How AI Assistance Affects Human Skill Development: A Study of Learning with Logic Puzzles

Shang Wu
shangw13@uci.edu
University of California, Irvine
USA

Catarina G Belem
cbelem@uci.edu
University of California, Irvine
USA

Shuyuan Fu
shuyuanf@uci.edu
University of California, Irvine
USA

Mark Steyvers
msteyver@uci.edu
University of California, Irvine
USA

Padhraic Smyth
smyth@uci.edu
University of California, Irvine
USA

Abstract

While AI assistance can improve human task performance in the short term, it may also undermine the development of skills in the longer term. We examine this tension in a controlled logic-puzzle experiment involving on-demand AI assistance, where participants complete tasks before, during, and after AI is available. By experimentally varying AI request costs, we find that lower-cost assistance induces more frequent AI use. We also find that participants who request AI assistance during the AI-access phase perform worse at the task after assistance is removed, and their subsequent unassisted performance is overestimated when predicted from earlier AI-assisted performance. We use a Bayesian latent ability model to separate initial ability, post-AI ability, and participant-specific skill change, while estimating how independent reasoning during the AI-access phase relates to skill development. The results show that greater independent problem-solving effort is associated with larger gains in latent ability, consistent with the interpretation that skill development is weaker when AI assistance substitutes for independent reasoning.

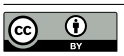
CCS Concepts

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

Keywords

human-AI interaction, individual skill development, reliance, Bayesian modeling

ACM Reference Format:
Shang Wu, Catarina G Belem, Shuyuan Fu, Mark Steyvers, and Padhraic Smyth. 2026. How AI Assistance Affects Human Skill Development: A Study of Learning with Logic Puzzles. In *2026 ACM Conference on Human-AI Complementarity and Alignment (HCOMP 2026), September 27–30, 2026, Alexandria, VA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3834580.3838741



HCOMP 2026, Alexandria, VA, USA

ACM ISBN 979-8-4007-2894-5/2026/09
https://doi.org/10.1145/3834580.3838741

1 Introduction

As AI systems become increasingly used as aids in everyday cognitive activities, an important question is not only whether they improve human cognitive performance but also whether they promote or hinder skill development. While AI assistance can improve immediate task performance in the short term [8, 15], it may also alter the degree to which people engage with a cognitive task. In particular, one significant concern is that readily available AI assistance could reduce the need for sustained independent reasoning and encourage humans to offload cognitive effort to the AI system [3, 17, 24]. This creates a central tension in AI-assisted cognition: the same assistance that helps users solve problems in the short run may weaken the development of the underlying skills needed to solve similar problems independently in the future.

This tension is especially important in settings where learning depends on active engagement. Skill development requires repeated opportunities to reason, test strategies, revise mistakes, and internalize problem structure [1]. If AI assistance provides a substitute for these processes, humans may appear to be successful at a task while the assistance is available, but have weaker performance once the assistance is removed [19, 35]. At the same time, AI use is not necessarily harmful in itself. Humans may request assistance while still engaging deeply with the task, and the consequences of AI use may depend on whether it complements or replaces independent problem-solving effort [18, 23, 25]. Understanding the effect of AI on learning, therefore, requires distinguishing between (a) the frequency of AI use, and (b) how users use it.

We study this issue through a controlled logic-puzzle experiment that tracks skill development across three phases: Phase 1, an initial AI-free assessment; Phase 2, an intermediate phase in which AI assistance may be available; and Phase 3, a final AI-free assessment (see Figure 1). During Phase 2, participants can request AI assistance on demand. We experimentally vary the costs of AI requests across conditions to induce different levels of AI engagement across participants. By comparing participants' performance in Phases 1 and 3, the design allows us to examine how the cost of AI shapes reliance behavior and how participants' engagement with AI during Phase 2 relates to subsequent performance after assistance is removed.

We investigate relationships between various factors related to individual behavior and performance by fitting a Bayesian latent ability model that treats pre- and post-AI accuracy and response

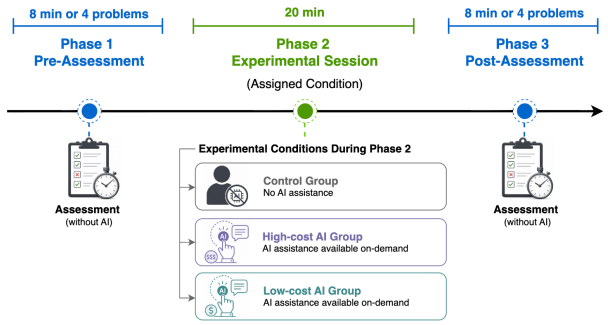


Figure 1: Overview of the three-phase study design. Participants completed Phase 1 and Phase 3 without simulated AI assistance, providing pre- and post-AI measures of unassisted performance. During Phase 2, participants were assigned either to a no-AI control condition or to an AI-access condition with experimentally varied AI request costs.

time as indicators of underlying ability, and that models changes in skill levels as a function of observed behavior, performance measures, and initial ability. Our findings suggest that weaker short-term skill development is associated with AI assistance that displaces independent reasoning. Lower-cost access leads participants to use AI more frequently, and participants who request AI perform worse once assistance is removed. The Bayesian model further suggests that short-term skill development is associated less with how often participants request AI help and more with whether they continue to engage in independent reasoning. Participants who devote more effort to independent problem solving show greater gains in latent ability, whereas the frequency of requests for AI assistance is not associated with gains in skill development after accounting for initial ability and independent effort.

This paper contributes to research on human-AI interaction by adding nuance to how AI assistance affects learning. Our findings suggest that, in a controlled logical puzzle task, the central issue is not AI use itself, but whether assistance displaces the independent reasoning through which skills are built. This distinction is particularly important in educational and training settings, where AI support should ideally improve current task performance without weakening future problem-solving ability.

2 Related Work

Although a large body of work examines human-AI complementarity in joint decision-making contexts [5, 31, 36], recent studies have begun to investigate how AI assistance influences subsequent performance once it is no longer available. A common approach uses pre- and post-assessment designs to compare behavior before AI exposure, during AI-assisted work, and after assistance is removed. In mathematical reasoning and reading-comprehension tasks, Liu et al. [19] find that AI assistance improves short-run performance but reduces persistence and impairs later independent performance. In programming, Shen and Tamkin [25] show that AI assistance can weaken conceptual understanding, code reading, and debugging skills, while also documenting cognitively engaged interaction patterns that preserve learning. Related evidence from writing tasks suggests that LLM assistance may reduce cognitive engagement and weaken later unaided performance [16]. These findings are consistent with broader work on AI as a "tool for thought", which emphasizes that AI systems can either scaffold reasoning or automate cognitive steps that users would otherwise perform themselves [33]. Together, these studies suggest that AI-assisted success does not necessarily translate into durable skill acquisition, and that AI's effects on human learning are context-dependent. Our work adds to this growing literature by focusing on whether learners continue to spend time on independent problem-solving during periods when AI assistance is available, and how this independent effort relates to subsequent skill development.

One mechanism linking AI assistance to skill development is cognitive offloading: the use of external tools to reduce internal cognitive effort [24]. Offloading can improve immediate performance, but it may also reduce learning when externally supported operations replace the reasoning, memory, or problem-solving processes through which skill is acquired [12, 26]. This connects directly to work on AI reliance, which studies when users request, accept, reject, or defer to AI assistance [13, 21]. Prior work shows that reliance depends on features of the task environment, such as AI accuracy, timing, and time pressure [4, 22, 32]. Reliance, however, is not necessarily harmful: on-demand assistance can act as a cognitive forcing mechanism that reduces over-reliance [3], and learning-oriented assistance can be beneficial when it scaffolds understanding rather than allowing users to bypass the task [2, 8]. Thus, a key distinction is not simply whether users rely on AI, but whether AI use preserves or displaces the cognitive work needed for learning.

Prior work also suggests that the consequences of AI assistance depend on how the system changes the user's role in the task. In chess, Poulidis et al. [23] show that the effects of AI assistance depend on whether it prescribes actions or instead directs users' attention, reinforcing the distinction between AI that substitutes for vs. complements human reasoning. In our earlier work [35], we explored the role of AI informativeness and usage in logical reasoning. In this paper we build on this framework to focus more directly on the behavioral mechanisms linking AI assistance to skill development. Specifically, we distinguish between the frequency and timing of AI requests and the extent to which participants preserve independent problem-solving effort when assistance is available. In this context, we develop a Bayesian latent ability model that allows us to disentangle the different factors affecting skill development over time. This allows us to examine whether weaker subsequent learning is better explained by AI use itself or by the displacement of independent reasoning during the AI-access phase.

3 Experiment

To understand how AI assistance affects human skill development, we conducted a controlled user study in the specific context of solving logic puzzles under time constraints, where participants could autonomously request AI assistance while solving problems. By experimentally varying the cost of AI requests, the design allows us to examine how different help-seeking environments shape AI engagement, task performance, and learning.

3.1 Task Description

We build on a logic-puzzle paradigm used in prior work on AI-assisted skill development [35], but modify the AI assistance format to allow for cost-driven help-seeking. In each problem, participants were presented with six objects and a set of five logical constraints that jointly define a unique ordering (Figure 2a). Participants solved the problem by arranging the six objects according to these constraints. An answer was deemed correct (from the user's perspective) only if all six objects were placed in the correct positions. Participants could submit up to two attempts per problem. After the first attempt, participants received feedback indicating the number of correctly positioned objects, but not which specific objects were correctly positioned. Given this feedback, participants could either revise and resubmit their response or accept the initial attempt as final and proceed to the next problem. The correct solution was revealed after each problem to facilitate learning across problems.

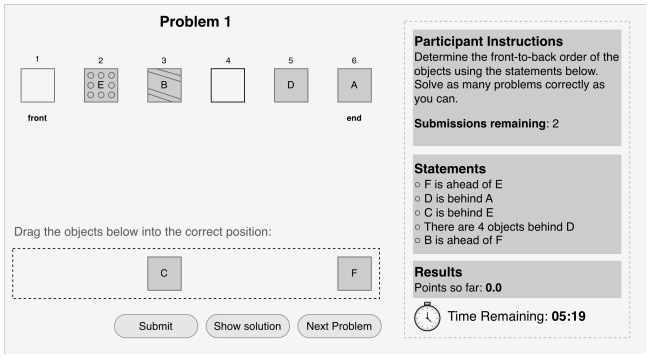
While more open-ended and more realistic tasks could in principle be used in a study like this, this more controlled setting reduces variability in terms of prior human knowledge and supports a clearer focus on problem-solving behavior under optional AI assistance. The task was designed to be intuitive for participants without specialized training, while still requiring logical reasoning under time constraints, reflecting decision settings in which both speed and accuracy are important. In addition, some objects contained recurring visual markers that were associated with specific positions, providing hidden structural cues that participants could learn to exploit across problems.

The study employed a three-phase design that separates baseline performance, AI-assisted task completion, and post-assistance performance, allowing us to examine how AI request costs relate to AI usage patterns and subsequent skill development (Figure 1). Phase 1 served as a *pre-AI assessment* and lasted 8 minutes or until participants completed at least four problems. Phase 2 lasted 20 minutes and introduced the experimental conditions, including optional AI assistance under different costs as described below in Section 3.2. Phase 3 mirrored Phase 1 as a *post-AI assessment* without AI assistance. Problem order was randomized within each phase to mitigate confounding effects of problem difficulty on learning.

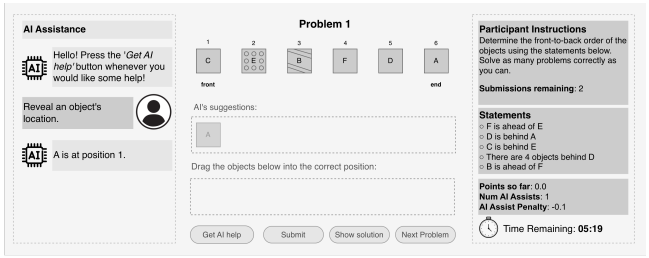
Participants earned one point for each correctly solved problem. During Phase 2, points were deducted for each AI request according to the assigned condition. Bonus payments were calculated at \$0.15 per point in Phases 1 and 2 (up to \$2 total) and \$0.18 per point in Phase 3 (up to \$1.5).

3.2 Conditions

We used a between-subjects design in which participants were randomly assigned to one of three conditions that varied the availability and cost of AI assistance during Phase 2. AI assistance was available only in Phase 2, and participants in AI-available conditions could request it at any time while solving the puzzles. Each request revealed the location of one randomly selected object and incurred a condition-specific point deduction, creating a trade-off between obtaining help and maximizing rewards. Participants were informed of the applicable AI request cost structure in the instructions before starting the study.



(a) Without AI assistance



(b) With AI assistance

Figure 2: User-interface examples from the logic-based puzzle task

Figure 2b shows the user interface for an AI-enabled condition. This interface is representative of all AI-enabled conditions, which share the same interaction design and differ only in the cost of AI requests. Participants could click "Get AI help" to reveal the location of a randomly selected object in the puzzle. Each request updated the right-hand panel by incrementing the number of assists and the deducted points.

Participants were randomly assigned to one of the following conditions:

- ***No-AI***: No AI assistance was available, and the AI panel was not shown (see Figure 2a).
- ***Low-cost AI***: Each AI request incurred a 0.1-point deduction.
- ***High-cost AI***: Each AI request incurred a 0.18-point deduction.

To ensure that variation in AI quality did not confound the effect of AI cost on participants' help-seeking behavior, we fixed the simulated AI agent to be perfectly correct (100%). Under this design, each AI request consistently returned the correct location of a randomly selected object, allowing us to attribute behavioral differences across AI-access conditions to request cost rather than output quality. This design choice aligns with prior work in human–AI interaction that uses simulated agents to isolate specific behavioral mechanisms (e.g., [3, 8, 27, 34]). Participants were not informed of the AI's accuracy and were instructed only that AI assistance would be available during Phase 2 at a cost.

3.3 Procedure

After providing informed consent, participants were randomly assigned to one of the three conditions and presented with task instructions. To ensure task understanding, participants completed two comprehension checks, and only those who passed both proceeded to the main study.

The main study consisted of three sequential phases, separated by short breaks, during which participants completed a set of logic puzzle problems under their assigned condition. To monitor data quality during task performance, an attention check was embedded in Phase 2. After completing all phases, participants completed a short post-study survey about their experience, perceived effort, strategy use, prior task familiarity, and, when applicable, their perceptions and timing of AI use (see Supplementary Material B.1 for details).

3.4 Participants

We recruited 150 English-speaking adults (18+) via Prolific, restricted to adults residing in the U.S. with at least an undergraduate degree. We excluded 26 participants: 1 failed the attention check, 14 showed inattentiveness (off-screen for ≥ 5 seconds on over three problems), 8 had inconsistent self-reported and logged AI use, 2 were extreme time outliers, and 1 showed low-effort responses with near-zero accuracy.[1] The final sample consisted of 124 participants (42 *No-AI*, 43 *Low-cost AI*, 39 *High-cost AI*).

Participants received $9 for participation (median completion time: 56 minutes) and could earn up to an additional $3.50 in performance-based bonuses. The study was approved by the Institutional Review Board (IRB) at the University of California, Irvine (protocol #7206).

The final sample consisted of 124 participants with a mean age of 40 years (SD = 12). The sample included 59 male participants, 64 female participants, and 1 participant who preferred not to disclose gender. In terms of education, 66 participants held an undergraduate degree and 58 held a master's degree or higher.

3.5 Measures

We study both observed task performance and latent skill development. For observed performance, we use **accuracy**, defined as the number of objects (0–6) placed in the correct position for a problem, and **response time**, measured in seconds. Both are computed per participant for each of the 3 phases, averaging over the different problems that a participant attempted during a phase. To summarize the speed-accuracy tradeoff, we also compute **reward rate**, as used in cognitive modeling [30]. For each participant in each phase, reward rate is defined as the average accuracy divided by the average response time, in units of correct objects per minute, where higher values indicate better performance.

To characterize reliance on AI during Phase 2, we measure **AI usage** as the total number of assistance requests made by a participant. We also measure independent problem-solving effort using **solo share**. At the problem level, this measure is defined as the fraction of response time spent working independently before requesting AI assistance. If no AI assistance is requested, the time until final submission is counted as the time working independently. Participant-level solo share is then defined as the average problem-level solo share across Phase 2 problems.

In Section 4.1 we use reward rate, accuracy, and response time to empirically summarize performance, to characterize AI reliance behavior, and to examine whether Phase 2 performance predicts subsequent unassisted performance differently across levels of AI use. Our Bayesian analyses in Section 4.2 use Phase 1 and Phase 3 performance as observed signals of latent ability, and use Phase 2 solo share as the main behavioral measure to predict the final (Phase 3) skill levels of participants, both with and without AI assistance.

4 User Study Results

We begin below with a high-level summary of results from our user study, followed by a more detailed Bayesian analysis to quantify the effect of AI assistance on skill development at both individual and population-wide levels.

4.1 Empirical Findings

Skill Development: To examine whether participants in general showed task-specific skill improvement over time, we show in Figure 3 the average reward rate across phases for participants who did not request any AI assistance. The unassisted task performance of this set of participants improved substantially from the pre-AI assessment (Phase 1) to the post-AI assessment (Phase 3). The mean reward rate increased from 2.03 correct objects per minute in Phase 1 to 3.86 in Phase 3, corresponding to a 90.2% increase in the number of correct objects per minute ($N = 75$, $t = 7.40$, $p < 0.01$), illustrating substantial short-term improvement in participants' skills on this task.

Effects of AI-cost condition: We next examine the effects of the randomized AI-cost manipulation. Phase 1 reward rates were well balanced across conditions (ANOVA: $F = 0.20$, $p = 0.82$). As shown in Figure 4, participants in the *Low-cost AI* condition made significantly more AI requests during Phase 2 than those in the *High-cost AI* condition (6.67 vs. 3.33 requests; one-sided Welch test: $t = 1.48$, $p < 0.10$). From Phase 1 to Phase 3, reward rate increased by 1.95 in the *No-AI* condition, 1.66 in the *High-cost AI* condition, and 1.32 in the *Low-cost AI* condition; the increase was smaller in the *Low-cost AI* than in the *No-AI* condition in a one-sided t-test ($t = 1.30$, $p < 0.10$). In Phase 3, accuracy did not differ across conditions (ANOVA: $F = 0.097$, $p = 0.91$), while response times were longer in the *Low-cost AI* condition ($M = 111.81$s, $SE = 6.13$) than in the *High-cost AI* ($M = 86.78$s, $SE = 4.18$; $p < 0.01$) or *No-AI* condition ($M = 92.21$s, $SE = 4.88$; $p = 0.01$).

Post-AI Performance: Descriptively, participants who requested AI assistance in Phase 2 had a lower average reward rate (3.42 correct items/minute) in Phase 3 (where no AI assistance was available) compared to the average reward rate in Phase 3 (3.86 correct items/minute) of participants who did not request AI in Phase 2. To explore this further, we use a regression model to predict Phase 3

[1] Off-screen time is total problem time minus active on-screen time. Inconsistent AI-use reporting refers to disagreement between post-study reports and usage logs. The low-effort participant expressed dislike for the study and answered randomly throughout.

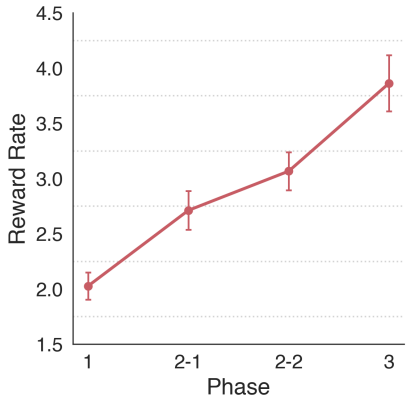


Figure 3: Reward rate across phases for participants with zero AI use in Phase 2. Labels 2-1 and 2-2 denote the first and second halves of Phase 2. Error bars indicate standard errors.

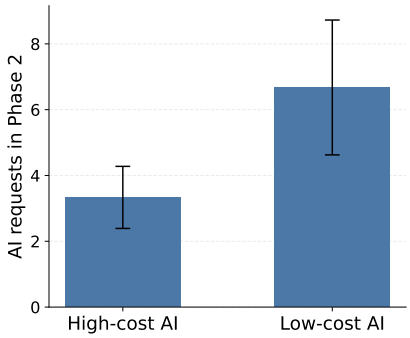


Figure 4: Average AI usage in Phase 2 by AI cost condition. Error bars indicate standard errors.

reward rate as a function of Phase 2 reward rate and find that predictions are systematically biased by AI use. Specifically, as shown in Figure 5, for participants who did not request AI, Phase 3 performance was underestimated by 0.15 reward-rate units on average. In contrast, for AI users, Phase 3 performance was overestimated by 0.22 reward-rate units on average. In other words, the apparent performance of participants while using AI (in Phase 2) systematically overestimates their subsequent performance when conducting the task on their own without AI assistance (in Phase 3).

These patterns suggest that Phase 2 AI-use behavior is associated with differences in short-term skill improvement from the pre-AI phase (Phase 1) to the post-AI phase (Phase 3). We investigate these relationships further below via a Bayesian modeling approach, separating initial ability, post-AI ability, and latent skill change while incorporating Phase 2 AI usage behavior.

4.2 Bayesian Analysis

We use a Bayesian model to infer latent individual ability and skill development, conditioned on observed individual-level data (accuracy, response time, AI usage, solo share) across the three experiment phases. This approach follows a long tradition in education

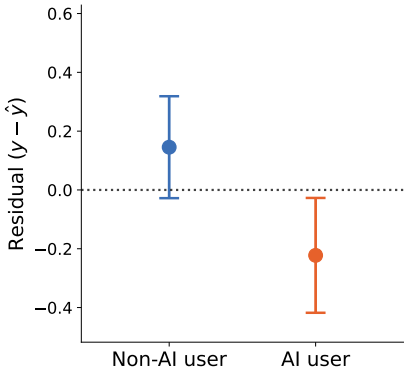


Figure 5: Residuals by AI-use status. Residuals are defined as observed minus predicted Phase 3 reward rate, $y - \hat{y}$; error bars indicate standard errors.

and psychometrics in modeling ability as a latent trait inferred from observed performance [7, 20]. It is useful in our setting because observed performance is a noisy proxy for ability: accuracy and response time provide complementary signals, and the same observed Phase 3 performance can correspond to different amounts of learning depending on participants' Phase 1 baseline performance and Phase 2 AI reliance behavior. A Bayesian latent ability framework allows us to separate initial ability from post-AI ability and participant-level skill change, while propagating uncertainty in both individual- and population-level quantities. In particular, we use the model to infer each participant's individual latent initial ability, latent post-AI ability, and latent skill change, as well as to estimate how short-term skill change is associated with independent problem-solving effort in Phase 2. We outline below the structure of the model in correspondence with the three phases of the experiment. We standardize all observed performance measures prior to fitting the model to improve the interpretability of coefficients across variables measured on different scales [10, 11], and we use weakly informative priors as outlined in the Supplementary Material A.1.

4.2.1 A Bayesian latent ability model of short-term skill development, with and without AI assistance. Beginning with Phase 1 (pre-AI, initial assessment), for individual i, let θ_{i1} denote latent initial ability, with a prior $\theta_{i1} \sim \mathcal{N}(0, 1)$. We model an individual's Phase 1 outcomes (accuracy A_{i1} and log response time R_{i1}) as conditionally independent noisy functions of their latent ability θ_{i1}:

$$\begin{aligned} A_{i1} \mid \theta_{i1}, \mu_A, \sigma_A &\sim \mathcal{N}(\mu_A + \theta_{i1}, \sigma_A) \\ R_{i1} \mid \theta_{i1}, \mu_R, \sigma_R &\sim \mathcal{N}(\mu_R - \theta_{i1}, \sigma_R) \end{aligned} \tag{1}$$

where the effect of θ_{i1} is reversed for the two variables (in terms of the signs in the mean terms) to reflect the expectation that higher values of ability will correspond (on average) to higher accuracy and to lower response times. Here, μ_A and μ_R are measurement intercepts for accuracy and log response time, respectively, and σ_A and σ_R capture residual measurement variation. Because A_{i1} and R_{i1} are phase-level measures aggregated over the problems completed by each participant, the model does not separately estimate problem-level effects. Problems were designed to be comparable in

difficulty and their order was randomized within each phase, so any remaining variation in problem difficulty, order, or the number of problems completed is treated as part of the residual measurement variation. These population-level parameters are inferred jointly during Bayesian inference along with the individual-level latent abilities and skill changes introduced below.

For Phase 2, where AI is available to some participants, we use solo share Solo_i as the observed behavioral measure in our main model, where Solo_i is the participant-level fraction of Phase 2 problem-solving time that participant i spends reasoning independently prior to (or without) AI assistance (see Section 3.5). We model participant-specific latent skill change, δ_i, as a function of initial ability and the Phase 2 engagement measure Solo_i:

$$\delta_i \mid \theta_{i1}, \text{Solo}_i \sim \mathcal{N}\left(\alpha_0 + \alpha_\theta \theta_{i1} + \alpha_{\text{solo}} \text{Solo}_i,\ \sigma_\delta\right). \tag{2}$$

Since δ_i is not constrained to be positive, the model can accommodate either skill improvement or skill decline after AI assistance. The intercept α_0 represents the expected latent skill change for a participant with average initial ability and average solo share. The coefficient α_θ allows latent skill change to vary with initial ability, accounting for the possibility that lower- and higher-ability participants have different room to improve or learn at different rates (e.g., allowing for saturation effects for individuals with very high initial ability). The α_{solo} coefficient captures the association of independent reasoning with latent skill change, conditioned on initial ability. Note that because solo share reflects observed participant behavior rather than a randomized treatment, this coefficient should be interpreted as associational rather than causal. The parameter σ_δ reflects residual heterogeneous variability in latent skill change that is not explained by initial ability or solo share. An alternative option would be to model latent skill change using AI usage as a Phase 2 behavioral predictor, either in addition to or instead of solo share. We use solo share in the main model because, as shown later, it led to better held-out predictive performance than AI request frequency.

For Phase 3, where AI is not available (as in Phase 1), let A_{i3} denote Phase 3 accuracy and R_{i3} denote Phase 3 log response time. Phase 3 outcomes are modeled analogously to Phase 1, i.e., as noisy functions of an individual's latent Phase 3 ability θ_{i3}:

$$\begin{aligned} A_{i3} \mid \theta_{i3}, \mu_A, \sigma_A &\sim \mathcal{N}(\mu_A + \theta_{i3}, \sigma_A) \\ R_{i3} \mid \theta_{i3}, \mu_R, \sigma_R &\sim \mathcal{N}(\mu_R - \theta_{i3}, \sigma_R) \end{aligned} \tag{3}$$

We model an individual's latent ability θ_{i3} as the deterministic sum of their initial ability θ_{i1} offset by the individual's change in ability δ_i, i.e., $\theta_{i3} = \theta_{i1} + \delta_i$, where posterior uncertainty about θ_{i3} is driven by uncertainty about θ_{i1} and δ_i.

We summarize the overall structure of the Bayesian model as a directed graphical model in Figure 6, where shaded nodes denote observed variables and unshaded nodes denote latent variables and unknown parameters in the model. The nodes inside the plate reflect replication over the N participants, where nodes inside the plate are at the individual level (with subscript i) and are independent (across the N individuals) conditioned on the population-level parameters outside the plate. The accuracy A and response time R in Phases 1 and 3 depend on the corresponding latent abilities θ_{i1} and θ_{i3} for these Phases. The dependence of θ_{i3} on δ_i and θ_{i1} represents the latent transition from initial ability to post-AI ability.

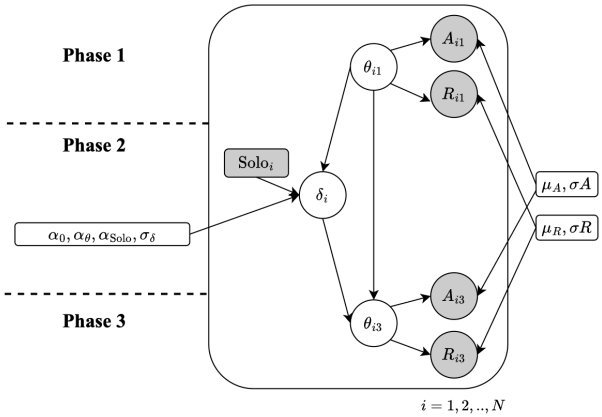


Figure 6: Bayesian model of latent skill development. Phase 1 and Phase 3 accuracy and response time are modeled as noisy measurements of latent abilities θ_{i1} and θ_{i3}. Participant-specific skill change δ_i depends on initial ability and solo share. Shaded nodes are observed variables, unshaded nodes are latent variables or unknown parameters, and the plate denotes participants $i = 1, \ldots, N$.

Finally, latent ability change, δ_i, is modeled as a function of initial ability θ_{i1} and solo share, allowing the model to capture associations between individual skill development and the amount of independent reasoning participants engage in during Phase 2. The graphical model represents the probabilistic dependencies assumed in the model and is not intended as a causal model of how Phase 2 behavior affects skill change. We note that the model does not include variables for observed participant Phase 2 accuracy A_{i2} or response time R_{i2}. We omit these measures to keep the specification focused on whether Phase 2 independent reasoning is associated with subsequent latent skill change beyond initial ability, rather than modeling Phase 2 task performance directly.

Given the model described above, Bayesian inference proceeds by conditioning on $\mathcal{D} = \{A_{i1}, A_{i3}, R_{i1}, R_{i3}, \text{Solo}_i\}_{i=1}^N$, and estimating the joint posterior distribution over all unknown quantities. Because the likelihood and priors do not yield a conjugate closed-form posterior, we approximate the posterior using Markov Chain Monte Carlo (MCMC) (see implementation details in Supplementary Material A.2). Standard MCMC diagnostics indicated satisfactory convergence, and we additionally assess model fit using posterior predictive checks and the uncertainty of held-out predictive comparisons; details are reported in Supplementary Material A.3.

We note that the randomized AI-cost condition (low-cost, high-cost) affects participants' opportunities and incentives to request assistance, and this condition could be modeled as an upstream parent of solo share in the model. Because the primary specification conditions directly on this observed behavioral measure, we omit the condition node from the graphical model in Figure 6 for simplicity. In Supplementary Material A.4 and Supplementary Material A.5, we report the results of alternative model specifications that add AI-cost condition controls and that replace solo share with AI usage, and find that these alternative model specifications do not add any additional explanatory power to the model.

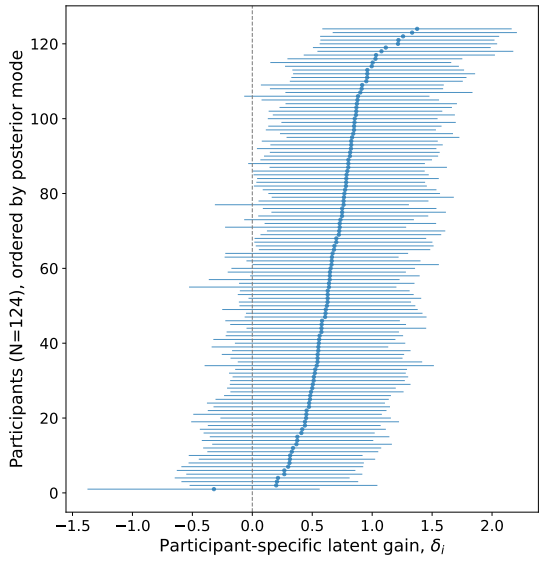
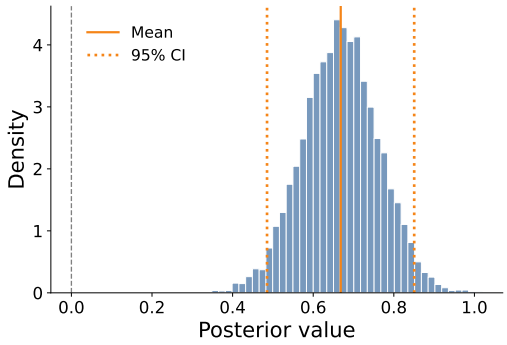


Figure 7: Posterior modes and 95% credible intervals of participant-specific latent ability changes, δ_i, ordered by posterior mode. The vertical dashed line indicates zero change.

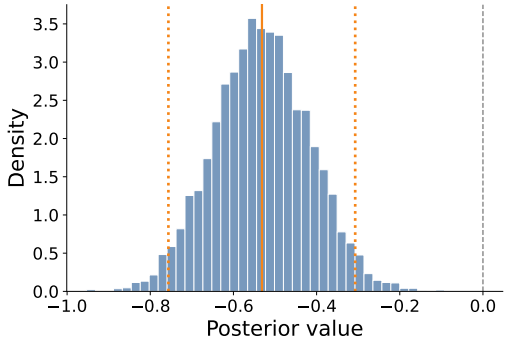
4.2.2 Independent reasoning is more predictive of short-term skill development than AI request frequency. We summarize the results obtained from the posterior samples after running MCMC inference for the model described in Figure 6. Figure 7 visualizes the posterior modes and 95% credible intervals of participant-specific latent ability changes, δ_i. Almost all participant-level posterior modes are positive, indicating positive estimated skill development for most participants. This is consistent with the earlier empirical observations of skill development at the population level in Figure 3. The credible intervals are wide, indicating substantial uncertainty and heterogeneity in individual-level skill development, but nonetheless broadly indicate positive skill gains.

The posterior distributions of the three population-level coefficients in the latent ability model are shown in Figure 8. The posterior distribution of the average expected latent skill change, α_0, is positive: for a participant with average initial latent ability and average solo share, the expected latent ability change is $\alpha_0 = 0.668$, with 95% CI $[0.485, 0.851]$. Because latent ability is measured jointly through accuracy and response time, a positive gain corresponds to improved accuracy and/or faster responses in the post-AI assessment. Figure 8b shows that the coefficient α_θ linking latent skill change to initial ability is strongly negative, with an expected posterior value of -0.531 and 95% CI $[-0.756, -0.307]$, with posterior probability $P(\alpha_\theta < 0) = 1.000$. This indicates that participants with lower initial latent ability tend to show larger gains, consistent with greater room for improvement among initially lower-performing participants.

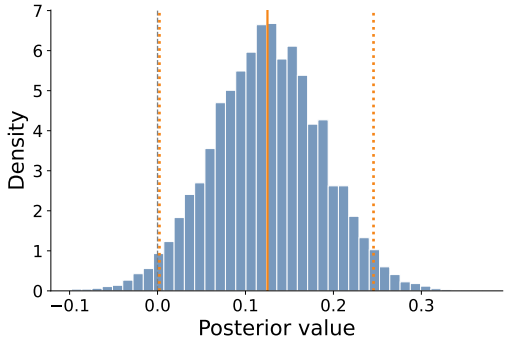
Importantly, Figure 8c shows that solo share is positively associated with latent skill change. The posterior mean of α_{solo} is 0.125, with 95% CI $[0.002, 0.246]$ and posterior probability $P(\alpha_{\text{solo}} > 0) = 0.977$. Thus, participants who spent more time independently solving problems (without AI assistance) during Phase 2 exhibited



(a) α_0: Average expected skill change



(b) α_θ: Impact of initial ability on skill change



(c) α_{solo}: Solo share

Figure 8: Posterior distributions of latent ability regression coefficients. Histograms show posterior draws; orange solid lines denote posterior means, orange dotted lines denote 95% credible intervals, and black dashed lines mark zero.

larger gains in latent ability from the pre-AI to the post-AI assessment. To interpret the magnitude of this association, we compute a model-implied contrast between two otherwise average participants whose Phase 2 solo share differs by one minute out of the 20-minute AI-access phase. Relative to the observed population mean value of solo share, this one-minute difference corresponds to a 1.5% increase in predicted Phase 3 accuracy and a 2.5% reduction in predicted response time. Together, these results suggest that, in

Table 1: Five-fold held-out prediction of Phase 3 unassisted performance across Bayesian latent-ability models. The main model uses Phase 2 solo share to predict latent skill change; the condition model adds AI-cost indicators; the AI-usage model replaces solo share with total AI requests; and the constant-change baseline assumes identical expected skill change across participants. Test LogL is the total held-out log predictive likelihood (higher is better). MSE is the mean squared prediction error for standardized Phase 3 accuracy and log response time (lower is better).

Model	Test LogL↑	Accuracy MSE↓	Log RT MSE↓
Solo share (main)	**-329.94**	**0.809**	**0.850**
Solo share + AI-cost conditions	-331.45	0.830	0.854
AI usage	-332.96	0.831	0.882
Constant change (baseline)	-339.02	0.893	0.915

this context, the key behavioral distinction is not how often participants use AI, but how much independent reasoning they preserve while using it. Participants who preserve more independent reasoning show greater short-term skill gains, consistent with weaker skill development when AI assistance is accompanied by greater cognitive offloading.

As an alternative specification, we also fit a model that replaces solo share with AI usage, measured as the total number of assistance requests during Phase 2 (see details in Supplementary Material A.5). We find that AI usage is not associated with latent skill change after accounting for initial ability, with a posterior for α_{usage} centered at a posterior mean of 0.0004, 95% CI $[-0.122, 0.124]$, and $P(\alpha_{\text{usage}} > 0) = 0.501$. Thus, in this setting, AI request frequency alone does not appear to explain latent skill change. The results are consistent with the hypothesis that time spent on independent reasoning is more predictive of skill development than simple frequency of AI use.

We further compare the main solo-share model discussed above against three alternative specifications: a solo-share model with direct AI-cost condition controls, a usage-only model that replaces solo share with AI usage (the count of the total number of AI requests), and a constant-change baseline in which all participants share the same expected latent skill change, i.e., $\delta_i = \alpha_0$, with no participant-level effects for latent skill change. We evaluate the four specifications using 5-fold cross-validation at the participant level. In each fold, the model is fit on four-fifths of participants and used to predict Phase 3 accuracy and log response time for the remaining participants based on their Phase 1 performance and the Phase 2 predictors included in the corresponding model. We evaluate predictive performance using the total held-out log-likelihood, where higher values indicate better overall predictive fit, and the mean-squared error (MSE) for standardized Phase 3 accuracy and log response time, where lower values indicate better point prediction. Table 1 summarizes the results. The solo-share model achieves the best predictive performance across all three metrics, with the highest held-out log-likelihood and the lowest MSE for both accuracy and log response time. Additional details of the evaluation procedure are provided in Supplementary Material A.6.

5 Discussion and Conclusion

This study examines the role of AI assistance in short-term, task-specific skill development. In a controlled three-phase logic-puzzle experiment, we find that lower-cost AI requests result in increased AI use, and that participants who request AI assistance perform worse after AI assistance is removed. Our Bayesian latent ability model further suggests that skill development is more closely associated with the amount of independent reasoning participants preserve during the AI-access phase than with AI request frequency itself. These findings contribute to the growing literature on AI-enabled deskilling, cognitive offloading, and learning. Prior research shows that external support can improve immediate performance while reducing the cognitive effort needed for later memory, reasoning, or skill formation [12, 24, 26]. More recent studies similarly raise concerns that AI assistance can reduce persistence, weaken conceptual understanding, or encourage cognitive disengagement [6, 16, 25]. Our results suggest that, in this setting, the central issue is not necessarily AI use per se, but how much independent thinking participants preserve while AI assistance is available. The practical implication is that AI systems should be designed not only to improve immediate performance, but also to preserve users' active engagement. Interfaces that encourage an initial attempt, delay assistance, provide partial hints, or prompt reflection may help users benefit from AI without fully offloading the independent reasoning process.

This study also points to several directions for future work. Our experiment uses a simulated AI assistant and focuses on short-term skill development in a controlled logic-puzzle task, which helps isolate the relationship between AI use, independent reasoning, and subsequent unassisted performance. Future studies could examine whether similar patterns persist over longer time horizons, in larger and more diverse samples, and in more realistic domains such as education, professional training, and other high-stakes settings. It would also be useful to study how variation in AI accuracy, explanation quality, and interaction design affects both immediate performance and longer-term skill development. In addition, because the Bayesian model uses participant-level phase aggregates rather than problem-level responses, it does not separately identify residual differences in problem difficulty or problem exposure, providing another potential dimension for future studies.

In conclusion, our work suggests that weaker short-term skill development is associated with greater AI use, particularly when that use involves less independent reasoning. In the context of skill development, designing AI systems that support rather than replace human thinking may therefore be important for improving not only immediate task performance but also subsequent unassisted performance.

Acknowledgments

We thank the reviewers for their valuable feedback that helped improve this paper. This work was supported by the National Science Foundation under awards NSF 2505006 and NSF IIS-2046873, by the Hasso Plattner Institute (HPI) Research Center in Machine Learning and Data Science at UCI, and by funding support from SAP.

References

[1] John R. Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*. 4, 2 (1995), 167–207.

[2] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. 2004. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 383–390.

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[4] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How time pressure in different phases of decision-making influences human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–26.

[5] Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Rodriguez. 2024. Towards human-AI complementarity with prediction sets. *Advances in Neural Information Processing Systems* 37 (2024), 31380–31409.

[6] John Ian Wilzon T Dizon, Norman B Mendoza, Dragan Gasevic, and Fraide A. Ganotice, Jr. 2026. Assessing AI-driven metacognitive offloading: Initial development and validation of the metacognitive laziness scale. *ECNU Review of Education* 9, 2 (2026). doi:10.1177/20965311261450994

[7] Susan E. Embretson and Steven P. Reise. 2025. *Item Response Theory: Foundations for Psychologists and Social Scientists*. Taylor & Francis.

[8] Krzysztof Z Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 794–806.

[9] Andrew Gelman. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* 1, 3 (2006), 515 –534. doi:10.1214/06-BA117A

[10] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis* (3 ed.). CRC Press.

[11] Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[12] Sandra Grinschgl, Frank Papenmeier, and Hauke S Meyerhoff. 2021. Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology* 74, 9 (2021), 1477–1496.

[13] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.

[14] Matthew D Hoffman, Andrew Gelman, et al. 2014. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* 15, 1 (2014), 1593–1623.

[15] Sheer Karny, Lukas William Mayer, Jackie Ayoub, Miao Song, Haotian Su, Danyang Tian, Ehsan Moradi-Pari, and Mark Steyvers. 2024. Learning with AI assistance: A path to better task performance or dependence?. In *Proceedings of the ACM Collective Intelligence Conference*. 10–17.

[16] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872* 4 (2025).

[17] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[18] Matthias Lehmann, Philipp B Cornelius, and Fabian J Sting. 2024. AI meets the classroom: When does ChatGPT harm learning? *arXiv preprint arXiv:2409.09047* (2024).

[19] Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A Bakker, and Rachit Dubey. 2026. AI assistance reduces persistence and hurts independent performance. *arXiv preprint arXiv:2604.04721* (2026).

[20] Frederic M Lord and Melvin R Novick. 2008. *Statistical Theories of Mental Test Scores*. IAP.

[21] Poornima Madhavan and Douglas A Wiegmann. 2007. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors* 49, 5 (2007), 773–785.

[22] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[23] Stefanos Poulidis, Haosen Ge, Hamsa Bastani, and Osbert Bastani. 2025. Action vs. attention signals for human-AI collaboration: Evidence from chess. *The Wharton School Research Paper* (2025).

[24] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in Cognitive Sciences* 20, 9 (2016), 676–688.

[25] Judy Hanwen Shen and Alex Tamkin. 2026. How AI impacts skill formation. *arXiv preprint arXiv:2601.20245* (2026).

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

[27] Divya K Srivastava, J Mason Lilly, and Karen M Feigh. 2022. Improving human situation awareness in AI-advised decision making. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–6.

[28] Stan Development Team. 2024. *Stan User's Guide*. https://mc-stan.org/docs/stan-users-guide/

[29] Stan Development Team. 2025. Prior Choice Recommendations. Stan Wiki.

[30] Dominic Standage, David H. Wang, Richard P. Heitz, and Patrick Simen. 2015. Toward a unified view of the speed-accuracy trade-off. *Frontiers in Neuroscience* 9 (2015), 139.

[31] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.

[32] Siddharth Swaroop, Zana Buçinca, Krzysztof Z Gajos, and Finale Doshi-Velez. 2024. Accuracy-time tradeoffs in AI-assisted decision making under time pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 138–154.

[33] Lev Tankelevitch, Elena L Glassman, Jessica He, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, and Hari Subramonyam. 2025. Understanding, protecting, and augmenting human cognition with generative AI: A synthesis of the CHI 2025 tools for thought workshop. *arXiv preprint arXiv:2508.21036* (2025).

[34] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[35] Shang Wu, Hongyu Yao, Catarina Belém, Shuyuan Fu, Mark Steyvers, and Padhraic Smyth. 2026. The impact of AI usage and informativeness on skill development in logical reasoning. In *Proceedings of the 5th International Conference on Hybrid Human-Artificial Intelligence (HHAI 2026)*. 145–159. doi:10.3233/FAIA260500

[36] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You complete me: Human-AI teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–28.

A Bayesian Modeling and Inference Details

A.1 Prior specification

We use weakly informative priors for all parameters in the Bayesian latent ability model. Because the observed performance measures and predictors are standardized before model fitting, these priors place coefficients on a common scale. For latent initial ability, we use a standard normal prior, $\theta_{i1} \sim \mathcal{N}(0, 1)$. For the latent skill-change component, we assign normal priors to the population-level regression coefficients, $\alpha_0, \alpha_\theta, \alpha_{\text{solo}} \sim \mathcal{N}(0, 1)$. We place an exponential prior on the residual skill-change scale, $\sigma_\delta \sim \text{Exponential}(2)$. For the measurement model, we use normal priors for the accuracy and response-time intercepts, $\mu_A, \mu_R \sim \mathcal{N}(0, 1)$, and exponential priors for the residual standard deviations, $\sigma_A, \sigma_R \sim \text{Exponential}(1)$. These exponential priors restrict scale parameters to positive values and provide weak regularization toward smaller residual variation, consistent with common recommendations for priors on scale parameters in hierarchical Bayesian models [9, 29].

A.2 MCMC inference

We fit the Bayesian latent ability model in Stan through the Python interface CmdStanPy, using the No-U-Turn Sampler (NUTS), an adaptive Hamiltonian Monte Carlo algorithm [14, 28]. For the main specification, we run four chains, each with 1,000 warmup iterations and 2,000 post-warmup sampling iterations, yielding 8,000 posterior draws in total. We set the target acceptance probability to 0.99 and the maximum tree depth to 12.

The model is estimated using a non-centered parameterization for participant-specific latent skill changes, which improves sampling efficiency in hierarchical models by separating the standard-normal skill-change residuals from the residual skill-change scale. Posterior summaries are computed empirically from the retained MCMC draws. Specifically, posterior means are computed as the average of posterior draws, 95% credible intervals are computed using the 2.5th and 97.5th posterior percentiles, and posterior probabilities such as $P(\alpha_{\text{solo}} > 0)$ are computed as the fraction of posterior draws satisfying the corresponding inequality. Participant-level summaries of latent initial ability, post-AI ability, and latent skill change are computed in the same way from the posterior draws of θ_{i1}, θ_{i3}, and δ_i.

A.3 Model Diagnostics and Predictive Checks

MCMC convergence. We assess MCMC convergence using standard sampling diagnostics. For the main solo-share model, all reported $\hat{R}$ values are below 1.01, and the rank-normalized effective sample sizes are satisfactory. We observe no divergent transitions, and the tree-depth and E-BFMI diagnostics indicate no sampling problems. Together, these diagnostics indicate satisfactory convergence and mixing of the posterior samples.

Posterior predictive checks. We assess posterior predictive fit by comparing observed Phase 1 and Phase 3 accuracy and log response-time distributions with replicated outcomes from the posterior predictive distribution. The observed means of all four measures and the standard deviations of three of the four fall within their corresponding 95% posterior predictive intervals; the model somewhat overpredicts the dispersion of Phase 1 log response time. Overall, the model reproduces the central tendencies and most of the observed dispersion well.

Uncertainty in cross-validation comparisons. In addition to the aggregate held-out results reported in Table 1, we examine uncertainty in predictive performance across held-out participants. The mean held-out log predictive density per participant (standard error) is −2.661 (0.090) for the main solo-share model, −2.673 (0.090) for the solo-share model with AI-cost condition indicators, −2.685 (0.090) for the AI-usage model, and −2.734 (0.088) for the constant-change baseline. Thus, although the solo-share model has the highest held-out predictive performance, the differences among the specifications are modest. The cross-validation results therefore provide additional predictive evidence in favor of the solo-share specification.

A.4 Alternative specification with solo share and AI-cost condition controls

As an alternative specification to the main solo-share model, we allow the assigned AI-cost condition to enter the Bayesian latent ability model directly in addition to solo share. This specification tests whether the randomized condition explains additional variation in latent skill change after accounting for initial ability and solo share. We use the no-AI condition as the reference group and define two indicator variables: $\text{LowCost}_i = 1$ for participants assigned to the low-cost AI condition, and $\text{HighCost}_i = 1$ for participants assigned to the high-cost AI condition. The latent skill-change component is then

$$\delta_i \sim \mathcal{N}\left(\alpha_0 + \alpha_\theta \theta_{i1} + \alpha_{\text{low}}\text{LowCost}_i + \alpha_{\text{high}}\text{HighCost}_i + \alpha_{\text{solo}}\text{Solo}_i,\ \sigma_\delta\right).$$

The remaining measurement model is unchanged from the main specification. Thus, Phase 1 and Phase 3 accuracy and log response time are modeled as noisy measurements of latent ability, and post-AI ability is defined as $\theta_{i3} = \theta_{i1} + \delta_i$.

The posterior estimates show that adding direct condition effects does not provide clear additional explanatory information. The coefficient for the low-cost AI condition is negative but highly uncertain, $\alpha_{\text{low}} = -0.084$, 95% CI $[-0.399, 0.225]$. The coefficient for the high-cost AI condition is positive but also highly uncertain, $\alpha_{\text{high}} = 0.062$, 95% CI $[-0.238, 0.359]$. Both credible intervals contain zero. In contrast, the initial-ability coefficient remains strongly negative, $\alpha_\theta = -0.531$, 95% CI $[-0.762, -0.304]$, consistent with the main model. The solo-share coefficient remains positive but is somewhat attenuated relative to the main specification, $\alpha_{\text{solo}} = 0.114$, 95% CI $[-0.017, 0.245]$.

Overall, this alternative specification does not indicate that the assigned AI-cost condition has an additional direct association with latent skill change once initial ability and solo share are included. This supports the main specification, which focuses on preserved independent reasoning during Phase 2 rather than modeling assigned condition as a direct predictor of skill development.

A.5 Alternative specification with AI usage

As an alternative specification, we replace solo share with AI usage in the Bayesian latent ability model. This specification tests whether the total number of AI assistance requests explains latent skill

change, rather than the amount of independent reasoning preserved during Phase 2. Let Usage_i denote standardized AI usage, measured as the total number of assistance requests made by participant i during Phase 2. The latent skill-change component is then

$$\delta_i \sim \mathcal{N}\left(\alpha_0 + \alpha_\theta \theta_{i1} + \alpha_{\text{usage}} \text{Usage}_i,\ \sigma_\delta\right).$$

The remaining measurement model is unchanged from the main specification. Phase 1 and Phase 3 accuracy and log response time are modeled as noisy measurements of latent ability, and post-AI ability is defined as $\theta_{i3} = \theta_{i1} + \delta_i$.

The posterior estimates show that AI usage does not explain latent skill change. The coefficient on AI usage is essentially zero, $\alpha_{\text{usage}} = 0.0004$, 95% CI $[-0.122, 0.124]$, with $P(\alpha_{\text{usage}} > 0) = 0.501$. Thus, higher request frequency is not associated with either higher or lower latent skill change after accounting for initial ability. The initial-ability coefficient remains negative, $\alpha_\theta = -0.495$, 95% CI $[-0.725, -0.268]$, consistent with the main specification and indicating that participants with lower initial latent ability tended to show larger skill gains.

Overall, this alternative specification suggests that latent skill development is not explained by AI request frequency itself. Together with the main solo-share model, this supports the interpretation that the relevant behavioral signal is the extent to which participants preserved independent reasoning during Phase 2, rather than how often they requested AI assistance.

A.6 Held-out model evaluation

We compare four Bayesian latent ability model specifications. The first is the main solo-share model described in Section 4.2.1, in which latent skill change depends on initial latent ability and Phase 2 solo share. The second adds direct AI-cost condition controls to the solo-share model, as described in Supplementary Material A.4. The third is the AI-usage model in Supplementary Material A.5, which replaces solo share with the total number of AI requests. The fourth is a constant-change baseline, in which all participants share the same expected latent skill change, $\delta_i = \alpha_0$, so post-AI ability differs only through baseline latent ability and measurement noise.

We evaluate these specifications using 5-fold held-out prediction of Phase 3 outcomes. In each fold, the model is fit using only the four-fifths of participants assigned to the training fold. For these training participants, the model observes the Phase 1 performance measures, Phase 3 outcomes, and any behavioral predictors used by the corresponding specification. The remaining one-fifth of participants are excluded from model fitting entirely. After fitting, we use only the held-out participants' prediction inputs to compute posterior predictive distributions for their Phase 3 outcomes. These inputs are Phase 1 performance and solo share for the main solo-share model; Phase 1 performance, solo share, and condition indicators for the condition-control model; Phase 1 performance and AI usage for the usage-only model; and Phase 1 performance only for the constant-change baseline.

The primary evaluation metric is held-out log likelihood, which measures out-of-sample predictive accuracy by evaluating the posterior predictive probability assigned to held-out observations. For each held-out participant, we compute the log predictive probability of the observed Phase 3 accuracy and log response time under the posterior predictive distribution, averaging over posterior draws from the model fit on the training fold. We then sum this quantity across held-out participants and folds, with higher values indicating better predictive fit. We also report held-out mean squared error (MSE), computed by comparing posterior mean predictions with the observed held-out Phase 3 accuracy and log response time; lower MSE values indicate better point prediction.

As reported in Table 1, the main solo-share model achieves the highest held-out log likelihood and the lowest held-out MSE for both Phase 3 accuracy and log response time. Thus, the held-out results favor the solo-share model on overall predictive fit.

B Experiment Details

B.1 Post-Study Survey

After completing the three-phase study, participants were asked to fill out a short post-study survey. This survey was originally used in the pilot study to gather feedback and adjust the game settings, and we kept it in the main experiments for consistency. In the analyses reported in this paper, we only use responses from participants in AI conditions regarding whether they reported using AI assistance, comparing these with actual usage logs to identify inconsistent behavior as a filtering condition. No other survey responses were used in our analyses.

The survey content is shown below.

Questions with 5-point Likert scale response options: *Strongly Disagree*, *Disagree*, *Neutral*, *Agree*, and *Strongly Agree*:

- I found the game to be boring.
- I felt confident in my performance throughout the game.
- I put in considerable effort to achieve my level of performance.
- I developed a strategy to improve my performance during the game.
- I find seeing the correct solution for each problem helps me learn.
- I've had prior experience with a similar task.

Two optional free-response questions:

- Did you notice any hidden rules during the game? If so, please describe them. Additionally, please share the strategy you used while solving the puzzles.
- Please share your overall experience with the game and any suggestions for improvement.

For participants in AI assistance conditions, we further asked:

- Rate the AI assistant's accuracy (0–100, or select Not Applicable).
- The AI assistant was helpful during the experiment (5-point Likert scale, with an additional choice *Not Applicable*).
- When did you typically use the AI assistant during the experiment? For example: Right after seeing the problem, after attempting a submission, to check your answer, etc. (optional free response question)