Journal of Contemporary Educational Research, 2026, Volume 10, Issue 7
http://ojs.bbwpublisher.com/index.php/JCER
Online ISSN: 2208-8474
Print ISSN: 2208-8466

Designing a Multi-Agent Personalized Learning Support System Grounded in Mastery Learning Theory

Chenwei Liu, Linshuo Xiao

Fujian Normal University, Fuzhou 350007, Fujian, China

Abstract: Generative artificial intelligence offers new ways to support personalized learning, yet many existing applications emphasize content generation without fully connecting learning evidence, instructional intervention, and subsequent evaluation. Adopting a design and development research approach, this study designs a multi-agent personalized learning support system with mastery learning theory as its instructional framework. Within the bounded context of a learning space, specialized agents share data and coordinate tasks to diagnose mastery from formative assessment evidence, generate corrective or enrichment support, and prepare a second parallel formative assessment after teacher review. This process creates an interpretable and traceable closed loop of personalized learning support. The study further derives five design principles and develops a learning-evidence model, a multi-agent collaboration architecture, personalized learning packages, and a staged evaluation framework. Its principal contribution is the translation of mastery learning into an executable and auditable multi-agent workflow, offering design guidance for personalized learning support systems. The proposed design and prototype establish feasibility claims only; their educational effectiveness remains to be examined empirically.

Keywords: Formative assessment; Generative artificial intelligence; Learning analytics; Teacher-in-the-loop; Design and development research

Online publication: August 5, 2026

1. Introduction

Learning management systems support a broad range of instructional, learning, and assessment activities [1]. The learner-interaction data generated within these environments can be analyzed to identify learning needs and inform personalized support [2]. Analyzing these data can help identify learning needs and inform personalized instructional support [2]. Nevertheless, the data often remain at the level of recording and visualization and are not readily translated into teachers' subsequent instructional decisions. Existing research indicates that some learning analytics systems continue to prioritize outcome displays, provide few actionable learning strategies grounded in learning theory, and employ limited evaluation designs [3]. In practice, learning data—such as assignment submissions, quiz responses, and error records—and related instructional resources are often dispersed across separate platform modules, leaving teachers to integrate the evidence and determine what support each student needs next.

Generative artificial intelligence (GenAI) and large language model (LLM)-based multi-agent systems provide technical affordances for content generation, feedback, and task coordination in personalized learning [4–8]. However, without explicit learning objectives, assessment evidence, and boundaries of responsibility, AI outputs may be disconnected from students' actual learning difficulties and may introduce factual errors, overreliance, bias, and privacy risks [5,9,10].

The challenge, therefore, is not merely to accumulate learning data or generate personalized content. It is to organize formative assessment, learning diagnosis, differentiated support, teacher review, and subsequent assessment as a coherent process guided by a unified instructional theory. Mastery learning provides such a logic through explicit objectives, formative assessment, corrective or enrichment activities, and a second parallel formative assessment [11–13].

Accordingly, this study designs a multi-agent personalized learning support system using mastery learning theory as its instructional framework. Through differentiated agent roles, coordinated support, and data sharing, the system organizes formative assessment evidence, teacher review, and a second parallel formative assessment into an interpretable and traceable closed loop. The study addresses three research questions:

1. How can mastery learning theory be translated into design principles and functional requirements for a multi-agent personalized learning support system?
2. How can multi-agent collaboration transform formative assessment evidence into a mastery diagnosis and corrective or enrichment support?
3. How can teacher review and a second parallel formative assessment be incorporated into the system workflow, and how should the system's technical feasibility, instructional quality, and usability be evaluated?

2. Literature review and theoretical foundation

2.1. Multi-agent systems for personalized learning

Personalized learning adapts learning content, support, or pathways to differences in learners' prior knowledge, pace, and performance, thereby improving the fit between instructional support and learning needs [14]. GenAI can summarize information, explain content, generate practice activities, and provide feedback, creating new technical opportunities for personalization [5–7]. Multi-agent systems can further divide roles, coordinate tasks, and share data to analyze learning states, diagnose mastery, generate resources, match tasks, and check output quality. These capabilities are well suited to personalized learning processes that require coordination across multiple stages and differentiated support [4]. For example, Vaccaro *et al.* generated personalized learning materials through collaboration between a learner-profile agent and a content-rewriting agent, demonstrating a practical use of multi-agent personalization.

Multi-agent collaboration, however, does not guarantee accurate diagnoses or generated content. GenAI may still produce factual errors, bias, overreliance, and privacy risks [5,9,10]. Although AI can support needs identification, resource adaptation, feedback generation, and pathway adjustment, its use must remain aligned with explicit instructional objectives and professional teacher judgment [15]. Teachers should therefore review AI outputs within the curriculum and classroom context and retain authority over instructional decisions [16,17].

2.2. Mastery learning theory

Mastery learning holds that when learning objectives and mastery criteria are explicit, students can reach high levels of achievement through varying amounts of time, feedback, and support. Failure to meet a criterion on the first attempt should not be treated as evidence of fixed inability [11,12]. The basic process comprises clarifying objectives, providing instruction, administering a formative assessment, offering corrective or enrichment activities, and reassessing mastery through a second parallel formative assessment [12,13].

Within this process, formative assessment identifies specific mastery gaps rather than merely recording scores. Corrective activities should address the identified difficulty by changing the explanation, learning material, or scaffold instead of simply repeating the original instruction [12,13]. Feedback should clarify the learning goal, current performance, and the next step for improvement [18,19]. Students who have not yet met the criterion receive corrective support, whereas those who have met it proceed to enrichment activities or the next objective.

Mastery learning thus organizes learning objectives, formative assessment, differentiated support, and reassessment into a continuous process. It can specify the sequence of tasks, the evidence required as input, and the criteria for judging outcomes within a multi-agent system. For these reasons, mastery learning serves as the central theoretical foundation for the system designed in this study.

3. Research design and methods

3.1. Research approach

This study adopts a design and development research approach focused on the analysis, design, development, and evaluation of educational products and tools [20]. It also draws on design science research, which emphasizes artifacts and design-knowledge contributions [21]. and on design-based research, which stresses implementation and iterative refinement in authentic settings [22,23]. Because the present study has not yet undertaken multiple iterations in an authentic classroom, it is not presented as a completed design-based research study.

3.2. Research process

The research proceeded through five stages. First, studies of personalized learning, multi-agent systems, teacher involvement, and mastery learning were reviewed to identify educational needs and theoretical requirements. Second, the functions, data flows, and permission relationships of an existing instructional platform were examined to identify fragmented learning data, disconnections between assessment and support, and unclear responsibility for AI outputs. Third, theoretical requirements were mapped to platform problems to derive system requirements and design principles. Fourth, these principles guided the design of the personalized learning support loop, formative assessment evidence model, multi-agent collaboration mechanism, personalized learning package, and teacher-review process; a prototype specification was then developed on the existing platform. Fifth, a staged evaluation plan was formulated to examine technical feasibility, instructional content quality, teacher usability, the incremental value of the multi-agent design, and classroom application.

3.3. Design basis and analytical method

The system design drew on three sources. The first comprised findings from personalized learning, multi-agent, and teacher–AI complementarity research regarding support needs, technical capabilities, and governance boundaries. The second was the instructional sequence prescribed by mastery learning: objectives, formative assessment, corrective or enrichment activities, and a second parallel formative assessment. The third was the existing platform's practical foundation in learning spaces, quizzes, question banks, resources, permissions, and historical records.

A structured mapping procedure was used. Core elements of mastery learning were first translated into observable instructional requirements. Platform problems were then categorized as evidence, process, content quality, authority and responsibility, or outcome confirmation. Finally, a mapping was constructed from "theoretical and research basis" through "design problem" and "system requirement" to "corresponding function." Only functions traceable to an explicit theoretical requirement or practical problem were retained in the core system proposal. This procedure provided an auditable rationale for the subsequent design principles and system modules.

3.4. Prototype development and functional inspection

The prototype extends an existing web-based instructional platform. The platform already supports role-based registration; the creation of and enrollment in learning spaces; notices and assignments; online quizzes; automatic scoring of objective items; manual scoring of constructed responses and attachment-based answers; question banks; learning resources; personalized tasks; review of incorrectly answered questions; quiz autosave; multiple attempts; and authorized access to attachments. These functions provide the technical basis for aggregating formative assessment evidence, publishing learning packages, and enforcing teacher permissions.

The proposed multi-agent workflow adds formative assessment evidence snapshots, agent input–output contracts, shared state, quality gates, teacher review, and a second parallel formative assessment. Functional inspection uses predefined learning spaces, role permissions, and anonymized cases to check data retrieval, state transitions, structured outputs, permission validation, and exception handling. This inspection is intended only to establish the feasibility of the workflow and technical design; it does not constitute evidence of instructional effectiveness.

3.5. Research scope and ethics

No new data from human participants were collected during the design and development stage reported in this paper. Future expert evaluations, teacher usability tests, and classroom studies should receive the appropriate ethics approval or exemption. Depending on participants' ages, informed consent and parental or guardian consent should be obtained. Claims in the present paper are limited to design rationale, prototype implementability, and the proposed evaluation plan; no causal claims are made regarding achievement, engagement, or teacher workload.

4. Design requirements and principles

4.1. Problem and requirements analysis

Drawing on the literature and theoretical analysis and on a review of the existing platform's functions, data flows, and permissions, five design problems were identified and translated into system requirements (**Table 1**).

Table 1. Problem–requirement mapping

Identified design problem	Instructional and system implication	System design requirement
Learning data and related resources are dispersed across modules.	A contextually bounded basis for personalized support is difficult to establish.	Aggregate formative assessment evidence within a defined learning space, mastery objective, and analysis period.
Scores and error records cannot be treated as mastery diagnoses in themselves.	A model may infer causes from insufficient evidence.	Link objectives, knowledge components, and items; calculate transparent indicators before generating a diagnostic interpretation.
Assessment, support, and reassessment lack a common sequence.	AI generation may deteriorate into resource recommendations detached from the instructional process.	Organize objectives, formative assessment, diagnosis, corrective or enrichment support, and reassessment according to mastery learning.
Generated content may be inaccurate or misaligned with curricular objectives.	Direct publication without review creates instructional and safety risks.	Apply output-quality checks and teacher review; retain unconfirmed content as drafts.
Task completion does not demonstrate mastery.	The effect of support lacks new assessment evidence.	Administer a second parallel formative assessment and retain pre- and post-support evidence and state transitions.

4.2. Theory–requirement–function mapping

To prevent system functions from being driven directly by technical affordances, mastery learning, personalized learning research, and the teacher-in-the-loop principle were mapped to specific system requirements and functions (**Table 2**).

Table 2. Theory–requirement–function mapping

Theoretical or research basis	System requirement	Corresponding function
Explicit learning objectives and mastery criteria	Organize diagnosis, resources, tasks, and assessment around the same objective.	Management of mastery objectives, knowledge components, and criteria.
Formative assessment for diagnosis	Aggregate objective-relevant evidence before judging mastery.	Formative assessment evidence snapshots and transparent indicators.
Parallel corrective and enrichment routes	Differentiate support by current mastery status.	Corrective or enrichment packages, resources, and item generation.
Feedback addressing goals, current performance, and next actions	State the diagnostic basis and provide actionable next steps.	Evidence citations, mastery-gap statements, and support rationales.
A second parallel formative assessment	Collect new evidence on the same objective after support.	Parallel assessment, status updating, and continuation rules.
Professional teacher judgment and responsibility	Ensure that AI outputs can be understood, revised, and controlled before publication.	Teacher-review workspace, version records, and publication permissions.

4.3. System design principles

The mapping yielded five theory-driven and operational design principles. P1–P4 derive primarily from mastery learning, whereas P5 integrates multi-agent collaboration, teacher–AI complementarity, and AI governance in education (**Table 3**).

Table 3. Design principles for the multi-agent personalized learning support system

Design principle	Primary basis	Operational meaning	Intended contribution
P1. Specify mastery objectives	Mastery learning	Bind every diagnosis, learning package, and assessment to a knowledge component, learning objective, and teacher-defined mastery criterion.	Align instruction, generation, and assessment.
P2. Diagnose from formative evidence first	Mastery learning; learning evidence	Aggregate formative assessment evidence before generating support and apply transparent rules to determine an initial mastery status.	Avoid treating a total score or model conjecture as a diagnosis.
P3. Provide corrective and enrichment routes	Mastery learning; personalized learning	Give targeted correction to students below the criterion and enrichment or the next objective to students who meet it.	Match support to the learner's current mastery status.
P4. Use a second parallel formative assessment	Mastery learning	Reassess with different items that target the same objective and have comparable cognitive demands and difficulty.	Distinguish task completion from mastery.
P5. Coordinate agents under teacher supervision	Multi-agent systems; teacher-in-the-loop	Exchange structured agent outputs and require teacher review of diagnoses, materials, and assessments.	Retain instructional responsibility and process traceability.

5. System design

5.1. A personalized support loop grounded in mastery learning

The process begins with teacher-defined learning objectives and mastery criteria. After instruction and the first formative assessment, the platform aggregates relevant performance data within the designated learning space and analysis period to create an assessment-evidence snapshot. Transparent, teacher-configured rules produce an initial mastery status. The agents then generate a draft mastery diagnosis and a corrective or enrichment learning package. After the teacher reviews the diagnosis, resources, and items, students complete the package and take a second parallel formative assessment of the same mastery objective. Students who meet the criterion proceed to the next objective; those who do not return to diagnosis and correction [11–13]. See **Figure 1**.

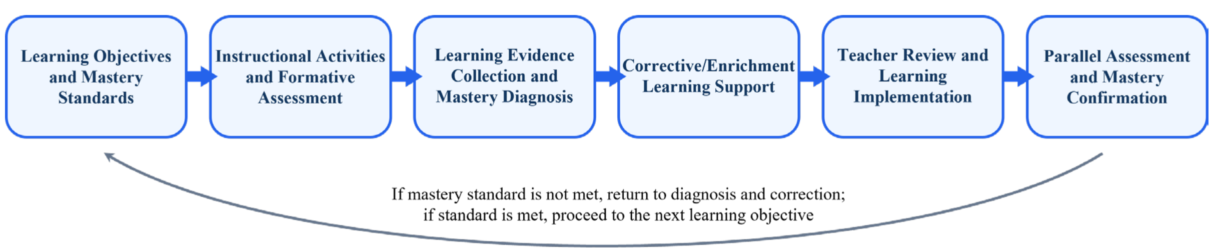

Figure 1. Personalized learning support loop grounded in mastery learning theory

The loop does not begin with AI generation; it is driven by objectives and formative assessment evidence. Interpretability means that a diagnosis can be traced to a specific mastery objective, item, and performance indicator. Traceability means that agent processing, teacher revisions, task publication, student completion, and reassessment are recorded as state transitions. The instructional workflow consequently constrains the evidence available to the AI, the purpose of its outputs, and the subsequent test of those outputs.

5.2. Learning spaces and the formative assessment evidence model

A learning space is a digital instructional unit created by a teacher for a particular group and set of tasks.

It has relatively independent membership, objectives, learning activities, data boundaries, and permission structures. Research on context-aware learning recommendations suggests that learning objectives, knowledge domains, activities, and use contexts influence the appropriateness of support [24]. Accordingly, the system uses only data from a specified learning space and analysis period when determining a learner's state, avoiding unbounded combinations of performance across courses or stages.

Assignment submissions, quiz responses, and error records become formative assessment evidence only after they are linked to objectives, knowledge components, and assessment tasks. Each item may be associated with one or more knowledge components and tagged with a primary component, cognitive demand, and difficulty. Historical quiz items are preserved as immutable snapshots so that later edits to the question bank cannot alter the evidence underlying an earlier diagnosis.

A learner-state snapshot is a time-stamped object for stage-specific analysis rather than a permanent ability label. It contains the mastery objective, evidence period, relevant activities, valid learning opportunities, performance by knowledge component, repeated errors, task-completion status, data recency, and confidence. Initial states may include *insufficient evidence*, *support needed*, *developing*, and *currently mastered*. This design follows a central principle of learner modeling and knowledge tracing: observable performance provides evidence about a latent knowledge state but does not constitute an absolute determination [25,26]. An LLM may summarize the snapshot and propose candidate explanations, but it may not overwrite computed evidence or invent reasons that are absent from the data.

5.3. Multi-agent collaboration mechanism

A coordinator agent manages task routing among specialized agents. Rather than relying on unrestricted conversation histories, each agent reads a typed task object and writes a structured result. The shared state records evidence citations, generated drafts, validation results, and teacher decisions. Missing required fields, invalid permissions, or failed quality checks prevent the workflow from advancing. Thus, the multi-agent design is not a collection of role prompts; it is a collaboration mechanism composed of task delegation, information sharing, state transitions, and quality gates. See **Figure 2** and **Table 4**.

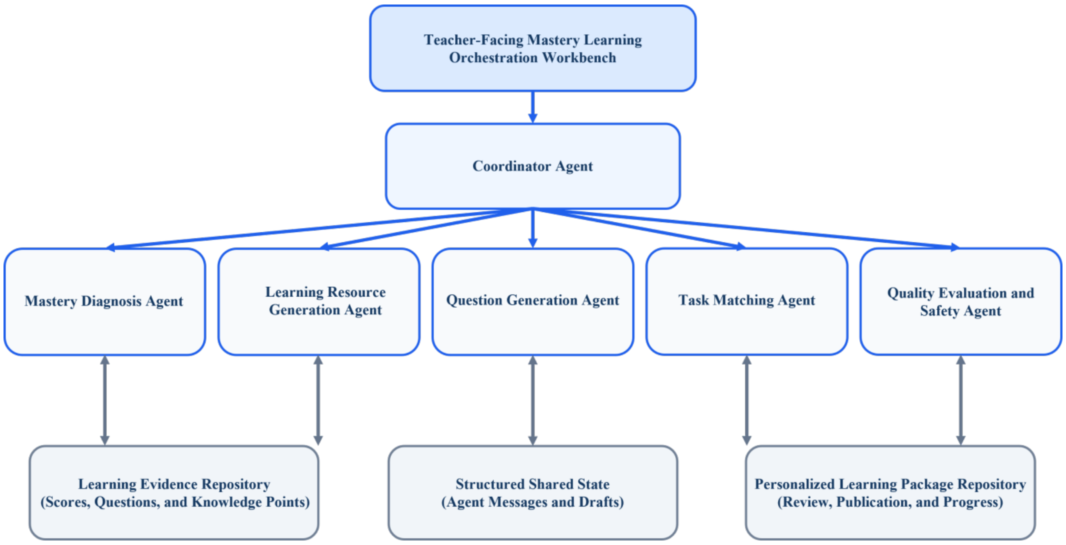

Figure 2. Mastery-learning-oriented multi-agent system architecture

Table 4. Agent roles, inputs, outputs, and principal constraints

Agent	Primary input	Structured output	Principal constraint
Coordinator agent	Teacher-defined objectives, mastery criteria, learning-space permissions, and snapshot identifier	Execution plan, task states, draft index, and workflow log	Does not publish content; manages state transitions and permission checks.
Mastery-diagnosis agent	Formative assessment evidence snapshot and teacher-defined criteria	Mastered objectives, mastery gaps, evidence citations, and confidence	Must cite source indicators; must not infer sensitive attributes.
Learning-resource generation agent	Mastery gaps or enrichment objectives, grade level, and curricular constraints	Draft corrective explanations, examples, or enrichment resources	Must not reveal assessment answers; must remain aligned with the objective.
Item-generation agent	Objective, error patterns, difficulty, and assessment purpose	Scaffolded practice, transfer items, parallel assessment items, and explanations	Checks answer consistency and alignment across the two assessments.
Task-matching agent	Mastery status, approved materials, and teacher strategy	Corrective or enrichment pathway, assignment rationale, and completion rules	Must not use permanent ability labels; retains the matching rationale.
Quality-evaluation and safety agent	All drafts, mastery criteria, and generation constraints	Alignment flags, revision requests, and risk status	Blocks unsupported, unsafe, inconsistent, or misaligned content.

5.4. Personalized learning packages

The personalized learning package is the principal instructional vehicle for corrective or enrichment support. It reuses the existing personalized-task fields for learning space, target students, deadline, publication status, and completion record while adding ordered learning components, response records, and evidence from the second parallel formative assessment. The goal is not to maximize the amount of generated content but to organize a coherent sequence around one mastery objective: gap explanation, resource support, scaffolded practice and transfer, reassessment, and reflection and feedback. See **Table 5**.

Table 5. Recommended sequence of a personalized learning package

Learning step	Function for the student	System record
1. Explanation of the mastery gap	Explains what the formative evidence indicates and why the package was assigned.	Evidence snapshot and teacher-confirmed mastery judgment.
2. Mastery objective	States the intended learning outcome and mastery criterion.	Objective identifier, knowledge-component links, and criterion.
3. Corrective or enrichment resource	Provides an explanation, example, or extension different from the initial instruction.	Resource type, access record, and teacher revision history.
4. Scaffolded practice	Moves progressively from demonstration to guided practice.	Responses, number of attempts, and feedback events.
5. Transfer task	Tests whether the target knowledge can be applied in a different context.	Response result and explanation.
6. Second parallel formative assessment	Reassesses the same mastery objective with different items.	Post-support formative assessment evidence.
7. Reflection and feedback	Prompts the student to identify remaining difficulties and next-step needs.	Learning reflection, system feedback, and teacher follow-up.

5.5. Teacher review and publication

The teacher workspace presents formative assessment evidence, the machine-generated diagnosis, generated learning components, and publication controls together. Teachers can revise the wording of a mastery gap, remove irrelevant resources, edit items, answers, and explanations, adjust difficulty and the target group, and choose whether to continue revising, return a draft for regeneration, or publish it. The system records differences among the AI draft, teacher revisions, and final version. Review is therefore not a single confirmation click on final text; it spans evidence interpretation, mastery judgment, content revision, and the publication decision.

Learning-space permissions remain in force for AI-assisted functions. A collaborating teacher may generate or edit drafts only while holding valid management permission, and the learning-space creator determines whether that teacher can publish. Teachers retain final authority over mastery criteria, diagnostic confirmation, and instructional action.

5.6. Safety, privacy, and accountability

To operationalize transparency, accountability, and privacy protection in educational AI [10], the system uses identity authentication, learning-space permission checks, service-to-service keys, authorized attachment access, and model-call audits. External models receive only the de-identified information necessary for a specified task, such as anonymized performance by knowledge component, error types, and teacher-defined objectives. Names, student numbers, contact details, original attachments, and profile attributes unrelated to the task are not transmitted.

Each model call should record the model and prompt-template versions, task identifier, evidence citations, generation time, validation result, teacher-review action, and final publication status. Students can access only packages published specifically to them within an active learning space. These records make it possible to trace how the system moves from formative assessment evidence to a diagnosis and support plan while avoiding the unnecessary long-term retention of complete, sensitive prompts.

6. Prototype system implementation

6.1. Technical architecture

The existing platform uses a browser/server architecture. The web layer is implemented with ASP.NET Core Razor Pages and cookie authentication; the REST API uses ASP.NET Core Controllers; asynchronous ADO.NET services access Microsoft SQL Server; and server-side HTTP clients call an LLM endpoint compatible with the OpenAI API specification. Separating the web, API, and database layers allows the multi-agent coordination service to be added without restructuring existing business functions.

Structured JSON contracts define the boundaries between multi-agent modules. Each response includes the contract version, agent type, task identifier, evidence citations, output content, warnings, and validation status. The coordination service persists state transitions and execution records. Deterministic validators check required fields, item types, answer consistency, score ranges, learning-space membership, and direct personal identifiers. A draft that fails validation is returned to the relevant agent or forwarded to the teacher; it is never published automatically.

6.2. Core functions

The prototype focuses on four functions central to the proposed loop. First, it reads assignments, quizzes, and error records within a learning space and generates a formative assessment evidence snapshot based on objective–knowledge component–item links. Second, the coordination service sequentially invokes mastery diagnosis, resource generation, item generation, task matching, and quality checking while persisting a structured shared state. Third, validated content is assembled into a learning-package draft for teacher revision and publication. Fourth, after a student completes the package, the system records the second parallel formative assessment and updates the stage-specific mastery status.

The platform's learning spaces, question bank, quizzes, personalized tasks, and permission controls already provide an operational foundation. The multi-agent closed loop currently exists as a workflow prototype, data contracts, and interface specifications; full integration still requires functional testing, expert evaluation, and teacher usability testing.

6.3. Integration with the existing platform

The design prioritizes reuse of the existing platform rather than creating a separate AI tool detached from classroom workflow. Teachers continue to create learning spaces, manage students and collaborating teachers, publish notices and assignments, administer quizzes, and manage questions and resources. Students continue to submit assignments; take quizzes with autosave and multiple attempts; review results and errors; and complete assigned tasks. The new workflow reads historical records through dedicated queries and does not modify submitted assignments or immutable quiz snapshots. See **Table 6**.

Table 6. Existing platform foundation and extensions proposed in this study

Module	Existing platform foundation	Extension proposed in this study	Current status
Learning spaces and permissions	Members, roles, collaborating teachers, and access control	Boundaries for evidence, task ownership, and review permissions	Implemented and reused
Quizzes, question bank, and error records	Items, responses, scores, and historical snapshots	Objective–knowledge component links and formative assessment evidence snapshots	Foundation implemented; extension designed
Personalized tasks	Target students, deadlines, status, and completion records	Ordered learning packages and a second parallel formative assessment	Foundation implemented; extension designed
LLM calls	Server-side model interface	Agent contracts, coordinated states, and quality gates	Prototype design
Teacher management	Learning-space administration and collaborating-teacher permissions	Integrated evidence–diagnosis–content–publication review	Interface and workflow design

6.4. Illustrative use scenario

Consider a middle-school mathematics learning space on linear equations in one variable. The teacher defines the mastery objective as "extracting an equality relationship from a verbal situation and expressing it correctly as an equation," establishes a mastery criterion, and links formative quiz items to concept identification, equation formulation, equivalence transformations, and computational accuracy. After submission, the system creates an evidence snapshot of performance by knowledge component and repeated errors. If a student's calculations are mostly correct but the student repeatedly fails to formulate an equation from a verbal relationship, the mastery-diagnosis agent writes that interpretation and its evidence citations to a draft.

The resource- and item-generation agents then produce a relationship diagram, worked examples, scaffolded practice, and a transfer item. The task-matching agent organizes their sequence, and the quality-evaluation and safety agent checks objective alignment and answer correctness. After reviewing the package and adjusting its linguistic difficulty, the teacher publishes it. The student completes the support and then takes a second parallel formative assessment targeting the same objective through different items and contexts. The resulting evidence updates the mastery status and informs whether the student should receive further correction or proceed to the next objective.

7. Staged evaluation plan

7.1. Evaluation rationale

An operational system does not in itself establish instructional quality or educational value. Evaluation should progress through technical and workflow validation, instructional content evaluation, teacher use, assessment of the incremental value of the multi-agent design, and classroom application. Each stage should meet basic requirements before the next begins. This sequence evaluates both the artifact and the design claims advanced in this paper. See **Table 7**.

Table 7. Staged evaluation plan

Evaluation stage	Participants or materials	Principal indicators	Analysis
Technical and workflow validation	Predefined spaces, roles, permissions, and anonymized cases	Task success, schema validity, permission enforcement, log completeness, latency, and failure recovery	Pass/fail matrix and descriptive statistics
Instructional content evaluation	Educational technology experts, subject teachers, and anonymized packages	Accuracy, objective alignment, difficulty, scaffolding, interpretability, and safety	Descriptive statistics and ICC, kappa, or Kendall's W
Teacher usability evaluation	Teachers completing a diagnosis–review–publication task	Completion, time, operational errors, System Usability Scale, workload, and interview themes	Descriptive statistics and thematic analysis
Multi-agent comparative evaluation	Matched cases generated by single- and multi-agent conditions	Evidence fit, coherence, error rate, extent of revision, response time, and cost	Blinded paired comparisons and effect sizes
Classroom pilot	Voluntarily participating students and teachers	Knowledge-component performance, transfer, completion, user experience, and teacher workload	Pre–post or controlled analysis with limitations reported

7.2. Technical and workflow validation

Technical validation should cover both normal and exceptional workflows, including evidence-snapshot generation, agent-call order, structured-output validation, teacher permissions, publication control, second-assessment records, and status updating. Test cases should include insufficient evidence, invalid permissions, missing model-output fields, inconsistent item answers, and injected sensitive information, with records showing whether each faulty workflow was successfully blocked. Such results can support claims of technical feasibility and process integrity only.

7.3. Instructional content and teacher usability evaluation

Educational technology experts and subject teachers can independently rate anonymized learning packages. Evaluation should examine whether mastery diagnoses cite sufficient evidence, corrective resources address specific difficulties, enrichment goes beyond repeated practice, the second assessment remains aligned with the original objective, and the content contains factual or safety problems. Interrater agreement should be reported rather than relying solely on mean ratings.

For teacher usability evaluation, participants should complete the full sequence of reviewing evidence, interpreting the diagnosis, revising the package, and publishing the task. Measures should include task-completion rate, time on task, operational errors, extent of teacher revision, System Usability Scale scores, and interview data. Particular attention should be paid to whether the system reduces the burden of integrating evidence and preparing differentiated materials or merely transfers work to the review stage.

7.4. Multi-agent comparison and ablation evaluation

The design claim to be tested is that, for a complex process involving evidence analysis, content generation, task matching, and quality checking, coordinated specialist agents can improve the evidence fit, instructional coherence, and traceability of learning packages. Matched anonymized learner cases can be processed using a single prompt or single-agent baseline and the complete multi-agent workflow. Evaluators unaware of the generation condition can then conduct a blinded assessment.

In addition to content quality, evaluation should record token use, response time, failure rate, and teacher editing time. If the multi-agent system does not demonstrate a stable quality advantage, or if its cost and review burden substantially exceed those of the baseline, unnecessary roles and calls should be removed.

7.5. Classroom evaluation and boundaries of inference

Before classroom implementation, researchers should specify the unit of analysis, the granularity of personalization, and the comparison condition. Pre-support evidence snapshots, package-completion records, second-assessment performance, and post-support states should be retained. Improvement in one group's scores indicates a change in performance but does not, by itself, demonstrate that the system caused the change. Stronger effectiveness claims require an appropriate comparison, comparable instructional time, baseline adjustment, and adequate sample size. Cases in which students do not complete packages, mastery does not improve, or teachers reject recommendations should also be reported.

8. Discussion

8.1. Theoretical and design contributions

The principal contribution of this study is the translation of mastery learning from a general instructional process into task sequences, state transitions, and quality constraints within a multi-agent system. Rather than starting with what a model can generate and then searching for an instructional use, the design first establishes a process of mastery objectives, formative assessment, corrective or enrichment support, and reassessment. Agents are then assigned to evidence analysis, content generation, task matching, and quality checking within that process.

At the design level, the study creates a traceable relationship among formative assessment evidence, multi-agent collaboration, teacher review, and a second parallel formative assessment. The five design principles are not tied to a particular model or programming framework and may serve as provisional design knowledge for subsequent system development and evaluation. Their applicability nevertheless requires testing and revision in authentic instructional contexts.

8.2. Implications for teaching and development

For teachers, the system is intended to shorten the distance between identifying a mastery gap and preparing differentiated support. Its value should not be judged solely by generation speed. Evaluation should also ask whether the diagnostic basis is intelligible, whether a learning package is easy to revise, whether the second assessment provides meaningful evidence of the updated mastery state, and whether professional teacher judgment is preserved.

For developers, a personalized AI system requires more than a model interface. It also depends on learning objectives, knowledge-component links, historical snapshots, permission structures, shared state, and review logs. Without these mechanisms, even fluent generated content cannot explain why it suits a particular student, who is accountable for publishing it, or how its effects will be assessed.

8.3. Conditions for using multiple agents

Multi-agent designs are most appropriate when a support process involves several distinct steps and roles and requires shared state and quality gates—for example, a continuous process from formative evidence through diagnosis, resource and item generation, matching, review, and reassessment. For a single explanation, a simple rewrite, or a one-off question, one agent may be sufficient.

The validity of a multi-agent system therefore cannot be inferred from the number of named roles. It depends on explicit task boundaries, structured information exchange, state transitions, tool use, and auditable records. Its incremental value must ultimately be established through comparative or ablation evaluation rather than inferred from an architecture diagram.

8.4. Limitations

This study reports the system design, the existing platform foundation, and a multi-agent workflow prototype. It has not yet completed empirical evaluation with experts, teachers, or students and therefore cannot establish improvements in achievement or engagement or reductions in teacher workload. Initial rule-based mastery states may oversimplify learning when item counts are small, knowledge-component annotations are inconsistent, or evidence is outdated.

A second parallel formative assessment can be designed for the same objective and comparable cognitive demand and difficulty, but strict equivalence cannot be claimed without pilot testing and psychometric analysis. LLM outputs will also vary with model version, prompt design, and the quality of curricular resources; multi-agent orchestration may increase latency, cost, and review burden. Finally, the illustrative scenario focuses on a clearly specified middle-school mathematics objective. Applicability to open-ended creation, affective support, and other subject areas remains to be investigated.

9. Conclusion

Grounding a multi-agent personalized learning support system in mastery learning makes it possible to treat personalization as an evidence-driven instructional cycle rather than as isolated content generation. The proposed design connects explicit objectives, formative evidence, corrective or enrichment support, teacher authority, and a second parallel formative assessment through structured agent contracts and auditable state transitions. Its contribution is therefore a theoretically derived and implementable design framework, not evidence of instructional effectiveness. The next step is a staged program of technical validation, expert content review, teacher usability testing, multi-agent comparison, and classroom evaluation, with particular attention to the sufficiency of diagnostic evidence, the instructional quality of differentiated support, the appropriateness of the second assessment, and teachers' actual review burden.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Turnbull D, Chugh R, Luck J, 2022, An Overview of the Common Elements of Learning Management System Policies in Higher Education Institutions. TechTrends, 66(5): 855–867.

[2] Khor ET, 2023, A Systematic Review of the Role of Learning Analytics in Supporting Personalized Learning. Education Sciences, 14(1): 51.

[3] Matcha W, Uzir NAA, Gašević D, et al., 2019, A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. IEEE Transactions on Learning Technologies, 13(2): 226–245.

[4] Chu Z, Wang S, Xie J, et al., 2025, LLM Agents for Education: Advances and Applications. arXiv preprint arXiv:2503.11733, 2.

[5] Kasneci E, Seßler K, Küchemann S, et al., 2023, ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 103: 102274.

[6] Lo CK, 2023, What is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences, 13(4): 410.

[7] Tlili A, Shehata B, Adarkwah MA, et al., 2023, What if the Devil Is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots In Education. Smart Learning Environments, 10(1): 15.

[8] Vaccaro Jr M, Friday M, Zaghi A, 2025, Multi-Agentic LLMs for Personalizing STEM Texts. Applied Sciences, 15(13): 7579.

[9] Akgun S, Greenhow C, 2022, Artificial Intelligence in Education: Addressing Ethical Challenges in K-12 Settings. AI and Ethics, 2(3): 431–440.

[10] Holmes W, Porayska-Pomsta K, Holstein K, et al., 2022, Ethics of AI in Education: Towards a Community-Wide Framework. International Journal of Artificial Intelligence in Education, 32(3): 504–526.

[11] Bloom BS, 1968, Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. Evaluation Comment, 1(2): n2.

[12] Guskey TR, 2007, Closing Achievement Gaps: Revisiting Benjamin S. Bloom's "Learning for Mastery." Journal of Advanced Academics, 19(1): 8–31.

[13] Guskey TR, 2010, Lessons of Mastery Learning. Educational Leadership, 68(2): 52–57.

[14] Bernacki ML, Greene MJ, Lobczowski NG, 2021, A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and For What Purpose (s)? Educational Psychology Review, 33(4): 1675–1715.

[15] Bhutoria A, 2022, Personalized Education and Artificial Intelligence in the United States, China, and India: A Systematic Review Using a Human-in-the-Loop Model. Computers and Education: Artificial Intelligence, 3: 100068.

[16] Holstein K, McLaren BM, Aleven V, 2019, Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher-AI Complementarity. Grantee Submission.

[17] Shankar SK, Pothancheri G, Sasi D, et al., 2025, Bringing Teachers in the Loop: Exploring Perspectives on Integrating Generative AI in Technology-Enhanced Learning. International Journal of Artificial Intelligence in Education, 35(1): 155–180.

[18] Hattie J, Timperley H, 2007, The Power of Feedback. Review of Educational Research, 77(1): 81–112.

[19] Shute VJ, 2008, Focus on Formative Feedback. Review of Educational Research, 78(1): 153–189.

[20] Richey RC, Klein JD, 2014, Design and Development Research: Methods, Strategies, and Issues. Routledge.

[21] Gregor S, Hevner AR, 2013, Positioning and Presenting Design Science Research for Maximum Impact. MIS Quarterly, 37(2): 337–355.

[22] Hoadley C, Campos FC, 2022, Design-Based Research: What It Is and Why It Matters to Studying Online Learning. Educational Psychologist, 57(3): 207–220.

[23] Wang F, Hannafin MJ, 2005, Design-Based Research and Technology-Enhanced Learning Environments. Educational Technology Research and Development, 53(4): 5–23.

[24] Abu-Rasheed H, Weber C, Fathi M, 2023, Context-Based Learning: A Survey of Contextual Indicators for Personalized and Adaptive Learning Recommendations–A Pedagogical and Technical Perspective. Frontiers in Education, 8: 1210968.

[25] Brusilovsky P, Millán E, 2007, User Models for Adaptive Hypermedia and Adaptive Educational Systems. In The Adaptive Web: Methods and Strategies of Web Personalization, 3–53, Berlin, Heidelberg: Springer Berlin Heidelberg.

[26] Corbett AT, Anderson JR, 1994, Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4(4): 253–278.