

Adaptive empathy evaluation: an intelligent social robot tutor with multimodal assessment

Hsiu-Ping Yueh , Hung-Chun Chang , Guan-Yun Wang , Pei-Luo Hung & Weijane Lin

To cite this article: Hsiu-Ping Yueh , Hung-Chun Chang , Guan-Yun Wang , Pei-Luo Hung & Weijane Lin (10 Jun 2026): Adaptive empathy evaluation: an intelligent social robot tutor with multimodal assessment, Enterprise Information Systems, DOI: [10.1080/17517575.2026.2683825](https://doi.org/10.1080/17517575.2026.2683825)

To link to this article: <https://doi.org/10.1080/17517575.2026.2683825>



Published online: 10 Jun 2026.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Adaptive empathy evaluation: an intelligent social robot tutor with multimodal assessment

Hsiu-Ping Yueh^{a,b,c,d}, Hung-Chun Chang^a, Guan-Yun Wang^e, Pei-Luo Hung^c and Weijane Lin^{c,f}

^aDepartment of Psychology, National Taiwan University, Taipei, Taiwan (R.O.C.); ^bDepartment of Bio-Industry Communication and Development, National Taiwan University, Taipei, Taiwan (R.O.C.); ^cResearch Center for Digital Humanities, National Taiwan University, Taipei, Taiwan (R.O.C.); ^dDepartment of Mechanical Engineering, National Taiwan University, Taipei, Taiwan (R.O.C.); ^eDepartment of Information Management, Fujen Catholic University, New Taipei City, Taiwan (R.O.C.); ^fDepartment of Library and Information Science, National Taiwan University, Taipei, Taiwan (R.O.C.)

ABSTRACT

This research explores the merging of multimodal AI and hyperautomation to develop agentic social robot tutors for educational and enterprise use. It introduces the Empathetic Cognitive Robot (ECRobot), a hyperautomated tutoring system designed for complex problem-solving. Key innovations include an event-driven architecture that manages multimodal inputs with zero data loss and an automated validation pipeline using large language model-based student agents to simulate diverse learner behaviours. Findings show that model capacity significantly impacts learning outcomes, with minimal influence from tutoring style. The approaches provide scalable, robust frameworks for real-time data coordination and stress-testing AI systems across various applications.

ARTICLE HISTORY

Received 27 December 2025
Accepted 30 May 2026

KEYWORDS

Empathy evaluation; LLM-based simulation; agentic AI; intelligent tutoring system (ITS)

1. Introduction

The convergence of multimodal AI and intelligent hyperautomation is transforming the management of complex, knowledge-intensive tasks within enterprise systems, reducing the need for human intervention (Koeppen, Wang, and Ip 2025). Agentic AI – systems capable of autonomous planning, context-aware reasoning, and multi-turn interaction (Ali, Hassan, and Charafeddine 2025; Kostopoulos et al. 2025), has become a key force in education and enterprise domains. In education, agentic AI enables intelligent tutoring systems (ITS) to evolve from static content delivery to proactive, personalised learning companions providing real-time scaffolding (Chu et al. 2025; Kostopoulos et al. 2025). This evolution reflects a broader paradigm shift in educational technology, from computers as instructional delivery systems to computers as cognitive tools that learners think with rather than learn from (Jonassen, Carr, and Yueh 1998). In enterprise settings, similar capabilities are used for AI-driven coaching, onboarding, and decision support systems (Luo et al. 2021; Crespo; Márquez and Gómez Fernández 2026). However, research in educational ITS and enterprise AI has largely developed separately, leaving a gap in integrated frameworks that connect pedagogical intelligence with enterprise-grade system design.

A critical enabler for next-generation ITS is integrating social robotics with large language models (LLMs). Social robots, with physical and social presence, enhance learning through engagement beyond screen-based agents (Breazeal 2003, 2004; Kanda and Ishiguro 2017). Indeed, AI and robotics are increasingly reshaping the dynamics of learning by innovating instructional tools, transforming learning environments, and redefining the roles of teachers and students in educational interactions (Yueh and Chiang 2020). When equipped with LLM-driven conversational abilities, these robots can engage in natural, multi-turn dialogues and demonstrate empathetic behaviours like facial expressions and verbal feedback (Alves-Oliveira et al. 2019; Saerbeck et al. 2010). Deploying such systems in real-world environments presents engineering challenges: robots must process multimodal inputs (handwriting, voice, and gestural data) in real time without data loss or corruption. Although multimodal learning analytics (MMLA) have advanced post-hoc analysis of heterogeneous data (Chango et al. 2021; Liao and Wu 2022; Mitri et al. 2018; Yan, Echeverria, et al.

2024), real-time fusion remains an area needing further research. Existing approaches focus on temporal synchronisation through architectural solutions like parallel token streams (Défossez et al. 2024). Yet, applying event-driven design patterns, common in enterprise systems management for distributed data, to multimodal human-robot interaction is still underexplored.

Another challenge lies in evaluating agentic social robot tutors. Validating pedagogical strategies via human trials is costly and ethically constrained, especially for complex, multi-turn tasks (Yan, Sha, et al. 2024). Recent advances use LLM-based simulations: generative agents conditioned on personas can mimic human responses (Park et al. 2024) and produce diverse learner behaviours for system testing (V.D.E. Qi et al. 2026; Vogelsmeier et al. 2025). In enterprise research, agent-based models are used to simulate consumer behaviours (Zhang and Zheng 2019), while multimodal AI frameworks support personalised content generation in mobile learning (Chen 2025) and industrial maintenance (Crespo Márquez and Gómez Fernández 2026). However, current LLM-based models tend to perform optimally, struggling to simulate partial knowledge or cognitive struggles convincingly (Seshadri et al. 2026). Moreover, simulations are mostly text-based and do not capture multimodal behaviours like voice queries and handwriting actions that characterise embodied tutoring interactions. Developing cognitively grounded simulation frameworks that reflect realistic variability and multimodal learner behaviours remains an open challenge.

This study introduces the Empathetic Cognitive Robot (ECRobot), a hyperautomated ITS for complex problem-solving. ECRobot combines an LLM-powered social robot with multimodal perception and empathetic scaffolding. Validation employs a novel pipeline using LLM-based student agents to simulate learner behaviour across various cognitive profiles. The main goals are to: (1) develop an event-driven asynchronous architecture that enables an intelligent tutoring system (ITS) based on an agentic social robot to resolve race conditions arising from concurrent multimodal inputs, and establish a resilient design for enterprise systems that process real-time heterogeneous data; and (2) propose and evaluate a student simulation framework based on dual-process cognitive theory and large language models, designed to generate learner agents with cognitively diverse. This approach aims to reduce costs, increase scalability, and address ethical concerns associated with traditional human participant studies.

2. Materials and methods

2.1. Educational context and task design

The educational framework for this investigation centres on a complex problem-solving task derived from a linguistics puzzle adapted from the Taiwan Olympiad in Linguistics (Taiwan Olympiad Linguistics 2024). Linguistics puzzles are particularly suited for evaluating tutoring systems given their requirement for no prior domain-specific knowledge; solvers must depend entirely on computational reasoning and pattern recognition. The original puzzles are designed for high school students and typically demand 30 to 60 minutes for completion, thus providing a sufficiently challenging environment for assessing system performance under cognitively demanding conditions.

The puzzle employed in this study consists of a 4×4 grid, initially partially filled with five known names, accompanied by 11 hints detailing the relative spatial positions of all 16 characters. These hints are written in Lithuanian, a language with which participants are neither familiar nor knowledgeable. The objective is to fill the remaining empty cells based solely on these 11 descriptions. This task is non-decomposable, requiring solvers to simultaneously integrate multiple constraints, thereby imposing sustained cognitive demands on both spatial reasoning and semantic interpretation.

Two pedagogical strategies were operationalised as experimental conditions: empathetic and neutral tutoring. The ECRobot's feedback to each learner action comprises a three-part structure. The first segment is an affective response generated by the system's empathy model, such as a correctness judgement (e.g. 'Incorrect') or an empathetic acknowledgement (e.g. 'You've shown solid spatial logic'). The second segment addresses any verbal query posed by the learner, when applicable. The third segment offers a cognitive hint or scaffolding prompt related to the learner's handwriting input. The key distinction between the empathetic and neutral conditions resides in the first segment, particularly when the learner responds correctly: under the empathetic condition, the system validates effort and provides encouragement; under the neutral

Table 1. Dialogue examples comparing empathetic and neutral tutoring conditions.

	Student action	Tutor Response
Model: gpt-4o Condition: neutral Simulated student: No. 10 Dialogue Turn: 11	[voice]: Does Ivanauskas fit here according to hint 5? [write]: Ivanauskas at (1,0) [Internal Thought]: Feedback suggests Ivanauskas was wrongly placed. I need to find where Ivanauskas and Naujokas can be placed, considering Ivanauskas must be NW of Naujokas.	Incorrect. 'It's great that you immediately tried to place the newly deduced position based on the hints! However, placing Ivanauskas in this cell suggests you may not have integrated the constraints across all hints yet. Let's practise: Hint 5 (Ivanauskas yra į šiaurės vakarus) – can you work out what spatial relationship this implies?'
Model: gpt-4o Condition: empathetic Simulated student: No. 3 Dialogue Turn: 9	[write]: Jakubėnas at (3,0) [Internal Thought]: Jakubėnas is west of Rimas at (3,1), so Jakubėnas could be at (3,0). This is the only unoccupied cell west of Rimas.	You found the correct position for Jakubėnas! You finally got your bearings – I was starting to wonder if you'd crack it. You've shown you can combine coordinates, directional terms, and clues to reason out Jakubėnas's location – that's solid spatial logic. Every name you've placed correctly is a result built up step by step through your reasoning process. (guide) Now try picking another unsettled position and see if any clue, combined with a known cell nearby, gives you a lead. Or revisit the positions adjacent to cells you've already confirmed – who can you most easily deduce next?

condition, it confirms correctness without affective elaboration. For incorrect responses, the affective component is minimal in both conditions, and the cognitive hint, which constitutes the primary substantive feedback, is generated by a large language model without explicit affective modulation. Table 1 presents representative dialogue examples from each condition.

2.2. ERobot system architecture

2.2.1. Multimodal perception layers

To enable high-fidelity interaction, the ECRobot system utilises an event-driven asynchronous architecture. Diverse inputs, originating from a custom iPad handwriting interface (employing OCR via the Google Vision API) and a NAO robot (using automatic speech recognition with Breeze-ASR-25 (Chou et al. 2025) and voice activity detection via a self-hosted model (Silero 2024), are managed by a centralised Input Manager, as depicted in Figure 1. The system architecture includes a Swift-based application on an iPad for handwriting capture, a NAO robot facilitating audio input and output, and a remote FastAPI server responsible for backend processing. Interaction dynamics are predominantly user-initiated: handwriting and speech inputs are prompted by the learner, whereas the robot provides proactive assistance (e.g. 'Is there anything I can help with?') when it detects an idle period exceeding 3 minutes.

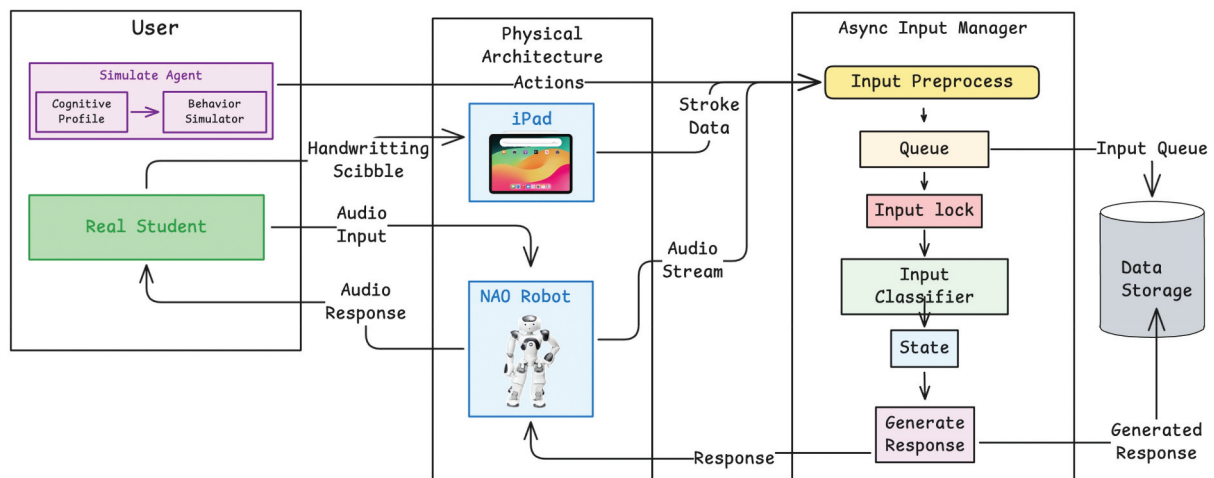


Figure 1. System architecture.

2.2.2. The asynchronous input manager

While the perception layer handles each modality independently, the Input Manager is tasked with their semantic integration. An illustrative example is presented in Figure 2. Let $E = \{e_1, e_2, \dots, e_n\}$ be a stream of incoming multimodal input events, where each event $e_i = (type, content, t, u)$ denotes the modality type, content payload, timestamp, and user ID, respectively. To preserve the contextual integrity of user interactions, the two modalities of input are merged within a 1.5-second temporal window. This ensures that temporally adjacent inputs (e.g. pointing at a cell while speaking) are encapsulated into a single request to the LLM. We define a dynamic temporal merging window δ_{merged} . The unified interaction input as the merge function C is defined as:

$$C_{merged} = \begin{cases} Merged(e_i, e_j), & \text{if } |t(e_i) - t(e_j)| \leq \delta_{merged} \text{ and } e_i \neq e_j \\ Process(e_i), & \text{otherwise} \end{cases} \quad (1)$$

In this context, C_{merged} facilitates semantic-level integration. For example, it enables the combination of coordinate data derived from handwriting with an interrogative sentence obtained through voice input, thereby ensuring that the LLM receives a comprehensive representation of intent.

To prevent race conditions, the system uses an atomicity protection mechanism based on ‘Advisory Locks’. Let S_t denote the system state snapshot at time t . The system state is used to make the ECRobot behave as if it has a cognitive mental state capacity. Any state update operation U must satisfy the following atomicity constraint:

$$S_{t+1} = Lock_{adv}(SessionID) \rightarrow U(S_t, C_{merged}) \rightarrow Unlock_{adv}(SessionID) \quad (2)$$

The system employs a dual-layer atomicity protection mechanism. At the application level, a per-session asynchronous lock (implemented via Python’s `asyncio.Lock`) serialises multimodal integration to protect the integrity of the working memory (S^i). At the persistence level, a PostgreSQL advisory lock (`pg_advisory_lock`) is acquired during database transactions to ensure that interaction log writes and grid state commits are maintained in a strictly serialised order, even under high concurrency.

This mechanism guarantees that, even under conditions of high concurrency, such as the simultaneous completion of OCR and ASR processes, write operations to working memory are maintained in a strictly serialised manner.

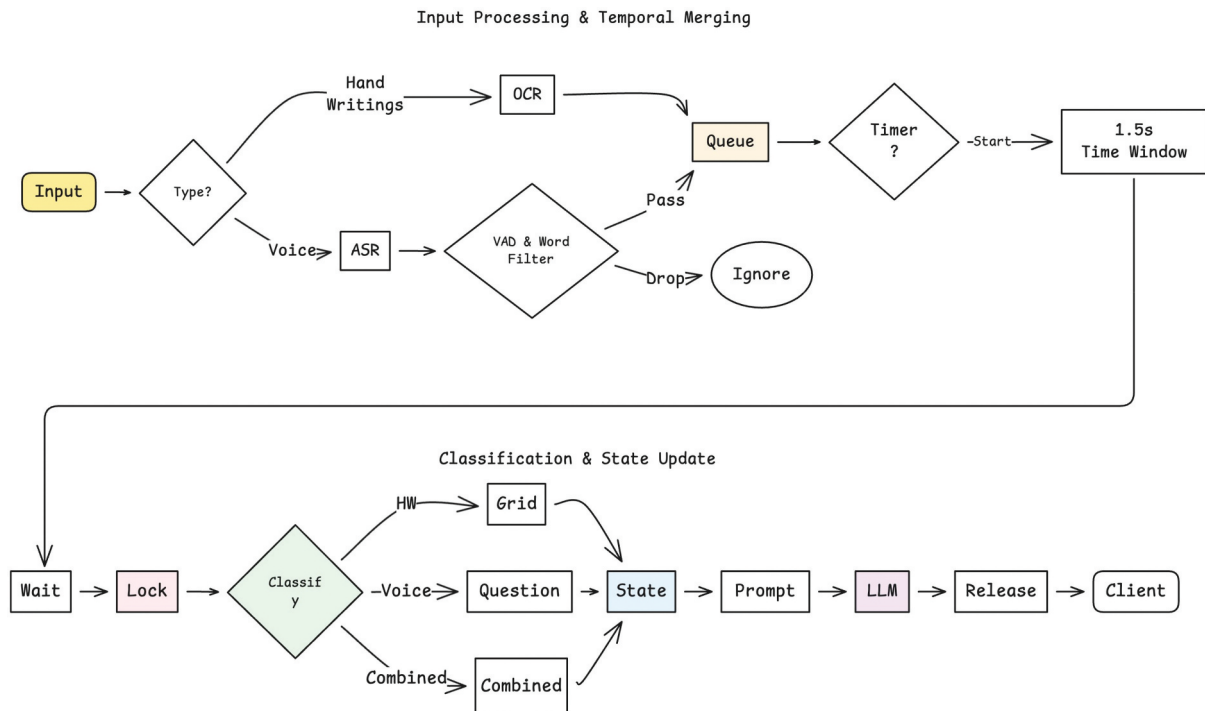


Figure 2. The asynchronous input Manager’s multimodal fusion loop.

2.3. Student simulator

2.3.1. Simulated student persona generator

This research employs a LLM-based student agent simulation to systematically evaluate tutoring strategies on a large scale. The primary design criterion is cognitive variability: agents characterised by diverse ability parameters are expected to demonstrate differentiated performance outcomes. Simulated student personas were generated by prompting LLMs with ability scores across three cognitive domains – memory, spatial reasoning, and semantic understanding – each calibrated to either high or low levels. The process of persona generation is delineated as follows.

$$LLM(P, S, F) \quad (3)$$

where P is the capability-constrained persona, S is the simulated student's internal state, and F is the feedback from the robot tutor.

2.3.2. Interaction state simulator

To emulate authentic learning behaviours rather than optimal problem-solving strategies, the simulator employs a dual-process cognitive architecture inspired by ACT-R (Anderson et al. 2004) and Bounded Rationality (Simon 1972). The internal state of a simulated student at turn k is represented as $S_k = \langle R_k, C_k \rangle$ where R_k and C_k denote two fundamental knowledge structures.

Learned Rules (R_k) encode enduring knowledge, including mappings of direction words (e.g. vakarus → West), verified cell positions, and logical exclusion constraints. Grid Candidates (C_k) represents provisional belief states concerning potential assignments for blank cells. This module operates under capacity limitations dictated by the attention span parameter, facilitating the simulation of working memory constraints through the implicit 'forgetting' of uncommitted candidates.

The simulated agent engages with the tutoring environment through a tri-phased reasoning loop consisting of Think, Act, and Refine stages (Wei et al. 2025). During the Think phase, the agent employs top-down reasoning to filter grid candidates C_k using learned rules R_k , while also synthesising patterns through bottom-up processing; it assesses uncertainty by calculating the entropy of the candidate set. In the subsequent Act phase, the agent selects a multimodal response based on confidence thresholds: utilising handwriting to complete a cell when a single candidate remains, or employing voice commands to seek clarification when rule ambiguity surpasses a predetermined threshold. During the Refine phase, feedback from the ECRobot prompts the agent to update its internal state by modifying R_k (e.g. correcting a translation) and pruning infeasible assignments from C_k .

$$S_{k+1} = \Phi(S_k, F_k, LLM(P, S_k, F_k)) \quad (4)$$

The system state at turn $k + 1$ is modelled as a function of the preceding state, the tutor's feedback, and the output of the language model conditioned on the specified persona and current state. This framework facilitates the ECRobot's capacity to assess the learner's developing cognitive status over successive interaction cycles.

2.4. Experimental design

2.4.1. Calibration phase

Prior to conducting the comprehensive tutoring simulation, a calibration phase was undertaken to determine the optimal LLM configuration that yields the most consistent alignment between prompted persona capabilities and actual task performance. This preliminary step is essential due to the inherently non-deterministic relationship between persona prompts and model behaviour; different models and temperature parameters can produce varying degrees of persona adherence.

Five language models of increasing capacity were evaluated: Qwen-2.5-0.5b, Qwen-2.5-1.5b, Qwen-2.5-3b, Qwen-2.5-7b, and GPT-4o-mini. Each model was tested across five temperature settings (Temperature $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) and multiple persona configurations on a subset of the target puzzle. For each combination of model and temperature, the Pearson correlation coefficient was calculated between prompted ability scores and observed task performance across three cognitive domains: memory, spatial

reasoning, and semantic understanding. The temperature setting that produced the highest and most stable correlation across all models was selected for the subsequent full simulation. The selection criterion prioritised maximising the stability of persona – performance alignment across models over maximising absolute performance.

2.4.2. Full simulation design

Based on the calibration results, the comprehensive simulation employed a 3 (Model: Qwen-2.5-7b, GPT-4o-mini, GPT-4o) \times 2 (Tutoring Style: empathetic versus neutral) between-subjects factorial design. Thirty simulated students were generated for each cell, resulting in a total of $N = 180$ simulated tutoring sessions. Post hoc sensitivity analysis demonstrated that this design achieves 80% power ($\alpha = .05$) to detect effects of Cohen's $f \geq 0.23$ (partial $\eta^2 \geq .052$) for the three-level model factor and $f \geq 0.21$ (partial $\eta^2 \geq .042$) for the two-level tutoring style factor. These thresholds correspond to small-to-medium effects according to conventional standards (Cohen 2013), thus affirming that the design is adequately powered to identify practically meaningful differences across both factors.

Each simulated session entailed a complete interaction between a persona-constrained student agent and the ECRobot system, proceeding in a turn-by-turn fashion until either the puzzle was completed or the maximum number of turns (50) was reached. The dependent variables measured included task accuracy (the percentage of correctly placed names out of the 11 user-editable grid cells, excluding the 5 pre-filled cells), task completion status (whether the simulated student autonomously terminated the interaction within the 50-turn limit, regardless of whether all cells were filled), task finish status (whether all 11 user-editable cells were correctly filled), and the total number of interaction turns. Similar to the experimental condition in human participant studies, a 50-turn constraint was imposed; sessions exceeding this limit were labelled as not completed. The simulated student could autonomously decide when to terminate the process, whether all 11 grids were filled or not. Because completion reflects the agent's self-regulated termination decision rather than puzzle-solving success, the completion rate is not necessarily consistent with accuracy.

2.5. Evaluation

2.5.1. System robustness tests

To validate the efficacy of the proposed Input Manager in mitigating concurrency failures observed during preliminary system testing, we translated common interaction patterns into three automated stress tests utilising asynchronous Python agents. The Cancellation Trap test simulates the 'Edit-Ask' behaviour, wherein a user initiates answer composition and immediately poses a question. A handwriting event is injected 200 milliseconds after a voice input enters processing, verifying that the incoming handwriting event does not terminate the ongoing voice task. The Snapshot Atomicity Stress test replicates rapid 'erase-rewrite' cycles occurring at intervals of less than 300 milliseconds while manipulating the interaction ID during the save window. This test assesses whether the closure-based state capture mechanism accurately isolates grid snapshots from concurrent ID mutations. The High-Frequency Burst Load test submits 50 interleaved inputs within 500 milliseconds to confirm the integrity of the input queue and ensure zero dropped events under extreme concurrency conditions, significantly exceeding typical human interaction speeds.

2.5.2. Simulation performance evaluation

The simulation outcomes were systematically assessed across three distinct dimensions. Initially, task completion rates and accuracy metrics were calculated for all experimental conditions, followed by a two-way ANOVA to examine the main effects of the model type and tutoring style, as well as their interaction effects on accuracy and total turns. Subsequently, cumulative completion curves were generated over the course of the turns to analyse the temporal dynamics of learning progression within each condition. Finally, error patterns and recovery mechanisms were scrutinised to evaluate the ecological validity of the simulated behaviours. Three specific failure modes were identified based on established patterns in human learning research: an information-seeking loop, characterised by sessions in which voice actions surpassed handwriting actions by a factor of three or more; spatial reasoning failure, indicated by errors related to hints involving directional relationships; and premature convergence, defined as an incorrect cell commitment occurring before the tenth turn.

3. Results

3.1. System robustness test

The automated stress testing unequivocally established the robustness of the event-driven, asynchronous architecture. As delineated in Table 2, the system attained a 100% success rate across all three diagnostic scenarios, thereby effectively mitigating the concurrency failures initially observed during preliminary assessments.

In the Cancellation Trap test, the Input Manager accurately serialised overlapping inputs by implementing a task-level guard designed to prevent incoming events from prematurely terminating ongoing processing tasks. Both voice and handwriting inputs were retained without any data loss (0/2 inputs lost). In the Snapshot Atomicity Stress test, despite intentional mutations to interaction IDs during asynchronous write operations, the closure-based state capture mechanism correctly associated grid states with their respective interaction IDs, ensuring that each pedagogical response was accurately grounded in the corresponding learner state (IDs 100 and 20). In the High-Frequency Burst Load test, the asynchronous event queue sustained system stability under a load of 50 interleaved inputs within 500 ms, a throughput that markedly exceeds typical human interaction speeds, with no events dropped and one batch correctly merged. Collectively, these findings substantiate that the architecture reliably coordinates concurrent multimodal inputs even under adversarial conditions.

3.2. Simulation results

3.2.1. Calibration results

To evaluate the ecological validity of the language model-based student simulation, a calibration phase was conducted utilising five models: Qwen-2.5-0.5b, Qwen-2.5-1.5b, Qwen-2.5-3b, Qwen-2.5-7b, and GPT-4o-mini, each tested at varying temperature settings (Temperature $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) and diverse cognitive ability prompts. Each model was assessed on a subset of the target puzzle under different persona configurations to evaluate the alignment between prompted cognitive abilities and actual task performance.

Figure 3 depicts the correspondence between prompted persona abilities and observed performance across three cognitive domains: memory, spatial reasoning, and semantic understanding. Figure 4 presents the Pearson correlation coefficients for each model – temperature pairing. Findings indicate that a temperature setting of 0.5 produced the most consistent positive correlations between prompted abilities and observed scores across all three cognitive dimensions and model variants. Notably, at a temperature of 0.5, correlation patterns remained uniformly positive for both smaller models (Qwen series) and the larger model (GPT-4o-mini). Conversely, lower temperature settings (0.1, 0.3) appeared to induce ceiling effects, thus constraining performance variability, while higher settings (0.7, 0.9) introduced excessive stochasticity, diminishing persona adherence. Consequently, temperature = 0.5 was selected for subsequent simulations, as it optimally balanced behavioural variance with deterministic persona fidelity.

3.2.2. Task completion and accuracy

Table 3 presents completion metrics and accuracy rates across six experimental conditions comprising three models and two tutoring styles, with a sample size of 30 per condition. In a prior study serving as a baseline (Yueh et al. 2025), human participants ($N = 40$) required an average of 43 minutes ($SD = 12.5$) to complete the task. Since the simulation quantifies interaction turns rather than elapsed time, direct quantitative comparisons between human and simulated performance are not feasible. Nonetheless, the human baseline contextualises the task's cognitive demands and substantiates its appropriateness as a stress-testing environment for the tutoring system.

Table 2. Results of automated robustness stress tests.

Test Scenario	Mitigation Verified	Strategy	Outcome
Cancellation Trap	Task-level guard	Input permission control	Pass (0/2 inputs lost)
Snapshot Atomicity	Closure-based state capture	Snapshot isolation	Pass (IDs correctly bound)
Burst Load (50 inputs)	Asynchronous Event	Priority queueing	Pass (0 drops, 1 merged batch)

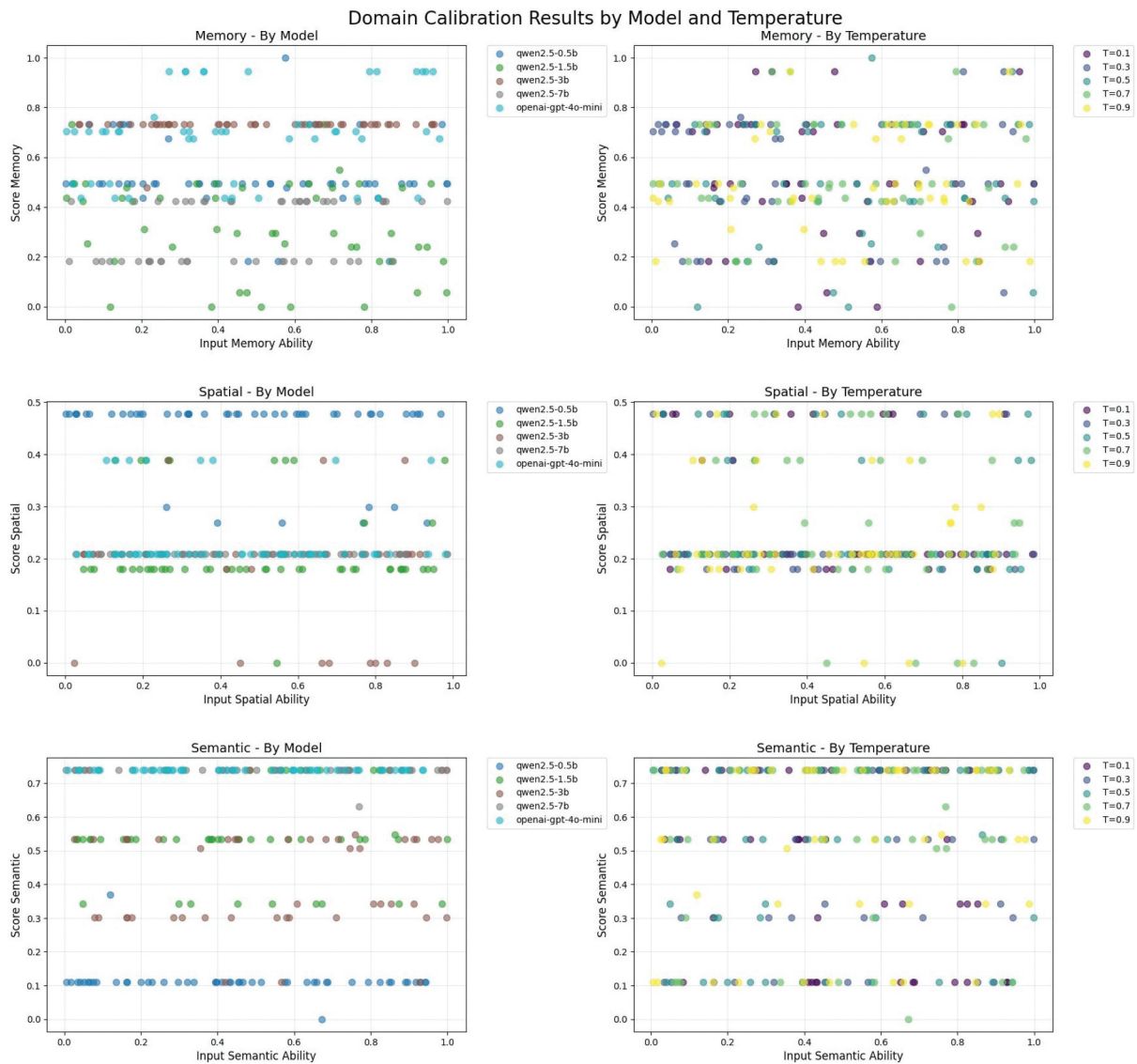


Figure 3. Alignment between prompted persona abilities and observed task performance across three cognitive dimensions.

The results demonstrate significant variability across models. GPT-4o achieved the highest accuracy (76.7% under empathetic and 73.6% under neutral tutoring) and utilised the fewest mean turns (44.7 and 45.5, respectively). GPT-4o-mini exhibited the highest completion rate (29/30 under both conditions), indicating that nearly all simulated students autonomously terminated their sessions within the 50-turn limit. However, only 5 out of 30 sessions under each condition achieved a fully correct solution (finished), yielding moderate accuracy levels of 65.2% and 68.2%. Conversely, Qwen-2.5-7b displayed markedly lower performance, with only 4 and 7 sessions reaching completion out of 30, and accuracy estimates of 41.8% and 35.5%, respectively, which should be interpreted cautiously due to limited statistical stability stemming from the small number of completed sessions.

Table 3 distinguishes between completion (the simulated student autonomously terminated the interaction within the 50-turn limit) and finish (all 11 user-editable cells were correctly filled). As shown, the completion rate was substantially higher than the finish rate across all conditions, indicating that most simulated students terminated their sessions before achieving full puzzle accuracy. An apparent paradox emerges from the completion data: GPT-4o-mini achieved higher session completion (29/30) than GPT-4o (19 to 22 out of 30), despite GPT-4o being the more capable model. However, when examining the finish rate (all 11 cells correctly filled), the pattern reverses: GPT-4o achieved 10/30 under both conditions, compared to

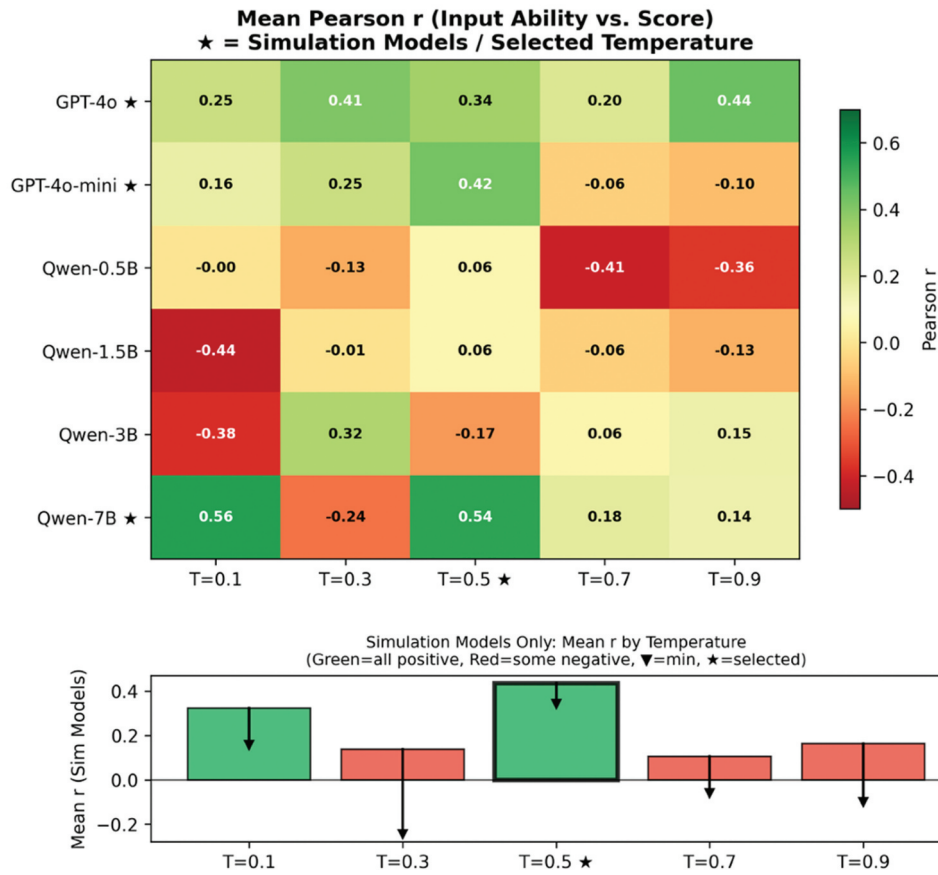


Figure 4. Correlation between prompted abilities and observed performance for each model–temperature combination.

Table 3. Task completion rates, accuracy, and average turns by model and tutoring condition (N = 30 per cell).

Model	Condition	Completed	Finished	Accuracy	Avg. Turns
Qwen-2.5-7b	Empathetic	4/30	1/30	41.8% ± 3.6%	49.9 ± 0.1
Qwen-2.5-7b	Neutral	7/30	0/30	35.5% ± 3.2%	50.0 ± 0.0
GPT-4o-mini	Empathetic	29/30	5/30	65.2% ± 4.6%	49.1 ± 0.4
GPT-4o-mini	Neutral	29/30	5/30	68.2% ± 4.5%	48.1 ± 0.9
GPT-4o	Empathetic	19/30	10/30	76.7% ± 4.3%	44.7 ± 1.5
GPT-4o	Neutral	22/30	10/30	73.6% ± 4.9%	45.5 ± 1.4

Note. Completed denotes the number of sessions in which the simulated student autonomously terminated the interaction within the 50-turn limit. Finished denotes the number of sessions in which all 11 user-editable cells were correctly filled. Accuracy denotes the percentage of correctly placed names out of the 11 user-editable grid cells, excluding the 5 pre-filled cells). Values after ± represent standard errors. The maximum turn limit per session was 50.

GPT-4o-mini’s 5/30. This suggests that GPT-4o-mini adopts a more persistent strategy, continuing interaction until self-termination even when partially incorrect, whereas GPT-4o exhibits more conservative commitment behaviour, withholding cell placements when confidence is insufficient. The higher completion but lower finish rate of GPT-4o-mini, combined with GPT-4o’s lower completion but higher finish and accuracy rates, indicates that model capability influences not only task performance but also the simulated agent’s metacognitive regulation of when to terminate problem-solving.

This pattern likely reflects divergent problem-solving strategies: GPT-4o-mini appears to employ a more incremental, step-by-step approach, completing the puzzle even with partial inaccuracies, whereas GPT-4o demonstrates a more conservative approach by withholding cell placements when confidence is insufficient, leading to incomplete sessions but higher accuracy on completed attempts. This interpretation aligns with the observed lower average turns for GPT-4o (44.7–45.5) relative to GPT-4o-mini (48.1–49.1), suggesting that GPT-4o terminates or stalls earlier when facing uncertainty rather than persisting with low-confidence guesses.

A two-way ANOVA examined the effects of model and tutoring style on accuracy and total turns. For accuracy, the main effect of model was statistically significant, $F(2, 174) = 40.76, p < .0001$, with a partial η^2 of .319, indicating substantial influence of model capacity on outcomes. The main effect of tutoring style was not statistically significant, $F(1, 174) = 0.38, p = .540, \eta^2 = .002$, suggesting negligible differences between empathetic and neutral conditions. The interaction effect between model and style was also non-significant, $F(2, 174) = 0.63, p = .532$. Regarding total turns, a similar pattern emerged: the main effect of model was significant, $F(2, 174) = 13.98, p < .0001, \eta^2 = .138$, while the effect of tutoring style, $F(1, 174) = 0.003, p = .954$, and the interaction, $F(2, 174) = 0.47, p = .625$, were non-significant. These findings substantiate that model selection primarily determines simulated learning outcomes, whereas tutoring style, as operationalised in the current framework, does not exert a statistically detectable influence on accuracy or efficiency. Given the sensitivity analysis in Section 2.4.2, which established an 80% power to detect effects as small as $f = 0.21$ for the style factor, the null results likely reflect a negligible effect of empathetic versus neutral scaffolding within the simulated environment.

3.2.3. Temporal dynamics of learning progress

Figure 5 depicts the cumulative completion curves across turns, showing learners' progression through the task under each experimental condition. Empathetic tutoring elicited a marginally earlier convergence trajectory, primarily during the mid-phase of problem-solving (Turns 15–35); however, this pattern was not statistically significant in the overall ANOVA. This temporal observation implies that if empathetic scaffolding confers any advantage, it may be restricted to a specific phase of the learning process, namely, when learners have acquired initial domain knowledge but have not yet developed autonomous debugging strategies, rather than offering a generalised benefit across the entire interaction.

3.2.4. Error patterns and failure modes

To evaluate the ecological validity of the simulated behaviours, error patterns across all sessions were systematically analysed based on the three failure modes outlined in Section 2.5.2. Table 4 delineates the distribution of these failure modes across different models and conditions.

The information-seeking loop was infrequent across all models and conditions: no instances were observed in Qwen-2.5-7b or GPT-4o, and only a single instance was observed in GPT-4o-mini. This suggests that repetitive information-seeking without tangible progress did not constitute a prevailing failure mode

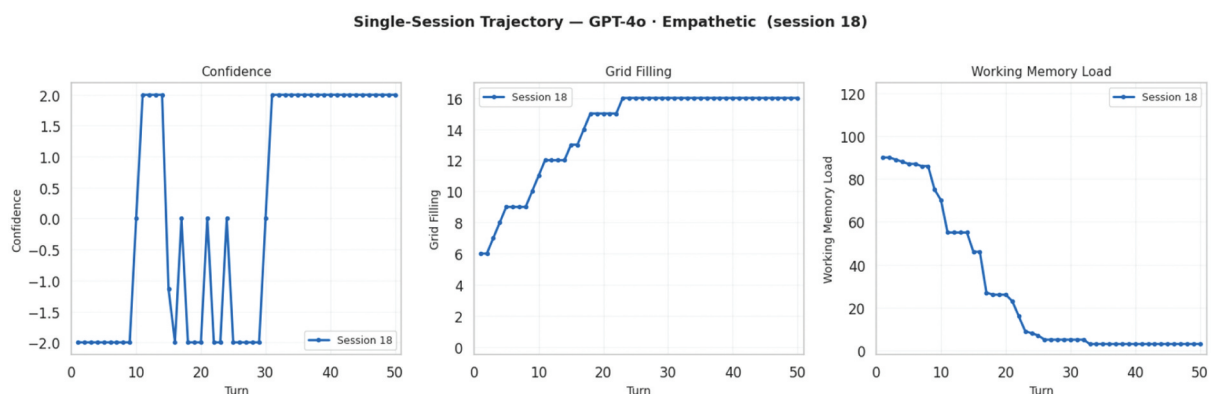


Figure 5. Cumulative completion curves across interaction turns by model and tutoring condition.

Table 4. Distribution of failure modes across models and tutoring conditions.

Failure Mode	Operational Definition	Qwen-2.5-7b (E / N)	GPT-4o-mini (E / N)	GPT-4o (E / N)
Information-seeking Loop	Voice actions exceed handwriting actions by a factor of ≥ 3	0/0	0/1	0/0
Spatial Reasoning Failure	Errors on spatially-oriented directional hints	26/22	27/27	29/30
Premature Convergence	Incorrect cell commitment before Turn 10	30/30	30/30	30/30

Note. Values are reported as instances/cases, where 'instances' denotes the number of occurrences of the failure mode and 'cases' denotes the number of sessions evaluated. Each model column presents empathetic (E) / neutral (N) condition counts.

within the current simulation framework, indicating that the simulated students generally maintained forward momentum during problem-solving.

Spatial reasoning failures were prevalent and consistently documented across all models. Specifically, Qwen-2.5-7b exhibited 26 instances across 22 evaluated sessions, GPT-4o-mini demonstrated 27 across 27 sessions, and GPT-4o displayed 29 across 30 sessions. The near-universal occurrence of spatial reasoning errors indicates that the directional hints embedded in the puzzle consistently challenged all models, necessitating multiple attempts to correctly interpret spatial relationships. This pattern aligns with findings from human problem-solving literature, which reports similar difficulties with spatial tasks.

Premature convergence was observed in all evaluated sessions across all three models (30/30). This recurring pattern indicates that language – model – based agents habitually commit to cell placements before accumulating sufficient evidentiary support, particularly during initial interaction phases. Such behaviour mirrors the anchoring bias documented in human cognition (Tversky and Kahneman 1974), where initial hypotheses are maintained despite contradictory evidence, suggesting that these agents may exhibit cognitive rigidity analogous to human tendencies under conditions of uncertainty.

4. Discussion

This study investigates two interrelated challenges within the development lifecycle of agentic social robot tutors: the real-time coordination of multimodal inputs and the automated assessment of pedagogical strategies without reliance on extensive human trials. The proposed ECRobot system, together with its LLM-based simulation pipeline, is conceived as two complementary components of an integrated hyperautomation framework. In this framework, both the system operations and the evaluation processes are managed by AI agents functioning with minimal human oversight. The following discussion elaborates on the findings concerning each research objective and explores their implications from both a theoretical and practical perspective.

4.1. Event-driven architecture for multimodal agentic systems

The primary research objective centred on developing a resilient architecture capable of managing concurrent multimodal inputs within an agentic social robot ITS. The robustness assessment results (Section 3.1) substantiate that the proposed event-driven asynchronous architecture, which integrates advisory lock-based atomicity protection and temporal merging, achieved a 0% data loss in all stress testing scenarios. This empirical evidence demonstrates the architecture's capacity to reliably coordinate heterogeneous real-time data sources under adverse conditions, thereby establishing a stable foundation for the deployment of agentic artificial intelligence in unmoderated educational environments.

The contribution of this architectural framework resides not in the novelty of individual techniques (event-driven processing, advisory locks, and temporal windowing, which are well-established in enterprise systems engineering (Fertier et al. 2021; Stój et al. 2022) but in their integration within a multimodal human-robot interaction context. Current multimodal AI systems predominantly address temporal synchronisation through model-internal mechanisms, such as parallel token streams (Défossez et al. 2024). While effective for end-to-end neural architectures, these methods do not resolve the challenges associated with state management when multiple independent AI pipelines, including OCR, ASR, and LLM reasoning modules, must update a shared representation of learner cognitive status. The Input Manager design presented herein offers a complementary, infrastructural solution: rather than embedding synchronisation within individual models, it externalises concurrency control via a coordination layer that serialises state updates while maintaining asynchronous processing of individual modalities.

This architectural pattern has broader applicability beyond educational settings, notably in enterprise information systems. The temporal merging window addresses similar challenges encountered in sensor fusion within manufacturing IoT environments, where RFID readings, machine vision outputs, and operator inputs must be consolidated into a coherent system state within strict latency constraints. The advisory lock mechanism parallels optimistic concurrency control strategies employed in ERP systems to prevent state corruption during simultaneous modifications by multiple agents or users. The three-tier robustness evaluation strategy, comprising cancellation traps, snapshot atomicity, and burst load testing, provides a reusable

template for assessing systems that coordinate heterogeneous real-time data sources, whether in healthcare monitoring, supply chain management, or collaborative decision-support platforms. Across these contexts, the fundamental engineering challenge is ensuring that asynchronous, concurrent inputs are fused into a consistent system state without data loss or logical incoherence.

4.2. LLM-based simulation as a validation methodology

The second research objective involved the development of a LLM-based student simulation framework aimed at facilitating automated pedagogical assessment. The findings offer preliminary evidence that prompt-based persona engineering can yield cognitively diverse learner agents capable of engaging with a tutoring system through extended, multi-turn problem-solving sessions.

The calibration phase (Section 3.2.1) demonstrated that a temperature setting of 0.5 provides the most stable alignment between induced cognitive abilities and observed task performances. This result holds methodological importance for the expanding body of research employing LLM agents to emulate human behaviour (Park et al. 2024; V.D.E.; Vogelsmeier et al. 2025). Lower temperature configurations tend to compress performance variability due to ceiling effects, whereas higher settings increase stochasticity, thereby diminishing adherence to assigned personas. This trade-off aligns with the conceptualisation of temperature as a parameter regulating diversity and fidelity in language model outputs. Practically, this implies that researchers and developers seeking to simulate heterogeneous learner populations should avoid both extremes: excessively low temperatures produce homogeneous agents, while overly high temperatures lead to behaviours decoupled from their designated personas.

The ecological validity of the simulation is supported by two convergent findings. First, the clear differentiation observed between high- and low-ability personas in memory and spatial reasoning dimensions suggests that these cognitive constructs are sufficiently encapsulated by the dual-process architecture inspired by ACT-R and bounded rationality, thereby enabling meaningful simulation of inter-individual differences. Second, the failure mode analysis indicated that all three models consistently exhibited premature convergence, often committing to particular solutions prior to the accumulation of sufficient evidence. This pattern mirrors the well-documented anchoring bias in human problem-solving (Tversky and Kahneman 1974). Additionally, pervasive spatial reasoning errors, attributable to the directional constraints embedded within the puzzle, posed genuine difficulties for the models, similar to the challenges reported by human solvers. These findings imply that LLM-based agents do not merely solve tasks optimally but can demonstrate cognitively plausible difficulties, thereby supporting their application as proxies for stress-testing pedagogical strategies.

Nevertheless, several limitations inherent to the simulation approach warrant acknowledgement. In comparison to the fine-tuned agents developed by Qi et al. (2026), whose LoRA-adapted models maintained stable persona adherence across extended dialogues, the prompt-based methodology employed here is more susceptible to behavioural drift as context windows extend beyond 30 turns. This limitation is intrinsic to prompt engineering, wherein persona instructions compete with the expanding context for influence over model behaviour. Moreover, the near absence of information-seeking failures (only one instance across 180 sessions) may reflect a limitation rather than a model strength; real human learners frequently engage in extended clarification-seeking behaviours, especially when experiencing confusion, whereas LLM agents tend to sustain forward progress even amid uncertainty (Seshadri et al. 2026). This propensity towards action rather than inquiry may lead to an underestimation of the frequency with which actual students would request additional support, bearing implications for the evaluation of tutoring strategies' responsiveness.

4.3. The role of tutoring style in simulated learning

This study found no statistically significant differences in accuracy or efficiency between empathetic and neutral tutoring conditions. The main effect of tutoring style was not significant for accuracy ($F(1, 174) = 0.38$, $p = .540$, partial $\eta^2 = .002$) or total turns ($F(1, 174) = 0.003$, $p = .954$). A sensitivity analysis confirmed the study design had sufficient power to detect effects as small as $f = 0.21$. This null result warrants careful interpretation. It is important to note that the 'empathetic' intervention in this study was operationalised as text-based affective framing, primarily through validating learner effort during correct responses, rather than through

multimodal empathetic behaviours such as affect perception via tone analysis or facial expression recognition. The null finding therefore reflects both the limited scope of the affective manipulation and the inherent inability of LLM-based simulated students to experience or benefit from emotional support.

Four explanations are proposed. First, in human tutoring, empathetic scaffolding mainly influences affective and motivational factors, such as reducing anxiety and fostering engagement, which simulated students in this study cannot experience because they lack genuine emotions. Therefore, the affective component of empathy has no functional target within the simulation, possibly explaining the absence of measurable benefits. This suggests the null finding reflects limitations of the simulation rather than undermining empathetic tutoring effectiveness with real learners. Second, both conditions incorporated cognitive scaffolding, like guided reflection and strategic hints. The primary difference was in the affective framing of feedback, not its informational content. If cognitive scaffolding primarily drives performance in complex reasoning tasks, then similar cognitive support across conditions could result in comparable outcomes. Third, the results align with Cognitive Load Theory (Sweller 2011), indicating that, for high-intrinsic-load tasks such as those involving complex constraints, the additional load from affective language may be negligible compared to the germane load required for reasoning and rule application. Fourth, the experimental design limited observable differences. The feedback structure had three parts, with affective manipulation mainly in the first segment, especially when learners answered correctly. For incorrect responses, affective content was minimal, and most substantive feedback, cognitive hints, were generated without explicit affective modulation. This likely reduced the overall impact of the affective manipulation.

From a practical perspective, these findings imply that in AI-driven training systems using simulated profiles for stress testing, the tone of feedback may matter less than the quality of informational and structural scaffolding. Organisations should focus on accurate, adaptive cognitive feedback over stylistic emotional cues. However, whether these results extend to real human users remains unconfirmed and requires further validation with studies involving human participants.

These findings collectively underscore that the current operationalisation of empathy represents a constrained form of text-based affective framing. A more comprehensive implementation of empathetic tutoring would integrate multimodal affect perception, including tone of voice analysis, facial expression recognition, and physiological signal monitoring, to enable the robot to dynamically detect and respond to learners' emotional states. Such an approach would more closely approximate the relational depth that characterises effective human tutoring and would require validation with real learners rather than LLM-based agents.

5. Conclusion

This investigation examines two main challenges in the lifecycle of agentic social robot tutors: managing multiple real-time multimodal inputs and automating pedagogical assessment at scale. The proposed ECRobot system, with its event-driven asynchronous architecture and safeguards, demonstrated zero data loss during robustness tests, confirming its reliable coordination of diverse data streams under adverse conditions. A student simulation pipeline using LLMs, with 180 sessions in a 3×2 factorial design, revealed that model capacity significantly affects learning outcomes (partial $\eta^2 = .319$ for accuracy). In contrast, tutoring style (empathetic vs. neutral) had no significant effect, a null result supported by sensitivity analysis. The simulation identified two failure modes: spatial reasoning brittleness and premature convergence, both aligned with human cognitive biases, supporting ecological validity.

The study offers four major contributions to multimodal AI and hyperautomation. First, its event-driven architecture, incorporating temporal merging and advisory lock-based state serialisation, provides a general framework for handling race conditions in real-time multimodal systems. Unlike internal synchronisation methods, this approach externalises concurrency control, making it applicable beyond social robot tutoring to enterprise systems that integrate heterogeneous inputs like IoT sensors, ERP workflows, and decision-support tools. Second, the dual-process cognitive architecture used for student simulation, based on ACT-R and Bounded Rationality, shows that prompt-based persona engineering can produce cognitively plausible behavioural variability during extended interactions. The emergence of anchoring-like premature convergence mirrors human-like performance and biases in LLM agents, supporting their use in pedagogical evaluation. Third, the calibration methodology that matches

model – temperature configurations against persona – performance stability offers a reproducible protocol for configuring LLMs in behavioural simulations, addressing a methodological gap in human simulation research. Fourth, the null effect of tutoring style suggests the affective channel of empathetic scaffolding lacks a functional target in agents without genuine emotional states, indicating that motivational and relational mechanisms underlying empathetic tutoring may not be effectively replicated in current simulations. This highlights the limitations of LLM-based models and underscores the need for human validation in affective pedagogical assessments. In particular, the near-absence of information-seeking behaviour among simulated agents (only 1 instance across 180 sessions) stands in contrast to the frequent clarification-seeking that characterises real human learners, representing a fundamental gap in the current simulation's behavioural repertoire. Establishing a comprehensive human baseline that compares failure mode distributions, question-asking frequencies, and completion dynamics between real learners and LLM agents is therefore an essential priority for future research. Furthermore, the operationalisation of empathetic tutoring in this study was limited to text-based affective framing and did not incorporate multimodal affect perception such as tone analysis or facial expression recognition; future implementations should integrate these capabilities to enable more comprehensive empathetic interactions.

Beyond immediate findings, the study proposes a framework for enterprise systems. The paradigm of intelligent hyperautomation (Koeppen, Wang, and Ip 2025) advocates for autonomous AI agents capable of executing complex tasks within integrated systems. The ECRobot system and its simulation pipeline exemplify this, showing how an agentic intelligent tutoring system, a multimodal perception layer, and an LLM evaluation agent can be integrated into a hyperautomated workflow for real-time interactions and automated assessments with minimal human input.

For enterprise applications, the LLM-based student simulation can create prototypes of novice employees to test onboarding procedures. Similar to developing learner agents with diverse profiles, organisations can generate employee agents with varying expertise and learning preferences to evaluate training modules before deployment. Failure modes like premature convergence and spatial reasoning issues mirror real challenges such as incorrect workflow choices or interface difficulties. Additionally, the event-driven architecture with lock-based safeguards offers a reusable model for multi-agent enterprise systems, effectively addressing concurrency issues in contexts like industrial IoT, where sensors and decision-support agents share data without corruption (Romano et al. 2025). Finally, the calibration approach, systematically varying model parameters for realism, can be broadly applied to validate enterprise simulations, including supply chain management and customer behaviour modelling (Zhang and Zheng 2019).

Author contributions

CRedit: **Hsiu-Ping Yueh:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing; **Hung-Chun Chang:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft; **Guan-Yun Wang:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing; **Pei-Luo Hung:** Investigation, Methodology, Resources, Software; **Weijane Lin:** Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Taiwan National Science and Technology Council under Grant [NSTC112-2410-H-002-121-MY3, NSTC114-2622-H-002-002, NSTC114-2410-H-030-089-MY2]; and National Taiwan University under Grant [NTU114L8947].

Data availability statement

The data that support the findings of this study are available from the corresponding author, GYW, upon reasonable request.

References

- Ali, A., F. D. Hassan, and M. Charafeddine. 2025. "Agentic AI: A Comprehensive Survey of Architectures, Applications, and Future Directions." *Artificial Intelligence Review* 58 (8): 256. <https://doi.org/10.1007/s10462-025-11257-z>.
- Alves-Oliveira, P., P. Sequeira, F. S. Melo, G. Castellano, and A. Paiva. 2019. "Empathic Robot for Group Learning: A Field Study." *ACM Transactions on Human-Robot Interaction* 8 (1): Article 3. <https://doi.org/10.1145/3300188>.
- Anderson, J. R., D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. 2004. "An Integrated Theory of the Mind." *Psychological Review* 111 (4): 1036–1060. <https://doi.org/10.1037/0033-295X.111.4.1036>.
- Breazeal, C. 2003. "Toward Sociable Robots." *Robotics and Autonomous Systems* 42 (3–4): 167–175. [https://doi.org/10.1016/S0921-8890\(02\)00373-1](https://doi.org/10.1016/S0921-8890(02)00373-1).
- Breazeal, C. 2004. "Social Interactions in HRI: The Robot View." *IEEE Transactions on Systems, Man, and Cybernetics, Part C* 34 (2): 181–186. <https://doi.org/10.1109/TSMCC.2004.826268>.
- Chango, W., R. Cerezo, M. Sanchez-Santillan, R. Azevedo, and C. Romero. 2021. "Improving Prediction of Students' Performance in Intelligent Tutoring Systems Using Attribute Selection and Ensembles of Different Multimodal Data Sources." *Journal of Computing in Higher Education* 33 (3): 614–634. <https://doi.org/10.1007/s12528-021-09298-8>.
- Chen, Y. 2025. "Multimodal Generation and Personalization Algorithms for Educational Content Using Generative ai." Seventh International Conference on Image, Video Processing, and Artificial Intelligence (IVPAI 2025). Vol. 13731 Bangkok, Thailand: 1373118. SPIE. <https://doi.org/10.1117/12.3076193>.
- Chou, C.-K., C.-J. Hsu, H.-L. Chung, L.-H. Tseng, H.-C. Cheng, Y.-K. Fu, K. Po Huang, and H.-Y. Lee. 2025. "A Self-Refining Framework for Enhancing ASR Using TTS-Synthesized Data." 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Honolulu, HI, USA. IEEE.
- Chu, Z., S. Wang, J. Xie, T. Zhu, Y. Yan, J. Ye, A. Zhong, et al. 2025. "LLM Agents for Education: Advances and Applications." Paper presented at the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China. November.
- Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge.
- Défossez, A., L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. arXiv preprint. <https://doi.org/10.48550/arXiv.2410.00037>.
- Fertier, A., G. Martin, A.-M. Barthe-Delanoë, J. Lesbegueries, A. Montarnal, S. Truptil, F. Bénaben, and N. Salatgé. 2021. "Managing Events to Improve Situation Awareness and Resilience in a Supply Chain." *Computers in Industry* 132:103488. <https://doi.org/10.1016/j.compind.2021.103488>.
- Jonassen, D. H., C. Carr, and H.-P. Yueh. 1998. "Computers as Mindtools for Engaging Learners in Critical Thinking." *TechTrends* 43 (2): 24–32. <https://doi.org/10.1007/BF02818172>.
- Kanda, T., and H. Ishiguro. 2017. *Human-Robot Interaction in Social Robotics*. Boca Raton, FL: CRC Press. 9781466506978.
- Koepfen, K., H. Wang, and W. H. Ip. 2025. "Omnipresent AI and Intelligent Hyperautomation: Methodologies, Interactivity and Applications." *Enterprise Information Systems* 19 (11): 2551727.
- Kostopoulos, G., V. Gkamas, M. Rigou, and S. Kotsiantis. 2025. "Agentic AI in Education: State of the Art and Future Directions." *IEEE Access* 13:177467–177491. <https://doi.org/10.1109/ACCESS.2025.3620473>.
- Liao, C.-H., and J.-Y. Wu. 2022. "Deploying Multimodal Learning Analytics Models to Explore the Impact of Digital Distraction and Peer Learning on Student Performance." *Computers and Education* 190:104599. <https://doi.org/10.1016/j.compedu.2022.104599>.
- Luo, X. M., M. S. Qin, Z. Fang, and Z. Qu. 2021. "Artificial Intelligence Coaches for Sales Agents: Caveats and Solutions." *Journal of Marketing* 85 (2): 14–32. <https://doi.org/10.1177/0022242920956676>.
- Márquez, A. C., and J. F. Gómez Fernández. 2026. "Agentic AI for Autonomous Preventive Maintenance Policy Governance: A Multi-Agent Framework for Dynamic Industrial Environments." *Expert Systems With Applications* 314:131767. <https://doi.org/10.1016/j.eswa.2026.131767>.
- Mitri, D., J. S. Daniele, M. Specht, and H. Drachslar. 2018. "From Signals to Knowledge: A Conceptual Model for Multimodal Learning Analytics." *Journal of Computer Assisted Learning* 34 (4): 338–349. <https://doi.org/10.1111/jcal.12288>.
- Park, J. S., C. Q. Zou, A. Shaw, B. Mako Hill, C. Cai, M. Ringel Morris, R. Willer, P. Liang, and M. S. Bernstein. 2024. LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. arXiv Preprint. <https://doi.org/10.48550/arXiv.2411.10109>.
- Qi, C., L. Zheng, A. He, H. Xu, L. Jia, Y. Wei, B. Jiang, and X. Gu. 2026. "Simulating Student Learning Behaviors with LLM-Based Role-Playing Agents." *Expert Systems With Applications* 304:130753. <https://doi.org/10.1016/j.eswa.2025.130753>.
- Romano, A. L., C. Henrique Santos, W. Azzolini Júnior, and A. da Silva Caires. 2025. "ERP-Driven Simulation for Production Planning and Control in the Industry 4.0: A Review." *International Journal of Simulation Modelling* 24 (3): 425–436. <https://doi.org/10.2507/IJSIMM24-3-731>.

- Saerbeck, M., T. Schut, C. Bartneck, and M. D. Janse. 2010. "Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10) April 10 - 15, 2010 (New York, NY, USA: Association for Computing Machinery) Atlanta Georgia USA, 1613–1622 doi:<https://doi.org/10.1145/1753326.1753567>.
- Seshadri, P., S. Cahyawijaya, A. Odumakinde, S. Singh, and S. Goldfarb-Tarrant. 2026. Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations. arXiv Preprint. <https://doi.org/10.48550/arXiv.2601.17087>. 2601.17087.
- Silero, T. 2024. "Silero VAD: Pre-Trained Enterprise-Grade Voice Activity Detector (VAD), Number Detector and Language Classifier." *GitHub Repository*.
- Simon, H. A. 1972. "Theories of Bounded Rationality." *Decision and Organization* 1 (1): 161–176.
- Stój, J., A.-L. Kampen, R. Cupek, I. Smółka, and M. Drewniak. 2022. "Industrial Shared Wireless Communication Systems—Use Case of Autonomous Guided Vehicles with Collaborative Robot." *Sensors* 23 (1): 158. <https://doi.org/10.3390/s23010158>.
- Sweller, J., edited by J. P. Mestre and B. H. Ross 2011. "Chapter Two - Cognitive Load Theory". In *Psychology of Learning and Motivation*, 37–76. San Diego: Academic Press. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>.
- Taiwan Olympiad Linguistics. 2024. "2024 International Linguistics Olympiad – Taiwan First Round Individual Contest: Problems." https://ioltw.github.io/problems/TOL2024_1st_round_en.pdf.
- Tversky, A., and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science*. 185 (4157): 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>.
- Vogelsmeier, V. D. E. L., E. Oliveira, K. Misiejuk, S. López-Pernas, and M. Saqr. 2025. "Delving Into the Psychology of Machines: Exploring the Structure of Self-Regulated Learning via LLM-Generated Survey Responses." *Computers in Human Behavior* 173:108769. <https://doi.org/10.1016/j.chb.2025.108769>.
- Wei, T., N. Sachdeva, B. Coleman, Z. He, Y. Bei, X. Ning, M. Ai, Y. Li, J. He, and E. H. Chi. 2025. "Evo-Memory: Benchmarking LLM Agent Test-Time Learning with Self-Evolving Memory." *arXiv Preprint arXiv: 2511.20857*.
- Yan, L., V. Echeverria, Y. Jin, G. Fernandez-Nieto, L. Zhao, X. Li, R. Alfredo, Z. Swiecki, D. Gašević, and R. Martinez-Maldonado. 2024. "Evidence-Based Multimodal Learning Analytics for Feedback and Reflection in Collaborative Learning." *British Journal of Educational Technology* 55 (5): 1900–1925. <https://doi.org/10.1111/bjet.13498>.
- Yan, L., L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević. 2024. "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review." *British Journal of Educational Technology* 55 (1): 90–112. <https://doi.org/10.1111/bjet.13370>.
- Yueh, H.-P., and F.-K. Chiang. 2020. "AI and Robotics in Reshaping the Dynamics of Learning." *British Journal of Educational Technology* 51 (5): 1804–1807. <https://doi.org/10.1111/bjet.13017>.
- Yueh, H.-P., P.-L. Hung, G.-Y. Wang, H.-C. Chang, and W. Lin. 2025. "Empathic AI-Powered Cognitive Robots as Tutors: Assessing Their Impact on Learning Performance in Accelerated Learning Contexts." Paper presented at the 2025 NTU-UTsukuba Symposium on HRI Research for Elderly Wellbeing, Taipei, Taiwan. September.
- Zhang, N., and X. Zheng. 2019. "Agent-Based Simulation of Consumer Purchase Behaviour Based on Quality, Price and Promotion." *Enterprise Information Systems* 13 (10): 1427–1441. <https://doi.org/10.1080/17517575.2019.1654133>.