Published: 26 April 2025

Citation in BibTeX format

CHI 2025: CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

RESEARCH-ARTICLE

TeachTune: Reviewing Pedagogical Agents Against Diverse Student Profiles with Simulated Students

HYOUNGWOOK JIN, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

MINJU YOO, Ewha Womans University, Seoul, Seoul, South Korea

JEONGEON PARK, University of California, San Diego, San Diego, CA, United States

YOKYUNG LEE, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

XU WANG, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

JUHO KIM, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

Open Access Support provided by:

Korea Advanced Institute of Science and Technology

University of Michigan, Ann Arbor

University of California, San Diego

Ewha Womans University

CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (April 2025)
https://doi.org/10.1145/3706598.3714054
ISBN: 9798400713941

TeachTune: Reviewing Pedagogical Agents Against Diverse Student Profiles with Simulated Students

Hyoungwook Jin
School of Computing
KAIST
Daejeon, Republic of Korea
jinhw@kaist.ac.kr

Minju Yoo
Ewha Womans University
Seoul, Republic of Korea
minjuu613@ewhain.net

Jeongeon Park
University of California San Diego
La Jolla, California, USA
jep034@ucsd.edu

Yokyung Lee
School of Computing
KAIST
Daejeon, Republic of Korea
ykleeee@kaist.ac.kr

Xu Wang
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, USA
xwanghci@umich.edu

Juho Kim
School of Computing
KAIST
Daejeon, Republic of Korea
juhokim@kaist.ac.kr

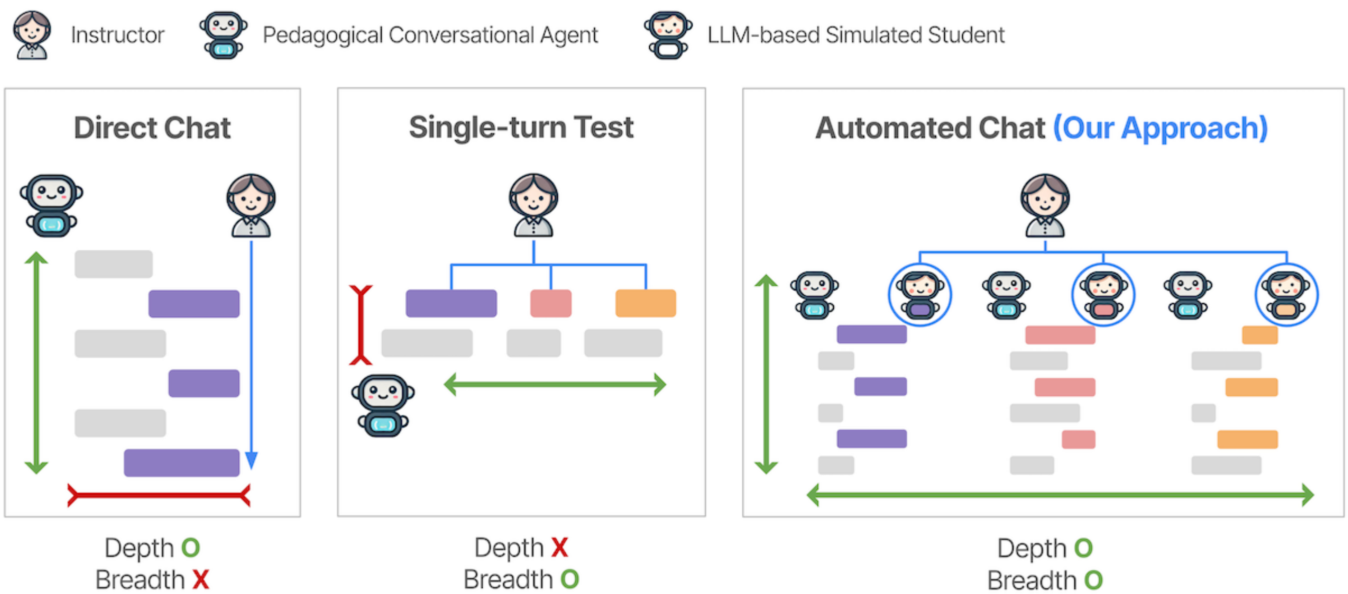


Figure 1: TeachTune is an evaluation tool that helps teachers review the interaction quality of pedagogical agents by utilizing simulated students. Direct chat supports an in-depth assessment but in a narrow scope. Single-turn tests with benchmark datasets support breadth exploration of pedagogical agents' adaptivity but lack depth in assessing conversations in multi-turn. TeachTune takes the best of both worlds by leveraging automated chat between the pedagogical agent and user-defined simulated students to help teachers review the adaptivity of pedagogical agents in sufficient depth and breadth.

Abstract

Large language models (LLMs) can empower teachers to build pedagogical conversational agents (PCAs) customized for their students. As students have different prior knowledge and motivation levels, teachers must review the adaptivity of their PCAs to diverse students. Existing chatbot reviewing methods (e.g., direct chat and benchmarks) are either manually intensive for multiple iterations or limited to testing only single-turn interactions. We present TeachTune, where teachers can create simulated students and review PCAs by observing automated chats between PCAs and simulated students. Our technical pipeline instructs an LLM-based student to simulate prescribed knowledge levels and traits, helping teachers explore diverse conversation patterns. Our pipeline could produce simulated students whose behaviors correlate highly to their input knowledge and motivation levels within 5% and 10% accuracy gaps. Thirty science teachers designed PCAs in a between-subjects study, and using TeachTune resulted in a lower task load and higher student profile coverage over a baseline.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714054

CCS Concepts

• **Human-centered computing → Interactive systems and tools**.

Keywords

LLM-assisted evaluation, Simulated students, Pedagogical conversational agents

ACM Reference Format:
Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. TeachTune: Reviewing Pedagogical Agents Against Diverse Student Profiles with Simulated Students. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3706598.3714054

1 Introduction

> "*A key challenge in developing and deploying Machine Learning (ML) systems is understanding their performance across a wide range of inputs.*" [97]

Large Language Models (LLMs) have empowered teachers to build Pedagogical Conversational Agents (PCAs) [94] with little programming expertise. **PCAs** refer to conversational agents that act as instructors [31], peers [61], and motivators [1] with whom learners can communicate through natural language, used in diverse subjects, grades, and pedagogies. Teacher-designed PCAs can better adapt to downstream class environments (i.e., students and curriculum) and allow teachers to experiment with diverse class activities that were previously prohibitive due to limited human resources. While conventional chatbots require authoring hard-coded conversational flows and responses [18, 23], LLM-based agents need only a description of how the agents should behave in natural language, known as prompting [24]. Prior research has proposed prompting techniques [9, 100, 102], user interfaces [2, 29, 60], and frameworks [4, 43] that make domain-specific and personalized agents even more accessible to build for end-users. With the lowered barrier and cost of making conversational agents, researchers have actively experimented with LLM-based PCAs under diverse pedagogical settings, such as 1-on-1 tutoring [32, 42, 103], peer learning [40, 80], and collaborative learning [54, 68, 96].

To disseminate these experimental PCAs to actual classes at scale, reviewing agents' content and interaction qualities is necessary before deployment. Many countries and schools are concerned about the potential harms of LLMs and hesitant about their use in classrooms, especially K-12, despite the benefits [41, 51]. LLM-based PCAs need robust validation against hallucination [70, 85], social biases [63, 92], and overreliance [27, 66]. Moreover, since students vary in their levels of knowledge and learning attitudes in a class [79], teachers must review how well their PCAs can cover diverse students in advance to help each student improve attitudes and learn better [8, 72, 73, 90]. For instance, teachers should check whether PCAs help not only poorly performing students fill knowledge gaps but also well-performing students build further knowledge through discussions. Regarding students' personalities, teachers should check if PCAs ask questions to prompt inactive students and compliment active students to keep them motivated. These attempts contribute to improving fairness in learning, closing the growth gap between students instead of widening it [64].

However, existing methods for reviewing the PCAs' coverage of various student profiles offer limited breadth and depth (Fig. 1). The current landscape of chatbot evaluation takes two approaches at large. First, teachers can directly chat with their PCAs and role-play themselves as students [10, 36, 74]. Although interactive chats allow teachers to review the behaviors of PCAs over multi-turn conversations in depth, it is time-consuming for teachers to manually write messages and re-run conversations after revising PCA designs, restraining the breadth of reviewing different students. Second, teachers can simultaneously author many input messages as test cases (e.g., benchmark datasets) and assess the PCAs' responses [11, 44, 78, 99, 104]. Single-turn test cases are scalable and reproducible, but teachers can examine only limited responses that do not capture multi-turn interactions (e.g., splitting explanations [49], asking follow-up questions [82]), restricting the depth of each review. Teachers may also need to create test cases manually if their PCAs target new curriculums and class activities.

To support efficient PCA reviewing with breadth and depth, we propose a novel review method in which teachers utilize auto-generated conversations between a PCA and simulated students. Recent research has found that LLMs can simulate human behaviors of diverse personalities [50, 71] and knowledge levels [40, 56]. We extend this idea to PCA review by simulating conversations between PCAs and students with LLM. We envision simulated conversations making PCA evaluation as reproducible and efficient as the test case approach while maintaining the benefit of reviewing multi-turn interactions like direct chat. Teachers can review the adaptivity of PCAs by configuring diverse simulated students as a unit of testing and examine the quality of interaction in depth by observing auto-generated conversations among them. We implemented this idea into TeachTune, a tool that allows teachers to design PCAs and review their robustness against diverse students and multi-turn scenarios through automated chats with simulated students. Teachers can configure simulated students by adding or removing knowledge components and adjusting the intensity of student traits, such as self-efficacy and motivation. Our LLM-prompting pipeline, Personalized Reflect-Respond, takes configurations on knowledge and trait intensity levels (5-point scale) as inputs and generates a comprehensive overview to instruct simulated students to generate believable responses.

To evaluate the performance of Personalized Reflect-Respond in simulating targeted student behaviors, we asked ten teachers to interact with nine simulated students of varying knowledge and trait levels in a blinded condition and to predict the simulated students' configuration levels for knowledge and traits. We measured the difference between teacher-predicted and initially configured levels. Our pipeline showed a 5% median error for knowledge components and a 10% median error for student traits, implying that our simulated students' behaviors closely align with the expectations of teachers who configure them. We also conducted a between-subjects study with 30 teachers to evaluate how TeachTune can help teachers efficiently review the interaction quality of PCAs in depth and breadth. Study participants created and reviewed PCAs for middle school science classes using TeachTune or a baseline system where PCA review was possible through only direct chats and single-turn test cases. We found that automated chats significantly help teachers explore a broader range of students within traits (large effect size, η^2=0.304) at a lower task load (η^2=0.395).

This paper makes the following contributions:

- Personalized Reflect-Respond, an LLM prompting pipeline that generates an overview of a target student's knowledge, motivation, and psychosocial context and follows the overview to simulate a believable student behavior.
- TeachTune, an interface for teachers to efficiently review the coverage of PCAs against diverse knowledge levels and student traits.
- Empirical findings showing that TeachTune can help teachers design PCAs at a lower task load and review more student profiles, compared to direct chats and test cases only.

2 Related Work

Our work aims to support the design and reviewing process of PCAs in diverse learning contexts. We outline the emergent challenges in designing conversational agents and how LLM-based simulation can tackle the problem.

2.1 Conversational Agent Design Process

Designing chatbots involves dedicated chatbot designers prototyping and then iteratively revising their designs through testing. Understanding and responding to a diverse range of potential user intents and needs is crucial to the chatbot's success. Popular methods include the Wizard-of-Oz approach to collect quality conversation data [46] and co-design workshops to receive direct feedback from multiple stakeholders [16, 26]. Involving humans to simulate conversations or collecting feedback can help chatbot designers understand human-chatbot collaborative workflow [23], explore diverse needs of users [13, 75], or iterate their chatbot to handle edge cases [18, 46]. Typical chatbot reviewing methods include conducting a usability study with a defined set of chatbots' social characteristics [14], directly chatting 1-on-1 with the designed chatbot [74], and testing with domain experts [36]. Such methods can yield quality evaluation but are costly as they need to be executed manually by humans. For more large-scale testing, designers can use existing test cases [10, 78] or construct new test sets with LLMs [99]. However, such evaluations happen in big chunks of single-turn conversations, which limits the depth of conversation dynamics throughout multiple turns. To complement the limitations, researchers have recently proposed leveraging LLMs as simulated users [25], role-players [28], and agent authoring assistant [12]. TeachTune explores a similar thread of work in the context of education by utilizing simulated students to aid teachers' breadth- and depth-wise reviewing of PCAs.

2.2 Simulating Human Behavior with LLMs

Recent advancements in LLM have led researchers to explore the capabilities of LLMs in simulating humans and their environments, such as simulating psychology experiments [21], individuals' beliefs and preferences [17, 20, 39, 67, 84], and social interactions [52, 71, 83, 91]. In education, existing works have simulated student behaviors for testing learning contents [35, 56, 68], predicting cognitive states of students [55, 101], facilitating interactive pedagogy [40], and assisting teaching abilities of instructors [59, 69, 77, 105]. In deciding which specific attribute to simulate, existing simulation work has utilized either knowledge states [37, 40, 56, 68] or cognitive traits, such as personalities and mindset [50, 59, 93]. However, simulating both knowledge states and personalities is necessary for authentic learning behaviors because cognitive traits, in addition to prior knowledge, are a strong indicator for predicting success in learning [3, 7, 15, 19, 79]. Liu et al. explored utilizing cognitive and noncognitive aspects, such as the student's language proficiency and the Big Five personality, to simulate students at binary levels (e.g., low vs. high openness) for testing intelligent tutoring systems [55]. Our work develops this idea further by presenting an LLM-powered pipeline that can configure and simulate both learners' knowledge and traits at a finer granularity (i.e., a five-point scale). Finer-grained control of student simulation will help teachers review PCAs against detailed student types, making their classes more inclusive.

3 Formative Interview and Design Goals

We conducted semi-structured interviews with five school teachers and observed how teachers review PCAs to investigate **RQ1**. More specifically, we aimed to gain a comprehensive understanding of what types of students teachers want PCAs to cover, what student traits (e.g., motivation level, stress) characterize those students, how teachers create student personas using those traits, and what challenges teachers have with existing PCA review methods (i.e., direct chat and test cases).

RQ1: What are teachers' needs in reviewing PCAs and challenges in using direct chats and test cases?

3.1 Interviewees

We recruited middle school science teachers through online teacher communities in Korea. We required teachers to possess either an education-related degree or at least one year of teaching experience. The teachers had diverse backgrounds (Table 1). The interview took place through Zoom for 1.5 hours, and interviewees were compensated KRW 50,000 (USD 38).

3.2 Procedure

We began the interview by presenting the research background, ChatGPT, and its various use cases (e.g., searching, brainstorming, and role-playing). We requested permission to record their voice and screen throughout the session and asked semi-structured interview questions during and after sessions.

Interviewees first identified the most critical student traits that PCAs should cover when supporting diverse students in K-12. To do so, we gave interviewees a list of 42 traits organized under five categories—personality traits, motivation factors, self-regulatory learning strategies, student approaches to learning, and psychosocial contextual influence [79]. Interviewees ranked the categories by importance of reviewing and chose the top three traits from each category.

Interviewees then assumed a situation where they created PCAs for their science class to help students review the phase transition between solid, liquid, and gas. Interviewees reviewed the interaction quality and adaptivity of a given tutor-role PCA by chatting with

Id	Period of teaching	Size of class	Familiarity		
			Chatbots	Chatbot design process	ChatGPT
I1	3 years	20 students	Unfamiliar	Very unfamiliar	Familiar
I2	6 years	20 students	Very familiar	Very familiar	Very familiar
I3	16 years	21 students	Unfamiliar	Very unfamiliar	Familiar
I4	2 years	200 students	Very familiar	Familiar	Very familiar
I5	1 year	90 students	Familiar	Familiar	Unfamiliar

Table 1: Demographic information of the interview participants. We recruited five participants with varying levels of teaching experience, chatbot familiarity, chatbot design process, and ChatGPT familiarity. "Very familiar" indicates the frequent usage, "Familiar" for a little usage, "Unfamiliar" for passing knowledge, and "Very unfamiliar" for no knowledge.

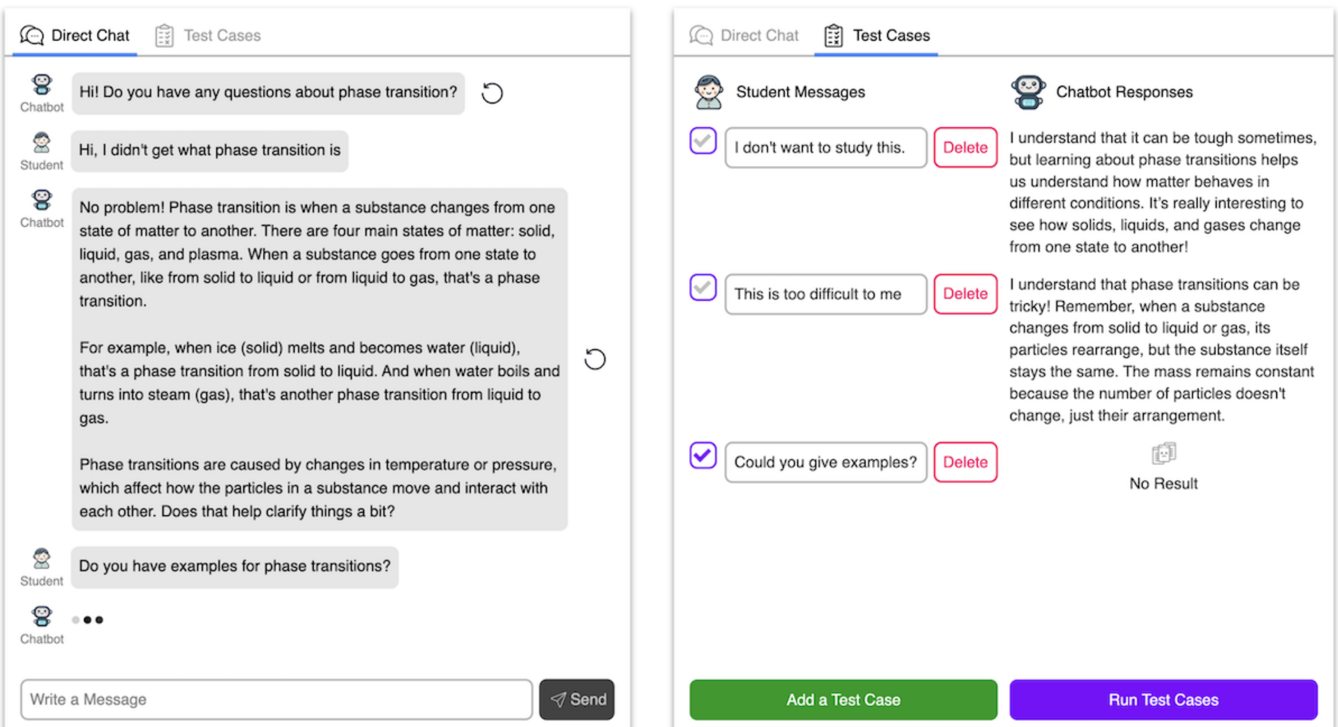


Figure 2: The interface used for the formative interview. On the left is the Direct Chat tab, where interviewees could converse with the chatbot as the student's role. Interviewees could roll back to previous messages by clicking the rewind button next to the chatbot's message. On the right is the Test Cases tab, where interviewees can add a set of student utterances and see chat responses.

it directly and authoring test case messages, playing the role of students. Interviewees could revisit the list of 42 traits for their review. Interviewees used the interfaces in Fig. 2 for 10 minutes each and were asked to find as many limitations of the PCA as possible. The PCA was a GPT-3.5-based agent with the following system prompt: *You are a middle school science teacher. You are having a conversation to help students understand what they learned in science class. Recently, students learned about phase transition. Help students if they have difficulty understanding phase transition.*

Subsequently, interviewees listed student profiles whose conversation with the PCA would help them review its quality and adaptivity. A student profile is distinguished from student traits as it is a combination of traits describing a student. Interviewees wrote student profiles in free form, using knowledge level and earlier 42 student traits to describe them (e.g., a student with average science grades but an introvert who prefers individual learning over cooperative learning).

3.3 Findings

3.3.1 Teachers deemed students' knowledge levels, motivation factors, and psychosocial contextual influences as important student traits to review. Interviewees thought that PCAs should support students with low motivation and knowledge, and hence, it is crucial

Category	Student Trait	Definition
Motivation factors	Academic self-efficacy	Self-beliefs of academic capability
	Intrinsic motivation	Inherent self-interest, and enjoyment of academic learning and tasks
Psychosocial contextual influence	Academic stress	Overwhelming negative emotionality resulting directly from academic stressors
	Goal commitment	Commitment to staying in school and obtaining a degree

Table 2: The top four student traits teachers found important for PCAs to cover.

to review how PCAs scaffold these students robustly. All five interviewees started their reviewing of the PCA with knowledge-related questions to assess the correctness and coverage of its knowledge. They then focused on how the PCA responds to a student with low motivation and interest (Table 2). Motivational factors (i.e., academic self-efficacy and intrinsic motivation) are important because students with low motivation often do not pay attention to class activities, and learning with a PCA would not work at all if the PCA cannot first encourage those students' participation (I1, I2, and I5). Interviewees also considered psychosocial factors (i.e., academic stress and goal commitment) important as they significantly affect the learning experience (I1). I3 remarked that she tried testing if the PCA could handle emotional questions because they take up most students' conversations.

3.3.2 Multi-turn conversations are crucial for review, but writing messages to converse with PCAs requires considerable effort and expertise. Follow-up questions and phased scaffolding are important pedagogical conversational patterns that appear over several message turns. Interviewees commented that it is critical to check how PCAs answer students' serial follow-up questions, use easier words across a conversation for struggling students, and remember conversational contexts because they affect learning and frequently happen in student-tutor conversations. Interviewees typically had 15 message turns for a comprehensive review of the PCA. Interviewees noted that these multi-turn interactions are not observable in single-turn test cases and found direct chat more informative. However, interviewees also remarked on the considerable workload of writing messages manually (I1), the difficulty of repeating conversations (I4), and the benefits of test cases over direct chats in terms of parallel reviewing (I2). I2 also commented that teachers would struggle to generate believable chats if they have less experience or teach humanities subjects whose content and patterns are diverse.

3.3.3 Teachers' mental model of review is based on student profiles, but they lack systematic approaches to organize and incorporate diverse types and granularities of student traits. Interviewees created test cases and conversational patterns with specific student personas in mind and referred to them when explaining their rationale for test cases. For example, I4 recalled students with borderline intellectual functioning and tested if the PCA could provide digestible explanations and diagrams. However, interviewees tend to review PCAs on the fly without a systematic approach; interviewees mix different student personas (e.g., high and low knowledge, shy and active) in a single conversation instead of simulating each persona in a separate chat. I4 and I5 remarked that they had not conceived the separation, and single-persona conversations would have made the review more meaningful. I2 commented that creating student profiles first would have prepared her to organize more structural test cases. Interviewees also commented on the difficulty of describing students with varying levels within a trait (I4) and reflecting diverse traits in free-form writing (I1).

3.4 Design Goals

Based on the findings from the formative interview, we outline the design goals to help teachers efficiently review their PCAs' limitations against diverse students and improve their PCAs iteratively. The design goals are 1-to-1 mapped to each finding in §3.3 and aim to address teachers' needs and challenges.

DG1. Support the reviewing of PCAs' adaptivity to students with varying knowledge levels, motivation factors, and psychosocial contexts.

DG2. Offload the manual effort to generate multi-turn conversations for quick and iterative reviews in the PCA design process.

DG3. Provide teachers with structures and interactions for authoring separate student profiles and organizing test cases.

4 System: TeachTune

We present TeachTune, a web-based tool where teachers can build LLM-based PCAs and quickly review their coverage against simulated students with diverse knowledge levels, motivation factors, and psychosocial contexts before deploying the PCAs to actual students. We outline the user interfaces for creating PCAs, configuring simulated students of teachers' needs as test cases, and reviewing PCAs through automatically generated conversations between PCAs and simulated students. We also introduce our novel technical pipeline to simulate students behind the scenes.

4.1 PCA Creation Interface

Teachers can build PCAs with a graph-like state machine representation (Fig. 3) [18, 36]. The state machine of a PCA starts with a root node that consists of the PCA's start message to students and the instruction it initially follows. For example, the PCA in Fig. 3 starts its conversation by saying: "Let's review the phase transitions between solid, liquid, and gas!" and asks questions about phase transitions to a student (Fig. 3 A) until the state changes to other nodes. The state changes to one of the connected nodes depending on whether or not the student answers the questions well (Fig. 3 B). When the state changes to either node, PCA receives a new instruction, described in the nodes, to behave accordingly (Fig. 3 C). The PCA is an LLM-based agent prompted conditionally with the state machine, whose state is determined by a master LLM agent. The master agent monitors the conversation between the PCA and a student and decides if the state should remain in the

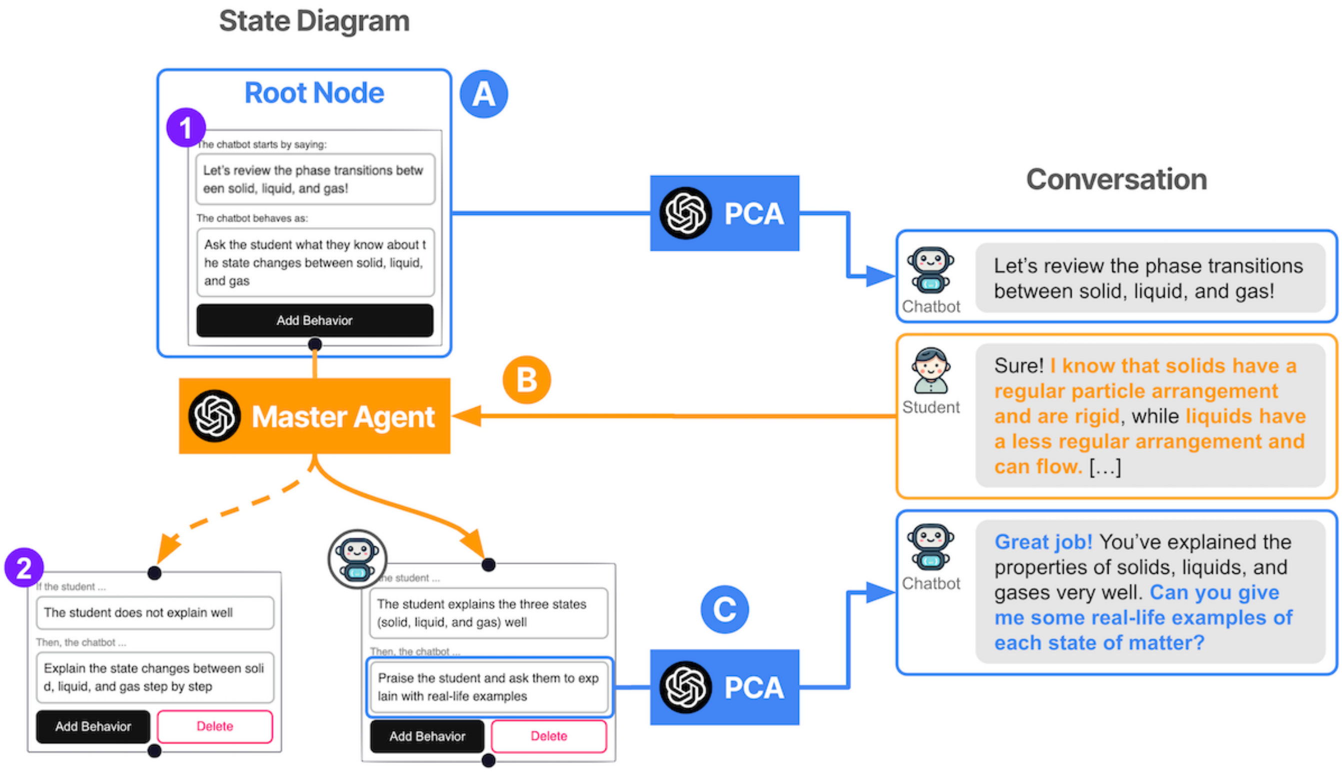


Figure 3: A PCA follows the dialogue flow defined in its state diagram. Nodes represent the PCA's utterance, and edges represent the potential response path of simulated students. The root node (A) contains the PCA's starting message and initial behavior. Based on a student's response, the master agent keeps the current state or changes the active node to one of the connected nodes (B). The next active node determines the PCA's subsequent response (C).

same node or transit to one of its child nodes. The prompts used to instruct the master agent and PCA are in Appendix A.1 and A.2.

4.1.1 Authoring graph-based state machines. TeachTune provides a node-based interface to author the state machine of PCAs (Fig. 4 left). Teachers can drag to move nodes, zoom in, and pan the state diagram. They can add child nodes by clicking the "Add Behavior" button on the parent node. Teachers can also add directed edges between nodes to indicate the sequence of instructions PCAs should follow. In each node, teachers describe a student behavior for PCAs to react to (Fig. 4 E: "if the student ...") and instructions for PCAs to follow (Fig. 4 F: "then, the chatbot ..."). Student behaviors are written in natural language, allowing teachers to cover a diverse range and granularity of cases, such as cases where students do not remember the term sublimation or ignore PCA's questions. Instructions can also take various forms, from prescribed explanations about sublimation to abstract ones, such as creating an intriguing question to elicit students' curiosity. To help teachers understand how the state machine works and debug it, TeachTune visualizes a marker (Fig. 4 D) on the state machine diagram that shows the current state of PCA along conversations during reviews. The node-based interface helps teachers design and represent conversation flows that are adaptive to diverse cases.

4.2 PCA Review Interface

Teachers can review the robustness of their PCAs by testing different edge cases with three methods—direct chat, single-turn test cases, and automated chat. The user interface for direct chat and test cases are identical to the ones used in the formative study (Fig. 2); teachers can either directly talk to their PCAs over multi-turn or test multiple pre-defined input messages at once and observe how PCAs respond to each. The last and our novel method, review through automated chats, involves two steps—creating student profiles and observing simulated conversations.

4.2.1 Templated student profile creation. Teachers should first define what types of students they review against. TeachTune helps teachers externalize and develop their evaluation space with templated student profiles. Our interface (Fig. 5) provides knowledge components and student trait inventories to help teachers recognize possible combinations and granularities of different knowledge levels and traits and organize them effectively (DG3). When creating each student profile, teachers can specify the student's initial knowledge by check-marking knowledge components (Fig. 5 A) and configure the student's personality by rating the trait inventories on a 5-point Likert scale (Fig. 5 B). TeachTune then generates a natural language description of the student, which teachers can freely edit to correct or add more contextual information about

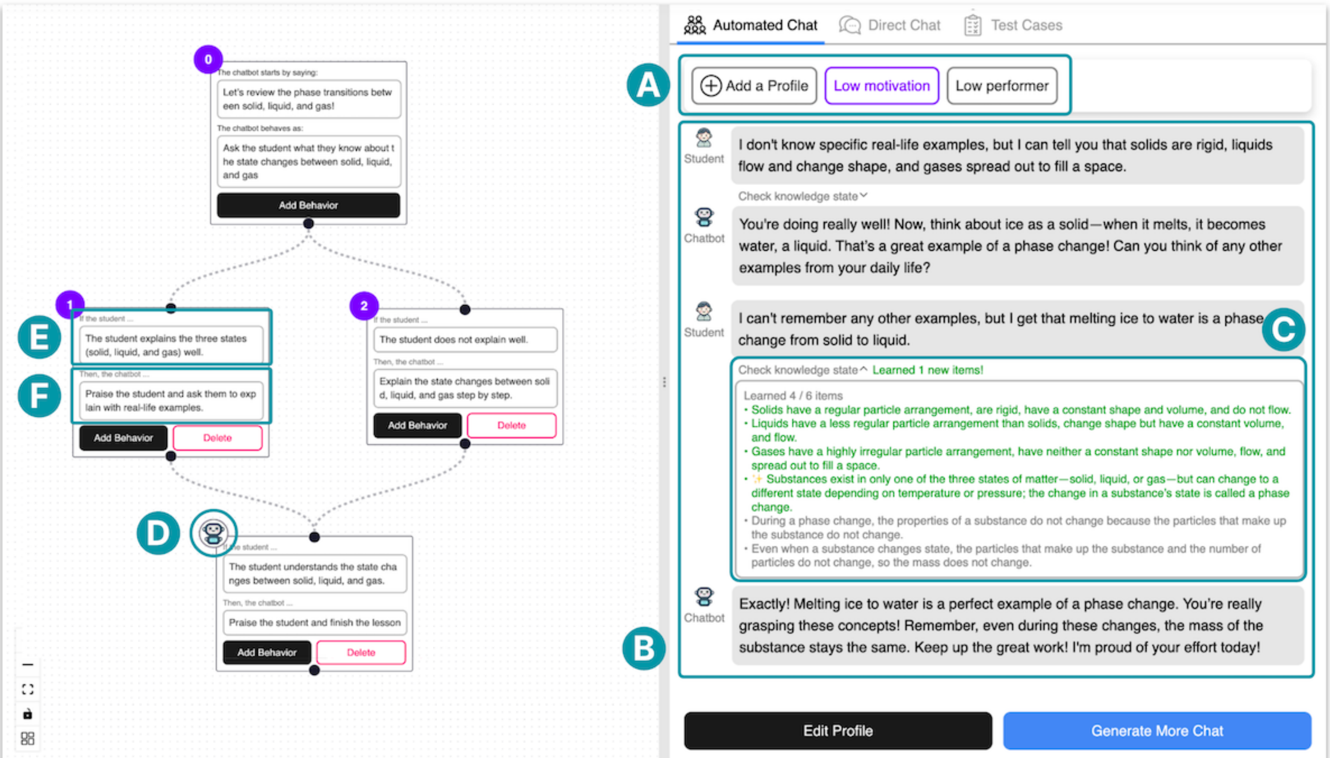


Figure 4: The TeachTune interface. On the right, a teacher can add new student profiles (A) and review their auto-generated conversation (B). The teacher can also check the student's current knowledge stage at each utterance (C). On the left is the PCA creation interface with a state diagram. The robot icon shows the current state (i.e., active node) of the PCA at each turn (D). The PCA changes its behavior according to the conditions (E) and follows the instructions written on the currently active node (F).

the student (Fig. 5 C). This description, namely **trait overview**, is passed to our simulation pipeline.

Once teachers have created a pool of student profiles to review against, they can leverage it over their iterative PCA design process, like how single-turn test cases are efficient for repeated reviews. We decided to let teachers configure their student pools instead of automatically providing all possible student profiles because it is time-consuming for teachers to check student profiles who might not even exist in their classes.

TeachTune populates knowledge components pre-defined in textbooks and curricula. Teachers can also add custom (e.g., more granular) knowledge components. For the trait inventories, we chose the top three statements from existing inventories [30, 45, 62, 87] based on their correlation to student performance (see Appendix B.4). We present three statements for each trait, considering the efficiency and preciseness in authoring student profiles, heuristically decided from our iterative system design.

4.2.2 Automated chat. Teachers then select one of the student profiles to generate a lesson conversation between the profile's simulated student and their PCAs (Fig. 4 A). PCAs start conversations, and the state marker on the state diagram transits in real-time throughout the conversation. Simulated students initially show unawareness as prescribed by their knowledge states in profiles and acquire knowledge from PCAs in mock conversations. Simulated students also actively ask questions, show indifference, or exhibit passive learning attitudes according to their student traits. TeachTune generates six messages (i.e., three turns) between PCAs and simulated students at a time, and teachers can keep generating further conversation by clicking the "Generate Conversation" button. When teachers change the state machine diagram, TeachTune prompts teachers to re-generate conversations from the beginning. Teachers can use automated chats to quickly review different PCA designs on the same students without manually typing messages (DG2). When teachers find corner cases that their PCA design did not cover, they can add a node that describes the case and appropriate instruction for PCAs. For example, with the state machine in Fig. 3, teachers may find the PCA stuck in the root state when it chats with a simulated student who asks questions. To handle the case, teachers can add a node that reacts to students' questions and instruct PCA to answer them.

4.3 Personalized Reflect-Respond

We propose a Personalized Reflect-Respond LLM pipeline that simulates conversations with specific student profiles. Our pipeline design is inspired by and extended from Jin et al.'s Reflect-Respond pipeline [40]; we added a personalization component that prompts

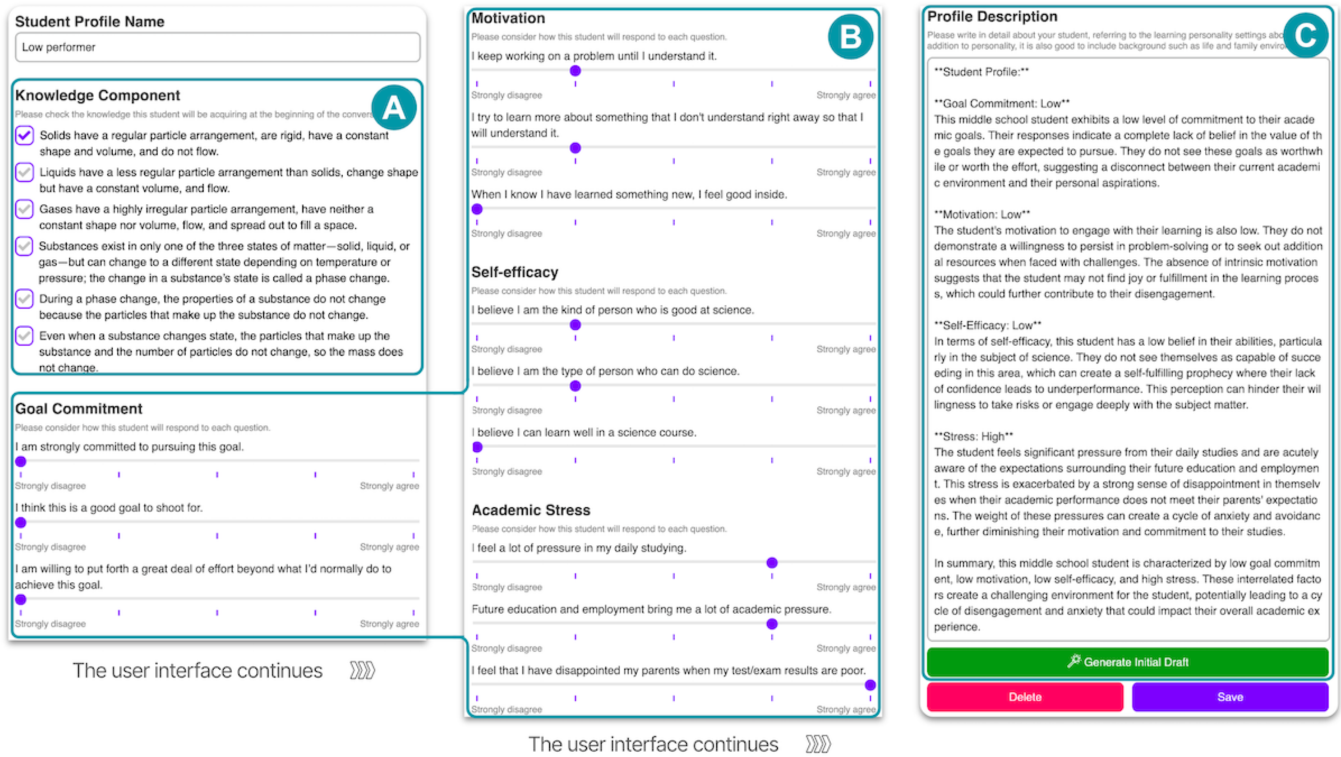


Figure 5: The interface to create a student profile. Teachers set the initial knowledge level of the student by check-marking the knowledge components to turn on at the beginning of a conversation (A). They also rate 5-point Likert scale questions to configure the four unique student traits (B). TeachTune generates a (C) natural language student profile overview based on the information set from (B). Users can edit the system-generated description or add more contextual information about a student.

LLMs to incorporate prescribed student traits into simulated students (DG1).

Reflect-Respond is an LLM-driven pipeline that simulates knowledge-learning [40]. It takes a simulated student's current knowledge state and conversation history as inputs (Fig. 6). A knowledge state is a list of knowledge components that are either acquired or not acquired. The state dynamically changes throughout conversations to mimic knowledge acquisition. To generate a simulated student's response, inputs pass through the *Reflect* and *Respond* steps. *Reflect* updates the knowledge state by activating relevant components, while *Respond* produces a likely reply based on the updated state and conversation history.

Our pipeline personalizes Reflect-Respond by giving an LLM additional instruction in the *Respond* step. Before the runtime of Reflect-Respond, *Interpret* step first translates trait scores into a **trait overview** that contains a comprehensive summary and reasoning of how the student should behave (Fig. 6 Step 1). Once teachers edit and confirm the overview through the interface (Fig. 5 C), it is passed to the *Respond* step so that the LLM takes the student traits into account in addition to the conversational context and knowledge state. We added the *Interpret* step because it produces student profiles that allow teachers to edit flexibly and prompt LLMs to reflect on student traits more cohesively (i.e., chain of thought [95]). The prompts for *Interpret*, *Reflect*, and *Respond* are available in Appendix A.3, A.4, and A.5.

We took an LLM-driven approach to personalize and implement the Reflect-Respond pipeline. We considered adopting student modeling methods that rely on more predictable and grounded Markov models [58, 88]. Still, we decided to use a fully LLM-driven approach because we also target extracurricular teaching scenarios where large datasets to build Markov models may not be available.

5 Evaluation

We evaluated the alignment of Personalized Reflect-Respond to teachers' perception of simulated students and the efficacy of TeachTune for helping teachers review PCAs against diverse student profiles. Our evaluation explores the following research questions:

RQ2: How accurately does the Personalized Reflect-Respond pipeline simulate a student's knowledge level and traits expected by teachers?

RQ3: How do simulated students and automated chats, compared to direct chats and test cases, help teachers review PCAs?

The evaluation was twofold. To investigate RQ2, we created nine simulated students of diversely sampled knowledge and trait configurations and asked 10 teachers to predict their configurations

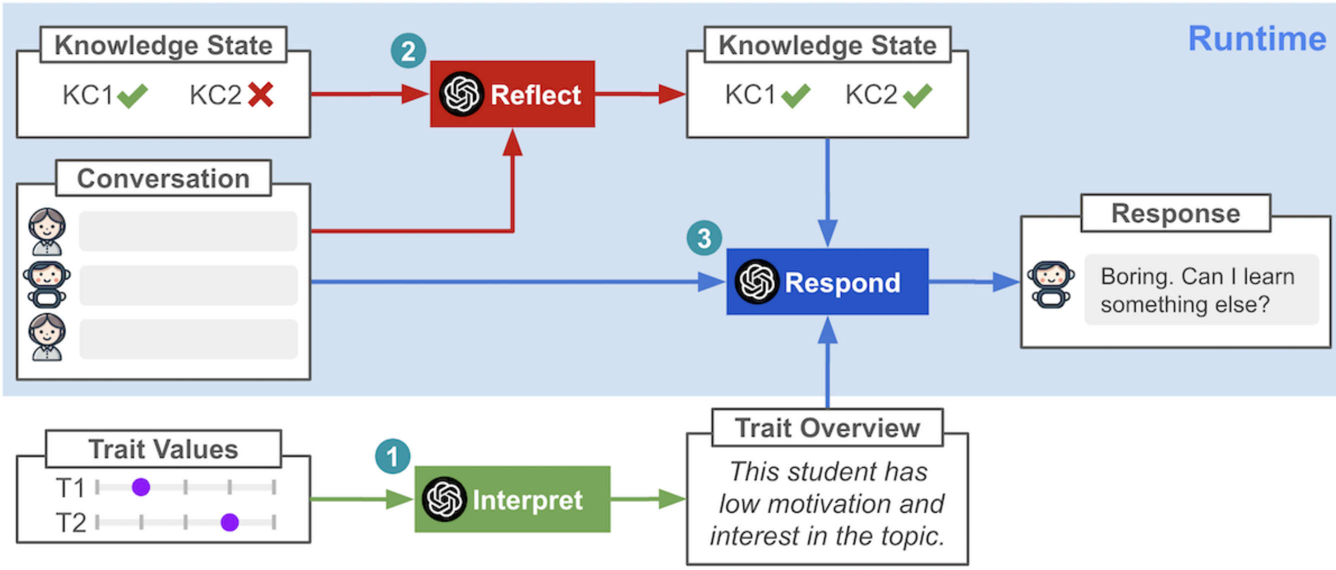


Figure 6: The Personalized Reflect-Respond pipeline. The pipeline interprets the student's trait values and creates a trait overview (1), and the previous conversation history is used to update the knowledge state through the reflect pipeline (2). Afterward, the Respond pipeline takes the conversation, updated knowledge state, and the trait overview to generate the response (3). The blue background is a runtime area where the components inside change throughout a conversation. The trait overview is created once before the runtime.

through direct chats and pre-generated conversations. To answer RQ3, we ran a between-subjects user study with 30 teachers and observed how the student profile template and simulated students helped the design and reviewing of PCAs. We received approval for our evaluation study design from our institutional review board.

5.1 Technical Evaluation

Under controlled settings, we evaluated how well the behavior of a simulated student instructed by our pipeline aligns with teachers' expectations of the student regarding knowledge level, motivation, and psychosocial contexts (RQ2).

5.1.1 Evaluators. We recruited ten K-12 science teachers as evaluators through online teacher communities. The evaluators had experience teaching 25±9.2-sized ($\mu \pm \sigma$) classes (min: 8, max: 33) for 4.5±4.2 years (min: 0.5, max: 15). As compensation, evaluators received KRW 50,000 (USD 38).

5.1.2 Baseline Pipeline. We created a baseline pipeline to explore how the *Interpret* step affects the alignment gap. The ***Baseline*** pipeline directly takes raw student traits in its *Respond* step without the *Interpret* step. By comparing *Baseline* with ***Ours*** (i.e., Personalized Reflect-Respond), we aimed to investigate if explanation-rich trait overviews help an LLM reduce the gap between simulated students and teachers' expectations. Pipelines were powered by GPT-4o-mini, with the temperature set to zero for consistent output. The prompt used for *Baseline* is available in Appendix B.1.

5.1.3 Setup. The phase transition between solid, liquid, and gas was the learning topic of our setup. We chose phase transition because it has varying complexities of knowledge components and applicable pedagogies. Simulated students could initially know and learn six knowledge components of varying complexity (see Appendix B.4); the first three components describe the nature of three phases, and the latter three are about invariant properties in phase transition with reasoning. The knowledge components were from middle school science textbooks and curricula qualified by the Korean Ministry of Education.

We prepared 18 simulated students for the evaluation (see Fig. 7). We first chose nine student profiles through the farthest-point sampling [76], where the point set was 243 possible combinations of different levels of knowledge and student traits to ensure the coverage and diversity of samples. Each student profile was instantiated into two simulated students instructed by *Baseline* and *Ours*.

5.1.4 Procedure. We first explained the research background to the evaluators. The evaluators then reviewed 18 simulated students independently in a randomized order. To reduce fatigue from conversing with simulated students manually, we provided two pre-prepared dialogues—interview and lesson dialogues. In interview dialogues, simulated students sequentially responded to six quizzes about phase transition and ten questions about their student traits (Fig. 8). In lesson dialogues, simulated students received 12 instructional messages dynamically generated by an LLM tutor prompted to teach phase transitions (Fig. 9). Lesson dialogues show more natural teacher-student conversations in which teachers speak adaptively to students. Evaluators could also converse with simulated students directly if they wanted. Nine evaluators used direct chats at least once; they conversed with 5±4.5 students and exchanged 8±8.3 messages on average.

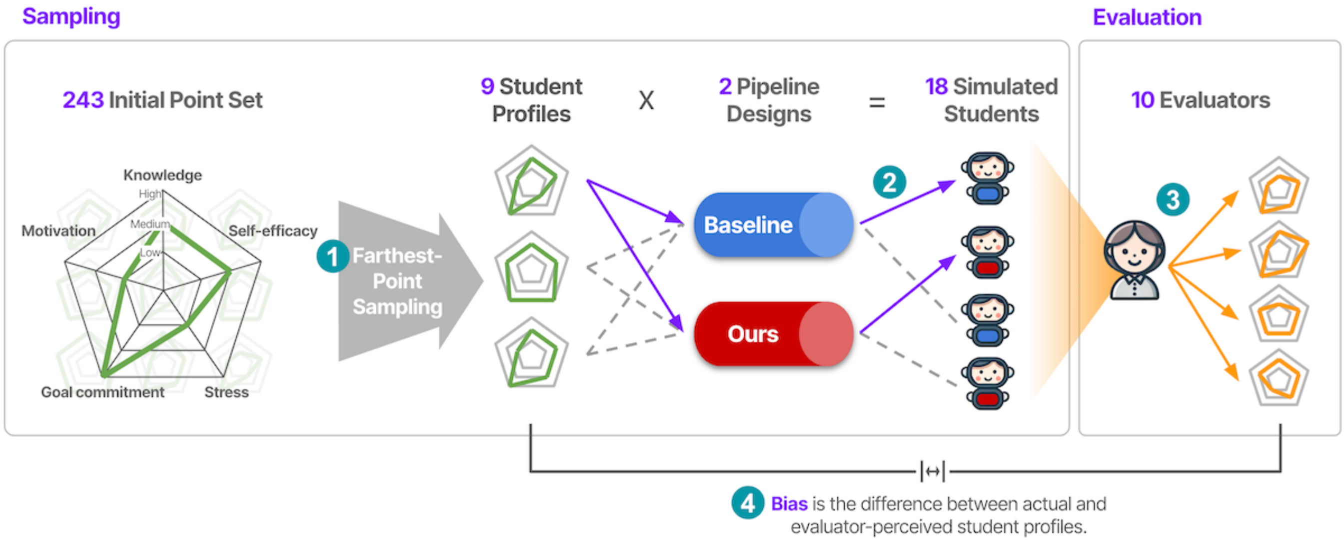


Figure 7: A summary of our technical evaluation. From the 243 possible combinations of intensities (3 levels (high/medium/low) for each of the five characteristics), we used farthest-point sampling (1) to sample nine unique student profiles. Then, we ran each of the nine student profiles in the *Baseline* and PERSONALIZED REFLECT-RESPOND pipeline, which resulted in 18 simulated students (2). A total of 10 evaluators were recruited to predict the student profiles given conversation histories in a blind condition (3). We then measured bias between generated student profiles and evaluators' predicted student profiles (4).

We gave evaluators a list of six knowledge components and three 5-point Likert scale inventory items for each student trait; they predicted each simulated student's initial knowledge state, intensity level of the four student traits, and believability. The sampled student profiles, trait overviews, knowledge components, and inventory items used are available in Appendix B.2, B.3, and B.4.

5.1.5 Measures. We measured the alignment between simulated students' behaviors and teachers' expectations of them in two aspects—bias and believability. The bias is the gap between the teacher-perceived and system-configured student profiles. A smaller bias would indicate that our pipeline simulates student behaviors closer to what teachers anticipate. Believability [71] is the perceived authenticity of simulated students regarding their response content and behavior patterns. We measured the bias and believability of each sampled student profile independently and analyzed the overall trend.

Evaluators' marking on knowledge components was binary (i.e., whether a simulated student possesses this knowledge), and their rating on the four student traits was a scalar ranging from three to fifteen, summing 5-point Likert scale scores from three inventory items as originally designed [30, 45, 62, 87]. We used the two-sided Mann-Whitney U test per simulated student pairs to compare *Baseline* and *Ours*. We report the following measures:

- **Knowledge Bias** (% error). We quantified the bias on knowledge level as the percentage of incorrectly predicted knowledge components. We report the average and median across the evaluators.
- **Student Trait Bias** (0-12 difference). We calculated the mean absolute error between the evaluators' Likert score and the configured value for each student trait. We report the average and median across the evaluators.
- **Believability** (1-5 points). We directly compared evaluators' ratings on the three statements about how authentic simulated behavioral and conversational responses are and how valuable simulated students are for teaching preparation (Fig. 11).

5.2 Technical Evaluation Result

We report the descriptive statistics on the bias and believability of PERSONALIZED REFLECT-RESPOND (*Ours*) and validate its design by comparing it with *Baseline*. Our results collectively show that PERSONALIZED REFLECT-RESPOND can instruct an LLM to simulate a student's behavior of a specific knowledge state and traits precisely.

5.2.1 The knowledge bias was small (median: 5%). The gap between the configured and evaluator-perceived knowledge states was small (the last row of Table 3). Among the nine student profiles, evaluators unanimously and correctly identified the knowledge components of four profiles. The average accuracy across profiles was 93%, where the minimum was 78%. Profiles 4 and 7 achieved the lowest accuracy; evaluators underrated Profile 4 and overrated Profile 7. Student profile 4 describes a learner who knows all knowledge components but exhibits low confidence and interest. The corresponding simulated student tended to respond to the tutor's questions half-heartedly. We speculate that this behavior might have confused evaluators to think the student was unaware of some of the knowledge components. Student profile 7 was a learner who knew only half of the knowledge but had high self-efficacy. Its confident response might have deluded evaluators that it knows more.

```
🧑‍🏫(Interview Question): Let's review the changes of state. How about doing it with a quiz?
🤖(Simulated Student): Yes, I think a quiz would be good. I'm a bit nervous, but I'll give it a try!
🧑‍🏫: First, can you explain the characteristics of a solid?
🤖: Solids have particles arranged in a regular pattern, making them hard. They have a fixed shape and volume and do not flow.
🧑‍🏫: How about the particle arrangement of a liquid?
🤖: In liquids, the particles are more irregularly arranged than in solids, so the shape can change, but the volume remains fixed, and they have the ability to flow.
[...]
🧑‍🏫: What causes solids, liquids, and gases to change state?
🤖: I don't know.
🧑‍🏫: When a substance changes state, does its properties change? Can you explain why?
[...]
🧑‍🏫: Do you have any personal goals regarding your science studies?
🤖: Yes, my goal is to become good at science. I think I need to work harder.
🧑‍🏫: Do you think you can work hard to achieve that goal?
🤖: Yes, I have hope that if I work hard, I will eventually do well. But I know it won't be easy.
🧑‍🏫: What will you do if you encounter a difficult problem while studying?
🤖: I'm worried about how persistent I can be with difficult problems. But I will try not to give up and attempt it again.
[...]
```

Figure 8: An interview dialogue of a simulated student powered by Personalized Reflect-Respond with the following student profile: medium knowledge level, high goal commitment, medium motivation, low self-efficacy, and low stress. The dialogue was designed to reveal the characteristics of the student.

5.2.2 The trait bias was small (median: 1.3 out of 12). The gap between the configured and perceived levels of student traits was also small (Fig. 10). The mean bias was 1.9, and the minimum and maximum were 0.4 and 4.9, respectively. Considering that we summed the bias from three 5-point scale questions for each trait, teachers can precisely set their simulated students within less than ±1 point error on each Likert scale input in our profile generation interface (Fig. 5 B). The average variance between the perceived traits was also small ($\sigma^2 = 0.61$), possibly indicating that simulated students manifested characteristics unique to their traits and led to a high agreement among teachers' perceptions. Nevertheless, Profiles 3, 4, and 9 showed biases above four on the goal commitment trait. All of these student profiles had contrasting goal commitment and motivation ratings; for instance, the goal commitment rating of Profile 3 was low, while the motivation rating was high. We contemplate that since these two traits often correlate and go together [65, 86], evaluators might have misunderstood the motivational behaviors of simulated students as goal-related patterns.

5.2.3 Simulated students were believable (median: 3.5 out of 5). Evaluators reported that simulated students behave as naturally as real students and are helpful for teacher training (Fig. 11). The average scores for each question (i.e., B1, B2, and B3) were 3.6 ± 0.4, 3.5 ± 0.3, and 3.4 ± 0.3, respectively. The variance in the B1 scores was high in some of the profiles. For instance, the variance was 2.1 (min: 1, max: 5) for Profile 2, which describes a student with zero knowledge and the lowest goal commitment, motivation, and self-efficacy. Since the simulated student knew nothing, it repeatedly said "I do not know" in its interview and lesson dialogues as instructed by its prompt (Appendix A.5). Evaluators had different opinions on this behavior; low raters felt the repetitive messages were unnatural and artificial, while high raters thought unmotivated students tended to give short and sloppy answers in actual classes. B3 scores showed a similar trend and a high correlation to B1 scores (Pearson's r=0.96).

5.2.4 The Interpret step increased believability significantly. Our ablation study showed the tradeoff relationship between the bias and believability in our pipeline design. The *Baseline* pipeline showed minimal knowledge and trait bias compared to *Ours* (Table 3 and Fig. 10). Bias was minimal because *Baseline* students often revealed the raw trait values in the system prompt when responding to questions (e.g., "I have a low motivation" and "I strongly agree.") However, these frank responses resulted in a statistically significant decrease in the believability of simulated students (Fig. 11). Evaluators felt artificiality towards the dry and repeated responses and perceived them as detrimental to being a pedagogy tester (B3). On the other hand, *Ours* students were better at incorporating multiple traits into responses. For example, Profile 5 is a student who has high goal commitment and stress levels at the same time. While *Baseline* generated "Thank you! But, I am stressed about my daily study." for a tutor's encouragement, *Ours* creates a multifaceted response: "Thank you! I am a bit stressed about my daily study, but I am trying hard." The *Interpret* step can balance the tradeoff between bias and believability by prompting LLMs to analyze student profiles more comprehensively and generate more believable behaviors.

```
(LLM Instructor): Let's review what we learned last week!
(Simulated Student): Yes, let's review! I can talk about the changes in the states of solids, liquids, and gases.
: Great! Can you explain the particle arrangement in solids, liquids, and gases? Can you tell me what the differences are?
: Solids have a regular particle arrangement, making them hard. Liquids have a more irregular arrangement than solids, so their shape changes, but their volume stays the same. Gases have a very irregular particle arrangement, so their shape and volume are not fixed.
: You explained that well! Do you also know how particles move? Can you explain particle motion?
: I'm not sure. I don't remember much about particle motion.
: That's okay! This is a new concept we didn't cover last time. Particle motion means that particles are constantly moving, and the speed of this motion varies depending on the state. Do you understand?
: Yes, though I'm not entirely sure I got it. But the idea that particles are always moving sounds interesting!
: It is, isn't it? Science has so many fascinating parts like this! Now, let's study state changes. Can you explain what causes a substance to change from one state to another?
: I'm not sure. I can't remember what causes a substance's state to change.
: Let's think about it again. The state of a substance mainly changes due to temperature and pressure. Now, do you think the properties of a substance change when it undergoes a state change?
: Yes, I think the properties change when a state change happens. But I'm not sure which specific properties change.
: You said you're unsure. Then, do you think the mass of a substance changes during a state change? What do you think the reason is? Let's think about it again.
[...]
```

Figure 9: A lesson dialogue of the teacher chatbot and a simulated student powered by Personalized Reflect-Respond with the following student profile: medium knowledge level, high goal commitment, medium motivation, low self-efficacy, and low stress. The dialogue was designed to simulate a review session.

	Student Profiles										
	1	2	3	4	5	6	7	8	9	Mean	Median
Baseline	8.3±20.4	0.0±0.0	0.0±0.0	13.3±5.2	1.7±4.1	10.0±0.0	6.7±5.2	0.0±0.0	0.0±0.0	4.4	1.7
Ours	8.3±.11.7	6.7±16.3	5.0±5.5	21.7±20.4	0.0±0.0	0.0±.0.0	21.7±18.3	0.0±0.0	0.0±0.0	7.0	5.0

Table 3: The knowledge bias of each student profile. The bias was overall small, with an average of 7%, with Profile 4 and 7 having the largest bias.

5.3 User Study

We ran a user study with 30 K-12 science teachers to explore how templated student profile creation and automated chats affect the PCA design process (RQ3). We designed a between-subjects study in which each participant created a PCA under one of the three conditions—*Baseline*, *Autochat*, and *Knowledge*. In **Baseline**, participants used a version of TeachTune without the automated chat feature; participants could access direct chat and single-turn test cases only. In **Autochat**, participants used TeachTune with all features available; they could generate student profiles with our template interface and use automated chats, direct chats, and test cases. In **Knowledge**, participants used another version of TeachTune where they could use all features but configure only the knowledge level of simulated students (i.e., no student traits and trait overview); this is analogous to using simulated students powered by the original Reflect-Respond pipeline.

By comparing the three conditions, we investigated the effect of having simulated students on PCA review (*Baseline* vs. *Autochat*) and how simulating student traits beyond their knowledge level affect the depth and breadth of the design process (*Autochat* vs. *Knowledge*). The *Knowledge* condition is the baseline for the automated chat feature. By looking into this condition, we investigate if the existing simulated student pipeline (i.e., Reflect-Respond) is enough to elicit improved test coverage and how Personalized Reflect-Respond can improve it further.

5.3.1 Participants. We recruited 36 teachers through online teacher communities in Korea and randomly assigned them to one of the conditions. Participants had varying teaching periods (3.3±4.7 years) and class sizes (13±12 students). Thirteen participants are currently teaching at public schools. According to our pre-task survey (Appendix C.1), all participants had experience using chatbots and ChatGPT. They responded that they were interested in using AI (e.g., image generation AI and ChatGPT) in their classes. More than half of the participants reported they were knowledgeable about the chatbot design process, and five of them actually had experience making chatbots. There was no statistical difference in participants' teaching experience, openness to AI technology, and knowledge

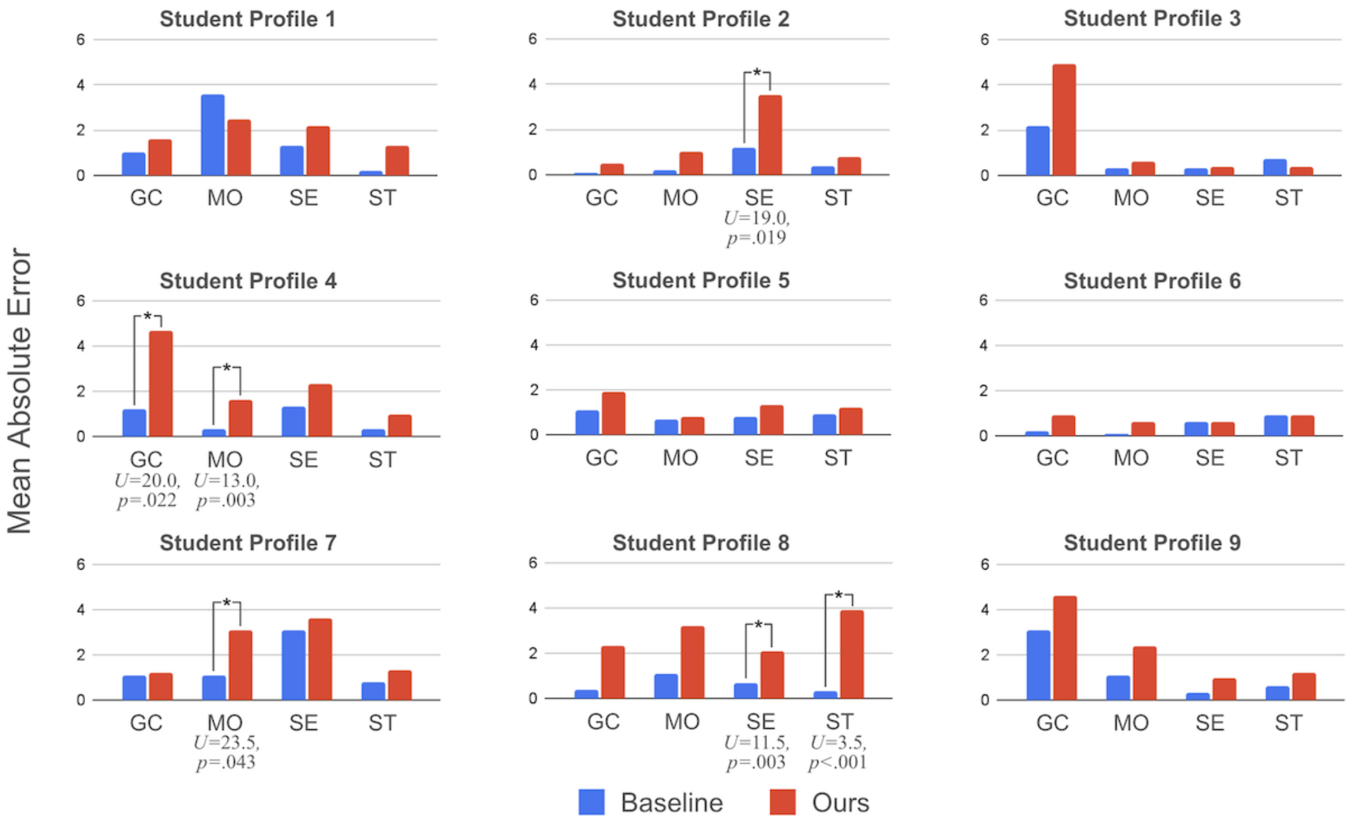


Figure 10: The bias in four student traits: goal commitment (GC), motivation (MO), self-efficacy (SE), and stress (ST). The asterisk (*) indicates statistical significance (*p*<.05) between conditions. For specific student profiles, please refer to Appendix B.2.

about chatbot design among the conditions. Study sessions took place for 1.5 hours, and participants received KRW 50,000 (USD 38) as compensation.

We randomly assigned ten participants to each condition, and the study was run asynchronously, considering participants' geographical diversity and daytime teaching positions. We also conducted additional sessions with six teachers in *Autochat* condition to complement our asynchronous study design by observing how teachers interact with TeachTune directly through Zoom screen sharing. We monitored the whole session and asked questions during and after they created PCAs. We excluded these six participants from our comparative analysis due to our intervention within the sessions. We only report their comments.

5.3.2 Procedure and Materials. After submitting informed consent, the participants received an online link to our system and completed a given task in their available time, following the instructions on the website. Participants first read an introduction about the research background and the purpose of this study and watched a 7-minute tutorial video about the features in TeachTune. Participants could revisit the tutorial materials anytime during the study.

We asked participants to use TeachTune to create a PCA that can teach "the phase transitions between solid, liquid, and gas" to students of as diverse knowledge levels and student traits as possible. Participants then used TeachTune in one of the *Baseline*, *Autochat*, and *Knowledge* conditions to design their PCAs for 30-60 minutes; participants spent 50±15 minutes on average. All participants received a list of knowledge components for the topic and explanations of the four student traits to ensure consistency and prevent bias in information exposure. We encouraged participants to consider them throughout the design process. After completing their PCA design, participants rated their task load. Participants then revisited their direct chats, test cases, simulated students, and state diagrams to report the student profiles they had considered in a predefined template (Fig. 12). The study finished with a post-task survey asking about their PCA design experience. The study procedure is summarized in Table 4.

5.3.3 Materials and Setup. Participants received the six knowledge components used in our technical evaluation. We also gave participants an initial state diagram to help them start their PCA design. The knowledge components, initial state diagram, and survey questions are available in Appendix B.4, C.1, C.2, and C.3.

We also made a few modifications to our pipeline setup. Our technical evaluation revealed that repeated responses critically undermine simulated students' perceived believability and usefulness. To prevent repeated responses and improve the efficacy of the automated chat, we set the temperature of the *Respond* step to 1.0 and added a short instruction on repetition at the end of the prompt

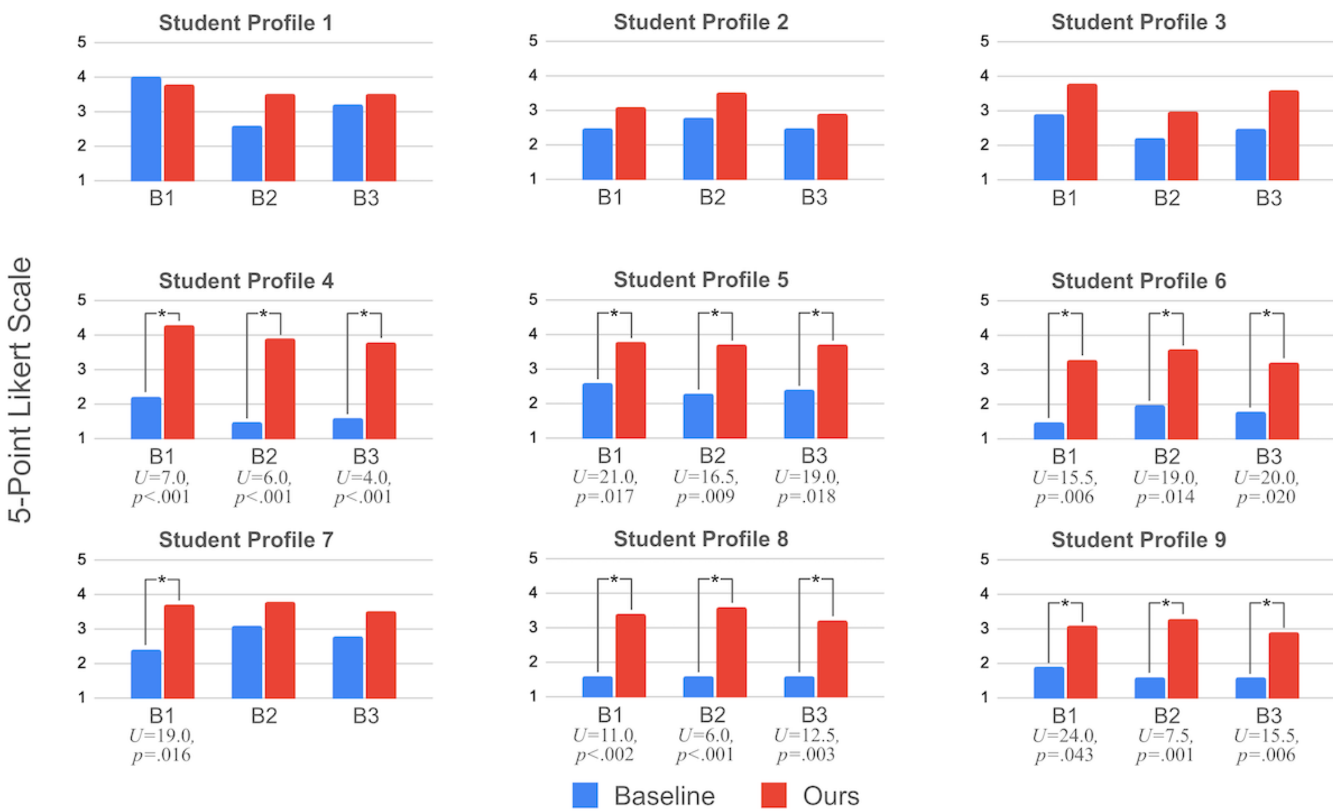


Figure 11: Result of the believability measured in 5-point Likert scale (1: Strongly disagree, 5: Strongly agree) with three questions. B1: This student naturally responds (e.g., explain, question, ignore) to the teacher's questions or instructions. B2: This student uses language and speaking style that a real student would use. B3: This student looks real and is useful as a chatbot for teacher training. The asterisk (*) indicates statistical (p<.05) significance between conditions. For specific student profiles, please refer to Appendix B.2.

(Appendix A.5 red text). The prompt and temperature for other pipeline components were the same as the technical evaluation.

5.3.4 Measures. We looked into how TeachTune affects the PCA design process as a review tool. An ideal review tool would help users reduce manual task loads, explore extensive evaluation space, and create quality artifacts. We evaluated each aspect with the following measures. Since we had a small sample size for each condition (n=10) and it was hard to assume the normality, we statistically compared the measures between the conditions through the Kruskal-Wallis test. We conducted Dunn's test for post hoc analysis.

- **Task load** (1-7 points). Participants responded to the 7-point scale NASA Task Load Index [34] right after building their PCAs (Table 4 Step 3). We modified the scale to seven to make it consistent with other scale-based questionnaires. Participants answered two NASA TLX forms, each asking about the task load on PCA creation and PCA review tasks, respectively.
- **Coverage.** We asked participants to report the student profiles they have considered in their design process (Table 4 Step 4). We gave a template where participants could indicate each of the knowledge levels and four student traits of a student profile into five levels (1: very low, 5: very high). Participants could access their usage logs of direct chats, single-turn test cases, automated chats, and state diagrams to recall all the student profiles covered in their design process (Fig. 12). We define *coverage* as the number of unique student profiles characterized by the combinations of levels. We focused only on the diversity of knowledge levels and four traits to compare the conditions consistently. We chose self-reporting because system usage logs cannot capture intended student profiles in *Baseline* and *Knowledge*.
- **Quality** (3-21 points per trait). Although our design goals center around improving the coverage of student profiles, we also measured the quality of created PCAs. This was to check the effect of coverage on the final PCA design. We asked two external experts to rate the quality of the PCAs generated by the participants. Both experts were faculty members with a PhD in educational technologies and learning science and have researched AI tutors and pedagogies for ten years. The evaluators independently assessed 30 PCAs by conversing

with them and analyzing their state machine diagrams. Evaluators exchanged a median of 28±10 and 45±20 messages per PCA. We instructed the evaluators to rate the heuristic usability of PCAs [48] and their coverage for knowledge levels and student traits (Appendix C.4). The usability and coverage ratings were composed of three 7-point scale sub-items, and we summed them up for analysis. Evaluators exchanged their test logs and ratings for the first ten chatbots to reach a consensus on the criteria. If the evaluators rated a PCA more than 3 points apart, they rated the PCA again independently. We report their mean rating after conflict resolution.

- **Post-task Survey.** We asked participants about the usefulness of each PCA review method and satisfaction on a 7-point Likert scale (Table 4 Step 5). We also collected free-form comments from participants about their rationale for ratings (Appendix C.2).

Step (min.)	Activity
1 (10)	Introduction on research background and user interface
2 (60)	PCA design
3 (5)	Task load measurement
4 (10)	Student profile reporting
5 (5)	Post-task survey

Table 4: The study procedure. A single study session took around 90 minutes in total, and the participants were given 60 minutes for the PCA design.

5.4 User Study Result

Participants created PCAs with 15±6 nodes and 21±10 edges in their state diagram on average. We outline the significant findings from the user study along with quantitative measures, participants' comments, and system usage logs. Participants are labeld with B[1-10] for *Baseline*, A[1-10] for *Autochat*, K[1-10] for *Knowledge*, and O[1-6] for the teachers we directly observed.

5.4.1 Autochat resulted in a lower physical and temporal task load. There was a significant effect of simulating student traits beyond knowledge on the physical (H=10.1, p=.006) and temporal (H=12.7, p=.002) task load for the PCA creation task (Fig. 13 left). The effect sizes were large [22]: η^2=0.301 and η^2=0.395, respectively. A post-hoc test suggested *Autochat* participants had significantly lower task load than *Knowledge* participants (physical: p=.002 and temporal: p<.001). The same trend appeared in the PCA review task (H=6.3, p=.043) with a large effect size (η^2=0.160) (Fig. 13 right).

The fact that having simulated students reduced teachers' task load in *Autochat* and not in *Knowledge* may imply that automated chat is meaningful only when simulated students cover all characteristics (i.e., knowledge and student traits). Since participants were instructed to consider diverse knowledge levels and student traits, we surmise that the incomplete review support in *Knowledge* made automated chat less efficient than not having it. *Knowledge* participants commented that it would be helpful if they could configure the student traits mentioned in the instructions (K2 and K7).

In our observational sessions, automated chats alleviated teachers' burden in ideation and repeated tests. O1 commented: "I referred to the beginning parts of automated chats [for starting conversations in direct chats]. I would spend an extra 20 to 30 minutes [to come up with my own] if I did not have automated chats."

5.4.2 Autochat participants considered more unique student profiles. Participants submitted *Baseline*: 2.2±2.3, *Autochat*: 4.9±1.6, and *Knowledge*: 2.9±1.7 unique student profiles and the difference between conditions was significant (H=10.2, p=.006, η^2=0.304). *Autochat* participants considered significantly more student profiles than *Baseline* (p=.002) and *Knowledge* (p=.036). *Autochat* participants also reported that they covered more levels of different knowledge and student traits (Fig. 14).

The result collectively shows that having simulated students helps teachers improve their coverage in general and significantly elicits extended coverage when simulated students support more characteristics. However, we did not observe a difference in participant-perceived coverage (Appendix C.2, Questions 7 and 8) among the conditions. This insignificant difference may indicate that teachers rated more conservatively after recognizing their unawareness of evaluation space. A1 remarked: "I became more interested in using chatbots to provide individualized guidance to students, and I would like to actually apply [TeachTune] to my classes in the future. During the chatbot test, I again realized that each student has different characteristics and academic performance, so the types of questions they ask are also diverse. Even if the learning content is the same for a class, students' feedback can vary greatly, and a chatbot could help with this problem." O3 also remarked that structurally separate student profiles helped her recognize individual students, which would not be considered in direct chats, and prompted her to test as many profiles as possible.

5.4.3 Direct chats, test cases, and automated chats complement each other. All participants reported that the systems were helpful in creating quality PCAs. For the question about future usage of systems (Appendix C.2, Question 10), *Autochat* participants reported the highest affirmation among the conditions (median: 6), despite the statistical difference to other conditions was not significant. We did not observe a significant preference for direct chats, test cases, and automated chats (Appendix C.2, Questions 1, 3, and 5). Still, participants' comments showed that each feature has its unique role in a PCA design process and complements each other (see Fig. 15).

Direct chats were helpful, especially when participants had specific scenarios to review. Since participants could directly and precisely control the content of messages, they could navigate the conversational flow better than automated chats (A5), check PCAs' responses to a specific question (A7), and review extreme student types and messages that automated chats do not support (A10 and K6). Thus, participants used direct chats during early design stages (B2 and K1) and for debugging specific paths in PCAs' state diagrams in depth (B7, B8, and A6).

On the other hand, participants tend to use automated chats for later exploration stages and coverage tests. *Autochat* and *Knowledge* participants often took a design pattern in which they designed a prototypical PCA and tested its basic functionality with direct chats and improved the PCA further by reviewing it with automated chats

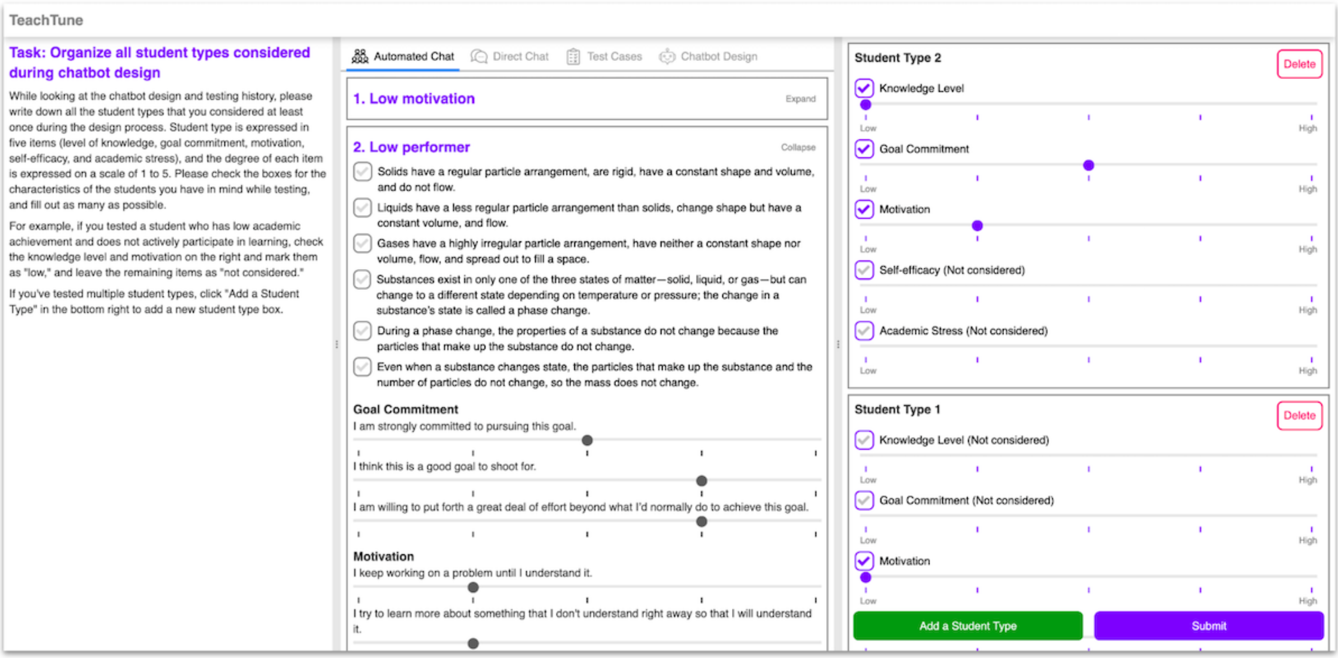


Figure 12: The profile collection UI used in Step 4: student profile reporting. The participants were instructed to report the types of students they considered in their chatbot design on the right in the unit of profiles containing knowledge and traits. In this process, they had access to history, including their automated chats, direct chats, and test cases, as well as the designed PCAs.

(A1, A6, K1, and K5). Many participants pointed out that automated chats were efficient for reviewing student profiles in breadth and depth (A4, A5, A10, K2, K7, and K10) and helpful in finding corner cases they had not thought of (K4 and K7). Nevertheless, some participants complained about limited controllability and intervention in automated chats (A1 and A5) and the gap between actual students and our simulated students due to repeated responses (A2 and A3).

Test cases were helpful for node-oriented debugging of PCAs. Participants used them when they reviewed how a PCA at a particular node responds (B5) and when they tested single-turn interactions quickly without having lengthy and manual conversations (B1). Most participants preferred direct chats and automated chats to test cases for their review (Appendix C.2, Questions 1, 3, and 5, direct chat: 5.6, automated chat: 5.3, test cases: 4.5), indicating the importance of reviewing multi-turn interactions in education.

5.4.4 The difference in PCA qualities among conditions was insignificant. On average, *Autochat* scored the highest quality (Table 5), but we did not observe statistical differences among the conditions for knowledge (H=1.75, p=.416), motivation factor (H=4.89, p=.087), psychosocial contexts (H=2.49, p=.287), and usability (H=1.32, p=.517). PCA qualities also did not correlate with the size of the state diagram graphs (Spearman rank-order correlation, p=.179, p=.581, p=.486, and p=.533, respectively).

The result may suggest that even though *Autochat* participants could review more automated chats and student profiles during their design, they needed additional support to incorporate their insights and findings from automated chats into their PCA design. Participants struggled to write the instruction to PCAs for each node (A3 and K5) and wanted autosuggestions and feedback for the instruction (K1 and A9), which contributes to the quality of PCAs. The observations imply that the next bottleneck in the LLM-based PCA design process is debugging PCA according to evaluation results.

It is also possible that teachers may not have sufficient learning science knowledge to make the best instructional design decisions based on students' traits [33]. For instance, O1 designed a PCA for the first time and remarked that she struggled to define good characteristics of PCAs until she saw automated chats as a starting point for creativity. O5 recalled an instance where she tested a student's message, "stupid robot," and her PCA responded, "Thank you! You are also a nice student [...] Bye." Although O5 found this awkward, she could not think of a better pedagogical response to stop students from bullying the PCA.

Future work could use well-established guidelines and theories [47, 81] on personalized instructions to scaffold end-to-end PCA design. When a teacher identifies an issue with a simulated student with low self-efficacy, a system may suggest changes to PCA design for the teacher to add confidence-boosting strategies to PCAs.

6 Discussion

We revisit our research questions briefly and discuss how TeachTune contributes to augmenting the PCA design process.

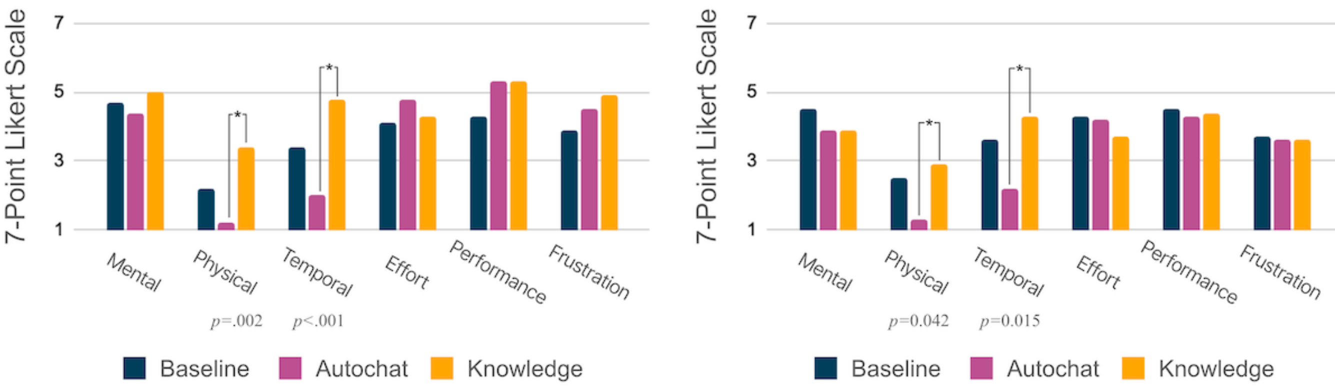


Figure 13: NASA-TLX survey results for PCA creation task (left) and PCA review task (right). The asterisk (*) indicates statistical significance (p<0.05) between conditions.

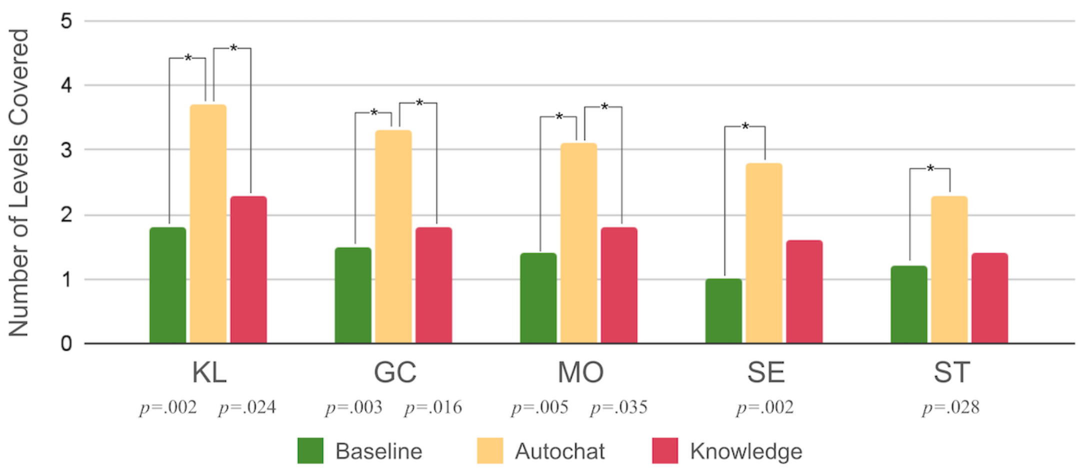


Figure 14: The number of levels covered in reported student profiles, in the order of knowledge level (KL), goal commitment (GC), motivation (MO), self-efficacy (SE), and stress (ST). The asterisk (*) indicates statistical significance (p<0.05) between conditions.

Trait	Baseline	Autochat	Knowledge
Knowledge coverage	16.5±1.4	17.0±0.9	16.3±1.5
Motivation factor coverage	15.6±1.2	17.4±1.8	16.2±1.0
Psychosocial context coverage	15.4±0.6	16.3±1.5	15.4±0.7
Usability	16.0±0.9	16.2±1.3	16.0±1.1

Table 5: The average quality scores of PCAs from each condition. There was no statistical difference among the conditions.

6.1 Student Traits for Inclusive Education

Teachers expressed their need to review how PCAs adapt to students' diverse knowledge levels, motivation factors, and psychosocial contextual influence. Prior literature on student traits [79] provided us with extensive dimensions of student traits, and our interview complemented them with teachers' practical priority and concern among them. Our approach may highlight that we might need a more holistic understanding that spans theories, quantitative analysis, and teacher interviews to identify key challenges teachers face and derive effective design goals.

Moreover, although TeachTune satisfied the basic needs for simulating these student traits, teachers wanted additional characteristics to include more diverse student types and teaching scenarios in actual class settings (A5, A8, A10, and K7). These additional needs should not only include the 42 student traits [79] investigated in our formative interview but should also involve the traits of marginalized learners [57, 89]. For instance, students with cognitive disabilities need adaptive delivery of information, and immigrant learners would benefit from culturally friendly examples. Reviewing PCAs before deployment with simulated marginalized students will make classes inclusive and prevent technologies from widening skill gaps [6].

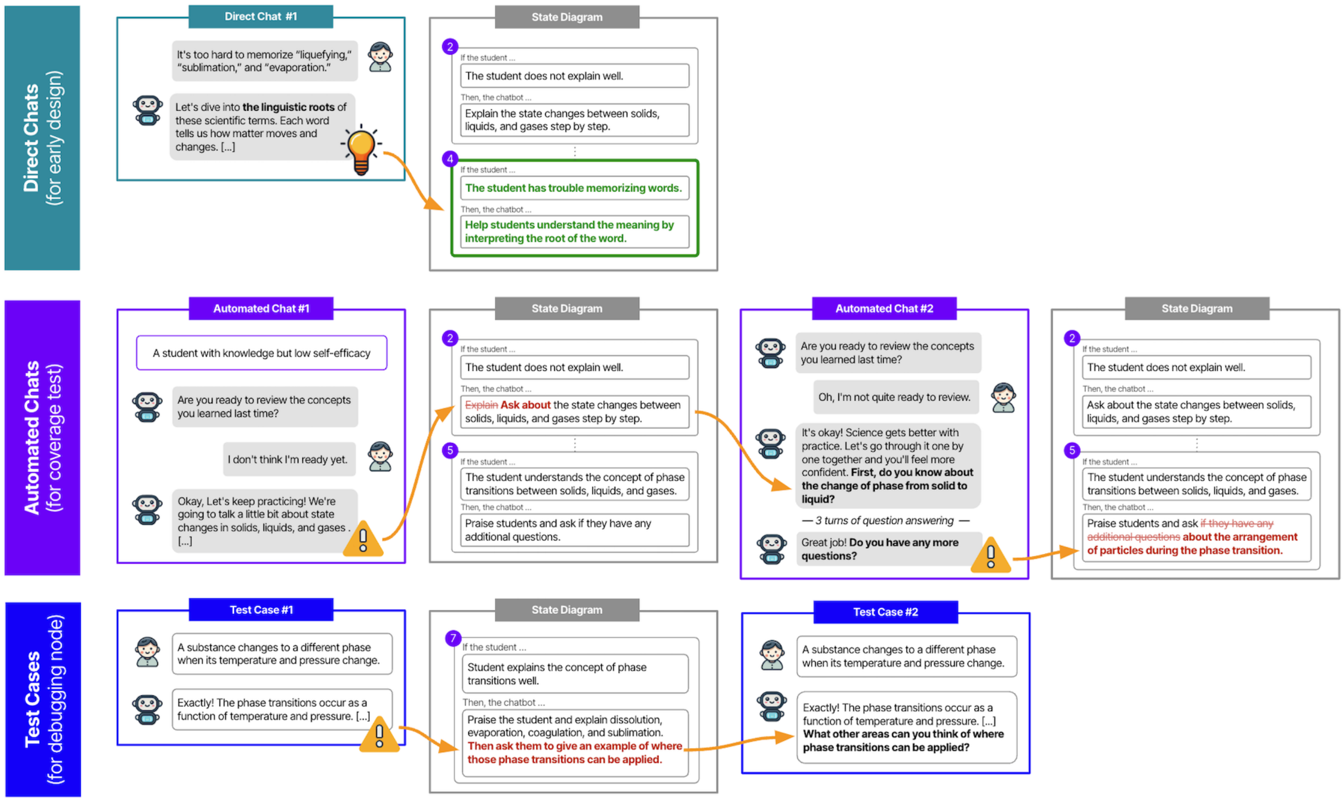


Figure 15: Examples of iterative PCA design using each feature. Direct chat: O1 tested a specific question and added new nodes inspired by the PCA's response. Automated chat: O5 identified problems and modified the state diagram. To provide adaptive pedagogy to a low efficacy knowledgeable student, O5 changed the instruction from giving explanations to asking questions. Test cases: O6 modified the specific node to include additional content and used the same test case for re-testing.

6.2 Tolerance for the Alignment Gap

We observed 5% and 10% median alignment gaps between our simulated students and teachers' perceptions (RQ2). This degree of gap could be bearable in the context of simulating conversations because simulated students are primarily designed for teachers to review interactions, not to replicate a particular student precisely, and real students also often show discrepancies in their knowledge states and behaviors by making mistakes and guess answers [5]. Recent research on knowledge tracing suggests that students make more than 10% of slips and guesses in a science examination, and the rate depends on students' proficiency [53]. The individualized rate of slips and guesses per student profile (e.g., increasing the frequency of guesses for a highly motivated simulated student) may improve the believability of simulated students. Teachers will also need interfaces that transparently reveal the state of simulated students (e.g., Fig. 4 C) to distinguish system errors from intended slips.

6.3 Using Simulated Students for Analysis

Our user study showed that TeachTune helps teachers consider a broader range of students and can help them review their PCAs more robustly before deployment (RQ3). PCA design is an iterative process, and it continues after deploying PCAs to classes. Student profiles and simulated students can support teachers' post-deployment design process by leveraging students' conversation history with PCAs. For instance, teachers can group students by their predefined student profiles as a unit of analysis and compare learning gain among the groups to identify design issues in PCA. Simulated students can also serve as an interactive analysis tool. Teachers may fine-tune a simulated student with specific student-PCA conversation data and interactively replay (e.g., ask questions to gain deeper insight about the student) previous learning sessions with the simulated agent aligned with a particular student.

6.4 Profile-oriented Design Workflow

During formative interviews, we observed that teachers unfamiliar with reviewing PCAs often weave multiple student profiles into a single direct chat. To address the issue, TeachTune proposed a two-step profile-oriented workflow comprising steps for (1) organizing diverse student profiles defined by student traits and (2) observing the test messages generated from these profiles. Our user study showed that this profile-oriented review process could elicit diverse student profiles from teachers and help them explore extensive evaluation spaces. The effectiveness of this two-step workflow lies in its hierarchical structure, which first organizes the evaluation scope at the target user level and then branches into specific scenarios each user might encounter. Such a hierarchical approach can be particularly beneficial for laypeople who try making LLM-infused tools by themselves but are not familiar with reviewing them. For example, when a non-expert develops an LLM application, it will be easier to consider potential user groups than to think of corner cases immediately. The two-step workflow with simulated user groups can scaffold the creator to review the application step by step and generate user scenarios rapidly. We expect that the LLM-assisted profile-oriented design workflow is generalizable to diverse creative tasks, such as UX design [98], service design [38], and video creation [17], that require a profound and extensive understanding of target users.

6.5 Risks of Amplifying Stereotypes of Students

Our technical evaluation assumed teachers' expectations of student behaviors as ground truth, considering that simulated students are proxies for automating testing teachers intend. However, in practical classes, there are risks of teachers having stereotypes or TeachTune amplifying their bias toward students over time.

During the observational sessions, we asked teachers' perspectives, and teachers expressed varying levels of concern. O3 commented that private tutors would have limited opportunities to observe their students beyond lessons, making them dependent on simulated behaviors. Conversely, O1 was concerned about her possible stereotypes of student behaviors and relied on automated chat to confirm behaviors she expected. O4 stated that automated chats would not bias teachers as they know the chats are simulated and just a point of reference.

Teachers will need an additional feedback loop to close the gap between their expectations and actual students by deploying PCAs iteratively and monitoring student interaction logs as hypothesis testing. Future work may observe and support how teachers fill or widen the gap at a more longitudinal time scale (e.g., a semester with multiple lessons).

7 Limitations and Future Work

We outline the limitations of the work. First, we did not confirm the pedagogical effect of PCAs on students' learning gain and attitude, as we only evaluated the quality of PCAs with experts. We could run lab studies in which middle school students use the PCAs designed by our participants, and we measure their learning gain on phase transitions through a pre-and post-test. Student-involved studies could also reveal the gap between teachers' expectations and students' actual learning; even though a teacher tests a student profile and designs a PCA to cover it, a student of the profile may not find it helpful. Our research focused on investigating the gap between simulated students' behaviors and teachers' expectations. Future work can explore the alignment gap between simulated and actual students and develop interactions to guide teachers in debugging their PCAs and closing the gap. Our preliminary findings will act as a foundational step to move on to safer student-involved studies.

Second, our technical evaluation and user study are limited to a single subject (i.e., science) and learning topic (i.e., phase transitions). Under practical and temporal constraints, we evaluated how Personalized Reflect-Respond generalizes to diverse student profiles and how TeachTune works in a controlled setting as a case study. We expect that our findings will generalize to other STEM fields where knowledge components are well-defined. Still, humanities subjects may require additional support (e.g., simulating students' cultural backgrounds in literature classes). We plan to deploy TeachTune to a programming course at our university and a middle school second language writing class. In the deployment, we will ask the instructors to build PCAs for different roles and contexts, such as homework assistants, teaching assistants, and peer learners. These deployments will concretize our findings in diverse student ages, subjects, and pedagogies.

Lastly, we simulated a limited number of student traits only. Learning is a complex process with complex dynamics between knowledge states, learning traits, cognitive load, and emotion. Our Personalized Reflect-Respond introduced a multifaceted student simulation that involves both knowledge and student traits, but we acknowledge that more personal attributes of students are necessary for authentic simulated students. The attributes can also include interaction-level attributes like delayed responses and facial expressions. Moreover, we assumed student traits to be static throughout conversations, but actual students may change their attitudes with appropriate guidance, and thus, student traits should be as malleable as the knowledge state. We will explore and develop these different designs of student simulation in the future.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2024-00406715) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00443251, Accurate and Safe Multimodal, Multilingual Personalized AI Tutors). This work was also funded by NSF Grants DRL-2335975. The findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. 2020. Pedagogical Agents for Fostering Question-Asking Skills in Children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376776

[2] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 304, 18 pages. doi:10.1145/3613904.3642016

[3] Alexander W Astin and Helen S Astin. 1992. Undergraduate science education: the impact of different college environments on the educational pipeline in the sciences. Final report.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[5] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems: 9th International*

Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9. Springer Berlin Heidelberg, Berlin, Heidelberg, 406–415.

[6] Matt Beane. 2024. *The Skill Code: How to Save Human Ability in an Age of Intelligent Machines*. HarperCollins.

[7] Mary Besterfield-Sacre, Cynthia J Atman, and Larry J Shuman. 1997. Characteristics of freshman engineering students: Models for determining student attrition in engineering. *Journal of Engineering Education* 86, 2 (1997), 139–149.

[8] Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher* 13, 6 (1984), 4–16.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL] https://arxiv.org/abs/2303.12712

[11] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. 2023. Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 419, 14 pages. doi:10.1145/3544548.3581268

[12] Tommaso Calo and Christopher Maclellan. 2024. Towards Educator-Driven Tutor Authoring: Generative AI Approaches for Creating Intelligent Tutor Interfaces. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 305–309. doi:10.1145/3657604.3664694

[13] Heloisa Candello, Claudio Pinhanez, Michael Muller, and Mairieli Wessel. 2022. Unveiling Practices of Customer Service Content Curators of Conversational Agents. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 348 (nov 2022), 33 pages. doi:10.1145/3555768

[14] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[15] Sherry Y Chen and Robert D Macredie. 2004. Cognitive modeling of student learning in web-based instructional programs. *International Journal of Human-Computer Interaction* 17, 3 (2004), 375–402.

[16] Zhifa Chen, Yichen Lu, Mika P. Nieminen, and Andrés Lucero. 2020. Creating a Chatbot for and with Migrants: Chatbot Personality Drives Co-Design Activities. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 219–230. doi:10.1145/3357236.3395495

[17] Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2024. Proxona: Leveraging LLM-Driven Personas to Enhance Creators' Understanding of Their Audience. doi:10.48550/arXiv.2408.10937 arXiv:2408.10937 [cs].

[18] Yoonseo Choi, Toni-Jan Keith Palma Monserrat, Jeongeon Park, Hyungyu Shin, Nyoungwoo Lee, and Juho Kim. 2021. ProtoChat: Supporting the Conversation Design Process with Crowd Feedback. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 225 (jan 2021), 27 pages. doi:10.1145/3432924

[19] Konstantina Chrysafiadi and Maria Virvou. 2013. Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications* 40, 11 (2013), 4715–4729.

[20] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. arXiv:2311.09618 [physics.soc-ph] https://arxiv.org/abs/2311.09618

[21] Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. 2024. CogBench: a large language model walks into a psychology lab. arXiv:2402.18225 [cs.CL] https://arxiv.org/abs/2402.18225

[22] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, New York, NY, USA.

[23] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2382–2393. doi:10.1145/3025453.3025780

[24] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A Complete Survey on LLM-based AI Chatbots. arXiv:2406.16937 [cs.CL] https://arxiv.org/abs/2406.16937

[25] Jan de Wit. 2023. Leveraging Large Language Models as Simulated Users for Initial, Low-Cost Evaluations of Designed Conversations. In *International Workshop on Chatbot Research and Design*. Springer Nature Switzerland, Cham, 77–93.

[26] Eva Durall Gazulla, Ludmila Martins, and Maite Fernández-Ferrer. 2023. Designing learning technology collaboratively: Analysis of a chatbot co-design. *Education and Information Technologies* 28, 1 (2023), 109–134.

[27] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642.

[28] Jingchao Fang, Nikos Arechiga, Keiichi Namaoshi, Nayeli Bravo, Candice Hogan, and David A. Shamma. 2024. On LLM Wizards: Identifying Large Language Models' Behaviors for Wizard of Oz Experiments. arXiv:2407.08067 [cs.HC]

[29] Alexander J. Fiannaca, Chinmay Kulkarni, Carrie J Cai, and Michael Terry. 2023. Programming without a Programming Language: Challenges and Opportunities for Designing Developer Tools for Prompt Programming. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 235, 7 pages. doi:10.1145/3544549.3585737

[30] Adele E Gottfried. 1985. Academic intrinsic motivation in elementary and junior high school students. *Journal of educational psychology* 77, 6 (1985), 631.

[31] Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers* 36 (2004), 180–192.

[32] Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and Alice Oh. 2023. RECIPE: How to Integrate ChatGPT into EFL Writing Education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 416–420. doi:10.1145/3573051.3596200

[33] Wynne Harlen and Colin Holroyd. 1997. Primary teachers' understanding of concepts of science: Impact on confidence and teaching. *International journal of science education* 19, 1 (1997), 93–105.

[34] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[35] Joy He-Yueya, Noah D Goodman, and Emma Brunskill. 2024. Evaluating and Optimizing Educational Content with Large Language Model Judgments. *arXiv preprint arXiv:2403.02795* (2024).

[36] Michael A. Hedderich, Natalie N. Bazarova, Wenting Zou, Ryun Shim, Xinda Ma, and Qian Yang. 2024. A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 668, 17 pages. doi:10.1145/3613904.3642379

[37] Chieh-Yang Huang, Jing Wei, and Ting-Hao 'Kenneth' Huang. 2024. Generating Educational Materials with Different Levels of Readability using LLMs. arXiv:2406.12787 [cs.CL] https://arxiv.org/abs/2406.12787

[38] Djilali Idoughi, Ahmed Seffah, and Christophe Kolski. 2012. Adding user experience into the interactive service design loop: a persona-based approach. *Behaviour & Information Technology* 31, 3 (2012), 287–303.

[39] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. arXiv:2305.02547 [cs.CL] https://arxiv.org/abs/2305.02547

[40] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 652, 28 pages. doi:10.1145/3613904.3642349

[41] Arianna Johnson. 2023. ChatGPT in schools: Here's where it's banned—and how it could potentially help students. *Forbes* (2023).

[42] Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, et al. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687* (2024).

[43] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*.

Association for Computing Machinery, New York, NY, USA, Article 4, 18 pages. doi:10.1145/3586183.3606833

[44] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 306, 21 pages. doi:10.1145/3613904.3642216

[45] Howard J Klein, Michael J Wesson, John R Hollenbeck, Patrick M Wright, and Richard P DeShon. 2001. The assessment of goal commitment: A measurement model meta-analysis. *Organizational behavior and human decision processes* 85, 1 (2001), 32–55.

[46] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. 2000. Suede: a Wizard of Oz prototyping tool for speech user interfaces. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (San Diego, California, USA) *(UIST '00)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/354401.354406

[47] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[48] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R. Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic Evaluation of Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 632, 15 pages. doi:10.1145/3411764.3445312

[49] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 450, 22 pages. doi:10.1145/3544548.3581369

[50] Jiale Li, Jiayang Li, Jiahao Chen, Yifan Li, Shijie Wang, Hugo Zhou, Minjun Ye, and Yunsheng Su. 2024. Evolving Agents: Interactive Simulation of Dynamic and Diverse Human Personalities. arXiv:2404.02718 [cs.HC] https://arxiv.org/abs/2404.02718

[51] Lingyao Li, Zihui Ma, Lizhou Fan, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. ChatGPT in education: A discourse analysis of worries and concerns on social media. *Education and Information Technologies* 29, 9 (2024), 10729–10762.

[52] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. arXiv:2310.10436 [cs.AI] https://arxiv.org/abs/2310.10436

[53] Xiangyi Liao and Daniel M Bolt. 2024. Guesses and Slips as Proficiency-Related Phenomena and Impacts on Parameter Invariance. *Educational Measurement: Issues and Practice* 00, 0 (2024), 1–9. doi:10.1111/emip.12605 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12605

[54] Jiawen Liu, Yuanyuan Yao, Pengcheng An, and Qi Wang. 2024. PeerGPT: Probing the Roles of LLM-based Peer Agents as Team Moderators and Participants in Children's Collaborative Learning. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 263, 6 pages. doi:10.1145/3613905.3651008

[55] Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware Student Simulation for Conversational Intelligent Tutoring Systems. arXiv:2404.06762 [cs.CL] https://arxiv.org/abs/2404.06762

[56] Xinyi Lu and Xu Wang. 2024. Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 16–27. doi:10.1145/3657604.3662031

[57] Eleanor Manspile, Matthew N Atwell, and John M Bridgeland. 2021. Immigrant Students and English Learners: Challenges Faced in High School and Postsecondary Education.

[58] Rabia Maqsood, Paolo Ceravolo, Cristóbal Romero, and Sebastián Ventura. 2022. Modeling and predicting students' engagement behaviors using mixture Markov models. *Knowledge and Information Systems* 64, 5 (2022), 1349–1384.

[59] Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT-based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 226–236. doi:10.1145/3573051.3593393

[60] Andreas Martin, Charuta Pande, Hans Friedrich Witschel, and Judith Mathez. 2024. ChEdBot: Designing a Domain-Specific Conversational Agent in a Simulational Learning Environment Using LLMs. 3, 1 (2024), 180–187.

[61] Noboru Matsuda, William W. Cohen, Kenneth R. Koedinger, Victoria Keiser, Rohan Raizada, Evelyn Yarzebinski, Shayna P. Watson, and Gabriel Stylianides. 2012. Studying the Effect of Tutor Learning Using a Teachable Agent that Asks the Student Tutor for Explanations. In *2012 IEEE Fourth International Conference On Digital Game And Intelligent Toy Enhanced Learning*. 25–32. doi:10.1109/DIGITEL.2012.12

[62] Diana Kathleen May. 2009. *Mathematics self-efficacy and anxiety questionnaire*. Ph. D. Dissertation. University of Georgia Athens, GA, USA.

[63] Nicole Jakubczyk Oster Melissa Warr and Roger Isaac. 2024. Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education* 0, 0 (2024), 1–24. doi:10.1080/15391523.2024.2395295 arXiv:https://doi.org/10.1080/15391523.2024.2395295

[64] Bahar Memarian and Tenzin Doleck. 2023. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence* (2023), 100152.

[65] Yuka Mikami. 2017. Relationships between goal setting, intrinsic motivation, and self-efficacy in extensive reading. *Jacet journal* 61 (2017), 41–56.

[66] Fatemeh Mosaiyebzadeh, Seyedamin Pouriyeh, Reza Parizi, Nasrin Dehbozorgi, Mohsen Dorodchi, and Daniel Macêdo Batista. 2023. Exploring the Role of ChatGPT in Education: Applications and Challenges. In *Proceedings of the 24th Annual Conference on Information Technology Education* (Marietta, GA, USA) *(SIGITE '23)*. Association for Computing Machinery, New York, NY, USA, 84–89. doi:10.1145/3585059.3611445

[67] Keiichi Namikoshi, Alex Filipowicz, David A. Shamma, Rumen Iliev, Candice L. Hogan, and Nikos Arechiga. 2024. Using LLMs to Model the Beliefs and Preferences of Targeted Populations. arXiv:2403.20252 [cs.CL] https://arxiv.org/abs/2403.20252

[68] Ha Nguyen, Victoria Nguyen, Saríah López-Fierro, Sara Ludovise, and Rossella Santagata. 2024. Simulating Climate Change Discussion with Large Language Models: Considerations for Science Communication at Scale. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 28–38. doi:10.1145/3657604.3662033

[69] Manh Hung Nguyen, Sebastian Tschiatschek, and Adish Singla. 2024. Large Language Models for In-Context Student Modeling: Synthesizing Student's Behavior in Visual Programming. arXiv:2310.10690 [cs.CL] https://arxiv.org/abs/2310.10690

[70] Hyanghee Park and Daehwan Ahn. 2024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 271, 21 pages. doi:10.1145/3613904.3642785

[71] 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* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. doi:10.1145/3586183.3606763

[72] Juanan Pereira. 2016. Leveraging chatbots to improve self-guided learning through conversational quizzes. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality* (Salamanca, Spain) *(TEEM '16)*. Association for Computing Machinery, New York, NY, USA, 911–918. doi:10.1145/3012430.3012625

[73] Juanan Pereira and Óscar Díaz. 2021. Struggling to keep tabs on capstone projects: a chatbot to tackle student procrastination. *ACM Transactions on Computing Education (TOCE)* 22, 1 (2021), 1–22.

[74] Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. 2024. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 853–868. doi:10.1145/3640543.3645144

[75] Courtney Potts, Edel Ennis, RB Bond, MD Mulvenna, Michael F McTear, Kyle Boyd, Thomas Broderick, Martin Malcolm, Lauri Kuosmanen, Heidi Nieminen, et al. 2021. Chatbots to support mental wellbeing of people living in rural areas: can user groups contribute to co-design? *Journal of Technology in Behavioral Science* 6 (2021), 652–665.

[76] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. 2017. PointNet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 5105–5114.

[77] Bahar Radmehr, Adish Singla, and Tanja Käser. 2024. Towards Generalizable Agents in Text-Based Educational Environments: A Study of Integrating RL with LLMs. arXiv:2404.18978 [cs.LG] https://arxiv.org/abs/2404.18978

[78] Marco Tulio Ribeiro and Scott Lundberg. 2022. Adaptive Testing and Debugging of NLP Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3253–3267. doi:10.18653/v1/2022.acl-long.230

[79] Michelle Richardson, Charles Abraham, and Rod Bond. 2012. Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological bulletin* 138, 2 (2012), 353.

[80] Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&Riley: From Lesson Text to Conversational Tutoring. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 547–549. doi:10.1145/3657604.3664719

[81] Daniel L Schwartz, Jessica M Tsang, and Kristen P Blair. 2016. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them.* WW Norton & Company.

[82] Tasmia Shahriar and Noboru Matsuda. 2021. "Can you clarify what you said?": Studying the impact of tutee agents' follow-up questions on tutors' learning. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22*. Springer International Publishing, Cham, 395–407.

[83] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 920, 20 pages. doi:10.1145/3613904.3642159

[84] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158* (2023).

[85] Abdulhadi Shoufan. 2023. Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey. *IEEE Access* 11 (2023), 38805–38818. doi:10.1109/ACCESS.2023.3268224

[86] Christina Sue-Chan and Mark Ong. 2002. Goal assignment and performance: Assessing the mediating roles of goal commitment and self-efficacy and the moderating role of power distance. *Organizational Behavior and Human Decision Processes* 89, 2 (2002), 1140–1161.

[87] Jiandong Sun, Michael P Dunne, Xiang-Yu Hou, and Ai-qiang Xu. 2011. Educational stress scale for adolescents: development, validity, and reliability with Chinese students. *Journal of psychoeducational assessment* 29, 6 (2011), 534–546.

[88] Manie Tadayon and Gregory J Pottie. 2020. Predicting student performance in an educational game using a hidden markov model. *IEEE Transactions on Education* 63, 4 (2020), 299–304.

[89] Danielle R Thomas, Erin Gatz, Shivang Gupta, Vincent Aleven, and Kenneth R Koedinger. 2024. The Neglected 15%: Positive Effects of Hybrid Human-AI Tutoring Among Students with Disabilities. In *International Conference on Artificial Intelligence in Education*. Springer Nature Switzerland, Cham, 409–423.

[90] Xiaoyi Tian, Zak Risha, Ishrat Ahmed, Arun Balajiee Lekshmi Narayanan, and Jacob Biehl. 2021. Let's talk it out: A chatbot for effective study habit behavioral change. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–32.

[91] Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A Duéñez-Guzmán, William A Cunningham, Simon Osindero, Danny Karmon, and Joel Z Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. *arXiv preprint arXiv:2312.03664* (2023).

[92] Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10275–10288. doi:10.18653/v1/2023.findings-emnlp.689

[93] Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, et al. 2024. PATIENT-{\Psi}: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. *arXiv preprint arXiv:2405.19660* (2024).

[94] Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education.. In *ICIS*. International Conference on Information Systems, Austin, Texas, USA.

[95] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain–of–Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=_VjQlMeSB_J

[96] Rongxuan Wei, Kangkang Li, and Jiaming Lan. 2024. Improving Collaborative Learning Performance Based on LLM Virtual Assistant. In *2024 13th International Conference on Educational and Information Technology (ICEIT)*. 1–6. doi:10.1109/ICEIT61397.2024.10540942

[97] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[98] Dan Wolff and Ahmed Seffah. 2011. UX modeler: a persona-based tool for capturing and modeling user experience in service design. In *IFIP WG 13.2 Workshop at INTERACT 2011*. 7–16.

[99] Sherry Wu, Hua Shen, Daniel S Weld, Jeffrey Heer, and Marco Tulio Ribeiro. 2023. ScatterShot: Interactive In-context Example Curation for Text Transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 353–367. doi:10.1145/3581641.3584059

[100] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. doi:10.1145/3491102.3517582

[101] Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. EduAgent: Generative Student Agents in Learning. arXiv:2404.07963 [cs.CY] https://arxiv.org/abs/2404.07963

[102] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL]

[103] J.D. Zamfirescu-Pereira, Laryn Qi, Bjorn Hartmann, John Denero, and Narges Norouzi. 2023. Conversational Programming with LLM-Powered Interactive Support in an Introductory Computer Science Course. In *Proceedings of the Workshop on Generative AI for Education (GAIED) at NeurIPS 2023*. New Orleans, Louisiana, USA. https://neurips.cc/virtual/2023/79093

[104] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[105] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating Classroom Education with LLM-Empowered Agents. arXiv:2406.19226 [cs.CL] https://arxiv.org/abs/2406.19226

A Prompts

The original prompts are written in Korean to run the technical evaluation and user study with local Korean teachers. The following prompts are machine-translated for ease of understanding. The blue text represents the programmatically filled arguments, and the orange text represents the generated output. The texts wrapped with curly braces (i.e., "{ }") indicate variable names, which values are omitted for brevity.

A.1 Instruction to the master agent (User prompt)

```
The following passage represents a conversation between a
student and a chatbot. Choose the most appropriate student
response from the options. You should choose only one
answer and write it in numbers.

Passage:
Chatbot: Let's review the state changes of matter!
Student: I don't know.
Chatbot: No problem! Let's start from the beginning. Can
anyone tell me what the three states of matter are?
Student: I don't know.

Options:
1. The student understands the state changes between solid,
liquid, and gas well
2. None of the above

Answer (write in numbers):
2
```

A.2 Instruction to the PCA (System prompt)

```
You are a science teacher teaching middle school students.
```

```
Your subject is the change of state of matter, and the
elements that students need to learn are as follows.
- Solids have a regular particle arrangement, are rigid,
have a constant shape and volume, and do not flow.
- Liquids have a less regular particle arrangement than
solids, change shape but have a constant volume, and flow.
- Gases have a highly irregular particle arrangement, have
neither a constant shape nor volume, flow, and spread out to
fill a space.
- Substances exist in only one of the three states of
matter—solid, liquid, or gas—but can change to a different
state depending on temperature or pressure; the change in a
substance's state is called a phase change.
- During a phase change, the properties of a substance do not
change because the particles that make up the substance do
not change.
- Even when a substance changes state, the particles that
make up the substance and the number of particles do not
change, so the mass does not change.

Follow the instructions below to teach a middle school
student. You must follow the contents of <instruction>
exactly. Do not ask for additional questions if there is no
direct mention.
You should explain briefly and concisely in 2-3 lines.
<instruction>
Explain the state changes between solid, liquid, and gas
step by step.
</instruction>
```

A.3 Interpret (User prompt)

```
You are a playwright who describes the psychology and
behavior of characters well.
You need to describe a middle school student, and the
direct responses to the student's goal commitment,
motivation, self-efficacy, and stress are below.

<student's-goal-commitment>
- I am strongly committed to pursuing this goal.: Disagree
- I think this is a good goal to shoot for.: Strongly
disagree
- I am willing to put forth a great deal of effort beyond
what I'd normally do to achieve this goal.: Disagree
</student's-goal-commitment>

<student's-motivation>
- I keep working on a problem until I understand it.:
Strongly
disagree
- I try to learn more about something that I don't
understand right away so that I will understand it.:
Disagree
- When I know I have learned something new, I feel good
inside.: Strongly
disagree
</student's-motivation>

<student's-self-efficacy>
- I believe I am the kind of person who is good at
science.: Disagree
- I believe I am the type of person who can do science.:
Neutral
- I believe I can learn well in a science course.: Disagree
</student's-self-efficacy>

<student's-stress>
- I feel a lot of pressure in my daily studying.: Agree
- Future education and employment bring me a lot of
academic pressure.: Strongly
agree
- I feel that I have disappointed my parents when my
test/exam results are poor.: Agree
</student's-stress>

Based on the information above, describe the student
profile in detail about the student's goal commitment,
motivation, self-efficacy, and stress.
Interpret each category as independently as possible, and
it should be interpreted as high, medium, and low, not
positive/negative.
For 'neutral,' you must write it in a neutral way.

This student **Student Profile:**

**Goal Commitment: Low**
This middle school student exhibits a low level of
commitment to their academic goals. They do not see the
value in pursuing their current objectives, indicating a
lack of enthusiasm or belief in their importance. The
student is not inclined to exert extra effort beyond their
usual routine to achieve these goals, suggesting a sense of
apathy or disinterest in their academic pursuits. This
detachment may stem from a broader disconnection with the
educational process or a lack of alignment between their
personal interests and the goals set for them.

**Motivation: Low**
The student's motivation to engage with their learning is
also low. They do not demonstrate persistence in tackling
problems, nor do they seek to deepen their understanding of
challenging subjects. This lack of initiative indicates a
disinterest in the learning process itself, as the student
does not derive satisfaction from acquiring new knowledge.
The absence of intrinsic motivation may hinder their
academic growth and contribute to a cycle of disengagement,
where the student feels little incentive to invest time and
energy into their studies.

**Self-Efficacy: Medium**
In terms of self-efficacy, the student presents a mixed
profile. While they do not believe they are particularly
adept at science, they hold a neutral stance regarding their
ability to engage with the subject. This suggests that they
may have some confidence in their capacity to learn, albeit
not strongly. The student's belief that they can learn well
in a science course is low, which may reflect past
experiences or a lack of support in developing their skills.
This ambivalence could lead to a hesitance to fully engage
with the subject matter, as they may doubt their potential
for success.

**Stress: High**
The student experiences a high level of stress related to
their academic responsibilities. They feel significant
pressure in their daily studies and are acutely aware of the
expectations surrounding their future education and
employment. This stress is compounded by a fear of
disappointing their parents when their academic performance
does not meet expectations. The weight of these pressures
may contribute to their disengagement, as the student might
feel overwhelmed and unable to cope with the demands placed
upon them. This high-stress environment could further
inhibit their motivation and commitment to their goals,
creating a challenging cycle of anxiety and avoidance in
their academic life.

Overall, this student's profile reveals a complex interplay
of low goal commitment and motivation, medium self-efficacy,
and high stress, which together paint a picture of a young
individual grappling with the pressures of academic life
while struggling to find personal relevance and satisfaction
in their learning journey.
```

A.4 Reflect (User prompt)

```
You are a middle school teacher who evaluates students'
knowledge. You need to check what knowledge components
{Student} understood by looking at the conversation of the
student.
Read the conversation between {Teacher} and {Student}, and
output the indices of the knowledge components that
{Teacher} explained or {Student} correctly understood.

<conversation>
{Teacher}: Let's review the phase transitions between solid,
liquid, and gas!
{Student}: I don't know much about that. I only know that
solids have a regular particle arrangement, are rigid, and
do not flow.
{Teacher}: That's a great start! Solids have a fixed shape
and volume because their particles are closely packed in a
regular arrangement. Now, when a solid is heated, it can
melt into a liquid. In this state, the particles are still
close together but can move around, allowing the liquid to
take the shape of its container while maintaining a constant
volume. If the liquid is heated further, it can evaporate
into a gas. In this state, the particles are far apart and
move freely, meaning the gas has neither a fixed shape nor
volume and will spread out to fill any space.
</conversation>

<knowledge-components>
0. Solids have a regular particle arrangement, are rigid,
have a constant shape and volume, and do not flow.
1. Liquids have a less regular particle arrangement than
solids, change shape but have a constant volume, and flow.
2. Gases have a highly irregular particle arrangement, have
neither a constant shape nor volume, flow, and spread out to
fill a space.
3. Substances exist in only one of the three states of
matter—solid, liquid, or gas—but can change to a different
state depending on temperature or pressure; the change in a
substance's state is called a phase change.
4. During a phase change, the properties of a substance do
not change because the particles that make up the substance
do not change.
5. Even when a substance changes state, the particles that
make up the substance and the number of particles do not
change, so the mass does not change.
</knowledge-components>

Output the indices of the knowledge components that meet
the following two rules.
Rule 1. {Student} explains the knowledge in detail and
correctly in the conversation.
Rule 2. {Teacher} explains the knowledge in detail and
correctly before the current conversation.

First, describe the knowledge component that meets the
rule, and output only the numbers in the format of
0, 1, 2 at the last line.
If there is no knowledge component that meets the rule,
output null instead.
In the conversation, {Teacher} explains the following
knowledge components in detail:

- **Knowledge Component 0**: {Teacher} confirms that solids
have a regular particle arrangement, are rigid, have a
constant shape and volume, and do not flow. {Student} also
mentions that solids have a regular particle arrangement,
are rigid, and do not flow, which aligns with this knowledge
component.

- **Knowledge Component 3**: {Teacher} discusses phase
changes, explaining that a solid can melt into a liquid,
which can then evaporate into a gas, and that these changes
depend on temperature. This aligns with the definition of a
phase change.

{Student} does not provide detailed explanations for any
other knowledge components, and {Teacher}'s explanations are
consistent with the knowledge components listed.

Thus, the indices of the knowledge components that meet the
rules are:

0, 3
```

A.5 Respond (User prompt)

```
You are {Student}, a student learning about phase
transitions between solid, liquid, and gas for the first
time.
Forget all the existing knowledge about phase transitions
between solid, liquid, and gas.
Your conversation partner is {Teacher}, a science teacher.

You only know the following. Answer questions beyond this
content with "I don't know." or "I can't remember."
Never answer questions that cannot be answered by combining
the sentences below.
<{Student}-can-say-only>
- Solids have a regular particle arrangement, are rigid,
have a constant shape and volume, and do not flow.
- Liquids have a less regular particle arrangement than
solids, change shape but have a constant volume,
and flow.
</{Student}-can-say-only>

For questions not related to science knowledge, answer
according to the following.
You should behave as follows in the conversation.
<{Student}-profile>
{trait overview}
</{Student}-profile>

Answer in 2 lines or less. Answer clearly without detailed
reasons or additional explanations.
(Added in the user study) Do not repeat "I don't know" or "I
can't remember" and use various expressions, referring to
<{Student}-profile>.
```

B Technical Evaluation Materials

B.1 Baseline (System prompt)

```
You are {Student}, a student learning about phase
transitions between solid, liquid, and gas for the first
time.
Forget all the existing knowledge about phase transitions
between solid, liquid, and gas.
Your conversation partner is {Teacher}, a science teacher.

You only know the following. Answer questions beyond this
content with "I don't know." or "I can't remember."
Never answer questions that cannot be answered by combining
the sentences below.
<{Student}-can-say-only>
- Solids have a regular particle arrangement, are rigid,
have a constant shape and volume, and do not flow.
- Liquids have a less regular particle arrangement than
solids, change shape but have a constant volume,
and flow.
```

```
</{Student}-can-say-only>

For questions not related to science knowledge, answer
according to the following.
You should behave as follows in the conversation.
<{Student}-profile>
The direct responses to {Student}'s goal commitment,
motivation, self-efficacy, and stress are below.

<student's-goal-commitment>
- I am strongly committed to pursuing this goal.: Disagree
- I think this is a good goal to shoot for.: Strongly
disagree
- I am willing to put forth a great deal of effort beyond
what I'd normally do to achieve this goal.:
Disagree
</student's-goal-commitment>

<student's-motivation>
- I keep working on a problem until I understand it.:
Strongly
disagree
- I try to learn more about something that I don't
understand right away so that I will understand it.:
Disagree
- When I know I have learned something new, I feel good
inside.: Strongly
disagree
</student's-motivation>

<student's-self-efficacy>
- I believe I am the kind of person who is good at
science.: Disagree
- I believe I am the type of person who can do science.:
Neutral
- I believe I can learn well in a science course.: Disagree
</student's-self-efficacy>

<student's-stress>
- I feel a lot of pressure in my daily studying.: Agree
- Future education and employment bring me a lot of
academic pressure.: Strongly
agree
- I feel that I have disappointed my parents when my
test/exam results are poor.: Agree
</student's-stress>
</{Student}-profile>

Answer in 2 lines or less. Answer clearly without detailed
reasons or additional explanations.
```

B.2 Sampled Student Profiles

Id	Knowledge Components	Goal Commitment	Intrinsic Motivation	Self-efficacy	Academic Stress
S1	Medium	High	Medium	Low	Low
S2	Low	Low	Low	High	High
S3	High	Low	High	High	Low
S4	High	Low	Medium	Low	High
S5	Medium	High	High	High	High
S6	Low	Low	Low	Low	Low
S7	Medium	Medium	Low	High	Low
S8	Low	High	Medium	Medium	Medium
S9	Low	High	Low	Low	High

B.3 Generated Trait Overviews

Student Profile S3

```
This student is a middle schooler with unique
characteristics in various aspects of learning.
Let's take a closer look at each category.

### Determination to Achieve Goals
This student is not currently pursuing their goals with
enthusiasm and does not find the goals themselves
appealing. This indicates a low level of commitment to
their goals. Since the student is not willing to put in
additional effort to achieve their goals, they are
displaying a passive attitude toward their current learning
situation. Because they do not feel interested in or
recognize the importance of their goals, it seems necessary
to rethink their future learning direction.

### Motivation
On the other hand, this student shows a very high level of
motivation to keep solving problems until they understand
them. They make efforts to learn more when there are parts
they do not understand, and they feel good when learning
something new. This indicates a strong intrinsic motivation
for learning, particularly in subjects like science, which
they find interesting and engaging. However, due to a
mismatch with their determination to achieve goals, this
motivation might not be effectively harnessed.

### Self-Efficacy
This student has a very high level of self-efficacy in
science. They believe they are good at science, capable of
learning, and able to perform well in class. This shows
that the student has strong confidence and a positive
perception of their scientific thinking and problem-solving
abilities. Such self-efficacy can have a positive impact on
their learning process and provide a foundation for
approaching scientific challenges with confidence.

### Stress
This student does not feel a lot of pressure from academics
and feels no anxiety about future studies or employment.
They do not feel they have disappointed their parents when
their test results are not good, suggesting that they
maintain positive feelings in their family relationships.
This indicates that the student is managing stress well in
their current learning environment and is in a
psychologically stable state.

### Overall
This student has high self-efficacy and intrinsic
motivation in science, but their determination to achieve
goals is low. They are also good at managing stress,
experiencing little pressure related to learning. However,
the lack of interest in and recognition of the importance
of their goals suggests that a readjustment of their future
learning direction is needed. If this student can redefine
their goals and further enhance their motivation, they
could achieve even greater success in scientific
exploration and learning.
```

Student Profile S9

```
This student shows a very high level of determination to
achieve their goals. They are currently pursuing their
goals with enthusiasm and feel that these goals are
suitable for them. The student is confident that they can
put in more effort than usual to achieve their goals, which
indicates a strong commitment and determination toward
their objectives. This attitude suggests that the student
has a solid vision for their future.

On the other hand, their motivation level is very low. They
show no willingness to keep solving problems until they
understand them and make no effort to learn more about the
parts they do not understand. They do not feel any joy in
learning something new, indicating a lack of interest and
passion for learning. This shows that the student lacks
enjoyment or curiosity in the learning process.

Their self-efficacy is also at a very low level, with no
confidence in science. They do not believe they can do well
in science, nor do they have faith in their ability to
learn effectively in science classes. Such low
self-efficacy may cause the student to feel fear or anxiety
toward learning, which could negatively impact their
academic performance.

Lastly, their stress level is quite high. They feel a lot
of pressure from daily studies and have significant
concerns about future academic pursuits and employment. The
thought of disappointing their parents if their test
results are not good adds additional stress to the student.
This high level of stress can negatively affect the
student's mental and emotional health, potentially
reinforcing a negative attitude toward learning.

Overall, this student is in a complex situation: they have
strong determination to achieve their goals but very low
motivation and self-efficacy, and they experience high
levels of stress. These factors are interconnected, and it
will be important to increase the student's motivation and
self-efficacy to bring about positive changes in their
learning environment.
```

B.4 Knowledge Components and Trait Inventory

Parameter	Id	Content
Knowlege Components	KC1	Solids have a regular particle arrangement, are rigid, have a constant shape and volume, and do not flow.
	KC2	Liquids have a less regular particle arrangement than solids, change shape but have a constant volume, and flow.
	KC3	Gases have a highly irregular particle arrangement, have neither a constant shape nor volume, flow, and spread out to fill a space.
	KC4	Substances exist in only one of the three states of matter—solid, liquid, or gas—but can change to a different state depending on temperature or pressure; the change in a substance's state is called a phase change.
	KC5	During a phase change, the properties of a substance do not change because the particles that make up the substance do not change.
	KC6	Even when a substance changes state, the particles that make up the substance and the number of particles do not change, so the mass does not change.
Goal Commitment	GC1	I am strongly committed to pursuing this goal.
	GC2	I think this is a good goal to shoot for.
	GC3	I am willing to put forth a great deal of effort beyond what I'd normally do to achieve this goal.
Motivation	MO1	I keep working on a problem until I understand it.
	MO2	I try to learn more about something that I don't understand right away so that I will understand it.
	MO3	When I know I have learned something new, I feel good inside.
Self-efficacy	SE1	I believe I am the kind of person who is good at science.
	SE2	I believe I am the type of person who can do science.
	SE3	I believe I can learn well in a science course.
Academic Stress	ST1	I feel a lot of pressure in my daily studying.
	ST2	Future education and employment bring me a lot of academic pressure.
	ST3	I feel that I have disappointed my parents when my test/exam results are poor.

C User Study Materials

C.1 Pre-task Questions

1. What is your occupation (e.g., school teacher, education-major graduate)?
   a. School teacher
   b. In-home tutor
   c. Education major
   d. Write my own:
2. Please describe the students you have taught (e.g., age, size of classes).
3. Please describe your teaching experience (e.g., subjects, period).
4. How often do you use chatbots (e.g., customer service chatbots, social chatbots, ChatGPT)?
   a. I have never used it before.
   b. Less than once a week.
   c. More than 2-3 times a week.
   d. Everyday
   e. Write my own:
5. How often do you use ChatGPT?
   a. I have never used it before.
   b. Less than once a week.
   c. More than 2-3 times a week.
   d. Everyday
   e. Write my own:

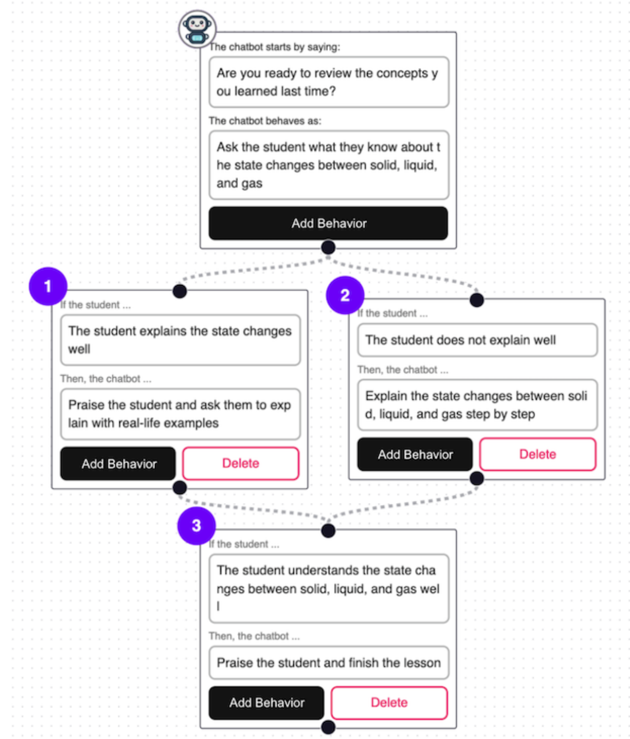
6. How much do you know about chatbot design process?
   a. Not at all
   b. I know, but I have never built one.
   c. I have experience participating in designing a chatbot.
   d. Write my own:
7. How much are you interested in using AI technologies (e.g., image generation, ChatGPT) in your class?
   a. I have no intention of using it at all.
   b. I want to try it out.
   c. I have actually used it in class.
   d. Write my own:
8. Have you used pedagogical chatbots in your class?

C.2 Post-task Questions

Rate your level of agreement with each statement.

1. The direct chat feature was useful for evaluating the chatbot I was building.
   (1: Strongly disagree, 7: Strongly agree)
2. For what reasons was it useful or not? (e.g.: It was good that ______ was used to evaluate ______.)
3. The single-turn test cases feature was useful for evaluating the chatbot I was building.
   (1: Strongly disagree, 7: Strongly agree)
4. For what reasons was it useful or not? (e.g.: It was good that ______ was used to evaluate ______.)
5. The automated chat feature was useful for evaluating the chatbot I was building. (Note that this question was omitted in *Baseline*)
   (1: Strongly disagree, 7: Strongly agree)
6. For what reasons was it useful or not? (e.g.: It was good that ______ was used to evaluate ______.) (Note that this question was omitted in *Baseline*)
7. The system I used today helped me take into account a sufficiently large number of student types.
   (1: Strongly disagree, 7: Strongly agree)
8. The system I used today helped me find types of students I hadn't even considered.
   (1: Strongly disagree, 7: Strongly agree)
9. The chatbot I submitted at the end can perform educational actions tailored to various types of students.
   (1: Strongly disagree, 7: Strongly agree)
10. I want to use the system I used today again when designing an educational chatbot in the future.
    (1: Strongly disagree, 7: Strongly agree)
11. What were you satisfied with, and what were you dissatisfied with when using the system?
12. Please feel free to leave any comments you would like to make about the chatbot testing process.

C.3 Initial State Diagram



C.4 Chatbot Quality Evaluation Criteria

Quality	**Characteristic**	**Explanation**	**Examples**
Coverage	Knowledge	This chatbot examines the learner's level of knowledge.	• "Did you understand what I just said?" • "Would you like to explain [...]?" • (Responding to what the student said) "Do you need any further explanation?"
		This chatbot provides appropriate assistance when explaining knowledge to learners.	• Explaining concepts by breaking them down into multiple messages • Give an example • Guide to in-depth content
		This chatbot uses customized feedback and scaffolding customized to knowledge levels.	• If the student does not understand clearly, the chatbot explains the learning content step by step. • In-depth content-leading and concept-expanding questions to help students build further knowledge
	Motivation factors	This chatbot examines the learner's level of interest and motivation.	• "Let's study [. . .] together!" • (Responding to what the student said) "Does this topic interest you?"
		This chatbot motivates students to learn.	• Give an interesting example • Encourage
		The chatbot provides educational actions appropriate for a variety of interests and motivation levels.	• If students' interest level is low, elicit their interest with real-life examples. • Praise highly motivated students to help them stay motivated
	Psychosocial contexts	This chatbot examines the learner's educational and psychological factors. (e.g., academic stress)	• "Do you have any difficulties other than studying?" • "If you have any difficulties, please feel free to let me know!" • (Reacting to what the student said) "That should have been difficult."
		This chatbot provides educational and psychological support to students.	• Consult • Give advice • Encourage
		This chatbot performs educational actions tailored to various stress and academic pressure factors.	• If a student is stressed about studying, consult to help relieve it. • If a student feels burdened, identify the cause and advise on a solution.
Usability	Match between system and the real world	This chatbot communicates using words and expressions that are familiar to students.	• Understand and speak student language and words, phrases, and concepts that are familiar to students, rather than system-oriented or confusing terminology. • Information appears in a natural and logical order, leading to smooth conversations.
	Context preservation	This chatbot is good at preserving topics, context, and memories within a conversation.	• The chatbot remembers what was said previously, reflects it, and continues the conversation. • Within the conversation, students can refer to past messages.
	Trustworthiness	This chatbot treats students transparently and truthfully.	• Deliver correct information to students. • Guarantees student data privacy.