coefficients (H1–H4) on arrows between constructs (circles) show how Trust and routine genAI Usage affect Reflection, Need for Understanding, and Critical Thinking, with significance ***p<.001, **p<.01, and *p<.05. Corresponding transitive effects are summarized in Tab. 5. Arrows from constructs to survey items (squares), and from HOCs (double circles) to their lower-order constructs, indicate factor loadings (in *blue*; all ≥ threshold of 0.6 [58]). Dotted arrows and the collapsed yellow region correspond to RQ2-Who results (omitted here for clarity), which are reported in Sec. 5.2.

meaningfully predicted need for understanding, stable individual differences (e.g., epistemic curiosity that students bring with them) remain important contributors to this construct. Still, the pattern is notable: participants who routinely used genAI reported significantly less intrinsic drive to grasp why's and how's.

(H4a) Usage and critical thinking. GenAI usage had a significant negative association with critical thinking ($\beta = -0.41$, p=.001; $f^2 = 0.20$), supporting **H4a**. From a dual-process perspective [66], critical thinking involves System 2 work—the deliberate questioning of claims, weighing of alternatives, and scrutiny of assumptions [44, 124]. This finding shows that routine genAI use predicted reductions in such deliberate processes, consistent with the tendency for fluent, seemingly "complete" genAI outputs to invite System 1 acceptance (as discussed in Sec. 3).

5.1.2 ***Trust and cognitive engagement.*** H1 hypothesized that students' trust in genAI would predict how frequently they used it. For our participants, this turned out to be the case: trust showed a large, significant positive association with routine genAI usage ($\beta = 0.64$, p<.001), with a moderate effect size ($f^2 = 0.25$), supporting **H1**. This finding is consistent with prior automation-bias accounts [16, 93], showing that trust lowers the reluctance towards (AI) delegation, making it feel reasonable rather than risky or provisional.

Since routine usage was negatively associated with all three dimensions of cognitive engagement **(H2a–H4a)**, we examined the extent to which trust affected the same three dimensions transitively through increased usage **(H2b–H4b)**.

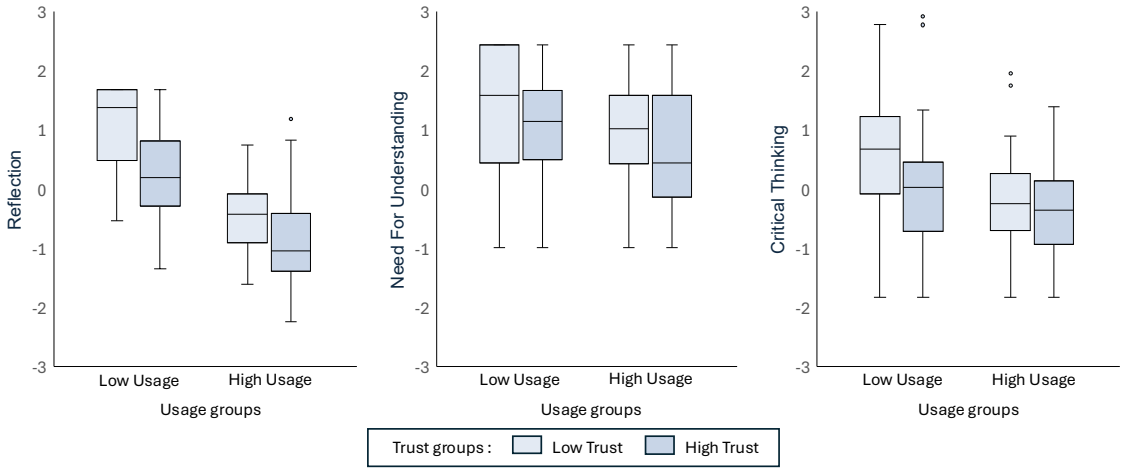

Fig. 4. Trust → usage → cognitive engagement distributions. Boxplots show standardized latent scores for Reflection, Need for Understanding, and Critical Thinking among participants grouped by genAI usage (median split) and trust in genAI (low vs. high; median split within each usage group).

Fig. 4 shows median-split box plot distributions by trust and usage. Across both low- and high-trust groups, participants with higher usage reported lower cognitive engagement (across all three dimensions). Moreover, within each usage group, participants in the "high-trust" group reported even lower cognitive engagement than their "low-trust" peers, indicating that trust amplified these outcomes.

Table 5. Effects of students' trust in genAI on routine usage and cognitive engagement habits (H1, H2b-H4b): Standardized path coefficients (β), standard deviations (SD), 95% confidence intervals (CI), p-values, and effect sizes (f^2). Transitive effects (H2b-H4b) capture how trust influences each cognitive engagement dimension mediated through routine usage. We consider $f^2 < 0.02$ no effect, $f^2 \in [0.02, 0.15)$ small, $f^2 \in [0.15, 0.35)$ medium, and $f^2 > 0.35$ large [31].

	B	SD	95% CI	p	f^2
H1 Trust→Usage	0.64	0.03	(0.58, 0.69)	**0.000**	0.25
Transitive effects: Trust to cognitive engagement mediated via usage					
H2b Trust→Usage→Reflection	−0.39	0.03	(-0.54, -0.45)	**0.000**	0.16
H3b Trust→Usage→Need for Understanding	−0.13	0.04	(0.07, 0.20)	**0.000**	0.11
H4b Trust →Usage→Critical Thinking	−0.26	0.05	(-0.32,-0.21)	**0.001**	0.15

Note: PLS-SEM estimates all structural paths jointly within a single model. Significance tests (and their Type-I error rates) are defined within this joint estimation; separate family-wise error corrections are not required [57, 58].

Mediation analysis confirmed these patterns (Tab. 5). Following standard PLS procedures [57], we tested the transitive effects of trust on each cognitive engagement dimension via routine usage. All three effects were significant, supporting **H2b–H4b**: participants with greater trust in genAI reported significantly less Reflection ($\beta = -0.39$, p<.001, $f^2 = 0.16$, VAF = 0.89), less Need for Understanding ($\beta = -0.13$, p<.001, $f^2 = 0.11$, VAF = 0.85), and less Critical thinking ($\beta = -0.26$, $p = .001$, $f^2 = 0.15$, VAF = 0.94). All VAF (Variance Accounted For) values were > 0.8, indicating *full mediation* [57]: trust in genAI affected each cognitive engagement dimension entirely through its effect on routine genAI usage.

Takeaway: RQ1-How's findings point to a consistent pattern of cognitive disengagement among students using genAI. Participants *who trusted and routinely used genAI* reported significantly lower levels of *reflection, need for understanding, and critical thinking* in STEM coursework.

5.2 RQ2-Who: Students' Cognitive Styles → Usage → Cognitive Engagement?

Trust was a strong predictor of routine genAI usage—and, transitively, of cognitive engagement. However, it alone did not fully account for the observed variance in participants' routine genAI use (R^2 = .54 in Fig. 3). Since individuals differ in how they approach and use technology [5, 18, 136], here we consider whether some students were particularly prone to genAI-related reductions in cognitive engagement via increased routine usage. If so, *which* students?

RQ2-Who investigates this question using cognitive styles: whether and how students' cognitive styles affect their routine use of genAI, and, in turn, their cognitive engagement.

5.2.1 ***Cognitive styles and routine genAI usage.*** Prior work has reported certain cognitive styles to significantly influence genAI adoption among knowledge workers [25, 28, 29]: intrinsic enjoyment of engaging with technology (*technophilic motivations*), comfort with uncertainty and technology-related risks (*risk tolerance*), and confidence in one's ability to use such tools effectively (*computer self-efficacy*) [5, 8, 18, 87, 90]. Thus, we investigated whether these same cognitive styles also predict *students'* routine usage of genAI.

The results were consistent with prior findings in knowledge work contexts. Participants with stronger technophilic motivations, higher risk tolerance, and greater computer self-efficacy reported significantly more routine genAI use than their peers (see Tab. 6, Fig. 5 for path associations).

Table 6. Effects of students' cognitive styles on routine genAI usage: Standardized path coefficients (β), standard deviations (SD), 95% confidence intervals (CI), p-values, and effect sizes (f^2). We consider $f^2 < 0.02$ no effect, $f^2 \in [0.02, 0.15)$ small, $f^2 \in [0.15, 0.35)$ medium, and $f^2 > 0.35$ large [31].

	B	SD	95% CI	p	f^2
Technophilic Motivations→Usage	0.14	0.06	(0.07, 0.26)	**0.011**	0.17
Risk Tolerance→Usage	0.29	0.05	(0.19, 0.38)	**0.000**	0.21
Computer Self-Efficacy→Usage	0.19	0.04	(0.06, 0.14)	**0.002**	0.11

Note: PLS-SEM estimates all structural paths jointly within a single model. Significance tests (and their Type-I error rates) are defined within this joint estimation; separate family-wise error corrections are not required [57, 58].

We find this result unsurprising. If a student is motivated by engaging with technology (cognitive style: technophilic motivations), it seems likely that they would likewise enjoy engaging with genAI, even when it is known to carry risks (cognitive style: risk tolerance), especially if the student is confident that they can handle those risks and use the tool effectively (cognitive style: computer self-efficacy). Indeed, these qualities are often touted in STEM education as indicators of career preparedness (e.g., [33, 88, 94, 107, 113, 125, 132]). Further, researchers have shown that these qualities *are* often helpful for STEM [52, 90]. However, it may be time to question whether this still holds with routine genAI use, which we investigate next.

5.2.2 ***Cognitive styles and cognitive engagement.*** If higher technophilic motivations, risk tolerance, and computer self-efficacy remain helpful for STEM education even under conditions of routine genAI use, then students with these

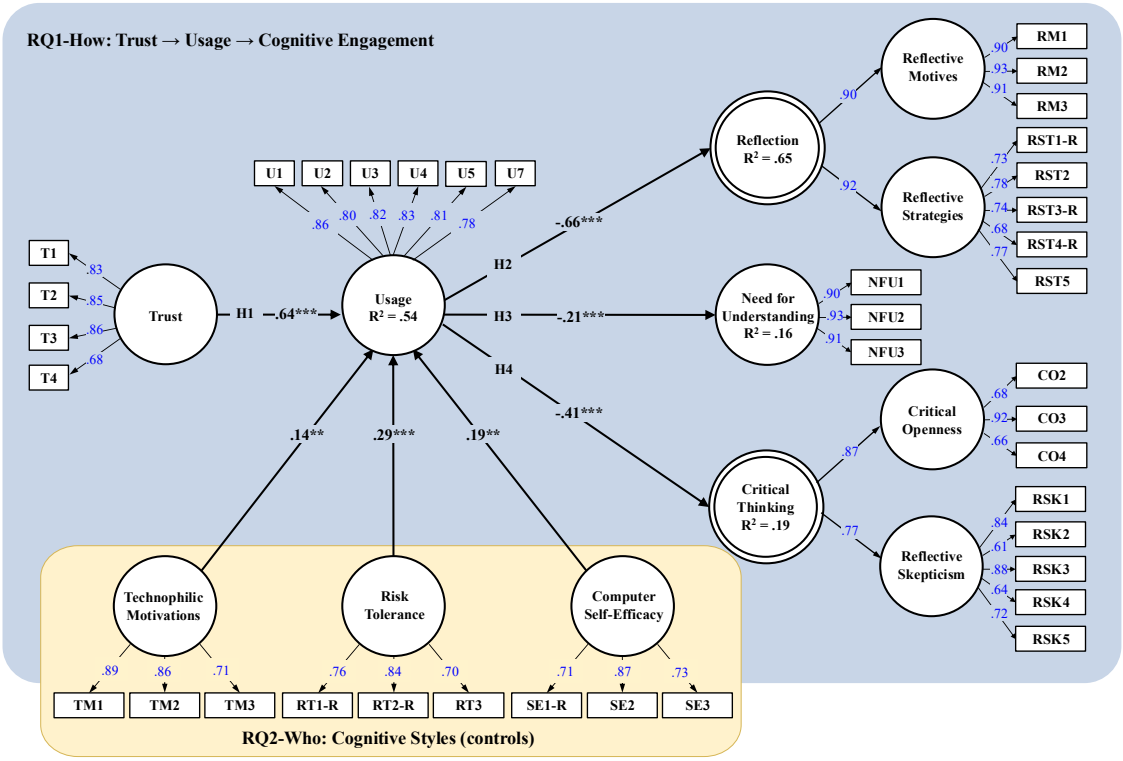

Fig. 5. Structural model path associations (full model: RQs1&2). Standardized path coefficients show direct effects of Trust and cognitive styles (Technophilic Motivations, Risk Tolerance, Computer Self-Efficacy; RQ2—yellow) on Usage, and of Usage on Reflection, Need for Understanding, and Critical Thinking, with significance ***p<.001, **p<.01, and *p<.05. Paths from LOCs (circles) to items (squares), and from HOCs (double circles) to LOCs, indicate factor loadings (in *blue*; all ≥ 0.6 [58]). Transitive effects of cognitive styles on cognitive engagement are reported in Tab. 7.

cognitive styles should also exhibit higher cognitive engagement in their coursework. For example, among RQ1-How's findings, the participants in the High Usage/Trust group (recall outliers above the box in Fig. 4) who exhibited fairly high reflection despite heavy genAI use would presumably be those with these cognitive styles.

Technophilic motivations: However, the opposite was true. As Fig. 6 (a) illustrates, higher technophilic motivations corresponded with *lower* median cognitive engagement across all three dimensions. Mediation analysis (Tab. 7) confirmed these patterns: participants with higher technophilic motivations reported significantly less Reflection ($\beta = -0.11$, p=.010, $f^2 = 0.15$), significantly less Need for Understanding ($\beta = -0.12$, $p = .014$, $f^2 = 0.07$), and significantly less Critical Thinking ($\beta = -0.05$, $p = .007$, $f^2 = 0.09$).

Risk tolerance and computer self-efficacy: Similar patterns emerged for the other two cognitive styles (Fig. 6 (b) and (c)). Participants with higher risk tolerance reported significantly less Reflection ($\beta = -0.15$, p<.001, $f^2 = 0.08$), significantly less Need for Understanding ($\beta = -0.11$, p=.002, $f^2 = 0.07$), and significantly less Critical Thinking ($\beta = -0.13$, p<.001, $f^2 = 0.08$). Research on attitude towards risk suggests one reason why [86]: risk tolerance captures willingness to proceed despite uncertainty. In HAI contexts, this can, alongside increasing usage, reduce users' inclination to question or scrutinize AI outputs [142], even when such scrutiny would be warranted.

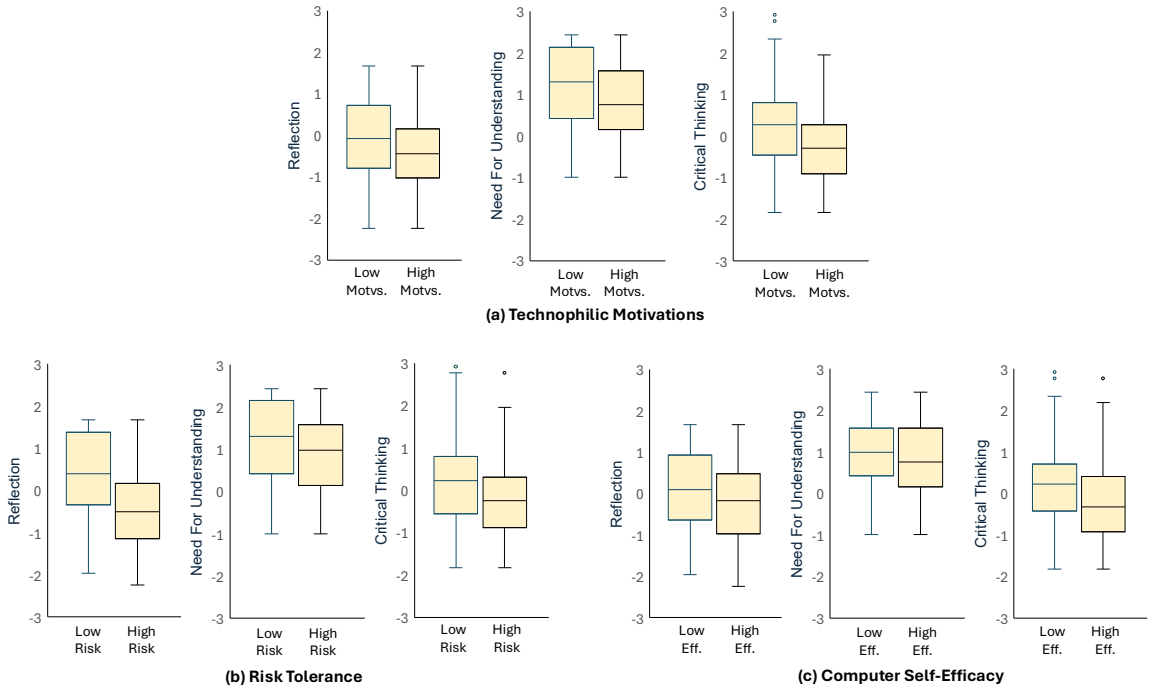

Fig. 6. Cognitive engagement scores by median-split groups on (a) Technophilic Motivations, (b) Risk Tolerance, and (c) Computer Self-Efficacy. Boxplots show standardized latent scores for Reflection, Need for Understanding, and Critical Thinking across all groups.

Table 7. Transitive effects of students' cognitive styles on their cognitive engagement habits via routine genAI use: Standardized path coefficients (β), standard deviations (SD), confidence intervals (CI), p-values, effect sizes (f^2), and Variance Accounted For (VAF $\geq$ 0.8 indicates full mediation). We consider $f^2 < 0.02$ no effect, $f^2 \in [0.02, 0.15)$ small, $f^2 \in [0.15, 0.35)$ medium, and $f^2 > 0.35$ large [31].

	B	SD	95% CI	p	f^2	VAF
Technophilic Motivations→Usage→Reflection	−0.11	0.05	(-0.21, -0.10)	**0.010**	0.15	0.83
Technophilic Motivations→Usage→Need For Understanding	−0.12	0.01	(-0.14, -0.11)	**0.014**	0.07	0.83
Technophilic Motivations→Usage→Crticial Thinking	−0.05	0.03	(-0.10, -0.02)	**0.007**	0.09	0.86
Risk Tolerance→Usage→Reflection	−0.15	0.04	(-0.19, -0.12)	**0.000**	0.08	0.87
Risk Tolerance→Usage→Need For Understanding	−0.11	0.02	(-0.12, -0.06)	**0.002**	0.07	0.85
Risk Tolerance→Usage→Crticial Thinking	−0.13	0.03	(-0.14, -0.09)	**0.000**	0.08	0.82
Computer Self-Efficacy→Usage→Reflection	−0.17	0.03	(-0.21, -0.12)	**0.013**	0.15	0.84
Computer Self-Efficacy→Usage→Need For Understanding	−0.12	0.01	(-0.17, -0.10)	**0.005**	0.07	0.86
Computer Self-Efficacy→Usage→Crticial Thinking	−0.14	0.02	(-0.16, -0.11)	**0.020**	0.13	0.87

Note: PLS-SEM estimates all structural paths jointly within a single model. Significance tests (and their Type-I error rates) are defined within this joint estimation; separate family-wise error corrections are not required [57, 58].

Participants with higher computer self-efficacy likewise reported significantly less Reflection ($\beta = -0.17$, p=.013, $f^2 = 0.15$), significantly less Need for Understanding ($\beta = -0.12$, p=.005, $f^2 = 0.07$), and significantly less Critical Thinking ($\beta = -0.14$, p=.020, $f^2 = 0.13$). While prior work suggests that higher self-efficacy leads to greater success (e.g.,

as per self-efficacy as an effective surrogate for ability research [8, 32]), some education research indicates that greater self-confidence in genAI contexts may lead students to overestimate learning gains from such interactions [47, 137].

Group analysis: Still, some research suggests that academic level and prior genAI experience may be important factors [27, 121, 133]. To investigate whether these factors affected the structural relationships in our model, we conducted standard PLS Multi-Group Analysis (MGA) [57, 112]. We defined groups by (i) academic level (lower- vs. upper-division) and (ii) prior genAI experience (median split on self-reported experience). The analysis revealed no statistically significant differences across any paths, confirming that the model relationships held consistently regardless of participants' academic seniority or prior experience with genAI.

Takeaway: Students that prevailing wisdom would classify as well-positioned for STEM work—those with *higher technophilic motivations, risk tolerance, and computer self-efficacy*—were also the ones who were particularly susceptible to cognitively disengaging when they routinely used genAI.

5.3 Model adequacy assessment

Statistical and practical significance establish that relationships exist, but they do not indicate how well a model accounts for observed variance, whether it is well-specified and consistent with the data (fit), or whether it exhibits predictive relevance beyond the estimation sample. To address these questions, we assessed model adequacy using coefficients of determination (R^2, Adj. R^2), model fit (SRMR), and out-of-sample predictive relevance (Q^2), following recommended PLS-SEM practices [58, 112].

Coefficients of determination values (R^2, Adj. R^2) capture how much variance in each outcome (e.g., Reflection) is explained by its predictors. As shown in Tab. 8, the model explains a substantial proportion of variance in *Usage* ($R^2 = 0.54$) and *Reflection* ($R^2 = 0.65$). The model explains more moderate variance in *Need for Understanding* ($R^2 = 0.16$) and *Critical Thinking* ($R^2 = 0.19$). While these constructs are influenced by a broader set of factors beyond the model, the observed values exceed the common benchmark of 0.15 for exploratory model adequacy in behavioral sciences [24, 57].

Table 8. Explanatory power of the structural model. R^2 and Adj. R^2 reflect the variance in endogeneous constructs explained by its predictors; $Q^2_{\text{predict}} > 0$ indicate the model's out-of-sample predictive relevance.

Construct	R^2	*Adj.* R^2	$Q^2_{predict}$
Usage	0.544	0.541	0.515
Reflection	0.652	0.649	0.478
Need for Understanding	0.164	0.163	0.131
Critical Thinking	0.192	0.188	0.267

Model fit was assessed using the Standardized Root Mean Square Residual (SRMR), which is recommended for detecting structural misspecification in PLS-SEM [112]. The observed SRMR of 0.067 falls below the conservative threshold of 0.08 [62], indicating a well-specified model, i.e., the proposed relationships are consistent with the covariance structure of the data.

Finally, we assessed out-of-sample predictive relevance to evaluate whether the model generalizes beyond the estimation sample. Using Stone-Geisser's Q^2 [131], computed via 10-fold, 10-repeat PLS-predict, all dependent constructs yielded positive values (see Tab. 8). This indicates that the model's predictions on held-out data outperform a naive mean-based benchmark, supporting its predictive relevance beyond in-sample explanatory fit.

Takeaway: The proposed model is adequate for its intended purpose: it captures sufficient variance in key outcomes, shows no evidence of structural misspecification, and demonstrates predictive relevance beyond in-sample fit.

6 Discussion: The Cognitive Debt Cycle

Our findings raise the possibility of a *cognitive debt cycle* emerging from students' trust-driven, routine use of genAI. Much like technical debt accumulates when short-term expedience substitutes for more principled design, cognitive debt may accrue when students repeatedly delegate relevant cognitive work onto genAI. The concern is that genAI may restructure the learning environment in ways that make delegation appear rational, gradually training the brain to reconsider what effort investment feels worthwhile. Further, these effects may be *self-reinforcing*: as students repeatedly delegate cognitive work, their ability to perform these tasks independently may decline. GenAI may then become increasingly necessary, further reducing the already limited opportunities to hone these skills. The result may be a downward spiral of overreliance, in which understanding becomes increasingly shallow, and the capacity to apply knowledge in new or dynamic contexts gradually erodes.

Unfortunately, our findings suggest that responsible genAI use did not automatically emerge with academic seniority or prior genAI experience, challenging the assumption that students will "grow into" reflective users. This raises a fundamental question: *How can students be expected to regulate their genAI use when the very capacities required to do so are progressively weakened by routine interaction with these tools?*

6.1 Interrupting the Cycle: What can educators do?

Recognizing the cognitive debt cycle is one thing; interrupting it is another. A common response has been to call for more AI literacy [56], with the expectation that better-informed learners will be more attuned to the risks of overreliance. Indeed, emerging frameworks, such as the Digital Education Council's AI literacy model [37], position critical thinking and judgment about AI as core competencies, emphasizing the ability to evaluate AI reasoning, identify bias, and recognize contextual gaps in AI outputs. We agree with the importance of these skills. Still, prior work suggests that awareness of AI's fallibility does not reliably deter reliance when answers are immediate and cognitively effortless [16, 130]. Our findings echo this pattern: neither prior AI experience nor academic seniority insulated students from cognitive disengagement, suggesting that *awareness alone does not counteract the incentives to disengage.*

Why might this be? Anderson's theory of rational analysis [7] suggests one reason why. According to this theory, humans exhibit a "local maximum of adaptedness" to their environments, optimizing for maximum immediate reward and minimum immediate cost. In educational contexts, this implies that delegating cognitive work to genAI may be a rational response to environments that reward efficiency while imposing few short-term costs for shallow engagement.

This theory then also suggests possibilities for adjusting students' perceived costs and rewards of AI delegation as a means to interrupt the cognitive debt cycle—affording educators a rethinking of educational experiences where delegating cognitive work becomes less attractive to students.

One possibility centers on increasing the intrinsic rewards of cognitive engagement itself. Tasks that are personally meaningful or inherently enjoyable could reduce the appeal of delegation [76, 84, 110], as learners continue to derive satisfaction from the work. For example, designing coursework as opportunities for exploration through storytelling [30], gamified elements [59, 83], or curiosity-driven puzzles (e.g., "Why did the AI get this wrong?" or "How might the answer differ depending on the point of view?") could tap into intrinsic motivations that invite more deliberate engagement [34]. Notably, curiosity-based approaches may also make the consequences of delegation more visible by surfacing AI errors, omissions, or mismatches within students' own contexts.

Another possibility is to consider increasing the costs of AI delegation through intentional task and activity design. One approach could involve alternating between AI-assisted and independent work. For instance, students might first generate code with AI assistance, then be required to independently critique, debug/repair, verify, and adapt that code to their project requirements without AI (e.g., during a supervised lab). Such designs can make delegation immediately costly, as knowledge gaps that were masked during the AI-assisted phase must be confronted during subsequent independent work. Cumulative assignments that build on previous assignments can amplify these costs further, as shallow cognitive engagement in early assignments could compound into visible struggles in later ones. Finally, combining these approaches with in-class assessments and/or collaborative/group activities [123] can further reinforce engagement by leveraging social accountability [80, 139], where shallow AI-generated contributions become quickly visible, potentially compromising individual and/or group reputation and outcomes.

Importantly, these strategies are not "silver bullets" nor miracle cures. As genAI capabilities, human practices, and mindsets co-evolve, responses to this challenge will need to adapt. A key ongoing question for *educators* remains: *How to design educational experiences to foster cognitive engagement in ways that improve both (1) students' learning of the subject matter and (2) their ability to use AI appropriately?*

6.2 Interrupting the Cycle: What can AI designers do?

Interrupting the debt cycle may require more than pedagogical adjustments; it raises questions about how genAI itself can be designed for educational use. One possibility, drawing on Engelbart's early vision of intellectual augmentation [40], is to design genAI as "bicycles for the mind". Just as bicycles amplify physical efficiency while requiring the rider's active control; genAI, designed as mind-bicycles, could amplify cognitive reach (e.g., by critiquing, counter-arguing, nudging the user [108]) while keeping humans responsible for exploration, steering, and judgment. The goal is for progress to emerge through *complementarity*: the human and the AI each need to contribute distinct forms of cognitive work to achieve meaningful outcomes.

What this framing seeks to avoid is a wholesale substitution of human agency. When genAI "thinks" on a student's behalf, they lose opportunities to build skill, and genAI, in turn, loses access to future high-quality human input shaping its evolution. Designing for complementarity is therefore not only pedagogically desirable; it is necessary for sustaining genAI-assisted knowledge work. A key ongoing challenge for *AI designers* then is: *How to design genAI for complementarity in ways that meaningfully amplify human thinking instead of replacing it?*

Incorporating "friction"—intentional points of difficulty [35]—could offer one possibility toward designing such systems for educational use. Designing for friction runs counter to conventional product metrics that prioritize maximal convenience (e.g., fewer clicks). However, prior work shows that friction can be productive when it slows users down to provoke reflection on their emerging work [16, 38]. One way to introduce friction, for example, could be to design education-oriented genAI as a "provocateur," repositioning assistance as questioning rather than confirming [38]. Such a provocateur could prompt students to revisit their assumptions and interrogate genAI's underlying rationales, enabling them to identify and correct flaws in that (and/or their own) reasoning. Another way could be to design systems that require users to deliberately engage with multiple AI perspectives: for instance, by comparing outputs generated under differing viewpoints or competing claims and reconciling disagreements among them. Prior work shows that such designs can indeed strengthen users' critical thinking and reasoning practices [85].

Furthermore, our findings suggest that reimagining genAI for complementarity should not assume a one-size-fits-all. Recall, students who might have been regarded as most "tech-ready"—those high in technophilic motivations, self-efficacy, and risk tolerance—were also the ones most prone to cognitive disengagement.

This raises possibilities for adaptivity in education-oriented genAI design to account for students' diverse cognitive styles. For instance, technophilic users may benefit from *interface nudges* that demand comparison and justification (e.g., selecting among multiple AI-generated solutions and justifying their choice, or attempting to "beat the AI"). Similarly, highly self-efficacious users may benefit from scaffolds that require an initial attempt (e.g., plans, assumptions, rationales) before AI assistance is provided, helping recalibrate their expectations and confidence in AI support.

That said, adaptivity that is too automatic risks undermining user agency and locking students into a fixed interaction style. Although human-AI interaction research is still relatively young, the principle of keeping users in control has been affirmed by many HAI researchers (e.g., [3, 5, 11, 26, 64, 92, 117, 118]). Accordingly, adaptivity should remain configurable, enabling students to steer how the system supports them as their goals, contexts, and skills evolve.

Ultimately, realizing this future will require coordinated effort among AI designers, researchers, and educators to support students in developing and sustaining essential intellectual habits when using genAI, rather than perpetuating the cognitive debt cycle. While the specific pathways forward remain open and evolving, the cost of inaction is clear: *a future where apparent "efficiency" rises even as sustained thinking habits quietly erode.*

7 Conclusion

We began this investigation by asking: *What happens to students' cognitive engagement when genAI becomes routine?* The answers we found were both clear and concerning. Our main findings were:

- **RQ1-How**: Students who routinely used genAI reported statistically significantly less cognitive engagement, i.e., less reflection, less need for understanding, and less critical thinking in their coursework. Higher trust in genAI further amplified these disengagement effects through increased routine genAI use.
- **RQ2-Who**: Ironically, students commonly touted as "tech-natives"—those with stronger technophilic motivations, greater computer self-efficacy, and higher risk tolerance—were the ones who were significantly more prone to genAI-related cognitive disengagement.

If routine reliance on genAI during formative years changes students' willingness to engage in effortful thinking, many may enter professional life without having developed the intellectual habits that earlier generations developed through practice. Yet this need not be inevitable. The cognitive debt cycle can be interrupted, but doing so will take more than exhortation or awareness campaigns. It calls for deliberate redesign of both learning environments and genAI systems to preserve human agency and support genuine complementarity—a division of cognitive labor in which both human and AI contribute meaningfully, and neither is displaced. The choices educators and AI designers make now will shape not only what students learn, but how they learn to think.

Acknowledgments

This work was supported in part by USDA-NIFA 2021-67021-35344 and by the National Science Foundation under Grant Nos. 2235601, 2236198, 2247929, 2303042, and 2303043. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] [n. d.]. Supplemental Package. https://zenodo.org/record/18420685.

[2] Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan. 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education* 21, 1 (2024), 10.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[4] Matin Amoozadeh, David Daniels, Daye Nam, Aayush Kumar, Stella Chen, Michael Hilton, Sruti Srinivasa Ragavan, and Mohammad Amin Alipour. 2024. Trust in Generative AI among Students: An exploratory study. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 67–73.

[5] Andrew Anderson, Jimena Noa Guevara, Fatima Moussaoui, Tianyi Li, Mihaela Vorvoreanu, and Margaret Burnett. 2024. Measuring user experience inclusivity in human-AI interaction via five user problem-solving styles. *ACM Transactions on Interactive Intelligent Systems* 14, 3 (2024), 1–90.

[6] Andrew Anderson, David Piorkowski, Margaret Burnett, and Justin Weisz. 2025. An LLM's Attempts to Adapt to Diverse Software Engineers' Problem-Solving Styles: More Inclusive & Equitable? *arXiv preprint arXiv:2503.11018* (2025).

[7] John R Anderson. 2013. *The adaptive character of thought*. Psychology Press.

[8] Albert Bandura. 1997. Self-efficacy the exercise of control. New York: H. *Freeman & Co. Student Success* 333 (1997), 48461.

[9] André Barcaui. 2025. ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open* 12 (2025), 102287.

[10] Jan-Michael Becker, Kristina Klein, and Martin Wetzels. 2012. Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models. *Long range planning* 45, 5-6 (2012), 359–394.

[11] Aditya Bhattacharya, Simone Stumpf, and Katrien Verbert. 2025. Importance of User Control in Data-Centric Steering for Healthcare Experts. *arXiv preprint arXiv:2506.18770* (2025).

[12] John Biggs, David Kember, and Doris YP Leung. 2001. The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology* 71, 1 (2001), 133–149.

[13] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

[14] Monique Boekaerts, Moshe Zeidner, and Paul R Pintrich. 1999. *Handbook of self-regulation*. Elsevier.

[15] Jennifer Bryce and Graeme Withers. 2003. Engaging secondary school students in lifelong learning. (2003).

[16] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. 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* 5, CSCW1 (2021), 1–21.

[17] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. 2016. Finding gender-inclusiveness software issues with GenderMag: A field investigation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2586–2598.

[18] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with Computers* 28, 6 (2016), 760–787.

[19] Jenna Butler, Steven Jaffe, Ralf Janßen, Nancy Baym, Brent Hecht, Jake Hofman, Sean Rintel, Bahareh Sarrafzadeh, Abigail M. Sellen, Teodor Vorvoreanu, and Jaime Teevan. 2025. *Microsoft New Future of Work Report 2025*. Technical Report MSR-TR-2025-58. Microsoft Research. https://aka.ms/nfw2025

[20] John T Cacioppo and Richard E Petty. 1982. The need for cognition. *Journal of personality and social psychology* 42, 1 (1982), 116.

[21] Xinyue Chen, Kunlin Ruan, Kexin Phyllis Ju, Nathan Yap, and Xu Wang. 2025. More AI assistance reduces cognitive engagement: Examining the AI assistance dilemma in AI-supported note-taking. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–29.

[22] Michelene TH Chi. 2009. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science* 1, 1 (2009), 73–105.

[23] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[24] Wynne W Chin et al. 1998. The partial least squares approach to structural equation modeling. *Modern Methods for Business Research* 295, 2 (1998), 295–336.

[25] Rudrajit Choudhuri, Carmen Badea, Christian Bird, Jenna Butler, Rob DeLine, and Brian Houck. 2025. AI Where It Matters: Where, Why, and How Developers Want AI Support in Daily Work. *arXiv preprint arXiv:2510.00762* (2025).

[26] Rudrajit Choudhuri, Dylan Liu, Igor Steinmacher, Marco Gerosa, and Anita Sarma. 2024. How Far Are We? The Triumphs and Trials of Generative AI in Learning Software Engineering. In *Proceedings of the IEEE/ACM 46th international conference on software engineering*. 1–13.

[27] Rudrajit Choudhuri, Ambareesh Ramakrishnan, Amreeta Chatterjee, Bianca Trinkenreich, Igor Steinmacher, Marco Gerosa, and Anita Sarma. 2025. Insights from the Frontline: GenAI Utilization Among Software Engineering Students. In *2025 IEEE/ACM 37th International Conference on Software Engineering Education and Training (CSEE&T)*. IEEE, 1–12.

[28] Rudrajit Choudhuri, Bianca Trinkenreich, Rahul Pandita, Eirini Kalliamvakou, Igor Steinmacher, Marco Gerosa, Christopher Sanchez, and Anita Sarma. 2025. What Guides Our Choices? Modeling Developers' Trust and Behavioral Intentions Towards GenAI. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. doi:10.1109/ICSE55347.2025.00087

[29] Rudrajit Choudhuri, Bianca Trinkenreich, Rahul Pandita, Eirini Kalliamvakou, Igor Steinmacher, Marco Gerosa, Christopher Sanchez, and Anita Sarma. 2025. What Needs Attention? Prioritizing Drivers of Developers' Trust and Adoption of Generative AI. *arXiv preprint arXiv:2505.17418* (2025).

[30] Ruth C Clark and Richard E Mayer. 2023. *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. john Wiley & sons.

[31] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

[32] Deborah R Compeau and Christopher A Higgins. 1995. Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly* (1995),

189–211.

[33] Holly C. Corbett. 2024. *Psychological Safety: Try This Tip to Know When to Take a Risk at Work.* Forbes. https://www.forbes.com/sites/hollycorbett/2024/09/27/psychological-safety-try-this-tip-to-know-when-to-take-a-risk-at-work/ Accessed: 2026-01-07.

[34] Arianna Costantini, Arnold B Bakker, and Yuri S Scharp. 2025. Playful Study Design: A Novel Approach to Enhancing Student Well-Being and Academic Performance. *Educational Psychology Review* 37, 2 (2025), 1–42.

[35] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1389–1397.

[36] John Dewey. 1933. How we think: A restatement of the relation of reflective thinking to the educative process. (1933).

[37] Digital Education Council. 2025. Digital Education Council AI Literacy Framework. https://www.digitaleducationcouncil.com/post/digital-education-council-ai-literacy-framework. Accessed December, 2025.

[38] Ian Drosos, Advait Sarkar, Neil Toronto, et al. 2025. " It makes you think": Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work. *arXiv preprint arXiv:2501.17247* (2025).

[39] John Dunlosky, Christopher Hertzog, M Kennedy, and Keith W Thiede. 2005. The self-monitoring approach for effective learning. *Cognitive Technology* 10, 1 (2005), 4–11.

[40] Douglas C Engelbart. 2023. Augmenting human intellect: A conceptual framework. In *Augmented education in the global age*. Routledge, 13–29.

[41] Robert H Ennis. 1993. Critical thinking assessment. *Theory into practice* 32, 3 (1993), 179–186.

[42] Robert H Ennis. 2018. Critical thinking across the curriculum: A vision. *Topoi* 37, 1 (2018), 165–184.

[43] Christina J Evans, John R Kirby, and Leandre R Fabrigar. 2003. Approaches to learning, need for cognition, and strategic flexibility among university students. *British Journal of Educational Psychology* 73, 4 (2003), 507–528.

[44] Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). (1990).

[45] Peter A Facione et al. 2011. Critical thinking: What it is and why it counts. *Insight assessment* 1, 1 (2011), 1–23.

[46] Peter A Facione, Carol A Sanchez, Noreen C Facione, and Joanne Gainen. 1995. The disposition toward critical thinking. *The Journal of general education* 44, 1 (1995), 1–25.

[47] Lei Fan, Kunyang Deng, and Fangxue Liu. 2025. Educational impacts of generative artificial intelligence on learning and performance of engineering students in China. *Scientific reports* 15, 1 (2025), 26521.

[48] Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology* 56, 2 (2025), 489–530.

[49] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behav Res Methods* 41, 4 (2009), 1149–1160.

[50] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.

[51] Michael Gerlich. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies* 15, 1 (2025), 6.

[52] Sonia C Giaccone and Mats Magnusson. 2022. Unveiling the role of risk-taking in innovation: antecedents and effects. *R&D Management* 52, 1 (2022), 93–107.

[53] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

[54] Yuhong Gong, Shang Zhang, Ting Zhang, and Xinfa Yi. 2025. The impact of feedback literacy on reflective learning types in Chinese high school students: based on latent profile analysis. *Frontiers in Psychology* 16 (2025), 1516253.

[55] Wolfgang L Grichting. 1994. The meaning of "I Don't Know" in opinion surveys: Indifference versus ignorance. *Aust Psychol* 29, 1 (1994).

[56] Xingjian Gu and Barbara J Ericson. 2025. AI literacy in K-12 and higher education in the wake of generative AI: An integrative review. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V. 1*. 125–140.

[57] Joseph F Hair. 2014. *A primer on partial least squares structural equation modeling (PLS-SEM)*. sage.

[58] Joseph F Hair, Jeffrey J Risher, Marko Sarstedt, and Christian M Ringle. 2019. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* (2019).

[59] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work? A literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[60] Md Montaser Hamid, Amreeta Chatterjee, Mariam Guizani, Andrew Anderson, Fatima Moussaoui, Sarah Yang, Isaac Escobar, Anita Sarma, and Margaret Burnett. 2024. How to measure diversity actionably in technology. In *Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights*. Apress Berkeley, CA, 469–485.

[61] Md Montaser Hamid, FatimA A Moussaoui, Jimena Noa Guevara, Andrew Anderson, Puja Agarwal, Jonathan Dodge, and Margaret Burnett. 2025. Inclusive design of AI's Explanations: Just for Those Previously Left Out? *ACM Transactions on Interactive Intelligent Systems* (2025).

[62] Jörg Henseler, Geoffrey Hubona, and Pauline Ash Ray. 2016. Using PLS path modeling in new technology research: updated guidelines. *Industrial Management & Data Systems* 116, 1 (2016), 2–20.

[63] Jörg Henseler, Christian M Ringle, and Marko Sarstedt. 2015. A new criterion for assessing discriminant validity in variance-based structural

equation modeling. *J. Acad. Mark. Sci.* 43, 1 (2015), 115–135.
[64] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with computers* 12, 4 (2000), 409–426.
[65] Matt C Howard. 2016. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction* 32, 1 (2016), 51–62.
[66] Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux* (2011).
[67] David Kember, Doris YP Leung, Alice Jones, Alice Yuen Loke, Jan McKay, Kit Sinclair, Harrison Tse, Celia Webb, Frances Kam Yuet Wong, Marian Wong, et al. 2000. Development of a questionnaire to measure the level of reflective thinking. *Assessment & Evaluation in Higher Education* 25, 4 (2000), 381–395.
[68] Barbara A Kitchenham and Shari L Pfleeger. 2008. Personal opinion surveys. In *Guide to advanced empirical software engineering*. Springer, 63–92.
[69] Amy J Ko, Thomas D LaToza, and Margaret M Burnett. 2015. A practical guide to controlled experiments of software engineering tools with human participants. *Empirical Software Engineering* 20, 1 (2015), 110–141.
[70] Aleksander Kobylarek, Kamil Błaszczyński, Luba Ślósarz, and Martyna Madej. 2022. Critical Thinking Questionnaire (CThQ)–construction and application of critical thinking test tool. *Andragogy Adult Education and Social Marketing* 2, 2 (2022), 1–1.
[71] Ned Kock. 2014. Advanced mediating effects tests, multi-group analyses, and measurement model assessments in PLS-based SEM. *International Journal of e-Collaboration* 10, 1 (2014).
[72] Ned Kock. 2015. Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of e-Collaboration (ijec)* 11, 4 (2015), 1–10.
[73] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872* (2025).
[74] Pia Kreijkes, Viktor Kewenig, Martina Kuvalja, Mina Lee, Sylvia Vitello, Jake M Hofman, Abigail Sellen, Sean Rintel, Daniel G Goldstein, David M Rothschild, et al. 2025. Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. *Available at SSRN* (2025).
[75] Susie Lamborn, Fred Newmann, and Gary Wehlage. 1992. The significance and sources of student engagement. *Student engagement and achievement in American secondary schools* (1992), 11–39.
[76] Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press.
[77] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. 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 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
[78] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[79] Marina Lepp and Joosep Kaimre. 2025. Does generative AI help in learning programming: Students' perceptions, reported use and relation to performance. *Computers in Human Behavior Reports* 18 (2025), 100642.
[80] Jennifer S Lerner and Philip E Tetlock. 1999. Accounting for the effects of accountability. *Psychological bulletin* 125, 2 (1999), 255.
[81] Wei Li, Ji-Yi Huang, Cheng-Ye Liu, Judy CR Tseng, and Shu-Pan Wang. 2023. A study on the relationship between student'learning engagements and higher-order thinking skills in programming learning. *Thinking Skills and Creativity* 49 (2023), 101369.
[82] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 1257–1268.
[83] Andreas Lieberoth. 2015. Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture* 10, 3 (2015), 229–248.
[84] Marjolein Lips-Wiersma and Lani Morris. 2009. Discriminating between 'meaningful work'and the 'management of meaning'. *Journal of business ethics* 88, 3 (2009), 491–511.
[85] Yiren Liu, Viraj Shah, Sangho Suh, Pao Siangliulue, Tal August, and Yun Huang. 2025. Perspectra: Choosing Your Experts Enhances Critical Thinking in Multi-Agent Research Ideation. *arXiv preprint arXiv:2509.20553* (2025).
[86] George F Loewenstein, Elke U Weber, Christopher K Hsee, and Ned Welch. 2001. Risk as feelings. *Psychological bulletin* 127, 2 (2001), 267.
[87] David Lyell and Enrico Coiera. 2017. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association* 24, 2 (2017), 423–431.
[88] Manufacturing.net (ASQ survey report). 2013. *U.S. Teens Fear Risk-Taking; STEM Careers Demand It.* Manufacturing.net. https://www.manufacturing.net/industry40/article/13215403/us-teens-fear-risktaking-stem-careers-demand-it Accessed: 2026-01-08.
[89] Lauren E Margulieux, James Prather, Brent N Reeves, Brett A Becker, Gozde Cetin Uzun, Dastyni Loksa, Juho Leinonen, and Paul Denny. 2024. Self-regulation, self-efficacy, and fear of failure interactions with how novices use llms to solve programming problems. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. 276–282.
[90] Mario Martínez-Córcoles, Mare Teichmann, and Mart Murdvee. 2017. Assessing technophobia and technophilia: Development and validation of a questionnaire. *Technology in Society* 51 (2017), 183–188.
[91] Jason Miklian and Kristian Hoelscher. 2025. A New Digital Divide? Coder Worldviews, the Slop Economy, and Democracy in the Age of AI. *arXiv preprint arXiv:2510.04755* (2025).
[92] Philippe Palanque, Fabio Paternò, Virpi Roto, Albrecht Schmidt, Simone Stumpf, and Jürgen Ziegler. 2023. A multi-perspective panel on user-centred transparency, explainability, and controllability in automations. In *IFIP Conference on Human-Computer Interaction*. Springer, 349–353.

[93] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.
[94] ParentsCanada Sponsored Content. 2022. *Getting Students Excited About STEM Requires Taking Risks.* ParentsCanada. https://parentscanada.com/sponsored/getting-students-excited-about-stem/ Accessed: 2026-01-08.
[95] Richard W Paul, Linda Elder, and Ted Bartell. 1997. California teacher preparation for instruction in critical thinking: Research findings and policy recommendations. (1997).
[96] Sheila A Paul. 2014. Assessment of critical thinking: a Delphi study. *Nurse Education Today* 34, 11 (2014), 1357–1360.
[97] Sebastian AC Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust issues with trust scales: examining the psychometric quality of trust measures in the context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.
[98] Paul R Pintrich and Elisabeth V De Groot. 1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology* 82, 1 (1990), 33.
[99] Philip M Podsakoff, Scott B MacKenzie, Jeong-Yeon Lee, and Nathan P Podsakoff. 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology* 88, 5 (2003), 879.
[100] James Prather, Juho Leinonen, Natalie Kiesler, Jamie Gorson Benario, Sam Lau, Stephen MacNeil, Narges Norouzi, Simone Opel, Vee Pettit, Leo Porter, et al. 2025. Beyond the hype: A comprehensive review of current trends in generative AI research, teaching practices, and tools. *2024 Working Group Reports on Innovation and Technology in Computer Science Education* (2025), 300–338.
[101] James Prather, Brent N Reeves, Paul Denny, Brett A Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2023. "It's weird that it knows what I want": Usability and interactions with copilot for novice programmers. *ACM transactions on computer-human interaction* 31, 1 (2023), 1–31.
[102] James Prather, Brent N Reeves, Juho Leinonen, Stephen MacNeil, Arisoa S Randrianasolo, Brett A Becker, Bailey Kimmel, Jared Wright, and Ben Briggs. 2024. The widening gap: The benefits and harms of generative AI for novice programmers. In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*. 469–486.
[103] Qiaolin Qin, Ronnie de Souza Santos, and Rodrigo Spinola. 2025. On the Role and Impact of GenAI Tools in Software Engineering Education. *arXiv preprint arXiv:2512.04256* (2025).
[104] Xiaodong Qu, Joshua Sherwood, Peiyan Liu, and Nawwaf Aleisa. 2025. Generative AI Tools in Higher Education: A Meta-Analysis of Cognitive Impact. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.
[105] Qualtrics. 2025. Qualtrics Survey Platform. https://www.qualtrics.com. Accessed August 11, 2025.
[106] Tenko Raykov and George A Marcoulides. 2011. *Introduction to psychometric theory.* Routledge.
[107] Miranda Talley Reagan. 2016. The Benefits of the STEM Mindset: Risk Taking and Resilience. (2016).
[108] Leon Reicherts, Zelun Tony Zhang, Elisabeth von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, help me think—but for myself: Assisting people in complex decision-making by providing different kinds of cognitive support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
[109] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in Cognitive Sciences* 20, 9 (2016), 676–688.
[110] Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research* (2001), 3–19.
[111] Daniel Russo. 2024. Navigating the complexity of generative AI adoption in software engineering. *ACM Transactions on Software Engineering and Methodology* (2024).
[112] Daniel Russo and Klaas-Jan Stol. 2021. PLS-SEM for Software Engineering Research: An Introduction and Survey. *ACM Comput Surv* 54, 4 (2021).
[113] Chris Sakellariou and Zheng Fang. 2021. Self-efficacy and interest in STEM subjects as predictors of the STEM gender gap in the US: The role of unobserved heterogeneity. *International Journal of Educational Research* 109 (2021), 101821.
[114] Marko Sarstedt and Jun-Hwa Cheah. 2019. Partial least squares structural equation modeling using SmartPLS: a software review.
[115] Marko Sarstedt, Joseph F. Hair, Jun-Hwa Cheah, et al. 2019. How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australas. Mark. J.* 27, 3 (2019), 197–211.
[116] Marko Sarstedt, Christian M Ringle, and Joseph F Hair. 2017. Treating unobserved heterogeneity in PLS-SEM: A multi-method approach. In *Partial Least Squares Path Modeling*. Springer, 197–217.
[117] Albrecht Schmidt. 2020. Interactive human centered artificial intelligence: a definition and research challenges. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces*. 1–4.
[118] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.
[119] Forrest Shull, Janice Singer, and Dag IK Sjøberg. 2007. Guide to Advanced Empirical Software Engineering.
[120] Naomi Silver, Matthew Kaplan, Danielle LaVaque-Manty, and Deborah Meizlish. 2023. *Using reflection and metacognition to improve student learning: Across the disciplines, across the academy.* Taylor & Francis.
[121] Karen E Singer-Freeman, Kristi Verbeke, and Betsy Barre. 2025. Generative AI Use Among University Students Depends on Academic Level and Task. *Higher Learning Research Communications* 15, 2 (2025), 8.
[122] SmartPLS Team. 2024. SmartPLS (Version 4.1.0) [Computer software]. https://smartpls.com/ Accessed: 2024-07-07.
[123] Karl A Smith. 1996. Cooperative learning: Making" groupwork" work. *New directions for teaching and learning* 67 (1996).
[124] Edward M Sosu. 2013. The development and psychometric validation of a Critical Thinking Disposition Scale. *Thinking Skills and Creativity* 9 (2013), 107–119.
[125] Rick Spair. 2025. *It Ain't Your Father's Tech Employer Anymore: The Shift in Job Security and Workplace Culture.* LinkedIn. https://www.linkedin.

com/pulse/aint-your-fathers-tech-employer-anymore-shift-job-security-rick-spair-hw3ze/ Accessed: 2026-01-07.

[126] Betsy Sparrow, Jenny Liu, and Daniel M Wegner. 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science* 333, 6043 (2011), 776–778.

[127] Matthias Stadler, Maria Bannert, and Michael Sailer. 2024. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior* 160 (2024), 108386.

[128] Robert J Sternberg and Elena L Grigorenko. 1997. Are cognitive styles still in style? *American Psychologist* 52, 7 (1997), 700.

[129] Klaas-Jan Stol and Brian Fitzgerald. 2018. The ABC of software engineering research. *ACM TOSEM* 27, 3 (2018).

[130] Brian W. Stone. 2024. Generative AI in Higher Education: Uncertain Students, Ambiguous Use Cases, and Mercenary Perspectives. *Teaching of Psychology* (2024). https://www.researchgate.net/publication/387284657_Generative_AI_in_Higher_Education_Uncertain_Students_Ambiguous_Use_Cases_and_Mercenary_Perspectives 41% of students reported using AI in ways explicitly banned by policies.

[131] Mervyn Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Series B Stat Methodol* 36 (1974).

[132] Kate Stringer. 2016. *Finding Success in Failure: STEM Educators Say Student Risk-Taking Is Key to Real-World Learning.* The 74. https://www.the74million.org/article/finding-success-in-failure-stem-educators-say-student-risk-taking-is-key-to-real-world-learning/ Accessed: 2026-01-08.

[133] Artur Strzelecki. 2025. ChatGPT in higher education: Investigating bachelor and master students' expectations towards AI tool. *Education and Information Technologies* 30, 8 (2025), 10231–10255.

[134] Bianca Trinkenreich, Klaas-Jan Stol, Anita Sarma, Daniel M German, Marco A Gerosa, and Igor Steinmacher. 2023. Do I belong? modeling sense of virtual community among linux kernel contributors. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 319–331.

[135] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[136] Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[137] Kathleen Walker and Mihaela Vorvoreanu. 2025. *Learning Outcomes with GenAI in the Classroom: A Review of Empirical Evidence.* Technical Report. Microsoft Aether Psi Working Group.

[138] Ruotong Wang, Ruijia Cheng, Denae Ford, and Thomas Zimmermann. 2023. Investigating and designing for trust in AI-powered code generation tools. *arXiv preprint arXiv:2305.11248* (2023).

[139] Kathryn R Wentzel. 1991. Social competence at school: Relation between social responsibility and academic achievement. *Review of educational research* 61, 1 (1991), 1–24.

[140] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[141] Lixiang Yan, Viktoria Pammer-Schindler, Caitlin Mills, Andy Nguyen, and Dragan Gašević. 2025. Beyond efficiency: Empirical insights on generative AI's impact on cognition, metacognition and epistemic agency in learning. 1675–1685 pages.

[142] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28.

[143] Ling Zhang and Junzhou Xu. 2025. The paradox of self-efficacy and technological dependence: Unraveling generative AI's impact on university students' task completion. *The Internet and Higher Education* 65 (2025), 100978.

[144] Esperanza Zuriguel-Pérez, María-Teresa Lluch-Canut, Montserrat Puig-Llobet, Luis Basco-Prado, Adrià Almazor-Sirvent, Ainoa Biurrun-Garrido, Mariela Patricia Aguayo-González, Olga Mestres-Soler, and Juan Roldán-Merino. 2022. The nursing critical thinking in clinical practice questionnaire for nursing students: A psychometric evaluation study. *Nurse Education in Practice* 65 (2022), 103498.