Evaluation of an LLM-powered student agent for teacher training

Saptarshi Bhowmik[1], Luke West[1], Alex Barrett[1], Nuodi Zhang[1], Chih-Pu Dai[2], Zlatko Sokolikj[1], Sherry Southerland[1], Xin Yuan[1], Fengfeng Ke[1]

[1]Florida State University, Tallahassee, FL
[2]University of Hawai'i at Mānoa, HI

Abstract. As technology continues to advance, there is a growing interest in exploring the potential of generative agents and large language model (LLM)-powered virtual students to revolutionize the field of education. In this work, we present Evelyn AI, a LLM-powered virtual student conversation agent that we developed for pre-service teacher training in a virtual environment. Students powered by Evelyn AI exhibit varying baseline conceptual understanding levels, dynamic cognitive-affective states, and short-term memory. These features enable personalized, adaptive training and promote a more engaging and immersive learning experience for pre-service teachers. We describe the design and implementation of Evelyn AI, and report results of alpha testing to assess the utility of Evelyn AI for pre-service teacher training.

Keywords: Teaching simulation · Teacher education · Large Language Model · Sentiment Analysis · Cognitive-affective states · Virtual Student Agent

1 Introduction

Virtual reality (VR) and artificial intelligence (AI) are two emerging technologies that are garnering considerable attention in teacher education. The combination of these two powerful technologies has broad implications for research in education and instructional design.

VR, including 3D virtual worlds, has been used to create simulations for teacher training [3, 10]. Virtual teacher training simulations afford users classroom teaching experience without the pragmatic difficulties of training in actual classroom environments. Specifically, with virtual simulations teachers can engage in perspective taking by pausing and reflecting [8], engage with active coaching from a trainer [1], and be exposed to diverse classroom compositions that can develop their competencies in inclusive education [2].

An example of successful virtual teacher training environments is TeachLive [10] where student avatars are puppeteered by human actors who simulate the student discourse and affective states through preset protocols. Due to the use of human actors, such a training environment is costly to use and maintain and limited in its scalability. To overcome these limitations, this research seeks to replace human actors with a generative AI large language model (LLM) [11] [9]. To this end, we developed Evelyn AI, a LLM-powered virtual student agent, for our virtual pre-service teacher training program. The main design objective of Evelyn AI is to support authentic student-teacher interactions and to enable dynamic classroom interactions with diverse students.

These dynamic, diverse classrooms feature students that challenge teachers to adapt instruction to (1) persistent student differences (static traits such as baseline conceptual understanding about a topic), (2) transient student differences more readily changing throughout class (dynamic cognitive-affective states), and (3) evolving verbal contexts of the classroom discussion. Supporting these features poses significant challenges in the design and development of Evelyn AI. We will discuss the design and implementation of Evelyn AI that addresses these challenges and report on a series of three investigations into Evenlyn IA utility for simulating classroom discussions.

2 Key Features of Evelyn AI

The key features of Evelyn AI are what enable it to generate realistic classroom conversation with a wide-range of student characteristics, which allows for personalized and adaptive training for pre-service teachers.

1. **Static trait: baseline conceptual understanding**. A key feature of Evelyn AI is the ability to simulate students who differ in static traits. Baseline understanding of a topic is a static student trait which we intend as a representative component of a student's personal epistemology, from which the student may draw productive resources that can be integrated in a classroom context [6]. We chose to represent differences in baseline understanding because it plays a key role in the orchestration of classroom discussion and because effectively teaching students with different baseline understandings is a fundamental skill that pre-service teachers need to develop.
2. **Dynamic cognitive-affective states**. Another key feature of Evelyn AI is the ability to simulate cognitive-affective states. A student in a particular cognitive-affective state (*e.g.*, bored or engaged) can exhibit significantly different learning interactions. As shown in Figure 1, Evelyn AI allows for transition among six cognitive-affective states: (1) Bored, (2) Fatigued, (3) Engaged, (4) Anxious, (5) Distressed, and (6) Fed Up. The six states are controlled by two variables: arousal and valence. These cognitive-affective states and the control variables are derived from earlier research on student cognitive-affective states [12, 4, 5]. By incorporating dynamic student cognitive-affective states in our simulation, preservice teachers can get valuable practice in identifying verbal and nonverbal cues, associating them with specific cognitive-affective states, and responding accordingly to shape productive learning experiences.

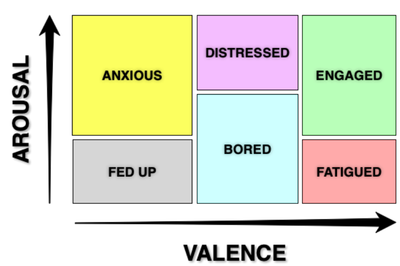

Fig. 1: Cognitive-affective states with arousal and valence as two variables

3. **Authentic context-sensitive student-teacher interaction**. The purpose of Evelyn AI is to foster realism in a virtual interactive classroom so that pre-service teachers can undergo challenging and effective training with the goal of seamless transition to real-life classrooms. One important part of realism in student-centered classroom interactions is the impact of previous classroom dialogue on the ongoing conversation, including how students recall and incorporate recent interactions. This kind of information requires registration in a memory; in naive terminology, we refer to this as short term memory [7]. Thus, to enhance the realism of the simulation and allow for referential dialogue, a certain degree of short term memory was incorporated into Evelyn AI.

2.1 Architectural overview of Evelyn AI

Figure 2 illustrates the major components in Evelyn AI as well as its workflow. Evelyn AI takes pre-service teacher's prompts or responses as the input and generates a student

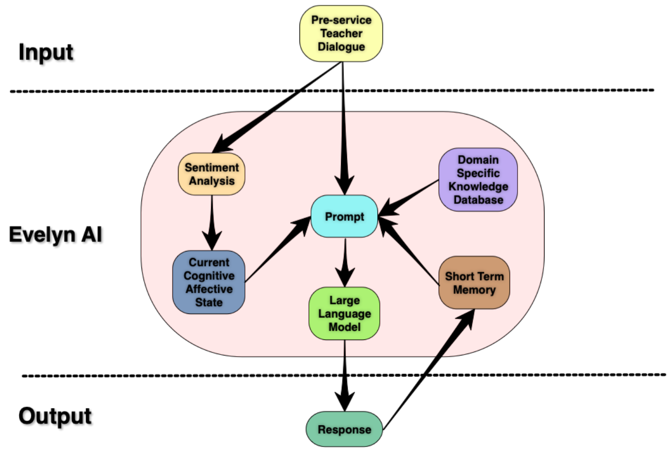

Fig. 2: Overview of Evelyn AI

response to the pre-service teacher. Each provided input is influenced by a variety of components.

1. Evelyn AI utilizes a LLM as its back-end response generator that produces realistic human-conversation-like responses. Evelyn AI can work with any open source LLM. But in our experiments, OpenAI GPT 3.5 [9] is used. To generate responses from a diverse set of students, Evelyn AI relies heavily on prompt engineering, which feeds multiple prompts to the LLM model in addition to the user input.
2. To simulate students with different prior baseline knowledge of the subject matter, Evelyn AI includes a domain-specific knowledge base of STEM subject matter. Prompt engineering allows Evelyn to simulate students of low, medium, and high levels of understanding of this knowledge.
3. To simulate students with different cognitive-affective states, Evelyn AI utilizes a sentiment analysis module to assess the verbal sentiment of the pre-service teacher. Based on the result of sentiment analysis, Evelyn AI dynamically adjusts student's cognitive-affective state.
4. To enhance the realism of teach-student interactions, we also incorporate an additional layer of short-term memory in Evelyn AI. This layer enables students to provide information and interact based on previous interactions in the classroom.

3 Detecting the Baseline Understanding, Cognitive-Affective States, and Contextual Adaptation in Education

3.1 Method

In this study, we analysed alpha-testing data of the Evelyn AI student agent. The Evelyn AI model for each student agent included the same custom knowledge base with 2,726 rounds (*i.e.*, turns-at-talk) of dialogue taken from transcripts of 19 middle-school classroom videos, including recordings from the United States, United Kingdom, and Japan (translated). We non-randomly recruited a total of 6 raters to evaluate the AI-student responses for both baseline understanding (Challenge 1), cognitive-affective states (Challenge 2) and Evelyn AI's ability to adapt to classroom discussion (Challenge 3). Each rater independently scored the responses.

3.2 Results

Challenge 1 The ICC for baseline understanding among six raters was .71 ($p < .001$), showing strong agreement. A Pearson's correlation analysis on 24 AI responses from three student agents revealed a moderately positive correlation of 0.47 ($p = .015$), indicating that differences in baseline understanding could be detected by humans with moderate consistency. Table 1 shows responses for the first two rounds of dialogue, highlighting correlations between intended and human-rated scores.

Teacher	Student 1 (Low Understanding)	Student 2 (Medium Understanding)	Student 3 (High Understanding)	Correlation Score
Angela, can you share something you remember about fungi from last class?	Um, yeah. I remember that fungi are plants and they reproduce through binary fission.	Um, yeah. Fungi can't move. They stay in one place.	Oh, um, I remember that fungi are not plants because they don't reproduce by binary fission like plants do. They reproduce through spores.	
Great. In your own words, what can you tell me about how fungi get nutrition?	Fungi get nutrition by eating bacteria and decomposing food. They break down the food and absorb the nutrients from it.	Um, fungi get nutrition by breaking down things like food or dead plants and animals, and then they absorb the nutrients from them.	Fungi get nutrition by breaking down organic matter, like dead plants or animals, and absorbing the nutrients from it. They use enzymes to break down the matter and then absorb the nutrients into their cells.	0.47 ($p = .015$)

Table 1: Verbal responses and correlation between human rater and intended scores for 3 Evelyn AI-powered student agents with low, medium or high prior knowledge.

Challenge 2 The ICC for cognitive-affective state ratings was .713 ($p < .001$), showing strong agreement among raters. A Pearson's correlation analysis on 27 AI responses revealed a strong positive correlation of 0.815 ($p < .001$), indicating that cognitive-affective differences were detectable by humans with moderate consistency. Table 2 presents responses for the first three rounds of dialogue, showing correlations between intended valence scores and human ratings.

Challenge 3 To assess Evelyn AI's adaptability in evolving classroom discussions, we explored its context-sensitive memory abilities. We examined whether the student agent could formulate, recall, and compare hypotheses after several dialogue rounds. Table 3 shows a dialogue with a student agent (low prior knowledge; bored state), demonstrating effective hypothesis formulation and recall, relevant to middle school life sciences. These abilities are crucial for facilitating engaging and equitable group discussions. Qualitative data on user experience with these dialogues was also collected.

4 Conclusion

In conclusion, this paper has presented the development of Evelyn AI, an innovative student agent driven by the powerful OpenAI GPT 3.5 Large Language Model for the purposes of preservice teacher training. Our work has addressed the critical challenge of

Teacher	Student 1 (Fed Up)	Student 2 (Bored)	Student 3 (Engaged)	Correlation Score
Welcome back to biology class! Today we'll keep talking about fungi and build on some of the concepts we learned last week.	Ugh, more fungi... I'm so tired of this.	Okay, sure. What do you want to know about fungi?	Great, I'm excited to learn more about fungi!	
Angela, can you share something you remember about fungi from last class?	Yeah, fungi are like mushrooms and stuff, right?	Um, yeah. Fungi are like, living things that can grow on trees and stuff.	Um, yeah. Fungi are like mushrooms and they can grow on trees and stuff.	0.82 ($p < .001$)

Table 2: Verbal responses and correlation between human rater and intended scores for 3 Evelyn AI-powered student agents in low-, medium- or high-valence cognitive-affective states.

Teacher Text	Student Medium Valence [Bored]; Low Prior Knowledge
Susan had the hypothesis that fungi only reproduce asexually. Juan, in your own words, how do you think fungi reproduce?	I think fungi reproduce by binary fission, just like plants.
Great thoughts on that. Can you tell us in your own words why you think that?	Um, well, I think fungi reproduce by binary fission because they, uh, split into two parts and each part becomes a new fungi. Yeah, that's it.
Ok. What effect do you think habitat might have on fungi reproduction?	Um, I think habitat might have an effect on fungi reproduction because different habitats have different nutrients and conditions that fungi need to reproduce. So, like, if the habitat has more nutrients and moisture, the fungi might reproduce more.
Why do you think nutrients might be important for reproduction?	Um, I think nutrients might be important for reproduction because, like, fungi need energy to make new fungi, and nutrients provide that energy. So, yeah, nutrients are important for reproduction.
Great. Let's think about our hypotheses so far. **What was Susan's hypothesis?**	Um, **Susan's hypothesis was that fungi only reproduce asexually.**
How do you think **her hypothesis** is different from yours?	Um, I think her hypothesis is different from mine because I think fungi reproduce by binary fission, while **she thinks they only reproduce asexually.**

Table 3: 6 rounds of dialogue with an Evelyn AI-powered student agent demonstrating short-term memory capabilities for classroom discussion.

achieving realistic classroom verbal interactions from multiple angles. We have striven to create a dynamic classroom environment where each student possesses domain-specific knowledge akin to long-term memory. Furthermore, we have introduced the concept of cognitive affective states, ensuring that student responses align with their current mental states. Our innovative approach introduces dynamic conversation agents within the simulated classroom, allowing them to adapt their cognitive affective states in response to the evolving dynamics of the simulation. This multifaceted approach marks a significant step forward in achieving lifelike classroom interactions within the realm of education and human-computer interaction, opening new avenues for the advancement of interactive educational technologies.

References

1. Cohen, J., Wong, V., Krishnamachari, A., Berlin, R.: Teacher coaching in a simulated environment. Educational Evaluation and Policy Analysis **42**(2), 208–231 (2020). https://doi.org/10.3102/0162373720906217, https://doi.org/10.3102/0162373720906217
2. Connolly, T., Tsvetkova, N., Hristova, P.: Gamifying teacher training: Simulated practice learning for future and practising teachers interacting with vulnerable learners. Games and Simulations in Teacher Education pp. 55–73 (2020)
3. Dai, C.P., Ke, F.: Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. Computers and Education: Artificial Intelligence p. 100087 (2022)
4. D'Mello, S., Taylor, R., Graesser, A.: Monitoring Affective Trajectories during Complex Learning, pp. 203–208. the Cognitive Science Society (2007)
5. Gobron, S., Ahn, J., Paltoglou, G., Thelwall, M., Thalmann, D.: From sentence to emotion: a real-time three-dimensional graphics metaphor of emotions extracted from text. The Visual Computer **26**, 505–519 (2010)
6. Hammer, D., Elby, A.: Tapping epistemological resources for learning physics. The Journal of the Learning Sciences **12**(1), 53–90 (2003)
7. Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S.: The mind and brain of short-term memory. Annu. Rev. Psychol. **59**, 193–224 (2008)
8. Ke, F., Xu, X.: Virtual reality simulation-based learning of teaching with alternative perspectives taking. British Journal of Educational Technology **51**(6), 2544–2557 (2020)
9. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
10. Peterson-Ahmad, M.B., Pemberton, J., Hovey, K.A.: Virtual learning environments for teacher preparation. Kappa Delta Pi Record **54**(4), 165–169 (2018)
11. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
12. Russell, J.A.: Core affect and the psychological construction of emotion. Psychological review **110**(1), 145 (2003)