An AI-Powered Framework for Personalized Coaching and Incentive Strategies to Enhance Performance of Secondary-Level Biology Students in Competitive Examinations

S. D. Abeysekara
University of Colombo
School of Computing
Colombo, Sri Lanka
sanuabeysekara@gmail.com

Prof. D. D. Karunaratne
University of Colombo
School of Computing
Colombo, Sri Lanka
ddk@ucsc.cmb.ac.lk

Dr. R.K. Abeysekara
Esymas (Pvt) Ltd
Colombo, Sri Lanka
contactruvan@gmail.com

Abstract—Secondary-level biology students preparing for competitive examinations require timely misconception diagnosis, adaptive practice, motivational support, and feedback that remains grounded in the syllabus. This paper presents an AI-powered coaching and incentive framework for Sri Lankan Grade 13 English-medium Biology students studying the Microbiology unit. The central contribution is a five-layer architecture that connects student activity evidence, a 27-submetric learner profile, a hybrid intervention decision system, and a guided LLM coaching layer. The intervention system uses an expert rule gate to enforce eligibility, threshold, sequencing, and fatigue constraints, while a contextual-bandit ranker adapts intervention priorities when outcome data become available. The LLM is therefore used as a bounded coaching delivery layer rather than as an unrestricted intervention selector. Its tool-calling module can retrieve student progress, search unit materials, recommend quizzes, manage study plans, inspect badge progress, and record topic-confidence checks. The evaluation retained 132 complete paired records from a 167-student registered sample frame. Readiness increased from 52.09 to 58.90 on a 0–100 scale, with a mean gain of 6.81 points, $t(131) = 6.27$, $p < .001$, and $d_z = 0.55$. The framework illustrates how expert knowledge, adaptive machine learning, measurable backend evidence, and LLM-based coaching can be combined without exposing students to an unguided chatbot.

Index Terms—AI in education, contextual bandits, large language models, intelligent tutoring systems, biology education, adaptive interventions, learning analytics, incentives

I. Introduction

Competitive examinations strongly influence academic progression for secondary-level science students in Sri Lanka. In Biology, this pressure is amplified by the need to combine factual recall, conceptual understanding, higher-order application, and timed assessment performance. Conventional classroom and digital learning systems often provide the same content and practice path to all learners, even when students differ in prior knowledge, misconception patterns, motivation, confidence, and response to feedback. Recent evidence on AI-driven intelligent tutoring systems indicates generally positive but design-sensitive learning effects in K–12 settings [1], while feedback research emphasizes that feedback quality depends on timing, focus, and actionability [9], [12]. However, many AI tutoring solutions still operate as isolated chat or content-delivery modules rather than as complete exam-preparation systems.

This research addresses the design problem of integrating personalized coaching, adaptive intervention selection, incentive support, learning analytics, and controlled LLM interaction for secondary-level Biology competitive examination preparation. The framework was developed in a web-based learning environment and targets Grade 13 English-medium students studying the Microbiology unit. The work follows a design-science orientation in which the artefact is evaluated through backend-aligned learning evidence and readiness outcomes [17]. Structured Biology content is not treated as a student-varying predictor because all students receive the same curriculum-scoped content and assessment-pack structure. Instead, structured content delivery is treated as a fixed implementation condition, while student-level variation is modeled through practice, mastery, cognitive skill, engagement, AI coaching interaction, and incentive constructs.

The main research contribution is an adaptive coaching framework that connects policy-level intervention selection with LLM-based coaching delivery. The LLM does not independently decide what a student should do next. Instead, normalized learner evidence is passed to a hybrid intervention decision system, and the selected intervention plan is supplied to the coaching orchestrator as the bounded action context. The coach then delivers that plan using approved tools, resource-grounded explanations, study-plan actions, recommendation cards, badge prompts, confidence checks, and feedback collection. This design preserves the conversational benefit of LLMs while keeping educational decisions constrained by measurable evidence and explicit policy rules.

The paper makes three contributions:

- a 27-submetric student-profile schema aligned with backend evidence for Biology exam readiness;
- a hybrid expert-rule-gated and contextual-bandit-ranked policy for adaptive coaching and incentive intervention selection;
- an evaluation model for pre-test/post-test exam-readiness analysis using student clusters, backend metrics, and intervention-response evidence.

A. Research Questions

The study is guided by the following research questions:

RQ1: What AI models are most effective for personalized coaching and incentives in secondary-level Biology competitive examination preparation?

RQ2: What constitutes an optimal AI-powered framework integrating coaching and incentives for secondary-level Biology competitive examinations?

RQ3: How do AI-driven coaching and incentives improve secondary-level Biology students' competitive examination readiness?

RQ4: What ethical, privacy, and data-security considerations must be addressed when implementing AI-powered coaching and incentive frameworks in secondary education?

B. Research Hypotheses

The dependent variable of this study is final exam readiness. The six independent variables are practice assessment performance, mastery and misconception diagnosis, cognitive skill performance, learning engagement and path progress, AI-powered coaching interaction and personalization, and incentive engagement. Initial academic preparedness is included as the control variable so that the evaluation can examine the contribution of these six independent variables after accounting for students' starting academic profile.

$H_{1,0}$: Practice assessment performance does not make a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{1,a}$: Practice assessment performance makes a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{2,0}$: Mastery and misconception diagnosis does not make a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{2,a}$: Mastery and misconception diagnosis makes a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{3,0}$: Cognitive skill performance does not make a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{3,a}$: Cognitive skill performance makes a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{4,0}$: Learning engagement and path progress does not make a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{4,a}$: Learning engagement and path progress makes a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{5,0}$: AI-powered coaching interaction and personalization does not make a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{5,a}$: AI-powered coaching interaction and personalization makes a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{6,0}$: Incentive engagement does not make a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

$H_{6,a}$: Incentive engagement makes a significant positive contribution to students' final exam readiness after controlling for initial academic preparedness.

II. Related Work

Recent intelligent tutoring research supports the use of AI-mediated tutoring, but it also shows why tutoring must be evaluated as a designed learning system rather than as a generic automation layer. Létourneau et al. reviewed AI-driven intelligent tutoring systems in K–12 education and found generally positive effects, while also noting variation in experimental design, intervention duration, and ethical reporting [1]. This motivates a framework in which coaching is connected to measurable learner evidence, controlled action selection, and explicit evaluation metrics.

AI-supported personalization extends this tutoring tradition by adapting learning paths to learner state, pedagogical goals, and contextual constraints. Bayly-Castaneda et al. synthesize recent work on AI-mediated personalized learning paths and emphasize that adaptive learning requires sound pedagogical design rather than only algorithmic matching [2]. In the proposed framework, this principle is implemented through a student-profile layer and a policy layer. Sequential action selection is represented using contextual bandits, which have been proposed for personalized learning action selection [10]; Thompson sampling provides a practical exploration–exploitation mechanism for such adaptive decisions [11].

Large language models add a conversational delivery layer, but recent education research stresses that they should be used with constraints. Kasneci et al. discuss LLM opportunities for educational content, engagement, and personalization alongside risks such as bias, brittle

outputs, and the need for human oversight [3]. Labadze et al. review AI chatbots in education and report benefits for study assistance and personalized learning, while also emphasizing reliability, accuracy, and ethical concerns [4]. Yan et al. identify practical and ethical challenges across LLM-supported education tasks, including feedback, grading, content generation, and recommendation [5]. Zhang et al. further show that generative AI research in K–12 education reports benefits for performance, cognition, motivation, and personalization, but also risks such as erroneous content and privacy concerns [6]. These findings support the decision to connect the intervention selector and the coaching LLM as a bounded decision-to-delivery pipeline.

Grounding is necessary when LLMs are used for syllabus-based tutoring. Retrieval-augmented generation (RAG) has been reviewed as a mechanism for combining language models with external knowledge stores to improve factuality and domain alignment [7]. In education, Henkel et al. show that RAG-based question answering can improve curriculum grounding but still involves trade-offs between groundedness and user preference [8]. Therefore, the proposed coach uses retrieval and approved tools, but does not allow the LLM to independently choose the educational intervention.

Feedback and motivation remain central to exam preparation. Hattie and Timperley define effective feedback in terms of where the learner is going, how the learner is progressing, and what action should come next [12]; recent GenAI feedback reviews likewise emphasize opportunities for timely personalized feedback and the need for careful integration [9]. Self-regulated learning theory emphasizes planning, monitoring, and strategic adjustment [13], while self-determination theory links sustained motivation to competence, autonomy, and relatedness [14]. Gamification research defines badges and related mechanisms as game-design elements in non-game contexts [15]; however, the present framework avoids relying on rewards alone. Incentives are treated as adaptive nudges tied to practice completion, badge thresholds, streak maintenance, and learning progress.

III. Proposed Framework

A. Design Scope

The framework is designed for Grade 13 English-medium Biology students in Sri Lanka preparing for competitive examinations, with the Microbiology unit used as the evaluation scope. All participants receive the same structured content and assessment delivery. This design choice makes content delivery a fixed condition rather than a regression predictor. Fidelity is checked through curriculum tagging coverage, MCQ/essay format coverage, unit-material availability, and assessment-pack structure coverage.

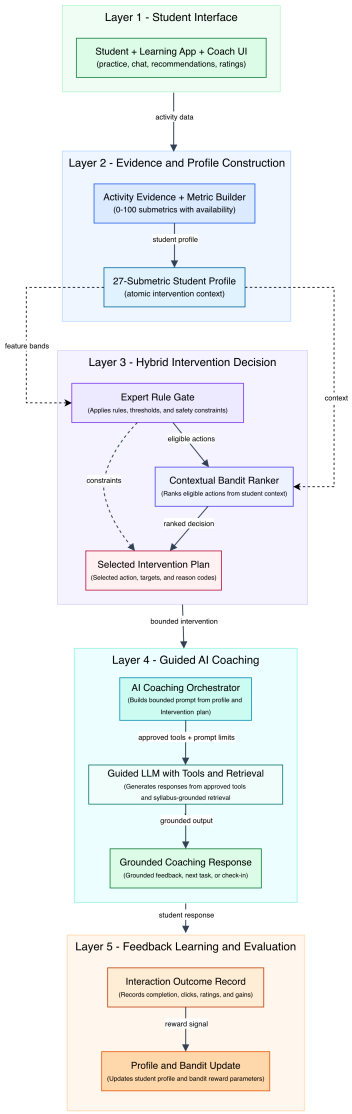

Fig. 1. Five-layer system interaction model for the proposed coaching framework. Student activity is transformed into a 27-submetric profile; the hybrid intervention decision layer selects a bounded plan; the guided AI coaching layer delivers that plan through approved tools and retrieval; and feedback outcomes update the profile and adaptive policy.

B. Layered System Interaction

Fig. 1 summarizes the operational design. Layer 1 is the student interface, where learners complete practice, use the coach, receive recommendations, and rate feedback. Layer 2 converts activity evidence into a research-ready profile. The metric builder produces normalized submetrics with explicit availability information and combines them into composite constructs. Layer 3 is the hybrid intervention decision layer. An expert rule gate applies eligibility, threshold, sequencing, and safety constraints, and a contextual-bandit ranker prioritizes eligible actions from the student context. Layer 4 is guided AI coaching. The coaching orchestrator converts the selected intervention

plan into a bounded prompt and exposes only approved tools and retrieval functions. Layer 5 records interaction outcomes, including completion, recommendation clicks, ratings, confidence checks, and subsequent metric gains, so that the student profile and bandit parameters can be updated.

This architecture connects the AI intervention suggestion system and the main coaching LLM as one decision-to-delivery framework. The intervention layer determines the educational action to be attempted; the LLM tool-calling layer delivers that action through progress-aware and resource-grounded interactions. The selected plan shapes the coach's allowable actions, and the coach's persisted artifacts become measurable outcome evidence for profile and policy updates.

C. Backend Evidence and Student Profile

Each student profile is represented as a three-layer metric snapshot:

$$S = \{M_{sub}, M_{comp}, M_{out}\}, \tag{1}$$

where M_{sub} contains 27 atomic backend submetrics, M_{comp} contains model-level control and explanatory constructs, and M_{out} contains outcome fields reserved for evaluation. Each metric entry stores a value, status, evidence count, and source. Values are normalized to a 0–100 scale. Status can be observed, inferred, missing, or not applicable. Missing measurements are not treated as zero; for decision making, unavailable features receive a neutral value and an explicit availability mask.

The dependent variable is final exam readiness:

$$FER = 0.5F_{mcq} + 0.5F_{essay}, \tag{2}$$

where F_{mcq} and F_{essay} are final unseen MCQ and essay/structured scores. The control variable is initial academic preparedness:

$$IAP = \frac{PK_{norm} + BER}{2}, \tag{3}$$

where IAP is initial academic preparedness, PK_{norm} is normalized prior Biology knowledge, and BER is baseline mock exam readiness before the intervention window.

Six student-varying explanatory constructs are used with this control variable: practice assessment performance, mastery and misconception diagnosis, cognitive skill performance, learning engagement and path progress, AI-powered coaching interaction and personalization, and incentive engagement.

D. System Components

The framework contains five connected operational components. First, the assessment module delivers MCQ, essay, and structured Biology practice with attempt timing, pass/fail status, Bloom-level cognitive tags [16], topic tags, and difficulty metadata. Second, the progress analytics module aggregates unit mastery, topic mastery, Bloom-level performance, repeated mistakes, assessment-pack progress, and student metrics such as streaks and AI coaching turns. Third, the intervention engine selects the next action using expert rules and adaptive bandit evidence. Fourth, the LLM coaching layer receives that selected action as bounded context and delivers it using controlled tools such as progress lookup, unit-material search, quiz recommendation, confidence recording, badge progress lookup, and study-plan creation. Fifth, the incentive module supports badges, badge progress, rewards, and learning streaks.

TABLE I
Student-Level Constructs and Measurement Sources

Construct	Submetrics
Initial academic preparedness	Prior Biology knowledge, baseline exam readiness
Practice performance	MCQ performance, essay performance, pass rate, timed efficiency
Mastery diagnosis	Unit mastery, topic mastery, repeated mistake reduction
Cognitive skill	Lower-order Bloom performance, higher-order Bloom performance, difficulty-adjusted performance
Engagement and path progress	Practice engagement, revision behaviour, learning consistency, assessment-pack progress
AI coaching personalization	Coaching turns, progress-aware tool use, resource-grounded tool use, recommendation uptake, message helpfulness, study-plan completion, topic confidence, AI marking helpfulness
Incentive engagement	Reward achievement, badge-tier progress, streak maintenance

IV. Hybrid Intervention Model

A. Intervention Catalogue

The intervention catalogue contains 22 policy actions grouped into assessment remediation, mastery and misconception support, AI coaching personalization, engagement and path progress, and incentive support. The implemented policy includes failed quiz remediation, targeted MCQ practice, higher-order essay practice, timed exam simulation, RAG-grounded misconception explanation, foundational concept review, progress-aware coaching check-in, short AI study plan, study-plan repair or simplification, topic confidence check, AI marking feedback review, recommendation uptake prompt, revision behaviour prompt, assessment-pack progress nudge, practice engagement activation, learning consistency nudge, badge achievement prompt, badge-tier progress prompt, streak maintenance prompt, challenge escalation, balanced mixed practice, and feedback-rating prompt.

Each intervention is linked to target profile features, eligibility conditions, reason codes, and delivery mechanisms. For example, weak topic mastery combined with repeated mistakes prioritizes a RAG-grounded misconception explanation, followed by targeted MCQ practice and a topic confidence check. Weak higher-order essay

performance prioritizes AI marking feedback review and higher-order essay practice. Low practice engagement and low consistency prioritize practice activation, consistency nudging, and a short AI study plan. This means the system selects a bounded intervention plan rather than a single generic message.

B. Expert Rule Gate and Cold Start

At the beginning of the study, insufficient intervention outcome data are available for a fully learned policy. The policy therefore begins with expert-gated cold-start behavior. Each feature is interpreted through four bands: critical (0–39), low (40–59), moderate (60–74), and strong (75–100). The expert gate first decides whether an intervention is eligible under the policy's threshold and safety rules. Only eligible interventions are ranked unless no eligible intervention exists, in which case the scorer exposes a fallback decision for diagnosis.

For an intervention a, the weighted expert score is computed over observed target features:

$$E(a,x) = \frac{\sum_{j \in O_a} w_{a,j} T_{a,j}(x_j)}{\sum_{j \in O_a} w_{a,j}}, \tag{4}$$

where O_a is the set of observed features used by intervention a, $w_{a,j}$ is an expert policy weight, and $T_{a,j}$ is one of three transforms. The deficit transform is

$$D(x_j) = \max(0, \min(1, (75 - x_j)/75)), \tag{5}$$

the strength transform is

$$S(x_j) = \max(0, \min(1, (x_j - 60)/40)), \tag{6}$$

and the balanced transform rewards moderate profiles near a target band. The score is then reduced by fatigue penalties for repeating the same intervention, repeating an intervention within the last seven days, or reselecting an intervention the student recently ignored. The rule layer also emits reason codes and sequencing hints, allowing the selected plan to be explained and audited.

C. Contextual Bandit Adaptation

After intervention outcomes are observed, the system uses Bayesian Linear Thompson Sampling to adapt rankings. The decision vector includes normalized feature values and availability masks. For intervention-specific learning, features are transformed in the same direction as the intervention's expert weights. Thus, a remediation intervention learns from deficits rather than from high raw performance values, reducing inappropriate generalization from one student profile to another.

The adaptive policy version uses expert-only ranking before any bandit update:

$$\alpha = 1.0, \quad \beta = 0.0. \tag{7}$$

Once outcomes exist, rankings blend expert and sampled bandit evidence:

$$Score(a) = \alpha E'(a,x) + \beta B(a,x), \tag{8}$$

where $E'(a,x)$ is the fatigue-adjusted expert score and $B(a,x)$ is the Thompson-sampled bandit score. The implemented adaptive weights are $\alpha = 0.60$ and $\beta = 0.40$ after the cold start. The selected sequence contains up to three interventions, with the primary intervention used as the first delivery target and the remaining interventions available as a bounded follow-on plan.

D. Reward Design

The bandit reward is based on short-term and medium-term evidence:

$$\begin{aligned} R =& 0.30G_t + 0.25G_r + 0.20C + \\ & 0.15G_e + 0.10H, \end{aligned} \tag{9}$$

where G_t is target-feature gain, G_r is predicted readiness gain, C is completion success, G_e is engagement gain, and H is student helpfulness feedback. Completion can distinguish ignored, viewed, started, and completed interventions. When final exam readiness is not yet available, predicted readiness gain is estimated from positive movement in beneficial decision features. This preserves final unseen exam readiness for evaluation while still giving the adaptive policy a practical online reward signal.

V. Guided LLM Coaching Layer

The LLM layer transforms selected interventions into student-facing coaching while remaining constrained by policy decisions and available tools. A direct LLM-to-student design would allow the model to decide content, difficulty, motivational strategy, and next action in one opaque step. In contrast, this framework connects the intervention decision output directly to the coaching orchestrator: the policy selects the plan, and the LLM executes that plan through approved tools and grounded explanations.

The coaching service uses a bounded tool loop. Its system instructions require age-appropriate Biology coaching, resource-grounded explanations when uploaded materials are available, and avoidance of hidden quiz answer keys. The available tools include listing unit topics, retrieving student progress, searching unit-material chunks, listing available quizzes, recommending quizzes, reading badge progress, finding badge opportunities, creating or updating study plans, marking study-plan steps as completed, reading topic-confidence history, and recording topic confidence. The tool loop is limited to a small number of rounds, and tool argument errors are returned to the model for correction rather than silently failing.

This tool-calling layer is the delivery side of the same intervention system. If the selected plan is misconception remediation, the coach searches unit materials and produces a cited explanation. If the plan is practice remediation, the coach recommends an MCQ or essay quiz without starting the attempt. If the plan is planning support, it creates or repairs a short study plan. If the plan is confidence monitoring, it asks for and records a 1–5 topic confidence score. If the plan is incentive support, it

retrieves badge progress or next badge opportunities and ties them to a learning action. The generated response is therefore progress-aware, resource-grounded, intervention-aligned, and measurable.

Assistant messages persist citations, recommendation cards, tool-call summaries, ratings, created study plans, and topic-confidence records. Recommendation clicks are recorded as separate events. These artifacts are used by the research metric builder to compute progress-aware coaching use, resource-grounded coaching use, recommendation uptake, message helpfulness, study-plan completion, topic confidence, and AI marking helpfulness. Thus, the LLM interaction is not merely conversational output; it becomes part of the measured learning system.

VI. Evaluation Design and Results

A. Participants and Attrition

The evaluation used a registered sample frame of 167 Grade 13 English-medium Biology students. The attrition flow is shown in Table II. Personally identifying fields are excluded from analysis exports.

TABLE II
Evaluation Attrition Flow

Stage	n
Registered	167
Enrolled	160
Completed baseline	154
Sufficient activity	143
Completed final test	132
Complete paired records	132

B. Study Design

The evaluation follows a single-group pre-experimental/quasi-experimental design [18]. The baseline window measures prior Biology knowledge and baseline exam readiness before students engage with the intervention activities. The intervention window measures practice, mastery, engagement, AI coaching, and incentive behavior. The final window measures unseen MCQ and essay/structured final exam readiness. The main improvement measure is:

$$\Delta ER = FER - BER. \tag{10}$$

The explanatory model used to evaluate the hypotheses is:

$$\begin{aligned} FER =& \beta_0 + \beta_c IAP + \beta_1 PAP + \beta_2 MMD + \beta_3 CSP \\ &+ \beta_4 LEP + \beta_5 AIC + \beta_6 IE + \epsilon, \end{aligned} \tag{11}$$

where FER is final exam readiness, IAP is initial academic preparedness, PAP is practice assessment performance, MMD is mastery and misconception diagnosis, CSP is cognitive skill performance, LEP is learning engagement and path progress, AIC is AI coaching interaction and personalization, and IE is incentive engagement. The model estimates the individual contribution of each independent variable to final exam readiness after controlling for initial academic preparedness. The 27 submetrics are inspected descriptively and through bivariate correlations, but they are not entered together into the regression model to avoid overfitting with $n = 132$.

TABLE III
Descriptive Evaluation Metrics on 0–100 Scale ($n = 132$)

Metric	Mean	SD
Baseline exam readiness	52.09	11.61
Final exam readiness	58.90	12.10
Readiness gain	6.81	12.46
Initial academic preparedness	52.98	10.70
Practice assessment performance	61.79	12.36
Mastery and misconception diagnosis	62.69	11.79
Cognitive skill performance	59.38	12.89
Learning engagement and path progress	58.58	14.55
AI coaching interaction and personalization	55.84	14.98
Incentive engagement	49.76	16.25

C. Evaluation Results

The paired analysis indicates a moderate improvement in readiness, from 52.09 at baseline to 58.90 at final assessment. The mean gain is 6.81 points. With $n = 132$ complete paired records, the paired-sample statistic is $t(131) = 6.27$, $p < .001$, with $d_z = 0.55$ and a 95% confidence interval of [4.66, 8.95] points. This is consistent with an educationally meaningful medium effect while avoiding the unsupported claim that the framework produces a very large effect in a single-group pilot.

Table IV shows bivariate correlations with final exam readiness. Initial academic preparedness is included as the control variable. Among the six independent variables, practice performance has the strongest association with final readiness, followed by mastery diagnosis, cognitive skill performance, engagement, AI coaching, and incentive engagement. The pattern is consistent with the framework's assumption that AI coaching is most useful when connected to practice, progress awareness, and sustained engagement rather than treated as a standalone chatbot variable.

TABLE IV
Construct-Level Correlations with Final Readiness

Construct	r
Initial academic preparedness	.58
Practice assessment performance	.49
Mastery and misconception diagnosis	.46
Cognitive skill performance	.43
Learning engagement and path progress	.40
AI coaching interaction and personalization	.39
Incentive engagement	.28

The OLS model included initial academic preparedness as the control variable and the six independent variables as explanatory constructs. The model produced $R^2 = .49$ and adjusted $R^2 = .46$. Standardized coefficients, significance levels, and variance inflation factors are shown in Table V. Initial academic preparedness remained the strongest positive predictor. Among the six framework-related independent variables, practice assessment performance and AI coaching interaction and personalization reached the .05 level, supporting $H_{1,a}$ and $H_{5,a}$. Mastery and misconception diagnosis, cognitive skill performance, learning engagement and path progress, and incentive engagement also showed positive coefficients, providing directional support for $H_{2,a}$, $H_{3,a}$, $H_{4,a}$, and $H_{6,a}$, although these effects did not reach the .05 significance level in the full model.

TABLE V
Regression Summary

Predictor	β	p	VIF
Initial academic preparedness	.30	.001	1.94
Practice assessment performance	.17	.026	1.43
Mastery and misconception diagnosis	.14	.080	1.55
Cognitive skill performance	.10	.234	1.69
Learning engagement and path progress	.13	.081	1.36
AI coaching interaction and personalization	.16	.042	1.42
Incentive engagement	.05	.510	1.33

VII. Ethical, Privacy, and Safety Considerations

The framework handles secondary-level learner data and therefore requires strict privacy and safety controls. Student identifiers should be de-identified in research exports, and personally identifying fields such as names, addresses, dates of birth, and phone numbers should not be included in analysis datasets. The intervention policy excludes final outcome fields from selection inputs to prevent target leakage. Missing values are explicitly represented to avoid unfairly penalizing students who did not receive a measurement opportunity.

LLM safety is addressed through guided interaction. The LLM does not independently choose interventions, expose unrestricted tools, or generate ungrounded Biology explanations when unit material is required. Resource-grounded answers include citations to processed learning material. The coach is instructed not to reveal hidden quiz answer keys and not to start quiz attempts on behalf of the student. Human review remains necessary for content quality, assessment validity, and ethical oversight, especially when AI marking or motivational prompts influence student behavior.

VIII. Limitations

The main limitation of the current evaluation is the single-group design, which can establish pre/post improvement but cannot fully isolate causal effects from maturation, external tuition, or concurrent study.

IX. Future Work

Future work should first extend the evaluation with a controlled comparison group so that the framework's contribution can be estimated more strongly than is possible in the current single-group pre/post design. Longer follow-up is also needed to examine whether readiness gains persist across multiple Biology units, different question formats, and longer preparation windows. The intervention layer should be evaluated with richer intervention-outcome logs, including whether recommended actions were viewed, accepted, completed, ignored, or rated as helpful. These data would allow the contextual bandit policy to move beyond expert-gated cold-start behavior toward more personalized ranking for different learner profiles.

Further development should broaden the content coverage beyond the Microbiology unit and add teacher-facing review workflows for validating AI-generated explanations, recommendations, study plans, and AI marking feedback. Before wider deployment in secondary education, future studies should also examine consent, de-identification, data minimization, audit logging, access control, and safe LLM tool-use controls in realistic school settings.

X. Conclusion

This paper presented an AI-powered framework for personalized Biology coaching and adaptive incentive delivery in competitive examination preparation. The core design connects intervention selection and LLM response generation in one bounded coaching pipeline. Expert educational rules provide safe eligibility gates and cold-start ranking, a contextual bandit adapts intervention priorities from observed outcomes, and the LLM delivers the selected intervention through approved tools and resource-grounded coaching actions. The evaluation model uses initial academic preparedness as the control variable and backend-aligned metrics for practice, mastery, cognitive performance, engagement, AI coaching, incentives, and final exam readiness. The evaluation results show moderate readiness improvement and indicate that AI coaching and engagement are meaningful mechanisms when connected to progress evidence.

References

[1] A. Létourneau, M. Deslandes Martineau, P. Charland, J. A. Karran, J. Boasen, and P. M. Léger, "A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education," npj Science of Learning, vol. 10, no. 29, 2025, doi: 10.1038/s41539-025-00320-7.

[2] K. Bayly-Castaneda, M.-S. Ramirez-Montoya, and A. Morita-Alexander, "Crafting personalized learning paths with AI for lifelong learning: A systematic literature review," Frontiers in Education, vol. 9, Art. no. 1424386, 2024, doi: 10.3389/feduc.2024.1424386.

[3] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, Art. no. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.
[4] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: Systematic literature review," International Journal of Educational Technology in Higher Education, vol. 20, no. 56, 2023, doi: 10.1186/s41239-023-00426-1.
[5] L. Yan et al., "Practical and ethical challenges of large language models in education: A systematic scoping review," British Journal of Educational Technology, vol. 55, no. 1, pp. 90–112, 2024, doi: 10.1111/bjet.13370.
[6] T. Zhang, Y. C. Lai, and L. H. P. Yu, "Generative artificial intelligence in K–12 education: A systematic review," Research and Practice in Technology Enhanced Learning, vol. 21, no. 34, 2026, doi: 10.58459/rptel.2026.21034.
[7] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2023.
[8] O. Henkel, Z. Levonian, M.-E. Postle, and C. Li, "Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference," in Proc. 17th International Conference on Educational Data Mining, 2024.
[9] S. S. Lee and R. L. Moore, "Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review," Online Learning, vol. 28, no. 3, 2024, doi: 10.24059/olj.v28i3.4593.
[10] A. S. Lan and R. G. Baraniuk, "A contextual bandits framework for personalized learning action selection," in Proc. 9th International Conference on Educational Data Mining, 2016, pp. 424–429.
[11] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in Proc. 30th International Conference on Machine Learning, 2013, pp. 127–135.
[12] J. Hattie and H. Timperley, "The power of feedback," Review of Educational Research, vol. 77, no. 1, pp. 81–112, 2007, doi: 10.3102/003465430298487.
[13] B. J. Zimmerman, "Becoming a self-regulated learner: An overview," Theory Into Practice, vol. 41, no. 2, pp. 64–70, 2002, doi: 10.1207/S15430421TIP4102_2.
[14] E. L. Deci and R. M. Ryan, "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior," Psychological Inquiry, vol. 11, no. 4, pp. 227–268, 2000, doi: 10.1207/S15327965PLI1104_01.
[15] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining gamification," in Proc. 15th International Academic MindTrek Conference, 2011, pp. 9–15, doi: 10.1145/2181037.2181040.
[16] D. R. Krathwohl, "A revision of Bloom's taxonomy: An overview," Theory Into Practice, vol. 41, no. 4, pp. 212–218, 2002, doi: 10.1207/s15430421tip4104_2.
[17] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.
[18] A. D. Harris, J. C. McGregor, E. N. Perencevich, J. P. Furuno, J. Zhu, D. E. Peterson, and J. Finkelstein, "The use and interpretation of quasi-experimental studies in medical informatics," Journal of the American Medical Informatics Association, vol. 13, no. 1, pp. 16–23, 2006, doi: 10.1197/jamia.M1749.