Engagement-Aware Response Generation for Educational Robots Using Grounded LLM Prompting

Nayanika Reddy Rajoli, Vennela Akula, Daniela Marghitu
Department of Computer Science and Software Engineering
Auburn University
Auburn, Alabama, USA
e-mail: nzr0054@auburn.edu, vza0024@auburn.edu, marghda@auburn.edu

***Abstract*—Educational robots could make learning more personal, but most current systems use pre-written responses that do not adjust to how a student is engaging or responding. Theory of Mind is the ability to infer what someone is paying attention to, confused by, or trying to do from observable behavior, and it is a necessary foundation for robots that tutor effectively in social settings. This paper presents a work-in-progress multimodal pipeline that implements a simplified, behavior-based approximation of Theory of Mind for an educational robot by observing a user's behavioral signals in real time and generating context-aware verbal responses through a small language model. The pipeline is designed to be configurable across domains and audiences, and is evaluated here on a single mathematics-quiz scenario: it uses MediaPipe for face detection, gaze estimation, head pose tracking, and hand-raise recognition, aggregates these signals into an engagement state through a priority-ordered trigger system, and constructs observation-grounded prompts for the model. We compare two prompt design strategies: the Grounded Mode, which provides both the sensor observation dictionary and an explicit trigger-type label, and the Ungrounded Mode, which provides the observation dictionary alone. The pipeline runs on both a desktop PC and a Raspberry Pi 5, with the model served from a remote graphics processing endpoint. This work falls under the conference Robotics track, specifically cognitive robotics and mobile assistive robots, combining real-time multimodal sensing with language-model reasoning on resource-constrained hardware. Preliminary results show that the Grounded Mode achieves 86.1% response relevance compared to 23.3% for the Ungrounded Mode, demonstrating that small language models require an explicit semantic-translation layer between numeric sensor output and the prompt to produce contextually appropriate responses.**

Keywords-theory of mind; educational robotics; engagement detection; large language models; multimodal sensing.

I. INTRODUCTION

Educational robots are increasingly being used as tutoring companions, but one major problem that still exists is that educational robots are still not able to perceive and respond to the mental state of a user during a session. A user who looks away, raises a hand, or stops responding may be signaling disengagement, confusion, or a desire to ask a question. The ability to perceive and respond to a user's mental state is something that a human tutor can accomplish effortlessly through a cognitive ability known as Theory of Mind (ToM) [1]. These behavioral cues are proxies: they are observable signals rather than direct measurements of an internal state. Robots that lack this capacity default to scripted interactions that cannot adapt to the learner's evolving needs.

Recent advances in Large Language Models (LLMs) offer a path toward more adaptive robot behavior. LLMs can generate contextually relevant natural language responses when provided with appropriate situational context [2]. However, integrating LLMs into real-time robotic systems requires solving two problems simultaneously: perceiving the user's behavioral state through sensors, and constructing prompts that convey that state to the LLM in a form it can act upon.

This paper addresses both problems. The system uses MediaPipe [3] for real-time face detection, gaze estimation, head pose tracking, and hand-raise detection via pose landmarks. These sensor signals are processed by an engagement manager that detects five types of disengagement in priority order: `hand_raised`, `face_absent`, `no_answer`, `head_turned`, and `no_movement`. When a trigger is detected, the current sensor readings are combined into a prompt and sent to a Phi-3-mini LLM (3.8B parameters) [4], which generates a short spoken response to re-engage the user or answer their question. While the pipeline architecture is intended to be configurable across domains and audiences, this work evaluates it only on a mathematics quiz. The prompts and quiz content can be adjusted for other target populations in future studies.

We evaluate the pipeline using an adaptive mathematics quiz for children aged 7 to 14 as the initial test scenario. The Grounded Mode provides the LLM with both the raw observation dictionary and an explicit trigger-type label (e.g., "The user's face is not visible" or "The user is looking away"). The Ungrounded Mode provides only the observation dictionary, requiring the LLM to infer the trigger type from sensor values alone. Preliminary results show that the Grounded Mode produces substantially more relevant responses, confirming that small LLMs cannot reliably infer engagement states from raw sensor data without explicit labeling.

The pipeline currently runs on both a desktop PC and a Raspberry Pi 5. Language-model inference is currently offloaded to a remote endpoint rather than run on the Pi itself, so the system is not yet fully self-contained at the edge. On-device inference is part of future work. The target deployment platform is a Marty V2 educational robot [5], which supports a Raspberry Pi as an onboard expansion module. The Raspberry Pi 5 runs Robot Operating System 2 (ROS 2) Jazzy [6], to control the robot's motors and sensors.

The remainder of this paper is organized as follows. Sec-

tion II reviews related work on ToM in robotics, engagement detection, and LLMs in educational settings. Section III describes the system architecture. Section IV presents the Grounded Mode versus Ungrounded Mode prompt design comparison. Section V reports preliminary results. Section VI concludes the paper and outlines future work.

II. Related Work

Theory of Mind was originally defined by Premack and Woodruff [1] as the ability to impute mental states to oneself and others and to use those attributions to predict behavior. In robotics, ToM has been explored primarily through probabilistic models, neural approaches, and reinforcement learning frameworks. Scassellati [7] proposed one of the earliest architectures for endowing a humanoid robot with ToM, focusing on joint attention and perspective-taking. More recently, Chen et al. [8] demonstrated that an observer robot could model another agent's behavior through visual processing alone, achieving 98.5% success in predicting future actions. Hellou et al. [9] surveyed the field and identified three main classes of computational ToM: probabilistic models using Bayesian inference, neural network approaches, and hybrid methods. They noted that implementations in Human-Robot Interaction (HRI) remain limited, particularly for educational contexts.

Engagement detection in educational settings has been studied using various sensor modalities. Wang et al. [10] developed a MediaPipe-based system for detecting behavioral and emotional engagement in e-learning environments, fusing features from gaze direction, head pose, and facial expressions. Their work demonstrated that MediaPipe provides a low-cost, low-latency framework suitable for real-time deployment on edge devices.

The integration of LLMs into social and educational robots is an active area of research. Studies have shown that LLM-powered robots can maintain open-domain interactions across multiple sessions [11], and that cloud-based LLMs significantly outperform scripted approaches in interaction quality and user satisfaction [12]. In the domain of child-robot interaction, LLMs have been fine-tuned to exhibit creative storytelling behavior with positive effects on engagement [13]. However, existing work has not combined real-time multimodal engagement sensing with observation-grounded LLM response generation in a single educational robot pipeline. This paper addresses that gap by building a pipeline that perceives behavioral cues, infers an engagement state, and generates contextually appropriate verbal responses through an LLM, implementing a simplified form of ToM.

III. System Architecture

The pipeline consists of seven modules that operate in a continuous loop during a tutoring session. Figure 1 provides an overview of the architecture. The architecture is designed to separate sensing, engagement detection, and response generation from the specific quiz content, although this paper evaluates only the mathematics-quiz configuration. For the current evaluation, the system is configured to run an adaptive mathematics quiz targeting children aged 7 to 14.

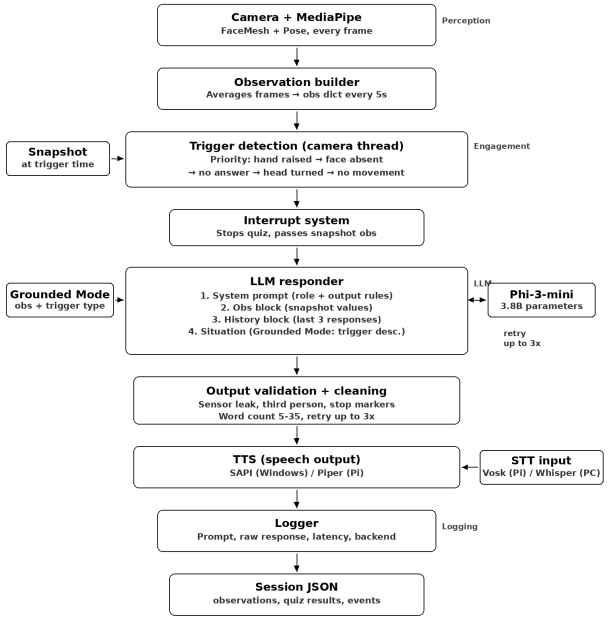

Figure 1. Pipeline architecture showing data flow from camera input through perception, engagement detection, LLM response generation, output validation, speech synthesis, and session logging.

A. Perception Layer

The MediaPipe Processor module captures video frames from a webcam and runs two MediaPipe task models in a background thread: FaceMesh for facial landmark detection and Pose Landmarker Lite for body and hand tracking [3]. Frames are processed continuously, and an observation dictionary is built by averaging sensor readings over a 5-second interval. The dictionary includes the following fields: `face_detected` (boolean), `gaze_on_robot` (0.0 to 1.0, estimated from head yaw as a proxy rather than true eye tracking), `head_yaw_deg` and `head_pitch_deg` (head orientation in degrees), `head_moving` (boolean, based on a 2-degree threshold), `body_detected` (boolean), `hand_raised` and `hand_raise_side` (detected from pose landmarks when a wrist is above shoulder level), `no_movement_sec` (seconds since last detected movement), and `speech_energy`. Several of these fields are deliberately coarse proxies for the underlying behavior, so the engagement states derived from them should be read as behavioral approximations rather than confirmed mental states. The Speech Input module uses Whisper on PC and Vosk on Raspberry Pi for speech-to-text (STT) transcription, with automatic platform detection.

B. Adaptive Quiz Engine

The current test scenario involves a quiz manager with 90 questions in mathematics for different levels of difficulty.

These levels are defined as Level 1 for easy questions, Level 2 for medium questions, and Level 3 for hard questions for ages between 7 and 14 years old. The difficulty level increases or decreases based on whether there are three consecutive correct or incorrect answers. The quiz context, including the current question, topic, difficulty level, and user's previous answers, is provided as input to the LLM prompt so that responses are grounded in the current learning context. The quiz module is replaceable according to different domains or ages.

C. Engagement Manager

The Engagement Manager aggregates sensor readings from the MediaPipe Processor into discrete engagement events. It monitors five trigger conditions in a hardcoded priority order. (1) `hand_raised`, the user's hand is detected above shoulder level. (2) `face_absent`, no face detected for a configurable duration (default 3 seconds). (3) `no_answer`, the user has not responded to a quiz question within a timeout period (default 15 seconds). (4) `head_turned`, the gaze score falls below a threshold (default 0.3 out of 1.0), and (5) `no_movement`, no body movement detected for a configurable duration (default 10 seconds). The priority order is a deliberate design decision: a hand raise always takes precedence over passive disengagement signals because it represents an active intent to communicate.

When the trigger fires, the Engagement Manager captures a snapshot of the current state of the sensor as an observation dictionary. This ensures the observation reflects conditions at the time the trigger occurred, rather than at the time the LLM responds, which could be several seconds later. A 15-second cooldown period prevents the same trigger from being repeated. The trigger priority order, the 5-second observation window, and the timeout, threshold, and cooldown values are untuned engineering defaults chosen from pilot observation. A systematic ablation of these choices is left for future work.

D. LLM Responder

The LLM Responder constructs a structured prompt from four components: (1) a system prompt defining the robot's role and strict output rules, (2) the observation snapshot with current sensor values, (3) a history block containing the last three LLM responses with an instruction not to repeat them, and (4) in the Grounded Mode, a plain-English situation description of the trigger type. The complete prompt is sent to a Phi-3-mini (3.8B parameter) LLM [4], served from a remote endpoint that is accessible from both the PC and the Raspberry Pi. The module enforces a word count between 5 and 35 words and validates the output to reject responses that contain raw sensor field names (e.g., `gaze_ratio` or `head_yaw`), which would be meaningless to the user. If a response fails validation, the module retries up to three times with increasing temperature values (0.7, 0.9, 1.1). If all retries fail, a pre-written fallback phrase specific to the trigger type is used instead.

E. Speech Output and Logging

The Speech Output module converts the LLM-generated response to speech using platform-specific text-to-speech (TTS) engines: Windows Speech API (SAPI) on PC and Piper TTS on Raspberry Pi. A stop method allows mid-speech interruption when a hand raise is detected during playback. The Logger module records all session data into a structured JSON file, including every quiz interaction, sensor reading, engagement trigger, the full prompt sent to the LLM, the raw LLM response, retry logs, response timing, and the LLM backend identifier for post-session analysis.

F. Privacy and Ethical Considerations

Because the intended end users are children and the system uses a webcam, a microphone, and session logging, privacy and safeguarding are first-order design concerns. Camera frames are processed in memory and are not stored. Only the derived observation dictionary, the transcribed text, and the generated responses are written to the session log, and no raw video or audio is retained. All testing reported in this paper used consenting adult volunteers, and no minors have participated. Any future study with children will proceed only under an approved Institutional Review Board (IRB) protocol, with informed parental consent and child assent, defined retention limits for the JSON logs, and the option to delete a session on request. We also acknowledge that face and pose detection can exhibit accuracy bias across skin tones, lighting, and seating geometry, which could systematically disadvantage some users. Auditing the perception layer for such bias is part of planned work. For safeguarding, the output validation layer (Section III-D) and the trigger-specific fallback phrases bound the range of responses the robot can speak, and the system neither requests nor stores personally identifying information.

IV. Prompt Design: Grounded Mode vs Ungrounded Mode

A central design question is how much contextual information the LLM needs to generate appropriate responses. We evaluated two prompt strategies using the Phi-3-mini model:

Grounded Mode (Observation Dictionary and Trigger Type): The prompt includes both the sensor observation snapshot and an explicit plain-English trigger description (e.g., "The user's face is not visible to the camera during the math quiz. Mention you can't see their face and ask them to come back."). This provides the LLM with both low-level sensor data and a human-interpretable summary of the engagement state.

Ungrounded Mode (Observation Dictionary Only): The prompt includes only the sensor observation snapshot without any trigger description. The LLM must infer the user's engagement state from the numerical sensor values alone.

Both modes include identical quiz context (current question, topic, user age), identical behavioral constraints (5 to 35 word limit, no sensor values, audience-appropriate language), and the same history block of previous responses. The comparison

isolates the effect of explicit trigger labeling on response relevance. By construction, the Grounded Mode receives information the Ungrounded Mode does not, so the comparison is not a like-for-like contest between equally informed systems. It is a measurement of how much the explicit label contributes to relevance.

We conducted a preliminary manual relevance review of all LLM responses across both modes. Each response was labeled relevant if it addressed the correct trigger context, and not relevant otherwise, following a written rubric. A single author performed the relevance scoring against this rubric. Independent scoring by multiple annotators, with a measure of inter-rater agreement, is planned as part of future validation. This review measures only whether a response matched the trigger context. It does not assess pedagogical quality, tone, or actual re-engagement effect.

TABLE I. GROUNDED MODE VS UNGROUNDED MODE RESPONSE RELEVANCE BY TRIGGER TYPE (MANUAL REVIEW).

	Grounded		Ungrounded	
Trigger Type	**R / Total**	**%**	**R / Total**	**%**
Hand raised	11 / 13	84.6	1 / 14	7.1
Face absent	11 / 13	84.6	0 / 12	0.0
No answer	9 / 11	81.8	3 / 10	30.0
No movement	14 / 16	87.5	9 / 19	47.4
Head turned	17 / 19	89.5	4 / 18	22.2
Total	**62 / 72**	**86.1**	**17 / 73**	**23.3**

R = relevant responses scored by manual review.

The difference is substantial. In the Ungrounded Mode, the LLM frequently misidentified the engagement trigger. For example, when a user's face was absent from the frame, Ungrounded Mode responses sometimes addressed gaze direction or encouraged continued focus, rather than acknowledging the user's absence. When a user raised their hand, the Ungrounded Mode could not connect the boolean flag `hand_raised: yes` to the concept of a user wanting to ask a question, because the model's training data contains no mapping between that field name and the associated intent. Grounded Mode responses consistently matched the trigger type because the explicit plain-English description removed the inference burden from the LLM.

This finding is consistent with known limitations of LLMs in ToM reasoning. Ullman [14] demonstrated that even large LLMs fail on trivial alterations to standard Theory of Mind tasks, suggesting that a 3.8B parameter model cannot be expected to infer engagement states from raw sensor values without explicit guidance. More broadly, this result is an empirical confirmation that an explicit semantic-translation layer is required between numeric sensor output and an LLM: the model behaves as a semantic reasoning engine rather than a numeric or boolean classifier, so feeding it raw sensor arrays bypasses its core strengths. The 86.1% relevance of the Grounded Mode is achieved precisely because the plain-English trigger description supplies that semantic abstraction. For our pipeline, the Grounded Mode is therefore the required configuration.

V. PRELIMINARY RESULTS

We conducted 8 test sessions with consenting adult volunteers using the Grounded Mode on both a desktop PC (Windows) and a Raspberry Pi 5 (Ubuntu 24.04, 8 GB RAM, Vosk for speech-to-text, Piper for text-to-speech). Table II compares platform-level performance.

TABLE II. PC VS RASPBERRY PI LATENCY (GROUNDED MODE, AVERAGED ACROSS SESSIONS).

Metric	**PC**	**Raspberry Pi 5**
Avg TTS Question (s)	3.91	5.44
Avg TTS Feedback (s)	1.72	3.29
Avg STT (s)	0.93 (Whisper)	0.08 (Vosk)
Avg LLM Response (s)	0.89	0.54

LLM latency weighted across 42 (PC) and 27 (Pi) events.

The pipeline achieved zero observation mismatches and zero sensor field leaks across all test sessions. This confirms that the snapshot approach (Section III-C) and the output validation layer (Section III-D) function correctly on both hardware configurations.

These results should be read as a small pilot rather than a controlled study. The relevance review covers 72 grounded and 73 ungrounded responses, and the latency figures aggregate 42 PC and 27 Pi trigger events. Responses within a single session are not fully independent, since they share the same user and learning context. The trigger-specific fallback phrases (Section III-D) also act as an implicit templated baseline, but a controlled comparison against a full templated-response system and against a cloud-hosted LLM is left for future work (Section VI).

VI. CONCLUSION AND FUTURE WORK

This paper presented a multimodal pipeline that implements a simplified, behavior-based approximation of Theory of Mind for educational robots by combining real-time engagement sensing with observation-grounded LLM response generation. The Grounded Mode vs Ungrounded Mode comparison confirms that small LLMs require explicit trigger labeling, achieving 86.1% relevance compared to 23.3% without it, which empirically validates the need for a semantic-translation layer between sensor output and the model. Although the architecture is configurable for other domains and audiences, this paper evaluates it only in a mathematics-quiz scenario for children aged 7 to 14. Therefore, such generalization remains to be demonstrated beyond the present scenario.

Several important challenges remain unresolved. First, the quality and pedagogical appropriateness of LLM responses have not been formally validated. While a manual relevance review confirmed that Grounded Mode responses address the correct trigger context at 86.1%, this does not assess whether responses are educationally effective, appropriate in tone, or helpful for re-engaging the user. Four validation approaches

are planned. The first is LLM-as-judge evaluation, using a larger LLM to score responses against pedagogical rubrics. The second is expert review by education specialists. The third is IRB-approved studies with participants in the target age range to measure actual re-engagement effects. The fourth is controlled baseline comparisons against a templated-response system and a cloud-hosted LLM, to quantify how much response quality the local small model gives up.

Second, the current system relies on a remote graphics processing unit (GPU) endpoint for LLM inference. Moving to on-device inference on the Raspberry Pi 5, potentially using a neural processing unit accelerator, such as the Hailo HAT+2, would eliminate network dependency and enable fully offline operation, which is important for classroom deployments with limited connectivity.

Third, the pipeline currently operates without a physical robot. The target platform, a Marty V2 robot [5] connected via ROS 2 Jazzy [6], will provide physical embodiment through walking and gesturing synchronized with verbal responses. Physical embodiment has been shown to increase engagement and trust in human-robot interaction [13].

Fourth, the current ToM model only detects behavioral engagement states, such as attentive, distracted, absent, and questioning. Adding the ability to infer emotional states, such as frustration, boredom, or confusion, through facial expression analysis and speech tone would allow the system to generate more targeted responses. For example, a frustrated student needs a different response than a bored student, even if both show similar gaze patterns. This would move the system from detecting where the user is looking to approximating how the user is feeling.

REFERENCES

[1] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[2] J. Wang *et al.*, "Large language models for robotics: Opportunities, challenges, and perspectives," 2024. arXiv: 2401.04334 [cs.RO].

[3] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," 2019. arXiv: 1906.08172 [cs.DC].

[4] M. Abdin *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," 2024. arXiv: 2404.14219 [cs.CL].

[5] Robotical Ltd., "Marty the robot V2," 2026, [Online]. Available: https://robotical.io/product/marty-the-robot-v2/ (visited on 03/31/2026).

[6] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, eabm6074, 2022.

[7] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.

[8] B. Chen, C. Vondrick, and H. Lipson, "Visual behavior modelling for robotic theory of mind," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[9] M. Hellou, S. Vinanzi, and A. Cangelosi, "A review of theory of mind and robotics: Mind reading in human-robot interaction for proactive social robots," in *Human-Friendly Robotics 2024 (HFR 2024)*, ser. Springer Proceedings in Advanced Robotics, vol. 35, Springer, 2025, pp. 197–211. DOI: 10.1007/978-3-031-81688-8_15.

[10] J. Wang, S. Yuan, T. Lu, H. Zhao, and Y. Zhao, "Video-based real-time monitoring of engagement in e-learning using MediaPipe through multi-feature analysis," *Expert Systems with Applications*, vol. 267, p. 128 239, 2025. DOI: 10.1016/j.eswa.2025.128239.

[11] M. Mauliana, A. Ashok, D. Czernochowski, and K. Berns, "Exploring LLM-powered multi-session human-robot interactions with university students," *Frontiers in Robotics and AI*, vol. 12, p. 1 585 589, 2025. DOI: 10.3389/frobt.2025.1585589.

[12] K. Smit, S. Leewis, H. Almoustafa, K. Yildrim, and T. Uymaz, "Enhancing educational dynamics: Integrating large language models with a social robot," in *Proceedings of the 2024 8th International Conference on Software and e-Business (ICSeB '24)*, ACM, 2024, pp. 87–94. DOI: 10.1145/3715885.3715889.

[13] M. Elgarf, H. Salam, and C. Peters, "Fostering children's creativity through LLM-driven storytelling with a social robot," *Frontiers in Robotics and AI*, vol. 11, p. 1 457 429, 2024. DOI: 10.3389/frobt.2024.1457429.

[14] T. Ullman, "Large language models fail on trivial alterations to theory-of-mind tasks," 2023. arXiv: 2302.08399 [cs.CL].