EduAgent: Generative Student Agents in Learning

Songlin Xu[1] **Xinyu Zhang**[1] **Lianhui Qin**[1,2]
[1]University of California, San Diego
[2]Allen Institute for Artificial Intelligence
soxu@ucsd.edu

Abstract

Student simulation in online education is important to address dynamic learning behaviors of students with diverse backgrounds. Existing simulation models based on deep learning usually need massive training data, lacking prior knowledge in educational contexts. Large language models (LLMs) may contain such prior knowledge since they are pre-trained from a large corpus. However, because student behaviors are dynamic and multifaceted with individual differences, directly prompting LLMs is not robust nor accurate enough to capture fine-grained interactions among diverse student personas, learning behaviors, and learning outcomes. This work tackles this problem by presenting a newly annotated fine-grained large-scale dataset and proposing EduAgent, a novel generative agent framework incorporating cognitive prior knowledge (i.e., theoretical findings revealed in cognitive science) to guide LLMs to first reason correlations among various behaviors and then make simulations. Our two experiments show that EduAgent could not only mimic and predict learning behaviors of real students but also generate realistic learning behaviors of virtual students without real data.

1 Introduction

Online education plays a crucial role not only as a strategic response to a wide variety of disruptions, including natural disasters and public health emergencies Uscher-Pines et al. [2018], but also as a universally accessible platform to promote inclusivity McLoughlin [2001] for students facing challenges in attending traditional in-person classes. However, online education suffers from several intrinsic limitations that hamper its effectiveness. In particular, online platforms lack effective mechanisms for the instructor to perceive the students' responses in real time as an ensemble. Such perceptions are needed for the instructors to gauge the students' understanding and decide on appropriate lecture adjustments Bond et al. [2020], Deslauriers et al. [2011]. "Intelligent tutor" systems have been proposed to provide feedback to student/teachers, but mostly following hand-crafted rules Syed et al. [2020], Craig et al. [2004], Zhao et al. [2021]. AI-models can be a powerful alternative, but they either lack real-time responses (e.g., only responding students' final test scores Bassen et al. [2020] or they can only offer "chatty" Markel et al. [2023] interactions which are atypical in video lecture-based online education.

Ideally, the online education should grant instructors a similar or even more granular perception of students' learning behavior, e.g., their engagement, cognitive load, modulated by the course content over time. To this end, AI models must overcome two fundamental barriers: (i) Fine-grained learning behavior modeling/prediction; (ii) Acquiring sufficient, labeled data of learning behaviors for model training. Behavior models can potentially overcome the data scarcity. However, existing student simulation models Piech et al. [2015], Beck and Woolf [2000], Hussain et al. [2019], Xu et al. [2017] themselves often need massive training data.

In this paper, we contribute a new $N = 310$ online education dataset (**EduAgent310**), consisting of fine-grained, multidimensional records of students' learning behaviors during slide-based lectures.

Preprint. Under review.

The dataset annotates the students' gaze trajectories, motor control behavior (moving a computer mouse), and 6 different cognitive states. These metrics are timestamped and mapped to the different content blocks on each slide. Each lecture ends with a comprehensive quiz to evaluate individual students' learning performance.

The EduAgent310 dataset can provide the ground-truth for modeling learning behaviors. However, a much larger dataset, with a larger student population and more diverse profiles, is needed for training AI-driven pedagogical models. Unlike common cyber-physical datasets, logging the elusive human cognitive states and behaviors can be extremely time-consuming and costly. We thus pose the question: *Can large generative models be used to produce realistic, fine-grained learning behavior data, similar to EduAgent310?* To answer this question, we develop a generative agent framework, called EduAgent, which enables us to benchmark the capabilities of state-of-the-art large language models (LLMs) in simulating the fine-grained learning behaviors in response to course content.

The challenge for the EduAgent framework lies in eliciting the LLM's capability to model sophisticated and dynamic correlation among the students' personas, behaviors, cognitive states, course content, etc. A straightforward prompt is obviously insufficient. Advanced LLM-based problem solving techniques, such as Chain of Thought (CoT) Wei et al. [2022] or few-shot demonstration cannot overcome the challenge either. The learning behavior simulation cannot be easily restructured into step-by-step subtasks which are amenable for CoT. On the other hand, few-shot demonstration may either fail to capture the dynamic, multi-faceted student profiles, or overfit to the demo itself.

Our EduAgent framework tackles the challenge by incorporating ***cognitive priors***, i.e., classical theories in cognitive science which delineate learning behaviors. Specifically, the correlation between personas, course content, and learning behaviors has been well established Karemera et al. [2003], but in a piecewise manner. Our EduAgent framework tries to capture the dynamics and the multi-dimensional relation simultaneously, in a modularized architecture which encompasses the different elements in an action space and memory space. First, we store different behaviors (such as gaze, cognitive states) in different layers of a memory module. We then prompt the LLM to reason how and why behaviors in each layer are modulated by personas (to capture the individual differences) and course contents (to model the learning process). We also prompt the LLM to reason the correlation between behaviors of different layers, preventing it from overfitting to the few-shot demonstration. By orchestrating the motor behaviors, persona information, and cognitive states following cognitive prior principles, EduAgent can model the learning process in a much finer-grained manner than prior works Chen et al. [2023], Jinxin et al. [2023], thus more accurately predict the learning performance.

Our experiments on the aforementioned EduAgent310 dataset show that the EduAgent framework can accurately predict a real student's learning behaviors and test results, even with a short history demonstration. Furthermore, using the EduAgent framework, we generate another dataset (**EduAgent705**) comprising $N = 705$ virtual students with more diverse personas. Our experiments verify that the simulated students exhibit behavioral patterns that are consistent with the real students', and with the cognitive priors, even when no real training data is provided.

In summary, this paper makes three main contributions:

- A large-scale and fine-grained newly annotated learning behavior dataset from real students ($N = 311$) and virtual students ($N = 705$).
- An open source generative agent framework [1], modularized following cognitive priors, to enable realistic simulation of learning behaviors in online education.
- Comprehensive experiments to verify the EduAgent framework and benchmark the capability of SoTA LLMs in modeling fine-grained learning behaviors.

Although our current dataset only contains 705 virtual students, the EduAgent framework can be used to generated an unlimited number of virtual students, bearing the cost of accessing the LLM APIs (e.g., \$0.2 or \$0.02 per-student through OpenAI GPT-4 or GPT-3.5). This can potentially overcome the data scarcity bottleneck, enabling the much-anticipated end-to-end human-in-the-loop training of intelligent tutor systems Bhutoria [2022].

[1] Data set and code are available: https://github.com/EduAgent/EduAgent

Table 1: Statistics of our dataset compared with existing student learning behavior datasets. **N**: participant number in the dataset, **Demo**, **Gaze**, **Motor**, **Cog**, **Test**, **Mat** represent whether the dataset contains student personas (demographics or characteristics), gaze behaviors, motor behaviors (such as moving computer mouse in online education or having any gestures in classroom), cognitive states, test question performance, and course/question materials, respectively. "–"means no explicit information of such data. "×"represents the lack of such data. "√"represents existence of such data.

DATASET	N	DEMO	GAZE	MOTOR	COG	TEST	MAT
KUZILEK ET AL. [2017]	32,593	√	×	√	×	√	×
BUI ET AL. [2022]	5,327	√	×	×	×	√	×
MARTÍN ET AL. [2015]	111	√	×	√	×	√	×
FAN [2023]	–	×	√	√	×	×	×
DELGADO ET AL. [2021]	19	×	√	×	√	–	×
KAUR ET AL. [2018]	78	×	√	×	√	–	×
HASAN ET AL. [2021]	326	√	×	√	×	√	×
RUIZ ET AL. [2022]	54	×	√	–	√	√	×
MAI ET AL. [2022]	400	–	×	√	×	√	×
SUN ET AL. [2021]	–	×	√	√	×	–	×
LIU ET AL. [2023]	18,066	–	×	√	×	√	√
CHOI ET AL. [2020]	1,677,583	–	×	√	×	√	√
WANG ET AL. [2021]	118,971	√	×	√	×	√	√
STAMPER AND PARDOS [2016]	1,146	–	×	√	×	√	√
POJEN ET AL. [2020]	247,606	√	×	×	×	√	√
STA	333	–	×	–	×	√	√
FENG ET AL. [2009]	4217	–	×	–	×	√	√
EDUAGENT	1,015	√	√	√	√	√	√

2 Related Work

2.1 Student Learning Behavior Modelling

Prior to the maturity of generative agents, substantial research has been devoted to modeling learning behaviors using generic deep learning methods. Student learning trace (i.e., records of a student's learning progress) can be modeled using RNN or similar structures Piech et al. [2015], Xiong et al. [2016], Chen et al. [2018], Minn et al. [2018]. Such "knowledge tracing" models can be improved by combining exercise contents as well Liu et al. [2019]. Other data-driven approaches have also been widely explored Lee et al. [2021], Waheed et al. [2020] for learning performance prediction.

2.2 Datasets for Learning Behavior Modeling

To support student behavior modelling, a variety of datasets have been created. Table. 1 makes a summary for comparison. Our EduAgent310 and EduAgent705 dataset is unique as it contains fine-grained records of students' cognitive processes in online education. Specifically, it incorporates diverse personas, gaze, motor behaviors, cognitive states, and post-test performance, all synchronized to related course/question materials. Mapping between such factors and learning performance has been a long standing problem in cognitive science Resnick [2017]. The dataset can enable the development of data-driven models for learning performance prediction, along with real-time feedback/intervention for the students and teachers Pardos et al. [2013], Liu et al. [2017].

2.3 LLMs and Agents in Education

The in-context learning capabilities of LLMs have been harnessed to create emergent agents Lin et al. [2023] in diverse applications. Examples include content recommendation Zhang et al. [2023], Jin et al. [2023], robotic control Ahn et al. [2024], web browsing (Yao et al. [2022], Deng et al. [2023]), game player (Gong et al. [2023]), communicative agents (CAMEL Li et al. [2023]), and so on. For human behavior simulation, Park et al. [2023] explored generative agents for interactive simulacra of human behaviors in a social system Gao et al. [2023]. Aher et al. [2023] demonstrated the feasibility to replicate human subject studies with LLMs. Finally, Wang et al. [2023] presented an agent framework to simulate user behaviors with memory/actions.

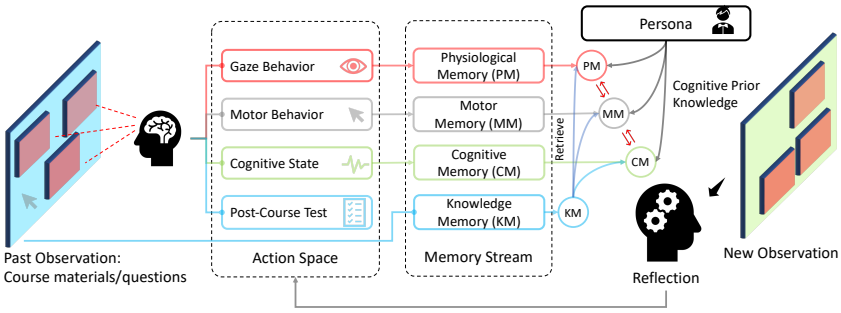

Figure 1: Our EduAgent framework.

In the education domain, although recent work has explored LLMs to provide feedback to students Kung et al. [2023], Matelsky et al. [2023], Cox [2023] and assist teachers Jeon and Lee [2023], there is limited research that uses LLM-powered agents to simulate learning behaviors. AgentVerse Chen et al. [2023] simulated classroom interactions on their open-sourced agent environments. CGMI Jinxin et al. [2023] simulated the speech interactions between students and teachers with different personas. Xu and Zhang [2023] leveraged LLMs to simulate cognitive states and learning performance. However, no prior work can simulate fine-grained student cognitive states and physiological behaviors. The prompts are usually designed to directly map an input state to learning outcome, abstracting out the correlations among diverse behaviors. By contrast, our EduAgent framework models finer-grained physiological and cognitive behaviors, and tackles the long lasting problem of creating realistic cognitive models in an online learning process by integrating cognitive prior knowledge. EduAgent captures the contextual learning history, students' personas and internal correlations among diverse learning behaviors, which are crucial for cognitive/learning science and AI-driven education.

3 Dataset

In this section, we elaborate on the **EduAgent310** and **EduAgent705** datasets.

The EduAgent310 is an augmented version from a recent dataset Xu et al. [2023] that contains raw behavior data of $N = 310$ students, where each student watched a slide-based lecture and answered test questions afterwards. The original dataset in Xu et al. [2023] only contains coarse-grained annotations of student behaviors corresponding to specific post-course questions.

In contrast, our EduAgent310 adds detailed annotation of gaze/motor behaviors, and cognitive states with precise timestamp. The timestamps are synchronized to the corresponding course materials and post-lecture test questions. Each lecture is a 5-min talk about one topic of machine learning and students have diverse educational backgrounds. Furthermore, for each slide, we annotate the potential Area of Interests (**AOIs**) in the format of bounding boxes. Each AOI corresponds to a text block, plot, figure illustration, etc. The gaze behavior is a time series of focal points on each slide, measured using the student's webcam and browser-based Webgazer model Xu et al. [2023]. Motor control behaviors refer to students' mouse moving activities while watching the online course. The time series of gaze and motor data samples are normalized into [0,1], to adapt to different screen size and then mapped into specific AOIs. The cognitive states include workload, curiosity, valid focus, course following, engagement, and confusion. More details about cognitive state measurement are available in Appendix. Each student watched one 5-min video lecture of 30-50 transcripts (sentences) and behaviors are annotated per second, so totally we obtained 100778 labelled samples in EduAgent310 dataset.

To increase the size and diversity of the dataset, we create a new dataset (EduAgent705) composed of $N = 705$ virtual students. The virtual students are simulated by our EduAgent framework and verified through experiments (Section 6). Before elaborating on the framework design, we first introduce the dataset itself. Inspired by Seidel et al. [2021], Nakayama et al. [2021] that shows the effect of the students' personas on learning performance, we consider the following *personas* when simulating virtual students: learning attitude, exam performance, focus, curiosity, interest in course, prior course knowledge, compliance in course, smartness, and family. Each characteristic has one

Table 2: Micro benchmark to show results in the first experiment. $\surd$: with cognitive prior knowledge, $\times$: no cognitive prior knowledge. **Cop.** means components used in the ablation study. **All**: all components are used, $\times$**M**: motor behaviors are removed in the memory, $\times$**P**: gaze behaviors are removed in the memory, $\times$**C**: cognitive states are removed in the memory, and $\times$**D**: the whole few-shot memory as example demonstrations are removed. For metrics, **Ga.**: gaze AOI distance, **Mo.**: motor AOI distance, **Wo.**: workload MAE, **Cu.**: curiosity MAE, **Foc.**: valid focus MAE, **Fol.**: course following MAE, **Eng.**: engagement MAE, **Co.**: confusion MAE, **CH.**: choice similarity, **Ac.**: accuracy similarity. Blue color means the configuration in the current row could achieve **better** simulation performance compared with the first row (GPT3.5 with cognitive prior knowledge with all components in the framework) in the specific metric while red color means the configuration in the current row results in **worse** performance compared with the first row.

MODEL	PRI.	COP.	GA.	MO.	WO.	CU.	FOC.	FOL.	ENG.	CO.	CH.	AC.
GPT3.5	$\surd$	ALL	0.35	0.34	0.17	0.23	0.25	0.35	0.11	0.07	0.61	0.66
GPT3.5	$\times$	ALL	0.35	0.34	0.25	0.29	0.27	0.39	0.18	0.11	0.60	0.65
GPT3.5	$\surd$	$\times$M	0.35	0.35	0.17	0.22	0.25	0.34	0.10	0.06	0.60	0.65
GPT3.5	$\surd$	$\times$P	0.37	0.35	0.18	0.23	0.25	0.34	0.12	0.07	0.60	0.65
GPT3.5	$\surd$	$\times$C	0.35	0.34	0.17	0.48	0.27	0.56	0.26	0.19	0.60	0.65
GPT3.5	$\surd$	$\times$D	0.36	0.34	0.17	0.50	0.27	0.55	0.28	0.19	0.56	0.63
GPT4	$\surd$	ALL	0.35	0.32	0.20	0.40	0.26	0.55	0.12	0.15	0.66	0.68
GEMINI	$\surd$	ALL	0.37	0.34	0.21	0.26	0.23	0.43	0.03	0.02	0.57	0.60
LLAMA2	$\surd$	ALL	0.36	0.32	0.28	0.32	0.27	0.34	0.04	0.14	0.39	0.52

positive item and one negative item, listed in Appendix Table. 3. We further include demographic information such as age, major and education levels, listed in Appendix Table. 3. Totally, there are $4 \times 3 \times 6 \times 4 \times 2^9 = 147456$ possible combinations of personas. Before running the EduAgent simulation, we instantiate the framework using one randomly selected persona. Our current simulation has created 705 virtual students, each going through one lecture session and one post-lecture test. But the framework itself is capable of generating unlimited amount of samples.

4 EduAgent Framework

4.1 Our Approach

The simulation pipeline is depicted in Fig. 1. Specifically, we first instantiate each student agent with one persona profile as mentioned before. Then we simulate the agent's learning process from the first to the last slide of the lecture in order. Each slide is one simulation step, where the agent receives the transcripts within this slide and outputs a trajectory of simulated learning behaviors (actions) for each transcript (each transcript is one sentence). The actions include the student's gaze behaviors, motor control behaviors to move computer mouse, and cognitive states (workload, curiosity, valid focus, course following, engagement, confusion), as well as answers of corresponding test questions related to the specific slide. Before making actions, agents first reflect the correlation among personas, past actions, and past course materials from the memory (demonstrations), guided by the integrated cognitive priors (depicted below). The demonstrations give the agent its past gaze/motor behaviors, cognitive states and past course materials in the time series format so that it can reason how different behaviors affect each other (reflection outcomes). Note that we only use past behaviors of the most recent past slide (not all past slides) for few-shot demonstration. After that, agents apply the reflection outcomes on new input course materials or test questions to make actions.

A key design principle of EduAgent is to incorporate cognitive priors Bourgin et al. [2019], which helps guide the LLMs to **first reason correlations** among learning behaviors and **then make simulations**. At a high level, we glean theoretical findings of student learning behaviors in cognitive science (e.g., correlations among cognitive states and learning performance), and embed such prior knowledge in the prompts, so that the LLM can stay grounded and get clear guidance regarding where to start reasoning. Instead of giving an explicit statement of the cognitive priors, we allow the LLM to reason by itself, regarding how and why behaviors in each memory layer are modulated by personas (to capture individual differences) and course contents (to model learning process). We

also prompt the LLM to reason how different behaviors affect each other, preventing it from directly copying few-shot history demonstration and mitigating the overfitting problem.

Gaze/Motor Behavior Simulation: It is difficult for the LLM to directly interpret raw gaze/motor sensor data, because they are usually noisy and massive, exceeding token limitation and lacking contextual information. Theoretical studies of online education Massaro et al. [2012] have established the correlation between gaze/motor behavior and the lecture content. Such correlation effects are multifaceted, vary over time, and do not admit any closed-form model. However, Mayer et al. advocate that effective online learning occurs when a student selects relevant elements, organizes the elements to form coherent mental representations, and integrates the new and existing representations Mayer [2009]. This process necessarily involves the interaction between gaze/motor following behaviors and the semantics within the course content, where LLM excels at. We thus propose to map gaze/motor coordinates into specific AOIs on slides, so that LLMs can correlate the semantic information to sensory behaviors. Instead of asking the agent to output the raw gaze/motor data, we prompt it to output the gaze/motor AOI ID on each slide. Gaze/motor changes across the AOIs in turn serve as the action for our gaze/motor simulation. During a reflection, we prompt the agent to first reason, based on its memory, regarding how these factors modulate gaze/motor behaviors. The agent then leverages the reasoning outcomes to perform the gaze/motor simulation according to new course materials.

Cognitive States Simulation: For each specific transcript, the agent generates a numeric value ranging from 0 to 1 to indicate the level of each cognitive state factor. Furnham et al. [2003] revealed that student cognitive states are not only modulated by course materials themselves, but also affected by students' own personas. For example, students who do not have strong academic background may have higher workload in course. Moreover, D'Mello et al. [2012] showed student gaze/motor behaviors can be indicators of cognitive states during learning. These cognitive priors inspire us to prompt the agent to first reason how its persona and past course contents modulate its past cognitive states and how past gaze/motor behaviors can indicate cognitive states from demonstrations in memory. The agent then applies such reasoning outcomes on current course materials and simulated gaze/motor behaviors to estimate the modulated cognitive states for output.

Learning Performance Simulation: For each post-lecture test question, the agent makes one selection from four choices. The goal is to mimic the question answering of each individual student instead of directly choosing the correct answer. Using two longitudinal university samples, Chamorro-Premuzic and Furnham [2003] revealed that personality can serve as important predictors of student academic performance. Moreover, Zhu et al. [2023] showed that gaze/motor behaviors are strongly correlated with student activities in e-learning to infer learning performance. Lei et al. [2018] further revealed that student engagement is apparently correlated with student learning performance. These cognitive priors inspire us to prompt the agent to reason how its persona and simulated cognitive states and gaze/motor behaviors could affect its question answer correctness according to course materials. If the agent reasons that the student is likely to choose the correct answer, it finds the correct answer based on transcripts. Otherwise, it should estimate the most likely but incorrect choice according to gaze/motor trajectories across the transcripts.

5 Experiment 1: Personalized Student Behavior Prediction

5.1 Task Settings and Metrics

We first evaluate EduAgent's ability to predict future student learning behaviors and outcomes, using students' past learning behavior as few-shot demonstration. In our experiment, we simulate 310 student agents, corresponding to the $N = 310$ dataset (**EduAgent310**). Each agent experiences 30-50 transcripts and takes actions per transcript. As depicted in Section. 4, we simulate students per slide and give corresponding course materials and post questions as input for student agents. After each slide, we log the real students' behaviors from previous slides into the agents' memory, to personalize their responses for future slide simulation.

To compare behaviors between agents and corresponding real students (ground truth), we use normalized **AOI distance** on screen for gaze/motor behaviors, **Mean Absolute Error** (MAE) for six cognitive states (normalized to [0,1]), **choice similarity** (whether both choices are similar) and **accuracy similarity** (whether both accuracy is similar) for question answering simulation. More details about metric design are depicted in Appendix.

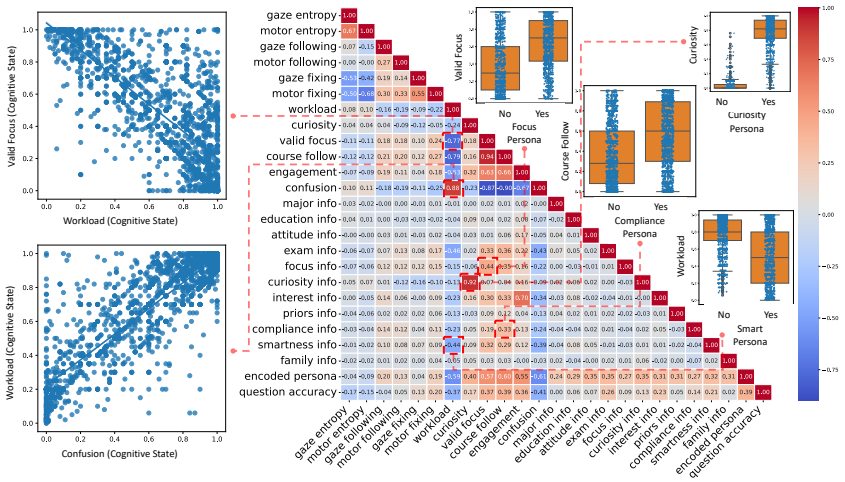

Figure 2: Correlation matrix heatmap and example relationships of generated behaviors in the second experiment.

5.2 Results and Analysis

For all language model generations, we set temperature to be 0 for more deterministic results.

Importance of Cognitive Prior Knowledge: We first verify whether cognitive prior knowledge can improve simulation performance using OpenAI GPT-3.5 modelOpe. As depicted in Section. 1, it is hard to directly apply recent advances of prompting techniques, such as Chain of Thought (CoT) Wei et al. [2022] into our problem. Therefore, we use standard prompt to serve as baseline that directly asks LLMs to give the output (actions) based on course contents, questions and memory. The results in Table. 2 show that integrating cognitive priors improves simulation performance including gaze behaviors, cognitive states and question answering. For motor behaviors, standard prompt is slightly better (AOI Distance = 0.336) compared with integrating cognitive priors (AOI Distance = 0.340). One potential reason is that, unlike gaze behaviors that indicate explicit student focus, mouse moving behaviors show weak correlations with other behaviors. Therefore integrating cognitive priors may not significantly enhance them. However, cognitive priors significantly improve simulation performance on cognitive states. These results indicate that, by incorporating cognitive prior knowledge to give clear guidelines for agents to reason from memory, agents can better capture potential correlations among diverse behaviors, and therefore further improve simulation. This also provides insights for future agent framework design (not limited to student agents but also other generative agents) that incorporating prior knowledge about correlations among actions and observations may boost agents' performance.

Importance of Different Components: We also run ablation studies to compare the importance of different components in our EduAgent framework. Specifically, we remove specific behavior data from memory and compare the performance difference. As depicted in Table. 2, we find that all behavior simulation performance drops when we remove past cognitive states from memory, indicating that cognitive states are highly correlated with other behaviors and therefore play an important role in our framework to provide contextual information for simulation in all behaviors. We also find that gaze/motor and question answering simulation performance drops more heavily when removing past gaze behaviors from memory compared with removing cognitive states. These results indicate that student gaze/motor and learning performance are more correlated with past gaze behaviors compared with past cognitive states, demonstrating the **importance of incorporating physiological behaviors** for student simulation, which is also a distinguishing feature of **our datasets** compared with other existing datasets. We also find that removing past motor behaviors from memory results in slightly better performance in four cognitive states simulation. This echoes the previous explanation comparing with standard prompts showing that mouse moving behaviors do not have that strong correlations with other behaviors. We also explore the effect when removing the whole few-shot memory as example demonstration and we find that the simulation performance

drops significantly for all behaviors, indicating that a student's historical data (the few-shot example demonstration) plays an important role in personalizing the student agent for behavior simulation.

Impact of Foundation Models: We also compare performance with different foundation models. As depicted in Table. 2, with Gemini Pro Gem and Llama 2 70B lla, we find slight improvement in motor and cognitive states simulation but obvious performance drop in gaze and question answering simulation compared with GPT-3.5, indicating the stronger ability of GPT-3.5 to capture correlation between student learning behaviors and learning performance. Moreover, the obvious improvement in gaze/motor and question answering simulation with GPT-4 Ope suggests its stronger ability than GPT-3.5 to capture such correlation.

6 Experiment 2: Virtual Generative Student Simulation

6.1 Task Settings and Metrics

Our second experiment tests whether EduAgent exhibits reasonable learning behavioral patterns without real demonstration data. The experiment is conducted using the EduAgent705 dataset, based on the optimal configuration we have identified in the first experiment (i.e., GPT-4 with cognitive priors with all components). Since there is no real student data, the agents directly use their own generated past behaviors as memory.

For virtual student datasets, we do not have ground truth for comparison. Therefore, we use Pearson coefficients to measure correlation among personas (demographics and characteristics) and behaviors, By doing so, we can verify whether such simulated behaviors echo related findings in cognitive science. To facilitate the Pearson coefficient calculation, we encode student specific personas into numeric values, and use entropy/following/fixing to represent different aspects of gaze/motor patterns (details in Appendix).

6.2 Results and Analysis

Detailed results of the 705 simulated agents are depicted in Fig. 2. Overall, most behaviors are consistent with well established principles in cognitive and learning science. We elabrate on a few representatives below (Pearson coefficients denoted by r).

Persona v.s. Gaze/Motor: We find that student agents with high GPA in exam have better gaze/motor following (gaze: $r = 0.07$, motor: $r = 0.13$) and fixing behaviors (gaze: $r = 0.08$, motor: $r = 0.17$) but low entropy (gaze: $r = -0.06$, motor: $r = -0.07$). Similar correlation can be found between focused/compliance persona and gaze/motor behaviors. By contrast, agents who have curious persona have larger gaze/motor entropy (gaze: $r = 0.05$, motor: $r = 0.07$) but smaller motor following ($r = -0.12$) and gaze/motor fixing (gaze: $r = -0.16$, motor: $r = -0.10$). Such results echo Kosel et al. [2021] about the correlation between gaze and student characteristics. Specifically, more compliant/attentive students have lower gaze entropy and better following/fixing since they are more focused in class. The results also verifies the effectiveness of the EduAgent framework in incorporating cognitive priors as well as knowledge of the students' personas.

Persona v.s. Cognitive States: We find that agents with high learning curiosity personas are also highly curious in cognitive states($r = 0.92$). Smart personas correspond to low course workload ($r = -0.44$) and attentive personas lead to better focus ($r = 0.44$). Finally, we find a positive correlation between agents GPA information and their valid focus ($r = 0.33$), course following ($r = 0.36$), and engagement ($r = 0.22$), but negative correlation between GPA and workload ($r = -0.46$), confusion($r = -0.43$). Such results verify the effectiveness of the EduAgent simulation, which echoes prior research Lau and Roeser [2002] in correlating learners' characteristics and cognitive states.

Persona v.s. Question Answering: We find correlation between answer accuracy and persona ($r = 0.39$), in education level ($r = 0.06$), attitude ($r = 0.07$), GPA ($r = 0.26$), focus ($r = 0.09$), curiosity ($r = 0.13$), interest ($r = 0.23$), prior knowledge ($r = 0.05$), compliance ($r = 0.14$), and smartness ($r = 0.21$). The results echo Karemera et al. [2003] in correlation between characteristics and academic performance and reflect cognitive prior knowledge design to consider personas impact on question answering.

Cognitive States v.s. Gaze/Motor: We find engagement is positively correlated with following (gaze: $r = 0.19$, motor: $r = 0.11$) and fixing behaviors (gaze: $r = 0.04$, motor: $r = 0.18$) but negatively correlated with entropy (gaze: $r = -0.07$, motor: $r = -0.09$). Similar relationships are also found between valid focus/course following and gaze/motor behaviors. By contrast, we can also find opposite relationships between workload/curiosity and gaze/motor behaviors. These results are aligned with theories in cognitive science D'Mello et al. [2012], and are **promising because** gaze/motor patterns are calculated from **actions of gaze/motor AOI** by comparing with course pace AOI while curiosity/workload/etc are obtained from **actions of cognitive states**. Such consistency suggests that our EduAgent successfully establishes the mapping between gaze/motor behaviors and cognitive states.

Gaze v.s. Motor: We find consistency between gaze and motor behaviors in most cases. Specifically, we find positive correlation between gaze and motor in entropy ($r = 0.67$), following ($r = 0.27$), and fixing ($r = 0.55$) patterns. Moreover, they do not fully overlap since gaze behaviors represent explicit watching focus while motor behaviors represent implicit focus by mouse moving Guo and Agichtein [2010]. The results echo Zhang et al. [2020], Zhu et al. [2023] about synchronization between students gaze and motor behaviors because mouse movement driven by cognitive states like curiosity or engagement could be indicated from gaze patterns Kwok et al. [2018].

Gaze/Motor v.s. Question Answering: We find that answer accuracy is negatively correlated with gaze ($r = -0.17$) and motor entropy ($r = -0.15$) but positively correlated with gaze ($r = 0.13$) and motor fixing ($r = 0.20$). The results reflect our cognitive prior knowledge design to consider such correlation and echo Zhu et al. [2023] since gaze/motor behaviors indicate cognitive states and therefore reveal learning success D'Mello et al. [2012]. However, we also find weak yet opposite effect of gaze ($r = -0.04$) and motor following ($r = 0.05$). This reflects the relationship between gaze and motor behaviors discussed above, i.e. gaze and motor behaviors do not exactly overlap although they have closed correlation Guo and Agichtein [2010].

Cognitive States v.s. Question Answering: We find that answer accuracy is negatively correlated with workload ($r = -0.37$) and confusion ($r = -0.41$), and positively correlated with curiosity ($r = 0.17$), valid focus ($r = 0.37$), course following ($r = 0.39$), engagement ($r = 0.36$). Such results echo Lei et al. [2018] about the correlations between cognitive states and academic performance and reflect our cognitive prior knowledge design to consider interactions between cognitive states and question answering. These results are **promising because** question accuracy is calculated from **actions of agent choices** by comparing with correct answers while curiosity/workload/etc are obtained from **actions of cognitive states**. Therefore, such consistency suggest that agents successfully map the correlation between cognitive states and learning success.

7 Limitations and Discussion

Student behavioral simulation is challenging as human behaviors are dynamic and come with noise in natureCziko [1989]. For example, it is easy for LLMs to answer questions correctly according to course materials. But it is hard to predict the same wrong choice of students. Moreover, except for precisely predicting student behaviors, we suggest that generating realistic learning behaviors like the second experiment is also one important research problem and may have broader impacts such as providing realistic course feedback to improve teaching strategy/course quality and facilitating hypotheses exploration in educational research. However, unlike directly asking LLMs to mimic specific personas or behaviors in individual cases, our problem mixes all student behaviors, personas and course contents together. Therefore, it is challenging for LLMs to cover all potential correlations while taking such massive information as input. Hence, there are also some inconsistent correlations as depicted in the second experiment. Future work could explore how to further improve the simulation performance.

8 Conclusion

We propose a novel generative agent framework (EduAgent) to simulate fine-grained and comprehensive student learning behaviors in online education. Two datasets are contributed to facilitate generative student agent research. Our experiments show that LLMs could not only predict student learning behaviors according to personalized history, but also generate realistic learning behavioral

patterns without real data. These results suggest a promising new line of research in student learning behavioral modelling and generative student agents. We believe that our work could serve as important groundwork and provide new insights in both student simulation and educational research.

References

Gemini pro. `https://ai.google.dev/tutorials/python_quickstart`. Accessed: January 24, 2024.

Openai. `https://platform.openai.com/docs/models`. Accessed: January 24, 2024.

Llama 2 70b. `https://huggingface.co/meta-llama/Llama-2-70b-chat-hf`. Accessed: January 24, 2024.

Statics2011 dataset. `https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507`. Accessed: January 24, 2024.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Sean Kirmani, Isabel Leal, Edward Lee, Sergey Levine, Yao Lu, Isabel Leal, Sharath Maddineni, Kanishka Rao, Dorsa Sadigh, Pannag Sanketi, Pierre Sermanet, Quan Vuong, Stefan Welker, Fei Xia, Ted Xiao, Peng Xu, Steve Xu, and Zhuo Xu. Autort: Embodied foundation models for large scale orchestration of robotic agents, 2024.

Stylianos Asteriadis, Paraskevi Tzouveli, Kostas Karpouzis, and Stefanos Kollias. Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment. *Multimedia Tools and Applications*, 41:469–493, 2009.

Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C Mitchell. How to train your learners: Reinforcement learning for the scheduling of online learning activities. 2020.

Joseph E Beck and Beverly Park Woolf. High-level student modeling with machine learning. In *International Conference on Intelligent Tutoring Systems*, pages 584–593. Springer, 2000.

Aditi Bhutoria. Personalized education and artificial intelligence in the united states, china, and india: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, 3:100068, 2022.

Melissa Bond, Katja Buntins, Svenja Bedenlier, Olaf Zawacki-Richter, and Michael Kerres. Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International journal of educational technology in higher education*, 17(1), 2020.

David D Bourgin, Joshua C Peterson, Daniel Reichman, Stuart J Russell, and Thomas L Griffiths. Cognitive model priors for predicting human decisions. In *International conference on machine learning*, pages 5133–5141. PMLR, 2019.

Jere E Brophy. *Teacher behavior and student achievement*. Number 73. Institute for Research on Teaching, Michigan State University, 1984.

Dien Thi Bui, Thuy Thi Nhan, Hue Thi Thu Dang, and Trang Thi Thu Phung. Online learning experiences of secondary school students during covid-19–dataset from vietnam. *Data in Brief*, 45: 108662, 2022.

Tomas Chamorro-Premuzic and Adrian Furnham. Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of research in personality*, 37(4): 319–338, 2003.

Penghe Chen, Yu Lu, Vincent W Zheng, and Yang Pian. Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 39–48. IEEE, 2018.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2023.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. Ednet: A large-scale hierarchical dataset in education. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 69–73. Springer, 2020.

Samuel Rhys Cox. The use of multiple conversational agent interlocutors in learning. *arXiv preprint arXiv:2312.16534*, 2023.

Scotty D. Craig, Arthur C. Graesser, Jeremiah Sullins, and Barry Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29, 2004.

Gary A Cziko. Unpredictability and indeterminism in human behavior: Arguments and implications for educational research. *Educational researcher*, 18(3):17–25, 1989.

Kevin Delgado, Juan Manuel Origgi, Tania Hasanpoor, Hao Yu, Danielle Allessio, Ivon Arroyo, William Lee, Margrit Betke, Beverly Woolf, and Sarah Adel Bargal. Student engagement dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3628–3636, 2021.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023.

Louis Deslauriers, Ellen Schelew, and Carl Wieman. Improved learning in a large-enrollment physics class. *Science*, 332(6031):862–864, 2011.

Carolina Diaz-Piedra, Hector Rieiro, Alberto Cherino, Luis J Fuentes, Andres Catena, and Leandro L Di Stasi. The effects of flight complexity on gaze entropy: An experimental study with fighter pilots. *Applied ergonomics*, 77:92–99, 2019.

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5):377–398, 2012.

Yang Fan. Scb-dataset: A dataset for detecting student classroom behavior. *arXiv preprint arXiv:2304.02488*, 2023.

Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 19:243–266, 2009.

Adrian Furnham, Tomas Chamorro-Premuzic, and Fiona McDougall. Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and individual Differences*, 14(1):47–64, 2003.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents, 2023.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, et al. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*, 2023.

Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17 (11):585–593, 2013.

Qi Guo and Eugene Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI'10 extended abstracts on human factors in computing systems*, pages 3601–3606. 2010.

Safiyeh Rajaee Harandi. Effects of e-learning on students' motivation. *Procedia-Social and Behavioral Sciences*, 181:423–430, 2015.

Raza Hasan, Sellappan Palaniappan, Salman Mahmood, Ali Abbas, and Kamal Uddin Sarker. Dataset of students' performance using student information system, moodle and the mobile application "edify". *Data*, 6(11):110, 2021.

Mushtaq Hussain, Wenhao Zhu, Wu Zhang, Syed Muhammad Raza Abidi, and Sadaqat Ali. Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, 52:381–407, 2019.

Jaeho Jeon and Seongyong Lee. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, pages 1–20, 2023.

Jiarui Jin, Xianyu Chen, Fanghua Ye, Mengyue Yang, Yue Feng, Weinan Zhang, Yong Yu, and Jun Wang. Lending interaction wings to recommender systems with conversational agents. *arXiv preprint arXiv:2310.04230*, 2023.

Shi Jinxin, Zhao Jiabao, Wang Yilei, Wu Xingjiao, Li Jiawen, and He Liang. Cgmi: Configurable general multi-agent interaction framework. *arXiv preprint arXiv:2308.12503*, 2023.

David Karemera, Lucy J Reuben, and Marion R Sillah. The effects of academic environment and background characteristics on student satisfaction and performance: The case of south carolina state university's school of business. *College Student Journal*, 37(2):298–309, 2003.

Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018.

Christian Kosel, Doris Holzberger, and Tina Seidel. Identifying expert and novice visual scanpath patterns and their relationship to assessing learning-relevant student characteristics. In *Frontiers in Education*, volume 5, page 612175. Frontiers Media SA, 2021.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4(1):1–8, 2017.

Cho Ki Kwok et al. Understanding user engagement level during tasks via facial responses, eye gaze and mouse movements. 2018.

Shun Lau and Robert W Roeser. Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment*, 8(2): 139–162, 2002.

Chia-An Lee, Jian-Wei Tzeng, Nen-Fu Huang, and Yu-Sheng Su. Prediction of student performance in massive open online courses using deep learning system based on learning behaviors. *Educational Technology & Society*, 24(3):130–146, 2021.

Hao Lei, Yunhuo Cui, and Wenye Zhou. Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, 46(3): 517–528, 2018.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *arXiv preprint arXiv:2305.17390*, 2023.

Min Liu, Emily McKelroy, Stephanie B Corliss, and Jamison Carrigan. Investigating the effect of an adaptive learning intervention on students' learning. *Educational technology research and development*, 65:1605–1625, 2017.

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115, 2019.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Tai Tan Mai, Marija Bezbradica, and Martin Crane. Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data. *Future Generation Computer Systems*, 127:42–55, 2022.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. GPTeach: Interactive TA Training with GPT Based Students. 2023.

Estefanía Martín, Manuel Gértrudix, Jaime Urquiza-Fuentes, and Pablo A Haya. Student activity and profile datasets from an online video-based collaborative learning experience. *British Journal of Educational Technology*, 46(5):993–998, 2015.

Davide Massaro, Federica Savazzi, Cinzia Di Dio, David Freedberg, Vittorio Gallese, Gabriella Gilli, and Antonella Marchetti. When art moves the eyes: a behavioral and eye-tracking study. *PloS one*, 7(5):e37285, 2012.

Jordan K Matelsky, Felipe Parodi, Tony Liu, Richard D Lange, and Konrad P Kording. A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439*, 2023.

Richard E. Mayer. *Multimedia Learning*. Cambridge University Press, 2nd edition, 2009. ISBN 0521514126.

Catherine McLoughlin. Inclusivity and alignment: Principles of pedagogy, task and assessment design for effective cross-cultural online learning. *Distance Education*, 22(1):7–29, 2001.

Sein Minn, Yi Yu, Michel C Desmarais, Feida Zhu, and Jill-Jenn Vie. Deep knowledge tracing and dynamic student classification for knowledge tracing. In *2018 IEEE International conference on data mining (ICDM)*, pages 1182–1187. IEEE, 2018.

Minoru Nakayama, Kouichi Mutsuura, and Hiroh Yamamoto. Impact of learner's characteristics and learning behaviour on learning performance during a fully online course. *Note taking activities in e-learning environments*, pages 15–36, 2021.

Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 117–124, 2013.

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

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

Chen Pojen, Hsieh Mingen, and Tsai Tzuyang. Junyi academy online learning activity dataset: A large-scale public online learning activity dataset from elementary to senior high school students. *Dataset available from https://www.kaggle.com/junyiacademy/learning-activity-public-dataset-by-junyi-academy*, 2020.

Lauren B Resnick. Toward a cognitive theory of instruction. In *Learning and motivation in the classroom*, pages 5–38. Routledge, 2017.

Nataniel Ruiz, Hao Yu, Danielle A Allessio, Mona Jalal, Ajjen Joshi, Tom Murray, John J Magee, Kevin Manuel Delgado, Vitaly Ablavsky, Stan Sclaroff, et al. Atl-bp: a student engagement dataset and model for affect transfer learning for behavior prediction. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022.

Tina Seidel, Katharina Schnitzler, Christian Kosel, Kathleen Stürmer, and Doris Holzberger. Student characteristics in the eyes of teachers: Differences between novice and expert teachers in judgment accuracy, observed behavioral cues, and gaze. *Educational Psychology Review*, 33:69–89, 2021.

John Stamper and Zachary A Pardos. The 2010 kdd cup competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics*, 3(2):312–316, 2016.

Bo Sun, Yong Wu, Kaijie Zhao, Jun He, Lejun Yu, Huanqing Yan, and Ao Luo. Student class behavior dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Computing and Applications*, 33:8335–8354, 2021.

Rohail Syed, Kevyn Collins-Thompson, Paul N Bennett, Mengqiu Teng, Shane Williams, Dr Wendy W Tay, and Shamsi Iqbal. Improving learning outcomes with gaze tracking and automatic question generation. In *Proceedings of The Web Conference 2020*, 2020.

Lori Uscher-Pines, Heather L Schwartz, Faruque Ahmed, Yenlik Zheteyeva, Erika Meza, Garrett Baker, and Amra Uzicanin. School practices to promote social distancing in k-12 schools: review of influenza pandemic policies and practices. *BMC public health*, 18(1):1–13, 2018.

Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Salem Alelyani, and Raheel Nawaz. Predicting academic performance of students from vle big data using deep learning models. *Computers in Human behavior*, 104:106189, 2020.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. When large language model based agent meets user behavior analysis: A novel user simulation paradigm, 2023.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. Results and insights from diagnostic questions: The neurips 2020 education challenge. In *NeurIPS 2020 Competition and Demonstration Track*, pages 191–205. PMLR, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Xiaolu Xiong, Siyuan Zhao, Eric G Van Inwegen, and Joseph E Beck. Going deeper with deep knowledge tracing. *International Educational Data Mining Society*, 2016.

Jie Xu, Kyeong Ho Moon, and Mihaela Van Der Schaar. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, 11(5):742–753, 2017.

Songlin Xu and Xinyu Zhang. Leveraging generative artificial intelligence to simulate student learning behavior. *arXiv preprint arXiv:2310.19206*, 2023.

Songlin Xu, Dongyin Hu, Ru Wang, and Xinyu Zhang. Peer attention enhances student learning. *arXiv e-prints*, pages arXiv–2312, 2023.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

An Zhang, Leheng Sheng, Yuxin Chen, Hao Li, Yang Deng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. *arXiv preprint arXiv:2310.10108*, 2023.

Zhaoli Zhang, Zhenhua Li, Hai Liu, Taihe Cao, and Sannyuya Liu. Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. *Journal of Educational Computing Research*, 58(1):63–86, 2020.

Jinjin Zhao, Weijie Xu, and Candace Thille. End-to-end question generation to assist formative assessment design for conceptual knowledge learning. 2021.

Rongrong Zhu, Liang Shi, Yunpeng Song, and ZhongMin Cai. Integrating gaze and mouse via joint cross-attention fusion net for students' activity recognition in e-learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–35, 2023.

9 Supplementary Material

9.1 EduAgent310 dataset

Here are the measurements we performed to get cognitive states. **Workload** is represented by gaze stationary entropy in specific duration according to Diaz-Piedra et al. [2019]. **Curiosity** is represented by gaze transition entropy according to Gottlieb et al. [2013]. For each second, **valid focus** is denoted as 1 if students' gaze falls into any AOIs on slides. Otherwise, it is 0. **Course following** is denoted as 1 if gaze falling AOI is the same as the AOI that the teacher is just talking about (course pacing AOI). Otherwise, it is 0. **Engagement** is denoted as 0 if the student face is not detected by the web camera. Otherwise, it is 1. **Confusion** is denoted as 1 if students click the mouse to report their confusion. Otherwise, it is 0. All cognitive states are first calculated within each second and then get averaged during specific transcripts so all states are continuous values.

Dataset distribution is depicted in Fig. 3 and Fig. 4.

9.2 EduAgent705 dataset

Dataset distribution is depicted in Fig. 5 and details of personas are depicted in Table. 3. The word cloud figure that contains all personas in the dataset is depicted in Fig. 6.

9.3 Additional information of experiment 1

Here we describe all metrics used in the first experiment in detail.

Gaze/Motor: As depicted in 4, the actions for gaze/motor are the simulated AOI ID on each slide. We compare the spatial **AOI distance** of the AOI center point location between agents and corresponding real students, serving as the metric. Note that all coordinates and AOI locations have been normalized into the range $[0, 1]$ by adapting different students' screen size. The reason why we use AOI distance instead of AOI accuracy (whether agent AOI and real student AOI are exactly the same) is that: First, closeby AOIs are acceptable even if they are not the same considering the potential errors caused by gaze tracking techniques. Additionally, we do not use Top-N accuracy because different slides have different number of AOIs (usually ranging from 5 to 12), making it not a general comparison across slides. Finally, our ultimate goal is still to simulate the focused location on slides where students are watching or moving the mouse so distance (continuous value) is a better metric to measure location difference compared with accuracy (categorical value).

Cognitive States: We use Mean Absolute Error (MAE) between the simulated agents' cognitive states (normalized to 1) and the ground truth as metrics.

Question Answering: We use *choice similarity* and *accuracy similarity* to quantify the answer choice difference and answer accuracy difference between agents and real students. Specifically, if a simulated agent and the real student make the same choice, then the choice similarity is 1 regardless of their choices are wrong or correct. Otherwise the choice similarity is 0. Whereas the accuracy similarity is 1 only when the agents' accuracy and real students' accuracy are the same by comparing with the correct question answers respectively. Otherwise the accuracy similarity is 0. Finally, we calculate the average results for both metrics.

Additional experiment results are depicted in Fig. 9, Fig. 8, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16.

9.4 Additional information of experiment 2

Here are the details of how we encode all personas and behaviors for evaluation.

For virtual student datasets, we do not have ground truth for comparison. Inspired by existing work Asteriadis et al. [2009], Brophy [1984] showing that student learning performance is affected by their personalities, we decide to use Pearson coefficients (similar with Harandi [2015]) to measure the correlation among personas (demographics and student characteristics) and all learning behaviors and outcomes. By doing so, we could measure whether the generated learning behaviors could echo related hypotheses and conclusions of existing student behavioral research to demonstrate the realism of generated behaviors. We examine the following specific aspects:

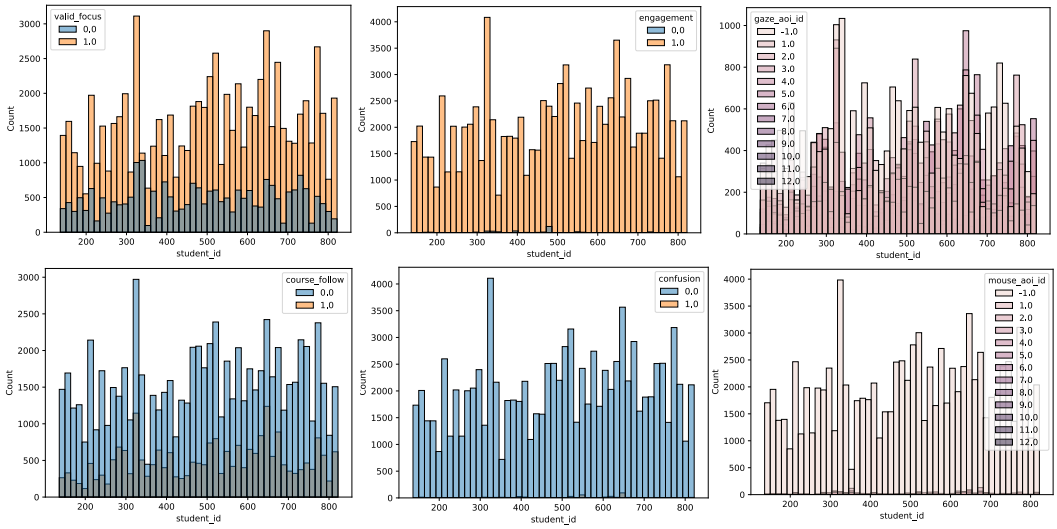

Figure 3: Data distribution in **EduAgent310**.

Persona: Each characteristic (from learning attitude to family) is either positive (denoted as 1) or negative (denoted as 0). In addition, major and education have several categories, which are normalized to 1. We also encode all characteristics and demographics into one aggregated persona measurement. Specifically, we first normalize each learning characteristic / demographic into the range from 0 to 1. Then we sum all characteristics and demographics, finally divided by the number of all characteristics and demographics, i.e. taking the average of them. The encoded overall persona is a continuous value from 0 to 1.

Gaze/Motor: We use the **entropy** of the gaze/motor AOI sequences to measure an agents' gaze/motor wandering behaviors. Moreover, we use **gaze/motor following** to measure whether an agent follows the pace of the lecture closely. For each transcript, gaze/motor following are set to be 1 if agents' gaze/motor AOIs are the same as AOIs of the teacher. Otherwise, they are 0. Additionally, we use **gaze/motor fixing** to measure the extent that agents keep their focus on specific AOIs across transcripts. Gaze/motor fixing are set to be 1 if current gaze/motor AOIs in the current transcript are the same as those in the previous one transcript. Otherwise, they are 0. We first calculate these measurements per transcript and then get average results of all transcripts per simulation step.

Cognitive States: We first get cognitive states (workload, curiosity, valid focus, course following, engagement, confusion) generated by agents per transcript and then get average results of all transcripts in one simulation step (slide).

Question Answering: By comparing agents' answers and correct answers, we calculate the average accuracy of all questions in specific simulation step (slide).

GPT4 v.s. Gemini: We also compare the correlation matrix generated by GPT4-powered student agents (Fig. 18) and Gemini-powered student agents (Fig. 17). As depicted in the two figures, GPT4 could achieve more realistic student behaviors than Gemini.

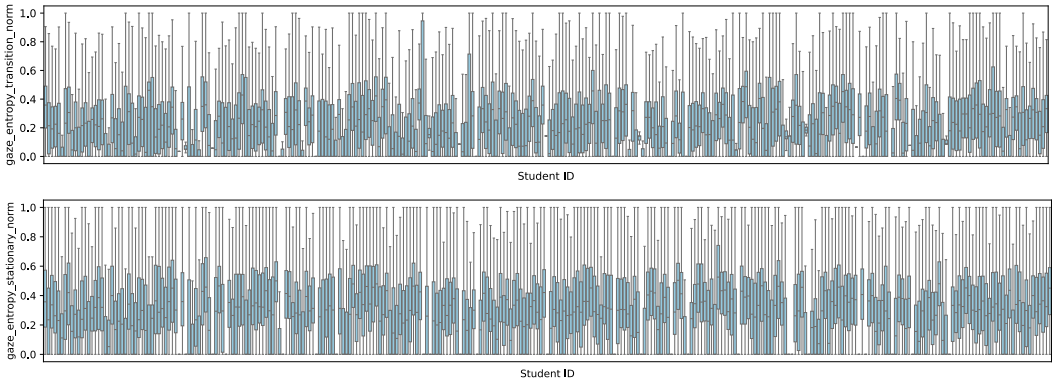

Figure 4: Distribution of gaze stationary entropy (used to represent workload) and transition entropy (used to represent curiosity) in **EduAgent310** dataset.

Table 3: Configurations of demographics and characteristics of virtual students

CATEGORY	ITEMS
AGE	0: 18-24, 1: 25-31, 2: 32-38, 3: $>$ 39
GENDER	0: FEMALE, 1: MALE, 2: OTHERS
MAJOR	0: HUMANITIES, 1: SOCIAL, 2: NATURAL, 3: TECHNOLOGY, 4: BUSINESS, 5: HEALTH
EDUCATION LEVEL	0: HIGH SCHOOL, 1: UNDERGRADUATE, 2: MASTER, 3: DOCTOR
LEARNING ATTITUDE	1: VERY MOTIVATED, 0: NOT MOTIVATED
EXAM PERFORMANCE	1: HIGH GPA, ANSWER TEST QUESTIONS CORRECTLY, 0: LOW GPA. MAKE MISTAKES IN POST-TEST
FOCUS	1: VERY FOCUS, 0: USUALLY ABSENT-MINDED
CURIOSITY	1: CURIOUS TO EXPLORE EVERYTHING IN THE COURSE, 0: NOT CURIOUS AT ALL
INTEREST IN COURSE	1: SUPER INTERESTED, 0: NOT INTERESTED AT ALL
PRIOR KNOWLEDGE	1: STRONG BACKGROUND WITH PRIOR KNOWLEDGE, 0: NO BACKGROUND WITHOUT PRIORS
COMPLIANCE	1: WELL-BEHAVED TO FOLLOW TEACHERS, 0: UNWILLING TO FOLLOW TEACHERS
SMARTNESS	1: SMART TO UNDERSTAND EVERYTHING FAST, 0: NOT SMART, UNDERSTAND THINGS SLOWLY
FAMILY	1: PARENTS HAVE A STRONG ACADEMIC BACKGROUND, 0: PARENTS DO NOT CARE ABOUT EDUCATION

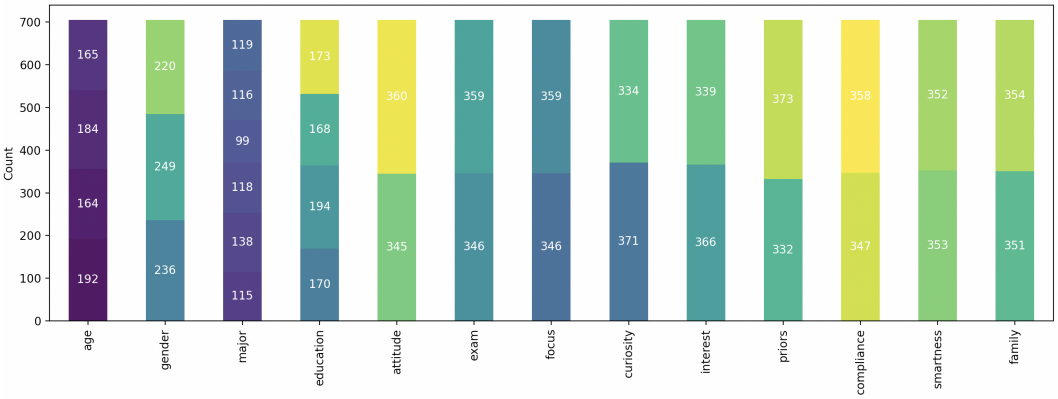

Figure 5: Distribution of each kind of persona in **EduAgent705** dataset.

Figure 6: Word cloud of personas of all agents in **EduAgent705** dataset.

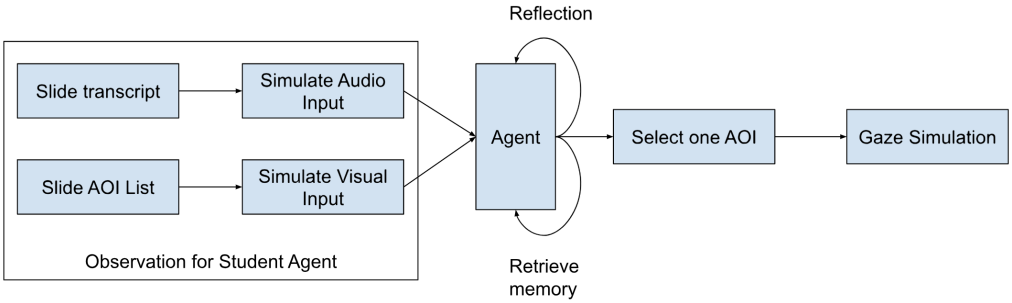

Figure 7: Illustration of our way to simulation gaze actions using AOIs.

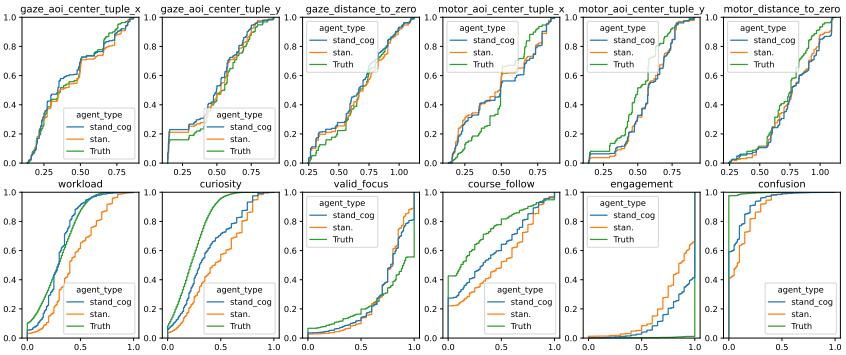

Figure 8: CDF (Cumulative Distribution Function) plots among all metrics by comparing standard prompt (stan.) with the prompt integrating cognitive prior knowledge (standard cog) in the first experiment.

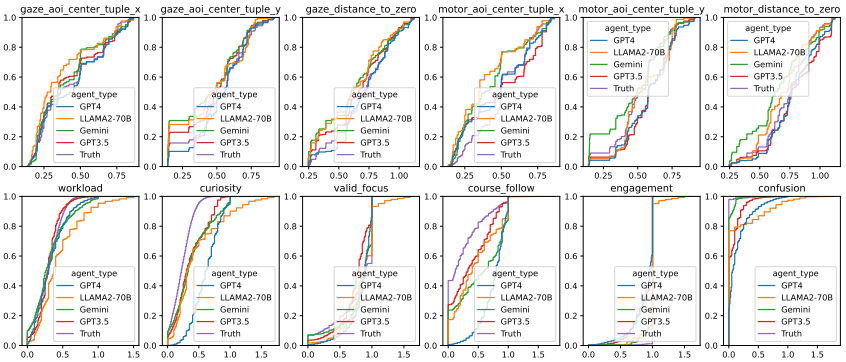

Figure 9: CDF (Cumulative Distribution Function) plots among all metrics compared with different foundation models in the first experiment.

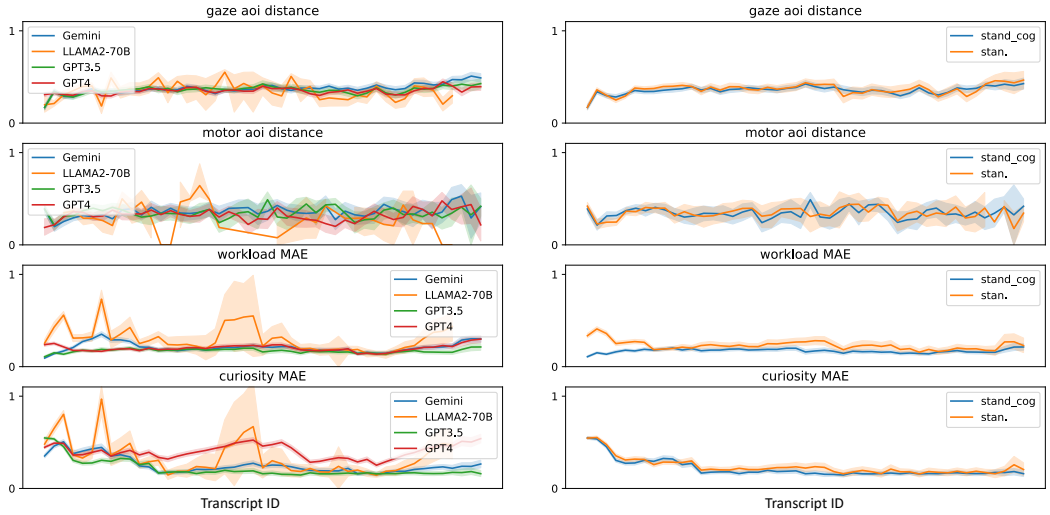

Figure 10: Simulation performance that changes with transcript ID by comparing different foundation models and by comparing standard prompt (stan.) with the prompt integrating cognitive prior knowledge (standard cog) in the first experiment.

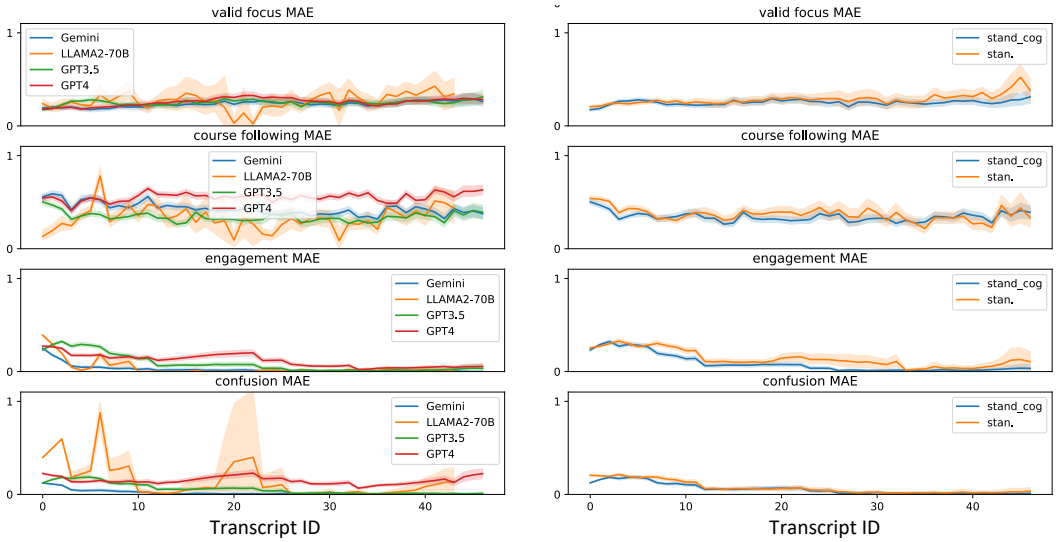

Figure 11: Simulation performance that changes with transcript ID by comparing different foundation models and by comparing standard prompt (stan.) with the prompt integrating cognitive prior knowledge (standard cog) in the first experiment.

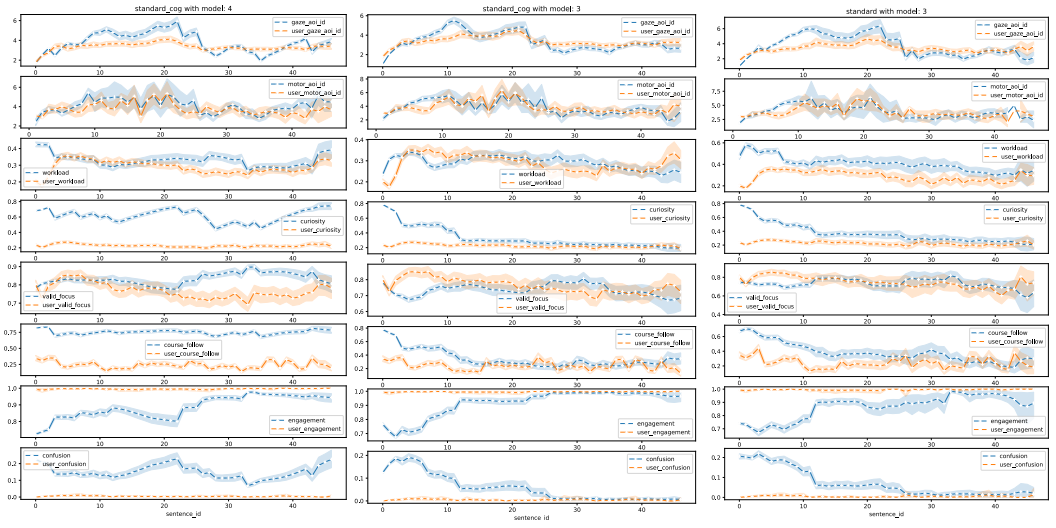

Figure 12: Simulation performance that changes with transcript ID (sentence ID in the figure) by comparing different foundation models and different prompts in the first experiment. Model 3 refers to GPT-3.5 and model 4 refers to GPT-4. Standard cog uses our cognitive priors. Blue curves are agents simulation behaviors and orange curves are real students behaviors as ground truth.

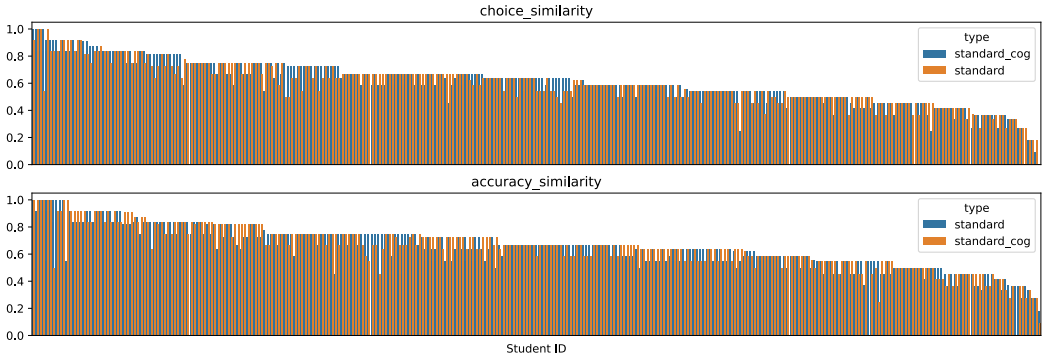

Figure 13: Simulation performance in question answering for different students with different prompts in the first experiment. Standard cog uses our cognitive priors.

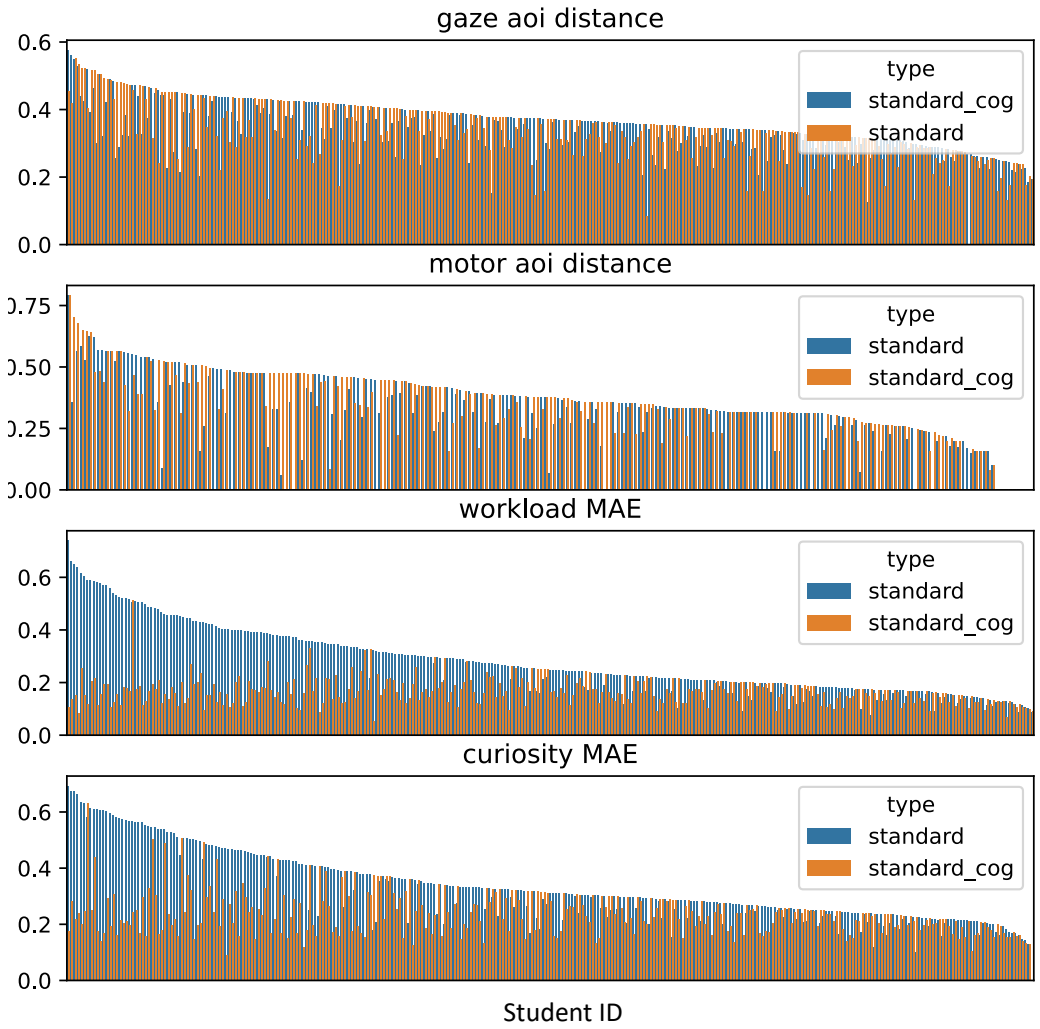

Figure 14: Simulation performance for different students by comparing different prompts in the first experiment. Standard cog uses our cognitive priors.

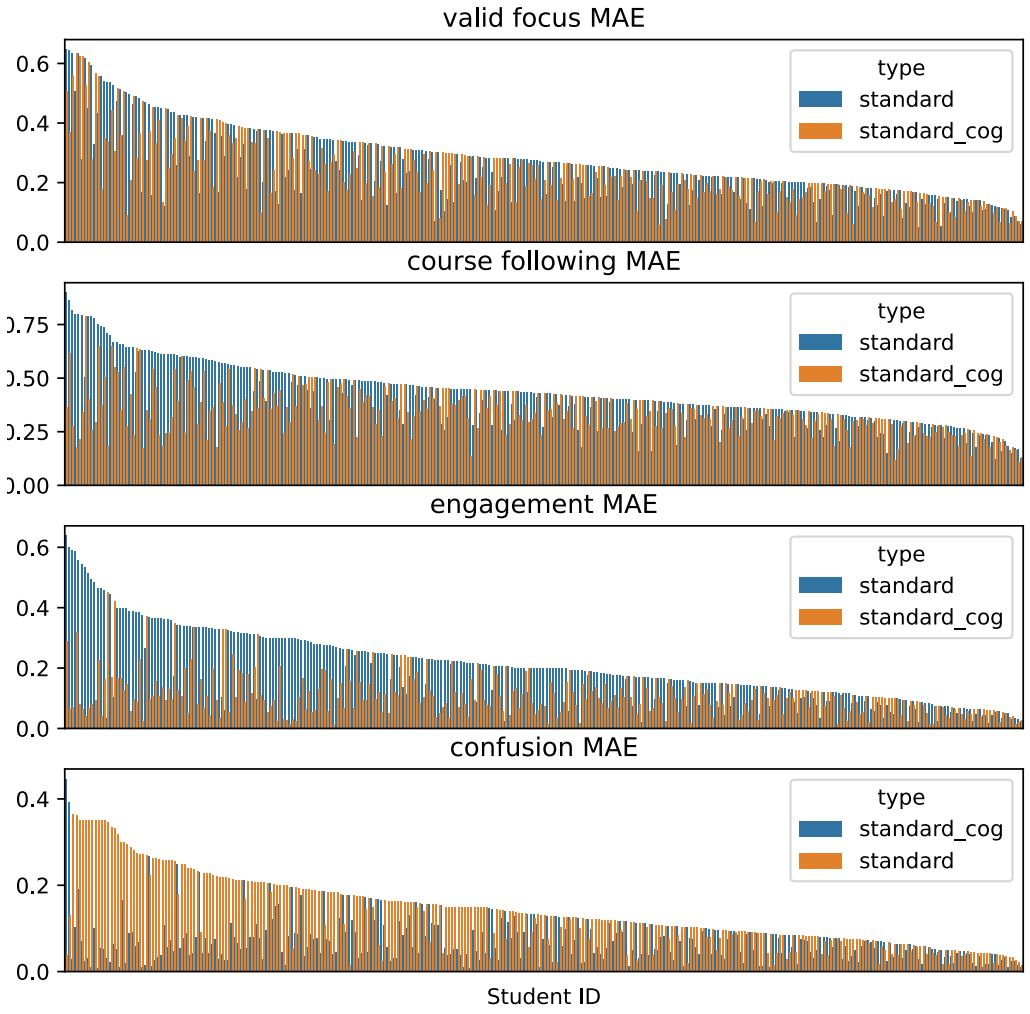

Figure 15: Simulation performance for different students by comparing different prompts in the first experiment. Standard cog uses our cognitive priors.

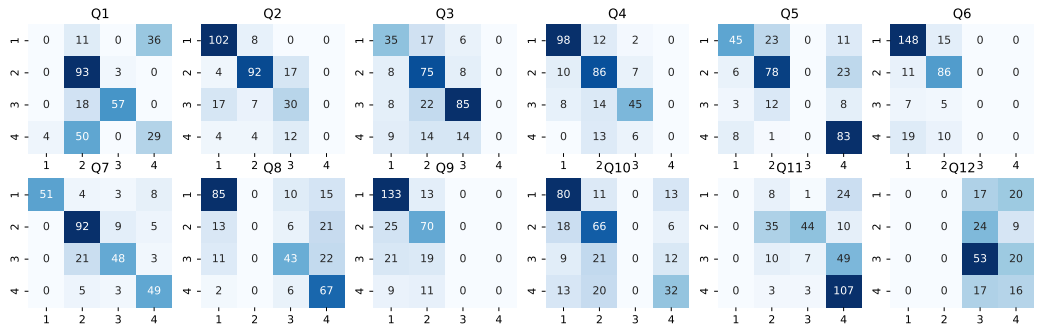

Figure 16: Confusion matrix in question answering simulation. Each matrix is one post-course question. X and Y axis represent agents' choices and corresponding real students' choices respectively. There are four choices per question. The confusion matrix shows that question answering simulation has good performance in some questions like Q7 but also bad performance in some questions like Q12.

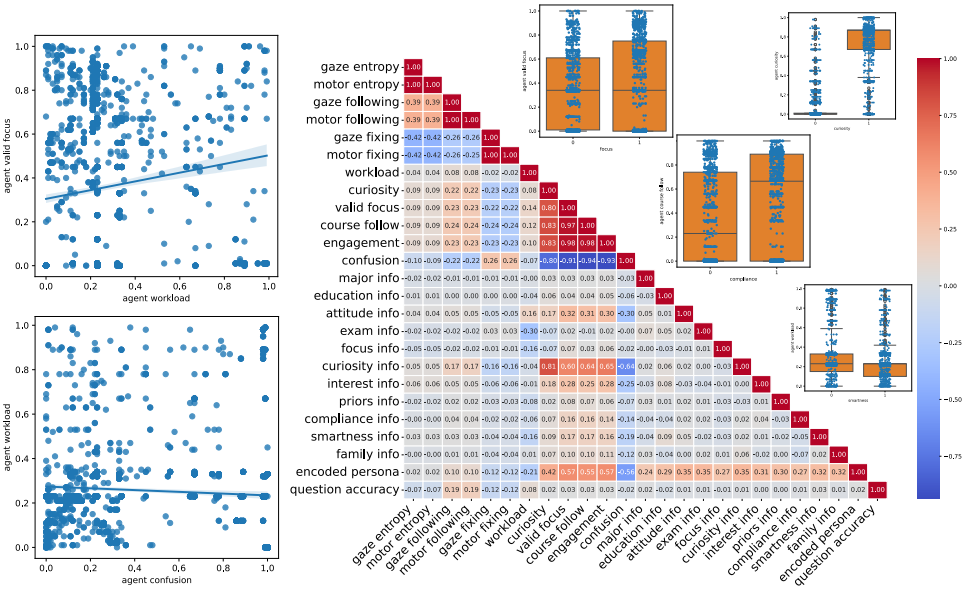

Figure 17: Heatmap of correlation matrix as well as examples of correlations of the second experiment using Gemini.

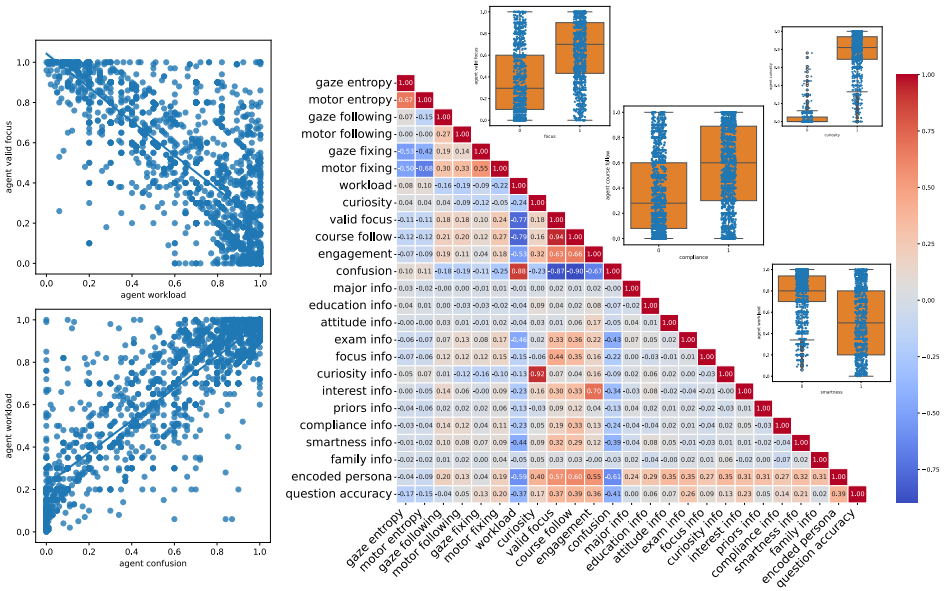

Figure 18: Heatmap of correlation matrix as well as examples of correlations of the second experiment using GPT 4.