Toward Universal Collaborative Intelligence for Explainable Pedagogy and Next-Generation Assessment

January 31, 2026

1 Illustrative Component: Socratic Playground (SPL)

The SPL is a representative component of UALS that exemplifies how philosophy-aware, multi-agent orchestration can support rich, dialogue-based learning. As illustrated in Figure 1, the SPL interface integrates dialogue-based tutoring, hinting, and progress indicators within a single unified workspace. SPL is designed as a multi-modal conversational environment that leverages AI-guided Socratic questioning and collaborative discovery to promote deep conceptual understanding, critical thinking, and metacognitive skill development. Within the broader UALS architecture, SPL inherits philosophy-aware prompts, ITS-agent recommendations, and learner-model updates, and instantiates them in the form of adaptive dialogue, targeted feedback, and fine-grained learning analytics.

1.0.1 Pedagogical Modes and Interaction Design

SPL implements five pedagogical modes that address diverse learning purposes and learner profiles:

- **Tutoring mode** provides guided instruction with graduated scaffolds that adapt to learner knowledge states. It offers strategic hints, worked examples,

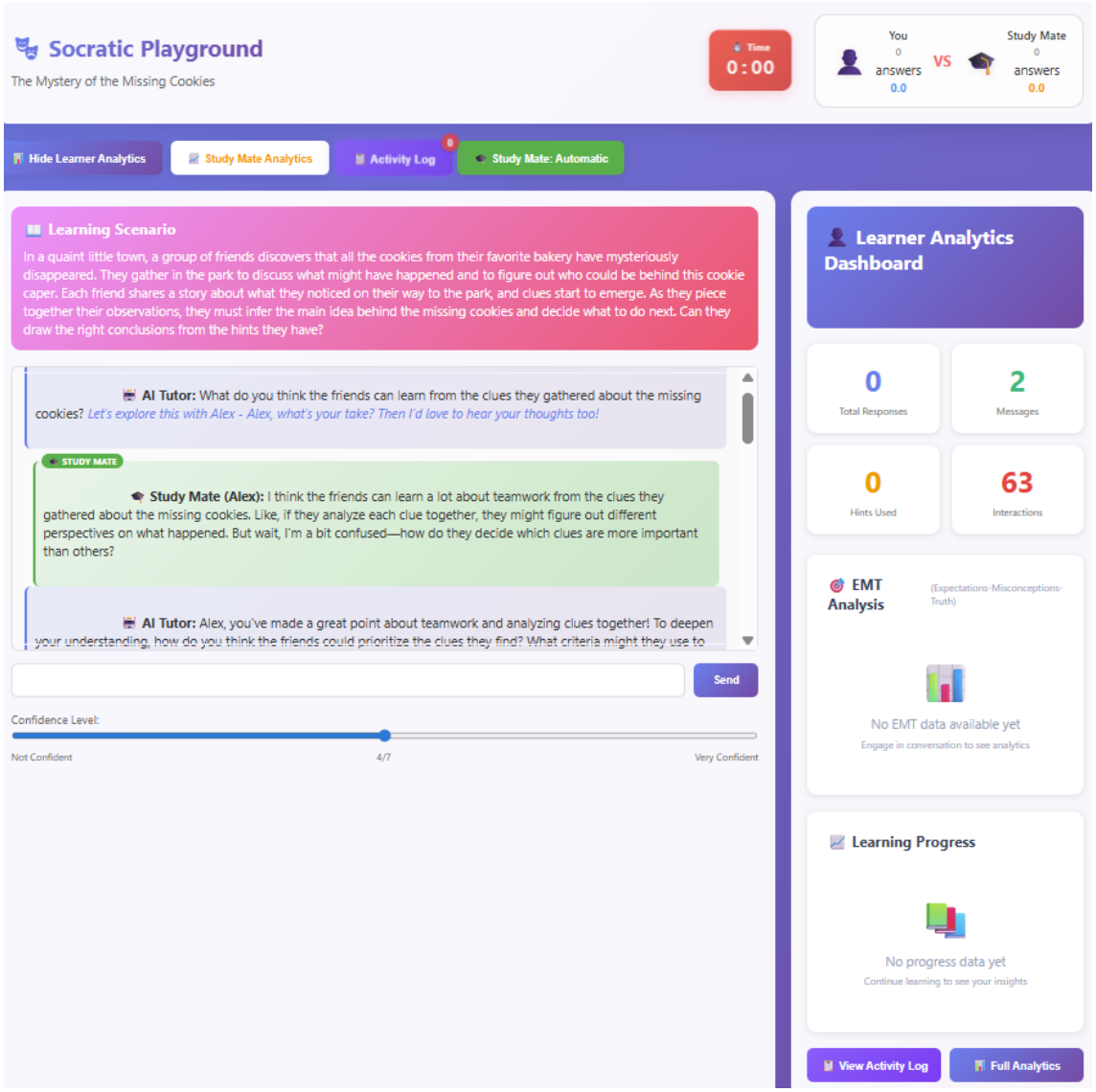

Figure 1: Screenshot of the SPL within UALS.

and step-by-step guidance calibrated to the learner's current level of understanding.

- **Assessment mode** focuses on eliciting and evaluating conceptual understanding through evidence-centered questioning. It probes reasoning processes, identifies knowledge gaps, and measures mastery using targeted diagnostic prompts rather than only surface-level correctness.

- **Vicarious Learning mode** supports learning through observation by presenting worked examples, expert demonstrations, and annotated solution paths. Learners can internalize effective reasoning strategies and discourse moves be-

fore attempting parallel tasks on their own.

- **Gaming mode** gamifies learning through challenge-based activities, reward structures, leaderboards, and competitive elements, with the goal of increasing engagement, persistence, and time-on-task while maintaining alignment with learning objectives.

- **Teachable Agent mode** reverses the traditional tutor–student relationship by positioning learners as teachers who explain concepts to an AI "student". By leveraging the principle that "to teach is to learn twice", it requires learners to articulate their reasoning, justify their decisions, and actively identify and correct the AI agent's intentional errors or misconceptions.

Across these modes, SPL flexibly adapts its interaction style to the active educational philosophy: for example, more open-ended Socratic questioning in constructivist contexts, more structured, stepwise guidance in behaviorist contexts, and more metacognitive prompts and reflection questions in cognitivist contexts.

1.0.2 Multi-Agent Orchestration in SPL

To realize these interaction patterns, SPL orchestrates a team of specialized agents that collaborate in real time:

- The **Tutor Agent** functions as the primary instructor, delivering explanations, conceptual overviews, worked examples, and scaffolded guidance. Its behavior is conditioned on the inferred philosophy and learner state (e.g., more inquiry-based prompts for advanced learners, more explicit guidance for novices).

- The **Hint Agent** implements a four-level hint progression: (i) *strategic hints* offer high-level guidance about problem-solving approaches without revealing specific steps; (ii) *tactical hints* point learners to relevant sub-problems, representations, or concepts; (iii) *operational hints* provide detailed, step-by-step

suggestions for immediate next actions; and (iv) *bottom-out hints* reveal complete solutions, reserved for cases where learners remain stuck after multiple attempts. This progression is designed to preserve productive struggle while preventing prolonged frustration.

- The **Socratic Agent** specializes in inquiry-driven dialogue, using carefully sequenced questions to challenge assumptions, probe reasoning, surface contradictions, and scaffold conceptual discovery rather than simply telling the answer.

- The **Study Mate Agent** simulates a peer collaborator, enabling trilogue interactions (learner + AI tutor + AI study mate). It offers alternative perspectives, poses clarifying questions from a peer viewpoint, and models collaborative problem-solving behaviors that support social learning and distributed cognition.

- The **Feedback Agent** delivers immediate, tailored feedback that goes beyond correctness flags. It generates explanatory feedback (why an answer is correct or incorrect), diagnostic feedback (which misconception or reasoning flaw is implicated), and corrective feedback (what to do next, including suggestions for targeted remediation).

These agents draw on shared context from the UALS learner model and the EMT framework (Section 1.0.4), ensuring that hints, questions, and feedback are consistent with the learner's history, current goals, and targeted competencies.

1.0.3 Learning Analytics and EMT-Enhanced Monitoring

The SPL learning analytics dashboard provides real-time visibility into dialogue-based learning processes. It tracks metrics such as session duration (total time engaged in SPL), engagement indicators (e.g., response latency, message length, question-asking frequency, hint requests), response counts (number and type of

learner utterances, distinguishing substantive explanations from brief acknowledgements), and message exchanges (dialogue turns between learners and agents). It also logs hint usage (frequency and hint level), interaction patterns over time (identifying stretches of intense reasoning versus disengagement), and EMT analyses that summarize how learner responses align with expert expectations, which misconceptions are active, and how understanding evolves across turns.

Progress visualizations aggregate these signals over multiple SPL sessions, highlighting concepts that have been mastered, misconceptions that persist and may require intervention, and difficulty zones that are optimal for future challenge calibration. These analytics inform both real-time adaptation (e.g., mode switching, hint escalation, agent role adjustment) and offline reflection for teachers and researchers. Figure 2 depicts the EMT feedback framework within SPL, showing how student responses are analyzed and transformed into adaptive feedback and long-term model updates.

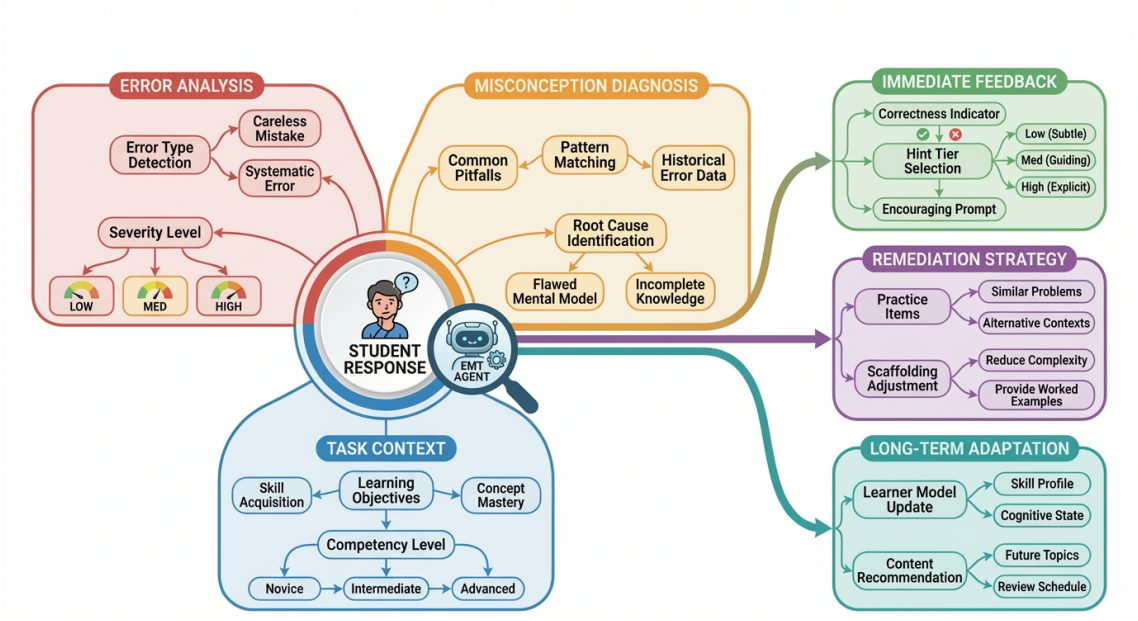

Figure 2: The EMT feedback framework within UALS SPL.

1.0.4 EMT Framework: Expectation–Misconception–Tailored Feedback

The EMT framework serves as the cognitive and diagnostic backbone of adaptive instruction within SPL. When a learner submits a response (e.g., explanation, solution attempt, or dialogue turn), EMT processes it through three interconnected diagnostic pathways:

- The **Error Analysis Pathway** identifies surface-level errors, such as computational mistakes, procedural missteps, and notational or syntactic errors. It classifies errors by severity (critical vs. minor) and systematicity (isolated slips vs. recurring patterns).
- The **Misconception Diagnosis Pathway** probes deeper conceptual understanding by comparing learner responses against a misconception library derived from educational research. Using pattern matching and semantic analysis, it infers specific misconceptions (e.g., "heavier objects fall faster") that may explain observed error patterns.
- The **Task Context Pathway** evaluates response appropriateness relative to task demands, learner history, and current learning objectives, incorporating contextual features such as time on task, prior performance on related concepts, and inferred affective states (frustration, boredom, confusion, engagement).

These pathways converge to generate three categories of adaptive outputs:

- **Immediate feedback**, which provides real-time correctness judgments (e.g., correct/ incorrect/ partially correct), brief explanations, and motivational messages that sustain engagement (e.g., "Good reasoning—now check this step" or "Not quite—let's work through this together").
- **Remediation strategies**, which specify targeted instructional responses based on diagnosed issues, such as presenting worked examples for procedural errors, offering conceptual explanations and analogies for misconception-driven

errors, invoking Socratic questioning sequences for shallow understanding, recommending practice items for consolidation, or triggering prerequisite review for foundational gaps.

- **Long-term adaptation**, which updates the learner model to reflect demonstrated knowledge, persistent misconceptions, and optimal challenge levels. These updates influence future content selection (topic and difficulty), pedagogical mode (e.g., shifting from Assessment to Tutoring mode when struggle is detected), and agent behavior (e.g., increasing Socratic Agent involvement for advanced learners, emphasizing the Hint Agent for novices).

EMT thus enables learner-state-responsive instruction. Advanced learners who respond quickly and accurately receive more challenging Socratic prompts that push them to justify reasoning and explore alternative solutions. Novice learners who make frequent errors or request many hints receive more structured scaffolds, including stepwise hints and annotated worked examples. Learners exhibiting behavioral signs of frustration or disengagement receive affective support, temporary difficulty reduction, and carefully managed re-entry into more challenging tasks. Through EMT, SPL operationalizes a principled linkage between diagnosis, feedback, and multi-agent behavior, providing a fine-grained, theory-grounded example of how UALS implements explainable adaptive learning.

2 The EMT Framework in the Era of Large Language Models

The EMT framework changes in important ways when implemented with LLMs, with significant implications for engineering practice, reliability expectations, and the evolving role of human expertise.

The Paradigm Shift from Engineering to Intelligence. The emergence of LLM-powered EMT represents a fundamental reconceptualization of adaptive

feedback—not merely an efficiency improvement but a qualitative transformation in how intelligent tutoring systems understand and respond to learner reasoning. Traditional ITS development operated under what might be called the "engineering paradigm": expert knowledge meticulously encoded into production rules, error taxonomies painstakingly constructed through empirical studies, and feedback templates carefully crafted for anticipated misconceptions. This approach achieved impressive results in constrained domains—Cognitive Tutors demonstrating learning gains in high school mathematics [1], AutoTutor improving physics comprehension through conversational dialogue [2]—yet it fundamentally could not scale beyond domains where extensive engineering investment proved economically viable. The critical insight of LLM-powered EMT is recognizing that foundation models' pretraining on internet-scale corpora has already absorbed much of the domain knowledge, misconception patterns, and pedagogical strategies that traditional ITS laboriously encoded by hand. This shifts the challenge from "how do we encode expert knowledge?" to "how do we prompt and guide LLMs to apply their latent pedagogical understanding appropriately?"—a dramatically more scalable proposition with profound implications for educational equity (enabling high-quality adaptive tutoring in under-resourced subjects and languages) and innovation velocity (reducing ITS development from years to weeks).

The Tension Between Generalization and Reliability. Yet this shift introduces a fundamental tension absent in rule-based systems: LLMs' remarkable generalization capabilities emerge from statistical patterns rather than verified logical rules, creating uncertainty about response reliability that traditional ITS never faced. A hand-crafted production rule stating "if student solves $2x+3=72x+3=7$ as $x=5x=5$, diagnose arithmetic error in final subtraction step" executes deterministically—always producing identical diagnoses for identical inputs, never inventing creative but incorrect explanations, never contradicting previous feedback. LLM-powered EMT sacrifices this deterministic reliability for probabilistic generalization: the same student error might receive subtly different diagnoses across

sessions, occasionally hallucinating confident but incorrect explanations (claiming, perhaps, that the student's error reflects a "common misconception about equation balancing" when it's simply arithmetic carelessness), and potentially providing inconsistent remediation strategies that confuse rather than clarify. This raises profound questions about trust and accountability in educational AI: Should we accept occasional diagnostic errors as an acceptable cost of unprecedented scalability? How do we audit and validate LLM feedback at scale when traditional approaches (expert review of hand-crafted rules) become infeasible? The UALS mitigation strategies—RAG grounding, multi-LLM consensus, human-in-the-loop oversight—represent pragmatic compromises, but the deeper question remains unresolved: Is probabilistic intelligence fundamentally compatible with the deterministic reliability traditionally expected of educational assessment and feedback systems, or must we reconceptualize educational AI evaluation criteria to accommodate inherent uncertainty?

Implications for Human Expertise and Educational Labor. Perhaps most provocatively, LLM-powered EMT forces reconsideration of human expertise's role in educational technology development and delivery. Traditional ITS positioned domain experts and educational researchers as indispensable architects—their knowledge crystallized into production rules, their pedagogical wisdom encoded into dialogue strategies, their understanding of misconceptions formalized into error taxonomies. LLM-powered EMT suggests an alternative model where human expertise shifts from encoding to orchestration: rather than manually constructing knowledge representations, experts design prompts, curate training examples, validate outputs, and intervene when LLMs fail. This transformation parallels broader shifts in knowledge work—from writing to editing, from creation to curation, from implementation to oversight—raising both opportunities (freeing experts from tedious rule-writing to focus on higher-level pedagogical design) and concerns (deskilling risks, quality control challenges, accountability gaps when errors emerge from opaque neural networks rather than traceable logical rules). For the educational technology research

community, this demands new evaluation methodologies: moving beyond traditional ITS metrics (knowledge tracing accuracy, hint effectiveness, completion rates) to assess LLM-specific qualities like feedback consistency, hallucination rates, pedagogical appropriateness of generated explanations, and long-term learning outcomes from probabilistic rather than deterministic tutoring. As UALS scales across diverse domains, populations, and learning contexts, empirical research must address fundamental questions: Does LLM-powered adaptive tutoring produce equivalent learning gains to hand-crafted ITS in well-engineered domains? Does it enable effective tutoring in previously inaccessible domains? How do learners perceive and respond to probabilistic feedback uncertainty compared to deterministic rule-based guidance?

Validation and Robust Deployment. As UALS scales to many agents and longer dialogues, errors such as misconception misclassification, drift from lesson goals, or plausible but weakly grounded feedback can become more consequential. Prior work documents systematic computation and reasoning failures in LLM outputs, underscoring the need for verification in educational use cases [3]. Accordingly, UALS can incorporate validation and cross-checking (e.g., verifier/critic agents, evidence requirements tied to rubrics or retrieved materials, and consistency checks against the learner model), along with fallback strategies such as clarification questions, conservative Socratic prompting, simplified response pathways, or educator review under high uncertainty.

References

[1] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. Intelligent tutoring goes to school in the big city. *International journal of artificial intelligence in education*, 8:30–43, 1997.

[2] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE*

Transactions on Education, 48(4):612–618, 2005.

[3] Liang Zhang and Edith Graf. Mathematical computation and reasoning errors by large language models. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 417–424, 2025.