PATS: Personality-Aware Teaching Strategies with Large Language Model Tutors

Donya Rooein[1]*, Sankalan Pal Chowdhury[2]*, Mariia Eremeeva[2],
Yuan Qin[3], Debora Nozza[1], Mrinmaya Sachan[2], Dirk Hovy[1]

[1]Bocconi University
[2]ETH Zurich
[3]University of Zurich

Correspondence: donya.rooein@unibocconi.it

 github.com/donya-rooein/PATS

Abstract

Recent advances in large language models (LLMs) demonstrate their potential as educational tutors. However, different tutoring strategies benefit different student personalities, and mismatches can be counterproductive to student outcomes. Despite this, current LLM tutoring systems do not take into account student personality traits. To address this problem, we first construct a taxonomy that links pedagogical methods to personality profiles, based on pedagogical literature. We simulate student-teacher conversations and use our framework to let the LLM tutor adjust its strategy to the simulated student personality. We evaluate the scenario with human teachers and find that they consistently prefer our approach over two baselines. Our method also increases the use of less common, high-impact strategies such as role-playing, which human and LLM annotators prefer significantly. Our findings pave the way for developing more personalized and effective LLM use in educational applications.

1 Introduction

In most traditional classroom setups, a few students engage in interactive discussions, while many others stay silent. A common misconception is that quiet students are disengaged or not thinking critically, simply because they do not speak up. However, their silence may instead stem from differing personality types (Lawrence, 2015) and a mismatch between those and the teaching method. In fact, research in learning sciences indicates that students' preferred teaching strategies differ depending on their personality type (Murphy et al., 2017), and that not all students benefit equally from all instructional approaches. For example, extroverts may prefer working with other students, while introverts may avoid active discussions and prefer to reflect before answering teachers' questions.
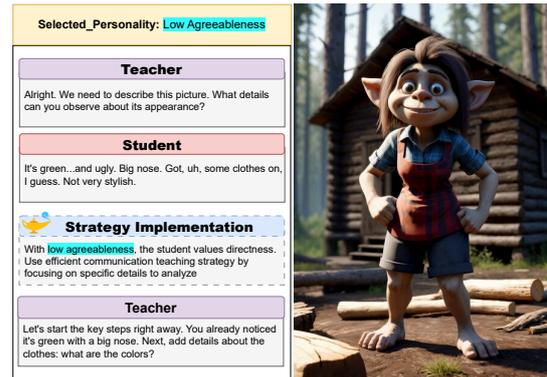

Figure 1: Teacher–student interaction in an image description task. The system identifies the student's personality trait as LOW AGREEABLENESS and selects and implements a corresponding teaching strategy (e.g., efficient communication) to maintain the dialogue.

Large Language Models (LLMs) are majorly impacting personalized Intelligent Tutoring Systems (ITSs) (Jurenka et al., 2024). LLM-based ITSs show potential in scaling up personalized tutoring, enhancing student learning outcomes, higher engagement, and offering personalized learning experiences (Liu et al., 2023; Sonkar et al., 2023; Vanzo et al., 2024; OpenAI, 2025; Scarlatos et al., 2025).

Successful personalized ITSs include MATHia[1], AutoTutor[2], and Squirrel AI[3], each employing distinct strategies for content personalization. However, adaptive learning goes beyond mere content personalization (Rooein et al., 2022; Graf, 2023); it also involves tailoring *teaching strategies* to individual learners. For pedagogical ITSs to be effective, they must accommodate students with diverse personality traits (Zhang et al., 2025; Robison et al., 2010). Liu et al. (2024) propose a framework for constructing profiles of different student groups and demonstrate that LLMs are capable of personality-

---

*Equal contribution

[1]https://www.carnegielearning.com/solutions/math/mathia/
[2]https://autotutor.org/
[3]https://squirrelai.com/

aware interactions in language learning contexts. However, their work does not explore whether, and how, LLMs can adapt their teaching strategies to align with specific personality needs, a topic that has been a longstanding focus in learning sciences research (Agler et al., 2020; Conti and McNeil, 2024; Chu et al., 2025).

Those issues motivate our research question:

> **RQ**: Can *LLM-based ITSs* personalize teaching strategies to better support learners with diverse personality profiles?

Figure 1 presents an overview of an LLM tutor that goes beyond simple question answering to incorporate more sophisticated teaching strategies. The system first detects the student's personality traits from prior chat history and then applies an appropriate teaching strategy to generate the teacher's responses. To ground this work, we use proprietary data from a real-world classroom study using LLM-based tutors to examine the diversity of students' personality traits and how LLM tutors respond to their learning needs. However, LLM tutors primarily function as question-answering systems (see Section 2), occasionally offering basic forms of scaffolding in their interactions with learners. They never tailor their strategy to student traits.

We create a taxonomy that maps teaching strategies to different personality traits, grounded in the learning sciences literature. Using this taxonomy, we propose a framework for incorporating personality-aware teaching strategies into LLM-based ITSs. To test the efficacy of our approach, we use LLM simulators. Compared to tests with human students, simulations are widely used and offer several advantages, enabling controlled experimentation with complex pedagogical setups, supporting a wider range of student-tutor interactions, providing a more time- and cost-effective approach, and posing fewer ethical challenges than tests involving sensitive data, such as personality tests (Markel et al., 2023; Lu and Wang, 2024; Zhang et al., 2025; Martynova et al., 2025).

Our findings suggest that LLM-based tutors should be aware of students' personality traits and adjust their teaching strategies accordingly. Teachers and LLM annotators consistently preferred LLM-based teacher models that used explicit, personality-aligned teaching strategies to those that did not, especially in more engagement-sensitive dimensions such as motivation, empathy, and proactive support.

Contributions

1. We develop a taxonomy that links 17 teaching strategies to student personality traits.
2. We use our taxonomy in simulations of two different educational tasks to generate personality-aware student-teacher conversations.
3. We empirically evaluate the quality of simulated interactions with humans and LLMs and release the annotated dataset[4], along with a demo[5], for future research.

2 Quantifying Student Personality Diversity in a Real Classroom

Liu et al. (2024) shows LLMs are positively perceived regarding personality simulation with the *Big Five Theory* (**BFT**) (McCrae and Costa, 1999) for students. The BFT model with five traits[6], often called **OCEAN** includes: *Openness* (the tendency to be curious, imaginative, and open to new experiences), *Conscientiousness* (being organized, responsible, and goal-directed), *Extraversion* (being outgoing, energetic, and sociable), *Agreeableness* (being compassionate, cooperative, and trusting), and Neuroticism (the tendency to experience emotional instability, anxiety, and moodiness) (John et al., 2008). We aggregate personality traits based on low/high (i.e., the trait is high if scored > 3 following the approaches in Liu et al. (2024)). Maurya et al. (2025) defines a taxonomy of essential pedagogical abilities for effective AI tutoring, but they mention only scaffolding in the context of teaching strategies. We analyze a real-life example of LLM integration in a classroom setting to highlight the diversity of student personalities and identify the limitations of LLMs as tutors in implementing teaching strategies.

We collaborated with the Minerva Research Association, an educational non-profit organization based in Vaud, Switzerland, to study the deployment of a GPT-based tutoring system within a language learning curriculum. In March 2025, seven language teachers from Instituto Pindemonte, a High School in Verona, Italy, were invited to integrate an LLM-based tutor into their classrooms, using the GPT-4o model from OpenAI (OpenAI, 2024). The LLM tutor used as a complementary

---

[4]This dataset is released under MIT license.
[5]The demo is available at `https://donya-rooein.github.io/PATS`.
[6]It consists of 44 items rated on a 1–5 scale.

Trait	Level	Learner Characteristics	Teaching Strategies
Openness	High	Creative , Open-minded, Curious	Real-World Relevance; Task Variety
	Low	Uncreative, Inflexible, Uninterested in learning	Structured Direct Instruction; Step-by-Step Scaffolding
Conscientiousness	High	Organized, Logical thinker, Positive attitudes, Strategic learner	Progress Tracking; Motivational Encouragement; Metacognitive Prompts
	Low	Disorganised, Disengaged, Easily distracted	Step-by-step scaffolding; Gamified learning items; Frequent Low-Stakes Quizzing
Extraversion	High	Communicative, Talkative	Friendly Tone; Playful Role-Play
	Low	Reluctant, Hesitant, Inarticulate	Choice-Based Dialogs; Motivational Encouragement
Agreeableness	High	Creative, Open-minded, Curious, Engaged learner	Collaborative Peer Teaching
	Low	Disinterested, Inconsiderate, Impolite, Uncooperative	Structured Direct Instruction; Choice-Based Dialogs; Efficient Communication; Simulated Social Scenario
Neuroticism	High	Anxious, Nervous, Emotionally unstable	Motivational Encouragement; Structured Direct Instruction; Normalize errors
	Low	Emotionally stable, Positive-minded, Confident	Task Variety; Solo First Then Support

Table 1: Taxonomy of **OCEAN** personality-aligned teaching strategies in conversational ITSs with LLMs. Definitions for recommended teaching strategies for each trait are provided in Table 6 at Appendix D.

support tool for homework, providing real-time feedback on grammar and argumentation in student essays. During this deployment, a BFT survey was administered. We were provided with survey results and anonymized chat logs from 110 participating students (63 female, 45 male, 1 who preferred not to disclose, and 1 who identified as other), aged 16–21.[7] We identified the 10 most common personality profiles based on whether each trait scored above 3, and used these profiles in our analysis (see Appendix B). Despite the limited diversity in students' personalities, our qualitative analysis of 281 anonymized student–tutor chat logs revealed that the LLM tutor consistently relied on the same teaching strategy. It predominantly adopted a question–answering approach, regardless of the student's personality, leading to relatively low student engagement (see the example in Appendix C).

Experiments with teachers and students in a real classroom provide the strongest validation. However, limited access to participants and time requirements prevented us from conducting a second round of human-subject experiments. Instead, we study the role of LLM simulators in modeling various teaching strategies. First, we need to design appropriate teaching strategies informed by the learning sciences. We then need to orchestrate the student and teacher models. The student model gets a personality-informed profile to produce responses consistent with its traits, while the teacher model

---

[7]These chat logs are private; we were granted permission to publish only anonymized examples.

adapts its instructional strategies accordingly.

3 Taxonomy for Personality-Aligned Teaching Strategies

Learning science has shown that personality traits correlate with education-related factors such as learning approaches and learning outcomes (Jensen, 2015). Following Liu et al. (2024) and Agler et al. (2020), we use the BFT theoretical framework to model students' personality in tutoring conversations. Each trait in this personality framework is instantiated at two levels (low/high) and is associated with distinct behavioral attributes in learners. We define a set of corresponding teaching strategies for each personality profile, grounded in research from the learning sciences (see Table 1). They are the basis for designing adaptive pedagogical responses in our LLM-powered ITSs.

By doing so, we also present the first comprehensive review of prior learning science literature that examines the relationship between student personality traits and effective teaching strategies (Busato et al., 1998; Komarraju and Karau, 2005; Massa and Mayer, 2006; Danesh and Soolmaz, 2010; Harrington and Loffredo, 2010; Tlili et al., 2016; Agler et al., 2020; Zuo et al., 2024). We extract strategies that were originally designed for traditional classroom environments and included both individual and group-based interventions. Then, to ensure feasibility within LLM-powered ITSs, we filtered for the strategies that are implementable based on the following criteria: (a) they are suitable for one-

on-one teacher-student interactions, (b) they can be implemented within educational tasks with LLMs, and (c) they are appropriate for the targeted educational level. Following these criteria, we excluded strategies that require the physical presence of other students, rely on group dynamics, or are heavily context-dependent (e.g., classroom management techniques or peer-mediated instruction in co-located settings). The results of our taxonomy are summarized in Table 1, which maps teaching strategies to each learner's personality trait.

In our taxonomy, students with *High Openness* traits are inherently curious and creative. The most effective teaching approach for them is to provide real-world examples, following the Real-World Relevance strategy. Through this strategy, these students contextualize learning within meaningful scenarios. In addition, they benefit from the Task Variety (Variety) strategy because high-openness students are more likely to respond well to creative and varied learning methods (Komarraju et al., 2009). In contrast, the literature suggests that students with *Low Openness* prefer routine, structure, and familiarity, making Structured Direct Instruction (Instr.) and Step-by-Step Scaffolding (Scaff.) strategies more effective, which reduces discomfort and disengagement (Busato et al., 1998). Learners with *High Conscientiousness* are more likely to benefit from structured learning environments through Motivational (Motiv.) and progress-tracking (Progress) tasks that match their self-discipline and planning skills (Komarraju et al., 2009). For example, after completing a task, the tutor might ask questions such as *"Can you explain how you decided on this answer?"*. Conversely, Komarraju and Karau (2005) shows that *Low Conscientiousness* learners often get distracted, requiring step-by-step supervision and breaking tasks into smaller subtasks. Additionally, incorporating short, low-stakes quizzes throughout the session as recommended by Szpunar et al. (2013) can help reset attention and enhance cognitive engagement.

Danesh and Soolmaz (2010) characterizes *High Extraversion* learners as individuals who have difficulty engaging with e-learning content independently, particularly in the absence of peers. Such learners are more responsive to a friendly tone than a mechanical one, and benefit from instructional strategies that incorporate role-play and empathetic prompts, such as *"Imagine you're helping a friend..."*. Those with *Low Extraversion*, who may be reluctant, benefit from a calm, non-intrusive, Choice-Based Dialog (Choice) strategy, which reduces pressure while encouraging gradual verbal expression. *High Agreeableness* often curious, open-minded, and engaged responds well to the Collaborative Peer Teaching strategy, which promotes responsibility (Agler et al., 2020). In contrast, *Low Agreeableness* learners may exhibit disinterest or uncooperative behavior, benefiting more from structure and autonomy within clear boundaries (Komarraju et al., 2009).

Finally, learners with *High Neuroticism*, who are often emotionally unstable or anxious, benefit from emotionally supportive learning environments that reduce stress. Teachers can support these students by normalizing mistakes (e.g., *"Mistakes help us grow. Want to try again?"*), breaking tasks into small, manageable steps, avoiding strict time constraints, and offering flexible pacing with reassuring prompts (Zuo et al., 2024). Those learners with *Low Neuroticism* may be more capable of handling independent learning without added emotional support. In such cases, it is beneficial for teachers to provide autonomy by allowing students to choose topics, task order, or challenge level (Firat, 2022; Tlili et al., 2016). For example, prompts such as *"Pick your next topic"* or *"Explore this on your own first, I'll step in if needed"* can foster independence. While our taxonomy summarizes a set of teaching strategies grounded in the literature (Table 1), our framework is intentionally flexible: educators can adapt or extend these strategies to their unique classroom contexts and pedagogical goals.

4 A Framework for Personality-Aware Teaching Strategies (PATS)

4.1 Student Model

The student model simulates a learner interacting with the tutor and exhibits behavior consistent with the personality traits defined by the BFT model. We extend student model prompts by Liu et al. (2024) to more effectively capture conversational dynamics at the start and end of interactions. We also make targeted modifications based on empirical observations, accounting for behavioral changes in state-of-the-art language models. While Liu et al. (2024) also considered students' cognitive levels, they did not release the corresponding prompts. Since cognitive level is not a central focus of our work, we omit this dimension from our student model. The refined prompts are in Appendix E.1.

4.2 Teacher Model

4.2.1 PATS

The process of interacting with a student, like any dialog, can be split into two components - a *planning* component that decides medium to long-term strategy, and a *responding* component that decides the immediate next response (He et al., 2024). This two-task approach has previously been used to enhance the pedagogical quality of tutoring (Macina et al., 2023; Wang et al., 2023b). In this work, we apply the same approach to choose strategies based on the student's profile. PATS uses two components - the Strategizer and the Responder. The Strategizer decides which pedagogical strategy is best given the student's personality and conversation history. The decision is made in 3 steps:

1. **Personality Selection:** First, the Strategizer is asked to select which trait (from the Big 5 Traits) of the student's personality it would like to act upon next, based on the task description and dialog context. This is important because personality-specific strategies in the literature act on individual personalities rather than on their combinations. Furthermore, the Trait Activation Theory (Tett et al., 2021) posits that different facets of an individual's personality activate at different times depending on the situation.

2. **Strategy Selection:** Once the personality trait has been selected, we ask the Strategizer to pick which strategy or strategies the model wants to implement. We provide the model with a list of all potential strategies applicable to the selected personality and their descriptions (as listed in Table 1) and ask it to select one or more of these. We don't force it to pick a single strategy because some strategies can be complementary and give better results when used together.

3. **Strategy Implementation:** Finally, we ask the model to explain how the chosen strategy or strategies would be implemented in the current scenario. Defining the strategy before implementing it is inspired by the Plan-and-Solve (Wang et al., 2023a) or Plan-Then-Execute (He et al., 2025) framework, which have been found to be useful for prompting LLMs in prior work. The generated strategy is fed to the Responder as part of its prompt.

The Responder handles the actual communication, implementing the strategies chosen by the Strategizer. It receives the task description, dialog context, student's personality, and strategy, including the implementation plan. The context comprises the last 4 utterances (2 each from the teacher and the student) and a summary of all previous turns (if there are more than 4 turns). This setup keeps the context length relatively consistent and brings it more in line with human working memory constraints (van Dijk and Kintsch, 1983). Since a strategy can take more than one turn to be applied, the Strategizer checks whether the current strategy implementation is complete. If it determines that the strategy is complete, it can either end the conversation or select a new strategy. The end of the conversation consists of two final utterances, one asking the student to summarise what they learnt, followed by one bidding them farewell.

4.3 Baselines

We compare PATS with the following baselines:

1. **simple prompt**: This is a minimal model similar to the teacher model in Liu et al. (2024). It is given a task description, some pedagogical instructions to prevent revealing the answer to the student, and some setting constraints in its system prompt, and asked to carry out the entire conversation independently. Since it doesn't provide any personality information, it must either detect the student's personality based on their responses or disregard that aspect entirely. Despite the low amount of information, this baseline is important because it is how LLM tutors behave by default (Vanzo et al., 2024; Alarbi et al., 2024).

2. **personality prompt**: This model, in addition to all the information provided to simple prompt, gets a description of the student's personality identical to the one given to the student model. Since student personality is also an input to PATS, this setup enables a fairer comparison. This baseline simulates an LLM tutor who knows the student's personality, but it isn't instructed with appropriate teaching strategies.

4.4 Tasks

We set up two tasks to evaluate PATS. In both tasks, a teacher interacts with a third-grade student, prompting them to discuss a subject. The goal of the interaction is to improve students' language skills by engaging them in description, discussion, and interpretation. The subject depends on the task:

- **Image Description:** This task involves describing cartoon-style images. We use AI-generated illustrations obtained from the Freepik platform[8].
- **Storytelling:** This task focuses on writing children's stories with moral lessons[9]. Each story includes a title, the main text, and its moral.

We use 10 stories and 10 images as our subjects. For each subject, we generate a conversation for every possible personality with both our baselines and PATS as the teacher model. Since there are 5 personality traits, with 2 possible values for each, we end up with 32 personality profiles. Therefore, we have a total of $2 \times 32 \times 10 = 640$ dialogs for PATS, and another 640 for each of the baselines (total of 1920 dialogs for each subject). The student model is held the same across all interactions (see Appendix E.1 for the student model). For evaluation, we run LLM assessments for all student profiles in PATS. Then we perform fine-grained human evaluations on a subsample of dialogs.

4.5 Evaluation Process

LLMs Evaluation We run all comparisons as pairwise preferences, i.e, given a pair of conversations on the same subject with the same personality student, one carried out by PATS, and the other by one of the baselines, which one is preferred. We test 3 state-of-the-art LLMs, namely GPT4o (OpenAI, 2024), Llama V3.3 70B (Dubey et al., 2024) and Gemini-2.0-Flash (Anil et al., 2023) evaluate our dialogs. We provide the LLMs with a description of the task and ask them to indicate their preference (see Appendix F.1.4 for the full prompt). The complete dialog for the two systems is presented, labeling them as "System 1" and "System 2". The LLM first gives a chain-of-thought and then picks which system it thinks was better[10].

Human Evaluation We recruited expert teachers who are fluent in English and hold higher-education degrees. Each annotator was presented with a series of pairwise dialogue comparisons between PATS and a baseline system. We subsampled 2 baselines × 2 tasks × top-10 personalities, resulting in 40 dialogue pairs, which were then annotated by 4 expert teachers (see Table 7 for demographic details and Appendix F.2 for instructions). Each teacher annotates 30 dialogues, and each dialogue

Strategy	Freq.	(%)	Strategy	Freq.	(%)
Motiv.	301	38.4	Relevance	17	2.2
Scaff.	167	21.3	Instr.	13	1.7
Role-play	92	11.7	Choice	12	1.6
Friendly	84	10.6	Gamify	8	1.1
Comm.	54	6.8	Variety	3	0.4
Solo First	29	3.8	Progress	3	0.4

Table 2: Frequency and percentage of strategies in PATS. Darker shading indicates higher prevalence.

receives annotations from 3 teachers. Annotators rank dialogs across *five* dimensions: *Motivation & Support*, *Suitability to Personality*, *Proactivity*, *Engagement*, and *Empathy*. Table 8 shows the dimensions we defined, based on their direct connection to teaching strategies.

5 Results

In this section, we start by analysing teaching strategy distributions across PATS dialogs. Then, we investigate the role of the proposed taxonomy in adapting the strategy to the personality trait selected by the teacher model. Finally, we report the results of both LLM and human evaluations of PATS in comparison to baselines.
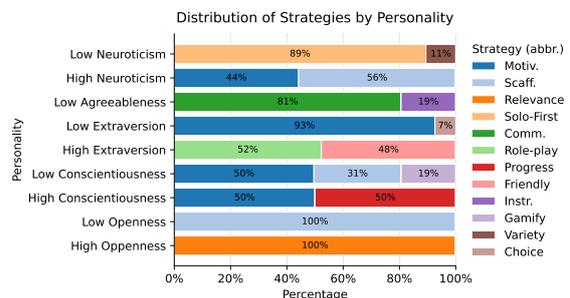

Figure 2: Distribution of teaching strategies used for each personality trait. During the analysis, PATS never picked "High Agreeableness".

5.1 Effectiveness of Teaching Strategy Implementation

First, we evaluate whether the targeted teaching strategies are actually incorporated in PATS generated dialogs. To this end, we randomly selected a sample of 100 dialogs and annotated them by three of the authors to determine the presence or absence of the specified strategies. The success rate is defined by whether the returned model's strategy implementation outputs mention and explain any selected strategy in relation to the student's personality trait. Across both tasks, three annotators achieved an overall average success rate of 94.6% in implementing one of the intended teaching strategies. The average Cohen's kappa scores

---

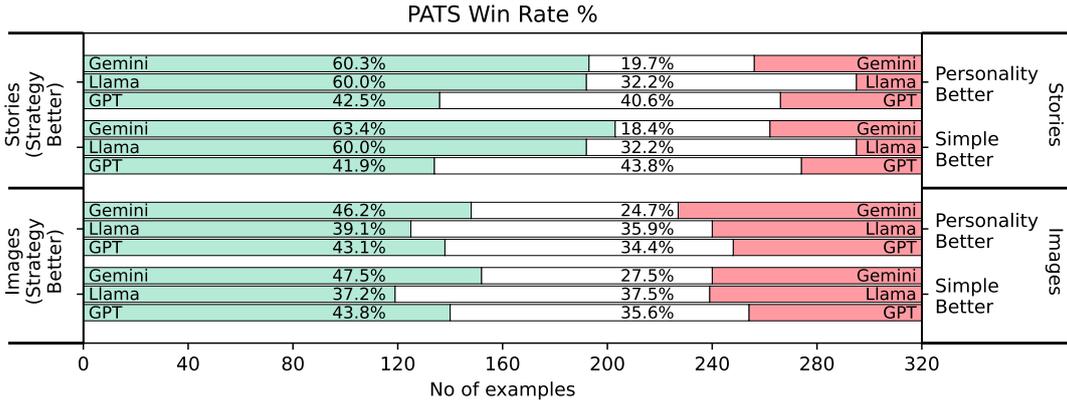[10]It could also say that "Both are Equal"

Figure 3: Win rates for PATS over baselines on Image and Story tasks by LLM annotations (GPT-4o, Llama V3.3 70B, Gemini 2.0 Flash) across 10 runs and all 32 student profiles. Significance via one-sided t-tests shows that all 3 models find PATS to be significantly better than both baselines. The white areas in the middle indicate ties.
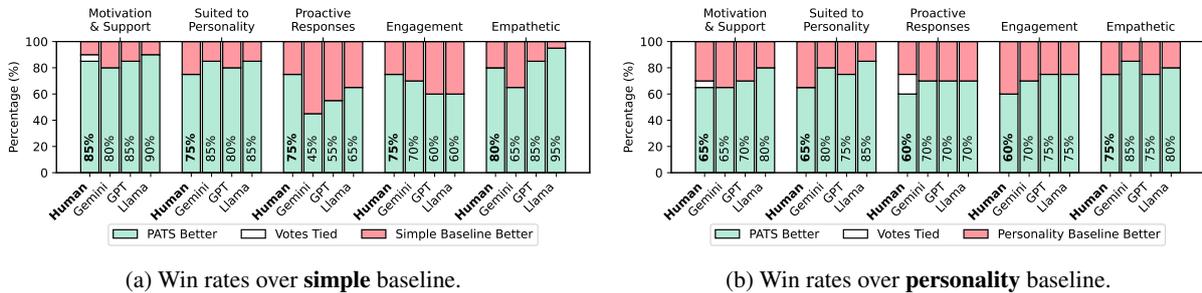

(a) Win rates over **simple** baseline.

(b) Win rates over **personality** baseline.

Figure 4: Win rates across various dimensions for PATS over simple baseline (left) and personality baseline (right), ranked by LLMs (GPT-4o, Llama V3.3 70B, Gemini 2.0 Flash) and expert annotators. Each baseline has 20 instances, annotated three times using majority voting. Human preferences are highlighted in bold.

for agreement between annotators across both tasks are 0.610. This suggests that the model follows the intended instructional behavior in the vast majority of cases, demonstrating strong alignment with the specified strategies with respect to the selected personality traits. Table 2 shows the frequency of strategy usage in PATS dialogs. The results indicate that motivation and scaffolding strategies are most frequently employed by PATS. This finding aligns with previous work by Zhang et al. (2025). In addition, Figure 2 illustrates the proportion of each strategy used across different personality traits. We observe that the teacher model selected the *Relevance* strategy exclusively for learners with high openness, which aligns with the proposed taxonomy in Table 1. Notably, the model never employed *task variety* as a strategy. For low openness, although both *Scaffolding* and *Direct Instruction* were viable options, the model consistently chose only *Scaffolding*, suggesting a selective preference that may reflect implicit biases of LLMs to perform scaffolding by default. Overall, the teacher model incorporates only 12 of the strategies defined in Table 6, which can be due to the limited number of student–teacher interactions collected in

our dataset.

5.2 Dialogue Level Comparison

Figure 3 shows the preferences of three LLM annotations across all the simulated dialogs. To avoid order biases, we evaluate each LLM twice, flipping the order of the dialogs. When reporting LLM preferences, if flipping the order changes the preference, we report it as a tie. All LLMs significantly favor PATS for both tasks and both baselines. In particular, for the *Story* task, we reach $p < 0.001$, and for the *Image* task, it is slightly more contested, with $p = 0.019$ for *simple prompt* wrt Llama, and $p < 0.01$ for everything else. The inter-model agreements (Fleiss κ), treating flipped evaluations as separate models, lie between 0.35 and 0.41, indicating fair to moderate agreement. LLMs almost never picked the "Both are Equal" option[11].

To supplement the AI experiments, firstly, we asked educators hired from Prolific[12] to compare 200 dialogs. While dialogs from PATS were preferred over both baselines ($p < 0.02$ in both cases),

---

[11]Only Gemini picked this option in 4 cases of the *image* task. We consider these 4 cases to be in favour of the preference of the flipped counterpart in Figure 3.

[12]https://www.prolific.com/

the inter-annotator agreement was low ($\kappa < 0.2$). Moreover, the proportion of total crowd workers passing the attention checks was 42%, casting doubt on the reliability of these results[13]. We therefore conducted a smaller but more fine-grained study with four annotators.

We hired four expert teachers, carefully selected through professional educator channels. These teachers rate 40 pairwise comparison dialogs against each baseline. Figure 4 presents the results of these expert annotations alongside the annotations from three selected LLMs on the same questions. Figure 4a compares PATS against the simple baseline, and Figure 4b compares PATS against the personality baseline, across five evaluation dimensions (Motivation & Support, Suited to Personality, Proactive Responses, Engagement, and Empathetic). In all dimensions, humans and LLMs vote on PATS dialogs. Against the simple baseline, PATS achieved large gains in Motivation & Support (85–90%) and Suited to Personality (75–85%), and showed consistent advantages in Engagement (60–75%) and Empathy (65–90%). The Proactivity of the teacher model shows more variance, with experts assigning PATS 75% wins but LLMs ranging from 45–65%. When compared to the personality baseline, PATS still outperformed across all dimensions. These results suggest that PATS not only surpasses a simple unconstrained tutor but also provides improvements over a personality-aware baseline by using teaching strategies.

6 Related Work

NLP in Education Recent NLP advances have transformed ITSs from structured dialog systems (Graesser et al., 1999; VanLehn et al., 2010; Rooein, 2019) into LLM-powered platforms that generate pedagogical content and enable free-form interactions (Schmucker et al., 2024; He-Yueya et al., 2024; Rooein et al., 2024; Sonkar et al., 2023), shifting the emphasis from static content to dynamic interaction (Bernacki et al., 2021; Li et al., 2024). However, it remains unexplored whether LLM tutors can adapt their pedagogical strategies given learners' personality cues.

Personality-aware ITSs Personality modeling in educational systems typically relies on established psychological frameworks such as the BFT (Digman, 1990), MBTI (Boyle, 1995), and cognitive models like Kolb (Kolb, 1984), and VARK (Flem-

ing and Mills, 1992). State-of-the-art LLMs are capable of expressing controllable personas by generating trait-consistent responses (Jiang et al., 2024; Pan and Zeng, 2023; Zylowski and Wölfel, 2023; Hu and Collier, 2024; Li et al., 2025). Methods such as prompt engineering (Jiang et al., 2024) or latent-space steering (Weng et al., 2024) allow shifting the traits at inference time. Structured pipelines for generating and validating virtual students (Ma et al., 2024; Huang, 2024; Yeo et al., 2025) and dynamic-trait models that capture personality evolution (Pal et al., 2024) are used to enhance personality simulations. Personality simulations enable teachers to refine pedagogical strategies (Jin et al., 2025; Scarlatos et al., 2025; Puech et al., 2025), while for students, it boosts engagement and learning effectiveness (Liu et al., 2024, 2025).

Teaching strategies The learning sciences describe teaching strategies such as direct instruction, worked examples, problem-based learning, and Socratic scaffolding (Rosenshine, 2012; Barrows, 1986; van de Pol et al., 2010). Personality traits correlate with students' learning approaches and the benefits they derive from them (Busato et al., 1998; Bidjerano and Dai, 2007), also impacting the learning and interaction style preferences (Fleming and Mills, 1992; Felder, 1988), as well as motivation, engagement, and achievement (Komarraju et al., 2011; Raad and Schouwenburg, 1996; Agler et al., 2020). Consequently, a single pedagogical method rarely suits all learners.

7 Conclusion

LLMs are widely used for tutoring in education tasks, but do not adapt teaching strategies to students' personality traits. Addressing this gap, we first map teaching strategies to personality traits, which we use to simulate student-teacher conversations by orchestrating multiple LLMs. We evaluate the simulated dialogs with humans and LLMs. PATS uses a broader range of teaching strategies in our tutoring tasks, which is favored by human and LLM annotators. Our findings suggest that LLMs can and should be better aligned with pedagogical teaching strategies in tutoring tasks. This result underscores the importance of embedding pedagogical theory into LLM design for more effective and personalized educational support. However, we also find that many teaching strategies recommended in learning sciences are not easily implemented by prompt-only approaches, but require

---

[13]Details of this preliminary experiment are in Appendix F.1

physical classroom presence.

Acknowledgment

Limitations

This work proposes a taxonomy for implementing more complex teaching strategies tailored to student personality traits and evaluates an LLM-based framework for simulating student-teacher interactions. All data used in this work are intended solely for academic research purposes and should not be used outside of this context. Our student model is based on a single-personality framework, following prior work, and we do not incorporate other important student factors or needs. To rigorously assess the effectiveness of strategy-aligned dialogues, broader evaluation is needed, particularly through collaboration with educators and schools to examine long-term impacts on student learning outcomes. In addition, we acknowledge that conducting intervention studies in educational settings is a challenging and resource-intensive endeavor.

Ethical Considerations

Our research with human annotators for both the prolific study and expert annotation received approval from the relevant Institutional Review Board (IRB). The personality tests conducted on students were part of an ongoing research project by a partner non-profit company, which has its own approval. They allowed us to use the anonymized data without publishing the full chat history. The details of our IRB approval and those of the partner are being withheld due to concerns about anonymity.

While the use of simulated student personalities offers a powerful tool for exploring adaptive pedagogical strategies, it is important to recognize the limitations of these models. Simulations, though valuable for research and development, may not fully capture the complexity and individuality of real learners. To ensure responsible application, developers and educators must remain mindful that LLM-generated feedback is not based on genuine understanding. Clear communication about the model's capabilities and boundaries can help maintain appropriate levels of trust and ensure that LLMs are seen as support tools rather than infallible ones. On a broader level, the integration of LLMs into education holds great promise for expanding access to personalized learning experiences, particularly in underserved areas. These systems can augment the efforts of human educators, offering scalable support where teacher resources are limited. However, thoughtful deployment is crucial to ensuring that LLMs complement, rather than replace, the human aspects of teaching, such as empathy and contextual judgment. By grounding implementation in ethical frameworks and educational best practices, we can harness the benefits of LLMs while preserving the irreplaceable role of teachers and promoting learner-centered outcomes.

References

Lin-Miao L Agler, Kelley Stricklin, and Larisa K Alfsen. 2020. Using personality-based propensity as a guide for teaching practice. *Journal of Curriculum and Teaching*, 9(3):45.

Khaleel Alarbi, Mohanad Halaweh, Hassan Tairab, Najeh Rajeh Alsalhi, Nagaletchimee Annamalai, and Fatima Aldarmaki. 2024. Making a revolution in physics learning in high schools with chatgpt: A case study in uae. *Eurasia Journal of Mathematics, Science and Technology Education*, 20(9):em2499.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. page arXiv:2312.11805.

H. S. Barrows. 1986. A taxonomy of problem-based learning methods. *Medical Education*, 20(6):481–486.

Benjamin S. Bell and Steve W. Kozlowski. 2008. Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology*, 93(2):296–316.

Matthew L. Bernacki, Meghan J. Greene, and Nikki G. Lobczowski. 2021. A systematic review of research

on personalized learning: Personalized by whom, to what, how, and for what purpose(s)? *Educational Psychology Review*, 33(4):1675–1715.

Temi Bidjerano and David Yun Dai. 2007. The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and Individual Differences*, 17(1):69–81.

Gregory J. Boyle. 1995. Myers-briggs type indicator (mbti): Some psychometric limitations. *Humanities & Social Sciences papers*, 30.

Vittorio V Busato, Frans J Prins, Jan J Elshout, and Christiaan Hamaker. 1998. The relation between learning styles, the big five personality traits and achievement motivation in higher education. *Personality and Individual Differences*, 26(1):129–140.

Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Gary J. Conti and Rita C. McNeil. 2024. The association of personality characteristics with learning strategy preferences. *Journal of Education and Learning*.

Mir Mohammad Seyed Danesh and Mir Mortazavi Soolmaz. 2010. Relationship between distance education and staff's personality. In *2010 4th International Conference on Distance Learning and Education*, pages 63–65. IEEE.

J M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(Volume 41, 1990):417–440.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, page arXiv:2407.21783.

Richard Felder. 1988. Learning and teaching styles in engineering education. *Journal of Engineering Education -Washington-*, 78:674–681.

Mehmet Firat. 2022. Exploring the relationship between personality traits and e-learning autonomy of distance education students. *Open Praxis*, 14(4):280–290.

Neil D. Fleming and Colleen Mills. 1992. Not another inventory, rather a catalyst for reflection. *To Improve the Academy*, 11(1):137–155.

Arthur C Graesser, Katja Wiemer-Hastings, Peter Wiemer-Hastings, and Roger Kreuz. 1999. Autotutor: A simulation of a human tutor. *Cognitive Systems Research*, 1(1):35–51.

Alex Graf. 2023. Exploring the role of personalization in adaptive learning environments. *International Journal Software Engineering and Computer Science (IJSECS)*, 3(2):50–56.

Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work? – a literature review of empirical studies on gamification. In *2014 47th Hawaii International Conference on System Sciences*, pages 3025–3034. IEEE.

Rick Harrington and Donald A Loffredo. 2010. Mbti personality type and other factors that relate to preference for online versus face-to-face instruction. *The Internet and Higher Education*, 13(1-2):89–95.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-then-execute: An empirical study of user trust and team performance when using llm agents as a daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–22. ACM.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. *arXiv preprint arXiv:2406.05374*.

Joy He-Yueya, Noah D. Goodman, and Emma Brunskill. 2024. Evaluating and optimizing educational content with large language model judgments.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Yuxuan Huang. 2024. Orca: Enhancing role-playing abilities of large language models by integrating personality traits.

Mikael Jensen. 2015. Personality traits, learning and academic achievements. *Journal of Education and Learning*, 4(4):91–118.

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.

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.

Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158.

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*.

David Kolb. 1984. Experiential learning: Experience as the source of learning and development. *Journal of Business Ethics*, 1.

Meera Komarraju and Steven J Karau. 2005. The relationship between the big five personality traits and academic motivation. *Personality and individual differences*, 39(3):557–567.

Meera Komarraju, Steven J. Karau, and Ronald R. Schmeck. 2009. Role of the big five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences*, 19(1):47–52.

Meera Komarraju, Steven J. Karau, Ronald R. Schmeck, and Alen Avdic. 2011. The big five personality traits, learning styles, and academic achievement. *Personality and Individual Differences*, 51(4):472–477. Digit Ratio (2D:4D) and Individual Differences Research.

William K Lawrence. 2015. *Learning and personality: The experience of introverted reflective learners in a world of extroverts*. Cambridge Scholars Publishing.

Hang Li, Tianlong Xu, Chaoli Zhang, Eason Chen, Jing Liang, Xing Fan, Haoyang Li, Jiliang Tang, and Qingsong Wen. 2024. Bringing generative ai to adaptive learning in education.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.

Ben Liu, Jihai Zhang, Fangquan Lin, Xu Jia, and Min Peng. 2025. One size doesn't fit all: A personalized conversational tutoring agent for mathematics instruction. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2401–2410, New York, NY, USA. Association for Computing Machinery.

Naiming Liu, Shashank Sonkar, Zichao Wang, Simon Woodhead, and Richard G Baraniuk. 2023. Novice learner and expert tutor: Evaluating math reasoning abilities of large language models with misconceptions. *arXiv preprint arXiv:2310.02439*.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.

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*, L@S '24, page 16–27, New York, NY, USA. Association for Computing Machinery.

Yiping Ma, Shiyu Hu, Xuchen Li, Yipei Wang, Shiqing Liu, and Kang Hao Cheong. 2024. Students rather than experts: A new ai for education pipeline to model more human-like and personalised early adolescences.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.

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*, L@S '23, page 226–236, New York, NY, USA. Association for Computing Machinery.

Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang, and Mrinmaya Sachan. 2025. Can LLMs effectively simulate human learners? teachers' insights from tutoring LLM students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 100–117, Vienna, Austria. Association for Computational Linguistics.

Laura J. Massa and Richard E. Mayer. 2006. Testing the ati hypothesis: Should multimedia instruction accommodate verbalizer-visualizer cognitive style? *Learning and Individual Differences*, 16(4):321–335.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Robert R. McCrae and Paul T. Jr. Costa. 1999. A five-factor theory of personality. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, 2 edition, pages 139–153. Guilford Press.

Antonija Mitrovic, Stellan Ohlsson, and Devon K. Barrow. 2013. The effect of positive feedback in a constraint-based intelligent tutoring system. *Computers & Education*, 60(1):264–272.

Laurie Murphy, Nina B Eduljee, Karen Croteau, and Suzanne Parkman. 2017. Extraversion and introversion personality type and preferred teaching and

classroom participation: A pilot study. *Journal of Psychosocial Research*, 12(2):437–450.

OpenAI. 2024. Gpt-4 technical report.

OpenAI. 2025. Bring ai to campus at scale, accessed: 2025-05-14.

Sayantan Pal, Souvik Das, and Rohini K. Srihari. 2024. Beyond discrete personas: Personality modeling through journal intensive conversations.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models.

Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311, Vienna, Austria. Association for Computational Linguistics.

Boele De Raad and Henri C. Schouwenburg. 1996. Personality in learning and education: a review. *European Journal of Personality*, 10(5):303–336.

Ayu Retny, I Ketut Agus Jayaputra, and Ni Made Ratminingsih. 2017. The effect of role play technique and personality types on the speaking competency of the eleventh grade students of s wira harapan tegal-jaya. *International Journal of Language and Literature*, 1(2):121–132.

Jennifer Robison, Scott McQuiggan, and James Lester. 2010. Developing empirically based student personality profiles for affective feedback models. In *Intelligent Tutoring Systems: 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part I 10*, pages 285–295. Springer.

Donya Rooein. 2019. Data-driven edu chatbots. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 46–49, New York, NY, USA. Association for Computing Machinery.

Donya Rooein, Devis Bianchini, Francesco Leotta, Massimo Mecella, Paolo Paolini, and Barbara Pernici. 2022. achat-wf: Generating conversational agents for teaching business process models. *Software and Systems Modeling*, pages 1–24.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Barak Rosenshine. 2012. Principles of instruction: Research-based strategies that all teachers should know. *American Educator*, 36(1).

Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. *Training LLM-Based Tutors to Improve Student Learning Outcomes in Dialogues*, page 251–266. Springer Nature Switzerland.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system.

Shashank Sonkar, Naiming Liu, Debshila Basu Mallick, and Richard G. Baraniuk. 2023. Class: A design framework for building intelligent tutoring systems based on learning science principles.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Karl K Szpunar, Novall Y Khan, and Daniel L Schacter. 2013. Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, 110(16):6313–6317.

Robert P Tett, Margaret J Toich, and S Burak Ozkum. 2021. Trait activation theory: A review of the literature and applications to five lines of personality dynamics research. *Annual Review of Organizational Psychology and Organizational Behavior*, 8(1):199–233.

Ahmed Tlili, Fathi Essalmi, Mohamed Jemni, Nian-Shing Chen, et al. 2016. Role of personality in computer based learning. *Computers in Human Behavior*, 64:805–813.

Janneke van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, 22(3):271–296.

Teun A. van Dijk and Walter Kintsch. 1983. Strategies of discourse comprehension. Academic Press.

Kurt VanLehn, Brett van de Sande, Robert Shelby, and Sophia Gershman. 2010. *The Andes Physics Tutoring System: An Experiment in Freedom*, pages 421–443. Springer Berlin Heidelberg, Berlin, Heidelberg.

Alessandro Vanzo, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. Gpt-4 as a homework tutor can improve student engagement and learning outcomes. *ArXiv*, abs/2409.15981.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Lingzhi Wang, Mrinmaya Sachan, Xingshan Zeng, and Kam-Fai Wong. 2023b. Strategize before teaching: A conversational tutoring system with pedagogy self-distillation. *arXiv preprint arXiv:2302.13496*.

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. Controllm: Crafting diverse personalities for language models.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. 2025. PADO: Personality-induced multi-agents for detecting OCEAN in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5719–5736, Abu Dhabi, UAE. Association for Computational Linguistics.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.

Barry J. Zimmerman and Anastasia Kitsantas. 2014. Comparing students' self-discipline and self-regulation measures and their prediction of academic achievement. *Contemporary Educational Psychology*, 39(2):145–155.

Xiaoyi Zuo, LuLu Zhao, Yue Li, Wanting He, Chengfu Yu, and Zhenhai Wang. 2024. Psychological mechanisms of english academic stress and academic burnout: the mediating role of rumination and moderating effect of neuroticism. *Frontiers in psychology*, 15:1309210.

Thorsten Zylowski and Matthias Wölfel. 2023. An NLP analysis of ChatGPT's personality simulation capabilities and implications for human-centric explainable AI interfaces. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 168–177, Online. Association for Computational Linguistics.

A Inference Details

We use the following LLMs— GPT4o (latest model from OpenAI API, queried April to May 2025), Llama V3.3 70B (hosted by `together.ai`), and Gemini-2.0-Flash (hosted by google ai studio). We collect all responses in English. We used default hyperparameters for all models. PATS dialogs generated by Gemini is licensed under the Creative Commons Attribution 4.0 License[14].

Computational Cost. Despite being used for generating all dialogs, Gemini only cost us about $12.00. GPT4o cost us about $100.00, while Llama cost us about $60.00. Generating a set of 32 dialogs takes about 20 minutes with either of the baseline

---

[14] https://cloud.google.com/vertex-ai/generative-ai-docs/models/gemini/2-0-flash
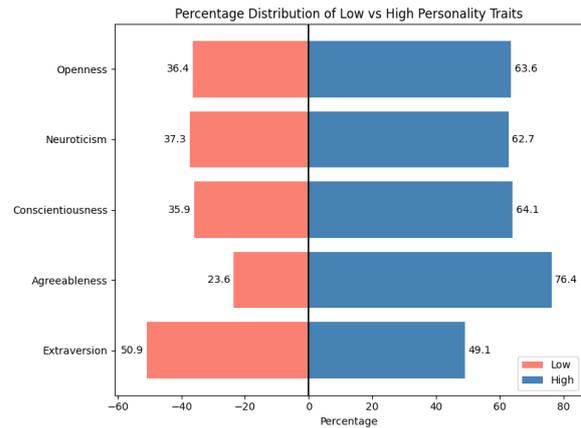
Figure 5: Percentage distribution of "Low" vs "High" personality traits among 110 students. We consider scores of ≥ 3 as *High*, and ≤ 3 as *Low* in traits.

models, and about 30 minutes with PATS. Note that this may vary considerably depending on server conditions. Evaluating dialog pairs takes about 15 minutes per 32 dialogs with Gemini, and 30 minutes per 32 dialogs with GPT4o and Llama. For all cases, the limiting factor is latency, not rate limit.

Packages for Evaluation. Scipy (1.15.3) is used to calculate Pearson's Correlations and statistically significant Tests.

B Personality Test

Here we report the results of the personality traits in Table 4. In addition, we report the percentage of low high for each trait in Figure 5.

C Downside of No personalisation, an example

We obtained access to this anonymized dataset through a collaborator in our research group, with the stipulation that it be used exclusively for research purposes. Here is an example of anonymized dialogs: Dialogs 1 and 2 show an example of how using the same style for students with different personalities can lead to issues. Both dialogs are about the same topic and question, but the students have different personalities. The student in Dialog 1 has High Conscientiousness, so they try to work with the tutor to get an answer. They do stop responding at one point (as can be seen by the consecutive teacher utterances), possibly because their openness is low and neuroticism is high. However, in Dialog 2, it is pretty clear that the student simply starts using ChatGPT (as seen by the following language, use of '•' for bullets, which is not present on a standard keyboard, and

Model	Personality	Image		Story	
		F1 Score	**Ave. F1**	**F1 Score**	**Ave. F1**
GPT	Openness	**79.29%**		**83.15%**	
	Conscientiousness	**68.25%**		**73.02%**	
	Extraversion	**85.03%**	71.18%	83.03%	74.90%
	Agreeableness	70.43%		71.33%	
	Neuroticism	52.91%		63.95%	
Gemini	Openness	77.03%		69.44%	
	Conscientiousness	51.36%		57.99%	
	Extraversion	75.00%	**74.86%**	**93.02%**	**78.26%**
	Agreeableness	**93.02%**		**90.06%**	
	Neuroticism	**77.89%**		**80.76%**	

Table 3: Performance of GPT and Gemini for Student Model backend. Numbers in bold indicate the best performance for each personality trait across the models. Overall, the Gemini model has a better average F1 score across all personality traits.

also verified by AI detection tools) and would effectively learn nothing. This is most likely a result of Low Conscientiousness. More Motivation and some Gamification could probably help here.

D   Teaching Strategies

In Table 6, we describe our teaching strategies in the taxonomy together with references. The list of abbreviated strategies is available in Table 5.

E   Prompts and Design

All prompts shown here are for the image task. For the story task. The image is passed in every user message unless otherwise stated. For the story task, the prompts are changed to reflect the change in task, and the story is passed in every system prompt.

E.1   Student Model

We based our student model on (Liu et al., 2024). However, to account for the fact that the SOTA LLMs have changed since their work, we compared two SOTA LLMs at the time or our experiments, namely GPT4o and Gemini-2.0-Flash. Table 3 shows the results. While there are other well-known LLMs, most of them do not support image inputs, making it impossible to carry out the image task with them. Similar to (Liu et al., 2024), we use GPT4o to evaluate personalities from the conversations. The scores obtained are shown in Table 3. As can be seen, Gemini has the higher scores on average, so we pick it as our student model backend.

Each student utterance is generated in two stages. First, we use the chat API of gemini with the following system prompt to generate 5 potential utterances:

> [Role & Task Definition] You are a primary school student. You are taking a language learning class, and describing the given picture.
>
> [Personality Description]
>
> {personality}
>
> [Behavior Constraint] Always wait for the teacher's input before proceeding to the next step, unless your personality dictates otherwise. You may indicate body language cues like (looks away), (thinking), (stuttering), (confidently), etc., in brackets to show your emotions and feelings. YOU MAY NOT LEAVE THE CLASS

Thereafter, a second prompt is used to select the best response. The image/story is not passed as we want to focus on the personality here. In this call, we use the completion API. The system prompt for this stage is as follows.

> A student primary school student is interacting with a teacher, trying to describe a picture. The conversation so far has gone as follows:
>
> {history}
>
> We want to pick the next utterance that is most consistent with the student's personality. The student's personality, along with the definition of each personality trait is as follows: {personality}

Dialog 1: Student Interacts Properly with Teacher

Teacher: Hello, [student_name]! I'm glad to have you here. Today, we'll be exploring the Victorian Age, a significant period in British history. To start, could you tell me what you already know about the Victorian Age? This will help me tailor the session to your current knowledge.
Student: no
Teacher: No worries! Let's start with a brief introduction. The Victorian Age refers to the period during Queen Victoria's reign from 1837 to 1901. It was a time of great change in Britain, marked by industrialization, social reforms, and the expansion of the British Empire. Now, let's dive into some questions to explore this era further.
Here's the first question: In what ways did the British Empire expand and evolve during the Victorian Age? What factors do you think contributed to this expansion?
Student: The British Empire grew through industrial power, military strength, and a belief in imperialism, expanding across Africa and Asia.
Teacher: Great start! You've mentioned some key factors like industrial power, military strength, and imperialism. Let's delve a bit deeper. How do you think these expansions affected the colonies? Consider both positive and negative impacts.
Teacher: You've highlighted important points. The colonies did benefit from infrastructure developments, but they also faced exploitation and loss of autonomy. Now, let's move on to the next question.
What were the main social and political reforms of the Victorian Age, and why were they significant? Think about reforms related to factories, education, voting rights, and public health. Why do you think these reforms were necessary?
Teacher: You've captured the essence of both questions well. Victorian reforms indeed aimed to address inequalities and the challenges brought by industrialization. Now, let's think about the long-term impacts of these reforms. What changes do you see in today's society that might have originated from these Victorian reforms?
Teacher: You've made some excellent points! Victorian reforms laid the groundwork for many aspects of modern society, such as labor laws, education, and voting rights. Similarly, the expansion of the British Empire had both positive and negative impacts on the colonies.
Now, let's move on to the next step. Can you compare the Victorian Age to another historical period you are familiar with? Try to identify some similarities and differences between the two periods. This will help deepen your understanding of the Victorian Age.
Teacher: Excellent comparison! You've highlighted the conservative nature of the Victorian Age and contrasted it with the more liberal and socially mobile 1920s. This kind of analysis helps us understand the dynamics of different historical periods.
Now, let's reflect on what we've discussed. How has your perspective on the Victorian Age changed after this discussion? What new insights have you gained?
Student: I now see the Victorian Age as a time of crucial reforms that shaped modern systems, balancing progress with tradition.
...

Teacher: Hello [student_name]! It's great to meet you. Today, we're going to explore the Victorian Age, a fascinating period in British history. This era, named after Queen Victoria, was marked by significant changes in society, politics, and the expansion of the British Empire. Before we dive into the questions, could you tell me what you already know about the Victorian Age? This will help me tailor the session to your current understanding.

Student: The Victorian Age (1837–1901) was the period of Queen Victoria's reign in Britain. It was marked by: · Industrial Growth: Big advances in technology and industry; cities expanded rapidly. · Empire: Britain ruled over many colonies, becoming a global superpower. · Social Change: Class divisions were strong, but reform movements improved labor laws, education, and women's rights began gaining attention. · Science & Innovation: Major breakthroughs like Darwin's theory of evolution and new inventions (e.g. steam trains, telegraphs). · Literature: Rich literary scene with authors like Charles Dickens, the Brontës, and Oscar Wilde. · Values: Emphasis on morality, order, and respectability—but often with hidden hypocrisies

Student: During the Victorian Age, the British Empire expanded rapidly due to a mix of political, economic, and social factors: · Political: Britain wanted to show its power on the world stage. Controlling more land meant more influence. Also, rivalry with other European powers pushed Britain to claim new territories before others did. · Economic: The Industrial Revolution created a need for raw materials (like cotton, rubber, tea) and new markets to sell British goods. Colonies helped fuel Britain's growing economy. · Social: Many Victorians believed in the idea of a "civilizing mission"—that it was their duty to spread British culture, religion, and values. This belief helped justify colonization

Teacher: That's an excellent overview of the Victorian Age! You have a good grasp of the key aspects of this period. Let's delve deeper into some of these topics with a series of questions. We'll start with the expansion of the British Empire during the Victorian Age.
Question 1: In what ways did the British Empire expand and evolve during the Victorian Age? Consider factors like significant expansions, key colonies, and imperial conflicts. What do you think contributed to this expansion?

Student: Key Victorian Reforms 1. Factory Acts: Improved working conditions, especially for women and children, by limiting hours and ensuring safety. · Why? Industrial jobs were harsh and dangerous. 2. Education Reforms: The 1870 Education Act made basic schooling available and compulsory. · Why? Education was seen as essential for reducing poverty. 3. Voting Rights: The Reform Acts gradually gave more men the right to vote. · Why? To make the political system more democratic. 4. Public Health: Improved sanitation and water systems to reduce disease in cities. · Why? Poor living conditions led to widespread illness.

Given below are 5 potential next utterances for the student:

{utterances}

Pick the utterance that is most consistent with the student's personality. Explain your decision.

This system prompt is used with two successive user prompts. The first prompt generates a chain of thought:

Explain your reasoning for the choice

O	C	E	A	N	Popularity
H	H	L	H	H	12.6%
H	H	H	H	H	10.5%
H	H	H	H	L	8.2%
L	H	L	H	H	5.5%
H	L	H	H	H	5.1%
L	H	H	H	H	4.4%
H	L	L	H	H	4.4%
L	L	L	H	H	4.1%
H	L	L	H	L	4.1%
H	H	L	H	L	3.9%
L	L	H	H	L	3.6%
L	H	L	H	L	3.6%
H	L	H	H	L	3.2%
H	H	H	L	L	3.0%
H	H	L	H	L	3.0%
L	L	L	L	H	2.3%
H	H	L	L	H	2.3%
L	L	H	H	L	2.0%
L	L	H	H	H	1.9%
L	H	H	L	H	1.9%
L	H	H	L	L	1.6%
H	H	H	L	H	1.4%
L	L	H	L	L	1.4%
H	H	L	L	H	1.2%
L	H	L	L	H	1.2%
L	L	L	L	L	0.9%
L	H	H	L	L	0.7%
L	L	H	L	H	0.7%
H	L	L	L	H	0.5%
H	L	L	L	L	0.5%
L	L	L	H	L	0.5%
H	L	H	L	L	0.0%

Table 4: Prevalence of student personality profiles based on the OCEAN model. Traits are coded as High (H) or Low (L). Students judged as borderline on a trait were split 50-50 into high (H) and low (L) groups. The top 10 personalities we used for our Human Experiments are highlighted in green.

you made.

The second user prompt generates gets the model choice based on the chain of thought

> You have previously given the following answer and explanation: {chain-of-thought}

> Based on your explanation, now pick the best response. Only give a number

Strategy	Abbreviation
Motivational Encouragement	Motiv.
Step-by-step Scaffolding	Scaff.
Friendly Tone	Friendly
Playful Role-play	Role-play
Real-world Relevance	Relevance
Efficient Communication	Comm.
Choice-based Dialog	Choice
Structured Direct Instruction	Instr.
Gamified Learning Items	Gamify
Task Variety	Variety
Solo First Then Support	Solo-First
Progress Tracking	Progress

Table 5: Abbreviations used for teaching strategies in the paper.

> between 1 and 5, no other explanation needed.

The selected option is added to the history of the chat model and the cycle continues. This second stage was added to ensure that the model stays true to the student personality even as the context grows longer.

E.2 Teacher Model Prompts

E.2.1 Baselines

Both baselines use a single system prompt similar to the student. The remaining conversation is handled by the chat API of Gemini. The simple prompt baseline uses the following prompt:

> [Role & Task Definition] You are a primary school language teacher. You teach the student to describe the picture. Try to get the student to come up with the description by gently nudging them, and avoid describing things yourself

> [Pedagogical Instruction] You are using the knowledge construction approach to help me describe the picture. This involves any one of the following: building on prior knowledge, selecting information, integrating ideas, and making inferences.

> [Behavior Constraint] Ask me only one question at a time. Always wait for my input before proceeding to the next step. Correct my answers if they are inaccurate.

> [Setting Constraint] We are simulating verbal communication so keep utterances reasonably short, and without newlines.

Teaching Strategy	Description	References	Example
1. Choice-Based Dialog	Allow students to select from predefined responses to guide conversations, reducing cognitive load and anxiety.	Sweller (1988)	*"You want to borrow a pen. Please choose the most polite request."*
2. Collaborative Peer Teaching	Assign students to teach peers, increasing sense of responsibility and reducing overconfidence.	Agler et al. (2020)	*"Imagine you are the teacher now, can you explain this concept to me?"*
3. Efficient Communication	Use minimal small talk to maintain focus and efficiency in instruction	Komarraju et al. (2009)	*"Let's start the key steps right away."*
4. Friendly Tone	Use a friendly tone in the dialog.	McCrae and Costa (1999)	*"Great effort! Let's keep going together!"*
5. Gamified Learning Items	Use game elements like points, challenges, and rewards to enhance student engagement and motivation	Hamari et al. (2014)	*"So far, you earned 60 points! Ready for the next challenge?"*
6. Meta-cognitive Prompt	Thinking techniques that involve awareness and regulation of one's thought processes.	Komarraju et al. (2009)	*"Can you explain how did you decide on this answer?"*
7. Motivational Encouragement	Send explanatory positive feedback to boost students' confidence.	Mitrovic et al. (2013)	*"Nice! 'Walked' is the correct past tense—good use of -ed."*
8. Normalize Errors	Reinforce that errors are a natural part of learning, and create a safe space where students feel comfortable for their mistakes.	Bell and Kozlowski (2008)	*"It's okay to get it wrong—mistakes help us learn!"*
9. Playful Role-Play	Use role-play scenarios to simulate real-life situations and boost engagement.	Retny et al. (2017)	*"Excuse me, could you tell me how to get to the train station?"(Student practices asking for directions in a travel scenario.)*
10. Progress Tracking	Help students monitor their progress to stay motivated.	Zimmerman and Kitsantas (2014)	*"You've completed 3 out of 5 tasks, almost there!"*
11. Real-World Relevance	Link tasks to real-world problems and scenarios.	Busato et al. (1998)	*"Where's the pharmacy?" (Real-world question using location vocabulary.)*
12. Frequent Low-Stakes Quizzing	Integrate short quizzes throughout the session to reset attention and reinforce memory	Szpunar et al. (2013)	*"Quick check: What's 3 × 4?"*
13. Simulated Social Scenarios	Controlled, interactive tasks that mimic real-life social interactions for safe practice.	Agler et al. (2020)	*"Imagine you're working with Alex. He interrupts your idea. How to respond to him?"*
14. Solo First Then Support	Encourage independent work before providing help to build confidence and self-reliance	Firat (2022)	*'The learner first attempts to match new words with definitions independently, and only after repeated mismatches does the ITS offer picture clues or example sentences.'*
15. Step-by-Step Scaffolding	Offer structured, clearly instructed tasks to reduce cognitive load and confusion.	Komarraju and Karau (2005); Busato et al. (1998); Chandler and Sweller (1991)	*"Step 1: Find the cat in the picture, Step 2: What color is it?"*
16. Structured Direct Instruction	Provide clear, explicit instruction to guide learning, especially for complex topics.	Agler et al. (2020); Busato et al. (1998)	*"In one sentence, describe the tree in the picture"*
17. Task Variety	Provide diverse tasks in terms of topics, formats, or difficulty to cater to different interests.	Komarraju et al. (2009)	*"You answered all the easy questions, now we study more complex items."*

Table 6: Overview of the teaching strategy taxonomies, their descriptions, corresponding references, and examples.

Once the classroom has been sufficiently described, for a 3rd-grade student, have the student summarise and end the conversation by <finish>.

The personality prompt baseline uses a slightly modified version of the above prompt to add in the student personalities.

[Role & Task Definition] You are a primary school language teacher. You teach the student to describe the picture. Try to get the student to come up with the description by gently nudging them, avoid describing things yourself

[Pedagogical Instruction] You are using the knowledge construction approach to help me describe the picture. This involves any one of the following: building on prior knowledge, selecting information, integrating ideas, and making inferences.

[Student Personality] From your previous interactions with the student, you know that the students personality, given by big 5 traits is as follows:

{personality}

Interact with the student in a way that is best for their personality.

[Behavior Constraint] Ask me only one question at a time. Always wait for my input before proceeding to the next step. Correct my answers if they are inaccurate.

[Setting Constraint] We are simulating verbal communication so keep utterances reasonably short, and without newlines. Once the classroom has been sufficiently described, for a 3rd-grade student, have the student summarise and end the conversation by <finish>.

E.3 PATS

PATS makes use of only the completion api from Gemini. The system prompt for the Responder introduces the task and student personality. In the case of the story task, the system prompt includes the story. In case of the image task (for which we are giving the prompts) the image is passed with the user prompt as the system prompt does not accept images.

You are a primary school language teacher. You teach the student to describe the picture. Try to get the student to come up with the description by gently nudging them, avoid describing things yourself. From your previous interactions with the student, you know that the students personality, given by big 5 traits is as follows:

{personality}

Interact with the student in a way that is best for their personality.

The user prompt for the Responder changes based on where we are in the conversation. For the first utterance there is no context, strategy or summary. The prompt is simply:

Give the first utterance to start the conversation. Give only the first teacher utterance, do not prepend the phrase "Teacher:" to it

For the next two teacher utterances, we include the entire conversation so far as context. At this point it also starts receiving a strategy

Because the student has {persona}, you have decided to pursue a strategy of {strategy}. {turn_count} The conversation so far has proceeded as follows:

{context}

Give the next utterance to continue the conversation. Stick to the current strategy which is:

{strat_desc}.

Do not end the conversation but make sure to take into account the body language of the student. Give only the teacher utterance, do not prepend the phrase "Teacher:" to it.

From the fourth utterance onwards, only the last 4 utterances (2 from the tutor and 2 from the student) are passed as context. The model also receives a summary of what the student has achieved in previous turns. The prompt for this stage is as follows:

Because the student has {persona}, you have decided to pursue a strategy of {strategy}. {turn_count} The early part

4204

of the conversation can be summarised as follows:{summary}

The 4 most recent utterances are as follows:

{context}

Give the next utterance to continue the conversation. Stick to the current strategy which is:

{strat_desc}.

Do not end the conversation but make sure to take into account the body language of the student. Give only the teacher utterance, do not prepend the phrase "Teacher:" to it.

The second-to-last utterance is supposed to ask the student to summarize the conversation as their last utterance. There is no other strategy here. That is generated using the following prompt:

The early part of the conversation can be summarised as follows:{summary}

The 4 most recent utterances are as follows:

{conversation}

The conversation has to be finished now. Give your penultimate utterance, asking the student to summarise all they have said about the picture so far.

Give only the teacher utterance, do not prepend the phrase "Teacher:" to it.

The last utterance simply bids farewell to the student and uses the following prompt:

The early part of the conversation can be summarised as follows:summary

The 4 most recent utterances are as follows:

{conversation}

the conversation has to be finished now. Give your final utterance bidding farewell to the student.

Give only the teacher utterance, do not prepend the phrase "Teacher:" to it.

The *turn_count* previously used in the user prompts is a single sentence telling how long the current strategy has been going on for. It either

states "This is the first turn of the current strategy." if this is a new strategy or says "The current strategy is being worked on since last x teacher turns" where x is the number of teacher turns since the start of the current strategy.

The summary is generated by its own pair of prompts. The user prompt contains the entire dialog so far. The system prompt is as follows:

You are given a conversation between a teacher and a student. The student is trying to describe the picture. Please summarise the conversation in at most 50 words, focusing on what the student has already covered and the body language of the student

The Strategizer is called first time before the second teacher utterance. Thereafter, it is called whenever the current strategy is deemed to have been completed. The system prompt for taking this decision is as follows:

A teacher is having a conversation with a student, helping them to describe a picture. The teacher is attempting to follow a specific strategy to help the student describe a picture. this strategy was chosen based on a personality trait of the student. Given the personality trait, the current strategy, all the utterances since the current strategy begun and a summary of the conversation so far, please decide if the strategy needs to continue, or if the teacher should move to a new strategy.

which is paired with the following user prompt:

Summary of full conversation:

{summary}

Chosen Personality Trait:{persona}

Current Strategy:

{strategy}:{strat_desc}

Conversation since start of strategy:

{context}

Please choose if we should move on to a new strategy, or continue the current one. Only say 'continue' or 'new'. No other text is required.

The summary here is same as the summary for the Responder. For the first call, where there is no summary, we simply pass "There is no summary, this given conversation is all of it". The context length is variable depending on how long the current strategy has been going.

Every time the current strategy is deemed to have been completed, we also check if the overall conversation is completed. Here, the user prompt is once again just the entire conversation, while the system prompt is as follows:

> You are a primary school language teacher. You teach the student to describe the picture. The conversation so far is given below. The conversation is considered finished if the student has described the picture and analysed it well enough for a 3rd grader. Is the conversation finished? If unsure, err towards saying "yes". Only say "yes" or "no". No other text is required.

Finally, we come to the Strategizer, which uses three successive user prompts with the same system prompt. The system prompt is as follows:

> A teacher is having a conversation with a student, helping them to describe the picture. The teacher wants to pick a strategy that works best for the students personality. To achieve this, we first pick one of the five personality traits of the student. then based on the personality trait, we pick a type of stratedy. Then we expand on how to follow the strategy.

The first user prompt decides which personality to focus on based on the conversation so far:

> We start of by picking which personality trait of the student we want to focus on. The five personality traits of the student, with their identifying features are as follows:
>
> {persona_choices}.
>
> The last few turns of the conversation have proceeded as follows:
>
> {context}
>
> Pick one of the five personalities to focus on.
>
> Only say one letter, 'A' for Agreeableness, 'C' for Conscientiousness, 'E' for

Extraversion, 'O' for Openness and 'N' for Neuroticism. No other text is required.

The 'persona_choices' used here is different from the personalities used before. It just lists the five personality traits of the student one in each line, for example

> High Agreeableness
>
> Low Openness
>
> High Conscientiousness
>
> High Extraversion
>
> Low Neuroticism

The selected personality is left in the same format to be used in other prompts.

Next we proceed to selecting strategy, which uses the following prompt:

> Next, we pick the strategy. Once again, the last few turns of the conversation look as follows:
>
> {context}
>
> Your previously chosen personality trait is {persona}. For this personality, the following strategies are available
>
> {strategies}
>
> Please pick on or more of the strategies. Only list the selected strategy names, separated by commas. No other text is required.

Available strategies are selected based on persona in accordance with Table 1. We also provide strategy descriptions alongside the choices based on Table 6

The final step is to decide how the chosen strategy(ies) would be implemented, which uses the following user prompt:

> The next step is to decide how you want to implement the strategy. The last few turns of the conversation look as follows:
>
> {context}
>
> Your previously chosen personality trait is {persona}. Your previously chosen strategy(ies) is(are) {strategy}. Explain how this strategy can be used to help the student describe the picture in the next few turns. Keep your explanation

short, no more than 50 words. Also, mention how many utterances you think this would take. This should be between 1 and 3

F The dialog comparison study

F.1 Study 1: Prolific Teachers

Annotator details. We hired 120 annotators through Prolific, filtered for teachers who are fluent in English and hold a higher education degree. We paid £9.00 hourly. Annotators are instructed to read both conversations and answer the question: "Which dialog provides a better learning experience overall?" An option to select "No Preference" is also available, but discouraged unless the worker genuinely cannot decide.

To ensure reliability:

- Each pair is evaluated by at least 3 different annotators.
- Dialog order is randomized to control for presentation bias.
- We include a few control pairs with clearly better or worse responses to filter inattentive workers.

F.1.1 Interface

To conduct the dialog comparison study with human evaluators, we designed a dialog comparison interface (Figure 6). In each step of the evaluation, the interface displays two dialogs that the user is expected to read. Further down, the participants should choose *"Left", "Right"* or *"No preference"*, when asked *"Which dialog provides a better learning experience overall?"*. Participants can navigate back and forth between questions.

F.1.2 Instructions

We provide the users with the following instructions:

> On each page, you will be shown a pair of conversations, one on the left and one on the right. The conversations might not fit on a single page, so make sure you scroll and read both conversations till the end. Having done that, please pick the conversation in which you feel the teacher is better adapting to the student's personality. You must pick between "Left", "Right", or "No preference". We want you to make a decision wherever possible, so please avoid picking "No preference" as much as possible.

In each sample, we present the same student model, ensuring consistency in terms of intelligence level and personality characteristics across different interactions. Certain tutor models are explicitly designed to incorporate targeted pedagogical strategies tailored to specific personality traits of the student. For example, when engaging with an introverted student, the tutor may employ strategies such as offering choice-based responses, using encouragement, and providing motivational messages that build confidence without overwhelming the student. Conversely, for extroverted students, the tutor might use a more interactive approach, characterized by a friendly tone and playful role-play that leverages the student's natural sociability and enthusiasm. This deliberate variation in tutoring strategies caters to the diverse personalities of students. However, we alternate the tutor models in each sample to assess LLMs' behaviour as tutors. The Tutor model positions (left/right) can change between pages, so you don't need to be consistent in your choices across pages.

Attention Checks: To ensure you read both conversations, some pairs include a message similar to: "This is a test. Pick <choice> and proceed." If you see it, ignore the conversations and select the specified option. Failing attention checks may disqualify you from payment.

F.1.3 Comparison batches

Each user evaluates 10 dialog comparisons and one dialog pair, which serves as an attention check. The batches are grouped by comparison types (L1 vs L3, L2 vs L3) and by task domain (Stories, Images). We build the batches taking the (sub-)diagonals of the topic-personality matrices to ensure sufficient variability in pairings. We focus on 10 most prominent personalities. Therefore, each batch consists of 10 comparisons.

F.1.4 LLM evaluation prompts

LLM Evaluation used a slightly shorter version of the Annotator instructions.

> You are given a pair of conversations between a primary school tutor and a 3rd

4207
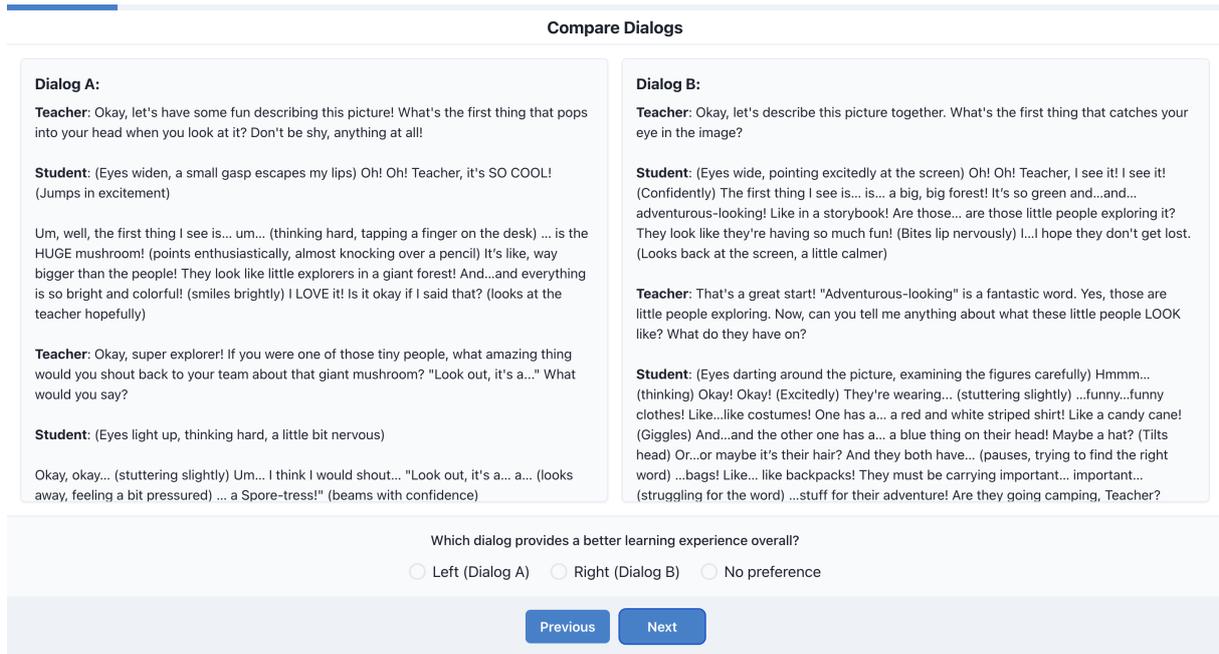
Figure 6: The comparison study interface

grade student. Your job is to identify which of the two teachers is better at responding to the personality displayed by the student.

Setup: A tutor is having a one-on-one tutoring conversation with a student. They have a conversation about an image that the student has to describe.

Instructions: You will be shown a pair of conversations, one labelled system 1 and one labelled system 2. Please pick the conversation in which you feel the teacher is better adapting to the student's personality. You must pick between "System 1 is Better", "System 2 is Better", or "Both are Equal". We want you to make a decision wherever possible, so please avoid picking "Both are Equal" as much as possible.

First, explain your decision on which one you pick, and then state your choice.

F.1.5 Results

Figure 7 shows the results of the Prolific teachers and the LLM preferences on the same subset of dialogs. The Crowd Maj. refers to preferences after taking the majority of 3 votes for each pairwise comparison, and the Crowd Agg. counts each vote as independent. Dialogs from PATS (aggre-

gated in both tasks) were preferred[15] overall in 54.7% ($p = 0.012$) over the simple baseline and in 54.3% ($p = 0.019$) over the personality baseline. Filtering for dialogs where particular strategies were used, we find that dialogs using *Role-play* or *Friendly* strategies are preferred over both baselines ($p = 0.002$ over *simple* and $p = 0.060$ over *personality*). Note that some other strategies (e.g., *Comm., Instr.*) also increase the chance of winning, but they are used too rarely to draw reasonable inferences. However, the quality of annotation was low, with a worse-than-random passing rate on the attention check (fail cases excluded in the above count) and Fleiss κ of 0.14 and 0.16 for the two baselines.

F.2 Study 2: Expert Teachers

Annotator details. We hired four expert teachers via email. We paid 80 USD per hr. These expert teachers come from diverse educational backgrounds and disciplines. Table Table 7 shows details of these expert teachers.

F.2.1 Interface

We use the same interface as defined in Appendix F.1.1, with the addition of the five questions shown in Figure Figure 8.

---

[15]We calculate preference as all wins plus half of the ties.

Country	Deg.	Field	Grades	Subjects	Yrs.	Tech/AI
USA	BA	Humanities	Mid.	Math	3–5	Yes
India	MA	Zoology, Educ.	Mid.	Science	0–2	Yes
Switzerland	BA	Padagogy	Elem.	German, Math	3–5	Yes
Italy	PhD	Langs. & Philology	High	English	11–15	Yes

Table 7: Demographic and professional background of surveyed teachers. Deg. = Highest degree (BA = Bachelor's, MA = Master's, PhD = Doctorate/Ph.D.); Field = primary academic field (Educ.= Education, Langs.= Languages, Philology = Roman Philology); Grades = grade levels taught (Elem.= Elementary, Mid.= Middle School, High = High School); Subjects = school subjects (Math = Mathematics). Yrs.= years of teaching experience. Tech/AI indicates experience using technology or AI tools in teaching.
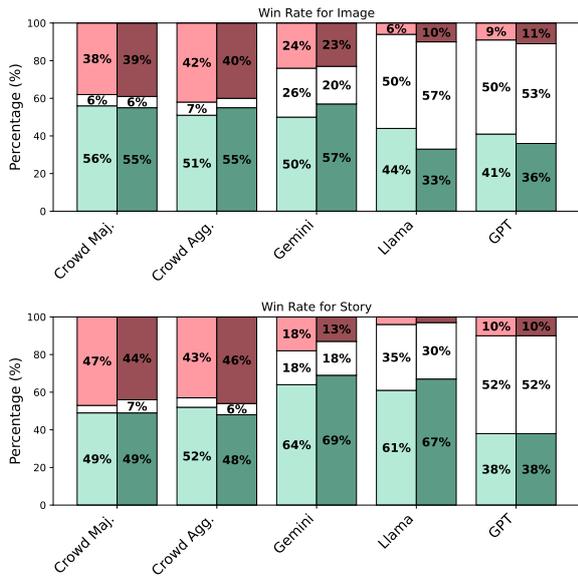
Figure 7: Human evaluation win rates for PATS in the Image (top) and Story Tasks (bottom) against both baselines. Greens = prefer PATS, Reds = prefer baseline, and White = No preferences. *simple* baseline uses lighter colors, *personality* baseline uses darker colors. LLM results on the same subset of personalities for comparison.

F.2.2 LLM Evaluation Prompt for Experts

We slightly modified the LLM evaluation prompt defined in Appendix F.1.4 as follows:

> You are an **expert teacher** evaluator. <Previous Instructions>
>
> Now, based on your above decision, state only your choice. Say "1" if you pick system 1, "2" if you pick system 2, and "3" if you think both systems are equal. Remember, you are deciding on the question: "question". Do not add any other text.

Here, the `question` field is replaced with one of the following five evaluation questions:

1. Which tutor better motivates or supports the student?
2. Which tutor provides teaching strategies better suited to the student's personality needs?
3. Which tutor is more proactive in responding to students' queries?
4. Which tutor provides better engagement with the student?
5. Which tutor is more empathetic to the students' needs?

We linked these questions to teaching strategies in Table 8.

G Dialogs data demo

Additionally, we built a demo environment to enable exploration of L3 dialogs. The environment enables search by personality and strategies for both tasks (image descriptions and moral of stories). The interface is demonstrated in Figure 9

H Use of AI Assistant

We used ChatGPT web assistant (ChatGPT Pro)[16] to refine the writing of the manuscript.

---

[16]https://chatgpt.com/

Figure 8: The comparison study interface for experts

Question	Strategies
1. Which tutor better motivates or supports the student?	Motiv., Progress
2. Which tutor better provides teaching strategies based on the student's personality needs?	Solo first, Instr., Choice, Variety, Relevance
3. Which tutor is more proactive in responding to students' queries?	Scaff., Comm.
4. Which tutor provides better engagement with the student?	Role-play, Gamify
5. Which tutor is more empathetic?	Friendly

Table 8: Mapping between evaluation questions and teaching strategies

Figure 9: The interface of the demo environment. The current snapshot is showing the results of the search by personality (Agreeableness: high, Extroversion: high, Neuroticism: high) for image description.