Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

The Multi-Level Regulation Engine: A Socially Adaptive Intelligent Tutoring System Architecture For Heterogeneous Collaborative Learning

Rym Aiouni[1] · Anis Bey[1]

[1]Higher School of Management Sciences – Annaba, Annaba, Algeria. Email: aiouni.rym@essg-annaba.dz
[2]Higher School of Management Sciences – Annaba, Annaba, Algeria. Email: anis.bey@essg-annaba.dz

Abstract

We introduce the Multi-Level Regulation Engine (MLRE), an Intelligent Tutoring System (ITS) architecture that regulates learning simultaneously at individual, dyadic, and group levels in heterogeneous collaborative groups, grounded in social learning theory. Most ITS adapt to learners as isolated individuals, leaving the quality of collaboration in diverse groups unregulated. Following a Design Science Research methodology, we validated two of the architecture's four layers offline prior to any classroom deployment. We benchmarked individual learner modeling on the public ASSISTments dataset (area under the curve = 0.78, consistent with the literature). We tested the MLRE's decision logic through a 1,000-session simulation (200 sessions × 5 scenarios), improving participation equity in every scenario, with the qualitative direction of this effect confirmed robust across a 135-combination parameter sweep. The group-interaction-modeling layer, identified as the architecture's central novelty, remains to be validated on authentic multi-party data. We propose a quasi-experimental field study to evaluate learning performance, participation equity, and socio-metacognitive development. This work contributes a formally specified, partially validated, and ethically grounded artifact for inclusive, explainable artificial intelligence in education.

Keywords: Intelligent Tutoring Systems; Collaborative Learning; Socially Adaptive AI; Heterogeneous Groups; Inclusive Education; Design Science Research.

1. Introduction

The increasing diversity of learner populations represents one of the most significant challenges facing contemporary education systems. Digital transformation, globalization, massification of higher education, and mobility have produced learning environments characterized by strong heterogeneity in learners' cognitive abilities, cultural backgrounds, linguistic repertoires, motivational profiles, and prior experiences. This diversity can enrich learning through multiple perspectives and forms of knowledge construction, but it also introduces substantial pedagogical complexity. Intelligent Tutoring Systems (ITS) have emerged as a promising response to the resulting demand for personalization, demonstrating positive effects on

learning outcomes across domains by adapting content and feedback to individual learners. Most ITS, however, remain grounded in a
learner-as-an-isolated-individual paradigm, at odds with decades of educational research showing that learning is a socially mediated activity. This article argues that next-generation ITS must evolve beyond individual adaptation toward
socially adaptive pedagogical regulation, capable of understanding and guiding collaborative processes at multiple levels. The central research question guiding this work is therefore:
How can Intelligent Tutoring Systems simultaneously support individual learning, regulate how learners interact, and help diverse groups collaborate inclusively?
To address this question, we propose, formally specify, and partially validate offline a conceptual framework that integrates social learning theory, collaborative learning research, and recent advances in artificial intelligence, while explicitly considering ethical and inclusivity concerns. Following a Design Science Research methodology (Section 5.0), the architecture itself is the research artifact; its pedagogical effectiveness remains to be established through the field study specified in Section 5.
The main contributions of this work are:
1. A novel socially adaptive ITS architecture supporting simultaneous regulation of individual, dyadic, and group-level learning processes in heterogeneous collaborative groups, formally specified in Section 4.
2. The formal definition of the Multi-Level Regulation Engine (MLRE), the architecture's core novel component, specified as an explicit state–action–constraint tuple rather than described only conceptually (Section 4.4.1), and positioned in Table 1 (Section 3.6) as the first framework to combine dynamic adaptivity with theory-derived pedagogical constraints and human oversight.
3. An explainable, human-in-the-loop pedagogical regulation mechanism that logs and justifies every regulation decision, gives educators override authority over the system's actions (Section 4.6), and surfaces the normative trade-off between individual performance and collective equity as an instructor-configurable parameter rather than a hidden optimization target (Section 4.4.1).
4. Offline validation of two architectural components, and a pilot of a third: individual learner modeling, benchmarked on a standard public knowledge-tracing dataset; the MLRE's decision logic, verified through a 1,000-session simulation robust across a 135-combination parameter sweep; and the Group Interaction Layer's participation-graph and discourse-analysis algorithms, piloted on 40 real multi-party meeting transcripts and 479 real K–12 classroom transcripts respectively (Section 5.9).
5. A complete Design Science Research evaluation framework specifying both the formative offline validation already conducted and the summative field study still required (Section 5).
The remainder of this paper is organized as follows. Section 2 reviews the theoretical foundations underpinning socially adaptive Intelligent Tutoring Systems, collaborative learning, and inclusive education. Section 3 analyzes related work and identifies the research gap addressed by this study. Section 4 presents the proposed multi-level architecture and the Multi-Level Regulation Engine (MLRE). Section 5 describes the Design Science Research methodology, the planned evaluation framework, and the offline validation conducted for key architectural components. Section 6 discusses the theoretical and practical implications, positioning the framework within recent advances in AI-enabled collaborative learning while acknowledging its limitations. Finally, Section 7 concludes the paper and outlines directions for future research.

2. Theoretical Framework: From Individual Adaptation to Inclusive Social Regulation

2.1. Intelligent Tutoring Systems: From Individual Adaptation to Social Awareness

Intelligent Tutoring Systems (ITS) are computer-based instructional environments that provide personalized learning through continuous modeling of learners' cognitive, behavioral, and, increasingly, affective and motivational states, aiming to reproduce the effectiveness of one-to-one human tutoring (Anderson et al., 1995; Woolf, 2010). Classical ITS architectures comprise a learner model, a domain model, a pedagogical model, and an interaction module, enabling fine-grained individual adaptation across a wide range of instructional domains (Lin, Huang, & Lu, 2023).

Recent advances in machine learning, natural language processing, affective computing, and large language models have significantly expanded these capabilities, enabling naturalistic dialogue, context-sensitive explanations, and dynamic strategy adjustment based on complex behavioral patterns (Graesser et al., 2018). Despite this progress, however, the dominant paradigm remains fundamentally individual-centric: adaptation mechanisms optimize individual performance indicators such as mastery or accuracy, with limited consideration for the social context of learning, an assumption at odds with decades of learning-sciences research on the socially mediated nature of knowledge construction (Vygotsky, 1978; Bandura, 1977).

This gap motivates a conceptual shift toward socially aware ITS capable of reasoning beyond the individual learner to incorporate group-level phenomena.

2.2. Social and Collaborative Learning Theories

Social and collaborative learning theories conceptualize learning as a process that is fundamentally embedded in social interaction. From this perspective, cognition is not solely located within the individual but is distributed across participants, tools, and cultural practices (Roschelle & Teasley, 1995). These theories therefore offer a critical foundation for rethinking the design of ITS intended to support collaborative learning.

Vygotsky's sociocultural theory emphasizes that learning is mediated through interaction with others and with cultural artifacts (Vygotsky, 1978). Central to this framework is the concept of the Zone of Proximal Development (ZPD), which defines learning potential as the distance between what learners can achieve independently and what they can accomplish through guided interaction. Scaffolding within the ZPD allows learners to progressively internalize strategies and concepts initially supported by social mediation. For ITS, this implies that adaptation should not be restricted to individual knowledge states but should also account for the availability and quality of social support within collaborative settings.

Bandura's social learning theory further reinforces the role of interaction by highlighting the importance of observation, modeling, and social reinforcement (Bandura, 1977). Learners acquire new behaviors and strategies by observing peers, evaluating outcomes, and regulating their actions based on social feedback. Motivation and self-efficacy are thus shaped by collective experience, suggesting that learning environments should facilitate productive peer modeling and socially meaningful feedback.

Building on these foundations, research in collaborative learning and Computer-Supported Collaborative Learning (CSCL) has demonstrated that collaboration is not inherently beneficial; its effectiveness depends on the nature of interaction. Dillenbourg (1999) argues that learning gains emerge through explanation, argumentation, and negotiation of meaning rather than mere co-presence. Similarly, Johnson and Johnson's social interdependence theory emphasizes that positive interdependence, individual accountability, and promotive interaction are essential conditions for effective cooperation (Johnson & Johnson, 2009).

More recently, the concept of socially shared regulation of learning has highlighted the collective nature of metacognitive processes. Hadwin et al. (2018) show that successful groups engage in joint planning,

monitoring, and adaptation of learning strategies. This body of work underscores that regulation is not solely an individual capability but a distributed, socially negotiated process.

2.3. Heterogeneity as a Pedagogical Resource

Differences in prior knowledge, cognitive styles, language proficiency, cultural background, and motivation are now commonplace in educational settings. Traditionally, such diversity has been viewed as a source of difficulty, associated with coordination problems, unequal participation, and disparities in learning outcomes. Research in the learning sciences suggests instead that heterogeneity can function as a powerful pedagogical resource when appropriately regulated.

From a sociocultural perspective, diversity within groups creates opportunities for cognitive conflict and perspective-taking, which are key mechanisms of conceptual development (Doise & Mugny, 1984). Heterogeneous groups encourage learners to articulate, justify, and revise their understanding, leading to deeper processing of knowledge. More advanced learners benefit from explaining concepts to peers, while less experienced learners gain access to scaffolding within their ZPD (Vygotsky, 1978).

Heterogeneity benefits learning not through group composition alone but through how groups interact. Webb (2009) demonstrates that learning gains are strongly linked to the quality of peer explanations and responsiveness. Conversely, without adequate regulation, heterogeneity can lead to participation asymmetries, dominance by confident learners, and marginalization of others (Barron, 2003).

These risks are particularly salient in technology-mediated environments, where social cues are attenuated and coordination demands are high, motivating the regulatory strategies (adaptive role assignment, targeted prompts, task redistribution) detailed in Section 4.4. This framing also aligns with inclusive education principles of equitable participation and recognition of diversity, positioning inclusive ITS as systems supporting agency and belonging rather than personalization alone.

3. 3. Related Work and Research Gaps

3.1. Intelligent Tutoring Systems and Individual-Level Adaptation

Early ITS such as Cognitive Tutor and ALEKS demonstrated that individualized feedback and mastery-based progression improve learning performance (Anderson et al., 1995; Koedinger & Corbett, 2006). Subsequent machine-learning advances strengthened learner modeling through probabilistic knowledge-state estimation and misconception detection (Desmarais & Baker, 2012), and NLP-based conversational agents later extended ITS toward dialogue-based tutoring (Graesser et al., 2018). As Section 2.1 discussed, this trajectory has remained individual-centric throughout, leaving the representational and computational mechanisms needed for collaborative contexts largely absent.

3.2. Collaborative Learning and Computer-Supported Collaborative Learning (CSCL)

In parallel with ITS research, Computer-Supported Collaborative Learning (CSCL) has extensively investigated how technology can mediate interaction, coordination, and shared knowledge construction among learners (Baker & Reimann, 2024), on the premise discussed in Section 2.2 that learning gains depend on the quality of interaction rather than on co-presence alone (Dillenbourg, 1999; Stahl, 2006).

CSCL environments have introduced tools for scripting collaboration, structuring roles, and visualizing group processes to support productive interaction. Empirical studies show that such scaffolds can enhance coordination and learning outcomes, particularly in heterogeneous groups (Kollar, Fischer, & Hesse, 2006). However, these systems typically rely on static or teacher-defined scripts, offering limited adaptivity to evolving group dynamics, although recent work has begun to compare adaptable with non-adaptable scripts delivered through conversational agents (Li, Hu, & Li, 2025).

Moreover, many CSCL systems lack the real-time adaptive intelligence characteristic of ITS, depending instead on instructor intervention (Section 3.4).

3.3. Toward Collaborative Intelligent Tutoring Systems

Recognizing the limitations of both individual-focused ITS and largely static CSCL environments, recent research has begun to explore collaborative Intelligent Tutoring Systems. These systems attempt to combine the adaptive power of ITS with the social focus of collaborative learning.

Several studies have investigated algorithmic group formation based on learner characteristics such as prior knowledge, learning style, or performance level, demonstrating that adaptive grouping can improve interaction quality and learning outcomes (Bernacki et al., 2014; Haq et al., 2021).

Other work has explored the use of pedagogical agents to facilitate collaboration by prompting discussion, encouraging participation, or mediating conflicts (Kim & Baylor, 2016; Edwards et al., 2025). Additionally, advances in reinforcement learning have been applied to adapt instructional interventions based on learner behavior in collaborative contexts (Abdelshiheed et al., 2023). Hybrid human–AI tutoring approaches, in which human tutors and automated systems jointly support learners, further illustrate the potential for blending AI adaptivity with human social judgment (Matsuda & VanLehn, 2016).

Despite these developments, adaptation in most collaborative ITS remains restricted to specific aspects of collaboration rather than treating it as a core, multi-level architectural concern (Section 3.4).

3.4. Limitations of Existing Approaches

A critical examination of the literature reveals several persistent limitations:

1. **Lack of multi-level modeling**
   Most ITS model learners individually and, at best, aggregate individual data to make group-level decisions. Few systems explicitly represent dyadic relationships, interaction patterns, or emergent group states such as cohesion or participation equity.
2. **Limited regulation of social dynamics**
   While some systems monitor interaction, they rarely intervene in a pedagogically principled manner to regulate participation, manage conflict, or support socially shared regulation of learning.
3. **Insufficient attention to heterogeneity and inclusion**
   Although heterogeneity is often acknowledged, few ITS explicitly address issues of inclusion, equity, and differential participation. This omission risks reinforcing existing inequalities within collaborative learning environments.
4. **Weak integration of theory and system design**
   Many collaborative ITS lack a strong grounding in social and collaborative learning theories, resulting in ad hoc design choices that are difficult to generalize or evaluate systematically.

3.5. Research Gap and Contribution of the Present Study

The review above highlights a clear gap at the intersection of Intelligent Tutoring Systems, collaborative learning, and inclusive education. Prior research has demonstrated the effectiveness of individual adaptation and explored isolated aspects of collaboration, but integrated frameworks that conceptualize ITS as agents of multi-level pedagogical regulation remain absent. The present study addresses this gap by proposing a socially adaptive ITS framework that models learning at the individual, dyadic, and group levels and treats heterogeneity as a pedagogical resource; Section 1 states the specific claims made.

3.6. Comparative Positioning Relative to Prior Frameworks

One question the architecture is likely to prompt is worth addressing directly: why not simply prompt a large language model (LLM) to regulate the group? Recent LLM-based multi-agent tutoring systems (Wang et al., 2025; Zhang et al., 2024) achieve conversational adaptivity but lack the explicit group-state modeling,

hard pedagogical constraints, and human-in-the-loop override mechanisms the MLRE provides. The architecture is therefore better read as a proposal for constraining and structuring LLM-based agents pedagogically, not as a competitor to them (Section 6.2).

To situate the contribution of the proposed architecture more precisely, Table 1 compares it against five reference points spanning the individual-adaptation and collaboration-support traditions reviewed above: classical individually-adaptive ITS, static CSCL scripts, first-generation collaborative ITS, and recent LLM-based multi-agent tutoring approaches. The comparison is organized around the dimensions most directly implicated in the research gap identified in Section 3.4: depth of individual and group modeling, adaptivity of the collaboration structure itself, real-time regulation of participation equity, explainability, human oversight, and empirical maturity.

Table 1 Comparison of the proposed framework with prior ITS and CSCL approaches

Dimension	Cognitive Tutor / ALEKS (Anderson et al., 1995; Koedinger & Corbett, 2006)	CSCL scripts (Kollar et al., 2006; Miller & Hadwin, 2024)	Collaborative ITS (Bernacki et al., 2014; Kim & Baylor, 2016)	LLM-based multi-agent tutors (Wang et al., 2025; Zhang et al., 2024)	Proposed multi-level socially adaptive ITS
Individual learner modeling	Fine-grained (mastery, KT)	Minimal / absent	Present, basic	Emergent, prompt-based	Fine-grained + affective/behavioral
Group/dyadic modeling	Absent	Structural only (fixed roles/phases)	Partial (group formation only)	Implicit, not explicitly modeled	Explicit (participation, cohesion, roles, conflict)
Adaptivity of collaboration structure	N/A	Static / teacher-defined	Limited (one-shot grouping)	Dynamic but unconstrained	Dynamic, theory-constrained (MLRE)
Real-time regulation of participation equity	No	No (scripted, not reactive)	No	No (reactive only)	Yes: explicit design principle
Explainability / rationale to teacher	Limited	N/A (transparent by design, but rigid)	Limited	Largely opaque	Explicit explainability module
Human-in-the-loop control	No	Implicit (teacher authors script)	No	No	Yes: teacher dashboard + override

Empirical validation status (as of this article)	Extensively validated	Extensively validated	Partially validated	Early-stage, few controlled studies	Partially validated offline (Section 5.9); field study proposed

As Table 1 shows, no prior framework combines dynamic, theory-constrained regulation of group-level dynamics with explicit explainability and human oversight. CSCL scripts achieve transparency and pedagogical groundedness but sacrifice adaptivity; LLM-based multi-agent tutors achieve adaptivity but at the cost of explainability and empirical maturity. The proposed architecture is positioned to combine the adaptivity of the latter with the interpretability and pedagogical accountability of the former, a combination that, as the final row of Table 1 records, has yet to be demonstrated in a classroom and remains the central claim to be tested by the evaluation plan in Section 5.

4. A Multi-Level Architecture for Socially Adaptive Intelligent Tutoring Systems

This section presents the proposed multi-level socially adaptive ITS architecture.

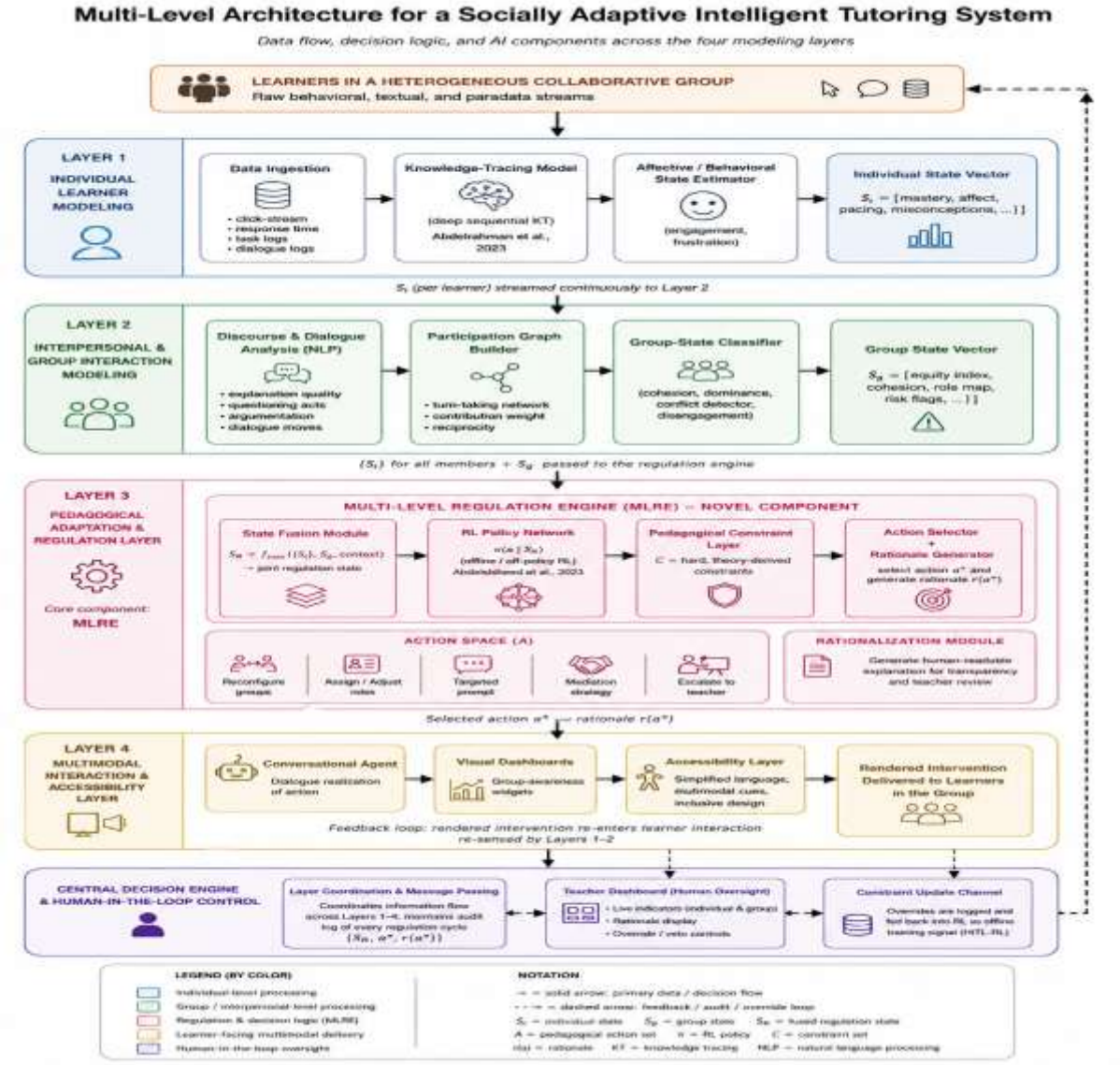


Fig. 1 Detailed multi-level architecture, showing data flow from raw learner interaction through individual and group modeling into the Multi-Level Regulation Engine, out to multimodal delivery, with human-in-the-loop oversight via the central decision engine

4.1. Design Principles

The architecture is guided by four core design principles derived from the literature. They are jointly operationalized by its core novel component, the Multi-Level Regulation Engine (MLRE, formally defined in Section 4.4.1), which fuses the individual- and group-level state representations introduced in Sections 4.2–4.3 into the regulation decisions described below:

1. **Multi-level pedagogical regulation**
   The architecture models learning as a process unfolding simultaneously at individual, dyadic, and collective levels. The system must therefore be capable of reasoning across these levels rather than aggregating individual data in a purely additive manner.
2. **Social awareness and interaction sensitivity**
   The system must model not only learner knowledge but also interaction patterns, participation equity, and relational dynamics within groups.
3. **Adaptive and inclusive orchestration**
   Heterogeneity is treated as a pedagogical resource. Adaptation mechanisms aim to promote equitable participation and leverage diversity through role distribution and task design.
4. **Human-centered and explainable AI**
   Pedagogical decisions made by the system should be transparent and interpretable by educators (Khosravi et al., 2022), preserving the teacher's role as pedagogical authority.

4.2. Individual Learner Modeling Layer

The individual modeling layer represents the traditional foundation of ITS while extending it to incorporate affective and behavioral dimensions. This layer continuously captures and updates representations of each learner's:

- Domain-specific knowledge and misconceptions,
- Learning progress and mastery level,
- Behavioral indicators (e.g., response time, task persistence),
- Affective and motivational states inferred from interaction traces.

Data are collected through learner interactions with tasks, system logs, and, where applicable, multimodal inputs such as textual dialogue or interaction frequency. Machine learning techniques are employed to infer latent learner states and predict learning trajectories.

4.3. Interpersonal and Group Interaction Modeling Layer

The interaction modeling layer is the architecture's principal point of departure from conventional ITS. Rather than treating learners as isolated entities, it aggregates individual data into representations of dyadic relationships and group-level dynamics**.**

Key constructs modeled at this level include:

- Participation distribution and equity,
- Interaction frequency and reciprocity,
- Role emergence and dominance patterns,
- Indicators of cohesion, disengagement, or conflict.

Natural language processing techniques can be applied to collaborative discourse to analyze explanation quality, responsiveness, and negotiation of meaning (Ouyang, Xu, & Cukurova, 2023). These indicators

allow the system to detect imbalances or breakdowns in collaboration that may hinder learning, particularly in heterogeneous groups.

4.4. Pedagogical Adaptation and Regulation Layer

The pedagogical regulation layer integrates information from both individual and collective models to determine context-sensitive instructional interventions**.** It operates as a social regulator, orchestrating learning activities across the individual, dyadic, and group levels.

Adaptive strategies at this level may include:

- Dynamic group formation or reconfiguration,
- Adaptive role assignment (e.g., explainer, moderator, synthesizer),
- Targeted prompts to encourage under-participating learners,
- Mediation strategies to resolve conflict or support shared regulation.

Decision-making can be supported by reinforcement learning approaches that optimize pedagogical interventions based on observed learning and interaction outcomes.

4.4.1. Formal Definition of the Novel Component: The Multi-Level Regulation Engine (MLRE)

The Multi-Level Regulation Engine (MLRE) is situated inside the Pedagogical Adaptation and Regulation Layer (Section 4.4) and depicted in the shaded core of Fig. 1. Where a conventional ITS pedagogical model selects an action from a state defined over a single learner, the MLRE selects an action from a state jointly defined over an entire heterogeneous group. This subsection gives the MLRE a precise, implementation-oriented definition so that the novelty claim is falsifiable rather than purely conceptual.

Definition (MLRE). Formally, the MLRE is defined as a tuple:

```
MLRE = < S_R, A, T, R, π, C >

where:
 S_R  : joint regulation state space,  S_R = f_fuse( {S_i}_(i=1..n) , S_g, ctx )
 A    : pedagogical action space (group reconfiguration, role
        reassignment, targeted prompt, mediation strategy, escalate)
 T    : transition dynamics of the group under action a (unknown;
        approximated via interaction logs / simulation)
 R    : reward signal combining individual learning gain,
        participation-equity delta, and socio-metacognitive proxy
 π    : policy π(a | S_R), the learned or rule-constrained mapping
        from state to action
 C    : hard constraint set derived from pedagogical theory that
        π may never violate (e.g., never isolate a struggling
        learner; never reassign roles more than once per session)
```

State fusion. The joint state S_R is not a simple concatenation of individual states. It is computed by the state fusion module (Fig. 1) as S_R = f_fuse({S_i}, S_g, ctx), where {S_i} are the per-learner vectors produced by Layer 1 (Section 4.2: mastery estimate, affective/behavioral indicators, misconception flags), S_g is the group-level vector produced by Layer 2 (Section 4.3: participation-equity index, cohesion score, dominance pattern, conflict/disengagement flags), and ctx encodes task-level context (task type, time remaining, group size). This fusion step is what allows the engine to reason about, for example, whether a single learner's disengagement (an individual signal) is a symptom of a group-level participation imbalance (a collective signal) rather than an independent event, a distinction a purely individual-level ITS cannot make.

Policy and constraints. The policy π is trained offline or off-policy, consistent with recent RL-based tutoring interventions (Abdelshiheed et al., 2023), using logged trajectories rather than live exploration, for safety. Critically, π does not act unconstrained: the constraint set C, derived from the theoretical foundations in Section 2 (e.g., Johnson & Johnson's positive-interdependence principle, Hadwin et al.'s socially shared regulation), filters the action distribution before execution. This operationalizes "human-centered and explainable AI" (design principle 4, Section 4.1) as a mechanism rather than an aspiration: C is inspectable, editable by instructors through the teacher dashboard (Section 4.6), and logged alongside every decision.

Reward signal. The reward R combines three terms, $R = \alpha \cdot \Delta(\text{learning gain}) + \beta \cdot \Delta(\text{participation equity}) + \gamma \cdot \Delta(\text{socio-metacognitive proxy})$, with weights (α, β, γ) set by the instructor or institution rather than fixed by the system designer. This weighting mechanism is itself a deliberate design choice: it surfaces, rather than conceals, the normative trade-off between individual performance and collective equity that is usually left implicit in adaptive-learning systems.

The specification is deliberately partial in one respect: the transition dynamics T are not assumed to be known analytically but are treated as a learning target, consistent with the design-stage status of this contribution (Section 6.4).

A cold-start question follows naturally: before enough interaction has been observed, S_g is undefined, so the MLRE enforces a minimum-observation guard, deferring to individual-level regulation (Layer 1) until a minimum number of turns has accumulated, then beginning the regular decision cycle (Algorithm 1). The Section 5.9.2 simulation was instrumented to measure this directly under a 3-turn guard: median first intervention at cycle 2, scenario means of 2.45–2.80 turns, and a maximum of 25 turns in the silent-learner scenario. Between 8% and 25.5% of sessions never intervened at all, correctly reflecting cases where the equity threshold was never crossed. Real-world warm-up latency remains an empirical question for the field study (Section 5).

4.4.2. Candidate Algorithms and a Simple MLRE Execution Workflow

Table 2 summarizes concrete algorithmic candidates for each component introduced above, to make the architecture's implementability more tangible without committing prematurely to a single technical stack.

Table 2 Candidate algorithms and techniques per architectural component

Architectural component	Candidate algorithms / techniques	Representative source(s)
Individual learner modeling (§4.2)	Deep / attention-based knowledge tracing; Bayesian knowledge tracing	Abdelrahman et al., 2023; Shen et al., 2024
Affective/behavioral state estimation (§4.2)	Multimodal engagement classifiers over log/paradata features	Section 4.2
Discourse & dialogue analysis (§4.3)	Transformer-based discourse tagging; LLM-based explanation-quality scoring	Kasneci et al., 2023; Ouyang & Zhang, 2024
Participation / group-state modeling (§4.3)	Social network analysis on turn-taking graphs; cohesion/conflict classifiers	Section 4.3
Policy learning (MLRE, §4.4.1)	Offline / off-policy reinforcement learning over logged trajectories	Abdelshiheed et al., 2023
Constraint enforcement (MLRE, §4.4.1)	Rule-based action filtering / shielding over the RL action distribution	Section 4.4.1
Explainability module (§4.4.1)	Post-hoc natural-language rationale generation (e.g., template- or LLM-based)	Türkmen, 2025; Maurya et al., 2025

Multimodal delivery agent (§4.5)	LLM-based conversational agent, subject to the MLRE constraint set	Cai et al., 2024; Wang et al., 2025; Zhang et al., 2024

To make the MLRE's operation concrete, Algorithm 1 formalizes a single regulation cycle, repeated at a fixed cadence or triggered by a threshold event such as a sustained participation-equity drop.

Algorithm 1. Multi-Level Regulation Engine (MLRE): single regulation cycle

```
Input:  learner states {S_i}, group state S_g, context ctx
Output: pedagogical action a, rationale r(a)

1. S_R <- Fuse({S_i}, S_g, ctx)        // state fusion (Sec. 4.4.1)
2. a_proposed <- Policy(S_R)           // pi(a | S_R), offline-trained RL
3. if Violates(a_proposed, C):         // pedagogical constraint check
4.     a_proposed <- NextBest(S_R, C)  // re-select under constraints
5. r(a) <- Explain(a_proposed, S_R)    // natural-language rationale
6. Deliver(a_proposed)                 // Layer 4: render to learner(s)
7. Log({S_R, a_proposed, r(a)})        // audit trail (Sec. 4.6)
return a_proposed, r(a)
```

A teacher override can enter the loop at step 4 (constraint re-selection) or after step 6 (delivery); step 7's audit log is also the source of the offline learning signal used to refine π. The workflow is intentionally simple, marking where oversight and learning signal enter rather than specifying a production-ready control loop.

4.5. Multimodal Interaction and Accessibility Layer

To support inclusive collaboration, the architecture includes a multimodal interaction layer designed to accommodate diverse learner needs and preferences. This layer manages communication between learners and the system through:

- Textual and conversational interfaces,
- Visual feedback and dashboards,
- Accessibility-oriented representations (e.g., simplified language, multimodal cues).

AI-driven pedagogical agents operate within this layer to facilitate interaction, provide feedback, and support regulation strategies initiated by the pedagogical layer. These agents are designed as mediators rather than instructors, reinforcing collaborative processes without replacing human facilitation.

4.6. Central Decision Engine and Human-in-the-Loop Control

At the core of the architecture lies a central decision engine that coordinates information flow across layers and selects appropriate pedagogical actions. This engine integrates predictive models with pedagogical constraints to balance adaptivity and explainability.

Crucially, the architecture adopts a human-in-the-loop approach, enabling educators to:

- Monitor individual and group indicators,
- Understand the rationale behind system interventions,
- Override or adjust system decisions when necessary.

We designed the system to function as a pedagogical assistant, supporting rather than supplanting human judgment.

5. Research Design and Methodology

This study adopts a design-based and quasi-experimental research approach to evaluate the pedagogical effectiveness of the proposed socially adaptive Intelligent Tutoring System (ITS) architecture (Fig. 1), assessing not only individual learning outcomes but also interaction quality, participation equity, and socio-metacognitive development in heterogeneous collaborative groups.

5.0. Overall Research Approach: A Design Science Research Perspective

We adopt a Design Science Research (DSR) methodology (Hevner et al., 2004; Peffers et al., 2007), treating the architecture itself as the research artifact and distinguishing the build activities reported in Section 4 from the evaluate activities specified here. The architecture has not yet been implemented as a working system, so the quasi-experimental design below is a specification rather than a completed empirical study: a research design section describing hypotheses and measures can otherwise be misread as reporting results that do not exist.

Following Peffers et al.'s (2007) DSR process model, the present work corresponds to the Awareness of Problem, Suggestion, and Development stages (Sections 1–4): the problem is the individual-centric bias of ITS in heterogeneous groups; the suggestion is the multi-level architecture; the development is its formal specification, including the Multi-Level Regulation Engine (Section 4.4.1). Sections 5.1–5.7 below specify the Evaluation stage of the DSR cycle, and Section 5.9 adds a formative evaluation phase that should logically precede it. The final DSR stage, Communication, corresponds to this article and its planned empirical follow-up.

The evaluation strategy is further structured using the Framework for Evaluation in Design Science (FEDS; Venable et al., 2016), which recommends selecting an evaluation strategy based on the trade-off between formative/summative timing and artificial/naturalistic setting. Given that the MLRE has not been implemented, an exclusively naturalistic, summative evaluation (i.e., a live classroom trial) would be premature and would risk exposing real heterogeneous learner groups to an unvalidated regulation policy.

5.1. Research Design

The research design follows a quasi-experimental pre-test/post-test model with two experimental conditions:

- **Control condition**: learners interact with a traditional ITS focused on individual-level adaptation (content sequencing, feedback, and pacing).
- **Experimental condition**: learners interact with a socially adaptive ITS implementing the multi-level architecture presented in Fig. 1, including group interaction modeling, adaptive role assignment, and pedagogical regulation.

This comparative design allows for the isolation of the added value of socially adaptive regulation beyond individual personalization.

5.2. Participants and Context

The study targets undergraduate or graduate students enrolled in domain-specific courses (e.g., computer science, engineering, or educational sciences) that require problem-solving and conceptual understanding. A sample of approximately 120 learners is envisaged, distributed into heterogeneous groups of three to four members.

We measure group heterogeneity across multiple dimensions:

- Prior knowledge (measured through a diagnostic pre-test),
- Preferences,
- Linguistic or cultural background,
- Self-reported motivational orientation.

5.3. Learning Environment and Experimental Conditions

Both conditions use the same instructional content and learning objectives to ensure comparability; they differ only in the adaptive mechanisms implemented.

In the experimental condition, adaptation is governed by the multi-level architecture illustrated in **Fig. 1**, integrating:

- Individual learner modeling,
- Interpersonal and group interaction analysis,
- Pedagogical regulation strategies targeting collaboration quality and participation equity.

AI-driven pedagogical agents mediate interactions by prompting explanation, redistributing roles, and supporting socially shared regulation of learning.

5.4. Research Hypotheses

We formulate the following hypotheses for the field study specified in Section 5.1, grounded in the theoretical framework and prior research:

H1: Learners using the socially adaptive ITS will demonstrate significantly higher learning gains than learners using an individual-focused ITS.

H2: Participation equity within heterogeneous groups will be significantly higher in the experimental condition.

H3: Learners in the experimental condition will show significantly greater pre-to-post improvement in socio-metacognitive skills (self-regulation, co-regulation, and collaborative awareness) than learners in the control condition.

H4: Learner satisfaction and perceived inclusiveness will be higher in the socially adaptive ITS condition.

5.5. Data Collection and Measures

We test these hypotheses using the following multi-level measures, combining quantitative and qualitative data to capture learning and interaction at each level:

Individual-level measures

- Pre-test and post-test performance scores,
- Learning gains computed through normalized gain indices,
- Task completion time and error patterns.

Interaction and group-level measures

- Participation distribution (e.g., contribution frequency, turn-taking balance),
- Discourse indicators derived from collaborative dialogue (e.g., explanation, questioning),
- Cohesion and engagement indices inferred from interaction logs.

Socio-metacognitive measures

- Self-report questionnaires assessing regulation strategies and collaboration awareness,
- System-derived indicators of shared regulation episodes.

5.6. Data Analysis

Quantitative data will be analyzed using mixed-effects statistical models to account for the nested structure of the data (learners within groups). Pre-test/post-test differences will be examined through repeated-measures ANOVA or linear mixed models.

Interaction data will be analyzed using descriptive and inferential statistics to compare participation equity and collaboration patterns across conditions. Where applicable, qualitative discourse analysis will complement quantitative findings to provide deeper insight into collaborative processes.

5.7. Validity and Ethical Considerations

Several measures are implemented to strengthen internal and external validity:

- Equivalence of learning content across conditions; random assignment of groups to conditions where feasible, or statistical control for pre-existing group differences where not; measures to prevent cross-condition contamination (e.g., separate class sections); and blinding of instructors to condition assignment where possible to reduce experimenter bias,
- Controlled group size and task complexity,
- Triangulation of data sources.

Ethical considerations include informed consent, anonymization of learner data, and transparency regarding AI-driven interventions. In line with the architecture presented in Fig. 1, a human-in-the-loop approach ensures that instructors retain oversight and pedagogical authority over system decisions.

5.9. Offline Validation of Core Architectural Components

Before any live classroom deployment, and consistent with the formative stage of the FEDS evaluation strategy adopted in Section 5.0, two of the architecture's four layers were validated offline using established methods and public data (Sections 5.9.1–5.9.2). Two further, deliberately narrow pilots test components of the Interpersonal and Group Interaction Modeling layer on authentic multi-party discourse (Sections 5.9.3–5.9.4). Taken together, these studies establish technical feasibility; they do not demonstrate pedagogical effectiveness, which requires the field study specified in Sections 5.1–5.7.

5.9.1. Validation of Individual Learner Modeling

The knowledge-tracing component of Layer 1 (Section 4.2) was validated on the ASSISTments 2009–2010 skill-builder dataset (Feng, Heffernan, & Koedinger, 2009), a standard public benchmark comprising 525,534 interaction records from 4,217 students across 124 skills, using the pre-defined student-level train/test split established in prior work (Khajah, Lindsey, & Mozer, 2016). A single-layer LSTM Deep Knowledge Tracing model (Piech et al., 2015) was trained to predict next-interaction correctness from each student's interaction history; no group-level or architecture-specific components were involved, since the purpose of this study is to establish that standard knowledge-tracing methods perform as expected within the proposed pipeline, not to propose a new knowledge-tracing algorithm. Table 3 reports performance on the held-out test students.

Table 3 Knowledge-tracing performance on the ASSISTments 2009–2010 benchmark (held-out test students)

Metric	Value	Notes
AUC	0.782	In the range reported in the knowledge-tracing literature for this dataset
Accuracy	0.730	Threshold = 0.5
RMSE	0.423	
Test predictions	53,701	From 834 held-out students (80/20 student-level split)

These results are consistent with published benchmarks for this dataset and confirm that the individual learner modeling layer's underlying approach is technically sound. One limitation should be stated plainly: this release of the ASSISTments 2009 dataset is documented to contain duplicate interaction records that can inflate AUC relative to a fully de-duplicated version (Xiong et al., 2016); the reported figures should therefore be read as consistent with the wider literature on this specific benchmark rather than as a bias-free estimate. Layer 1 also rests on a capability well established in prior ITS (Section 3.1), so this study confirms its technical feasibility rather than any architectural novelty.

5.9.2. Simulation-Based Validation of the Multi-Level Regulation Engine

The MLRE (Section 4.4.1) was evaluated through a controlled simulation rather than live deployment, consistent with the safety rationale in Section 5.0. One thousand synthetic collaborative sessions (200 sessions × 5 scenarios, replicated across 5 random seeds) were generated, each modeling a 4-learner group over 30 regulation cycles under one of five scenario conditions: balanced participation, a dominant learner, a silent learner, sustained conflict between two learners, and mid-session disengagement. Under the unregulated condition, participation share was allowed to evolve under a reinforcing ("rich-get-richer") dynamic; under the regulated condition, the MLRE's policy and constraint set C (Section 4.4.1) were active, selecting between targeted-prompt and mediation-strategy actions subject to an intervention cooldown and a participation floor. Table 4 reports the resulting participation-equity index (0–1, higher is more equal) under each condition.

Table 4 Participation equity index by scenario, with and without the MLRE active (mean ± SD across 1,000 sessions)

Scenario	Without MLRE	With MLRE	Mean interventions / session
Balanced	0.852 ± 0.108	0.969 ± 0.023	1.9
Dominant learner	0.642 ± 0.200	0.929 ± 0.046	2.7
Silent learner	0.708 ± 0.137	0.880 ± 0.055	2.9
Conflict	0.767 ± 0.149	0.937 ± 0.044	2.1
Disengagement	0.843 ± 0.111	0.943 ± 0.041	1.9

Because the magnitude of this effect depends on parameters chosen by the authors (the boost and dampening strength of each action, and the equity threshold that triggers intervention), a sensitivity analysis was run across a 3×3×3 grid of these three parameters (135 combinations), re-evaluating all five scenarios at each grid point. The direction of the effect, MLRE improving the equity index relative to the unregulated condition, held in 135 of 135 combinations (100%), with the mean improvement ranging from +0.084 (balanced) to +0.273 (dominant learner) depending on scenario and parameter setting, as shown in Fig. 2. This indicates that the qualitative finding is not an artifact of one hand-picked parameter setting, though it does not establish the correct real-world values for these parameters, which would require calibration against authentic interaction data.

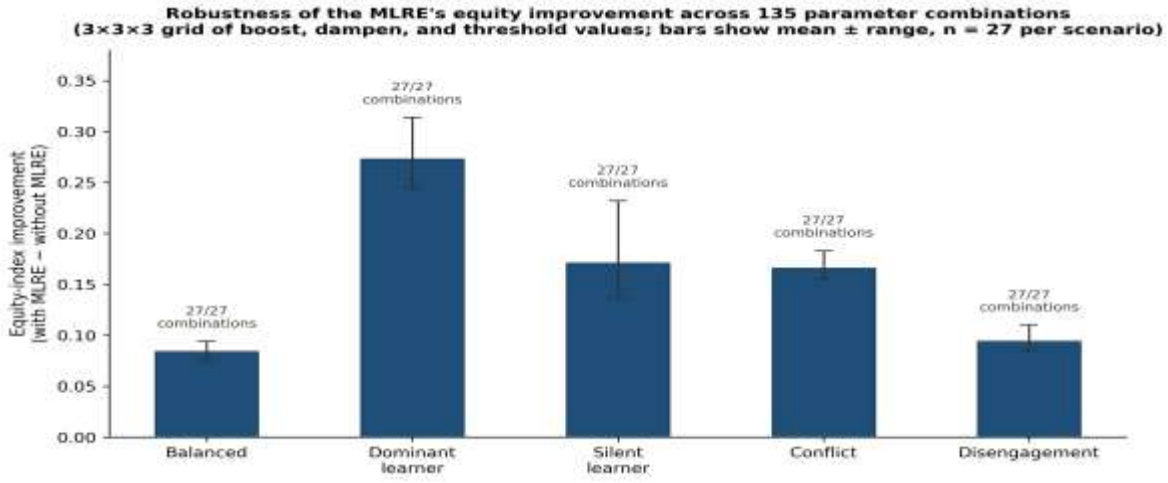


Fig. 2 Robustness of the MLRE's equity-index improvement across the 135-combination parameter

sensitivity sweep, by scenario (mean improvement, min–max range across 27 combinations per scenario)
This simulation demonstrates that the MLRE's policy and constraint logic behave as intended under controlled, synthetic conditions; it does not demonstrate, and should not be read as demonstrating, that the MLRE improves participation equity in real collaborative groups. The synthetic dynamics, including the reinforcing unregulated baseline, are modeling assumptions made by the authors rather than estimates derived from empirical group behavior.

5.9.3. Pilot Validation of the Group Interaction Layer on Authentic Multi-Party Discourse

Sections 5.9.1 and 5.9.2 leave the Interpersonal and Group Interaction Modeling layer (Section 4.3) entirely untested against real data. As a partial, deliberately narrow response, the Participation Graph Builder component was piloted on the AMI Meeting Corpus (Carletta et al., 2005), a public corpus of 3–5-participant workplace design meetings with manually transcribed, speaker-labeled utterances. This pilot targets a specific, falsifiable question: does the participation-graph extraction pipeline (turn segmentation, contribution weighting, reciprocity computation, and the equity-index formula used throughout Section 5.9.2) run successfully on authentic, noisy multi-party discourse, or does it break down on real conversational phenomena such as backchannel responses, filler words, and marked interruptions?

The pilot deliberately does not claim domain match: AMI meetings are professional workplace design discussions among adults, not educational collaborative-learning sessions among students, and this distinction matters for exactly the reasons Section 6.2 discusses regarding LLM-based systems trained on out-of-domain discourse. What this pilot can speak to is narrower and more mechanical: whether the algorithmic machinery of Layer 2 is robust to real conversational noise at all, prior to any claim about its validity in an educational setting.

The pipeline was run on all 40 publicly available transcribed meetings in the corpus's train, development, and test splits, totaling 24,000 raw utterances (merged into 15,496 speaker turns after consecutive same-speaker utterances were combined). Table 5 reports the results.

Table 5 Participation Graph Builder pilot results on 40 real AMI meetings

Metric	Value
Meetings processed successfully	40 / 40 (100%)
Raw utterances processed	24,000
Speaker turns after merging	15,496
Backchannel/filler-only turns encountered (e.g., "okay," "yeah")	2,923 (18.9% of turns)
Marked interruptions encountered	4,316
Participation equity index (turn-count basis)	0.966 ± 0.031 (mean ± SD)
Participation equity index (word-count basis)	0.916 ± 0.066 (mean ± SD)
Reciprocity index	0.931 ± 0.048 (mean ± SD)

The pipeline completed successfully on all 40 meetings with no manual pre-cleaning, including meetings containing nearly 19% backchannel-only turns and over 4,000 marked interruptions, which supports the narrow claim above: the participation-graph algorithm does not require artificially clean input to run. Two further observations are informative rather than conclusive. First, equity computed on word-count is consistently lower and more variable than equity computed on turn-count (0.916 vs. 0.966), indicating that turn-count alone, the simpler and more common operationalization in the CSCL literature, can understate participation imbalance when some speakers take many short turns and others take fewer but longer ones; this is a concrete methodological reason to prefer a word- or content-weighted equity measure over a raw

turn count in the eventual field study (Section 5.5). Second, the relatively high mean equity (0.966) and reciprocity (0.931) observed in AMI likely reflect the corpus's professional workplace setting rather than any property of the algorithm; these values should not be read as a baseline for what to expect in heterogeneous student groups, where prior evidence reviewed in Section 2.3 (Barron, 2003) suggests imbalance is more pronounced.

This pilot provides evidence that the participation-graph component of Layer 2 is mechanically robust to authentic conversational noise; a complementary pilot of the discourse-analysis component follows in Section 5.9.4.

5.9.4. Pilot Validation of the Discourse Analysis Component on Real K–12 Classroom Data

Section 5.9.3 tested the participation-graph component of Layer 2 on real but out-of-domain (workplace) discourse. As a complementary pilot, the discourse-analysis component, specifically questioning-act detection, one of the two indicators listed for this component in Section 4.3, was piloted on the TalkMoves dataset (Suresh et al., 2019): 567 publicly available, de-identified transcripts of real K–12 mathematics lessons (CC BY-NC-SA 4.0) and, unlike AMI, educational-domain data.

A first attempt used the dataset's bundled automated dialogue-act tags (a DAMSL-style classifier) as ground truth for evaluating a simple, transparent questioning-act heuristic (utterances ending in "?" or beginning with a wh-word or auxiliary-inversion pattern). This attempt was abandoned after inspection showed the bundled tagger, evidently trained on adult telephone conversation data (the Switchboard corpus, the standard DAMSL training source), performed poorly on classroom speech: clear questions such as "What was it about?" and "Do you guys remember that problem?" were labeled "Statement-non-opinion" far more often than any question category. This is itself a useful methodological finding, consistent with the caution raised in Section 6.2 about applying off-the-shelf NLP tools to educational discourse without domain adaptation, but it meant the bundled tags could not serve as a fair evaluation of the heuristic.

A structural check was used instead, relying only on reliable fields (speaker identity and turn order) rather than the unreliable automated tags. If the heuristic captures genuine elicitation function, teacher utterances it flags as questioning acts should be followed by a student turn more often than teacher utterances it does not flag, since eliciting a response is a defining function of a question in classroom dialogue. Table 6 reports the result across all successfully processed files.

Table 6 Structural validity check for questioning-act detection on 479 real TalkMoves classroom transcripts

Metric	Value
Files processed successfully	479 / 567 (84%)
Teacher utterances scanned	122,570
Teacher utterances flagged as questioning acts	35,800
→ followed by a student turn	26.5%
Teacher utterances not flagged as questioning acts	86,682
→ followed by a student turn	22.3%
Difference in student-response rate	+4.3 percentage points (z = 15.6)

The difference is small in absolute terms but, given the sample size, is not attributable to chance. It provides modest structural evidence that the heuristic captures some genuine elicitation signal rather than firing at random, consistent with the direction predicted by discourse theory. The effect size should not be overstated: a large share (22.3%) of non-flagged utterances are also followed by a student turn, which is expected in classroom Initiate-Respond-Evaluate discourse patterns where student turns regularly follow

teacher talk of any kind, not only questions; "followed by a student turn" is therefore an imperfect proxy for genuine elicitation, and this check establishes plausibility rather than classification accuracy.

Taken together with Section 5.9.3, these two pilots now cover two of Layer 2's three sub-components (the participation-graph builder and the discourse-analysis heuristic) using real, if domain-imperfect, data. Neither pilot tests the group-state classifier, and, with the partial exception of this TalkMoves pilot, neither uses genuine small-group student collaboration as opposed to whole-class or workplace discourse. The priority limitation identified in Section 6.4 is accordingly narrowed further but not resolved.

6. Discussion and Implications

This section situates the architecture's contribution for inclusive education, positions it relative to the fast-moving landscape of LLM-enabled collaborative learning, and sets out the deployment and methodological limitations that follow from its current, design-stage status.

6.1. Implications for Inclusive and Equitable Education

Reframing heterogeneity as a pedagogical asset rather than a constraint (Section 2.3), the architecture is designed to detect and regulate participation imbalances, though whether this translates to improved equity in real classrooms remains untested (Section 5.9.2). These benefits depend on careful attention to algorithmic bias, privacy, and cognitive overload (Memarian & Doleck, 2023), discussed further as practical limitations in Section 6.3.

6.2. Positioning within the Emerging Landscape of LLM-Enabled Collaborative Learning

Since this architecture was first conceptualized, the rapid diffusion of LLMs has materially changed the technical landscape in which a socially adaptive ITS would be built (Grassini, 2023), and the framework's positioning relative to this landscape needs to be made explicit rather than left implicit. Three developments are particularly consequential.

First, LLMs have made several components of the architecture considerably more tractable than they were when comparable collaborative ITS were first proposed. The discourse-analysis functions required by the Interpersonal and Group Interaction Modeling Layer (Section 4.3), detecting explanation quality, questioning acts, and negotiation of meaning, are precisely the kind of open-ended language-understanding task at which LLMs now perform credibly, whereas earlier collaborative ITS relied on brittle, rule-based or shallow statistical discourse classifiers (Kasneci et al., 2023; Ouyang & Zhang, 2024). This suggests that the interaction-modeling layer, previously the architecture's most speculative component (Section 6.4), is now the component with the clearest implementation path.

Second, as anticipated in Section 3.6, a fast-growing line of work now uses LLM-based conversational agents directly as collaborative-learning facilitators (Cai et al., 2024; Wang et al., 2025; Zhang et al., 2024) as retrieval-augmented course tutors (Modran et al., 2024), as simulated students for teacher training (Markel et al., 2023), or as co-orchestrators of transitions between individual and group work (Yang et al., 2023). The multimodal interaction layer (Section 4.5) could plausibly be implemented using such an agent, provided its outputs remain subject to the MLRE's constraint set C.

Third, this same tractability raises the stakes on the explainability and human-oversight requirements already built into design principle 4 (Section 4.1) and the central decision engine (Section 4.6). LLM-based tutoring agents are documented to be capable of fluent, confident-sounding pedagogical interventions whose underlying reasoning is opaque even to their designers (Kasneci et al., 2023; Türkmen, 2025), and evaluation methodologies for LLM-powered tutors are themselves still maturing (Maurya et al., 2025). Were the multimodal layer of this architecture implemented with an LLM-based agent, the explainability module specified in Section 4.4.1 and the teacher override mechanism in Section 4.6 would shift from being

a desirable design principle to a load-bearing safety requirement, since the fluency of LLM output can otherwise mask exactly the kind of participation-equity failure (Barron, 2003) the architecture is designed to detect and correct.

Two studies published while this article was in preparation reinforce this positioning. Uchoa et al. (2025) propose an Advisory Governance Layer for LLM-powered ITS that negotiates conflicting stakeholder values through specialized agents, addressing an accountability gap similar to the one the MLRE's constraint set C targets, though through multi-agent negotiation rather than a single hard-coded policy filter. Separately, in a randomized experiment, Zhang, Meng, et al. (2026) found that GenAI availability shifts group regulation away from socially shared forms toward more hybrid, AI-directed ones, empirical evidence for the risk Section 4.1's design principles are intended to guard against: an ungoverned conversational agent can crowd out the socially distributed regulation the architecture is designed to support rather than displace. Zhang and Lin (2026) propose a related system pairing a process-focused support agent with a teacher-facing analytics dashboard, though without the MLRE's constraint-and-override mechanism or any offline validation of its components.

In sum, the emergence of LLM-enabled collaborative learning tools does not obsolete the proposed architecture; if anything, it sharpens the case for it, since the harder problem is no longer generating fluent group-facing dialogue but constraining it pedagogically, making it explainable, and keeping a human able to intervene, precisely the function the MLRE and the central decision engine are designed to serve.

6.3. Practical Implications for Deployment

While the architecture has not yet been deployed, its design carries implications, and limitations, for how institutions, teachers, and existing infrastructure might engage with it. For institutions, adoption would most plausibly proceed incrementally: the individual learner modeling layer (Section 4.2) could integrate into platforms that already perform knowledge tracing, with the interaction-modeling and regulation layers (Sections 4.3–4.4) added once Section 5.9's Layer 2 gap has been closed. This path is not free: real-time discourse analysis and RL-based decision-making across many simultaneous groups carry a computational cost not benchmarked here, and whether commodity classroom hardware can support this at scale, or whether it requires cloud inference with attendant latency and connectivity dependence, remains an open engineering question. Deployment and support costs may also be prohibitive for under-resourced schools, meaning a system designed to promote equity within groups could, at the institutional level, disproportionately reach already better-resourced classrooms.

For teachers, the system is intended as a pedagogical assistant rather than an autonomous decision-maker (Section 4.6), with the dashboard surfacing live indicators and rationale usable during, not only after, a class session. This intent carries its own risk: the dashboard adds a monitoring task to an already demanding role, and without deliberate interface design and training on interpreting and overriding suggested interventions, not merely operating the software, a well-intentioned oversight mechanism risks becoming an ignored notification stream, undermining the safeguard it is meant to provide.

Technically, the architecture is deliberately agnostic about the specific learning-management platform it would sit within; no compatibility claims are made here. The multimodal layer's reliance on a conversational agent (Section 4.5) makes it a natural candidate for eventual LLM-based delivery, provided, as argued in Section 6.2, that agent's outputs remain constrained by the MLRE. Finally, continuous capture of group discourse for real-time analysis raises data-governance obligations (e.g., FERPA in the United States, GDPR in the European Union) beyond those discussed in Section 5.7 for the proposed field study, since sustained operational logging differs from a bounded research protocol in scope and retention.

6.4. Methodological Reflections and Limitations

While the proposed research design provides a robust foundation for empirical evaluation, several limitations must be acknowledged. First, the quasi-experimental design, although appropriate for authentic educational settings, limits causal inference compared to randomized controlled trials.

Second, the operationalization of group-level constructs such as cohesion or shared regulation relies on proxies derived from interaction data, which may not fully capture the richness of social processes.

A further limitation concerns the offline validation reported in Section 5.9, which establishes technical feasibility without establishing pedagogical effectiveness. It substantively grounds two of the architecture's four layers, individual learner modeling (Section 5.9.1) and the Multi-Level Regulation Engine's decision logic (Section 5.9.2), and offers two narrow algorithmic pilots of a third, the Interpersonal and Group Interaction Modeling layer (Sections 5.9.3–5.9.4), which is the architecture's central novelty. Those pilots are not a substitute for domain-matched validation: they show that the participation-graph algorithm runs on authentic noisy workplace discourse and that a questioning-act heuristic has modest structural validity on real K–12 classroom transcripts, but neither establishes whether Layer 2's outputs are valid indicators of collaboration quality among students working in small groups, and neither tests the group-state classifier. This is not a minor omission: because knowledge tracing is a capability well established in prior ITS, validating that discourse and participation-equity indicators can be reliably extracted from authentic small-group educational interaction matters more than either of the two layers already tested. Validation on authentic collaborative-learning traces (e.g., the VMT corpus or PISA 2015 CPS logs) therefore remains a necessary next step before the architecture's central claim can be considered even partially supported.

7. Conclusion and Future Research

This article advances research on educational artificial intelligence by proposing a socially adaptive Intelligent Tutoring System architecture, centered on the Multi-Level Regulation Engine, that moves beyond the individual-centric paradigm dominating much of the ITS literature toward regulating learning at the individual, dyadic, and group levels. Two of the architecture's four layers are already empirically grounded through offline validation (Section 5.9), establishing, together with the Design Science Research evaluation plan (Section 5), a concrete pathway from formal specification to full empirical validation.

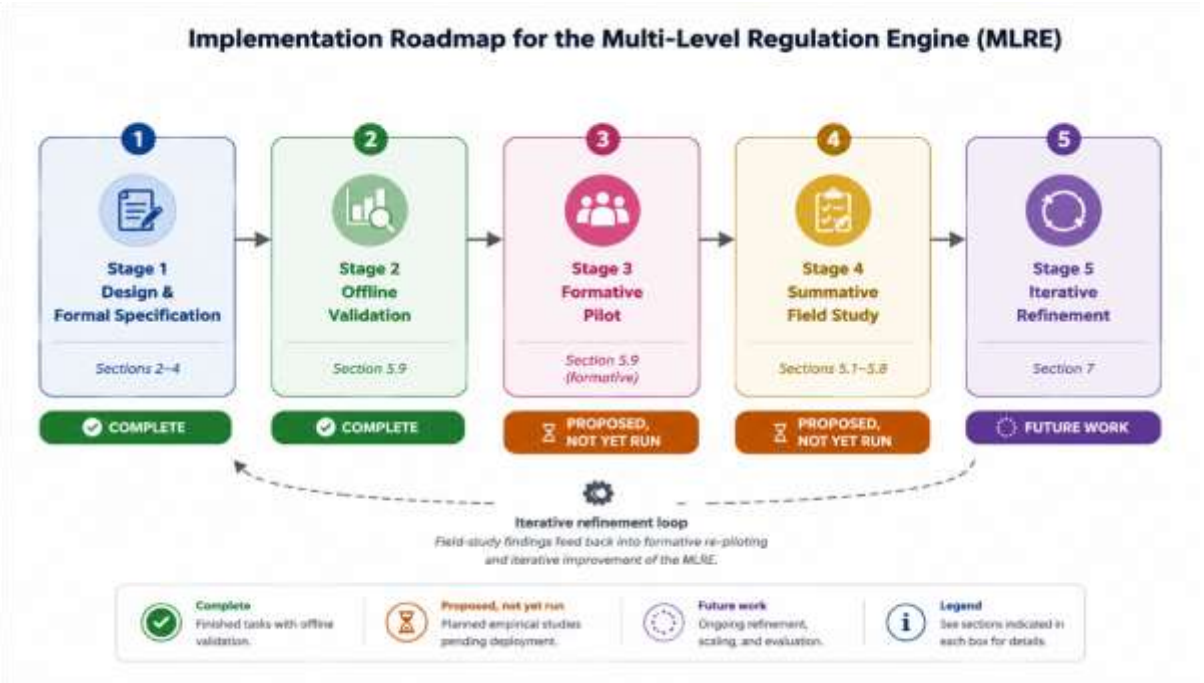


Fig. 3 Implementation roadmap from formal specification to classroom deployment. Stages 1–2 (design and offline validation) are complete and reported in this article; Stages 3–5 (formative pilot, summative field study, and iterative refinement) are proposed future work

Future research will focus on four main directions. First, and most urgently, offline validation of the Interpersonal and Group Interaction Modeling layer using authentic multi-party interaction traces (e.g., the VMT corpus or PISA 2015 Collaborative Problem Solving logs) is needed, since this layer remains the architecture's central novelty and its least empirically grounded component (Section 6.4). Second, the formative and summative field studies specified in Section 5 are needed to validate the pedagogical effectiveness of the full, integrated system in diverse educational contexts and disciplines. Third, further work is required to refine explainable AI mechanisms that allow educators and learners to understand and trust system decisions. Fourth, longitudinal research should examine how sustained exposure to socially adaptive ITS influences learners' collaborative competencies, self-regulation, and sense of inclusion.

Ultimately, the promise of educational AI lies not in replacing human educators but in augmenting their capacity to orchestrate complex, inclusive learning environments (Jeon & Lee, 2023). By embedding social awareness, ethical oversight, and human-centered design into ITS architectures, this work contributes to a vision of future systems that act as collaborative pedagogical orchestrators: intelligent, but also equitable, collaborative, and profoundly human, regulating group learning while preserving transparency and human oversight over every decision they make.

References

- Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. ACM Computing Surveys, 55(11), 1–37. https://doi.org/10.1145/3569576
- Abdelshiheed, M., Hostetter, J. W., Barnes, T., & Chi, M. (2023). Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems. In Proceedings of the 24th International Conference on Artificial Intelligence in Education (AIED 2023) (pp. 291–303). Springer.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4(2), 167–207.
- Baker, M., & Reimann, P. (2024). CSCL: A learning and collaboration science? International Journal of Computer-Supported Collaborative Learning, 19(3), 273–281. https://doi.org/10.1007/s11412-024-09428-1
- Bandura, A. (1977). Social learning theory. Prentice Hall.
- Barron, B. (2003). When smart groups fail. Journal of the Learning Sciences, 12(3), 307–359.
- Bernacki, M. L., Byrnes, J. P., & Cromley, J. G. (2014). Creating collaborative learning groups in intelligent tutoring systems. Educational Technology Research and Development, 62(4), 529–548.
- Cai, Z., Park, S., Nixon, N., & Doroudi, S. (2024). Advancing knowledge together: Integrating large language model-based conversational AI in small group collaborative learning. In Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24). ACM. https://doi.org/10.1145/3613905.3650868
- Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI meeting corpus: A pre-announcement. In S. Renals & S. Bengio (Eds.), Machine Learning for Multimodal Interaction (MLMI 2005), Lecture Notes in Computer Science, vol. 3869 (pp. 28–39). Springer. https://doi.org/10.1007/11677482_3

- Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling. User Modeling and User-Adapted Interaction, 22(1–2), 9–38.
- Dillenbourg, P. (1999). Collaborative learning: Cognitive and computational approaches. Pergamon.
- Doise, W., & Mugny, G. (1984). The social development of the intellect. Pergamon.
- Edwards, J., Nguyen, A., Lämsä, J., Sobocinski, M., Whitehead, R., Dang, B., Roberts, A. S., & Järvelä, S. (2025). Human-AI collaboration: Designing artificial agents to facilitate socially shared regulation among learners. British Journal of Educational Technology, 56(2), 712–733. https://doi.org/10.1111/bjet.13548
- Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. User Modeling and User-Adapted Interaction, 19(3), 243–266. https://doi.org/10.1007/s11257-009-9063-7
- Graesser, A. C., Hu, X., & Sottilare, R. (2018). Intelligent tutoring systems. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (2nd ed., pp. 451–473). Cambridge University Press.
- Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. Education Sciences, 13(7), 692. https://doi.org/10.3390/educsci13070692
- Hadwin, A. F., Järvelä, S., & Miller, M. (2018). Self-regulation, co-regulation, and socially shared regulation of learning. Educational Psychologist, 53(4), 1–14.
- Haq, I., Khan, S., & Ahmed, R. (2021). Dynamic group formation with intelligent tutor collaborative learning. Computers & Education, 175, 104339.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625
- Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. Education and Information Technologies, 28(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1
- Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. Educational Researcher, 38(5), 365–379.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274
- Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing? In Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016) (pp. 94–101). International Educational Data Mining Society.
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074
- Kim, Y., & Baylor, A. L. (2016). Social agents for learning: Design and evaluation of pedagogical conversational agents. Educational Technology Research and Development, 64(5), 765–785.
- Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (pp. 61–78). Cambridge University Press.
- Kollar, I., Fischer, F., & Hesse, F. W. (2006). Collaboration scripts – A conceptual analysis. Educational Psychology Review, 18(2), 159–185.

- Li, X., Hu, W., & Li, Y. (2025). Effect of adaptable and non-adaptable collaboration scripts through conversational agents on students' engagement in online collaborative learning. Educational Technology Research and Development, 73, 875–907. https://doi.org/10.1007/s11423-024-10442-3
- Lin, C.-C., Huang, A. Y., & Lu, O. H. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. Smart Learning Environments, 10(1), 41. https://doi.org/10.1186/s40561-023-00260-y
- Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). GPTeach: Interactive TA training with GPT-based students. In Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23) (pp. 226–236). ACM. https://doi.org/10.1145/3573051.3593393
- Matsuda, N., & VanLehn, K. (2016). Improving student learning with hybrid human–AI tutoring systems. In Proceedings of the 17th International Conference on Artificial Intelligence in Education (pp. 145–156). Springer.
- Maurya, K. K., Srivatsa, K. A., Petukhova, K., et al. (2025). Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 1234–1251). Association for Computational Linguistics.
- Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence, 5, 100152. https://doi.org/10.1016/j.caeai.2023.100152
- Miller, M., & Hadwin, A. F. (2024). Comparing the effectiveness of CSCL scripts for shared task perceptions in socially shared regulation of collaborative learning. International Journal of Computer-Supported Collaborative Learning, 19(4), 455–478. https://doi.org/10.1007/s11412-024-09433-4
- Modran, H. A., Bogdan, I. C., Ursuțiu, D., de Fátima Chouzal, M., & Restivo, M. T. (2024). LLM intelligent agent tutoring in higher education courses using a RAG approach. In Proceedings of the International Conference on Interactive Collaborative Learning (pp. 589–599). Springer. https://doi.org/10.1007/978-3-031-71906-5_58
- Ouyang, F., Xu, W., & Cukurova, M. (2023). An artificial intelligence-driven learning analytics method to examine the collaborative problem-solving process from the complex adaptive systems perspective. International Journal of Computer-Supported Collaborative Learning, 18, 39–66. https://doi.org/10.1007/s11412-023-09387-z
- Ouyang, F., & Zhang, L. (2024). AI-driven learning analytics applications and tools in computer-supported collaborative learning: A systematic review. Educational Research Review, 44, 100616. https://doi.org/10.1016/j.edurev.2024.100616
- Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In Advances in Neural Information Processing Systems 28 (NeurIPS 2015) (pp. 505–513). Curran Associates.
- Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. E. O'Malley (Ed.), Computer-supported collaborative learning (pp. 69–97). Springer.
- Shen, S., Liu, Q., Huang, Z., Zheng, Y., Yin, M., Wang, M., & Chen, E. (2024). A survey of knowledge tracing: Models, variants, and applications. IEEE Transactions on Learning Technologies, 17. Advance online publication. https://doi.org/10.1109/TLT.2024.3383325

- Stahl, G. (2006). Group cognition: Computer support for building collaborative knowledge. MIT Press.
- Suresh, A., Sumner, T., Huang, I., Jacobs, J., Foland, B., & Ward, W. (2019). Automating analysis and feedback to improve mathematics teachers' classroom discourse. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 9721–9728. https://doi.org/10.1609/aaai.v33i01.33019721
- Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. Journal of Educational Computing Research. Advance online publication. https://doi.org/10.1177/07356331241310915
- Uchoa, A. P., Oliveira, C. E. T., Motta, C. L. R., & Schneider, D. (2025). Multi-stakeholder alignment in LLM-powered collaborative AI systems: A multi-agent framework for intelligent tutoring. arXiv preprint arXiv:2510.23245. https://doi.org/10.48550/arXiv.2510.23245
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. European Journal of Information Systems, 25(1), 77–89. https://doi.org/10.1057/ejis.2014.36
- Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.
- Wang, T., Zhan, Y., Lian, J., Hu, Z., Yuan, N. J., Zhang, Q., Xie, X., & Xiong, H. (2025). LLM-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. In Companion Proceedings of the ACM Web Conference 2025 (pp. 510–519). ACM. https://doi.org/10.1145/3701716.3715244
- Webb, N. M. (2009). The teacher's role in promoting collaborative dialogue. Journal of Educational Psychology, 101(1), 1–17. https://doi.org/10.1037/a0013848
- Woolf, B. P. (2010). Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. Morgan Kaufmann.
- Xiong, X., Zhao, S., Van Inwegen, E. G., & Beck, J. E. (2016). Going deeper with deep knowledge tracing. In Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016) (pp. 545–550). International Educational Data Mining Society.
- Yang, K. B., Echeverria, V., Lu, Z., Mao, H., Holstein, K., Rummel, N., & Aleven, V. (2023). Pair-Up: Prototyping human-AI co-orchestration of dynamic transitions between individual and collaborative learning in the classroom. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–17). ACM. https://doi.org/10.1145/3544548.3581512
- Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Hao, Z., Jiang, J., Cao, J., Liu, H., Liu, Z., Hou, L., & Li, J. (2024). Simulating classroom education with LLM-empowered agents. arXiv preprint arXiv:2406.19226. https://doi.org/10.48550/arXiv.2406.19226
- Zhang, Y., & Lin, J. (2026). Building regulation capacity in human-AI collaborative learning: A human-centred GenAI system. arXiv preprint arXiv:2604.10221. https://doi.org/10.48550/arXiv.2604.10221
- Zhang, Y., Meng, X., Feng, S., & Lin, J. (2026). Human-AI collaboration reconfigures group regulation from socially shared to hybrid co-regulation. arXiv preprint arXiv:2604.08344. https://doi.org/10.48550/arXiv.2604.08344