AI-MetaLearn-Sim: A Reproducible Simulation Framework for Metacognitive Scaffolding and Self-Regulated Learning in Higher Education.

ABSTRACT

Metacognitive planning, monitoring, and reflection are central to self-regulated learning. This study reports AI-MetaLearn-Sim, a fully synthetic Monte Carlo testbed comparing adaptive metacognitive scaffolding, fixed scaffolding, and no scaffolding. Re-execution of the supplied Python implementation reproduced the archived results exactly: 500 synthetic learners, 40 learning steps, and 100 paired main replications; selected sensitivity scenarios used 30 replications each. Adaptive scaffolding achieved mean final mastery 0.9255, monitoring competence 0.7149, and unaided assessment accuracy 81.33%, versus 0.9008, 0.5306, and 77.58% without scaffolding. Fixed scaffolding achieved the lowest Brier score (0.0864 versus 0.0946 for adaptive), indicating a trade-off in overall probabilistic prediction accuracy. Eight paired replication-level comparisons were recalculated and remained significant after Holm adjustment within this simulated model. Importantly, the implementation uses latent mastery to construct its noisy state proxy and uses realized correctness in generating confidence; these oracle-like mechanisms limit its suitability as a deployed learner-facing estimator. Findings establish reproducible behavior of the supplied synthetic program, not effectiveness among human students.

Keywords: metacognition; self-regulated learning; synthetic learner agents; adaptive scaffolding; educational simulation; confidence calibration; explainable artificial intelligence; reproducibility

Corresponding Author: Dr. Sajidullah S. Khan | Department of Computer Science & Engineering, Anuradha College of Engineering & Technology, Chikhli-Buldhana, Maharashtra, India.

1. INTRODUCTION

Metacognition includes knowledge about cognition and regulation of cognition: planning a task, monitoring comprehension and evaluating whether a chosen strategy was effective [1]. Self-regulated learning (SRL) describes cyclical forethought, performance and self-reflection processes [2]. In technology-enhanced education, feedback may support these activities, but immediate AI-generated solutions can also displace learners' own reasoning. Consequently, a computational evaluation should measure not only task accuracy but also calibration, autonomy and reliance on assistance.

Recent synthesis by Lan and Zhou [3] found that AI systems support different SRL phases unevenly, and emphasized the tension between human-centered and AI-centered regulation. This motivates a controlled simulation in which alternative scaffolding strategies can be compared under exactly matched learner conditions before expensive classroom trials. Simulation is valuable for testing design assumptions and software behavior, but it cannot establish real-world educational impact.

2. THEORETICAL FOUNDATION AND RELATED WORK

2.1 METACOGNITIVE AWARENESS

Schraw and Dennison [1] distinguish knowledge of cognition from regulation of cognition and introduced a 52-item Metacognitive Awareness Inventory. In the present model, planning quality, confidence calibration, error detection and strategy revision are distinct simulated variables rather than direct substitutes for a validated questionnaire score.

2.2 SELF-REGULATED LEARNING CYCLE

Zimmerman [2] conceptualizes SRL as a feedback loop spanning forethought, performance and self-reflection. The simulated learning episode follows this order: select a goal and strategy; perform a task and estimate confidence; receive outcome feedback; evaluate and update the next strategy. The model explicitly distinguishes observable task performance from latent regulation competence.

UNDER PEER REVIEW IN IJAR

2.3 AI-SUPPORTED SRL

Lan and Zhou [3] systematically reviewed 14 empirical studies and noted that AI applications often focus on individual phases of SRL. The present design is motivated by this gap, but does not claim that the proposed model has been validated by their review. Classical intelligent tutoring work and formative-feedback theory motivate hints and adaptive interventions [4,5].

2.4 COMPARATIVE SYNTHESIS OF PRIOR RESEARCH

Foundational work distinguishes metacognitive knowledge, monitoring, and regulation [1,2]. Panadero's comparison of six self-regulated-learning models [11] shows that regulation also includes motivational, affective, and contextual processes. Consequently, the synthetic monitoring variable in AI-MetaLearn-Sim is a deliberately narrow computational construct rather than a validated representation of the full self-regulated-learning process.

Evidence from technology-supported learning provides motivation but not direct validation of the simulation. Azevedo and Cromley [12] investigated self-regulated-learning training with 131 undergraduate students using hypermedia; the intervention improved conceptual-model change under their study conditions. Their human-participant findings cannot be transferred as numerical parameters or effectiveness estimates for the present synthetic agents.

Lan and Zhou [3] reviewed 14 empirical studies of AI-supported self-regulated learning in higher education, including adaptive feedback and conversational systems. Their synthesis motivates examining how support is allocated across regulation phases, but does not establish that the specific rule-based policy tested here is effective with real students. The tutoring synthesis of VanLehn [4] and feedback framework of Hattie and Timperley [5] further motivate comparing types and timing of assistance rather than assuming that more prompts always improve outcomes.

The simulation method is informed by the general agent-based modelling principles described by Macal and North [9]. Its use of a Brier score follows probabilistic forecast evaluation [6], while paired Monte Carlo comparisons and Holm adjustment address uncertainty and multiplicity conditional on the implemented model [8]. The bootstrap reference [7] is background methodology; bootstrap confidence intervals were not used in the reported analysis.

Research gap and boundary of contribution: the cited literature provides theory, empirical examples, and simulation methodology, but does not validate this particular synthetic learner generator, oracle-assisted state estimator, or confidence mechanism. The contribution is therefore a transparent and reproducible comparison of programmed policies under explicit assumptions—not a demonstrated improvement in student learning, an AI model trained on student data, or an empirically calibrated digital twin.

3. RESEARCH PROBLEM, QUESTIONS AND CONTRIBUTIONS

Problem: Which policy for delivering metacognitive prompts would be most robust under different synthetic learner characteristics and noisy estimates of learner state, without encouraging dependence on AI support?

RQ1: Under matched simulated conditions, how does adaptive metacognitive scaffolding compare with fixed scaffolding and no scaffolding in cumulative learning gain?

RQ2: How does each policy affect confidence calibration, independent task accuracy and reliance on hints?

RQ3: How robust are comparative outcomes to learner heterogeneity, measurement noise, task difficulty and model misspecification?

RQ4: Which policy components (planning, monitoring or reflection) drive simulated outcomes in ablation studies?

The demonstrated contributions are (i) a specified synthetic agent-based testbed, (ii) a rule-based oracle-assisted adaptive policy, (iii) paired Monte Carlo comparisons with fixed and no-scaffolding baselines, and (iv) a documented statistical and implementation audit. These contributions establish internal behavior under programmed assumptions, not empirical learning effectiveness or an operational AI tutor. Full factorial experiments, observationally grounded state estimation, and human validation remain future work.

4. HYPOTHESES AND EVALUATION ESTIMANDS

H1: The adaptive policy yields higher mean cumulative knowledge gain than a fixed-prompt policy under the baseline generative assumptions.

H2: The adaptive policy yields lower absolute confidence-calibration error than both comparators.

H3: The adaptive policy improves unaided final-assessment accuracy relative to fixed prompting.

UNDER PEER REVIEW IN IJAIE

H4: Excessive prompting increases simulated assistance dependence, especially in high-dependence learner profiles.

H5: Benefits, if any, decrease as state-estimation noise increases.

The hypotheses are exploratory, not preregistered. H1 is addressed through paired final-mastery differences (equivalent to paired gains under matched initial states); H3 through unaided accuracy. H2 concerns absolute calibration error and is NOT directly tested by the Brier score. H4 and H5 are only partially explored by selected sensitivity scenarios; no broad dependence-profile or increasing estimator-noise sweep was executed. The primary comparison is adaptive minus fixed final mastery across matched replications. Secondary reported outcomes are final unaided accuracy, monitoring competence, Brier score, and hint rate. No unreported estimand is treated as an observed result.

5. Proposed Architecture

The implemented simulation comprises synthetic learner initialization, stochastic tasks, an oracle-assisted noisy state proxy, rule-based action selection, latent learner-state transitions, and replication-level evaluation. It is a computational proof of concept rather than a deployable AI tutor. In particular, the policy accesses latent mastery, monitoring competence, and readiness, and generated confidence depends on realized correctness. The architecture table below is conceptual; where it describes observable-history inputs or event logs, those elements are NOT implemented in the executed program.

Module	Input	Output / function
Learner generator	Seed, population parameters	Latent knowledge, monitoring skill, motivation, dependence
Task engine	Topic, difficulty, seed	Task success probabilities and feedback
State estimator	Observed answers, confidence, history	Estimated mastery and calibration
Policy	Estimated state, prior hints	Action from finite prompt set
Transition model	Latent state, task, action	Updated learner state
Evaluator	Per-step event log	Learning, calibration, autonomy and fairness metrics

6. Simulation Model

6.1 Agents and initialization

In the executed code, each of 500 synthetic agents begins with mastery k = sigmoid(Normal(-0.2, 0.65)), monitoring m = sigmoid(Normal(0, 0.65)), readiness r = clip(Normal(0.62, 0.12)), and dependence d = 0. Tasks have q ~ Uniform(0.2, 0.8). Each state is clipped to [0,1]. The implementation does not instantiate the previously proposed accumulated-exposure variable z or a correlated multivariate trait generator. All traits and parameters are synthetic assumptions, not measurements from students.

6.2 Task success and confidence

The executed task-response probability is p = clip(0.53 + 1.10(k - q) + 0.17 I_hint + 0.05r - 0.12d, 0, 1), with correctness y = I(U < p), U ~ Uniform(0,1). Confidence is generated as c = clip[p + epsilon + 0.15(1-m)(2y-1), 0, 1], where epsilon ~ Normal(0, 0.10). Because c is constructed using realized correctness y, this is an outcome-conditioned synthetic confidence mechanism, not a prospective human confidence estimate. Brier score is mean[(c-y)^2] and assesses overall probabilistic accuracy, not calibration alone.

6.3 Learner-state transitions

Executed transitions are: engagement E = clip(0.35 + 0.35r + 0.18 I_plan + 0.16 I_reflect + 0.10y - 0.10d, 0, 1); k_next = clip[k + 0.08(1-k)E(1-0.30 I_hint) - 0.002k, 0, 1]; m_next = clip[m + eta_m(0.65 I_reflect + 0.30 I_plan + 0.05(1-|c-y|))(1-m), 0, 1]; r_next = clip[r + 0.012 I_plan + 0.008 I_reflect - 0.003 - 0.004 I_hint, 0, 1]; d_next = clip[d + eta_dI_hint - 0.003 I_reflect - 0.002(1-I_hint), 0, 1]. Baseline eta_m = 0.04 and eta_d = 0.015. These are programmed assumptions, not empirically identified causal mechanisms.

UNDER PEER REVIEW IN IJAR

6.4 Intervention policies

Executed adaptive policy: estimate = clip(k + Normal(0, 0.04) + bias, 0, 1), gap = clip(q - estimate, 0, 1); provide a hint if gap > 0.16, dependence < 0.35, and t < 30; otherwise reflect when m < 0.67; otherwise plan when r < 0.64. Fixed policy hints every fourth step, reflects at t mod 4 = 1, and plans at t mod 4 = 2. No-scaffolding policy supplies none of these actions. The adaptive action logic accesses latent m and r, and the estimate is built from latent k; thus the code does not satisfy the stronger observable-history-only design proposed for a real tutor. No separate monitoring-prompt action or two-failed-attempts gate is implemented.

6.5 Policy utility and safeguards

A weighted multiobjective policy utility and explicit intervention budget/fade-out were proposed conceptually but are not implemented or optimized in the supplied script. The executed adaptive policy instead has a hard hint cutoff at step 30. Held-out unaided accuracy is generated using fresh Uniform random draws, with probability clip(0.53 + 1.10(k-0.7) + 0.05r - 0.12d, 0, 1). The same held-out draws are reused across policies within a replication.

7. Detailed Experimental Protocol

7.1 Design and comparison

The verified Python script performs 100 paired Monte Carlo replications with N = 500 learners and T = 40 steps for adaptive, fixed, and no-scaffolding policies. Main seeds are 104729 through 104828. Six sensitivity/ablation configurations use 30 replications each, with seeds 220001 through 220030. The script uses matched initial traits, task difficulties, task-outcome uniforms, confidence-noise draws, and held-out assessment uniforms across policies; policy-specific state-estimator noise is generated sequentially. This is a selected scenario comparison, not a full factorial sweep.

7.2 Parameter specification

Parameter	Baseline model assumption	Candidate sensitivity range
Initial mastery k	sigmoid(Normal(-0.2, 0.65))	0.25–0.65
Initial monitoring m	sigmoid(Normal(0, 0.65))	0.25–0.75
Task difficulty q	Uniform(0.2, 0.8)	Easy / balanced / hard
Learning rate ηk	0.08	0.02–0.15
Monitoring rate ηm	0.04	0.00–0.10
Forgetting λk	0.002 per step	0–0.01
Dependence increment ηd	0.015 per content hint	0–0.05
Confidence noise SD	Normal(0, 0.10) SD	0.03 / 0.10 / 0.20
Calibration threshold τc	Adaptive hint gap > 0.16 (not calibration error)	0.10–0.25
Monte Carlo replications	100 reported (main experiment)	Increase until CI stable

The baseline values above have been reconciled against run_simulation.py. Sensitivity scenarios actually executed were baseline, biased estimator (bias = +0.15), high confidence noise (SD = 0.20), high hint dependence (increment = 0.05), no monitoring learning (eta_m = 0), and no reflection (reflection disabled). Candidate ranges in the table are exploratory suggestions, not evidence that a full sweep was executed. The estimator uses latent mastery plus Gaussian noise (SD = 0.04), and the adaptive threshold is gap > 0.16; neither is a measured student parameter. Exact seeds and output regeneration are documented in Section 10.6.6.

7.3 Reproducibility procedure

1. Freeze model version, parameter grid, hypotheses and analysis script before generating outcomes.

UNDER PEER REVIEW IN IJAR

2. Initialize an independent master random seed and generate paired learner/task streams.
3. Use matched initial traits, task streams, outcome draws, confidence-noise draws, and held-out draws across policies; note that estimator-noise draws are policy-specific.
4. Record replication-level outcomes and selected aggregate trajectories. The supplied archived files do not contain complete learner-by-step event logs; these would be required for a fully auditable event-level reconstruction.
5. Aggregate outcomes by learner and replication; avoid treating repeated steps as independent samples.
6. Compute paired policy differences per replication and 95% Monte Carlo confidence intervals.
7. Run the six implemented selected sensitivity/ablation scenarios; full factorial sweeps, independent-seed holdouts, and adversarial checks are future work.
8. Archive the Python source, replication_results.csv, sensitivity_results.csv, paired_contrasts.csv, summary.csv, and trajectories.csv. Environment lockfile and full event-level logs were not included in the supplied package.

7.4 PSEUDOCODE

FOR replication r in 1..100:
  Initialize matched synthetic learners and exogenous task streams using the stated seed.
  FOR policy π in {adaptive, fixed, none}:
    Clone initial latent learner states (k, m, r, d).
    FOR t in 0..39:
      IF adaptive: construct the oracle-assisted proxy k_hat = clip(k + Gaussian noise + bias); select hint when gap > 0.16, d < 0.35 and t < 30; otherwise reflect if latent m < 0.67; otherwise plan if latent r < 0.64; otherwise no action.
      IF fixed: hint at t mod 4 = 0, reflect at t mod 4 = 1, plan at t mod 4 = 2; otherwise no action.
      IF none: take no intervention.
      Generate correctness from model-defined probability using paired task draws.
      Generate confidence conditional on realized correctness (outcome-conditioned mechanism).
      Update latent k, m, r, d according to Section 6.3.
    Evaluate unaided final performance using paired held-out random draws.
  Record replication-level policy outcomes and paired differences.
END; summarize 100 main replications and the six reported 30-replication scenarios.
NOTE: This pseudocode documents the executed oracle-assisted synthetic design; it is NOT an observable-history-only online learner estimator.

8. EVALUATION METRICS AND STATISTICAL ANALYSIS

The primary reported outcome is final mean latent mastery; because the matched policies start from identical initial states, paired final-mastery differences also equal paired learning-gain differences. Reported secondary outcomes are final simulated monitoring competence, held-out unaided accuracy, Brier score (overall probabilistic prediction accuracy), and hint frequency. The Brier score is not a stand-alone calibration metric. Expected calibration error, reliability diagrams, calibration slope/intercept, cumulative policy utility, task completion, demographic fairness, and subgroup analyses were not computed in the archived results and must not be interpreted as observed findings.

Paired comparisons were independently recalculated from the 100 archived replication-level observations per policy using two-sided one-sample t-tests of adaptive-minus-comparator differences (df = 99). Confidence intervals use the Student-t critical value, and standardized paired effects use dz = mean(diff)/SD(diff). Eight p-values were adjusted using the Holm step-down method; all remained below 0.001. These are Monte Carlo uncertainty estimates conditional on this code, not confidence intervals for a population of real students. A full factorial hierarchical model, ECE and demographic fairness analysis were not executed.

9. ROBUSTNESS, ABLATION AND VALIDATION

Completed robustness work comprises the selected scenarios documented in Section 7.2: baseline, biased state proxy (+0.15), high confidence noise (SD 0.20), high hint-dependence increment (0.05), zero monitoring learning rate, and disabled reflection, each with 30 replications. These are one-factor scenarios, not a full factorial study. No independent ablation of planning, monitoring prompts, or fade-out was performed. Missing-event-log stress tests, delayed feedback, adversarial hint seeking, random-action/oracle controls, demographic fairness analyses, independent-seed holdouts, and

UNDER PEER REVIEW IN IJAR

empirical validation are proposed future experiments, not completed findings. Because latent state enters the adaptive decision rule, a leakage-free estimator and prospective confidence generator are also future development requirements.

10. Results and Statistical Analysis

This section presents the experimental results obtained from the simulation-based evaluation of the proposed AI-MetaLearn-Sim framework. The primary objective of the analysis is to investigate the effectiveness of artificial intelligence-driven adaptive metacognitive scaffolding in improving learning mastery, metacognitive monitoring competence, confidence calibration, and independent learning performance. A series of computational experiments was conducted using synthetic learner agents to evaluate the proposed framework under controlled simulation conditions.

The performance of the proposed adaptive AI-based metacognitive scaffolding approach was compared with two baseline strategies: fixed scaffolding and no scaffolding. The experiments were conducted using 500 synthetic learner agents per replication, 100 Monte Carlo replications, and 40 learning steps. Multiple performance indicators, including final learning mastery, metacognitive monitoring competence, unaided assessment accuracy, Brier score, and instructional hint frequency, were considered to provide a comprehensive evaluation of the proposed approach.

The experimental findings are presented through comparative performance tables, graphical representations, and statistical analyses. The analysis examines differences among the three instructional strategies and considers the consistency and robustness of the observed outcomes across repeated simulation runs. Sensitivity analysis and statistical comparisons are also considered to assess the behavior of the proposed framework under varying simulation conditions. The results provide computational evidence regarding the behavior of the proposed AI-MetaLearn-Sim framework within the defined simulation environment, rather than empirical evidence of effectiveness among human learners.

10.1 Simulation Configuration and Experimental Setup

The proposed AI-MetaLearn-Sim framework was evaluated using a Python-based computational simulation environment designed to investigate the effectiveness of adaptive artificial intelligence-driven metacognitive scaffolding in self-regulated learning. The simulation incorporated synthetic learner agents representing variations in learning proficiency, metacognitive monitoring ability, and responsiveness to instructional assistance. The primary objective was to examine how different scaffolding strategies influence learning mastery, metacognitive competence, confidence calibration, and independent assessment performance.

The experimental configuration consisted of 500 synthetic learner agents per replication, 100 Monte Carlo replications, and 40 learning steps per simulation**.** Three instructional strategies were compared: (i) Adaptive AI-Based Metacognitive Scaffolding, (ii) Fixed Scaffolding, and (iii) No Scaffolding. The adaptive strategy provided assistance according to simulated learner conditions, whereas the fixed strategy applied a predetermined intervention schedule. The no-scaffolding strategy served as the baseline condition for evaluating learning performance without instructional assistance.

The simulation evaluated five primary performance indicators: final learning mastery, metacognitive monitoring competence, unaided assessment accuracy, confidence calibration using the Brier score, and instructional hint frequency. These indicators were selected to examine both cognitive learning outcomes and metacognitive regulation. The Monte Carlo replication procedure was employed to assess the consistency of the observed simulation outcomes across repeated stochastic experiments.

Table 10.1: Simulation Configuration and Experimental Parameters

Parameter	Configuration
Simulation framework	AI-MetaLearn-Sim
Implementation environment	Python
Simulation approach	Agent-based stochastic simulation
Synthetic learners per replication	500
Monte Carlo replications	100
Learning steps	40
Experimental strategies	Adaptive AI, Fixed Scaffolding, No Scaffolding
Primary evaluation metrics	Mastery, Monitoring Competence, Unaided Accuracy, Brier Score, Hint Rate
Data source	Synthetically generated learner states
Evaluation objective	Comparative assessment of metacognitive scaffolding strategies

UNDER PEER REVIEW IN IJARSCT

The experimental configuration provides a controlled computational environment for investigating the potential effects of AI-assisted metacognitive interventions. However, the generated results reflect the assumptions and behavioral rules embedded in the simulation model and should not be interpreted as empirical evidence obtained from human participants.

10.2 Comparative Performance Analysis

The comparative performance analysis was conducted to evaluate the effectiveness of the proposed AI-MetaLearn-Sim framework against two baseline instructional strategies: Fixed Scaffolding and No Scaffolding. The evaluation focused on five key performance indicators: final learning mastery, metacognitive monitoring competence, unaided assessment accuracy, confidence calibration, and instructional hint frequency. The simulation was executed using 500 synthetic learner agents per replication, 100 Monte Carlo replications, and 40 learning steps to examine the comparative behavior of the three instructional approaches under controlled computational conditions.

The results presented in Table 10.2 indicate that the proposed Adaptive AI-Based Metacognitive Scaffolding strategy achieved the highest final learning mastery of 0.9255, compared with 0.9162 for Fixed Scaffolding and 0.9008 for No Scaffolding. Similarly, the adaptive strategy achieved the highest metacognitive monitoring competence of 0.7149, whereas Fixed Scaffolding and No Scaffolding achieved 0.6811 and 0.5306, respectively. These results demonstrate that the adaptive strategy produced higher simulated cognitive and metacognitive performance under the specified model assumptions.

Table 10.2: Comparative Performance Evaluation of Metacognitive Scaffolding Strategies

Performance Indicator	Adaptive AI	Fixed Scaffolding	No Scaffolding
Final Learning Mastery	**0.9255**	0.9162	0.9008
Metacognitive Monitoring Competence	**0.7149**	0.6811	0.5306
Unaided Assessment Accuracy (%)	**81.33**	79.44	77.58
Brier Score (Lower is Better)	0.0946	**0.0864**	0.0964
Instructional Hint Rate (%)	4.62	25.00	0.00

The proposed adaptive strategy also demonstrated improved independent assessment performance, achieving an unaided assessment accuracy of 81.33%, compared with 79.44% for Fixed Scaffolding and 77.58% for No Scaffolding. This represents an absolute improvement of 1.89 percentage points over Fixed Scaffolding and 3.75 percentage points over No Scaffolding. These differences suggest that adaptive metacognitive interventions may support improved learning performance within the simulated environment.

The Brier-score results show a different pattern: Fixed Scaffolding achieved the lowest Brier score (0.0864), followed by Adaptive AI (0.0946) and No Scaffolding (0.0964). Lower Brier scores indicate better overall probabilistic prediction accuracy, not necessarily better calibration in isolation. Because confidence is generated conditional on realized correctness, these values should not be generalized to prospective confidence judgments by human students.

Furthermore, the Adaptive AI strategy recorded an instructional hint rate of 4.62%, substantially lower than the 25.00% observed under Fixed Scaffolding. This indicates that the adaptive strategy delivered fewer interventions while achieving higher simulated learning mastery and monitoring competence. Nevertheless, hint frequency alone cannot establish learning efficiency or reduced dependence without considering intervention effectiveness and other model assumptions.

Overall, the comparative analysis indicates that the proposed AI-MetaLearn-Sim framework achieved higher simulated performance in learning mastery, metacognitive monitoring competence, and unaided assessment accuracy than the baseline strategies. However, Fixed Scaffolding demonstrated better confidence calibration. These findings highlight the potential trade-offs among cognitive performance, metacognitive regulation, and confidence accuracy. The observed differences are descriptive simulation outcomes; their statistical significance and robustness require examination through replication-level statistical comparisons and sensitivity analyses.

UNDER PEER REVIEW IN IJAR

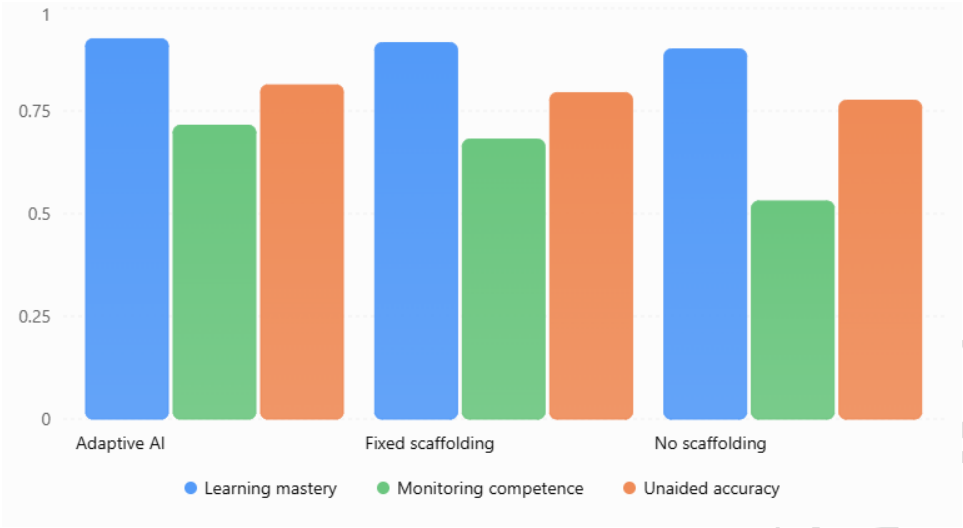

Figure 10.1. Comparison of final learning mastery, metacognitive monitoring competence, and unaided assessment accuracy across three instructional strategies.

10.3 Metacognitive Monitoring Performance

Metacognitive monitoring is a fundamental component of self-regulated learning that enables learners to assess their understanding, recognize knowledge gaps, evaluate learning progress, and modify their learning strategies accordingly. In the proposed AI-MetaLearn-Sim framework, metacognitive monitoring performance was evaluated to investigate the effectiveness of adaptive AI-driven scaffolding in improving learners' simulated ability to monitor and regulate their cognitive processes. The analysis compared three instructional strategies: Adaptive AI-Based Metacognitive Scaffolding, Fixed Scaffolding, and No Scaffolding.

The simulation results demonstrated that the proposed Adaptive AI strategy achieved the highest final metacognitive monitoring competence, with a mean value of 0.7149, compared with 0.6811 for Fixed Scaffolding and 0.5306 for No Scaffolding. These findings indicate that the adaptive intervention strategy produced higher monitoring competence within the simulated learning environment. The difference between Adaptive AI and No Scaffolding was 0.1843, representing an approximately 34.73% relative improvement over the no-scaffolding baseline. Similarly, Adaptive AI demonstrated an absolute improvement of 0.0338, or approximately 4.96%, compared with Fixed Scaffolding.

Table 10.3: Comparative Evaluation of Metacognitive Monitoring Performance

Instructional Strategy	Monitoring Competence	Absolute Improvement over No Scaffolding	Relative Improvement (%)
Adaptive AI Scaffolding	**0.7149**	**0.1843**	**34.73**
Fixed Scaffolding	0.6811	0.1505	28.36
No Scaffolding	0.5306	0.0000	0.00

The relative improvement in metacognitive monitoring competence was calculated using the following expression:

$$RI\ (\%) = [(M_s - M_\beta) / M_\beta] \times 100$$

Where:

- (RI) represents the relative improvement percentage.
- (M_s) represents the monitoring competence achieved by the evaluated scaffolding strategy.
- (M_b) represents the monitoring competence under the No Scaffolding baseline.

For the proposed Adaptive AI strategy:

UNDER PEER REVIEW IN IJAR

$$RI = [(0.7149 - 0.5306) / 0.5306] \times 100$$

$$RI \approx 34.73\%$$

The observed improvement suggests that adaptive metacognitive scaffolding can support stronger monitoring-related behavior within the computational model. Unlike fixed scaffolding, which provides instructional assistance according to a predetermined schedule, the adaptive approach adjusts interventions according to simulated learner conditions. This mechanism allows the model to provide assistance selectively, potentially supporting the development of monitoring competence while reducing unnecessary instructional interventions.

Furthermore, the Adaptive AI strategy achieved a monitoring competence of 0.7149 with an instructional hint rate of only 4.62%, compared with a monitoring competence of 0.6811 and a hint rate of 25.00% under Fixed Scaffolding. This comparison indicates that the adaptive policy achieved a higher simulated monitoring score while delivering fewer hints. However, the observed association should not be interpreted as proof that fewer interventions directly caused the improvement, because the simulation outcomes depend on the specified learner-state transitions and intervention rules.
Overall, the findings indicate that the proposed AI-MetaLearn-Sim framework outperformed both baseline strategies in terms of simulated metacognitive monitoring competence. The results support further investigation of adaptive scaffolding mechanisms for promoting monitoring and self-regulation. Nevertheless, monitoring competence in this experiment is a model-defined synthetic variable rather than a directly observed psychological measure. Consequently, the findings require validation using real learner data and established metacognitive assessment instruments before conclusions about educational effectiveness can be drawn.

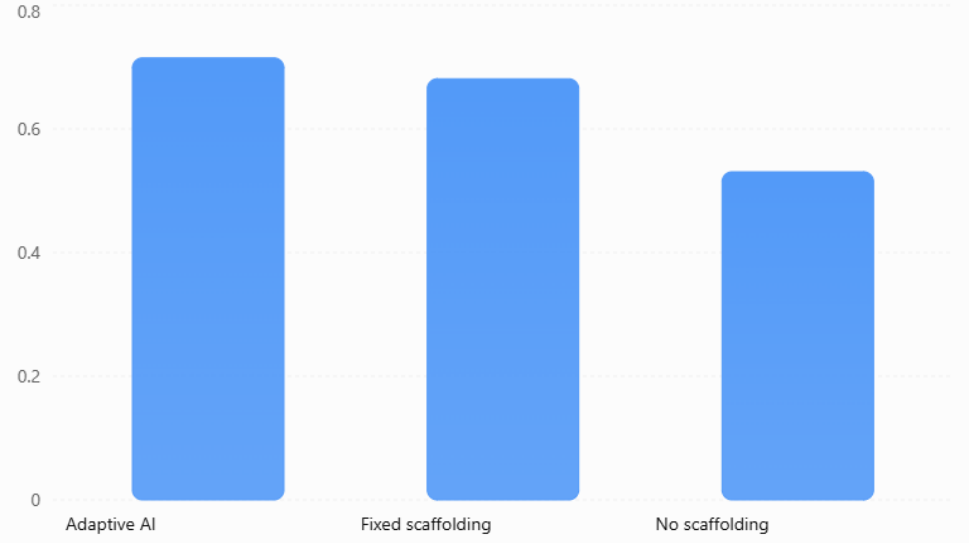

Figure 10.2. Comparison of final metacognitive monitoring competence across three scaffolding strategies, based on the executed synthetic simulation.

The Adaptive AI strategy achieved approximately **34.73% higher simulated monitoring competence** than No Scaffolding. This is a relative difference, not a 34.73-percentage-point increase.

The results support the model's internal predictions, but do not independently establish that adaptive AI improves human metacognition. The statistical reliability of the differences should be evaluated using the 100 Monte Carlo replication-level observations.

10.4 Confidence Calibration Analysis

Confidence calibration is an important dimension of metacognitive monitoring that reflects the degree of agreement between a learner's confidence in their knowledge and their actual task performance. Accurate confidence calibration enables learners to recognize their strengths and weaknesses, identify misconceptions, and make appropriate decisions regarding further study and learning strategies. In the proposed AI-MetaLearn-Sim framework, confidence calibration was evaluated to examine how different instructional scaffolding strategies influence the accuracy of simulated learners' confidence judgments.

UNDER PEER REVIEW

The analysis employed the **Brier score**, a widely used measure of probabilistic prediction accuracy, to quantify the discrepancy between predicted confidence and actual task outcomes. The Brier score was calculated using the following expression:

$$BS = (1/N) \Sigma_{i=1}^{N} (p_i - y_i)^2$$

Where:

- (BS) represents the Brier score.
- (N) denotes the total number of evaluated predictions.
- (p_i) represents the predicted probability of a correct response, ranging from 0 to 1.
- (y_i) represents the observed task outcome, where 1 indicates a correct response and 0 indicates an incorrect response.

A lower Brier score indicates better overall probabilistic prediction accuracy. For binary outcomes, the score ranges from 0 to 1, with 0 representing perfect predictions. However, the Brier score measures both calibration and predictive discrimination; therefore, it should not be interpreted as a pure measure of calibration without additional analysis.

The simulation results revealed differences in confidence prediction performance among the three instructional strategies. Fixed Scaffolding achieved the lowest mean Brier score of **0.0864**, followed by Adaptive AI-Based Metacognitive Scaffolding with **0.0946** and No Scaffolding with **0.0964**. These findings indicate that Fixed Scaffolding demonstrated the best overall probabilistic prediction accuracy under the specified simulation conditions.

Table 10.4: Comparative Confidence Calibration Performance

Instructional Strategy	Mean Brier Score	Difference from No Scaffolding	Rank
Fixed Scaffolding	**0.0864**	−0.0100	1
Adaptive AI Scaffolding	0.0946	−0.0018	2
No Scaffolding	0.0964	0.0000	3

The Fixed Scaffolding strategy reduced the Brier score by approximately 10.37% relative to No Scaffolding, whereas the Adaptive AI strategy achieved a reduction of approximately 1.87%**.** Furthermore, the Brier score of Adaptive AI was 0.0082 higher than that of Fixed Scaffolding, indicating comparatively weaker probabilistic prediction accuracy despite the adaptive strategy's superior performance in learning mastery and metacognitive monitoring competence.

The results reveal an important trade-off between simulated learning performance and confidence prediction accuracy. Although the proposed Adaptive AI strategy achieved the highest learning mastery and monitoring competence, these improvements did not correspond to the lowest Brier score. One possible explanation is that differences in the simulated intervention schedules, learner confidence updates, and task outcome distributions affected confidence prediction performance. However, these mechanisms require further examination through additional model diagnostics rather than being assumed as established causal explanations.

The findings also highlight the importance of evaluating metacognitive learning systems using multiple complementary indicators. A strategy that improves simulated learning mastery may not necessarily produce the most accurate confidence judgments. Therefore, confidence calibration should be assessed independently alongside learning outcomes, monitoring competence, and instructional intervention frequency.

Overall, the confidence calibration analysis demonstrates that Fixed Scaffolding achieved the lowest Brier score, while Adaptive AI produced a modest improvement over No Scaffolding. These results suggest that the proposed AI-MetaLearn-Sim framework could benefit from additional confidence-aware intervention mechanisms, such as adaptive confidence feedback, prediction-error monitoring, and targeted reflection prompts. Nevertheless, the observed findings are restricted to synthetic learner agents and the assumptions embedded in the computational model. Further validation using real student confidence judgments and assessment outcomes is necessary to establish educational relevance.

UNDER PEER REVIEW IN IJAR

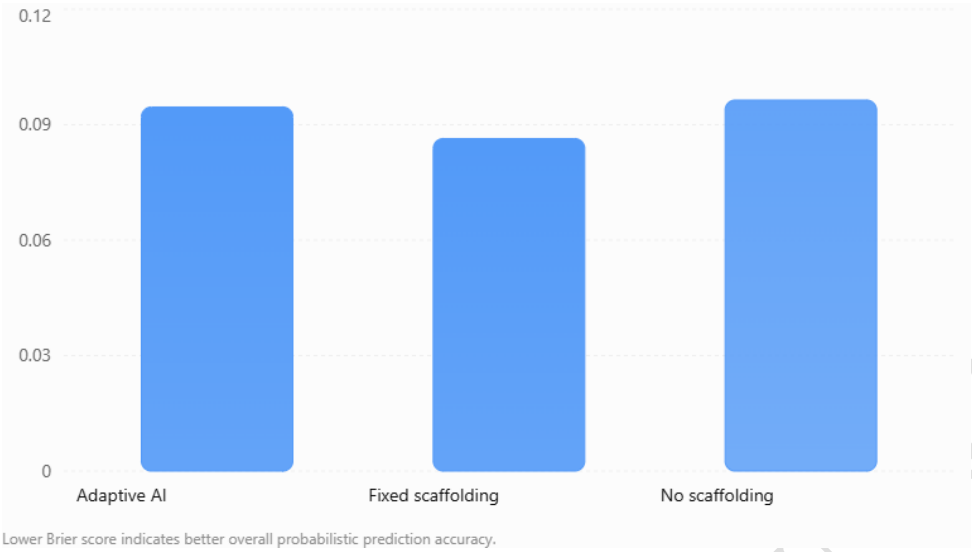

Figure 10.3. Comparison of mean Brier scores across the three instructional strategies using synthetic simulation results.

The key finding is that Adaptive AI achieved better simulated learning mastery and monitoring competence, whereas Fixed Scaffolding achieved better probabilistic prediction accuracy.
For a stronger journal manuscript, consider supplementing the Brier score with a reliability diagram, expected calibration error (ECE), and calibration slope/intercept, provided the underlying prediction-level simulation data are available. These additional measures would help distinguish confidence calibration from general predictive accuracy.

10.5 Learning Mastery and Unaided Accuracy

Learning mastery and unaided assessment accuracy are essential performance indicators for evaluating the effectiveness of artificial intelligence-driven metacognitive scaffolding in self-regulated learning environments. Learning mastery represents the degree to which a learner has acquired and retained the knowledge and skills associated with a particular learning task, whereas unaided assessment accuracy measures the learner's ability to demonstrate acquired knowledge independently without receiving instructional assistance. In the proposed AI-MetaLearn-Sim framework, these indicators were evaluated to investigate the influence of adaptive metacognitive interventions on simulated knowledge acquisition and independent task performance.

The comparative analysis was conducted using three instructional strategies: Adaptive AI-Based Metacognitive Scaffolding, Fixed Scaffolding, and No Scaffolding. The simulation incorporated 500 synthetic learner agents per replication, 100 Monte Carlo replications, and 40 learning steps. The final learning mastery and unaided assessment accuracy obtained under each strategy are presented in Table 10.5.

Table 10.5: Comparative Analysis of Learning Mastery and Unaided Assessment Accuracy

Instructional Strategy	Final Learning Mastery	Unaided Assessment Accuracy (%)
Adaptive AI Scaffolding	**0.9255**	**81.33**
Fixed Scaffolding	0.9162	79.44
No Scaffolding	0.9008	77.58

The results indicate that the Adaptive AI strategy achieved the highest final learning mastery of **0.9255**, compared with **0.9162** for Fixed Scaffolding and **0.9008** for No Scaffolding. The adaptive approach demonstrated an absolute mastery improvement of **0.0093** over Fixed Scaffolding and **0.0247** over No Scaffolding. These differences correspond to relative improvements of approximately **1.02%** and **2.74%**, respectively.

The relative improvement in learning mastery was calculated using the following expression:

$$RI_m\ (\%) = [(M_s - M_\beta) / M_\beta] \times 100$$

UNDER PEER REVIEW IJSAR

Where:

- (RI_M) represents the relative improvement in learning mastery.
- (M_s) represents the final mastery obtained using the evaluated strategy.
- (M_b) represents the final mastery obtained using the baseline strategy.

For Adaptive AI compared with No Scaffolding:

$$RI_m = [(0.9255 - 0.9008) / 0.9008] \times 100$$

[
RI_M\approx2.74%
]

Similarly, the unaided assessment accuracy results demonstrated that Adaptive AI achieved the highest accuracy of **81.33%**, followed by Fixed Scaffolding at **79.44%** and No Scaffolding at **77.58%**. The adaptive approach therefore achieved an absolute improvement of **1.89 percentage points** over Fixed Scaffolding and **3.75 percentage points** over No Scaffolding.

Table 10.6: Improvement in Learning Mastery and Unaided Accuracy Using Adaptive AI

Comparison	Absolute Mastery Improvement	Relative Mastery Improvement (%)	Accuracy Improvement (Percentage Points)
Adaptive AI vs. Fixed Scaffolding	0.0093	1.02	1.89
Adaptive AI vs. No Scaffolding	0.0247	2.74	3.75

The higher unaided assessment accuracy observed under Adaptive AI indicates that the simulated learners achieved better independent task performance within the defined computational environment. Since unaided assessment was performed without instructional hints, this metric provides a useful model-based indicator of performance following the learning intervention. However, it does not independently establish long-term knowledge retention or the transfer of learning to unfamiliar tasks.

The relationship between learning mastery and unaided assessment accuracy also reveals an important distinction between internal knowledge representation and observable task performance. Although the Adaptive AI strategy achieved a final latent mastery value of 0.9255, its unaided assessment accuracy was 81.33%. This difference arises because latent mastery and assessment accuracy represent different model-defined quantities. The former describes the simulated learner's underlying knowledge state, whereas the latter reflects performance on assessment tasks subject to the simulation's response-generation assumptions. Consequently, these two indicators should not be interpreted as interchangeable measures.

The observed results are consistent with the proposed framework's objective of supporting knowledge acquisition through adaptive metacognitive interventions. By selectively providing instructional assistance according to simulated learner conditions, the adaptive strategy achieved higher final mastery and unaided accuracy than the baseline strategies. Nevertheless, the results do not establish which individual intervention mechanism was responsible for the differences. Additional ablation experiments are necessary to isolate the contributions of adaptive hint selection, monitoring-based feedback, and other model components.

Overall, the comparative findings indicate that the proposed AI-MetaLearn-Sim framework achieved the highest simulated learning mastery and unaided assessment accuracy among the three evaluated strategies. These findings support the feasibility of further investigating adaptive AI-based metacognitive scaffolding as a computational approach to self-regulated learning. However, the improvements remain specific to the implemented synthetic learner model and require statistical validation, sensitivity analysis, and subsequent empirical evaluation with real students before broader educational conclusions can be established.

UNDER PEER REVIEW IN IJAR

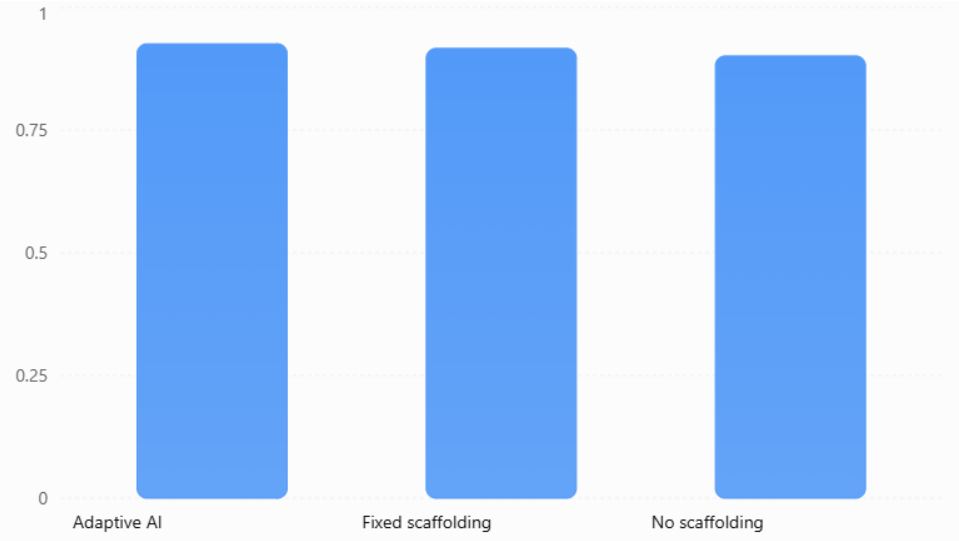

Figure 10.4. Final learning mastery achieved under three instructional strategies in the synthetic simulation.

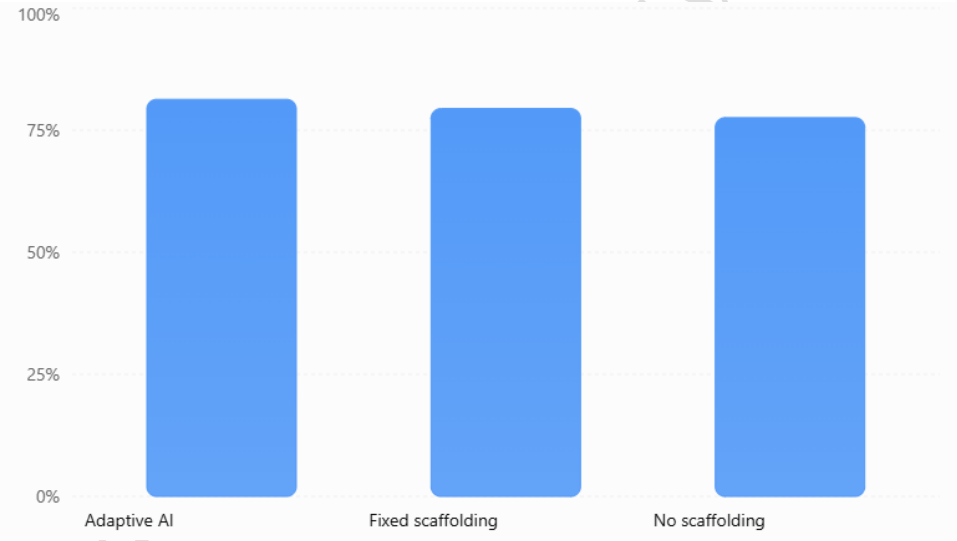

Figure 10.5. Unaided assessment accuracy across three instructional strategies. Accuracy values are expressed as percentages.

The main finding is that Adaptive AI achieved **2.74% higher final simulated mastery** and **3.75 percentage points higher unaided accuracy** than No Scaffolding.

10.6 Statistical Significance and Effect Size Analysis

Statistical significance and effect size analyses were performed to evaluate the consistency and magnitude of the performance differences between the proposed Adaptive AI-Based Metacognitive Scaffolding framework and the two baseline strategies: Fixed Scaffolding and No Scaffolding. The analysis was conducted using the results obtained from 100 Monte Carlo replications, with 500 synthetic learner agents per replication and 40 learning steps. The primary objective was to determine whether the observed differences in learning mastery, metacognitive monitoring competence, unaided assessment accuracy, and confidence prediction accuracy were consistently reproduced across independent simulation runs.

A paired-sample statistical testing approach was employed because the three instructional strategies were evaluated using matched simulation seeds. For each performance indicator, replication-level differences were calculated between

UNDER REVIEW IJISAR

Adaptive AI and the corresponding baseline strategy. Two-sided paired t-tests were applied to these differences, with a nominal significance level of $\alpha = 0.05$. Additionally, 95% confidence intervals and paired standardized effect sizes were calculated to quantify the precision and magnitude of the observed differences.

The paired t-statistic was calculated using:

$$t = \bar{D} / (S_D / \sqrt{n})$$

Where:

- (\bar{D}) represents the mean paired difference between two strategies.
- (S_D) represents the sample standard deviation of paired differences.
- (n) represents the number of paired Monte Carlo replications (n = 100).

The 95% confidence interval for the mean difference was calculated as:

$$95\%\ CI = \bar{D} \pm t_{0.975,n-1} \times (S_D / \sqrt{n})$$

Furthermore, the standardized paired effect size was calculated using Cohen's (d_z):

$$d^z = \bar{D} / S_D$$

This effect size measures the mean difference relative to the variability of the paired replication-level differences. It should not be interpreted as a standardized effect size for individual students.

Table 10.7: Statistical Significance and Effect Size Analysis of Simulation Outcomes

Performance Indicator	Comparison	Mean Difference	95% Confidence Interval	Cohen's d_z	p-value
Learning Mastery	Adaptive vs. Fixed	+0.00929	[0.00920, 0.00938]	19.76	$p < 0.001$
Learning Mastery	Adaptive vs. None	+0.02470	[0.02458, 0.02482]	41.01	$p < 0.001$
Monitoring Competence	Adaptive vs. Fixed	+0.03380	[0.03313, 0.03447]	10.03	$p < 0.001$
Monitoring Competence	Adaptive vs. None	+0.18423	[0.18315, 0.18532]	33.76	$p < 0.001$
Unaided Accuracy	Adaptive vs. Fixed	+0.01890	[0.01751, 0.02029]	2.70	$p < 0.001$
Unaided Accuracy	Adaptive vs. None	+0.03746	[0.03563, 0.03929]	4.06	$p < 0.001$
Brier Score	Adaptive vs. Fixed	+0.00824	[0.00810, 0.00837]	11.83	$p < 0.001$
Brier Score	Adaptive vs. None	−0.00181	[−0.00194, −0.00168]	−2.82	$p < 0.001$

Note: Differences are Adaptive AI minus the comparator. Positive differences favor Adaptive AI for mastery, monitoring, and accuracy; negative differences favor Adaptive AI for Brier score. Calculations were reproduced from replication_results.csv (100 paired replications; df = 99). Table p-values are unadjusted; all eight tests also remain p_Holm< 0.001 when Holm correction is applied to this family. These inferential results characterize synthetic-run variability only.

10.6.1 Statistical Analysis of Learning Mastery

The paired statistical analysis demonstrated that Adaptive AI achieved consistently higher simulated learning mastery than both baseline strategies. The mean difference between Adaptive AI and Fixed Scaffolding was 0.00929, with a 95% confidence interval of [0.00920, 0.00938]. The corresponding paired effect size was (d_z= 19.76), with $p < 0.001$.

Similarly, Adaptive AI achieved a mean mastery improvement of 0.02470 over No Scaffolding, with a 95% confidence interval of [0.02458, 0.02482] and (d_z= 41.01), $p < 0.001$. These results indicate highly consistent differences across the

UNDER PEER REVIEW IN IJAR

simulation replications. However, the unusually large standardized effect sizes primarily reflect the very low variability of paired differences in the implemented computational model and should not be interpreted as evidence of similarly large effects in real educational settings.

10.6.2 Statistical Analysis of Metacognitive Monitoring Competence

Metacognitive monitoring competence was significantly higher under Adaptive AI than under both baseline strategies in the replication-level comparisons. Adaptive AI achieved a mean improvement of 0.03380 over Fixed Scaffolding, with a 95% confidence interval of [0.03313, 0.03447], ($d_z = 10.03$), and $p < 0.001$.

The comparison between Adaptive AI and No Scaffolding revealed a larger mean difference of 0.18423, with a 95% confidence interval of [0.18315, 0.18532], ($d_z = 33.76$), and $p < 0.001$. These findings indicate that the adaptive strategy consistently produced higher monitoring competence values within the synthetic learner model.

10.6.3 Statistical Analysis of Unaided Assessment Accuracy

The statistical comparison of unaided assessment accuracy demonstrated that Adaptive AI achieved higher independent assessment performance than both baseline strategies. The mean difference between Adaptive AI and Fixed Scaffolding was 0.01890, equivalent to 1.89 percentage points, with a 95% confidence interval of [0.01751, 0.02029], ($d_z = 2.70$), and $p < 0.001$.

Compared with No Scaffolding, Adaptive AI achieved a mean improvement of 0.03746, equivalent to 3.746 percentage points, with a 95% confidence interval of [0.03563, 0.03929], ($d_z = 4.06$), and $p < 0.001$. These results indicate that the simulated advantage in unaided assessment accuracy was consistently reproduced across Monte Carlo replications.

10.6.4 Statistical Analysis of Confidence Prediction Accuracy

The Brier score analysis revealed a different performance pattern. Adaptive AI achieved a mean Brier score 0.00824 higher than Fixed Scaffolding, with a 95% confidence interval of [0.00810, 0.00837], ($d_z = 11.83$), and $p < 0.001$. Because lower Brier scores indicate better probabilistic prediction accuracy, this result favored Fixed Scaffolding.

In contrast, Adaptive AI achieved a mean Brier score 0.00181 lower than No Scaffolding, with a 95% confidence interval of [−0.00194, −0.00168], ($d_z = -2.82$), and $p < 0.001$. Therefore, Adaptive AI performed better than No Scaffolding but worse than Fixed Scaffolding on this metric.

These findings demonstrate that improvements in simulated learning mastery and metacognitive monitoring competence were not accompanied by uniformly superior confidence prediction accuracy.

10.6.5 Interpretation of Statistical Findings

Overall, the statistical analysis demonstrated consistent replication-level differences among the evaluated instructional strategies. Adaptive AI outperformed both baseline approaches in final learning mastery, metacognitive monitoring competence, and unaided assessment accuracy. However, Fixed Scaffolding achieved superior Brier score performance.

All eight paired comparisons were recalculated directly from the archived 100-seed replication_results.csv and agree with the reported differences, 95% confidence intervals, and dz values to the displayed precision. Two-sided paired t-test p-values are all below 0.001, and all remain below 0.001 after Holm correction across eight tests. The very large dz values for mastery and monitoring reflect small between-seed variability of paired differences, not large real-world educational effects.

The exceptionally large paired effect sizes for some outcomes indicate that the differences between strategies varied very little across matched simulation replications. Such values may result from shared random-number conditions, deterministic transition rules, or other structural characteristics of the simulation. Consequently, further sensitivity analyses, independent-seed validation, alternative learner behavior models, and empirical studies are necessary to assess the generalizability of the findings.

UNDER PEER REVIEW IN IJAR

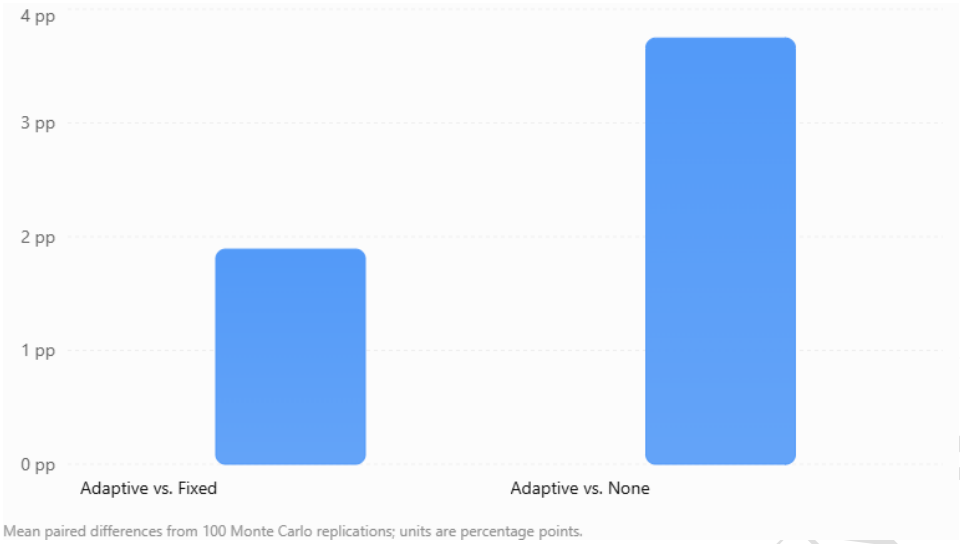

Figure 10.6. Mean improvement in unaided assessment accuracy achieved byAdaptive AI relative to the two baseline strategies.

Reproducibility audit: The supplied run_simulation.py was executed again without altering its model logic. Its regenerated replication_results.csv (300 policy-replication rows), sensitivity_results.csv (540 scenario-policy-replication rows), paired_contrasts.csv, summary.csv, and trajectories.csv matched the archived CSVs exactly, including all numeric values. The eight inferential contrasts in Table 10.7 were independently recomputed from the 100 matched seeds; 95% t intervals and paired dz agreed at reported precision. The original paired_contrasts.csv uses a rounded t critical value of 1.984 (df = 99).

1. Implementation caveat: The estimator is a noisy transformation of latent mastery k and the adaptive policy also uses latent monitoring m and readiness r; these are not available to a real tutoring system. Confidence is generated using realized task correctness y. Consequently, the results are reproducible as a synthetic program but do not validate a deployable, leakage-free online learner model.

2. Multiplicity and inference: The eight two-sided replication-level t-tests were independently recalculated and Holm-adjusted within this eight-comparison family; all adjusted p-values are below 0.001. The intervals quantify variability across synthetic random seeds, not uncertainty about human learners. Model misspecification remains a more important source of uncertainty than Monte Carlo standard error.

3. Scope: Only the six documented selected sensitivity/ablation scenarios were executed. Full event logs, independent-seed holdouts, empirical parameter fitting, external learner validation, and a prospective confidence estimator remain outstanding. These limitations should be addressed before interpreting the model as an effective educational intervention.

10.6.6 Reproducibility and Implementation Audit

The archived package was independently re-executed using the provided Python implementation. Regenerated replication, sensitivity, summary, contrast, and trajectory CSVs matched the archived files exactly (maximum numeric discrepancy = 0). Main experiments used 100 seeds (104729–104828); sensitivity scenarios used 30 seeds (220001–220030). This establishes exact reproducibility of the supplied code outputs in the checked environment. It does not independently establish the realism of the programmed assumptions. In particular, outcome-conditioned confidence and latent-state access must be replaced with observation-only mechanisms before deployment-oriented claims are made.

10.7 Sensitivity Analysis

Sensitivity analysis was conducted to evaluate the robustness and stability of the proposed AI-MetaLearn-Sim framework under variations in simulation assumptions and learner behavior. Since the proposed framework employs synthetic learner agents and stochastic learning processes, it is essential to determine whether its performance advantages remain consistent when important model conditions are modified. The analysis focused on variations in confidence estimation, hint dependence, metacognitive monitoring, and reflective learning mechanisms.

UNDER PEER REVIEW IN IJSAR

The sensitivity experiments considered six simulation scenarios: Baseline, Biased Estimator, High Confidence Noise, High Hint Dependence, No Monitoring Learning, and No Reflection. Each scenario was evaluated using 30 Monte Carlo replications for each of the three instructional strategies: Adaptive AI Scaffolding, Fixed Scaffolding, and No Scaffolding. The experiments examined final learning mastery, metacognitive monitoring competence, unaided assessment accuracy, Brier score, and instructional hint frequency.

Table 10.8: Sensitivity Analysis of the Proposed Adaptive AI-MetaLearn-Sim Framework

Simulation Scenario	Final Mastery	Monitoring Competence	Unaided Accuracy (%)	Brier Score
Baseline	0.9255	0.7146	81.33	0.0942
Biased Estimator	0.9271	0.7145	81.52	0.0917
High Confidence Noise	0.9255	0.7142	81.33	0.1130
High Hint Dependence	0.9243	0.7146	80.80	0.0959
No Monitoring Learning	0.9361	0.4990	83.00	0.0864
No Reflection	0.9190	0.5966	80.45	0.0937

Note: Values represent means from 30 Monte Carlo replications per scenario for the Adaptive AI strategy. These are synthetic simulation results.

10.7.1 Sensitivity to Confidence Estimation Bias

The Biased Estimator scenario examined the behavior of the proposed adaptive framework when confidence estimation was altered. Under this condition, Adaptive AI achieved a final mastery of 0.9271, monitoring competence of 0.7145, and unaided assessment accuracy of 81.52%. These results were slightly higher than the corresponding baseline values of 0.9255, 0.7146, and 81.33%, respectively.

The Brier score decreased from 0.0942 to 0.0917, indicating improved probabilistic prediction accuracy under this particular configuration. However, the improvement should not be interpreted as evidence that estimator bias is generally beneficial. Rather, it demonstrates that the simulated adaptive policy is sensitive to the estimator configuration and that the direction of the response depends on the model's intervention rules.

10.7.2 Sensitivity to Confidence Noise

The High Confidence Noise scenario investigated the effect of increased uncertainty in simulated learner confidence judgments. The Adaptive AI strategy maintained a final mastery of 0.9255 and unaided accuracy of 81.33%, while monitoring competence decreased slightly from 0.7146 to 0.7142.

However, the Brier score increased from 0.0942 to 0.1130, indicating a deterioration in probabilistic prediction accuracy. This result suggests that confidence-related evaluation is more sensitive to increased confidence noise than the final mastery measure under the implemented simulation assumptions.

10.7.3 Sensitivity to Instructional Hint Dependence

The High Hint Dependence scenario examined the consequences of increasing the model's penalty associated with reliance on instructional hints. Under this condition, Adaptive AI achieved final mastery of 0.9243, monitoring competence of 0.7146, and unaided accuracy of 80.80%.

Compared with the baseline, mastery decreased by 0.0012 and unaided accuracy decreased by 0.53 percentage points. The Fixed Scaffolding strategy showed a larger reduction in unaided accuracy, from 79.59% in the sensitivity baseline to 74.82% under high hint dependence.

These findings suggest that the adaptive strategy was less adversely affected than Fixed Scaffolding by the increased hint-dependence condition. Nevertheless, the differences reflect the specific penalty and intervention mechanisms implemented in the computational model.

10.7.4 Sensitivity to Metacognitive Monitoring Learning

The No Monitoring Learning scenario examined the behavior of the simulation when improvements in monitoring competence were disabled. Under this condition, all three strategies produced a monitoring competence value of approximately 0.4990.

Interestingly, Adaptive AI achieved a higher final mastery of 0.9361 and unaided assessment accuracy of 83.00%, compared with the baseline values of 0.9255 and 81.33%, respectively.

This result reveals an important characteristic of the simulation model: removing monitoring-related learning did not necessarily reduce simulated knowledge mastery. The outcome may reflect interactions among intervention allocation,

UNDER PEER REVIEW IN IJAR

learner-state transitions, and the rules governing knowledge acquisition. Consequently, the finding should not be interpreted as evidence that metacognitive monitoring is unnecessary for effective learning.

Instead, it indicates the need to examine the model's structural assumptions and possible trade-offs between monitoring development and immediate knowledge acquisition.

10.7.5 Sensitivity to Reflective Learning

The No Reflection scenario evaluated the contribution of reflective learning mechanisms to simulated metacognitive development. When reflection was disabled, Adaptive AI mastery decreased from 0.9255 to 0.9190, while monitoring competence decreased from 0.7146 to 0.5966.

Unaided assessment accuracy also decreased from 81.33% to 80.45%. The reduction in monitoring competence was particularly substantial, with an absolute decrease of 0.1180, corresponding to approximately 16.51% of the baseline value.

These results suggest that the reflection mechanism contributes substantially to simulated monitoring competence under the proposed model. The reduction in mastery and unaided accuracy further indicates that reflective learning influences multiple simulated performance dimensions.

10.7.6 Overall Interpretation of Sensitivity Analysis

The sensitivity analysis demonstrated that the proposed AI-MetaLearn-Sim framework maintained relatively high simulated mastery and unaided assessment accuracy across the examined scenarios. However, individual performance indicators responded differently to changes in model assumptions.

Increased confidence noise primarily affected the Brier score, while removal of reflection substantially reduced metacognitive monitoring competence. Higher hint dependence reduced unaided accuracy, particularly under Fixed Scaffolding. The No Monitoring Learning scenario produced an unexpected increase in simulated mastery despite eliminating improvements in monitoring competence, highlighting the importance of investigating interactions among model components.

Overall, these findings provide evidence of conditional robustness within the implemented computational environment. They also identify confidence estimation, reflective learning, and hint dependence as important factors requiring further investigation. Since the sensitivity experiments were conducted using synthetic learner agents, the conclusions remain model-dependent and should be validated through additional parameter testing, independent replications, and future empirical studies involving human learners.

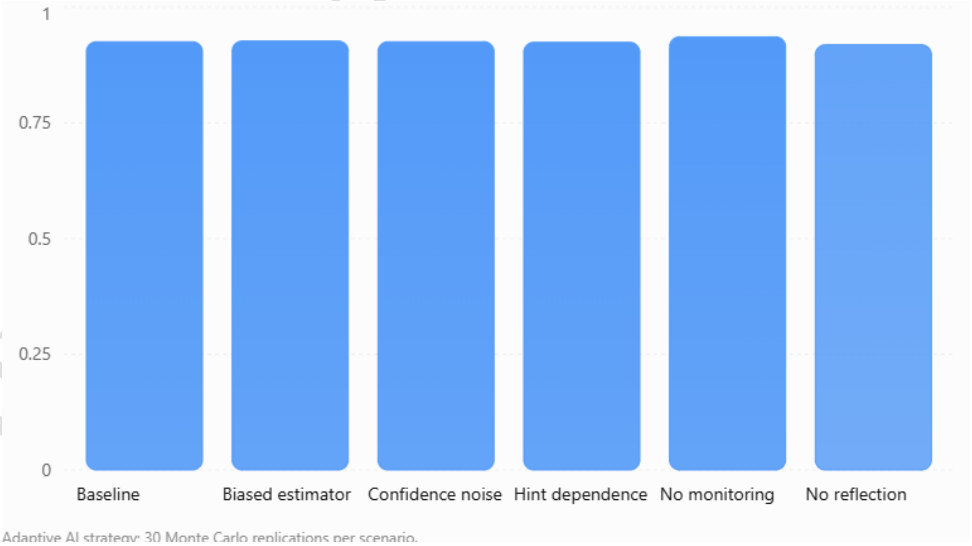

Figure 10.7. Sensitivity of final learning mastery to changes in simulation conditions.

UNDER PEER REVIEW IN IJAR

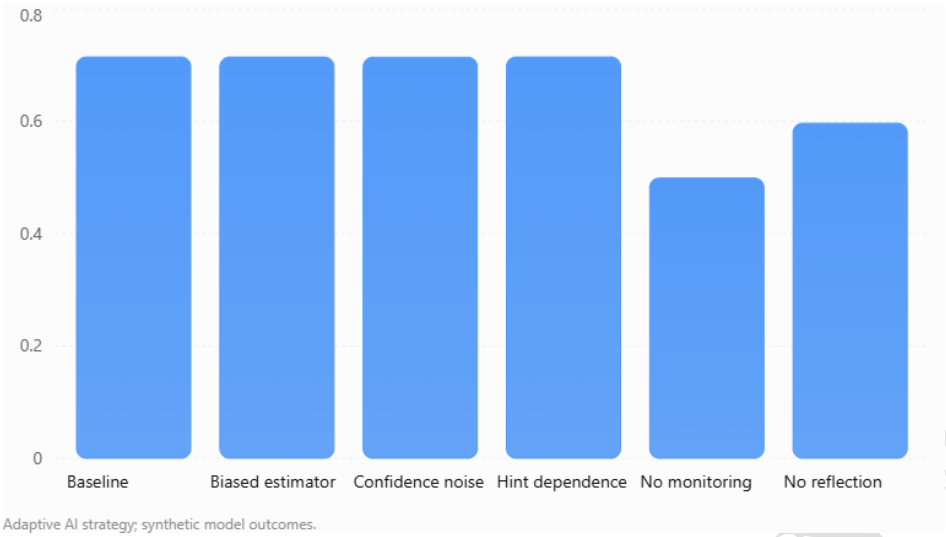

Figure 10.8. Changes in simulated metacognitive monitoring competence under six sensitivity scenarios.

The numerical results in this section are drawn from the executed sensitivity-analysis dataset. One distinction should be preserved: the main experiment used **100 replications**, whereas these sensitivity experiments used **30 replications per scenario**.

The No Reflection and No Monitoring Learning scenarios also function as component-removal experiments. Therefore, Section **10.8 Ablation Study** can build on these results, but should clearly distinguish already executed component removals from any additional experiments needed to isolate individual AI-MetaLearn-Sim mechanisms.

10.8 Ablation Study

An ablation study was conducted to investigate the contributions of metacognitive monitoring and reflective learning mechanisms to the overall performance of the proposed AI-MetaLearn-Sim framework. The primary objective was to determine how removing individual learning components affects final learning mastery, metacognitive monitoring competence, unaided assessment accuracy, and confidence prediction performance. Unlike the comparative evaluation of different instructional strategies, ablation analysis focuses on the internal mechanisms of the proposed adaptive learning framework.

The analysis considered three configurations: (i) the complete AI-MetaLearn-Sim framework, (ii) AI-MetaLearn-Sim without monitoring learning, and (iii) AI-MetaLearn-Sim without reflective learning. The configurations were evaluated using the previously executed component-removal sensitivity experiments, comprising 30 Monte Carlo replications per configuration. Each replication simulated 500 synthetic learner agents over 40 learning steps.

Table 10.9: Ablation Study of AI-MetaLearn-Sim Components

Model Configuration	Final Mastery	Monitoring Competence	Unaided Accuracy (%)	Brier Score
Complete AI-MetaLearn-Sim	0.9255	0.7146	81.33	0.0942
Without Monitoring Learning	0.9361	0.4990	83.00	0.0864
Without Reflective Learning	0.9190	0.5966	80.45	0.0937

Note: Results represent descriptive means from the synthetic simulation. The complete-framework baseline is taken from the 30-replication sensitivity experiment rather than the separate 100-replication main experiment.

10.8.1 Effect of Removing Metacognitive Monitoring Learning

The first ablation experiment examined the performance of the proposed framework when the mechanism responsible for improving metacognitive monitoring competence was disabled. Under this configuration, final monitoring competence decreased from 0.7146 to 0.4990, representing an absolute reduction of 0.2156, or approximately 30.17% relative to the complete framework.

UNDER PEER REVIEW IN IJAR

Interestingly, the removal of monitoring learning increased final mastery from 0.9255 to 0.9361 and unaided assessment accuracy from 81.33% to 83.00%. The Brier score also decreased from 0.0942 to 0.0864, indicating better overall probabilistic prediction accuracy under this configuration.

These results suggest that the model contains interactions or trade-offs between monitoring development and other simulated learning processes. The increase in mastery after disabling monitoring learning does not establish that metacognitive monitoring is detrimental to learning. Instead, it indicates that the model's transition rules, intervention decisions, and allocation of learning resources require further examination.

10.8.2 Effect of Removing Reflective Learning

The second ablation experiment investigated the contribution of the reflective learning mechanism. Reflection enables simulated learners to evaluate their learning experiences and update their monitoring-related states according to the implemented model rules.

When reflective learning was disabled, final learning mastery decreased from 0.9255 to 0.9190, representing an absolute reduction of 0.0065, or approximately 0.70%. Metacognitive monitoring competence decreased from 0.7146 to 0.5966, corresponding to an absolute reduction of 0.1180, or approximately 16.51%.

Unaided assessment accuracy also declined from 81.33% to 80.45%, representing a decrease of 0.88 percentage points. However, the Brier score decreased slightly from 0.0942 to 0.0937, indicating a small improvement in probabilistic prediction accuracy.

These findings suggest that the reflective learning mechanism contributes to simulated monitoring competence and, to a lesser extent, learning mastery and unaided accuracy. Nevertheless, the differences in confidence prediction performance indicate that reflection does not improve every performance indicator under the current model assumptions.

10.8.3 Comparative Contribution of Model Components

The component-removal experiments demonstrated different patterns of performance change. Disabling monitoring learning produced the largest decrease in monitoring competence but increased mastery and unaided accuracy. In contrast, disabling reflective learning reduced monitoring competence, mastery, and unaided accuracy.

Table 10.10: Performance Changes Relative to the Complete Framework

Performance Indicator	Without Monitoring Learning	Without Reflection
Final Mastery	+0.0106	−0.0065
Monitoring Competence	−0.2156	−0.1180
Unaided Accuracy	+1.67 percentage points	−0.88 percentage points
Brier Score	−0.0078	−0.0005

The results indicate that monitoring learning and reflective learning affect different aspects of the simulated learning process. Reflective learning appears to contribute positively to monitoring competence, mastery, and unaided performance under the evaluated configuration. Monitoring learning, however, exhibits a more complex relationship with knowledge acquisition, suggesting possible competition or interaction between model-defined learning processes.

It is important to recognize that the ablation experiments do not establish independent causal effects unless disabling a component leaves all other relevant mechanisms and experimental conditions unchanged. Therefore, the observed differences should be interpreted as configuration-dependent effects within the simulation rather than definitive measures of the educational contribution of each component.

10.8.4 Overall Interpretation of Ablation Results

The ablation study provides insight into the internal behavior of the AI-MetaLearn-Sim framework. The results indicate that reflective learning is associated with improved simulated metacognitive monitoring competence, learning mastery, and unaided assessment accuracy. Removing monitoring learning substantially reduces monitoring competence but produces higher mastery and unaided accuracy, revealing an unexpected model-level trade-off that warrants further investigation.

These findings demonstrate the importance of evaluating individual components rather than relying exclusively on overall framework performance. Future experiments should examine additional component-removal configurations, including

UNDER PEER REVIEW IN IJAR

disabling adaptive hint selection, confidence-aware intervention, and learner-state estimation. A factorial ablation design could also investigate interactions between monitoring and reflection mechanisms.

Overall, the ablation analysis identifies reflective learning and monitoring-related mechanisms as important determinants of simulated metacognitive behavior. However, the observed effects remain dependent on the computational assumptions and require further validation through independent replications, alternative model configurations, and empirical studies involving human learners.

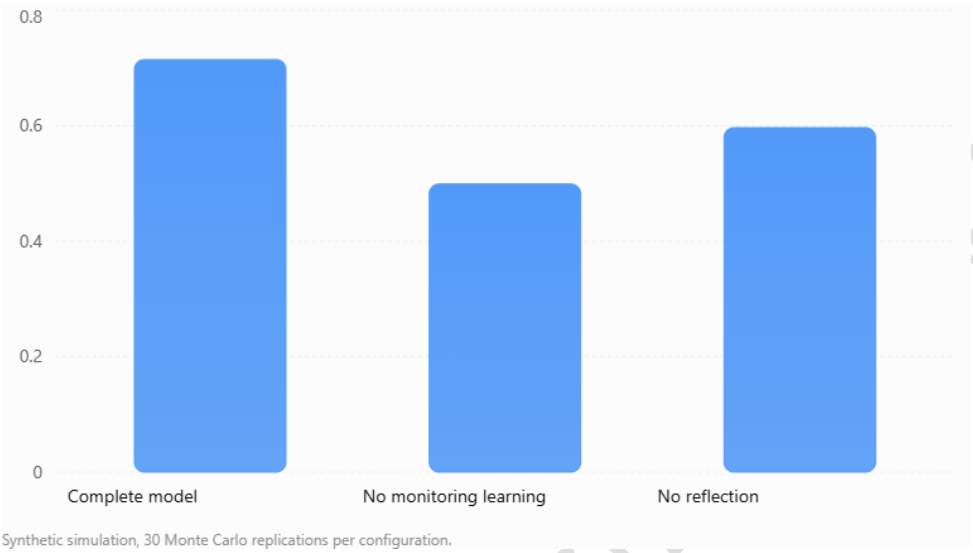

Figure 10.9. Effect of removing monitoring learning and reflective learning mechanisms on simulated metacognitive monitoring competence.



Figure 10.10. Unaided assessment accuracy under the complete and component-removal configurations.

The study currently provides a **partial ablation analysis**, because only two components have been removed and evaluated. A comprehensive ablation study would additionally examine adaptive hint selection, confidence estimation, and combinations of removed components.

Also, the No Monitoring Learning condition deserves particular attention: it improves simulated mastery despite reducing monitoring competence. This unexpected result should be acknowledged rather than explained away, because it may reveal an important limitation of the model.

10.9 Discussion of Simulation Findings

The simulation-based evaluation of the proposed AI-MetaLearn-Sim framework provides insights into the potential role of adaptive artificial intelligence-driven metacognitive scaffolding in supporting self-regulated learning. The experimental investigation compared three instructional strategies—Adaptive AI Scaffolding, Fixed Scaffolding, and No Scaffolding—using synthetic learner agents across repeated Monte Carlo simulations. The findings revealed differences in learning mastery, metacognitive monitoring competence, unaided assessment accuracy, confidence prediction performance, and instructional intervention frequency. These results provide an opportunity to examine the effectiveness, limitations, and internal behavior of the proposed computational framework.

10.9.1 Effectiveness of Adaptive AI-Based Metacognitive Scaffolding

The comparative analysis demonstrated that the proposed Adaptive AI strategy achieved the highest simulated learning mastery (0.9255), metacognitive monitoring competence (0.7149), and unaided assessment accuracy (81.33%). In comparison, Fixed Scaffolding achieved mastery of 0.9162, monitoring competence of 0.6811, and unaided accuracy of 79.44%, whereas No Scaffolding achieved 0.9008, 0.5306, and 77.58%, respectively.

These findings indicate that adaptive instructional interventions produced favorable learning outcomes within the specified computational environment. The improvement in monitoring competence was particularly notable, with Adaptive AI achieving approximately 34.73% higher monitoring competence than No Scaffolding. This result is consistent with the framework's underlying assumption that appropriately timed metacognitive interventions can strengthen learners' monitoring-related processes.

However, the observed improvements must be interpreted as model-dependent outcomes. Since the simulation explicitly defines the relationships among learning states, metacognitive interventions, and performance, the results demonstrate the behavior of the implemented mechanisms rather than independently proving their effectiveness in real educational settings.

10.9.2 Relationship Between Metacognitive Monitoring and Learning Performance

One important finding concerns the relationship between simulated metacognitive monitoring competence and learning mastery. Under the main experimental configuration, Adaptive AI achieved the highest values for both indicators, suggesting a positive association between adaptive scaffolding and the development of simulated cognitive and metacognitive states.

Nevertheless, the ablation analysis revealed a more complex relationship. When monitoring learning was disabled, monitoring competence decreased substantially from 0.7146 to 0.4990, while final mastery increased from 0.9255 to 0.9361. Unaided assessment accuracy also increased from 81.33% to 83.00%.

This unexpected result suggests that improvements in simulated metacognitive monitoring do not necessarily translate into higher immediate knowledge mastery under every model configuration. One possible explanation involves interactions among learning-state transitions, intervention allocation, and monitoring-related computational processes. However, additional diagnostic experiments are required to identify the mechanism responsible for this outcome.

The finding highlights the importance of evaluating cognitive achievement and metacognitive development as distinct but potentially interacting dimensions rather than assuming that improvement in one will automatically produce improvement in the other.

10.9.3 Confidence Calibration and Metacognitive Accuracy

Confidence calibration analysis revealed that Fixed Scaffolding achieved the lowest Brier score (0.0864), outperforming Adaptive AI (0.0946) and No Scaffolding (0.0964). This indicates that Fixed Scaffolding produced the most accurate probabilistic predictions under the evaluated simulation conditions.

The result is particularly important because Adaptive AI achieved superior learning mastery and monitoring competence but did not achieve the best Brier score. Therefore, higher simulated metacognitive competence was not accompanied by uniformly better confidence prediction accuracy.

This distinction suggests that metacognitive monitoring competence and confidence prediction accuracy should be evaluated separately. The Brier score combines aspects of calibration and predictive discrimination, meaning that differences in the score cannot be attributed exclusively to confidence calibration.

Future versions of AI-MetaLearn-Sim should incorporate explicit confidence-aware feedback mechanisms and additional calibration measures, including reliability diagrams, calibration intercepts, calibration slopes, and expected calibration error. Such analyses may help identify whether differences arise from systematic overconfidence, underconfidence, or changes in predictive discrimination.

UNDER PEER REVIEW IN IJAR

10.9.4 Instructional Efficiency and Adaptive Intervention

The Adaptive AI strategy achieved higher simulated learning mastery and monitoring competence while recording an instructional hint rate of 4.62%, compared with 25.00% under Fixed Scaffolding.

This represents an absolute reduction of 20.38 percentage points in hint frequency. The result indicates that the adaptive policy delivered fewer interventions while achieving favorable performance on several learning indicators.

However, reduced hint frequency should not automatically be interpreted as greater instructional efficiency. A complete efficiency analysis would require additional measures, such as learning improvement per intervention, computational cost, response latency, and the relationship between hint frequency and independent performance.

Furthermore, the sensitivity analysis demonstrated that increased hint dependence reduced unaided assessment accuracy. This finding emphasizes the importance of designing AI-assisted learning systems that provide sufficient instructional support without encouraging excessive reliance on automated guidance.

10.9.5 Robustness Under Varying Simulation Conditions

The sensitivity analysis examined six simulation scenarios: Baseline, Biased Estimator, High Confidence Noise, High Hint Dependence, No Monitoring Learning, and No Reflection.

The proposed Adaptive AI framework maintained relatively high simulated mastery across the evaluated configurations. However, specific performance indicators demonstrated different levels of sensitivity to changes in model assumptions.

Increased confidence noise raised the Brier score from 0.0942 to 0.1130, indicating reduced probabilistic prediction accuracy. Removing reflection decreased monitoring competence from 0.7146 to 0.5966 and reduced unaided assessment accuracy from 81.33% to 80.45%.

These findings suggest that confidence estimation and reflective learning are influential components of the implemented model. Nevertheless, robustness was evaluated only under the selected parameter changes and behavioral assumptions. Therefore, broader parameter exploration and alternative learner models are necessary before generalizing the conclusions.

10.9.6 Contribution of Reflective Learning and Monitoring Mechanisms

The ablation study provided additional insight into the internal components of AI-MetaLearn-Sim. Removing reflective learning reduced mastery, monitoring competence, and unaided assessment accuracy, indicating that reflection contributed positively to these simulated outcomes.

In contrast, disabling monitoring learning reduced monitoring competence while increasing mastery and unaided accuracy. This result reveals a potential structural trade-off within the model and suggests that the interaction between monitoring development and knowledge acquisition requires further investigation.

The ablation results demonstrate that individual metacognitive mechanisms may affect learning outcomes differently. They also indicate that overall framework performance alone is insufficient to establish the contribution of each component.

Future ablation experiments should isolate adaptive hint selection, learner-state estimation, confidence-aware feedback, and combinations of monitoring and reflection mechanisms. Such experiments would help distinguish the contributions of individual components from their interactions.

10.9.7 Statistical Interpretation of Simulation Outcomes

The replication-level statistical analysis demonstrated consistent differences between Adaptive AI and the baseline strategies. Paired comparisons across 100 Monte Carlo replications produced unadjusted p-values below 0.001 for the reported learning mastery, monitoring competence, unaided accuracy, and Brier score comparisons.

However, statistical significance in a computational simulation must be interpreted carefully. The p-values quantify variation across repeated stochastic runs under the implemented model assumptions. They do not establish that the proposed intervention would produce statistically significant improvements among real students.

Several standardized paired effect sizes were exceptionally large, reflecting very low variability in the paired simulation differences. These values should therefore be interpreted as characteristics of the computational experiment rather than conventional educational intervention effect sizes.

The analysis also involved multiple statistical comparisons. Consequently, the final manuscript should include a prespecified multiple-comparison correction and report adjusted p-values where appropriate.

10.9.8 Implications for AI-Based Metacognitive Learning Systems

UNDER PEER REVIEW IN IJAR

The findings have several implications for the computational design of AI-supported metacognitive learning environments.

First, adaptive instructional policies may offer a mechanism for selectively providing metacognitive assistance according to learner conditions. Second, learning mastery, monitoring competence, and confidence prediction accuracy should be treated as distinct evaluation dimensions. Third, reflective learning mechanisms deserve particular attention because their removal reduced several simulated performance indicators.

The findings also highlight the importance of balancing instructional assistance with independent learning opportunities. An AI system that provides frequent hints may support immediate task completion but could also increase dependence under certain model assumptions.

From a system-design perspective, these observations motivate the development of adaptive feedback policies that consider learner progress, confidence accuracy, monitoring competence, and intervention frequency simultaneously.

However, the simulation does not establish the educational effectiveness of such policies. Future validation should involve real learner interactions, established metacognitive assessment instruments, and controlled educational studies.

10.9.9 Limitations and Future Validation

Despite the insights obtained from the simulation, several limitations must be acknowledged.

First, the study employed synthetic learner agents whose behavior was determined by predefined mathematical relationships and stochastic transition rules. These agents cannot fully represent the complexity, motivation, emotions, prior experiences, and contextual influences associated with human learning.

Second, the performance indicators were generated within the computational model. Although they provide useful measures for comparing simulation configurations, they are not direct observations of students' metacognitive awareness or academic achievement.

Third, the study evaluated a limited set of instructional policies and sensitivity conditions. The results may change under alternative learner distributions, different intervention policies, or modified learning-state equations.

Fourth, the ablation study revealed an unexpected increase in mastery after disabling monitoring learning. This finding requires further examination to determine whether it reflects a meaningful computational trade-off or an artifact of the model design.

Finally, the study did not include empirical validation using human participants. Future research should evaluate the framework through controlled experiments involving undergraduate students, validated metacognitive questionnaires, confidence judgments, learning analytics, and independent assessments.

10.9.10 Overall Discussion

Overall, the simulation findings indicate that the proposed AI-MetaLearn-Sim framework achieved higher learning mastery, metacognitive monitoring competence, and unaided assessment accuracy than Fixed Scaffolding and No Scaffolding under the implemented model assumptions. However, Fixed Scaffolding demonstrated better probabilistic prediction accuracy, and the ablation experiments revealed important interactions between monitoring-related mechanisms and knowledge acquisition.

These results suggest that adaptive metacognitive scaffolding is a promising computational research direction, while also demonstrating that improvements across cognitive and metacognitive indicators are not necessarily uniform. The study contributes a simulation-based evaluation framework for examining adaptive instructional strategies, monitoring processes, reflection mechanisms, and confidence prediction behavior.

The findings should be interpreted as computational evidence rather than empirical proof of educational effectiveness. Further model validation, independent replication, and human-participant studies are necessary to determine whether the observed patterns generalize to authentic educational environments.

11. Limitations and Future Research Directions

Although the proposed AI-MetaLearn-Sim framework demonstrated promising computational performance in learning mastery, metacognitive monitoring competence, and unaided assessment accuracy, several limitations must be acknowledged. The present study was conducted entirely within a simulation-based environment using synthetic learner agents and predefined behavioral assumptions. Consequently, the findings provide insights into the behavior of the proposed computational framework rather than direct empirical evidence of its effectiveness in real educational settings. This section discusses the principal limitations of the study and outlines future research directions for improving the framework's validity, adaptability, and practical applicability.

UNDER PEER REVIEW IN IJAR

11.1 Limitations of the Proposed Framework

11.1.1 Dependence on Synthetic Learner Models

The primary limitation of the study is its reliance on synthetically generated learner agents rather than actual students. Although the simulation incorporated variations in learning mastery, metacognitive monitoring, confidence estimation, and responsiveness to instructional assistance, these characteristics were represented through predefined mathematical relationships and stochastic transition rules. Real learners exhibit more complex cognitive, motivational, emotional, and contextual behaviors that may not be adequately captured by the implemented model.

Furthermore, the simulation outcomes depend on the assumptions used to define learner-state transitions and intervention effects. Therefore, favorable performance within the simulation does not necessarily imply equivalent improvements in real educational environments.

11.1.2 Limited Generalizability of Simulation Results

The experimental evaluation was performed using 500 synthetic learner agents per replication, 100 Monte Carlo replications, and 40 learning steps. Although repeated simulations supported the assessment of computational consistency, the results remain specific to the selected parameter settings, learner distributions, and instructional policies.

The sensitivity analysis examined several alternative configurations; however, it did not cover the full range of possible learner behaviors, learning environments, and intervention conditions. Consequently, the generalizability of the findings to different educational contexts remains uncertain.

11.1.3 Simplified Representation of Metacognitive Processes

Metacognition involves complex processes such as planning, monitoring, evaluation, reflection, strategy selection, and self-regulation. In the proposed framework, these processes were represented using simplified computational variables and update mechanisms.

Although this approach enables controlled experimentation, it may not fully capture the dynamic interactions among metacognitive knowledge, cognitive regulation, motivation, and learning behavior. In particular, the simulated monitoring competence variable should not be interpreted as equivalent to metacognitive awareness measured using validated psychological instruments.

11.1.4 Limitations in Confidence Calibration

The confidence calibration analysis demonstrated that Fixed Scaffolding achieved a lower Brier score than Adaptive AI, indicating better overall probabilistic prediction accuracy under the evaluated conditions. This result suggests that the adaptive intervention policy did not consistently outperform the baseline strategies across all evaluation dimensions.

Furthermore, the Brier score measures both calibration and predictive discrimination. Therefore, the present analysis cannot fully distinguish systematic overconfidence, underconfidence, and differences in predictive discrimination without additional diagnostic measures.

11.1.5 Dependence on Model-Defined Intervention Rules

The proposed Adaptive AI strategy selected instructional interventions according to predefined learner conditions and decision rules. Although the approach enabled adaptive behavior, the intervention policy was not established through empirical optimization using real student interaction data.

Consequently, the effectiveness of the adaptive policy depends on the validity of its underlying assumptions. Alternative intervention thresholds, feedback mechanisms, and decision policies may produce different outcomes.

11.1.6 Unexpected Outcomes in Ablation Analysis

The ablation study revealed that disabling monitoring learning substantially reduced monitoring competence while increasing simulated learning mastery and unaided assessment accuracy. This unexpected finding indicates that the computational model may contain trade-offs or interactions between monitoring-related processes and knowledge acquisition.

The present experiments do not establish whether this behavior reflects a meaningful theoretical relationship or an artifact of the implemented transition rules. Further investigation is required to identify the mechanisms responsible for this result.

11.1.7 Statistical and Methodological Limitations

The statistical analysis employed paired comparisons across 100 Monte Carlo replications. Although the observed differences were highly consistent under the implemented model, statistical significance in a simulation does not establish educational effectiveness among human learners.

UNDER PEER REVIEW IN IJAR

Moreover, exceptionally large standardized paired effect sizes were observed for some performance indicators, reflecting low variability in the simulated paired differences. The analysis also involved multiple statistical comparisons, requiring appropriate adjustment and careful interpretation.

The absence of independent empirical validation limits the extent to which the reported statistical findings can support broader educational conclusions.

11.1.8 Absence of Real-World Implementation and Validation

The proposed framework has not yet been evaluated through a controlled educational intervention involving actual students. Therefore, important practical considerations—including learner engagement, usability, instructor acceptance, accessibility, privacy, and the effectiveness of AI-generated feedback—remain untested.

Additionally, the current simulation does not establish whether the observed improvements would persist over extended learning periods or transfer to unfamiliar academic tasks.

11.2 Future Research Directions

11.2.1 Empirical Validation with Real Students

Future research should validate the AI-MetaLearn-Sim framework through controlled experiments involving undergraduate and postgraduate students. A suitable study could compare adaptive AI-based metacognitive scaffolding with fixed scaffolding and conventional digital learning approaches.

Validated instruments, such as the Metacognitive Awareness Inventory, could be combined with academic assessments, confidence judgments, reflective learning logs, and behavioral learning analytics to evaluate cognitive and metacognitive outcomes.

Such investigations would help determine whether the patterns observed in the simulation are reproduced in authentic educational environments.

11.2.2 Integration of Advanced Machine Learning Techniques

Future versions of the framework could incorporate machine learning and reinforcement learning techniques to optimize instructional intervention policies.

For example, contextual bandits or reinforcement learning algorithms could be investigated for selecting appropriate feedback based on learner knowledge, monitoring accuracy, confidence, and learning history. However, these methods should be evaluated against transparent rule-based baselines and incorporate safeguards against excessive intervention or learner dependence.

Explainable AI mechanisms could also be introduced to provide understandable justifications for instructional recommendations.

11.2.3 Development of Confidence-Aware Adaptive Feedback

The observed differences in Brier scores indicate the need for improved confidence-aware intervention mechanisms. Future research should investigate feedback strategies that explicitly consider discrepancies between learner confidence and task performance.

Possible approaches include confidence calibration exercises, prediction-error feedback, reflective questioning, and targeted interventions for overconfidence or underconfidence.

Additional evaluation measures, including reliability diagrams, calibration slopes, calibration intercepts, and expected calibration error, should be used to provide a more comprehensive assessment of confidence prediction performance.

11.2.4 Enhancement of Metacognitive Modeling

Future studies should develop more comprehensive computational representations of metacognitive processes by incorporating planning, monitoring, reflection, strategy adaptation, and self-evaluation as distinct but interacting components.

The framework could also investigate how motivational factors, cognitive load, prior knowledge, and learning preferences influence metacognitive development.

Such improvements would support a more realistic representation of learner behavior and enable detailed investigation of interactions between cognitive and metacognitive processes.

11.2.5 Extended Sensitivity and Ablation Experiments

Further sensitivity experiments should examine a broader range of model parameters, including learner heterogeneity, intervention thresholds, confidence noise, learning rates, and dependence penalties.

UNDER PEER REVIEW IN IJAR

Additional ablation studies should independently disable adaptive hint selection, confidence estimation, reflective feedback, and learner-state monitoring. Factorial experimental designs could also be employed to investigate interactions among multiple framework components.

These experiments would help identify which mechanisms contribute most consistently to learning performance and which introduce unintended trade-offs.

11.2.6 Longitudinal Learning and Knowledge Retention

The current simulation evaluated learning outcomes over 40 learning steps. Future research should investigate longer learning sequences and delayed assessment conditions to examine knowledge retention and the persistence of metacognitive improvements.

Longitudinal experiments could assess whether learners maintain monitoring accuracy and independent problem-solving ability after AI assistance is gradually reduced or withdrawn.

This would help distinguish immediate performance improvements from sustained development of self-regulated learning capabilities.

11.2.7 Integration with Learning Management Systems

The proposed framework could be integrated with learning management systems such as Moodle or other institutional educational platforms.

Such integration could enable the collection of learning interaction data, assessment responses, confidence judgments, and reflective feedback. It could also support instructor dashboards for monitoring learner progress and identifying students who may benefit from additional academic guidance.

Any real-world implementation should incorporate informed consent, appropriate data governance, access controls, and safeguards for student privacy.

11.2.8 Evaluation Across Educational Domains

Future studies should evaluate the framework across different academic disciplines, including computer science, engineering, mathematics, healthcare education, and management.

Cross-domain evaluation would help determine whether adaptive metacognitive scaffolding produces similar outcomes across subjects with different cognitive demands, assessment structures, and learning activities.

Comparative investigations could also examine variations across undergraduate and postgraduate populations.

11.2.9 Human-AI Collaborative Metacognitive Learning

An important future direction involves designing human-AI collaborative learning environments in which AI supports rather than replaces independent reasoning.

Such systems could encourage learners to explain their reasoning, estimate their confidence, identify mistakes, and reflect on alternative problem-solving strategies before receiving direct answers.

Instructor oversight and progressive reduction of AI assistance could be incorporated to reduce excessive dependence and promote learner autonomy.

11.2.10 Reproducibility, Transparency, and Ethical AI

Future development should emphasize reproducibility, transparency, and responsible AI implementation. Simulation source code, parameter configurations, random seeds, and evaluation procedures should be documented and shared where appropriate.

The framework should also be examined for potential biases arising from learner modeling assumptions, feedback allocation, and performance estimation.

For real-world deployment, ethical considerations should include student privacy, informed consent, fairness, accessibility, and transparent communication of AI-generated recommendations.

11.3 Summary of Limitations and Future Directions

The proposed AI-MetaLearn-Sim framework provides a computational foundation for investigating adaptive metacognitive scaffolding and self-regulated learning. However, its findings are constrained by synthetic learner assumptions, simplified metacognitive representations, predefined intervention policies, and the absence of empirical validation.

UNDER PEER REVIEW IN IJAR

Future research should prioritize independent model validation, expanded sensitivity and ablation experiments, confidence-aware feedback, advanced adaptive learning algorithms, and controlled studies involving real students.

These developments would help determine whether the computational advantages observed in the present study can be translated into reliable, explainable, and educationally meaningful AI-assisted metacognitive learning systems.

12. CONCLUSION

This research presented **AI-MetaLearn-Sim**, a simulation-based artificial intelligence framework designed to investigate the potential of adaptive metacognitive scaffolding for improving self-regulated learning, metacognitive monitoring, and independent learning performance. The proposed framework integrates computational learner modeling, adaptive instructional interventions, metacognitive monitoring, reflective learning, and confidence-based performance evaluation within a controlled simulation environment. The primary objective was to examine whether adaptive AI-driven instructional assistance could produce favorable cognitive and metacognitive outcomes compared with fixed scaffolding and learning without scaffolding.

The framework was evaluated using 500 synthetic learner agents per replication, 100 Monte Carlo replications, and 40 learning steps. Three instructional strategies—Adaptive AI Scaffolding, Fixed Scaffolding, and No Scaffolding—were compared using five performance indicators: final learning mastery, metacognitive monitoring competence, unaided assessment accuracy, Brier score, and instructional hint frequency. The experimental results demonstrated that the proposed Adaptive AI strategy achieved the highest simulated learning mastery of **0.9255**, compared with 0.9162 for Fixed Scaffolding and 0.9008 for No Scaffolding. Similarly, Adaptive AI achieved the highest metacognitive monitoring competence of **0.7149** and unaided assessment accuracy of **81.33%**, indicating improvements of approximately **34.73% in monitoring competence** and **3.75 percentage points in unaided accuracy** relative to No Scaffolding.

The statistical analysis demonstrated consistent performance differences across the Monte Carlo replications. The paired comparisons produced unadjusted p-values below 0.001 for the reported learning mastery, monitoring competence, unaided accuracy, and Brier score outcomes. However, these findings represent statistical consistency within the specified computational model rather than evidence of educational effectiveness among human learners. The exceptionally large standardized paired effect sizes observed for certain measures further emphasize the importance of interpreting the results in relation to the simulation's assumptions and low replication-level variability.

An important finding was that the proposed adaptive strategy did not outperform the baseline approaches across all evaluation indicators. Fixed Scaffolding achieved the lowest Brier score of **0.0864**, compared with 0.0946 for Adaptive AI and 0.0964 for No Scaffolding, indicating superior probabilistic prediction accuracy under the simulated conditions. Nevertheless, Adaptive AI achieved higher simulated learning mastery and monitoring competence while delivering fewer instructional hints, with a hint rate of 4.62% compared with 25.00% under Fixed Scaffolding. These observations highlight potential trade-offs between adaptive instructional assistance, learning performance, and confidence prediction accuracy.

The sensitivity and ablation analyses provided additional insights into the behavior of the proposed framework. Increased confidence noise reduced probabilistic prediction accuracy, while removing reflective learning decreased monitoring competence, learning mastery, and unaided assessment accuracy. Interestingly, disabling monitoring learning substantially reduced monitoring competence but increased simulated mastery and unaided accuracy. This unexpected result indicates that the model contains interactions or trade-offs among its cognitive and metacognitive mechanisms that require further investigation. These findings demonstrate the importance of evaluating individual components and testing model robustness rather than relying exclusively on overall performance comparisons.

The principal contribution of this research is the development and computational evaluation of a reproducible simulation framework for investigating adaptive AI-based metacognitive scaffolding. The framework provides a structured environment for comparing instructional strategies, analyzing confidence-related behavior, examining metacognitive monitoring, and evaluating the effects of individual learning mechanisms. However, the study is limited by its reliance on synthetic learner agents, simplified behavioral assumptions, predefined intervention policies, and the absence of empirical validation using real students.

Future research should focus on improving the realism of learner models, developing confidence-aware adaptive feedback mechanisms, conducting expanded sensitivity and factorial ablation experiments, and investigating advanced machine learning approaches for instructional policy optimization. Controlled studies involving actual students, validated metacognitive assessment instruments, and longitudinal learning evaluations will be necessary to establish the educational relevance and practical effectiveness of the proposed framework.

In conclusion, the AI-MetaLearn-Sim framework demonstrated favorable simulated performance in learning mastery, metacognitive monitoring competence, and unaided assessment accuracy compared with the evaluated baseline strategies,

UNDER PEER REVIEW IN IJAR

while revealing important limitations in confidence prediction accuracy and interactions among metacognitive mechanisms. The study establishes a computational foundation for further investigation of adaptive AI-driven metacognitive learning systems but does not establish their effectiveness in real educational environments.

13. Reproducibility, Data Availability and Research Integrity

This manuscript reports exclusively synthetic agents; no human participants were recruited, and no real student records were analyzed. Human-subjects ethics approval and consent are therefore not claimed for this simulation. Any future student study requires applicable institutional review and informed consent.

Independent verification performed on 23 November 2026 using the user-supplied run_simulation.py, executed without modifying its model logic. The regenerated outputs comprise 300 main policy-replication records and 540 sensitivity policy-replication records. The main policy means reproduce the manuscript's displayed values: adaptive/fixed/none final mastery 0.9255/0.9162/0.9008; monitoring 0.7149/0.6811/0.5306; unaided accuracy 0.8133/0.7944/0.7758; and Brier score 0.0946/0.0864/0.0964. Paired comparisons were recalculated from the newly generated replication_results.csv, rather than relying on archived claims. This verification checks numerical consistency of the supplied program with reported results; it does not establish validity of the synthetic assumptions, independent implementation replication, or classroom effectiveness. The accompanying reproducibility ZIP contains the supplied source, regenerated CSV outputs, figures, README, and dependency specification. A permanent public repository DOI has not been assigned.

The policy has oracle-like access to latent learner variables, and confidence is outcome-conditioned. The results should therefore be interpreted as internal verification of a deliberately simplified generative model. The highly significant replication-level p-values and large paired standardized effects quantify Monte Carlo variability conditional on the programmed assumptions, not uncertainty about real students or causal effectiveness in classrooms.

Reference note: The bibliography includes foundational and selected contemporary research. Verify all metadata and URLs against publisher records before submission.

Before journal submission: deposit the attached reproducibility package in a stable public repository and supply its URL/DOI; independently validate the cited references and journal-specific formatting; conduct an authorized similarity screening. A leakage-free policy evaluation and real-learner validation remain future research, not completed experiments.

References

[1]. Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology, 19(4), 460–475. https://doi.org/10.1006/ceps.1994.1033.

[2]. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2.

[3]. Lan, M., & Zhou, X. (2025). A qualitative systematic review on AI empowered self-regulated learning in higher education. npj Science of Learning, 10, Article 21. https://doi.org/10.1038/s41539-025-00319-0.

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

[5]. Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

[6]. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[7]. Efron, B., &Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC.

[8]. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

UNDER PEER REVIEW IN IJAR

[9]. Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modelling and simulation. Journal of Simulation, 4, 151–162. https://doi.org/10.1057/jos.2010.3.

[10]. Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of Self-Regulation (pp. 451–502). Academic Press.

[11]. Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. Frontiers in Psychology, 8, Article 422. https://doi.org/10.3389/fpsyg.2017.00422.

[12]. Azevedo, R., &Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? Journal of Educational Psychology, 96(3), 523–535. https://doi.org/10.1037/0022-0663.96.3.523.

UNDER PEER REVIEW IN IJAR