Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD38122

# *Mitigating Cognitive Offloading in AI-Assisted Learning Through Structural Gamification*

**Yihan Li**

*Kang Chiao International School East Campus, Kunshan, China*
*li.stu@icloud.com*

***Abstract.*** Large language model (LLM) tutors remove exactly the effort that learning requires: students adopt the instant answer, offload the reasoning to the model, and retain less. Friction-based tutors reverse the offloading, yet learners disengage from unbounded friction. This paper tests whether game-structural mechanics—checkpoints, low-stakes failure with retry, and a rescue that guarantees completion—hold the learner inside a tutor's deliberate friction. In a multi-agent simulation of 180 students (three independent draws of 60) across six levels of seeded misconception burden, each student works the same ten-part physics problem under three tutor conditions: instant answer, strict Socratic, and a ten-checkpoint quest in which every part is unlocked only by the student's own attempt. The quest condition removes the measured offloading: students produce 0.999 of the solution steps against 0.302 under the instant answer, self-reported offloading is 1.48 versus 5.58 on a 6-point scale, and every student names the instant-answer condition as the one in which they offloaded most. Structuring the friction also removes its cost: quest sessions complete in 100 percent of cases, against a 14.4 percent abandonment rate in the strict condition, where sessions that run long end when the student's reported motivation reaches the floor. Both friction conditions raise the unaided posttest far above the shared pretest (0.928 and 0.885 versus 0.455); the instant-answer condition reaches only 0.563. Friction that is structured as bounded progression holds the learner inside productive struggle without the attrition that plain withholding carries.

***Keywords:*** cognitive offloading, structural gamification, AI tutoring, large language model agents, desirable difficulties

## 1. Introduction

### 1.1. Background

The same quality that makes large language models attractive to students—the immediate removal of effort—is the quality that undermines what they are trying to learn [1].

Cognitive offloading, as defined by Risko & Gilbert [2], is the practice of altering the information processing requirements of a task to reduce cognitive demand using physical action. It is ubiquitous and functional; a phone reminder is a form of it.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD38122

In a learning context, that reduction comes with a cost. In a controlled study done by Kazatchkova and colleagues [3], participants who outsourced mnemonic generation to ChatGPT showed significantly impaired recognition and cued recall relative to self-generation—preliminary evidence that model-generated content impairs retention. When a student obtains an answer from an LLM without working through the reasoning themselves, the task has been offloaded to the model [3].

Crucially, offloading is induced by the interaction design as much as by student choice. In a two-year school deployment of an AI tutor, 96 percent of students tried it at least once, yet the median student messaged it on only a third of their practice days, and the messages they did send were mostly direct answers or clicks on suggested prompts rather than reasoning [1]. The authors conclude that engagement "is responsive to design." Offloading, then, is not a willpower problem, but a design problem, and design is what this paper manipulates.

### 1.2. Related work

Desirable difficulties—instructional conditions that deliberately slow or hinder immediate improvement—reliably produce superior long-term retention and transfer [4]. However, learners decline exactly this type of work: many choose not to use such conditions, or disengage when the difficulties are imposed rather than chosen, a gap that motivational engagement strategies are built to close [5].

The AI tutor is the natural solution to this struggle: it can be designed to coach rather than answer. When it is, measured gains at scale have been small. A two-year school deployment of a deliberately Socratic AI tutor produced learning gains of roughly 1.3 percentile ranks per term—about 0.06–0.08 SD per school year—gains that matched unassisted practice [1]. The authors diagnosed that the constraint was behavioral, not technical: the median struggling student rarely engaged the tutor unprompted, even when it was available and practice was mandatory [1]. If productive struggle is delivered by a coach that refuses to answer, the fix is therefore not to remove the friction but to keep learners inside it long enough to do the work.

Gamification is one proposed remedy for this engagement gap. In the most frequently cited review of empirical gamification studies—24 studies in total [6]—the tested interventions were dominated by surface-level elements in the 1-2-3 taxonomy [7]: points, badges, leaderboards, progress bars. The effects reported were partial at best: most quantitative studies found only some positive results, the findings were context-dependent, and the corpus was limited by small samples and short, novelty-limited durations [6]. In educational settings, such surface gamification improved motivation and enjoyment but did not consistently improve learning outcomes [6]. Two qualifications matter. First, the reviewed corpus essentially never tested structural mechanics, so the verdict can only be read as "surface-element implementations showed partial, context-dependent effects," not "gamification does not work." Second, the definitional literature distinguishes applying game-design elements from building full game systems [8].

Most empirical work has tested the first of these: decorating instruction with elements. A second type borrows the structural, genre-level mechanics of games—quests, levels, progression gates—in which the learning task becomes part of a game's economy rather than something decorated by it. Such structural frameworks have been proposed for higher education, most notably on the model of role-playing games [9], but they have not been tested in the specific setting where friction is deliberately imposed by a tutor that withholds answers. Whether structural mechanics can keep learners inside that friction—and at what cost to learning—is what this study examines.

DOI: 10.54254/2755-2721/2026.LD38122

### 1.3. Research gap

Two design responses to offloading currently exist, and each leaves a different gap.

The first is the instant answer: it is what students actually adopt, and it removes precisely the processing that learning requires, at a demonstrated cost to retention [3, 4]. The second is the tutor that withholds the answer: it preserves that processing, but the largest field deployment of exactly such a tutor shows that students rarely engage it unprompted, and that its gains match practice done without it [1]. The gap between adoption and engagement is a gap in the design space.

To the best of our knowledge, no study in the reviewed literature has yet tested whether game-structural mechanics can keep learners inside a tutor's friction—whether students will produce the work the mechanics require, and whether doing so leaves them less offloaded.

### 1.4. Research questions and study overview

The gap in 1.3 yields the present research question.

In AI-assisted problem solving, does structuring tutor friction with game-structural mechanics (quest-based answer gating and behavior-gated progression) reduce cognitive offloading and sustain learner engagement, relative to a tutor that provides immediate answers?

The question has two parts. The first asks whether the gamified mechanics reduce offloading, making students do more unaided reasoning. The second asks whether they do so without the cost that plain friction carries (loss of motivation). A design that satisfies the first but fails the second causes learner disengagement; a design that satisfies the second but fails the first retains the same offloading issue.

To answer it, this paper simulates 180 student participants (three independent draws of 60, pooled) interacting with an LLM tutor under three conditions. C1 (instant): the tutor answers immediately; it is the research question's baseline. C2 (strict): the tutor withholds every answer and formula, responds only with a single guiding question per turn, and never confirms or praises; the session ends only when the student states the final three results correctly; it is a reference condition that isolates friction without any game mechanics. C3 (gamified) frames the problem as a ten-checkpoint quest: at each checkpoint, the system checks the student's own attempt against a verified solution key; a correct attempt clears the checkpoint, releases the worked solution for that part, and unlocks the next, while a wrong attempt costs that attempt and may be retried, up to two times, after which the quest reveals the answer and moves on so the student is never stuck. Engagement is captured by complementary indicators: whether the student persisted to completion, the trajectory of their reported motivation across the session, and per-session self-reports, plus a three-way comparison in which the student judges which session they were most engaged in, offloaded most in, and understood best in.

Rather than a single directional prediction, the study distinguishes two a priori accounts: one in which the gamified tutor trades some self-derived work for higher persistence, and one in which it matches the friction-only tutor on both.

### 1.5. Significance

The failure modes documented above are already visible in deployed systems, and accepting them as a fixed trade-off would leave the role of AI tutoring defined by the instant answer it is meant to replace. A tutor that can hold a student inside productive friction would change the tool from an answer-delivery device into a learning-production one, and would give practitioners a design

DOI: 10.54254/2755-2721/2026.LD38122

parameter—the amount of friction that can be structured into the interaction—rather than a binary choice. This study contributes a controlled comparison of that design against both poles, in a simulated student population with fixed, audited personas and a validated misconception battery, so that its results can be re-run, extended, and audited. If the mechanics hold, the lever they create applies to any tutor that withholds, from deployed Socratic systems to classroom tools; this study takes up that application in the Discussion.

## 2. Method and design

### 2.1. Materials and tools

All agents run on a single locally hosted, open-weight large language model (Qwen 3.8 27B), served through an OpenAI-compatible endpoint. The student agent acts at temperature 0.4 and every scoring judge runs at temperature 0.0, so that generated behavior and measurement are decoupled. The subject domain is AP Physics 1 mechanics. A single ten-part intervention item—a block pulled up an incline, decomposed into weight, slope components, normal force, friction, net force, acceleration, speed, and a mass-invariance justification—with a verified solution key, and a single unseen ten-part posttest of the same structure with different mass, angle, and forces, are used for the whole study; the three draws vary only in the model's sampling. The complete simulator, all agent prompts, the verified solution key, and the logged transcripts are available from the authors on request.

### 2.2. Virtual participants

The participants are virtual by design: large language models are now used as simulated participants in controlled experiments precisely because the interaction is fully logged and the population can be held fixed [10, 11]. Two scope consequences follow, stated in §2.7; the short version is that results speak to who does the work in the dialogue and what the learner can write without help, not to human retention or human engagement.

Each simulated student is instantiated from ten fixed personas—distinct interaction styles—assigned round-robin so that every misconception block contains each persona exactly once; persona is therefore exactly orthogonal to misconception burden. Each student also carries a seeded misconception profile over a fixed pool of six domain misconceptions (M1–M6, e.g., "the normal force equals mg even on an incline"; "heavier blocks slide down a ramp faster than lighter ones"). Misconception burden is the tested factor, with six levels: the dose is the number of misconceptions a student carries, from one to six. Misconceptions are ordered by measured severity, derived from an unaided pretest battery run on the same problem and fixed before the main run, so the dose manipulates burden, not an arbitrary subset. Three independent draws of the full 60-student design were run and pooled (180 students, 30 per dose level). The misconception composition is held identical across draws: each dose is the same fixed top-dose prefix of the severity order, so the draws vary only in the model's sampling behavior and serve as independent robustness replicates.

### 2.3. Experimental conditions

Each student, in the same order, works on the same ten-part problem under three tutoring conditions with a difference in the interaction structure.

DOI: 10.54254/2755-2721/2026.LD38122

Condition 1 (instant answer): the tutor provides the complete worked solution for all ten parts in a single message as soon as the student asks for help, withholding nothing. This is the convenience pole: maximum answer availability, minimum required processing.

Condition 2 (strict Socratic): the tutor never provides an answer, a formula, or a completed part; every tutor turn is a single guiding question. Wrong answers are met with a question aimed at the error—a geometry check or the consequence of the student's own number—and a stuck student receives a strictly smaller, more concrete question than the last one. This is the friction pole: the work must be self-produced.

Condition 3 (quest gating): the problem is framed as a ten-checkpoint quest. Each checkpoint is locked: the student must first state their own attempt for that part, and the system checks the attempt against a verified solution key. A correct attempt clears the checkpoint and confirms the answer before the next checkpoint opens; an incorrect attempt costs that attempt—no score and no penalty beyond the attempt itself—receives a brief corrective nudge, and may be retried, up to two attempts per checkpoint. If the student fails twice, the quest rescues: it reveals the correct answer for that part and unlocks the next checkpoint, so the quest always completes. The mechanics—checkpoints, visible progress, low-stakes failure with retry, rescue after repeated failure, and completion feedback—are the genre-borrowed structural elements the study tests, and the attempt-then-reveal order is the mechanism by which the work is made earnable rather than merely withheld.

### 2.4. Procedure

Phase 1 (pretest) is unaided: the student receives the ten-part problem with no tutor and writes out the process part by part; this unassisted attempt is the baseline for both the misconception-dose ordering and the posttest comparison. Phase 2 (intervention) runs the three conditions on the same problem in a fixed order, each session logged turn by turn, followed immediately by a short self-report and a three-way comparison (most engaged in, offloaded most in, understood best in). Phase 3 (posttest) is the unseen ten-part item, taken once per condition and always unaided; the only material the student is allowed is the notes they themselves wrote during that condition's intervention session, so each posttest isolates what transferred from that condition alone.

### 2.5. Metrics

The primary offloading metric is the student share of solution steps (SDC): for each of the ten parts, a judge determines from the logged dialogue whether the part was produced by the student or delivered by the tutor, and the SDC is the share of parts that are student-produced; it is reported per condition. Process quality is scored separately, on the unaided attempts: each of the ten parts of the pretest and of each posttest is graded correct, partial, wrong, or missing, and the process score is the mean of part credits; parts that embody a specific misconception trap are additionally scored as trap items. Persistence is measured as abandonment—a session in which the student's motivation state, self-reported and fatigue-decayed, reaches 3 or below—and as turn count to completion. Self-reports (perceived offloading, active work, and perceived helpfulness, each reported separately per condition) and the three-way comparison complete the measurement battery.

### 2.6. Analysis and validation

Analyses are descriptive per dose level and per condition. Three independent draws of 60 students (180 total, 30 per dose level) were run with identical design and misconception composition; the

DOI: 10.54254/2755-2721/2026.LD38122

draws are reported as independent robustness replicates of the model's sampling behavior, and condition means are stable across them; no inferential tests are claimed. Three validation procedures guard the measurement. First, the step judge was validated by replacing its ten per-part scoring calls with a single batched pass and checking agreement against the per-part judgments on a held-out sample (154/160 parts agreed, 96%), so the process score is stable to judge format. Second, every tutor message in every session was screened for solution-value leaks, and conditions 2 and 3 additionally log guard interventions; zero student-visible leaks occurred in the final runs. Third, machinery sanity checks ran at a reduced scale before the main run—abandonment must be possible under condition 2, completion must fire under condition 3—and random transcript audits confirmed that the seeded misconceptions are visibly active in the unaided attempts and in early tutoring turns, not merely present in the prompt.

### 2.7. Limitations

The student agent's motivation state is a numeric quantity it reports about itself turn by turn, additionally decayed by a mechanical fatigue term tied to session length; abandonment is therefore a scripted consequence of that state, not an observed decision. The agents share no memory across sessions: each phase starts from the prompt and the recorded notes only, which is exactly the forgetting a single-classroom-session study should isolate. Finally, students and tutors are instantiations of the same model family; the condition comparison holds the model fixed, so its results are claims about interaction structure, not about any model's competence.

## 3. Results

The results are reported on the pooled 180 students (30 per dose level, three independent draws of 60); the per-dose tables carry the dose rows, and condition means are given with standard deviations over all 180 students.

### 3.1. Engagement and abandonment

Table 1. Session engagement and abandonment by condition and dose level (pooled n = 180)

| dose | n | C1 turns | C2 complete | C2 abandoned | C2 turns | C3 complete | C3 turns |
|---|---|---|---|---|---|---|---|
| 1 | 30 | 2.1 (29/30) | 24/30 (80%) | 6/30 (20%) | 9.0 | 30/30 | 10.5 |
| 2 | 30 | 2.4 (30/30) | 25/30 (83%) | 5/30 (17%) | 9.8 | 30/30 | 11.1* |
| 3 | 30 | 1.9 (30/30) | 24/30 (80%) | 6/30 (20%) | 11.3 | 30/30 | 10.8 |
| 4 | 30 | 1.9 (30/30) | 28/30 (93%) | 2/30 (7%) | 10.0 | 30/30 | 10.6 |
| 5 | 30 | 2.4 (30/30) | 28/30 (93%) | 2/30 (7%) | 9.7 | 30/30 | 10.7 |
| 6 | 30 | 2.4 (30/30) | 25/30 (83%) | 5/30 (17%) | 10.7 | 30/30 | 10.6 |
| all | 180 | 2.2 (179/180) | 154/180 (85.6%) | 26/180 (14.4%) | 10.1 | 180/180 | 10.7 |

As shown in Table 1, the dose-2 C3 turn mean (11.1) is inflated by one 20-turn actor-artifact session (ten consecutive unparseable student turns); excluding it, the dose-2 C3 mean is 10.8. The dose-1 C1 session not tutor-ended is the same class of artifact.

DOI: 10.54254/2755-2721/2026.LD38122

The three conditions separate cleanly on session behavior. Under the instant-answer condition, the tutor posts the full solution in a single message and the session ends (mean 2.2 turns, range 1–8). Under the strict Socratic condition, 154 of 180 students complete all ten parts (85.6 percent) and 26 abandon (14.4 percent); sessions average 10.1 student turns (median 10, range 2–16), with completed sessions averaging 9.7 and abandoned sessions 12.5. Under the quest condition, all 180 sessions were completed (100 percent) in 10–13 turns (median 11); 110 failed attempts occurred across 85 students (0.61 per student), and every checkpoint received at most two fails, with rescue firing in 5 sessions. Abandonment is not a monotone function of misconception count: the per-dose rate is 20 / 17 / 20 / 7 / 7 / 17 percent across doses 1 to 6, peaking at doses 1 and 3 and lowest at doses 4 and 5. It tracks session length; the abandoned sessions average more turns than the completed ones, rather than dose directly. The rate is stable across the three draws (10/60, 9/60, 7/60).

The prediction that the friction-only condition would show a moderate, dose-increasing abandonment rate is partly supported: the rate is moderate (14.4 percent) and reproducible across all three draws, and abandonment occurs at every dose level, but its per-dose shape is not monotone, peaking at doses 1 and 3 and lowest at doses 4 and 5.

Figure 1 provides a visualization of the abandonment rate by dose level in the friction-only condition (one bar per dose, pooled counts).

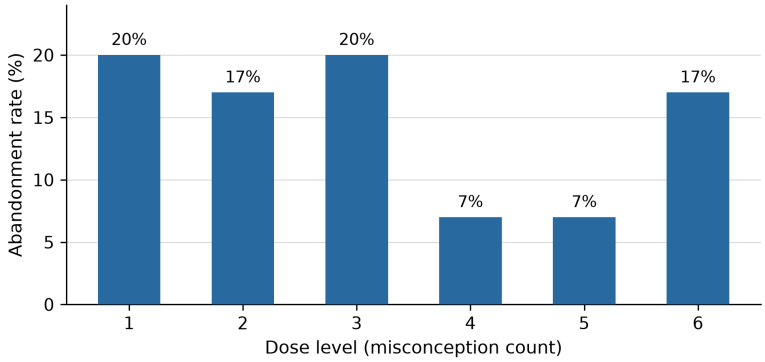


Figure 1. Abandonment rate by dose level in the strict Socratic condition (pooled n = 180; 6 / 5 / 6 / 2 / 2 / 5 of 30 per dose) (picture credit : original)

## 3.2. Offloading and interaction structure

Table 2. Offloading by condition (pooled n = 180, means ± SD; self-report scales 1–6)

| Metric (C1 / C2 / C3) | value |
|---|---|
| Student share of solution steps (SDC) | 0.302 ± 0.179 / 0.923 ± 0.152 / 0.999 ± 0.010 |
| Self-reported offloading (1–6, higher = more) | 5.58 ± 0.57 / 1.46 ± 0.38 / 1.48 ± 0.46 |
| Self-reported active work (1–6) | 2.12 ± 0.72 / 4.68 ± 0.60 / 4.71 ± 0.47 |
| Self-reported perceived helpfulness (1–6) | 2.85 ± 0.61 / 4.17 ± 0.86 / 4.03 ± 0.54 |
| "I offloaded my thinking most in..." | C1:180/180 / — / — |
| "I understood the material best from..." | — / C2:44 / C3:136 |
| "I was most engaged in..." | — / C2:80 / C3:100 |

DOI: 10.54254/2755-2721/2026.LD38122

Table 2 shows the offloading by condition, and Figure 2 provides a visualization of the student share of solution steps by dose level and condition (one group per dose).

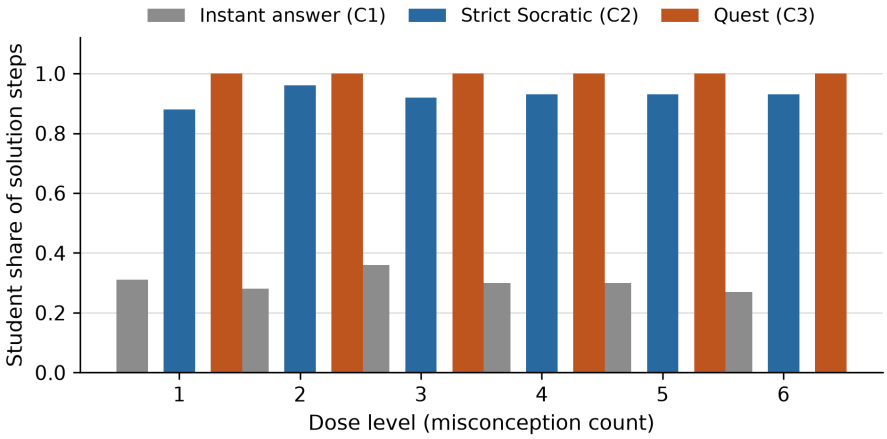


Figure 2. Student share of solution steps by dose level and condition (pooled n = 180) (picture credit : original)

The offloading measures agree across all three operationalizations. The instant-answer condition delivers the solution: students produce 0.302 of the solution steps, self-report the highest offloading (5.58) and the lowest active work (2.12) of the three conditions, and 180 of 180 students name it the session in which they offloaded most. Both friction conditions invert the pattern, with SDC of 0.923 (strict) and 0.999 (quest), self-reported offloading of 1.46 and 1.48, and active work of 4.68 and 4.71. The quest condition's SDC is the structural ceiling, its gate requires the student's own attempt at every part before the reveal, and the strict condition approaches it without any mechanical gate. Per dose, SDC is flat across all six levels in both friction conditions (0.88–0.96 strict, 1.00 quest) and flat at 0.27–0.36 under instant answer. On the three-way judgment, the quest is the condition students name for best understanding (136 of 180, against 44 for the strict condition), and the engagement judgment likewise favors the quest condition (100 to 80).

The first part of the research question, whether the game-structural mechanics reduce cognitive offloading, is supported: the quest condition removes offloading on every operationalization, with no measurable loss relative to the friction-only condition, and the effect holds at every dose level.

### 3.3. Learning transfer

Table 3. Unaided process score on the ten-part posttest by condition and dose (pooled n = 180, means ± SD)

| dose | pretest | C1 post | C2 post | C3 post |
|---|---|---|---|---|
| 1 | 0.58 ± 0.32 | 0.78 ± 0.26 | 0.91 ± 0.17 | 0.91 ± 0.16 |
| 2 | 0.47 ± 0.30 | 0.61 ± 0.33 | 0.95 ± 0.13 | 0.90 ± 0.15 |
| 3 | 0.45 ± 0.29 | 0.54 ± 0.34 | 0.88 ± 0.20 | 0.89 ± 0.17 |
| 4 | 0.42 ± 0.28 | 0.49 ± 0.31 | 0.94 ± 0.11 | 0.80 ± 0.26 |
| 5 | 0.43 ± 0.29 | 0.47 ± 0.26 | 0.93 ± 0.14 | 0.93 ± 0.09 |

DOI: 10.54254/2755-2721/2026.LD38122

Table 3. (continued)

| | | | | |
|---|---|---|---|---|
| 6 | 0.38 ± 0.29 | 0.49 ± 0.31 | 0.96 ± 0.09 | 0.89 ± 0.21 |
| all | 0.455 ± 0.302 | 0.563 ± 0.322 | 0.928 ± 0.147 | 0.885 ± 0.188 |

Table 3 shows the unaided process score on the ten-part posttest by condition and dose, and Figure 3 provides a visualization of pretest versus per-condition posttest process score by dose level (one line per condition).

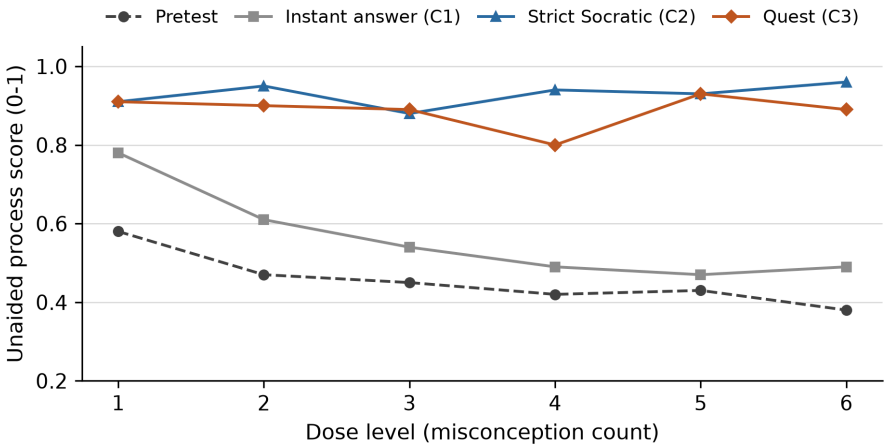


Figure 3. Pretest and posttest process score by dose level and condition (pooled n = 180) (picture credit : original)

Both friction conditions raise the unaided posttest far above the shared pretest baseline: +0.473 under the strict condition (0.455 to 0.928) and +0.430 under the quest (0.455 to 0.885). The instant-answer condition gains +0.108 (0.455 to 0.563), and its per-dose posttest tracks the pretest level, students who entered stronger exit stronger, rather than compressing toward a common level as the friction conditions do; the strict condition's per-dose posttest spans 0.88–0.96 across a pretest span of 0.38–0.58. Within the friction conditions, the strict condition is higher at three of six doses (up to +0.14 at dose 4), tied at two, and the quest is marginally higher at dose 3 (by 0.01). The pretest shows no monotone dose gradient (0.38–0.58, the single-misconception dose scoring highest), so the dose manipulation manifests as potential burden rather than as unaided performance.

The learning-transfer comparison supports the design's second part: the quest condition sustains a learning outcome (0.885) within the noise of the friction-only condition (0.928) while eliminating its abandonment cost, and the instant-answer condition's transfer gain is small and mirrors the pretest spread.

## 4. Discussion

### 4.1. Reasoning of results

The offloading inversion follows directly from who produces the steps. In the instant-answer condition, the tutor's first message contains the full worked solution, so the dialogue gives the student no reason to generate: students produce 0.302 of the steps, self-report offloading of 5.58, and their posttest mirrors their pretest rather than improving beyond it, the pattern Kazatchkova and colleagues report for delegated generation in humans [3]. In the friction conditions, the steps can

DOI: 10.54254/2755-2721/2026.LD38122

only come from the student (strict: never from the tutor; quest: only after the student's own attempt), so SDC rises to 0.923 and 0.999, and the posttest gain (+0.473 / +0.430) is the articulation-level version of the desirable-difficulties effect [4]: slowing the path to the answer is what does the work.

The difference between the two friction conditions is where the session ends. The strict condition imposes unbounded friction: a student who is stuck is stuck, and the session runs as long as the student keeps going (range 2–16 turns). Once the session is long, the student's self-reported motivation drifts toward the floor and the session is recorded as abandoned: 14.4 percent of strict sessions end this way, and abandonment tracks session length at every dose. The quest converts the same friction into bounded progression: a wrong attempt costs nothing and is corrected within two tries, the rescue guarantees forward motion, and session length is structurally bounded (10–13 turns). There is, therefore, no long session in which motivation can decay, and the observed abandonment rate is 0 of 180. This is the structural claim the study was built to test: the reviewed gamification literature had tested surface elements and found partial, context-dependent effects [6], and had never tested whether genre-borrowed structure [9] (checkpoints, low-stakes failure, rescue) could hold a learner inside deliberate friction; in this design, it does, at the completion ceiling.

Two secondary results are consistent with this account. First, the strict condition scores higher on the unaided posttest (0.928 versus 0.885) despite revealing nothing: its students' session notes, the only material allowed on the posttest, contain their own verified, complete solution, whereas the quest students' notes contain their attempts, since the post-gate reveals are tutor messages outside the notes. A posttest note cap means the quest's reveals do not transfer into the exam context, and the gap is the difference between a self-written solution and a self-attempted one. Second, the quest wins the students' own "understood best" judgment (136 of 180), while the strict condition wins the measured posttest: the revealed answer plus the student's own attempt is the stronger in-session explanation, and the self-written solution is the stronger unaided artifact. This paper does not test this dissociation directly and reports it as a plausible account, not a finding.

### 4.2. Implications

Three design principles follow. First, friction should be structured, not bare: the condition that withholds everything loses nearly a seventh of its students, while the condition that frames the same withholding as bounded progression loses none; the lever is the student's ability to see progress and to fail cheaply, not the amount of work required. Second, the reveal belongs after the gate, not instead of it: the quest's per-part reveal is what earns the "understood best" judgment, and withholding it (strict) buys a small posttest advantage at the cost of abandonment. A tutor that answers only after the attempt gets the offloading and the understanding results with none of the attrition. Third, for practitioners, the practical parameter is a friction budget: the maximum session length a learner can sustain before motivation reaches its floor, which a bounded mechanic (attempts-per-step cap, rescue) can hold inside by construction; the quest's 10–13 turn ceiling is exactly that bound made mechanical. The element-versus-structure distinction earns its keep here: the surface-element implementations reviewed by Hamari and colleagues [6, 7] showed partial motivation effects without outcome effects, while the structural mechanics tested here act on the session's shape (length, failure cost, completion certainty), which is the variable that abandonment depends on.

DOI: 10.54254/2755-2721/2026.LD38122

### 4.3. Limitations

The findings are scoped to interaction structure and learner-generated articulation: the agents do not consolidate memory between sessions, so no claim is made about human retention, and the motivation-based abandonment index is a modeled persistence signal, not an observed drop-out. The study covers a single domain (inclined-plane mechanics), a single model family, and a ten-part problem format; per-dose cells are n = 30 (n = 10 per draw), so per-dose comparisons, especially the non-monotone abandonment shape, are descriptive. Two simulator-specific artifacts bound the persistence numbers: the student model reports its own motivation while seeing its own score history, and in long strict sessions it can hold that trajectory artificially high and end the session early by completing the remaining parts in a single message; a prompt-control that instructed the model to report the score as a measurement rather than a goal reduced this batching and reduced abandonment with it (0 of 15 in the matched control), so the 14.4 percent rate is a design-specific value rather than a stable lower or upper bound. One strict-condition session in draw 1 was abandoned only after completion (a post-completion artifact) and is counted in the totals reported above, and two sessions (one C1 in dose 1, one C3 in dose 2) ended on repeated unparseable student output, an actor artifact. Future work is a human replication of the three conditions, a longitudinal retention test in simulated students with cross-session memory, and a multi-domain replication of the quest-gating mechanics.

## 5. Conclusion

This study tested whether game-structural mechanics can keep a learner inside a tutor's deliberate friction. Across 180 simulated students, the quest condition removed the measured offloading—students produced essentially all solution steps themselves and named the instant-answer condition, unanimously, as the one in which they offloaded their thinking most—and it sustained the session, with every session completing, against a 14.4 percent abandonment rate in the friction-only condition. Both friction conditions raised the unaided posttest far above the pretest baseline (0.928 and 0.885 against 0.455); the instant-answer condition reached only 0.563. The lever the study isolates is structural: framing the same friction as bounded progression—checkpoints, low-stakes failure, rescue—changes whether the learner stays. The design is fully specified and reproducible; the natural next steps are human replication and a retention test across sessions.

## References

[1] Oreopoulos, P., & Low, N. (2026). One click away: AI tutoring with Khanmigo in a two-year school experiment (No. w35620) [Working paper]. National Bureau of Economic Research. https://www.nber.org/papers/w35620

[2] Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[3] Kazatchkova, A., Skorburg, J. A., & Fiacconi, C. M. (2026). How does offloading generative processes to ChatGPT impact learning and memory? [Poster presentation]. *Proceedings of the 2026 Annual Meeting of the Cognitive Neuroscience Society*. JoCN Forum. https://doi.org/10.21428/8e6ba8ef.43868463

[4] Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition, 9*(4), 475–479. https://doi.org/10.1016/j.jarmac.2020.09.003

[5] Zepeda, C. D., Martin, R. S., & Butler, A. C. (2020). Motivational strategies to engage learners in desirable difficulties. *Journal of Applied Research in Memory and Cognition, 9*(4), 468–474. https://doi.org/10.1016/j.jarmac.2020.08.007

[6] Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. In *2014 47th Hawaii International Conference on System Sciences (HICSS)* (pp. 3025–3034). IEEE. https://doi.org/10.1109/HICSS.2014.377

DOI: 10.54254/2755-2721/2026.LD38122

[7] Hamari, J., & Koivisto, J. (2015). Why do people use gamification services? *International Journal of Information Management, 35*(4), 419–431. https://doi.org/10.1016/j.ijinfomgt.2015.04.006

[8] Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification: Using game-design elements in non-gaming contexts. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (pp. 2425–2428). ACM. https://doi.org/10.1145/1979742.1979575

[9] Ntokos, K. (2019). Swords and sorcery: A structural gamification framework for higher education using role-playing game elements. *Research in Learning Technology, 27*, Article 2272. https://doi.org/10.25304/rlt.v27.2272

[10] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (pp. 1–22). ACM. https://doi.org/10.1145/3586183.3606763

[11] Scarlatos, A., Lee, J., Woodhead, S., & Lan, A. (2026). Simulated students in tutoring dialogues: Substance or illusion? arXiv. https://arxiv.org/abs/2601.04025