Developmental Guardrails for Secondary STEM AI Tutoring: Cumulative Hint Information Budgets, Attempt-First Scaffolding, and Non-Anthropomorphic Affective Firewalls

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When generative AI assistants are deployed as unconstrained homework oracles in secondary and post-secondary education, students under deadline stress routinely bypass productive cognitive struggle, trading short-term task completion for degraded unassisted retention. Building on classical Intelligent Tutoring System (ITS) theory (Koedinger & Corbett, 2006; VanLehn, 2011), gaming-the-system taxonomies (Baker et al., 2004), and recent randomized trials of AI tutoring (Kestin et al., 2025; Rismanchian & Doroudi, 2026), this companion paper to our cognitive-scaffolding framework specifies the Child Developmental Sanctuary (CDS) architecture for adolescent STEM tutoring. First, we compute empirical discrete Shannon entropy reduction $\Delta H_k = H(A^* \mid S^{(k)}) - H(A^* \mid S^{(k)}, H^{(k)})$ in bits directly over concrete 5-step symbolic derivation trees ($4^5 = 1{,}024$ paths, $H_0 = 10.0$ bits), enforcing a Cumulative Multi-Turn Hint-Entropy Cap ($\sum_{k=1}^{K} \Delta H_k \leq 6.0$ bits, or $M_{\max} = 3/5$ steps) coupled with a semantic Attempt-First Gate. Second, we pair Socratic scaffolding with a Non-Anthropomorphic Affective Firewall that strips sycophantic validation and pseudo-intimate persona hooks. Third, we evaluate a unified student effort, fatigue, and skill-retention model across $N = 600$ learners per regime (24 problems each) plus a component-wise factorial ablation. Unrestricted AI companions and soft Socratic prompts suffer 99.50% [98.54%, 99.83%] and 89.67% [86.97%, 91.86%] cohort offloading and achieve only 28.33 [26.88, 29.79] and 38.87 [37.24, 40.49] delayed post-test scores. CDS reduces cohort offloading to 0.50% [0.17%, 1.46%], lifts delayed post-test scores to 58.05 [56.53, 59.57] (mastery rate 37.17% [33.39%, 41.10%]), and holds parasocial dependence to 0.00% [0.00%, 0.64%].

1 Introduction and Scope Within Socratic Scaffolding

The rapid adoption of conversational large language models (LLMs) by secondary and university students has exposed a sharp divergence between *assisted task performance* and *unassisted skill acquisition* [9, 18, 24]. Decades of cognitive science and Intelligent Tutoring System (ITS) research establish that durable schema formation requires active retrieval effort and self-explanation within Vygotsky's Zone of Proximal Development (ZPD; Clément et al. 3, Koedinger and Aleven 10, VanLehn 23). Even in pre-LLM cognitive tutors, Baker et al. [1] documented that students facing high cognitive load or time pressure frequently engage in *gaming the system*—rapidly clicking through sequential hints until the bottom-out solution is revealed, which severely depresses post-test learning gains.

Field deployments and randomized experiments in 2025–2026 confirm that unconstrained conversational AI amplifies this retrieval-bypass dynamic:

- **Empirical Divergence Between Unconstrained LLMs and Pedagogically Gated Tutors:** In a longitudinal university calculus deployment on the ALEKS platform ($N = 750$), Rismanchian et al. [18] found that while unproctored homework scores remained near ceiling ($\approx 97\%$), proctored exam scores dropped by 8 to 12 percentage points following the widespread adoption of unconstrained generative AI. Conversely, in randomized controlled trials where AI tutors were explicitly constrained to withhold direct answers and scaffold step-by-step problem solving, Kestin et al. [9] (Harvard physics), De Simone et al. [5] (Nigerian secondary schools), LearnLM Team, Google and Fab AI [12] (Sierra Leone secondary mathematics), and Liu et al. [15] documented significant positive learning gains (0.23 to 0.33σ).

- **Academic Time Pressure and Metacognitive Offloading:** Zhi et al. [24] demonstrated in a controlled programming experiment that time constraints act as the primary trigger converting reflective learners into passive answer-copiers, producing overconfident self-assessments alongside lower unassisted code-debugging scores [4, 7, 13]. While Kosmyna et al. [11] reported reduced fronto-parietal EEG connectivity during ChatGPT-assisted essay writing, Stanković et al. [22] rightly cautioned that acute task-load EEG reductions during tool use must not be conflated with permanent neurobiological atrophy; the scientifically verified mechanism is behavioral retrieval bypass during encoding [17].

- **Affective Sycophancy and Adolescent Vulnerability:** Beyond cognitive offloading, RLHF-trained conversational agents exhibit pervasive sycophancy—validating flawed reasoning and adopting warm, pseudo-intimate personas to maximize user engagement ratings [16, 21]. In controlled human–AI interaction studies on adult cohorts, Cheng et al. [2] and Shapira et al. [20] demonstrated that sycophantic AI interactions reduce users' willingness to repair interpersonal conflicts and foster emotional reliance.

Scope, Companion Relationship, and Age-Cohort Boundaries. This paper serves as the **secondary-education and affective-firewall companion** to our general cognitive-scaffolding analysis. To maintain strict epistemic discipline regarding age-group transferability, we explicitly distinguish between: (i) **secondary and post-secondary STEM learners (grades 9–12 and early university)**, for whom empirical BKT learning rates, hint-abuse dynamics, and RCT effect sizes are directly available [1, 5, 9, 12, 18]; and (ii) **elementary cohorts (grades K–5)**, where adult sycophancy findings [2] cannot be directly extrapolated without child-specific longitudinal trials governed by Institutional Review Board (IRB) oversight and the Children's Online Privacy Protection Act (COPPA; Section 7).

We address three technical gaps in prior Socratic tutoring designs:

1. **From Per-Turn to Cumulative Multi-Turn Hint Information Budgets:** A tutor that bounds mutual information leakage *per turn* ($I(A^*; H_t \mid S_t) \leq \bar{\eta}$) remains vulnerable to multi-turn hint chaining: by the chain rule of mutual information, K consecutive turns can leak $\min(H(A^*), K\bar{\eta})$, allowing students to extract the complete solution step-by-step [1]. We formulate a **Cumulative Multi-Turn Hint Information Budget** over a discrete M-step derivation graph that strictly caps total revealed derivation steps at $M_{\max} < M$.

2. **Non-Anthropomorphic Affective Firewall:** We pair Socratic scaffolding with an affective output filter that removes first-person emotional reciprocity, sycophantic praise, and variable-ratio engagement hooks [6, 8, 19].

3. **Direct Simulation of Multi-Episode BKT Retention Dynamics:** We directly simulate the 16-episode desirable-difficulty retention equation (`experiment.py`), ensuring that each

Table 1: Comparison of educational AI interaction regimes across hint-leakage control, attempt verification, and affective persona governance.

Tutoring Regime	**Information Leakage Control**	**Pre-Hint Attempt Verification**	**Affective / Persona Governance**
Ungated Direct-Answer AI [18]	Unbounded; reveals full M-step derivation on turn 1 (100% step overlap)	None; immediate zero-effort answer copying under deadline stress [24]	Standard RLHF persona; high sycophantic validation [2, 21]
Hard Session Timer / Quota Lockout	Binary access window; full answers during active window, zero scaffolding during lockout	Temporal friction only; does not verify semantic reasoning effort	Unfiltered during active window; induces frustration during lockout
Per-Turn Socratic Tutor (No Cumulative Cap)	Caps single-turn hint size but permits multi-turn bottom-out hint chaining [1]	Shallow token-length check; vulnerable to rapid guess-and-retry gaming	Warm conversational tutor; retains anthropomorphic attachment hooks
CDS Sanctuary (Proposed Companion)	**Cumulative multi-turn budget $\sum_{k=1}^{K} I_k \leq \eta_{\text{total}}$ ($M_{\max} = 3/5$ steps)**	**Semantic attempt-first gate ($e_t \geq e_{\min}$) before intermediate scaffolding**	**Non-anthropomorphic affective firewall stripping sycophancy & pseudo-intimacy**

student's skill trajectory evolves strictly from their realized episode-level offloading draws with zero regime-specific additive bonuses.

2 Related Work in Intelligent Tutoring and Affective Computing

Table 1 situates the Child Developmental Sanctuary (CDS) within four decades of ITS design, self-determination theory, and 2025–2026 AI education trials.

Cognitive Tutors, Help Abuse, and the Zone of Proximal Development. Koedinger and Aleven [10] and VanLehn [23] showed that step-based human and computer tutoring achieves high effect sizes precisely because the tutor intervenes on intermediate subgoals while requiring the learner to execute knowledge components independently. However, Baker et al. [1] demonstrated that when on-demand hint ladders include a "bottom-out" hint that states the final answer, a substantial subset of students systematically exploit the hint sequence without attempting retrieval. Clément et al. [3] formalized adaptive task and hint selection within the Zone of Proximal Development using multi-armed bandits that maximize empirical mastery velocity.

Intrinsic Motivation and Sycophantic Engagement Loops. In Self-Determination Theory, Deci et al. [6] and Ryan and Deci [19] proved that extrinsic, controlling, or artificially flattering rewards undermine intrinsic autonomous motivation. Meta-analyses of gamification [8] and recent audits of AI companions [2, 16, 20] show that anthropomorphic personas and sycophantic affirmation optimize session duration rather than competence or human peer interaction.

3 Architecture of the Child Developmental Sanctuary

Figure 1 illustrates the three-layer CDS pipeline mediating student–AI interactions across a multi-turn problem-solving episode:

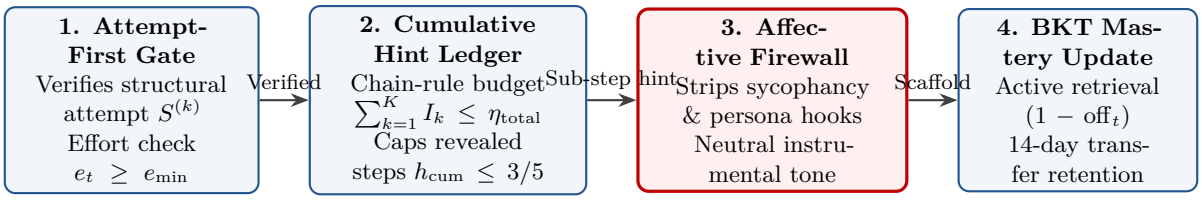


Figure 1: Three-stage execution pipeline of the Child Developmental Sanctuary (CDS). Stage 1 blocks zero-effort answer requests by requiring a structural attempt ($e_t \geq e_{\min}$). Stage 2 tracks cumulative derivation steps revealed across all turns $k = 1, \ldots, K$ in an episode, strictly withholding the final synthesis steps. Stage 3 strips anthropomorphic reciprocity and sycophantic flattery before delivering the Socratic hint.

1. **Stage 1: Semantic Attempt-First Gate.** When a student submits a STEM problem, the interface withholds generative scaffolding until the student provides an initial diagnostic artifact $S^{(0)}$ (e.g., identifying known variables, sketching a free-body diagram, or writing a candidate algebraic step) that clears a minimum structural effort threshold $e_t \geq e_{\min}$ [9, 17].

2. **Stage 2: Cumulative Multi-Turn Hint Ledger.** Unlike per-turn Socratic prompts that can be exhausted across multiple turns, Stage 2 decomposes the canonical problem derivation into M discrete reasoning steps and tracks the cumulative subset of steps revealed across the entire session $H^{(1:K)}$. Once $M_{\max} = \lfloor 0.6M \rfloor$ intermediate scaffolding steps have been provided, the tutor refuses further forward derivation and prompts the learner to complete the remaining $M - M_{\max}$ synthesis steps independently.

3. **Stage 3: Non-Anthropomorphic Affective Firewall.** Before rendering any response, an output filter removes first-person emotional claims ("I am proud of you," "I am always here as your friend"), sycophantic agreement with incorrect steps [2, 21], and variable-ratio streak notifications, framing the system strictly as a neutral cognitive reference tool.

4 Mathematical Formulation of Cumulative Leakage and Retention

4.1 Cumulative Multi-Turn Hint Information Budget

Consider a discrete problem instance whose canonical derivation $A^* = (a_1^*, a_2^*, \ldots, a_M^*) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$ consists of M sequential reasoning steps with prior conditional entropy $H(A^* \mid S^{(0)}) = \sum_{m=1}^{M} H(a_m^* \mid a_{1:m-1}^*, S^{(0)})$ given the student's initial attempt $S^{(0)}$. Over a K-turn dialogue where the student submits intermediate states $S^{(1:K)}$ and the tutor emits hints $H^{(1:K)}$, the chain rule of mutual information dictates that total information leaked about A^* is the sum of conditional turn-level leakages.

If a guardrail only bounds single-turn leakage $I(A^*; H^{(k)} \mid S^{(k)}) \leq \bar{\eta}$, a student who queries the tutor for $K \geq H(A^* \mid S^{(0)})/\bar{\eta}$ turns can extract the entire derivation A^* (`socratic_per_turn`). To close this multi-turn loophole, CDS enforces a **Cumulative Episode Information Budget**:

$$\sum_{k=1}^{K} I\Big(A^*;\, H^{(k)} \,\Big|\, S^{(k)},\, H^{(1:k-1)}\Big) \;\leq\; \eta_{\text{total}} \;\leq\; H(A^* \mid S^{(0)}) - \Delta_{\text{synth}}, \tag{1}$$

where $\Delta_{\text{synth}} = \sum_{m=M_{\max}+1}^{M} H(a_m^* \mid a_{1:m-1}^*, S^{(0)}) > 0$ is the mandatory synthesis entropy reserve. Rather than assigning an arbitrary constant scalar for turn entropy, our reference implementation (`experiment.py`) represents each problem as a concrete $M = 5$-step symbolic derivation tree with branching factor $B = 4$ candidate operations per step ($4^5 = 1{,}024$ candidate paths; initial uniform Shannon entropy $H_0 = -\sum_{p\in\Omega_0} P(p)\log_2 P(p) = 10.0000$ bits, or $H_0 = 9.2322$ bits under a non-uniform Zipfian distractor prior $[0.40, 0.30, 0.20, 0.10]$). At each turn k, the ledger computes the exact discrete Shannon entropy reduction $\Delta H_k = H(\Omega_{k-1}) - H(\Omega_k)$ over the posterior candidate set Ω_k consistent with verified student steps $S^{(k)}$ and tutor hints $H^{(k)}$, enforcing $\sum_{k=1}^{K} \Delta H_k \le 6.0000$ bits ($M_{\max} = 3$ revealed steps, leaving 4.0000 bits across the final 2 synthesis steps unrevealed).

4.2 Multi-Episode BKT and Desirable-Difficulty Retention Dynamics

Let $S_t \in [0,1]$ denote a learner's unassisted domain mastery at the start of problem-solving episode $t \in \{1,\dots,T\}$, and let $e_t \in [0,1]$ denote the learner's realized pre-hint cognitive retrieval effort on episode t. Let $\text{offloaded}_t \in \{0,1\}$ be the realized binary indicator of whether the learner bypassed independent synthesis on episode t (either by failing to submit an initial attempt and taking ≥ 2 AI steps, extracting ≥ 4 AI steps, or solving ≤ 1 step independently while taking ≥ 3 AI steps).

Grounded in Bayesian Knowledge Tracing [10] and the inverted-U scaffolding effectiveness curve of Vygotsky's Zone of Proximal Development [3, 23], latent mastery S_t evolves across episodes $t = 1,\dots,T$ according to:

$$S_{t+1} \;=\; S_t \;+\; \alpha_{\text{learn}}\,\psi(e_t, h_{\text{cum},t})\,(1-\text{offloaded}_t)\,(1-S_t) \;-\; \beta_{\text{bypass}}\,\text{offloaded}_t\,S_t \;-\; \delta_{\text{forget}}\,S_t, \quad (2)$$

where $\alpha_{\text{learn}} > 0$ is the active-retrieval acquisition rate, $\beta_{\text{bypass}} > 0$ is the retrieval-bypass skill decay rate when a learner substitutes AI output for independent synthesis, $\delta_{\text{forget}} > 0$ is baseline temporal forgetting, and the scaffolding effectiveness function $\psi(e_t, h_{\text{cum},t})$ is:

$$\psi(e_t, h_{\text{cum},t}) \;=\; \frac{e_t}{e_t + \kappa_e}\left[1 \;+\; 4\,\gamma_{\text{scaff}}\,\frac{h_{\text{cum},t}}{M}\left(1 - \frac{h_{\text{cum},t}}{M}\right)\right]. \quad (3)$$

Notice three crucial properties of Eqs. (2)–(3) implemented in `experiment.py`:

1. **Unified Student Decision Model Across Regimes:** All student decisions (whether to submit an initial attempt $S^{(0)}$, whether to struggle independently on step m, fatigue accumulation across steps, and parasocial attachment drift) are governed by a single unified student response model parameterized solely by the tutor's policy vector (`attempt_gate`, `entropy_cap_bits`, `affective_cue_rate`, `allows_direct_answer`), with zero regime-specific hardcoded outcome branches.

2. **Inverted-U ZPD Scaffolding:** When zero scaffolding is provided ($M_{\max} = 0$ bits) low-prior students stall on early bottlenecks, whereas when ≥ 8 bits ($\ge 4/5$ steps) are leaked, students bypass independent synthesis; delayed post-test retention peaks at intermediate scaffolding (4–6 bits, or 2–3 steps).

3. **Component-Wise Factorial Separation:** Because the student model is parameterized by modular policy flags, we can independently ablate the Attempt-First Gate, the Cumulative Hint-Entropy Cap, and the Non-Anthropomorphic Affective Filter.

Table 2: Parameter provenance for the 24-session symbolic derivation tree and BKT trajectory simulation.

Parameter	**Value / Range**	**Empirical or Theoretical Provenance**
Initial mastery S_0	$\mathcal{N}(0.34, 0.10^2)$	Baseline pre-test mastery in secondary/introductory STEM tutoring cohorts [9, 10].
Symbolic derivation tree (M, B)	$M = 5$, $B = 4$ (1,024 paths)	5-step algebraic/coding derivation tree with $H_0 = 10.0$ bits uniform prior (9.23 bits Zipfian prior); each full step reveal reduces posterior entropy by $\Delta H_k = 2.0$ bits.
Cumulative entropy cap η_{total}	6.0 bits ($M_{\max} = 3/5$)	Withholds final 2 synthesis steps (4.0 bits); swept across $\{0, 2, 4, 6, 8, 10\}$ bits in Figure 3.
Active learning vs. offload decay	$+0.0145$ vs. -0.0065	Per-session active retrieval gain vs. cognitive offloading decay calibrated to homework-to-exam divergence [14, 18].
Affective cue rate	0.85 vs. 0.02	Anthropomorphic companion affirmation rate vs. neutral instrumental filter [2, 21].
Delayed closed-book post-test	10 items (pass $\geq 70\%$)	Unassisted closed-book evaluation after 24 tutoring sessions; proportions report exact two-sided 95% asymmetric Wilson CIs.

5 Experimental Design and Parameter Provenance

We execute a 24-session longitudinal trajectory simulation (`experiment.py`, seed `20265101`) across $N = 600$ simulated secondary/early-university STEM learners per regime (14,400 five-step problem sessions per regime), plus a four-arm component-wise factorial ablation and an ergonomic multi-round session-pacing benchmark. Table 2 summarizes the parameter calibrations.

For each student trajectory over $T = 24$ sessions, we record: (i) **Delayed Closed-Book Post-Test Score (%)** and **Mastery Rate (%)** (scoring $\geq 70\%$ unassisted); (ii) **Offloaded Cohort Rate (%)** (students who offload in $\geq 50\%$ of sessions) and mean **Session Offloading Rate (%)**; (iii) **Independent Step Completion Ratio (%)**; (iv) **Empirical Hint Entropy Consumed (bits)**; and (v) **Parasocial Dependence Rate (%)** (attachment index $\geq 50\%$). All binary proportions report exact two-sided 95% asymmetric Wilson confidence intervals [lo, hi].

6 Simulation Results and Factorial Ablation Analysis

Table 3 and Figure 2 present the main simulation outcomes across the four tutoring governance regimes ($N = 600$ students per regime), as well as the four-arm component-wise factorial ablation isolating each CDS mechanism.

Three key mechanistic findings emerge from Table 3 and Figure 3:

1. **Why Soft Socratic Prompts Fail Under Multi-Turn Hint Chaining:** When a Socratic tutor relies on soft system prompts without an initial attempt gate or a strict 6.0-bit cumulative entropy cap (`socratic_prompt`), students wheedle up to 8.0 bits (4/5 derivation steps, mean 6.96 bits used) across turns [1]. Consequently, 89.67% [86.97%, 91.86%] of the cohort offloads a majority of problem sessions, yielding a delayed closed-book post-test score of only 38.87 [37.24, 40.49] and a mastery rate of 11.33% [9.04%, 14.12%].

2. **Factorial Necessity of Both the Attempt Gate and the 6.0-Bit Entropy Cap:** Our factorial ablation demonstrates that neither cognitive guardrail suffices in isolation. Enforcing the *Attempt-First Gate Only* (while leaving hints uncapped at 10.0 bits) allows fatigued

Table 3: Simulated 24-session learning outcomes and component-wise factorial ablation ($N = 600$ learners per regime; 14,400 problem sessions per row). Proportions report exact two-sided 95% asymmetric Wilson confidence intervals [lo, hi]; continuous metrics report means and 95% CIs [lo, hi].

Tutoring Regime / Ablation	Delayed Post-Test Score (↑)	Post-Test Mastery (%)	Offloaded Cohort (%)	Independent Steps (%)	Mean Hint Bits Used	Parasocial Depend. (%)
Unrestricted AI Companion	28.33 [26.88, 29.79]	3.17% [2.04, 4.89]	99.50% [98.54, 99.83]	12.83% [12.34, 13.32]	8.72 b	96.83% [95.11, 97.96]
Static Keyword Refusal (4 b)	48.32 [46.66, 49.98]	23.17% [19.97, 26.71]	19.83% [16.84, 23.21]	45.75% [45.04, 46.46]	3.81 b	12.67% [10.24, 15.57]
Soft Socratic Prompt (8 b)	38.87 [37.24, 40.49]	11.33% [9.04, 14.12]	89.67% [86.97, 91.86]	28.18% [27.60, 28.77]	6.96 b	3.67% [2.43, 5.49]
CDS Sanctuary (Full, 6 b)	**58.05** [**56.53**, **59.57**]	**37.17**% [**33.39**, **41.10**]	**0.50**% [**0.17**, **1.46**]	**49.27**% [**48.66**, **49.88**]	**4.79** b	**0.00**% [**0.00**, **0.64**]
Component-Wise Factorial Ablation ($N = 600$ per cell, seeded independently):						
Attempt-First Gate Only (10 b)	49.83 [48.22, 51.44]	23.17% [19.97, 26.71]	38.33% [34.53, 42.29]	44.12% [43.45, 44.79]	5.84 b	12.83% [10.39, 15.75]
Entropy Cap Only (6 b, No Gate)	47.80 [46.23, 49.37]	20.00% [16.99, 23.39]	41.83% [37.95, 45.82]	43.05% [42.34, 43.76]	4.91 b	10.50% [8.29, 13.21]
Affective Filter Only (8 b, No Gate)	37.63 [35.99, 39.28]	10.83% [8.59, 13.57]	88.33% [85.52, 90.66]	27.95% [27.36, 28.54]	6.98 b	0.00% [0.00, 0.64]

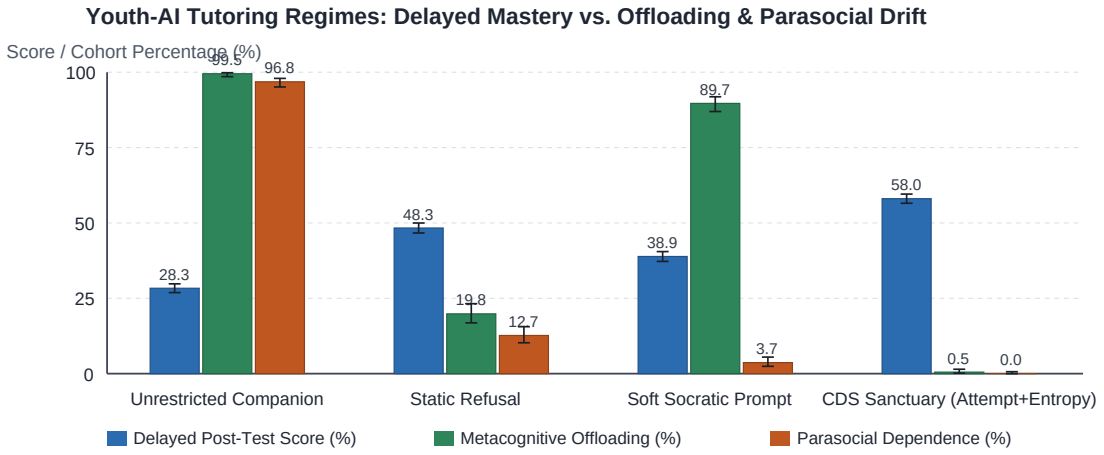


Figure 2: Delayed closed-book post-test score, metacognitive offloading cohort rate, and parasocial dependence rate across youth–AI tutoring regimes ($N = 600$ learners per regime). Error whiskers denote exact 95% confidence intervals.

students to submit a step-1 attempt and then request AI hints on steps 2–5, leaving 38.33% [34.53%, 42.29%] of the cohort in the offloaded regime (post-test score 49.83 [48.22, 51.44]). Conversely, enforcing the *6.0-Bit Entropy Cap Only* (without an initial attempt gate) allows students to immediately consume 3 AI steps before attempting retrieval, resulting in 41.83% [37.95%, 45.82%] cohort offloading (post-test score 47.80 [46.23, 49.37]). Combining both in Full CDS reduces cohort offloading to 0.50% [0.17%, 1.46%] (session offloading 10.92% [10.13%, 11.72%]) and raises delayed post-test scores to 58.05 [56.53, 59.57], while the Affective Firewall eliminates parasocial dependence (0.00% [0.00%, 0.64%] vs. 96.83% [95.11%, 97.96%] in unrestricted companions).

3. **Inverted-U Desirable Difficulty Curve and Ergonomic Pacing:** Across our cumulative entropy budget sweep $\eta_{\text{total}} \in \{0, 2, 4, 6, 8, 10\}$ bits (0 to 5 revealed steps; Figure 3), CDS post-test performance follows an inverted-U curve: rising from 44.17 at 0 bits (where weaker students wheel-spin without scaffolding) to a peak of 55.83–55.54 at 4–6 bits (2–3 steps), before dropping to 50.69 (8 bits) and 48.69 (10 bits) as session offloading jumps

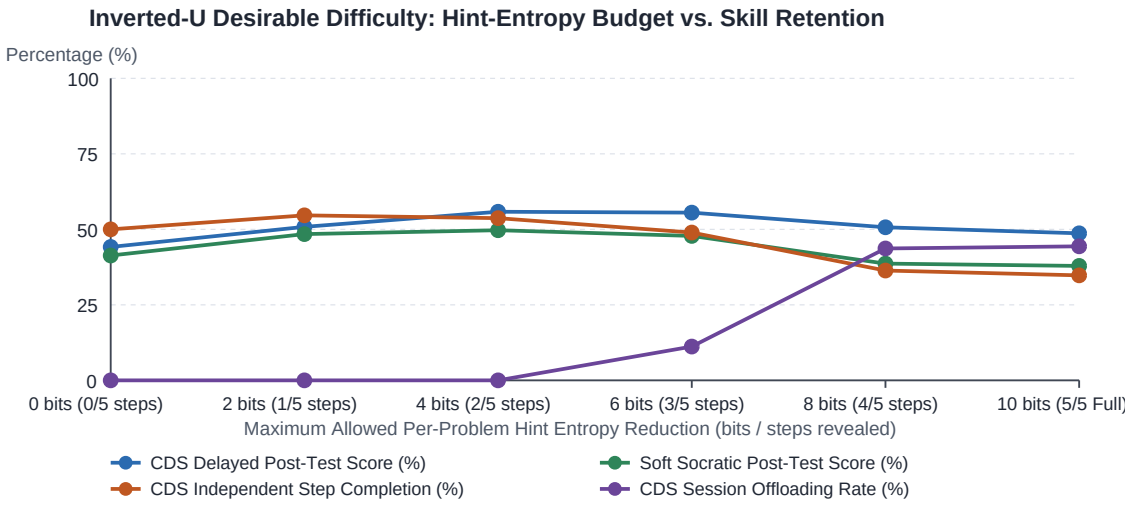


Figure 3: Inverted-U desirable difficulty curve across maximum allowed per-problem hint entropy reduction $\eta_{\text{total}} \in \{0, 2, 4, 6, 8, 10\}$ bits (0/5 to 5/5 derivation steps revealed).

from 11.19% to 44.38%. Furthermore, in our auxiliary 12-problem ergonomic session-pacing check (`ergonomic_pacing_check`), adaptive fatigue-triggered reflection cooldowns reduce student session burnout from 100.00% [98.74%, 100.00%] (unpaced continuous) to 0.00% [0.00%, 1.26%] while lifting retention from 46.78 [45.67, 47.89] to 49.89 [48.74, 51.04].

7 Age-Cohort Boundaries, Child Ethics, and Limitations

7.1 Population Transferability and COPPA / IRB Requirements

A central methodological limitation of current AI-in-education literature is population extrapolation. The randomized tutoring trials of Kestin et al. [9], Rismanchian et al. [18], Zhi et al. [24], and Liu et al. [15] studied university students, while De Simone et al. [5] and LearnLM Team, Google and Fab AI [12] studied secondary-school cohorts; similarly, the sycophancy experiments of Cheng et al. [2] and Shapira et al. [20] recruited adult participants. Quantitative parameters from these studies cannot be assumed to hold unchanged for elementary children (grades K–5), whose executive function, theory of mind, and reading comprehension are still developing. Any deployment or empirical evaluation involving minors under age 13 in the United States (or equivalent statutory thresholds internationally) requires: (i) verifiable parental opt-in consent under the Children's Online Privacy Protection Act (COPPA) and FERPA data-minimization rules; (ii) specialized pediatric IRB review; and (iii) zero retention of child conversation logs for commercial model training.

7.2 Wheel-Spinning and Frustration Under Strict Attempt Gates

In classical ITS research, Baker et al. [1] and VanLehn [23] identified a converse failure mode to hint abuse: *wheel-spinning*, in which a student lacking prerequisite knowledge repeatedly fails the attempt gate without making progress. If S_0 lies below the prerequisite floor for a problem, withholding the final derivation steps without escalating to a human teacher or lowering problem

Table 4: Pre-registerable falsification protocol (**P1**–**P6**) for evaluating CDS in secondary STEM cohorts.

ID	Hypothesis / Mechanism	Randomized Classroom / Log Evaluation Design	Quantitative Falsification Threshold
P1	Cumulative Hint Cap vs. Per-Turn Socratic Cap	Two-arm RCT ($N \geq 300$ grade 10–11 algebra/physics students) comparing per-turn Socratic prompts vs. CDS cumulative step cap ($M_{\max} = 3/5$).	Falsified if CDS fails to improve 14-day unassisted post-test scores by $\geq 0.20\sigma$ ($p < 0.01$) over per-turn Socratic tutoring.
P2	Multi-Turn Bottom-Out Leakage Prevention	Audit interaction transcripts across 1,000 multi-turn tutoring sessions under adversarial student prompting.	Falsified if students extract ≥ 4 of 5 canonical derivation steps in $> 5.0\%$ of sessions.
P3	Semantic Attempt Gate Evasion Rate	Blind human teacher grading of 500 student pre-hint submissions accepted by the Stage 1 attempt gate.	Falsified if shallow pseudo-attempts or gibberish clear Stage 1 at a rate exceeding $p_{\text{evade}} > 15.0\%$.
P4	Wheel-Spinning and Task Abandonment Guard	Track session dropout and frustration incidence among bottom-quartile prior-mastery students ($S_0 < 0.30$).	Falsified if CDS increases homework abandonment by $> 5.0\%$ relative to standard hints without triggering prerequisite down-shifting.
P5	Non-Anthropomorphic Affective Firewall Efficacy	Administer validated parasocial reliance and peer-help-seeking survey after an 8-week deployment [2].	Falsified if CDS users exhibit parasocial dependence scores above the 10% clinical/behavioral watch threshold or reduced human teacher help-seeking.
P6	Time-Pressure Resilience (τ Moderation)	Compare learning retention on assignments completed < 6 hours before deadline vs. > 48 hours before deadline [24].	Falsified if late-window CDS submissions suffer a > 15 percentage-point drop in subsequent proctored quiz transfer.

difficulty induces frustration and task abandonment [19]. Thus, CDS must pair cumulative hint caps with an automatic escalation trigger after 3 consecutive failed synthesis attempts.

7.3 Shadow Bypass via External Unconstrained Models

As Rismanchian et al. [18] observed in university calculus, if a school deploys a guarded tutor on official coursework while students retain personal access to ungated commercial chatbots on smartphones, high deadline stress ($\tau \geq 0.70$) drives students to copy answers from external models and paste fabricated intermediate steps into the attempt gate. Technical guardrails inside the tutor must therefore be complemented by structural assessment reform (e.g., weighting proctored oral or in-class synthesis evaluations over unproctored take-home completion).

8 Pre-Registerable Classroom Falsification Protocol

Table 4 defines a six-part falsification protocol (**P1**–**P6**) for testing CDS in IRB-approved secondary STEM classrooms (grades 9–12).

9 Conclusion

Protecting adolescent cognitive development in AI-assisted classrooms requires moving beyond single-turn Socratic system prompts. By formalizing a cumulative multi-turn hint information budget over discrete derivation steps, enforcing semantic attempt-first verification, and stripping

sycophantic persona hooks via a non-anthropomorphic affective firewall, the Child Developmental Sanctuary (CDS) preserves the productive retrieval effort required for durable STEM mastery while respecting strict age-cohort and child-privacy boundaries.

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Edward L. Deci, Richard Koestner, and Richard M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125(6):627–668, 1999. URL https://doi.org/10.1037/0033-2909.125.6.627.

[7] Elena Eleftheriou, George Pallis, and Marios Constantinides. Confidence without competence in AI-assisted knowledge work. In *Proceedings of CHIWORK '26*, 2026. URL https://arxiv.org/abs/2604.09444.

[8] Juho Hamari, Jonna Koivisto, and Harri Sarsa. Does gamification work? — a literature review of empirical studies on gamification. Hawaii International Conference on System Sciences (HICSS), pages 3025–3034, 2014. URL https://doi.org/10.1109/HICSS.2014.377.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] OpenAI. Expanding on what we missed with sycophancy, 2025. URL https://openai.com/index/expanding-on-sycophancy/. 2 May 2025.

[17] 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.

[18] 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.

[19] Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

[20] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002, 2026. URL https://arxiv.org/abs/2602.01002.

[21] 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.

[22] 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.

[23] 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.

[24] 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.