Intelligent Robotic Tutoring: Integrating Verbal Input for Personalizing Learning Responses

Puming Jiang
Imperial College London
London, UK
o.jiang23@imperial.ac.uk

Nicole Salomons
Imperial College London
London, UK
n.salomons@imperial.ac.uk

ABSTRACT

Previous research has demonstrated that robots can be effective teaching aids, yet they lack the ability to understand and respond to students' speech in the way human teachers do. In this paper, we introduce an innovative approach that utilizes Large Language Models (LLMs) to enable more intelligent teaching robots, thereby enhancing the learning experience. Initially, we explore the capability of LLMs to evaluate students' learning progress through their speech, surpassing traditional student progress models like Bayesian Knowledge Tracing, which depend solely on the correctness of students' answers. We then proceed to discuss the second phase, which involves comparing various mechanisms of using LLMs to generate personalized feedback, and incorporating the most appropriate one into the final framework.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Computing methodologies** → *Natural language processing*; Robotics; • **Human-centered computing** → Interaction paradigms.

KEYWORDS

Large Language Model, Intelligent Tutoring System

ACM Reference Format:
Puming Jiang and Nicole Salomons. 2018. Intelligent Robotic Tutoring: Integrating Verbal Input for Personalizing Learning Responses . In *Proceedings of Human – Large Language Model Interaction Workshop (HRI '24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION AND MOTIVATION

Social robots have emerged as educational tools, acting as peers or tutors to enhance student learning [3]. Although the effectiveness of teaching robots has been confirmed [1, 2, 6, 9], they have not yet achieved the capabilities of human teachers, such as continuously monitoring student progress, interpreting speech, answering queries, and providing personalized feedback. While research indicates that encouraging students to verbalize their thoughts while answering questions can boost their learning gains [8], robots currently have a limited understanding of such speech and gauge their progress subsequently.

Instead, teaching robots typically rely on user modelling frameworks like Bayesian Knowledge Tracing (BKT) to track students [4]. BKT predicts students' likelihood of understanding the skill necessary to answer a question. Variations of the BKT model have also been proposed to achieve more personalized modeling [11]. However, such estimations overlook a crucial aspect: understanding students' verbal inputs. This gap is a significant difference between human and robot teachers, where the latter lacks the capability to provide explanations based on students' verbal questions.

Large Language Models (LLMs), leveraging the transformer network architecture proposed by [10], have shown remarkable capabilities in interpreting natural languages and offer the potential to enhance the intelligence of teaching robots. In this paper, we present our approach for integrating LLMs into teaching robots, inspired by research indicating that encouraging students to verbalize their thoughts while solving problems can lead to increased learning gains [8]. We refine this concept by employing LLMs to analyze the textual data obtained from students' spoken responses after conversion by a speech-to-text module, thereby developing a more precise model of student progress. Additionally, we employ another LLM to generate personalized feedback, considering not only the correctness of students' answers but also the nuances of their verbal responses.

2 PROPOSED FRAMEWORK

We are proposing two novel contributions using students' speech during tutoring. In other words, we utilize LLMs to a) create a more accurate personalized model of the student using their speech while learning; and b) generate feeback according to the students personal learning needs.

2.1 LLMs for User Modeling

Traditional methods like BKT estimate skill mastery based solely on answer correctness [4], but LLMs can enhance accuracy by also considering students' verbal inputs. These verbal inputs are captured by encouraging students to think aloud while answering questions, as explored by [8]. As illustrated in Figure 1 and inspired by [7], LLM could assign a 0 to 10 score reflecting the likelihood of skill comprehension. Inputs to LLM include the question, answer options, correct answer, student's choice, solution text, and the student's spoken responses, converted via speech-to-text. Such predictions can then either be averaged with BKT's prediction or act as a direct substitution.

HRI '24, March 11-15, 2024, Boulder, CO

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

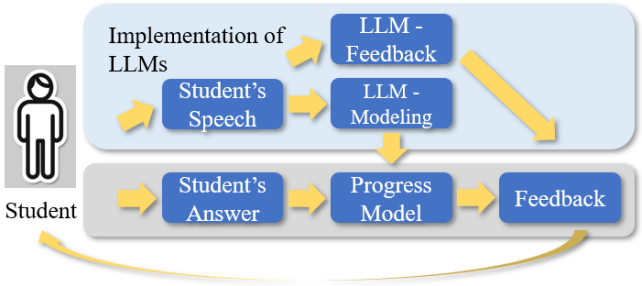

Figure 1: The Implementation of LLMs. "LLM - Modeling", as referenced in section 2.1, is utilized for user modeling. "LLM - Feedback", as mentioned in section 2.2, is employed to generate more personalized feedback.

Another innovative approach involves integrating LLM with the traditional BKT method. For traditional BKT, we calculate the probability of mastery given the student's correctness of the answer, as illustrated when a student answers correctly:

$$P(L|correct) = \frac{P(L) \cdot (1 - P(slip))}{P(L) \cdot (1 - P(slip)) + (1 - P(L)) \cdot P(guess)}$$

Where:

- $P(L|correct)$ is the updated probability of mastery given a correct answer.
- $P(L)$ is the prior probability of mastery.
- $P(slip)$ and $P(guess)$ are the probabilities of slipping and guessing, respectively.

Incorporating LLMs to consider verbal inputs, we adjust observation from correctness to LLM's binary prediction of skill mastery, $P(pL)$. The equation below represents the scenario when the LLM predicts True:

$$P(L|pL) = \frac{P(L) \cdot P(pL|L)}{P(L) \cdot P(pL|L) + (1 - P(L)) \cdot P(pL|\neg L)}$$

Where:

- $P(L|pL)$ is the updated probability of mastery based on LLM prediction.
- $P(pL|L)$ and $P(pL|\neg L)$ are the probability of LLM predicting mastery when the skill is mastered and not mastered, respectively.

$P(pL|L)$ and $P(pL|\neg L)$ are pre-determined True Positive and False Positive Rate. This model is likely to offer a more precise estimation of mastery, given the added understanding of speech.

2.1.1 Evaluation. We plan to conduct within-subject user studies with approximately twenty adult participants, who will learn probability with the help of robots. All participants will receive the same set of questions. We will simultaneously employ three user modeling techniques (three conditions) to track progress: the traditional BKT model, LLM for direct evaluation, and LLM's binary prediction, which will serve as input to Bayes' rule. At the conclusion of the teaching session, each model will generate predictions about each participant's skills. We will then compare the accuracy of these predictions against the ground truth.

Generating ground truth is challenging, and experienced teachers will be recruited to determine whether students have understood each skill by observing the entire teaching session. Uncertain predictions should be double-checked by posing questions to the students after the session.

2.2 LLMs for Robot Feedback

LLMs have the potential to offer personalized feedback to students by interpreting their speech, thereby addressing each student's unique needs and responses. If implemented successfully, this could significantly enhance teaching efficiency.

The main challenge, however, is ensuring the accuracy of the feedback. Directly inputting students' responses into LLMs could result in "hallucinations" [5], leading to feedback that is either inaccurate or irrelevant. To counter this, one approach involves LLM accessing a predefined solution corpus to retrieve personalized feedback, which reduces but doesn't eliminate the risk. The most cautious method selects feedback from a set of predefined options, which avoids inaccuracies but limits the depth of personalization. The central research question revolves around finding the optimal balance between personalization and accuracy.

It is important to note that the feedback discussed above, namely informational feedback, is not the sole type of feedback that can be offered. Another aspect worth exploring involves affective feedback, focusing on how to employ LLMs to mimic human teachers' praise, encouragement, comfort, and even criticism—either before, during, or after providing explanations on solving problems. Although such feedback can be readily generated through narrowly defined prompts and few-shot prompting, the effectiveness and strategy need to be verified beforehand.

LLMs also offer the potential for more nuanced support by providing continuous assistance throughout the question-solving process. Specifically, if a student shows confusion or poses a question while attempting to solve the problem, the LLM can detect this state and generate appropriate assistance. However, such advanced applications must be approached carefully to avoid interrupting the student's thought process. The key research question then becomes how to optimally time the provision of help so that it is perceived as beneficial and not as an unwelcome interruption.

2.2.1 Evaluation. For feedback generated by LLMs, we are going to first generate the rejection rate, which includes rejections due to hallucinations or rejections due to unhelpful/unrelated feedback. Rejections resulting from hallucinations should be regarded as significantly more critical than those due to unhelpfulness. This evaluation could be a preliminary study before conducting a large-scale user study, allowing us to pre-filter some conditions. We are planning to recruit at least ten human evaluators who assess or rate the responses generated by different approaches, with a condition of human-generated feedback serving as the baseline.

A between-subject study can then assess whether students learning with LLM-assisted robots outperform those receiving predefined feedback that is either fixed or dependent on their selections (for multiple-choice questions). Post-study interviews will collect student feedback on the robot's effectiveness, while engagement levels—potentially measured by eye contact with the robot—offer additional insights into the learning experience.

REFERENCES

[1] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 701–706.

[2] Paul Baxter, Emily Ashurst, Robin Read, James Kennedy, and Tony Belpaeme. 2017. Robot education peers in a situated primary school study: Personalisation promotes child learning. *PloS one* 12, 5 (2017), e0178126.

[3] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3, 21 (2018), eaat5954. https://doi.org/10.1126/scirobotics.aat5954

[4] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4 (1994), 253–278.

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[6] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 34.

[7] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[8] Aditi Ramachandran, Chien-Ming Huang, Edward Gartland, and Brian Scassellati. 2018. Thinking aloud with a tutoring robot to enhance learning. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 59–68.

[9] Nicole Salomons, Kaitlynn Taylor Pineda, Adérónké Adéjàre, and Brian Scassellati. 2022. We Make a Great Team!: Adults with Low Prior Domain Knowledge Learn more from a Peer Robot than a Tutor Robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 176–184.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[11] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*. Springer, 171–180.