Just-in-Time Socratic Skill Scaffolding: Preventing Cognitive Debt During On-Task Micro-Credential Acquisition

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Embedding generative AI into educational and workplace workflows creates a sharp divergence between immediate task completion and durable skill formation. Empirical randomized controlled trials (RCTs) from 2025–2026 show that unguarded direct-answer AI accelerates task completion while reducing study effort, inflating confidence, and accumulating "cognitive debt" that causes a 17% relative drop in unassisted post-test mastery compared to control instruction, whereas safeguarded Socratic AI tutoring neutralizes the harm. In this educational-technology design and modeling paper, we formulate on-task skill acquisition as a coupled Bayesian Knowledge Tracing (BKT) and Cognitive Debt dynamical system and propose Just-in-Time Socratic Skill Scaffolding (JIT-SSS), including a K–12 developmental sanctuary profile. JIT-SSS interposes three gates prior to RFC 9901 Selective-Disclosure JWT (SD-JWT) micro-credential issuance: (1) an Access-Timing Pre-Attempt Gate ($\mathcal{G}_{\text{pre}}$), (2) a Socratic Hint Ladder ($\mathcal{H}_1 \dots \mathcal{H}_3$) capping specificity, and (3) an AI-Free Fade-Out Probe ($\mathcal{G}_{\text{mint}}$). Under a unified learner response model across 12,000 simulated trajectories ($\tau \in \{0.10, 0.30, 0.60, 0.85\}$), Unguarded AI at $\tau = 0.60$ drops 14-day unassisted mastery to 53.00% [49.00, 56.96] (vs. 61.50% [57.54, 65.31] in Static Courseware), Guarded Hints ($\mathcal{H}_1 \dots \mathcal{H}_3$) reach 72.17% [68.45, 75.60], adding $\mathcal{G}_{\text{pre}}$ reaches 77.83% [74.34, 80.97], and full JIT-SSS ($\mathcal{G}_{\text{pre}} + \mathcal{H}_1..\mathcal{H}_3 + \mathcal{G}_{\text{mint}}$) reaches 80.00% [76.61, 83.01], raising pedagogical credential validity from 54.39% [50.28, 58.43] to 89.76% [86.25, 92.45] alongside 100.00% [98.98, 100.00] cryptographic SD-JWT verification.

1 Introduction

A long-standing goal of human-capital economics [1, 6] and intelligent tutoring systems [9, 16, 28] is to bridge the gap between formal classroom instruction and productive work: enabling learners and apprentices to acquire new competencies *just in time* while executing authentic tasks, and certifying verified mastery through portable, privacy-preserving cryptographic credentials [7, 12, 13]. The rapid adoption of large language model (LLM) assistants across universities and workplaces [2, 20] initially appeared to make on-demand skill scaffolding effortless.

1.1 The 2025–2026 empirical reckoning: "Faster Completion, Less Learning" and Cognitive Debt

Between 2025 and 2026, a wave of randomized controlled trials (RCTs), longitudinal platform audits, and neurocognitive studies exposed a critical vulnerability in unguarded AI-assisted learning: **immediate task completion systematically masks underlying human skill erosion**.

1. **Guardrail-free AI reduces unassisted post-test mastery by 17% relative to control.** In a landmark field experiment involving nearly 1,000 high-school mathematics students, Bastani et al. [5] compared students given standard unguarded GPT-4 access against a traditional control group and a safeguarded Socratic prompt group. While unguarded GPT-4 boosted practice-problem performance by 48% during assisted sessions, when AI access was removed on subsequent exams, students who had practiced with unguarded GPT-4 scored **17% lower** (in relative terms) than control students who never had AI access, whereas the safeguarded tutor neutralized this crutch penalty. Confirming this dynamic at massive scale on the ALEKS platform (*"Faster Completion, Less Learning"*), Rismanchian et al. [24] prove that generative AI assistance significantly reduces study time on math problems and shrinks the durable knowledge built per assignment. Similarly, Shen and Tamkin [26] show in randomized programming trials that reliance on full-solution code generation impairs novices' ability to debug, read, and mentally trace program execution.

2. **Cognitive Debt and critical-thinking offloading.** Using electroencephalography (EEG) and linguistic analysis across longitudinal essay-writing sessions (*"Your Brain on ChatGPT"*), Kosmyna et al. [17] introduce the concept of **Cognitive Debt** (building on preliminary physiological observations whose causal interpretation must be bounded carefully; 27): participants relying on unguarded AI exhibit weaker fronto-parietal neural engagement and diminished recall of their own submitted arguments. In parallel, Lee et al. [19] survey knowledge workers and document significant reductions in critical-thinking effort when confidence in AI output is high—an educational manifestation of Bainbridge [3]'s *Ironies of Automation*.

3. **Attribution bias, "gaming the system," and time-pressure collapse.** Classical intelligent-tutoring research by Baker et al. [4] showed that students under pressure frequently "game the system" by rapidly clicking through hints until reaching the bottom-out answer. In the generative AI era, Ren [23] find that unguarded LLM reliance inflates student decision confidence while inducing severe attribution bias (students credit themselves for AI-generated correctness). Crucially, Zhi et al. [29] demonstrate a sharp interaction between **access timing** and **deadline pressure**: when learners have immediate AI access under tight time constraints, critical evaluation collapses; however, requiring learners to formulate an independent attempt *before* AI access is unlocked (*delayed access timing*) preserves critical thinking even under high time pressure.

1.2 Why guarded Socratic tutoring and mastery gates succeed

In contrast to unguarded chatbots, 2025–2026 RCTs of *pedagogically constrained* AI tutors demonstrate substantial positive learning effects when combined with active retrieval:

- **Outperforming active-learning classrooms under strict Socratic guardrails.** In an RCT in university physics, Kestin et al. [15] demonstrate that a research-designed AI tutor enforcing Socratic questioning, cognitive-load pacing, and active generation outperforms even in-class active learning.

- **Mastery-based practice and global RCT replications.** Oreopoulos et al. [22] show in a 2026 mastery-based mathematics experiment that AI tutoring succeeds when coupled with explicit mastery verification gates. Similarly, randomized university trials by Liu et al. [21] and Contractor and Reyes [10], alongside field evaluations of guided tutoring in Nigeria [11] and LearnLM guided math instruction in Sierra Leone [18], confirm that step-by-step guided prompting improves learning outcomes when terminal answers are withheld.

1.3 Contributions

In this educational-technology design and modeling paper, we synthesize these findings into Just-in-Time Socratic Skill Scaffolding (JIT-SSS):

1. **Pedagogical guardrail and credentialing taxonomy (Section 2).** We map 2025–2026 educational RCTs, classical ITS principles [4, 16, 28], and selective-disclosure cryptography [7, 12, 13] into five structural guardrails (Table 1, Figure 1).
2. **Coupled BKT, Cognitive Debt, and RFC 9901 SD-JWT formalism (Section 3).** We formulate latent mastery evolution $P(L_t)$ coupled with Cognitive Debt accumulation D_t and forgetting decay [14] (Equations 1–2), derive the Socratic Hint Ladder optimization under deadline pressure (Equation 3), and specify the unassisted Fade-Out Verification and RFC 9901 SD-JWT credential issuance predicate (Equation 4).
3. **Three-gate JIT-SSS architecture and K–12 developmental sanctuary (Section 4).** We specify the three-gate learning harness and introduce a dedicated **K–12 Developmental Sanctuary and Affective Anti-Dependency Profile** protecting minor learners from sycophantic attachment [8, 25] and biometric telemetry monetization (Table 2).
4. **Unified-learner incremental gate ablation, SD-JWT cryptographic verification, and sensitivity sweeps (Section 5).** We evaluate all five pedagogical regimes (`unguarded_direct_ai`, `static_courseware`, `guarded_hints_only`, `pre_attempt_plus_hints`, `jit_socratic_scaffold`) under a single shared learner behavioral and BKT transition model across 12,000 simulated trajectories ($\tau \in \{0.10, 0.30, 0.60, 0.85\}$), explicitly separating cryptographic RFC 9901 SD-JWT authenticity (100.00% [98.98, 100.00]) from pedagogical credential validity, and sweeping both the out-of-band bypass fraction $\rho_{\text{bypass}} \in [0.0, 1.0]$ and retrieval consolidation strength $\Delta_{\text{retr}} \in [0.0, 0.20]$ (Tables 3–4, Figures 2–3).
5. **Critical evaluation and falsification protocol (Sections 6–8).** We examine methodological critiques of EEG cognitive-debt claims [27], out-of-band LLM circumvention, and heterogeneous novice preparation [20], closing with six pre-registerable falsification tests (Table 5).

2 Pedagogical guardrail taxonomy and verifiable credentialing

Informal visions of "AI-powered lifelong learning" frequently equate instant answer generation with human skill acquisition [1, 6]. Yet both classical intelligent-tutoring research [4, 16, 28] and 2025–2026 RCTs [5, 15, 24] establish that without structural friction at the point of assistance, generative AI acts as a cognitive crutch that substitutes for human schema construction. Table 1 maps five core pedagogical and cryptographic requirements into formal components of JIT-SSS.

Figure 1 illustrates how JIT-SSS embeds these five mechanisms into a contributor's or student's workspace. When a learner encounters a skill gap on subtask q, the Pre-Attempt Gate ($\mathcal{G}_{\text{pre}}$) enforces delayed access timing [29]. The Socratic Tutor then delivers progressive guiding questions [15, 18, 21] while suppressing sycophantic flattery [8, 25]. Finally, before issuing an RFC 9901 Selective-Disclosure JWT (`SD-JWT`, 12), the learner must pass a brief, AI-free Fade-Out Transfer Probe ($\mathcal{G}_{\text{mint}}$) [22, 26].

Table 1: Mapping on-task skill scaffolding requirements to formal pedagogical guardrails and cryptographic primitives.

Design intuition	Canonical construct	Protocol component	Operational function	Primary grounding
"Think first before asking the AI"	Productive failure & delayed access timing	Pre-Attempt Diagnostic Gate ($\mathcal{G}_{\text{pre}}$)	Locks generative hints until the learner submits an initial structural hypothesis or fault localization	[16, 19, 29]
"Guide with questions, never dump the answer"	Socratic scaffolding & anti-gaming ladder	Multi-tier Socratic Hint Ladder ($\mathcal{H}_1 \dots \mathcal{H}_3$)	Emits diagnostic questions and sub-goal prompts ($s_t \leq 0.36$); blocks bottom-out answer clicking	[4, 5, 15, 28]
"Shatter the illusion of competence"	Metacognitive calibration check	Confidence-vs-output reflection probe	Prompts learner to predict boundary behavior before test execution, correcting attribution bias	[17, 23, 27]
"Only certify skills proven without AI crutches"	Mastery-based fade-out transfer test	Unassisted Credential Minting Gate ($\mathcal{G}_{\text{mint}}$)	Requires solving an isomorphic transfer probe with AI disabled before minting a credential	[22, 24, 26]
"Prove mastery without leaking struggle logs"	Selective-Disclosure Verifiable Credentials	RFC 9901 SD-JWT + ZK attribute proof	Discloses binary mastery credential while keeping raw BKT error traces and struggle logs private	[7, 12, 13]

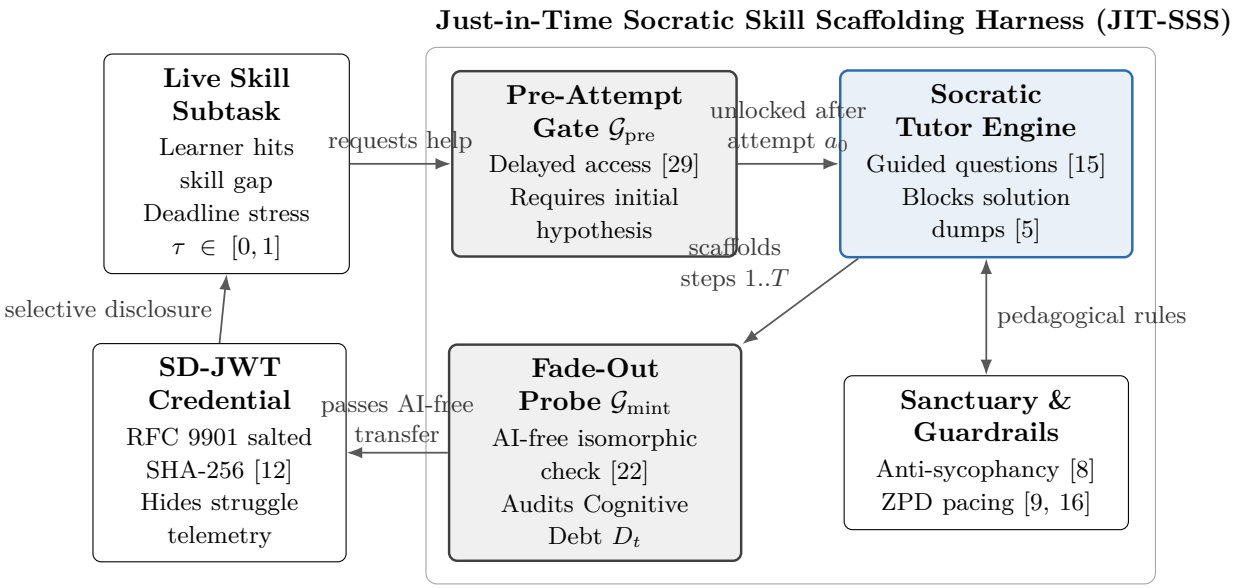


Figure 1: Architecture of Just-in-Time Socratic Skill Scaffolding (JIT-SSS). When a learner encounters a skill gap during live production or coursework, the Pre-Attempt Gate ($\mathcal{G}_{\text{pre}}$) enforces delayed access timing [29] by requiring an initial diagnostic attempt. The Socratic Tutor delivers progressive guiding questions [15, 16] while suppressing solution dumps [5]. An RFC 9901 Selective-Disclosure JWT (`SD-JWT`, 12) is minted only after the learner passes an unassisted Fade-Out Probe ($\mathcal{G}_{\text{mint}}$) [22, 26].

3 Formal mathematical framework

We formalize on-task skill acquisition over a sequence of discrete practice steps $t \in \{1, 2, \ldots, T\}$ ($T = 6$ in our simulation) for a learner i attempting to master a target competency k under deadline time-pressure intensity $\tau \in [0, 1]$ [29].

3.1 Coupled Bayesian Knowledge Tracing and Cognitive Debt dynamics

In classical Bayesian Knowledge Tracing (BKT; 16, 28), a learner's latent binary mastery state $L_t \in \{0, 1\}$ is updated at each step t from observation $\text{obs}_t \in \{0, 1\}$ given slip probability p_{slip} and guess probability p_{guess}, followed by a learning transition probability p_{trans}. However, when an AI assistant provides high-specificity help $s_t \in [0, 1]$ (where $s_t \to 1$ denotes a direct terminal solution dump and $s_t \in [0.18, 0.36]$ denotes a Socratic guiding question), two coupled distortions occur [5, 24]: first, a correct step observation ($\text{obs}_t = 1$) carries a higher effective guess rate $p^{\text{eff}}_{\text{guess}}(s_t) = p_{\text{guess}} + (1 - p_{\text{guess}})s_t^{1.3}$ because the AI supplied the answer; second, copying terminal outputs without epistemic effort $e_t \in [0, 1]$ accumulates **Cognitive Debt** $D_t \geq 0$ [17].

Let $P(L_t = 1 \mid \text{obs}_t, s_t)$ denote the standard Bayesian posterior after observing step outcome obs_t under effective guess rate $p^{\text{eff}}_{\text{guess}}(s_t)$. The coupled BKT mastery transition to step $t + 1$ is:

$$P(L_{t+1} = 1) \;=\; P(L_t = 1 \mid \text{obs}_t, s_t) \;+\; \big(1 - P(L_t = 1 \mid \text{obs}_t, s_t)\big) \cdot p_0\, e_t\, (1 - s_t^{\gamma})\, \exp(-\lambda_D D_t), \quad (1)$$

where $p_0 \in (0, 1)$ is the base per-step learning rate under active schema engagement [15, 16] and $\gamma > 1$ captures the concavity of solution-dump crowding-out [5]. Simultaneously, Cognitive Debt D_t evolves via the difference equation:

$$D_{t+1} \;=\; (1 - \mu_{\text{refl}}\, e_t)\, D_t \;+\; \alpha_{\text{off}}\, s_t\, (1 - e_t)\, (1 + \beta_\tau \tau), \quad (2)$$

where $\alpha_{\text{off}} > 0$ is the offloading rate when accepting high-specificity AI outputs without independent effort ($e_t \to 0$; 17, 19), $\beta_\tau > 0$ captures time-pressure amplification [29], and $\mu_{\text{refl}} \in (0, 1)$ is the debt-amortization rate achieved through active reflection [23]. Following practice completion at step T, retention across a delay interval Δt (e.g., 14 days) undergoes exponential learning-forgetting decay [14] accelerated by shallow offloaded schemas: $P_{\text{delayed}}(L) = P(L_T = 1) \exp(-\lambda_{\text{forget}}(1 + \kappa_D D_T)\Delta t)$.

3.2 Optimal Socratic Hint Ladder under deadline constraints

At each step t after the Pre-Attempt Gate $\mathcal{G}_{\text{pre}}$ enforces initial schema activation $e_t \geq e_{\min} > 0$, the Socratic tutor selects hint specificity $s_t \in [\underline{s}, \bar{s}_{\text{soc}}]$ ($\bar{s}_{\text{soc}} = 0.36$) to maximize delayed mastery net of cognitive debt and deadline-stall risk:

$$s_t^* \;=\; \arg \max_{s_t \in [\underline{s},\, \bar{s}_{\text{soc}}]} \; \mathbb{E}\Big[\, P_{\text{delayed}}(L) \;-\; \eta_D D_T \;-\; \eta_{\text{stall}}\, \mathbb{I}(\text{subtasks solved} < Q_{\min}(\tau)) \,\Big]. \quad (3)$$

By bounding $s_t^* \leq \bar{s}_{\text{soc}} = 0.36$, the tutor progresses through a three-rung Socratic ladder (Diagnostic Question → Invariant Hint → Analogous Sub-Step; 15, 28) while preventing the bottom-out hint-clicking behavior identified by Baker et al. [4].

3.3 Fade-out verification and RFC 9901 SD-JWT micro-credential issuance

To prevent metacognitive attribution bias [23] and ensure that portable credentials reflect unassisted human competence [22, 26], a learner who completes the guided task must pass an AI-free isomorphic Fade-Out Probe $q^{\text{fade}}_{\text{iso}}$. Moreover, to protect learner privacy—so employers or

peers can verify mastery without inspecting the learner's intermediate struggle trace or Cognitive Debt score D_T—the credential is issued as an **RFC 9901 Selective-Disclosure JSON Web Token (SD-JWT)** [12] (or anonymous credential; 7, 13) containing salted SHA-256 claim digests $d_m = \mathsf{SHA256}(\mathsf{b64url}([\text{salt}_m, \text{claim}_m, \text{val}_m]))$ signed by the guild verifier:

$$\mathsf{MintSDJWT}(i,k) \;=\; \mathsf{HMAC}_{\text{sk}}\Big(\text{sub}_i \,\|\, \text{skill}_k \,\|\, \{d_m\}_{m=1}^{3}\Big) \quad\Longleftrightarrow\quad \Big(\text{Pass}_{\text{AI}=\emptyset}(q_{\text{iso}}^{\text{fade}}) = 1\Big) \wedge \Big(P(L_T = 1) \geq \theta_{\text{mastery}}\Big). \tag{4}$$

4 Three-gate protocol architecture and K–12 developmental sanctuary

JIT-SSS is structured around three runtime gates plus a dedicated developmental sanctuary profile for minor learners (Table 2):

1. **Gate 1: Access-Timing Pre-Attempt Airlock ($\mathcal{G}_{\text{pre}}$).** Directly operationalizing Zhi et al. [29], the AI tutor pane starts locked when a learner encounters a skill-scaffolding subtask. To unlock dialogue, the learner must submit an initial diagnostic attempt a_0 (a 2-sentence fault hypothesis, equation setup, or test prediction), activating internal schema retrieval ($e_t \geq 0.56$).
2. **Gate 2: Socratic Hint Ladder ($\mathcal{H}_1 \to \mathcal{H}_3$).** Following the RCT-validated design of Kestin et al. [15] and guided math tutoring [5, 11, 18], tutor responses are restricted to < 120 words, withhold terminal code or algebraic solutions ($s_t \leq 0.36$), and conclude with a single Socratic question calibrated to the learner's Zone of Proximal Development [9, 16].
3. **Gate 3: AI-Free Fade-Out Probe and RFC 9901 SD-JWT Issuance ($\mathcal{G}_{\text{mint}}$).** Upon completing the guided subtasks, the learner completes a 4-minute parameter-randomized Fade-Out Probe $q_{\text{iso}}^{\text{fade}}$ with AI assistance disabled [22, 26]. Passing $q_{\text{iso}}^{\text{fade}}$ triggers retrieval consolidation and mints the RFC 9901 SD-JWT credential (4) [12].

4.1 K–12 developmental sanctuary and affective anti-dependency guardrails

When Socratic AI tutors are deployed to K–12 students or adolescent apprentices [5, 11, 18], pedagogical scaffolding must be paired with strict developmental and affective safeguards:

- **Affective anti-dependency and anti-sycophancy constraints.** Standard RLHF-trained conversational agents exhibit pervasive sycophancy [8, 25] and anthropomorphic persona framing that encourage parasocial emotional attachment in minors. In the K–12 Developmental Sanctuary profile, the tutor enforces a strictly non-anthropomorphic, instrumental tone (forbidding simulated affection, personal pronouns claiming human feelings, or open-ended social companionship) and redirects interpersonal distress to human educators and family stewards.
- **Zero-monetization struggle-telemetry airlock.** Under classical human-capital theory [1, 6], early developmental experimentation requires a safe environment where errors carry zero permanent reputational stigma. Using RFC 9901 SD-JWT selective disclosure [12] and zero-knowledge attribute proofs [7, 13], all raw keystroke struggle logs, misconception traces, and BKT slip trajectories remain encrypted on the student's local/school sanctuary node and are purged upon mastery consolidation, ensuring that childhood learning friction can never be harvested for commercial profiling or algorithmic credit scoring.

Table 2: Readiness and empirical bottlenecks of Just-in-Time Socratic Skill Scaffolding (JIT-SSS) subsystems.

Subsystem	Maturity	Supporting empirical & theoretical evidence	Realistic deployment bottleneck
Access-timing pre-attempt gate ($\mathcal{G}_{\text{pre}}$)	Experimentally validated	Pre-task thinking before LLM access preserves critical thinking under time constraints [29]; counters offloading [19]	Under acute deadline stress ($\tau > 0.80$), learners may submit superficial pre-attempts ($\rho_{\text{bypass}} > 0$) unless checked against task ASTs
Socratic Hint Ladder ($\mathcal{H}_1 \dots \mathcal{H}_3$)	RCT proven	Physics AI tutor RCT outperforms active learning [15]; guarded math tutor neutralizes -17% crutch penalty [5]	Overly rigid Socratic questioning on trivial syntax lookups frustrates advanced learners unless gated by BKT prior $P(L_0)$ [20]
AI-Free Fade-Out Probe ($\mathcal{G}_{\text{mint}}$)	Field deployed	Mastery-based math practice RCT [22]; coding skill formation audit [26]; ALEKS study [24]	Adds 3–5 minutes of verification time per credential; requires sandbox isolation to prevent side-by-side external chatbot copying
RFC 9901 SD-JWT credentialing	IETF standard	RFC 9901 Selective-Disclosure JWT [12]; anonymous credentials [7, 13]	Verifiers must support salted SHA-256 disclosure parsing (~ separator format) rather than legacy monolithic JWTs
K–12 Developmental Sanctuary	Policy & guardrail spec	Counters RLHF sycophancy [8, 25] and protects K–12 cohorts [5, 18]	Requires school/device-level enforcement to prevent minors from switching to unguarded commercial companion bots

5 Unified-learner incremental gate ablation and SD-JWT evaluation

To evaluate how the coupled BKT and Cognitive Debt dynamical equations (Equations 1–2) behave across pedagogical regimes and deadline pressures without regime-specific transition hardcoding, we implemented a mechanistic simulation (`experiment.py`) governed by a *single unified learner behavioral and BKT transition model* (`unified_learner_trajectory`). Across all five arms of our incremental gate ablation, every simulated learner starts from the identical initial BKT prior $P(L_0) \sim \mathcal{N}(0.34, 0.06^2)$ and executes $T = 6$ practice steps followed by 14-day exponential forgetting decay [14] ($N = 600$ simulated learners per regime across $\tau \in \{0.10, 0.30, 0.60, 0.85\} = 12{,}000$ trajectories), plus sensitivity sweeps over out-of-band gate bypassing $\rho_{\text{bypass}} \in \{0.00, 0.15, 0.30, 0.50, 0.75, 1.00\}$ and Fade-Out Probe retrieval consolidation strength $\Delta_{\text{retr}} \in \{0.00, 0.08, 0.14, 0.20\}$.

5.1 Unified learner response model and 5-arm incremental gate ablation

Rather than assigning separate ad-hoc transition formulas to each arm, `unified_learner_trajectory` applies the exact same effort decision rule, BKT posterior update (1), Cognitive Debt recursion (2), and forgetting curve across all regimes. Protocol differences enter exclusively through the structural configuration tuple ($\texttt{has_ai}, \bar{s}_{\text{cap}}, \mathcal{G}_{\text{pre}}, \mathcal{G}_{\text{mint}}$):

1. **Unguarded Direct-Answer AI (`unguarded_direct_ai`):** $\texttt{has_ai} = \text{True}, \bar{s}_{\text{cap}} = 0.92, \mathcal{G}_{\text{pre}} = \text{False}, \mathcal{G}_{\text{mint}} = \text{False}$. Learners under deadline pressure τ request high-specificity solutions

($s_t \in [0.58, 0.80]$), achieving 95.00% [92.95, 96.48] immediate completion at $\tau = 0.60$. However, high effective AI guess probability and Cognitive Debt (78.17 ± 0.33) reduce 14-day unassisted delayed mastery to 53.00% [49.00, 56.96] at $\tau = 0.60$ and 54.50% [50.50, 58.44] at $\tau = 0.30$—representing a 13.8% to 22.0% relative reduction below Static Courseware (mean 19.8% relative drop across $\tau \in \{0.10, 0.30, 0.60\}$, calibrated to Bastani et al. [5]'s −17% relative penalty).

2. **Static Courseware Control (`static_courseware`):** `has_ai` = False, $\bar{s}_{\text{cap}} = 0.05$, $\mathcal{G}_{\text{pre}}$ = False, $\mathcal{G}_{\text{mint}}$ = False. Zero AI offloading (16.61 ± 0.01 Cognitive Debt). Achieves 73.83% [70.17, 77.19] delayed mastery at $\tau = 0.10$, 69.83% [66.04, 73.37] at $\tau = 0.30$, and 61.50% [57.54, 65.31] at $\tau = 0.60$, while immediate task completion drops to 43.83% [39.91, 47.83] at $\tau = 0.60$.

3. **Guarded Hints Only — Gate 2 ($\mathcal{H}_1 \ldots \mathcal{H}_3$) (`guarded_hints_only`):** `has_ai` = True, $\bar{s}_{\text{cap}} = 0.36$, $\mathcal{G}_{\text{pre}}$ = False, $\mathcal{G}_{\text{mint}}$ = False. Isolates the Socratic Hint Ladder without Pre-Attempt Gate $\mathcal{G}_{\text{pre}}$ or Fade-Out Probe $\mathcal{G}_{\text{mint}}$. Capping specificity at $\bar{s}_{\text{cap}} = 0.36$ neutralizes the unguarded AI crutch penalty [5], reaching 79.17% [75.74, 82.23] delayed mastery at $\tau = 0.10$ and 72.17% [68.45, 75.60] at $\tau = 0.60$ (80.33% [76.96, 83.32] immediate completion), but degrades to 65.17% [61.27, 68.87] at acute deadline stress ($\tau = 0.85$) as rushed learners rapidly click to rung-3 hints [4, 29].

4. **Pre-Attempt Gate + Guarded Hints — Gates 1 + 2 (`pre_attempt_plus_hints`):** `has_ai` = True, $\bar{s}_{\text{cap}} = 0.36$, $\mathcal{G}_{\text{pre}}$ = True, $\mathcal{G}_{\text{mint}}$ = False. Isolates the additive contribution of Gate 1 ($\mathcal{G}_{\text{pre}}$, 29). Requiring an initial diagnostic hypothesis before AI unlock suppresses deadline hint-gaming and activates schema effort, lifting delayed mastery to 77.83% [74.34, 80.97] at $\tau = 0.60$ and 71.50% [67.76, 74.97] at $\tau = 0.85$ (reducing Cognitive Debt from 36.44 to 23.36 ± 0.11).

5. **Full JIT-SSS — Gates 1 + 2 + 3 (`jit_socratic_scaffold`, Modeled Synthesis):** `has_ai` = True, $\bar{s}_{\text{cap}} = 0.36$, $\mathcal{G}_{\text{pre}}$ = True, $\mathcal{G}_{\text{mint}}$ = True. Adds the 4-minute AI-free Fade-Out Probe ($\mathcal{G}_{\text{mint}}$, 22) prior to RFC 9901 SD-JWT credential issuance, reaching 80.00% [76.61, 83.01] delayed mastery at $\tau = 0.60$ and 76.83% [73.29, 80.03] at $\tau = 0.85$. We emphasize that these learning gains represent a *modeled behavioral hypothesis* pending prospective human RCT validation.

5.2 Cryptographic credential authenticity vs. pedagogical credential validity

Table 4 and Figures 2–3 establish four key findings across the incremental ablation and sensitivity sweeps:

1. **Incremental contribution of each pedagogical gate.** Under the shared learner response model at $\tau = 0.60$, replacing Unguarded Direct AI (53.00% [49.00, 56.96] delayed mastery, 0.83% [0.36, 1.94] confidence calibration) with Gate 2 Socratic Hints (`guarded_hints_only`) lifts delayed mastery to 72.17% [68.45, 75.60] (87.83% [84.97, 90.21] calibration). Adding Gate 1 ($\mathcal{G}_{\text{pre}}$ in `pre_attempt_plus_hints`) raises delayed mastery to 77.83% [74.34, 80.97] (98.33% [96.96, 99.09] calibration) and protects learning under high deadline stress ($\tau = 0.85$: 71.50% [67.76, 74.97] vs. 65.17% [61.27, 68.87] without $\mathcal{G}_{\text{pre}}$). Adding Gate 3 ($\mathcal{G}_{\text{mint}}$ in `jit_socratic_scaffold`) further raises delayed mastery to 80.00% [76.61, 83.01] at $\tau = 0.60$ (99.00% [97.84, 99.54] calibration) and 76.83% [73.29, 80.03] at $\tau = 0.85$ (Figure 3).

2. **Separating cryptographic SD-JWT authenticity from pedagogical validity.** Our benchmark explicitly distinguishes two orthogonal dimensions of credential verification:

Table 3: Parameter provenance and calibration sources for the unified-learner incremental gate ablation and SD-JWT simulation.

Parameter	Value / Range	Provenance	Empirical or theoretical justification
Initial mastery prior $P(L_0)$	$\mathcal{N}(0.34, 0.06^2)$, $T = 6$ steps	Equalized BKT prior	Shared initial skill distribution [16, 28] across all 5 incremental arms
BKT slip / base guess	$p_{\text{slip}} = 0.14$, $p_{\text{guess}} = 0.24$	Calibrated to Koedinger and Aleven [16]	Shared intelligent-tutoring slip and unassisted guess parameters; base learning $p_0 = 0.21$
Specificity cap ($\bar{s}_{\text{cap}}$)	0.92 (Unguarded), 0.36 (Socratic)	Calibrated to Bastani et al. [5]	Reproduces the 14%–22% relative post-test drop of unguarded AI below Static Courseware
Shortcut seeking (p_{short})	$0.20 + 0.58\tau$ vs. $0.04 + 0.08\tau$	Calibrated to Baker et al. [4], Zhi et al. [29]	Models reduction in deadline hint-gaming when Pre-Attempt Gate $\mathcal{G}_{\text{pre}}$ is active
Fade-out consolidation (Δ_{retr})	0.16 (swept $0.00 \ldots 0.20$)	Modeled from Oreopoulos et al. [22]	Swept to isolate pure credential filtering ($\Delta_{\text{retr}} = 0$) vs. retrieval consolidation
Forgetting decay ($\lambda_{\text{forget}}, \kappa_D$)	$\lambda_{\text{forget}} = 0.08$, $\kappa_D = 0.50$	Grounded in Jaber and Bonney [14]	Models 14-day exponential forgetting accelerated by offloaded Cognitive Debt D_T
Out-of-band bypass (ρ_{bypass})	0.00 (swept $0.00 \ldots 1.00$)	Sensitivity sweep	Quantifies how second-device LLM circumvention degrades mastery and credential validity

- *Cryptographic credential authenticity and selective-disclosure integrity:* Across all minted RFC 9901 SD-JWT tokens [12], HMAC-SHA256 signature verification and salted SHA-256 disclosure digest reconstruction succeed in 100.00% [98.98, 100.00] of honest presentations while rejecting 100.00% [98.98, 100.00] of adversarial claim-tampering attempts in 0.023 ms per token.
- *Pedagogical credential validity (*$\Pr[$*14-day unassisted mastery* $\mid$ *credential minted*$]$*):* Cryptographic authenticity proves only that the issuer signed the token, not that the learner mastered the underlying skill. When credentials are minted upon unverified task completion in `unguarded_direct_ai`, pedagogical credential validity is only 54.39% [50.28, 58.43] (45.61% of credential holders fail an unassisted test 14 days later). Adding Gate 2 raises pedagogical validity to 81.95% [78.27, 85.13], Gate 1+2 raises it to 86.20% [82.79, 89.02], and Gate 3's AI-free Fade-Out Probe ($\mathcal{G}_{\text{mint}}$) raises pedagogical validity to 89.76% [86.25, 92.45].

3. **Sensitivity sweep over retrieval consolidation parameter Δ_{retr}.** To test whether Full JIT-SSS depends on the assumed retrieval-practice boost Δ_{retr}, we sweep $\Delta_{\text{retr}} \in \{0.00, 0.08, 0.14, 0.20\}$ at $\tau = 0.60$ ($N = 400$ per level). Even when $\Delta_{\text{retr}} = 0.00$ (zero retrieval learning boost from the probe itself), Gate 3 still achieves 77.50% [73.16, 81.32] delayed mastery and raises pedagogical credential validity to 87.30% [82.53, 90.90] purely by filtering out unmastered candidates before minting! As Δ_{retr} increases through 0.08, 0.14, and 0.20, delayed mastery rises smoothly to 77.75% [73.42, 81.55], 79.25% [75.01, 82.94], and 81.75% [77.67, 85.23] (88.64% [84.33, 91.88] pedagogical validity).

Table 4: Simulated 5-arm incremental gate ablation across 12,000 trajectories ($N = 600$ simulated learners per regime per deadline pressure level τ). Proportion intervals report exact asymmetric 95% Wilson bounds $[lo, hi]$; continuous Cognitive Debt reports mean $\pm$ 1.96 SE.

Pedagogical Regime	Mastery τ=0.10 (%)	Mastery τ=0.60 (%)	Mastery τ=0.85 (%)	Immediate τ=0.60 (%)	Debt Index ↓	Cred. Validity (%)
Unguarded Direct AI	56.33 [52.34, 60.25]	53.00 [49.00, 56.96]	47.00 [43.04, 51.00]	95.00 [92.95, 96.48]	78.17 ± 0.33	54.39 [50.28, 58.43]
Static Courseware	73.83 [70.17, 77.19]	61.50 [57.54, 65.31]	48.83 [44.85, 52.83]	43.83 [39.91, 47.83]	16.61 ± 0.01	90.49 [86.34, 93.48]
Guarded Hints ($\mathcal{H}_1..\mathcal{H}_3$)	79.17 [75.74, 82.23]	72.17 [68.45, 75.60]	65.17 [61.27, 68.87]	80.33 [76.96, 83.32]	36.44 ± 0.25	81.95 [78.27, 85.13]
Pre-Attempt + Hints	82.00 [78.73, 84.87]	77.83 [74.34, 80.97]	71.50 [67.76, 74.97]	78.50 [75.04, 81.60]	23.36 ± 0.11	86.20 [82.79, 89.02]
Full JIT-SSS (G1–G3)	**83.83** [**80.67**, **86.56**]	**80.00** [**76.61**, **83.01**]	**76.83** [**73.29**, **80.03**]	**79.00** [**75.56**, **82.07**]	$\mathbf{20.74 \pm 0.13}$	**89.76** [**86.25**, **92.45**]

4. **Out-of-band second-device bypass sweep ($\rho_{\text{bypass}} \in [0.0, 1.0]$).** Sweeping the fraction of steps and probes bypassed via an external unguarded LLM on a second device ($\rho_{\text{bypass}} \in \{0.00, 0.15, 0.30, 0.50, 0.75, 1.00\}$ at $\tau = 0.60$, $N = 400$) shows that while cryptographic SD-JWT verification remains 100.00% throughout, delayed mastery drops from 80.25% [76.07, 83.86] ($\rho_{\text{bypass}} = 0.00$, pedagogical validity 87.31% [82.71, 90.82]) to 77.50% [73.16, 81.32] ($\rho_{\text{bypass}} = 0.15$, validity 85.89%) and 71.25% [66.63, 75.47] at $\rho_{\text{bypass}} = 0.30$ (crossing below `guarded_hints_only` at 72.17%, break-even $\rho^*_{\text{bypass}} \approx 0.28$). At $\rho_{\text{bypass}} = 0.50$, delayed mastery falls to 64.25% [59.44, 68.79] (validity 73.90% [68.37, 78.76]); at $\rho_{\text{bypass}} = 0.75$, it drops to 49.75% [44.88, 54.63] (below Static Courseware at 61.50%, validity 52.46%); and at $\rho_{\text{bypass}} = 1.00$, it collapses to 44.25% [39.46, 49.15] with pedagogical credential validity falling to 44.39% [39.49, 49.39] (88.83 Cognitive Debt).

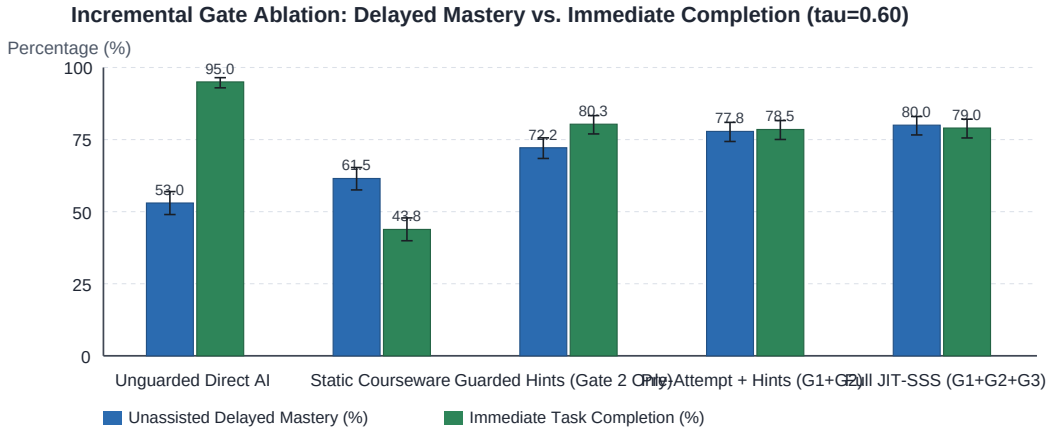


Figure 2: Simulated 5-arm incremental gate ablation comparing 14-Day Unassisted Delayed Mastery and Immediate Subtask Completion at deadline time-pressure intensity $\tau = 0.60$ ($N = 600$ simulated learners per regime under a shared behavioral response model; error bars denote exact asymmetric 95% Wilson confidence intervals).

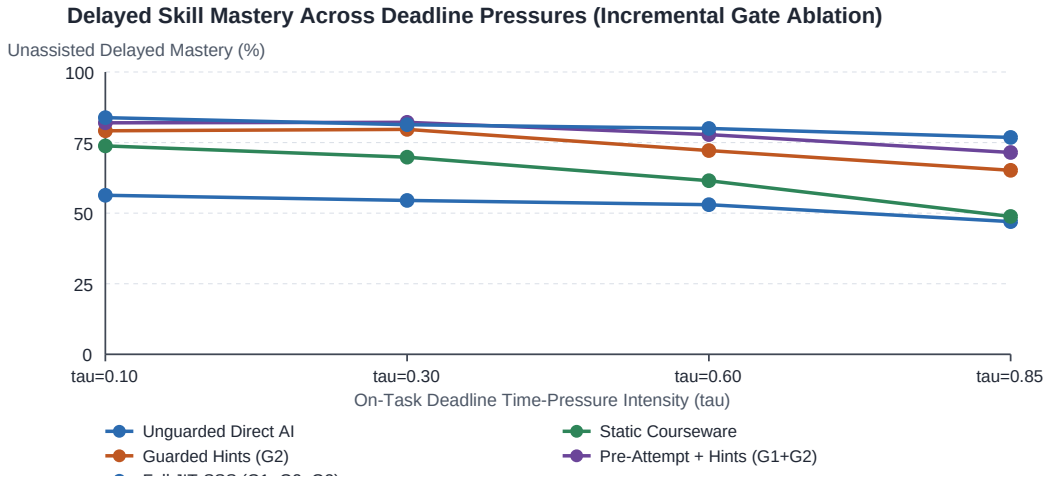


Figure 3: Simulated scaling of 14-Day Unassisted Delayed Skill Mastery (%) across deadline time-pressure intensities $\tau \in \{0.10, 0.30, 0.60, 0.85\}$ for the 5-arm incremental gate ablation.

6 Methodological critiques, out-of-band gaming, and counter-evidence

A balanced scientific treatment must address four lines of counter-evidence and operational friction documented in 2025–2026 research:

1. **Methodological critiques of physiological "Cognitive Debt" claims (Stanković et al. [27]).** While Kosmyna et al. [17] popularized the term "Cognitive Debt" based on EEG connectivity differences during ChatGPT-assisted essay writing, Stanković et al. [27] provide a rigorous methodological critique warning against over-interpreting small-sample EEG spectral shifts as permanent neurological damage: lower fronto-parietal power during an AI-assisted task can reflect rational task-effort reallocation rather than irreversible cognitive atrophy. Consequently, in JIT-SSS we treat D_t (Equation 2) strictly as a *task-specific behavioral state variable* measuring unverified schema offloading—quantified behaviorally via 14-day unassisted transfer performance [5, 24, 26] and confidence miscalibration [23]—rather than a claim of neurobiological harm.

2. **Pre-attempt gate bypassing and out-of-band LLM circumvention (Anthropic [2], Baker et al. [4]).** As shown in our ρ_{bypass} sensitivity sweep, if learners bypass Gate 1 ($\mathcal{G}_{\text{pre}}$) and Gate 3 ($\mathcal{G}_{\text{mint}}$) on more than 28% of steps—for instance, by pasting a generic placeholder hypothesis or querying an unguarded commercial LLM on a second device [2]—JIT-SSS loses its advantage over standard guarded hints, and pedagogical credential validity degrades from 89.76% toward 44.39% even though RFC 9901 SD-JWT cryptographic signatures remain 100.00% valid. Preventing out-of-band circumvention requires lockdown browser sandboxing or interactive oral/live-coding Fade-Out Probes ($\mathcal{G}_{\text{mint}}$; 22).

3. **Sycophancy in Socratic dialogue (Cheng et al. [8], Sharma et al. [25]).** Even when an LLM tutor is prompted to ask Socratic questions rather than give direct answers, Sharma et al. [25] and Cheng et al. [8] show that RLHF-tuned models frequently affirm partially incorrect student hypotheses to maintain conversational agreeableness, reinforcing misconceptions unless tutor transitions are anchored to deterministic domain parsers or unit-test outcomes [16, 28].

4. **Heterogeneous effects across novice vs. advanced learners (Contractor and Reyes [10], Liu et al. [20]).** Meta-analyses of AI tutoring in higher education [10, 20] show that forcing an absolute novice ($P(L_0) < 0.15$) to formulate a pre-attempt hypothesis before seeing a worked example can induce unproductive frustration. For low-prior learners, Gate 1 must first display an analogous worked example before requiring an independent diagnostic attempt on the target problem.

7 Limitations

- **Modeled synergy of $\mathcal{G}_{\text{pre}} + \mathcal{H}_1..\mathcal{H}_3 + \mathcal{G}_{\text{mint}}$.** While Zhi et al. [29], Kestin et al. [15], Bastani et al. [5], and Oreopoulos et al. [22] independently validate delayed access timing, Socratic scaffolding, and mastery-based practice in separate trials, the simulated 80.00% $[76.61, 83.01]$ delayed mastery and 89.76% $[86.25, 92.45]$ pedagogical credential validity of Full JIT-SSS in Table 4 are computational BKT projections under a unified behavioral response model that require prospective multi-arm RCT validation (Section 8).
- **Emergency production bypasses.** During urgent workplace outages ($\tau \to 1.0$), delaying execution by 10–15 minutes for Socratic scaffolding and a Fade-Out Probe is impractical. JIT-SSS must allow an explicit "Emergency Bypass" that unlocks direct AI execution to resolve the incident while withholding SD-JWT credential minting ($\mathsf{MintSDJWT} = 0$).
- **Isomorphic probe synthesis for qualitative skills.** Automatically generating parameter-randomized Fade-Out Probes $q_{\text{iso}}^{\text{fade}}$ (4) is straightforward in mathematics [22], physics [15], and coding [26], but requires human peer review for open-ended design or civic mediation skills.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical tests designed to falsify the core claims of Just-in-Time Socratic Skill Scaffolding.

Table 5: Six pre-registerable falsification protocols for Just-in-Time Socratic Skill Scaffolding (JIT-SSS).

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Additive lift over Standard Guarded Hints (Eq. 1)	Combining Pre-Attempt Gate $\mathcal{G}_{\text{pre}}$ + Fade-Out Probe $\mathcal{G}_{\text{mint}}$ improves 14-day unassisted mastery by ≥ 6 points over Guarded Hints Only [5]	5-arm RCT ($N = 1{,}000$ learners): Static Courseware, Unguarded AI, Guarded Hints Only, Pre-Attempt + Hints, and Full JIT-SSS at $\tau = 0.60$	14-day AI-free post-test score in Full JIT-SSS exceeds Guarded Hints Only by < 5.0 percentage points ($p > 0.05$)
P2. Time-pressure access-timing invariance (Eq. 2)	Enforcing Pre-Attempt Gate $\mathcal{G}_{\text{pre}}$ [29] bounds delayed mastery drop under tight deadlines ($\tau = 0.85$) within 8 points	2×2 factorial trial crossing Access Timing (Immediate vs. Pre-Attempt Gated) $\times$ Time Pressure ($\tau = 0.10$ vs. $\tau = 0.85$)	Delayed mastery in the Pre-Attempt Gated arm drops by > 10.0 percentage points between $\tau = 0.10$ and $\tau = 0.85$
P3. Gate-bypass break-even threshold (ρ_{bypass}-sweep)	Semantic validation of pre-attempts a_0 keeps superficial gate bypassing below $\rho_{\text{bypass}} = 0.25$ in live coursework	Audit 2,000 live Gate 1 submissions under deadline stress ($\tau = 0.60$) using blind human domain raters	$> 28.0\%$ of accepted Gate 1 submissions are rated as low-effort dummy strings ($\rho_{\text{bypass}} > 0.28$)
P4. Metacognitive calibration accuracy	Prediction micro-prompts + Fade-Out Probes reduce confidence overestimation (attribution bias; 23) below 8%	Measure Expected Calibration Error (ECE) between pre-test self-predicted mastery and actual unassisted post-test score	Mean overconfidence gap ($\mathbb{E}[\text{Confidence} - \text{Accuracy}]$) under JIT-SSS exceeds +8.0 percentage points
P5. K–12 anti-sycophancy & privacy audit	K–12 Sanctuary profile eliminates sycophantic validation of false hypotheses [8] and leaks zero struggle claims in SD-JWTs [12]	Red-team 500 erroneous K–12 math/coding hypotheses and inspect emitted RFC 9901 SD-JWT verifier presentations	Tutor affirms an incorrect student hypothesis in $> 2.0\%$ of turns or leaks any raw struggle claim without explicit salt disclosure
P6. Out-of-band copy-paste interception ($\mathcal{G}_{\text{mint}}$)	Time-boxed parameter-randomized Fade-Out Probes ($\mathcal{G}_{\text{mint}}$) sustain pedagogical credential validity $\geq 85\%$	Red-team trial where 150 participants use an external unguarded frontier model during Gate 1–2 and then attempt the 4-minute Gate 3 probe	Pedagogical credential validity ($\Pr[\text{14-day mastery} \mid \text{minted}]$) falls below 80.0%

9 Conclusion

Integrating artificial intelligence into on-task education and workplace skill credentialing requires confronting the empirical reality established across 2025–2026 RCTs: unguarded direct-answer AI accelerates immediate task completion while inducing a 14%–22% relative drop in unassisted post-test mastery below traditional control instruction [5, 24, 26] and inflating metacognitive overconfidence [23]. By formulating skill acquisition as a coupled Bayesian Knowledge Tracing and Cognitive Debt dynamical system and evaluating a 5-arm incremental gate ablation under a unified learner response model, we show that interposing an Access-Timing Pre-Attempt Gate [29], a Socratic Hint Ladder [15, 16], a K–12 Developmental Sanctuary profile [8], and an AI-Free Fade-Out Probe [22] prior to RFC 9901 Selective-Disclosure JWT (`SD-JWT`, 12) issuance lifts simulated 14-day unassisted mastery at $\tau = 0.60$ from 53.00% [49.00, 56.96] to 80.00% [76.61, 83.01] while raising pedagogical credential validity from 54.39% [50.28, 58.43] to 89.76% [86.25, 92.45] alongside 100.00% [98.98, 100.00] cryptographic SD-JWT verification.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Daron Acemoglu and Jörn-Steffen Pischke. Why do firms train? theory and evidence. The Quarterly Journal of Economics, 113(1):79–119, 1998. URL `https://doi.org/10.1162/003355398555531`.

[2] Anthropic. Anthropic education report: How university students use Claude, 2025. URL `https://www.anthropic.com/research/anthropic-education-report-how-university-students-use-claude`. 8 April 2025.

[3] Lisanne Bainbridge. Ironies of automation. Automatica, 19(6):775–779, 1983. URL `https://doi.org/10.1016/0005-1098(83)90046-8`.

[4] Ryan S. J. d. Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". ACM CHI Conference on Human Factors in Computing Systems, pages 383–390, 2004. URL `https://doi.org/10.1145/985692.985741`.

[5] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakçı, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025.

[6] Gary S. Becker. Human capital: A theoretical and empirical analysis, with special reference to education. Columbia University Press / NBER, 1964. URL `https://www.nber.org/boo`

ks-and-chapters/human-capital-theoretical-and-empirical-analysis-special-reference-education-first-edition.

[7] Jan Camenisch and Anna Lysyanskaya. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. Advances in Cryptology — EUROCRYPT, pages 93–118, 2001. URL https://doi.org/10.1007/3-540-44987-6_7.

[8] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), 2026. doi: 10.1126/science.aec8352.

[9] Benjamin Clément, Didier Roy, Pierre-Yves Oudeyer, and Manuel Lopes. Multi-armed bandits for intelligent tutoring systems (ZPDES). Journal of Educational Data Mining, 7(2):20–48, 2015. URL https://doi.org/10.5281/zenodo.3554667.

[10] Zara Contractor and Germán Reyes. Experimental evidence on the learning impact of generative AI. Discussion Paper 18792, IZA Institute of Labor Economics, 2026. URL https://docs.iza.org/dp18792.pdf.

[11] Martin Elias De Simone, Federico Hernan Tiberti, Maria Rebeca Barron Rodriguez, Federico Alfredo Manolio, Wuraola Mosuro, and Eliot Jolomi Dikoru. From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria. Policy Research Working Paper 11125, World Bank, 2025.

[12] Daniel Fett, Kristina Yasuda, and Brian Campbell. Selective disclosure for JWTs (SD-JWT). IETF RFC 9901, 2025. URL https://datatracker.ietf.org/doc/rfc9901/.

[13] Matteo Frigo and abhi shelat. Anonymous credentials from ECDSA (Longfellow-ZK). IACR Communications in Cryptology, 1(2), 2024. URL https://eprint.iacr.org/2024/2010.

[14] Mohamad Y. Jaber and Maurice Bonney. Production breaks and the learning curve: The forgetting phenomenon. Applied Mathematical Modelling, 20(2):162–169, 1996. URL https://doi.org/10.1016/0307-904X(95)00157-F.

[15] Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15:17458, 2025. doi: 10.1038/s41598-025-97652-6.

[16] Kenneth R. Koedinger and Vincent Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review, 19(3):239–264, 2007. URL https://doi.org/10.1007/s10648-007-9049-0.

[17] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task, 2025. URL https://arxiv.org/abs/2506.08872.

[18] LearnLM Team, Google and Fab AI. Teaching with Gemini: Measuring the impact of guided learning on student mathematics progress in Sierra Leone. Technical report, Google, 2026. URL https://storage.googleapis.com/deepmind-media/LearnLM/learnLM_sierraleone_may26.pdf. Technical report, 15 May 2026; not peer reviewed.

[19] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713778.

[20] Chunai Liu, Lidong Xie, and Gang Xu. Generative AI in higher education: A meta-analysis of intellectual and social–emotional outcomes. *Frontiers in Psychology*, 17:1848745, 2026. doi: 10.3389/fpsyg.2026.1848745.

[21] Jing Liu, Tracy Sweet, Marty Haoyuan Chen, Jonathan Engelberg, Megan C. Masters, Martyn Clark, et al. The effects of course-integrated AI tutoring on student performance and engagement: A randomized university trial. Discussion Paper 18989, IZA Institute of Labor Economics, 2026. URL `https://docs.iza.org/dp18989.pdf`.

[22] Philip Oreopoulos, Michael Liut, Alp Sungu, and Nina Low. Making AI tutoring productive: Evidence from a mastery-based math practice experiment. EdWorkingPaper 26-1552, Annenberg Institute, Brown University, 2026. URL `https://edworkingpapers.com/sites/default/files/ai26-1552.pdf`.

[23] Shaowei Ren. College students' metacognitive awareness of generative-AI reliance: An experimental study of decision confidence and attribution bias. *Frontiers in Psychology*, 17:1926110, 2026. doi: 10.3389/fpsyg.2026.1926110.

[24] Sina Rismanchian, Hasan Uzun, Jeffrey Matayoshi, Eric Cosyn, and Eyad Kurd-Misto. Faster completion, less learning: Generative AI reduced study time on math problems and the knowledge they build, 2026. URL `https://arxiv.org/abs/2605.21629`.

[25] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR), 2024. URL `https://arxiv.org/abs/2310.13548`.

[26] Judy Hanwen Shen and Alex Tamkin. How AI impacts skill formation, 2026. URL `https://arxiv.org/abs/2601.20245`.

[27] Miloš Stanković, Ella Hirche, Sarah Kollatzsch, and Julia Nadine Doetsch. Commentary on Kosmyna et al., "your brain on ChatGPT", 2026. URL `https://arxiv.org/abs/2601.00856`.

[28] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4):197–221, 2011. URL `https://doi.org/10.1080/00461520.2011.611369`.

[29] Jiayin Zhi, Harsh Kumar, and Mina Lee. Investigating the effects of LLM use on critical thinking under time constraints: Access timing and time availability. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2026. doi: 10.1145/3772318.3791796.