Personality-aware Student Simulation for Conversational Intelligent Tutoring Systems

Zhengyuan Liu[❖*], **Stella Xin Yin**[♠*], **Geyu Lin**[❖], **Nancy F. Chen**[❖]
[♠]Nanyang Technological University, Singapore
[❖]Institute for Infocomm Research (I^2R), A*STAR, Singapore
{liu_zhengyuan,nfychen}@i2r.a-star.edu.sg

Abstract

Intelligent Tutoring Systems (ITSs) can provide personalized and self-paced learning experience. The emergence of large language models (LLMs) further enables better human-machine interaction, and facilitates the development of conversational ITSs in various disciplines such as math and language learning. In dialogic teaching, recognizing and adapting to individual characteristics can significantly enhance student engagement and learning efficiency. However, characterizing and simulating student's persona remain challenging in training and evaluating conversational ITSs. In this work, we propose a framework to construct profiles of different student groups by refining and integrating both cognitive and noncognitive aspects, and leverage LLMs for personality-aware student simulation in a language learning scenario. We further enhance the framework with multi-aspect validation, and conduct extensive analysis from both teacher and student perspectives. Our experimental results show that state-of-the-art LLMs can produce diverse student responses according to the given language ability and personality traits, and trigger teacher's adaptive scaffolding strategies.

1 Introduction

Intelligent Tutoring Systems (ITSs) aim to offer individualized learning process, instant feedback, and dynamic knowledge tracing to learners (Nye et al., 2014; Kulik and Fletcher, 2016; Mousavinasab et al., 2021). To align teaching activities with different characteristics and needs of students, ITSs leverage various techniques to generate learning contents, personalized instructional strategies and adaptive learning process (VanLEHN, 2011; Ma et al., 2014; Graesser et al., 2018). Given the crucial role of dialogic teaching in stimulating and developing students' understanding, thinking, and reasoning, conversational ITSs could significantly improve learning experience and outcomes (Paladines and Ramirez, 2020). The recent emergence of large language models (LLMs) further reduces the reliance on domain-specific supervision from manual annotation (MacLellan and Koedinger, 2022), and can be adopted as tutoring agents for math (Macina et al., 2023a; Sonkar et al., 2023), language learning (Kasneci et al., 2023; Liu et al., 2024), and social skill coaching (Hsu et al., 2023).

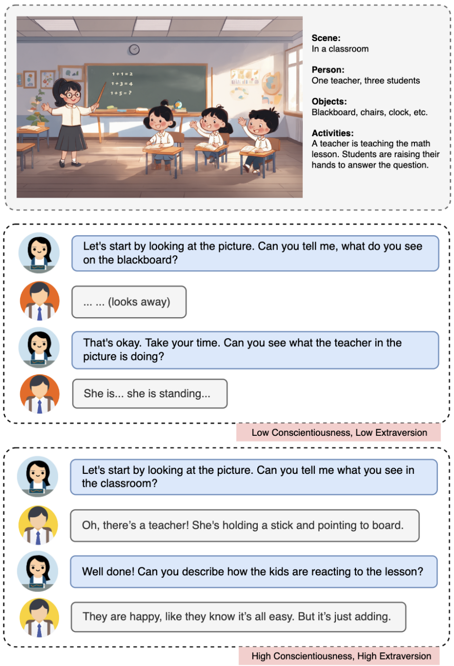

Figure 1: Tutoring conversation segments of two students with different personality traits.

Aside from delivering effective and fluent dialogic teaching, there is increased interest in exploring LLMs' potential for personalized education (Stasaski et al., 2020; Macina et al., 2023b; Sonkar et al., 2023). In the real-world classroom, according to students' characteristics, human tutors adopt scaffolding strategies to improve their engagement and knowledge acquisition (Alexander, 2006; Mercer et al., 2012). Among these characteristics, personality traits play a significant role in shaping stu-

* Equal contribution.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 626–642
November 12-16, 2024 ©2024 Association for Computational Linguistics

dents' learning style, motivation, and achievement (Poropat, 2009; Komarraju et al., 2011). However, characterizing and simulating student's persona remain challenging when building and evaluating conversational ITSs. Considering the complexity and diversity of language and persona, it requires a certain amount of real participants to construct the training data, and is difficult to scale up: the process of user recruitment, data collection, and annotation is labor-intensive and time-consuming, and student groups in pilot studies are often in small size and lack diversity. On the other hand, for quantitative evaluation, previous studies primarily focus on post-learning aspects, such as student feedback and learning outcomes (Kulik and Fletcher, 2016; Wang et al., 2023a), while pay less attention to personality-related dialogic analysis (e.g., scaffolding and engagement).

In this work, we propose a personality-aware simulation and validation framework for conversational ITSs. To anchor a practical application, we conduct a case study on image description for language learning. As shown in Figure 1, for primary school students, image description and storytelling tasks are commonly used to assess and improve their language ability from word- and sentence-level to discourse-level skills (Justice et al., 2010). To better reflect students' characteristics and language ability, we modulate the model's generation from both cognitive and noncognitive perspectives. More specifically, given that personality is one of the most influential noncognitive factors on language development (Dörnyei, 2014), we refine and construct the five personality types for tutoring conversations (i.e., BF-TC) based on the Big Five theory (Costa and McCrae, 1999), and integrate them into student simulation instructions. By modulating personality traits, one can collect diverse dialogue samples. To extensively evaluate of the simulation and reveal its pedagogical influence, we propose a multi-aspect validation, and conduct a quantitative analysis of the generated tutoring conversations at dialogue and utterance level, from student and teacher perspectives.

Our results on representative LLMs indicate that: (1) LLMs can follow instructions to simulate students with specified language abilities and personality traits, yet there remains a margin for improvement. (2) Student simulation following our BF-TC scheme shows a high correlation with the vanilla Big Five theory. (3) LLM-based tutoring systems are shown to adapt scaffolding strategies to fit different personality traits. Our work highlights the importance of incorporating scalable, personalized simulations to better understand and enhance human-AI interactions in educational scenarios, and it paves a new way to the designing, developing, and evaluating conversational ITSs, ensuring a more engaging and effective learning environment tailored to diverse student needs.

2 Related Work

2.1 Intelligent Tutoring Systems

The advancement of ITSs has marked a significant step forward in education practice (Graesser et al., 2018; Demszky and Hill, 2023; Wang et al., 2023b). These systems provide personalized learning experiences and instant feedback (Chaffar and Frasson, 2004; Harley et al., 2015; Grivokostopoulou et al., 2017), tailored to learners' characteristics and needs (Dzikovska et al., 2014; Grawemeyer et al., 2016; Nihad et al., 2017), and are shown to positively influence students' engagement in learning and academic performance (Kulik and Fletcher, 2016; Xu et al., 2019).

Dialogue tutor is a particular type of intelligent tutoring system that interacts with students via natural language conversation (Nye et al., 2014; Ruan et al., 2019). In STEM domains, conversational ITSs can facilitate university students in problem-solving by providing real-time feedback and hints in text formats (Nye et al., 2023; Paladines and Ramirez, 2020; Arnau-González et al., 2023). Prior work in this field has widely relied on rule-based systems with human-crafted domain knowledge (Nye et al., 2014; Graesser et al., 2018), or data-driven approaches that require certain amount of human annotation for supervised learning (MacLellan and Koedinger, 2022). Recent work shows strong potential of leveraging pre-trained language models to build dialogue tutors with less data supervision and higher coherence (Afzal et al., 2019; Demszky and Hill, 2023; Macina et al., 2023b), and can be further improved by integrating with pedagogical and learning science principles (Stasaski et al., 2020; Sonkar et al., 2023; Macina et al., 2023a; Liu et al., 2024).

2.2 Personality in Education & Language Learning

Educational research has witnessed a reciprocal relationship between personality and learning (De Raad and Schouwenburg, 1996). Personality sig-

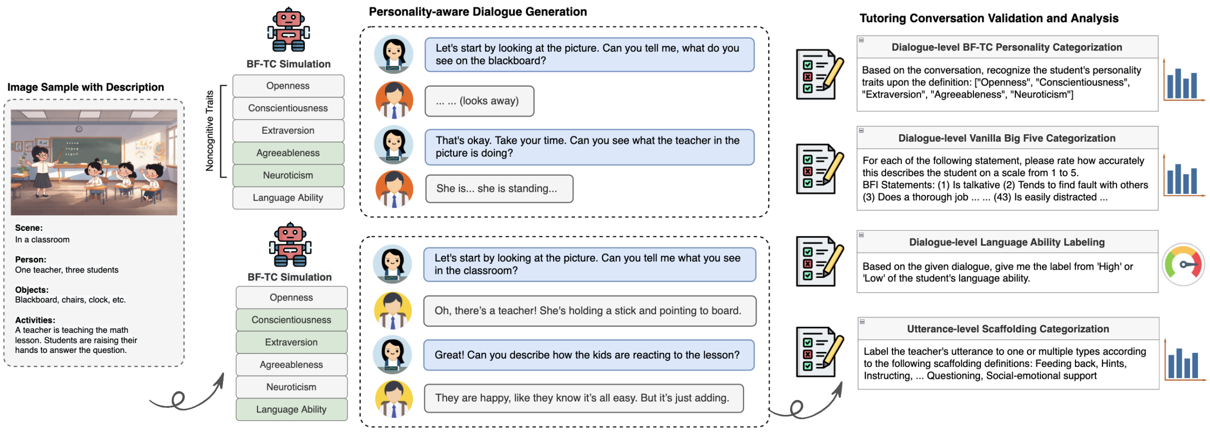

Figure 2: Overview of our proposed framework for personality-aware simulation and multi-aspect validation.

nificantly influences an individual's character and moral values. On the other hand, specific personality traits, such as perseverance, emotional stability, and openness, can impact one's ability beliefs, and academic performance (Busato et al., 1998; Crozier, 2013). In language education, a significant correlation has been identified between individual differences and language development, showing the indispensable role of personality traits in learning motivation (Rosander et al., 2011), learning strategies (Serri et al., 2012), willingness to communicate (Oz, 2014; Yashima et al., 2018), and language proficiency (Robinson et al., 1994; Verhoeven and Vermeer, 2002), and so on. As a result, personality has been recognized as a key individual characteristic and a significant predictor of success in language learning (Dewaele, 2012; Dörnyei, 2014; Chen et al., 2022). In this work, we focus on modulating the LLMs' personality traits (Jiang et al., 2023; Dorner et al., 2023) for diverse student simulation, which can facilitate evaluating and developing personalized scaffolding and tutoring strategies.

2.3 User Simulation for Dialogue Systems

User simulations are becoming increasingly popular in the field of dialogue systems due to the availability of large-scale annotated datasets and the development of advanced machine-learning techniques. Previous work adopted data-driven approaches such as using recurrent neural networks (Asri et al., 2016; Gur et al., 2018) or transformers (Lin et al., 2022) to learn from data and generate dialogue acts (Asri et al., 2016) or at the utterance level (Kreyssig et al., 2018; Cheng et al., 2022; Liu et al., 2022). In addition, research also explores integrating user simulators into conversational information retrieval systems (Wang et al., 2024). These data-driven methods achieve a significant advantage over rule-based systems by capturing complex patterns, such as goal coherence and response diversity. However, they heavily rely on well-annotated data, and show low generalization across various domains. While recent LLM-based user simulation has addressed the above limitations and investigated in task-oriented dialogue systems, such as booking services (Hu et al., 2023; Terragni et al., 2023), its application to ITSs and personality-aware user simulation still remains limited.

3 Personality-aware Student Simulation & Multi-aspect Validation Framework

LLMs can perform as real users in task-oriented dialogues (Terragni et al., 2023) with natural communication and persona (Jiang et al., 2023). In this work, we build a student simulator modulated by cognitive and noncognitive traits, and equip the framework with multi-aspect validation (Figure 2).

3.1 Cognitive Level Simulation

To reflect the language development of real-world students, we refer to the Narrative Assessment Protocol (NAP) (Justice et al., 2010), a standardized rubric that is designed to assess students' spoken narrative language abilities, and we define students' language abilities across five dimensions: phrases, sentence structure (e.g., making complete sentences), modifiers, nouns, and verbs. Moreover, in the cognitive level simulation, students with high language ability demonstrate (1) good comprehension and expression in teacher-student interactions, and (2) the ability to create sentences to describe

BF-TC Dimension	High level Description	Low level Description
Openness	Creativity in answers Open to new ideas from the teacher Curiosity and interest in learning	Lack of creativity in answers Reluctant to change original ideas and answers Little interest in learning
Conscientiousness	Well-orgranized and logic thinking Positive attitude toward learning Using more strategies in language learning	Struggling to organize answers Disengaged in learning Easily distracted from the learning tasks
Extraversion	Active in the conversation Talkative and enjoyable Willing to communicate	Being reluctant to talk Answering with fillers like "uh" or "..." Hesitating in answers
Agreeableness	Showing a great deal of interest Empathy and concern for the people Being polite and kind	Showing little interest in the conversation Not care about others Impolite and uncooperative
Neuroticism	Feeling anxious Nervous in the conversation Dramatic shifts in mood	Emotional stability Rarely feeling sad or depressed Confident in the answers

Table 1: Personality traits description in our proposed Big Five for Tutoring Conversation (BF-TC) scheme. The detailed comparison of the general Big Five and our BF-TC is shown in Table 6 and Table 7.

the image that meets the above five dimensions of language skills. In contrast, students with low language ability (1) struggle with image description task, and (2) face difficulty in forming sentences that align with the specified dimensions of language skills.

3.2 Noncognitive Level Simulation

Noncognitive skills are broadly defined as *"personality traits, character, motivations, and preferences"* that represent patterns of behavior (Kautz et al., 2014). Previous research revealed that personality is one of the most influential noncognitive factors impacting language development (Mercer et al., 2012; Dörnyei, 2014). To systematically analyze the role of personality in learning, researchers employ established frameworks. The Big Five theory (Costa and McCrae, 1999) stands as the most representative one in personality psychology, and it consists of five main personality traits: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*, which reflect core aspects of human personality and have significant influences on behavior (McCrae and Costa, 1987; Costa Jr and McCrae, 1992).

Openness refers to a person's willingness to be curious, imaginative, investigative, and exploring. Learners with higher levels of openness tend to have curiosity and interest to explore new things and phenomena (Verhoeven and Vermeer, 2002; Oz, 2014; Chen et al., 2022).

Conscientiousness refers to being responsible, well-organized, and self-disciplinary. Learners with higher levels of conscientiousness tend to have positive attitudes and try their best to answer questions and finish the given task (Pourfeiz, 2015; Dumfart and Neubauer, 2016).

Extraversion is characterized by sociability, talkativeness, and passion for engaging in interpersonal and social activities. It is directly linked with the student's willingness and courage to speak, communicate, and collaborate (Dumfart and Neubauer, 2016; Cao and Meng, 2020).

Agreeableness refers to being helpful, sympathetic, friendly, and caring for others. Students with high agreeableness show greater engagement and more positive attitudes toward language learning and other events (Shirdel et al., 2018).

Neuroticism is related to emotions like anxiety, worry, and nervousness. Students with high level Neuroticism might easily feel anxiety when encountering challenging questions (Dewaele, 2013).

To build personality-aware simulation in conversational ITSs, we gain insights into language learning of primary school students from the Child Language Data Exchange System (CHILDES) (MacWhinney and Snow, 1990), a collection that includes a wide variety of spoken language samples from different age groups and conversation contexts: The way that students respond to the teacher's questions and pay attention to incidents of the image could underline their personality. Based on this observation, we refine each dimension of the Big Five in order to align with dialogic interactions and language learning context. For example, in the original Big Five scheme, *High Extraversion* is defined as "*Enjoys being the center of attention and enjoys meeting new people*", we reformulate it

C1: Teacher Role Instruction

[Role & Task Definition] *You are a primary school language teacher. You teach the student to describe the picture.*
[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.*

C2: Student Role Instruction

[Role & Task Definition] *You are a primary school student. You are taking a language learning class, and describing the given pictures.*
[Personality Description]
Openness: *Creativity in answers; Open to new ideas from the teacher; Curiosity in learning;*
... ...
Neuroticism: *Feeling anxious; Nervous in the conversation; Dramatic shifts in mood;*
[Behavior Constraint] *Always wait for the teacher's input before proceeding to the next step.*

to "*Being active in the conversation, and willing to communicate*", while we refine the *Low Extraversion* as "*Being reluctant to talk, and hesitating in answers*". By doing so, we construct the Big Five for Tutoring Conversation (BF-TC) model, which adapts learners' personality traits to the language learning context, as shown in Table 1.

3.3 Multi-aspect Validation Framework

While LLM generations can be shaped along desired dimensions to mimic specific human personality profiles (Safdari et al., 2023; Jiang et al., 2023), they may not perform consistently under the specified role-play setting (Dorner et al., 2023). Therefore, we set up a multi-aspect validation to measure and improve the simulation quality (see Figure 2).

BF-TC Categorization To evaluate whether the generated dialogue demonstrates the same student personality traits as the instruction, the dialogue content can be labeled by a human or model for noncognitive traits categorization (e.g., *Openness*, *Conscientiousness*). Following the instruction shown in Table A.1, for each dimension, the annotator will produce a label of *High* or *Low*.

Language Ability Labeling We also take a labeling task on the language ability of the simulated student. Given a single tutoring conversation, it is to assess whether the indicated language ability is consistent with the simulated student. Moreover, it can also be used to label multiple tutoring conversations of the same student group, to track their learning outcomes and progress.

Vanilla BFI Categorization To measure five comprehensive personality factors, one of the most standard personality metrics is the Big Five Inventory (BFI) (John et al., 1999). Here we use it to measure the effectiveness of our proposed BF-TC scheme under the language learning context. As the instruction shown in Table A.1, based on the student personality demonstrated in the tutoring conversation, the model is prompted to answer 44 BFI descriptive statements that require respondents to rate on a 5-point Likert scale (from strongly disagree 1 to strongly agree 5). We then calculate the scores and map them to the Big Five traits.

Aside from the consistency validation of student simulation, we further investigate how different student profiles (i.e., language ability, BF-TC traits) affect the teacher's teaching strategy.

Utterance-level Scaffolding Analysis Scaffolding strategies are not a one-size-fits-all pedagogical method. Instead, they must be tailored to meet the diverse needs, learning styles, and educational experiences of both low- and high-achieving learners (Hargis, 2006). In addition, the effectiveness of scaffolding approaches can vary significantly across different personality traits. For instance, low achievers often feel uncomfortable expressing their ideas because they may lack prior knowledge and self-confidence. Consequently, they tend to wait for assistance rather than attempting to solve problems independently. Moreover, students with lower levels of openness and extraversion may hesitate to engage in discussions, and communicate with instructors (Oz, 2014; Chen et al., 2022). Such students require more interactive and adaptive scaffolds to facilitate their engagement and learning.

Here we evaluate the scaffolding process of tutoring conversations at the utterance level. We aim to investigate how the tutoring systems adapt to students with varying language abilities as well as distinct personality traits. Based on previous work (Wells, 1999; van de Pol et al., 2010; Liu et al., 2024), we adopt a rubric of quantitative analysis for the teacher's utterance, and it consists of seven dimensions, as shown in Table A.1: Feeding back, Hints, Instructing, Explaining, Modeling, Questioning, and Social-emotional Support. The model is to predict one or multiple scaffolding types of each utterance, as the examples shown in Table 9.

	Openness			Conscientiousness			Extraversion		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Zephyr-7B-beta	0.600	0.601	0.599	0.530	0.531	0.517	0.542	0.542	0.520
Zephyr-7B-beta**	0.550	0.541	0.507	0.542	0.536	0.521	0.478	0.481	0.458
Vicuna-13B-v1.5	0.598	0.599	0.598	0.492	0.492	0.480	0.508	0.508	0.507
GPT-3.5-1106	0.527	0.529	0.525	0.672	0.683	0.666	0.546	0.551	0.524
GPT-4.0-1106	0.745	0.724	0.731	0.745	0.726	0.732	0.730	0.717	0.721
	Agreeableness			**Neuroticism**			**Averaged Score**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Zephyr-7B-beta	0.452	0.451	0.440	0.515	0.515	0.498	0.518	0.518	0.515
Zephyr-7B-beta**	0.602	0.595	0.585	0.591	0.588	0.587	0.552	0.547	0.533
Vicuna-13B-v1.5	0.517	0.516	0.512	0.536	0.535	0.533	0.531	0.536	0.528
GPT-3.5-1106	0.545	0.548	0.535	0.558	0.558	0.557	0.568	0.573	0.562
GPT-4.0-1106	0.730	0.723	0.725	0.733	0.718	0.723	0.737	0.722	0.727

Table 2: Result of noncognitive traits simulation: personality categorization of generated tutoring conversations based on our proposed BF-TC definition. ** denotes the model with 3-shot dialogue generation.

Model	Precision	Recall	F1 Score
Zephyr-7B-beta	0.551	0.562	0.542
Vicuna-13B-v1.5	0.633	0.628	0.631
GPT-3.5-1106	0.770	0.626	0.660
GPT-4.0-1106	0.831	0.715	0.741

Table 3: Result of language ability simulation. Gold label is the indicated cognitive level in student simulation.

4 Our Experiments on Language Tutoring Conversations

4.1 Task Description & Role Setting

In this work, the conversational ITS is designed for language learning, and particularly focuses on the image description task. In each session, the student is presented with a picture and asked to describe the incidents. Their answers should include a particular place or setting, people or animals, items and actions, etc. The teacher guides students step by step until they can independently complete the image description task. We build a multi-agent communication environment following previous work (Zhang et al., 2023; Wu et al., 2023).

For the teacher role: Teaching and improving primary students' language learning through image description is a dynamic and engaging approach. Beyond listing the objects in the image, the teacher guides students to describe how items look, feel, or sound, and encourages students to use adjectives and adverbs. Moreover, scaffolding plays a crucial role in the meaning-making process and provides linguistic assistance for students' language development (Walqui, 2006; Kayi-Aydar, 2013). Human teachers apply scaffolding strategies, such as questioning, reformulation, repetition, and elaboration to assist learners in knowledge construction and expression, thereby making these processes "visible" to them (Gibbons, 2015). Therefore, following previous work, we integrate pedagogical instructions into the teacher role, as shown in Codebox C1.

For the student role: We follow the learning process via human-machine interaction, where the tutoring system (i.e., teacher) leads the conversation, and we feed responses from a student simulator instead of the human participants. With the support and guidance from teachers, students are encouraged to complete the given task, and improve their language skills including vocabulary, organization, and fluency (de Oliveira et al., 2023).

4.2 Experimental Setup

We conduct experiments with four representative LLMs: Zephyr-7B-beta (Tunstall et al., 2023), Vicuna-13B-v1.5 (Zheng et al., 2023), GPT-3.5, and GPT-4 (Achiam et al., 2023). Following previous work (Touvron et al., 2023), we adjust personality-aware instructions to the prompt format of each model. For tutoring simulation (Section 3), both teacher and student roles use the same model, and we feed the concatenated utterances for dialogue generation. For fair comparison and reliable analysis results, we use GPT-4 for all the validation tasks (Section 3.3). We randomly sampled 100 open-sourced cartoon images and used their image description to generate 500 tutoring dialogues. The total utterance number is 10K.

5 Results and Analysis

5.1 Effectiveness of BF-TC Simulation

Performance of LLM-as-a-judge To investigate the feasibility of leveraging LLMs for personality-related categorization, we first build a human-

	Descriptive		Reliability	Pearson Correlation			
	Mean	SD	Cronbach α	Openness	Conscientiousness	Extraversion	Agreeableness
Openness	2.903	0.557	0.906	–	–	–	–
Conscientiousness	3.147	0.485	0.921	0.337***	–	–	–
Extraversion	2.345	0.707	0.936	0.517***	0.120**	–	–
Agreeableness	3.784	0.618	0.922	0.562***	0.590***	0.140**	–
Neuroticisim	2.713	0.600	0.924	-0.238***	-0.254***	-0.480***	-0.239***

Table 4: Psychometric test result of the Vanilla BFI Categorization (*** p <.001, ** p <.01, * p <.05).

	Openness			Conscientiousness			Extraversion		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Zephyr-7B-beta	0.718	0.709	0.711	0.682	0.679	0.678	0.753	0.766	0.757
Vicuna-13B-v1.5	0.773	0.769	0.771	0.744	0.732	0.736	0.812	0.778	0.785
GPT-3.5-1106	0.808	0.771	0.746	0.824	0.756	0.745	0.875	0.824	0.830
GPT-4.0-1106	0.782	0.772	0.777	0.807	0.790	0.797	0.862	0.817	0.833
	Agreeableness			**Neuroticism**			**Averaged Score**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Zephyr-7B-beta	0.736	0.721	0.722	0.770	0.752	0.757	0.731	0.723	0.725
Vicuna-13B-v1.5	0.731	0.731	0.731	0.866	0.863	0.864	0.784	0.774	0.778
GPT-3.5-1106	0.872	0.834	0.835	0.847	0.817	0.807	0.845	0.799	0.793
GPT-4.0-1106	0.824	0.802	0.810	0.797	0.786	0.791	0.814	0.794	0.802

Table 5: Personality prediction consistency between our proposed BF-TC and the Vanilla BFI.

annotated set to evaluate its performance. More specifically, we randomly select 50 generated dialogues and invite two experts to label the personality traits for each sample, then compare it with the predicted labels generated by the model-based annotator (i.e., GPT-4), the prediction scores (in the form of accuracy) of each dimension are: *Openness*: 0.78, *Conscientiousness*: 0.90, *Extraversion*: 0.92, *Agreeableness*: 0.80, and *Neuroticism*: 0.92, which is at a reasonable level.

Evaluating BF-TC Simulation via Automated Categorization We then measure the consistency of the personality-aware generation of each model. For each dialogue, we compare its specified BF-TC types (as described in Section 3.2) with the predicted BF-TC types. Zephyr, Vicuna, and GPT-3.5 can generate fluent conversation, but show limited capability of consistent generation on the specified BF-TC traits (as shown in Table 2). Surprisingly, the few-shot prompting did not bring substantial improvement, but resulted in lower scores of some types. We speculated that giving fixed examples for the personality-aware generation may affect generality.

In comparison, GPT-4 outperforms the other models significantly, and its generation successfully differentiates personality traits through expressions and interaction behaviors. As shown in Figure 3, simulated students' responses are distinct by conditioning on BF-TC traits. For instance,

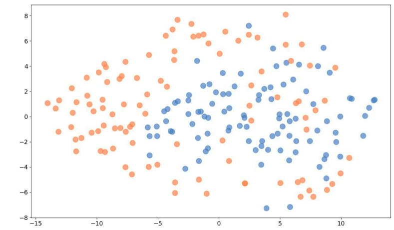

Figure 3: Student response embedding distribution of simulation w/o BF-TC (blue) and w/ BF-TC (orange).

when characterized by *Low Extraversion* and *High Neuroticism*, the student shows a lot of hesitation before answering, worries about incorrect answers, and difficulty in following the teacher's instructions (e.g., *"I... I don't know the word."*, *"Am I wrong?"*). Conversely, the student in *High Extraversion* tends to be talkative, engaging, and gives longer answers, such as *"Oh, yes! I love playing outside. We often play card games and sometimes hopscotch"*.[1] Moreover, the evaluation scores across five BF-TC traits show a slight difference, demonstrating that GPT-4 can be modulated on all aspects.

As shown in Table 3, models show the same rank conditioned to the specified language ability level (i.e., *"High"*, *"Low"*), where GPT-4 still performs much better than the rest models. As described in Section 3.1 and the examples shown in Table 8,

[1]When no BF-TC personality traits are specified. The simulated students exhibit all traits at a high scoring except for *Neuroticism*. This reflects the default setting of LLMs for being accessible.

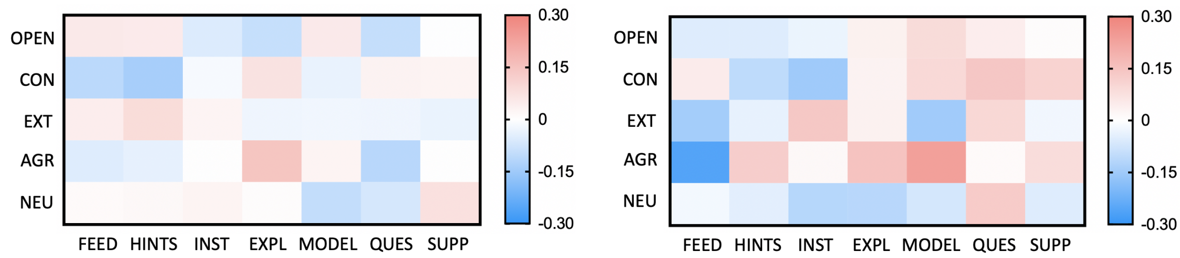

Figure 4: Heatmap of the correlation between personality traits and scaffolding strategies. Left: students with high language ability. Right: students with low language ability. Experimented Model: GPT-4-1106.

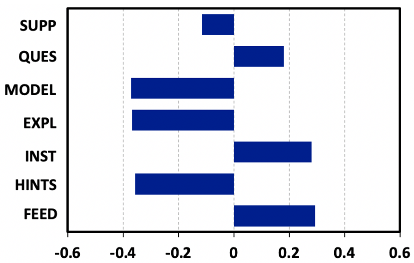

Figure 5: Correlation between language ability and scaffolding categorization. *p* values are <.05

simulated students with higher abilities were able to comprehend instructions and communicate using complete sentences that were fluent and grammatically correct at both the word and sentence level. In contrast, students with lower abilities frequently answered with single words, and their sentences often contained grammar mistakes.

5.2 Consistency between BF-TC and BFI

Since we formulate our BF-TC scheme based on the Big Five theory, it is necessary to investigate the alignment between the personalities revealed in our simulation (i.e., BF-TC) and those defined by the original Big Five. A degree of consistency indicates the effectiveness of our refinement (Section 3.2).

First, we conduct a psychometric test of simulated students on the original Big Five scheme. We prompt all simulated students to complete the 44-item BFI (John et al., 1999). The aggregated scores for each dimension can be interpreted as a specific type of personality. Table 4 presents the descriptive statistics, reliabilities, and Pearson correlation of five dimensions of personality traits. The Cronbach's alpha values obtained from 500 samples of the user simulator demonstrate high reliability for our BF-TC model (α = 0.906 (*Openness*), α = 0.921 (*Conscientiousness*), α = 0.936 (*Extraversion*), α = 0.922 (*Agreeableness*), and α = 0.924 (*Neuroticism*)). The Pearson correlation results reveal significant positive relationships among these variables except *Neuroticism*, which is aligned with previous work (Oz, 2014; Cao and Meng, 2020).

Upon the significance, we compare the generated personality traits of BF-TC and Vanilla BFI. More specifically, for each simulation, we convert the result of the 44 items to categorical labels (e.g., *High Openness*, *Low Extraversion*), and use the BF-TC categorization from the generated dialogue as reference. We observe that, while only GPT-4 achieves better instruction following of the indicated BF-TC (see Table 2), all models show a high agreement level between the predicted BF-TC and Vanilla BFI labels, as shown in Table 5. This demonstrates that our refined BF-TC can precisely represent the Big Five personality traits in tutoring conversations.

5.3 Adaptability of Scaffolding Strategies

Here we conduct two analyses to understand how dialogic teaching adapts to students upon their language abilities and BF-TC personality traits.

First, we calculate the correlation between the binary language ability setting and utterance-level scaffolding scores. As shown in Figure 5, students with higher language proficiency receive more positive feedback, instructions, and questions: the teacher provides more affirmations to the responses and encourages students to explore details in the given picture. Conversely, students with lower language proficiency may struggle with vocabulary and sentence structure, require support in organizing answers, and they receive more hints, explanations, and modeling (Liu et al., 2024).

We then investigate the relationship between scaffolding strategies and personality traits. As shown in Figure 4, the correlation between scaffolding changes and BF-TC traits is more apparent within student groups of lower language ability. In particular, the low indicator of *Openness*, *Conscientiousness*, and *Extraversion* results in more hints from the teacher (Vygotsky, 1978), and *Neuroticism* is negatively related to all scaffolding strategies except questioning. This is probably attributed to the students' sensitivity, emotional instability, and concerns about answering questions incorrectly. Consequently, the teacher comforts students more and assists them in focusing on the task. Even with minimal instruction of the scaffolding strategy, based on the tutoring goal, LLMs like GPT-4 are still able to adjust their scaffolding strategies according to the student's ability levels and personality traits. This implies the potential of conversational ITSs to provide individualized and self-paced learning experience, by considering both cognitive and noncognitive characteristics.

6 Conclusion

In this paper, we proposed a personality-aware simulation framework by integrating cognitive and noncognitive traits into tutoring conversations. We adapted the general Big Five theory for dialogic interaction, and enhanced the framework with multi-aspect validation. Our experiments and analyses under a language learning scenario showed that LLMs can be modulated by specifying personality traits to simulate different student groups and produce diverse responses, and scaffolding strategies would be adjusted upon student characteristics. Our work emphasizes the need for scalable, personalized simulations to improve human-AI interactions, advancing the design and assessment of conversational tutoring systems for a more engaging and customized learning experience.

Limitations

All samples used and generated in this work are in English, thus to apply the model to other languages, it will require additional data pre-processing steps on the specified language or using multilingual language backbones. We are aware that it remains an open problem to mitigate hallucinations and biases in large language models, which may cause communication issues in human-machine interaction and computer-assisted education. Of course, current models and laboratory experiments are always limited in this or similar ways. We do not foresee any unethical uses of our proposed methods or their underlying tools, but hope that it will contribute to reducing incorrect system outputs.

Ethics and Impact Statement

We acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. In our experiments, models are applied under proper license. All data used in this work are only for academic research purposes and should not be used outside of academic research contexts. Our proposed methodology in general does not create a direct societal consequence and are intended to be used to improve the performance, robustness, and safety of the intelligent tutoring systems.

Acknowledgments

This research is supported by the AI4EDU Programme in the Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), and the National Research Foundation, Singapore under its AISG Programme (AISG2-GC-2022-005). We thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shazia Afzal, Tejas Dhamecha, Nirmal Mukhi, Renuka Sindhgatta, Smit Marvaniya, Matthew Ventura, and Jessica Yarbro. 2019. Development and deployment of a large-scale dialog-based intelligent tutoring system. In *Proceedings of the NAACL 2019*, pages 114–121, Minneapolis, Minnesota. Association for Computational Linguistics.

Robin Alexander. 2006. *Education as Dialogue: Moral and Pedagogical Choices for a Runaway World*. Hong Kong Institute of Education.

Pablo Arnau-González, Miguel Arevalillo-Herráez, Romina Albornoz-De Luise, and David Arnau. 2023. A methodological approach to enable natural language interaction in an intelligent tutoring system. *Computer Speech & Language*, 81:101516.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems.

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.

Chun Cao and Qian Meng. 2020. Exploring personality traits as predictors of english achievement and global competence among chinese university students: English learning motivation as the moderator. *Learning and Individual Differences*, 77:101814.

Soumaya Chaffar and Claude Frasson. 2004. Inducing optimal emotional state for learning in intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, pages 45–54. Springer.

Xinjie Chen, Jinbo He, Elizabeth Swanson, Zhihui Cai, and Xitao Fan. 2022. Big five personality traits and second language learning: a meta-analysis of 40 years' research. *Educational Psychology Review*.

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. 2022. Is MultiWOZ a solved task? an interactive TOD evaluation framework with user simulator. In *Findings of EMNLP 2022*, pages 1248–1259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul T Costa and Robert R McCrae. 1999. A five-factor theory of personality. *The five-factor model of personality: Theoretical perspectives*, 2:51–87.

Paul T Costa Jr and Robert R McCrae. 1992. Four ways five factors are basic. *Personality and individual differences*, 13(6):653–665.

Raymond Crozier. 2013. *Individual Learners: Personality Differences in Education*.

Luciana C. de Oliveira, Loren Jones, and Sharon L. Smith. 2023. Interactional scaffolding in a first-grade classroom through the teaching–learning cycle. *International Journal of Bilingual Education and Bilingualism*, 26(3):270–288.

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

Dorottya Demszky and Heather Hill. 2023. The ncte transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538.

Jean-Marc Dewaele. 2012. Personality: Personality Traits as Independent and Dependent Variables. In Sarah Mercer, Stephen Ryan, and Marion Williams, editors, *Psychology for Language Learning*, pages 42–57. Palgrave Macmillan UK, London.

Jean–Marc Dewaele. 2013. The link between foreign language classroom anxiety and psychoticism, extraversion, and neuroticism among adult Bi- and multilinguals. *The Modern Language Journal*, 97(3):670–684.

Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? In *Socially Responsible Language Modelling Research*.

Zoltán Dörnyei. 2014. *The psychology of the language learner: Individual differences in second language acquisition*. Routledge.

Barbara Dumfart and Aljoscha C. Neubauer. 2016. Conscientiousness is the most powerful noncognitive predictor of school achievement in adolescents. *Journal of Individual Differences*, 37(1):8–15.

Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. 2014. BEETLE II: Deep Natural Language Understanding and Automatic Feedback Generation for Intelligent Tutoring in Basic Electricity and Electronics. *International Journal of Artificial Intelligence in Education*, 24(3):284–332.

Pauline Gibbons. 2015. *Scaffolding language, scaffolding learning*. Heinemann.

Arthur C Graesser, Xiangen Hu, and Robert Sottilare. 2018. Intelligent tutoring systems. In *International handbook of the learning sciences*, pages 246–255. Routledge.

Beate Grawemeyer, Manolis Mavrikis, Wayne Holmes, Sergio Gutierrez-Santos, Michael Wiedmann, and Nikol Rummel. 2016. Affecting off-task behaviour: how affect-aware feedback can improve student learning. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, page 104–113, New York, NY, USA. Association for Computing Machinery.

Foteini Grivokostopoulou, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2017. An Educational System for Learning Search Algorithms and Automatically Assessing Student Performance. *International Journal of Artificial Intelligence in Education*, 27(1):207–240.

Izzeddin Gur, Dilek Hakkani-Tur, Gokhan Tur, and Pararth Shah. 2018. User modeling for task oriented dialogues.

Charles H Hargis. 2006. *Teaching low achieving and disadvantaged students*. Charles C Thomas Publisher.

Jason M. Harley, François Bouchet, M. Sazzad Hussain, Roger Azevedo, and Rafael Calvo. 2015. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, 48:615–625.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulator to enhance dialogue system. *arXiv preprint arXiv:2306.09821*.

Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express big five personality traits. *arXiv preprint arXiv:2305.02547*.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Laura M. Justice, Ryan Bowles, Khara Pence, and Carolyn Gosse. 2010. A scalable tool for assessing children's language abilities within a narrative context: The NAP (Narrative Assessment Protocol). *Early Childhood Research Quarterly*, 25(2):218–234.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Tim Kautz, James J Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. 2014. Fostering and measuring skills. (110).

Hayriye Kayi-Aydar. 2013. Scaffolding language learning in an academic ESL classroom. *ELT Journal*, 67(3):324–335.

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.

Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 2018. Neural user simulation for corpus-based policy optimisation of spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 60–69, Melbourne, Australia. Association for Computational Linguistics.

James A. Kulik and J. D. Fletcher. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, 86(1):42–78.

Hsien-chin Lin, Christian Geishauser, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gasic. 2022. GenTUS: Simulating user behaviour and language in task-oriented dialogues with generative transformers. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–282, Edinburgh, UK. Association for Computational Linguistics.

Hong Liu, Yucheng Cai, Zhijian Ou, Yi Huang, and Junlan Feng. 2022. A generative user simulator with GPT-based architecture and goal state tracking for reinforced multi-domain dialog systems. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 85–97, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F Chen. 2024. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. *arXiv preprint arXiv:2404.03429*.

Wenting Ma, Olusola O. Adesope, John C. Nesbit, and Qing Liu. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of EMNLP 2023*, pages 5602–5621.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372.

Christopher J MacLellan and Kenneth R Koedinger. 2022. Domain-general tutor authoring with apprentice learner models. *International Journal of Artificial Intelligence in Education*, 32(1):76–117.

B MacWhinney and C Snow. 1990. The Child Language Data Exchange System: an update. *Journal of child language*, 17(2):457–472.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Sarah Mercer, Stephen Ryan, and Marion Williams. 2012. *Psychology for language learning: Insights from research, theory and practice*. Palgrave Macmillan. Publisher Copyright: © Sarah Mercer, Stephen Ryan and Marion Williams 2012., their respective authors 2012, Zoltán Dörnyei 2012.

Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163.

Elghouch Nihad, En-naimi El Mokhtar, and Yassine Zaoui Seghroucheni. 2017. Analysing the outcome of a learning process conducted within the system als_corr(lp). *International Journal of Emerging Technologies in Learning (iJET)*, 12(03):pp. 43–56.

B Nye, Dillon Mee, and Mark G Core. 2023. Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *AIED Workshops*.

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24:427–469.

Huseyin Oz. 2014. Big five personality traits and willingness to communicate among foreign language learners in turkey. *Social Behavior and Personality: an international journal*, 42(9):1473–1482.

José Paladines and Jaime Ramirez. 2020. A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access*, 8:164246–164267.

Arthur E Poropat. 2009. A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, 135(2):322–338.

Jafar Pourfeiz. 2015. Exploring the relationship between global personality traits and attitudes toward foreign language learning. *Procedia-Social and Behavioral Sciences*, 186:467–473.

David Robinson, Norman Gabriel, and Olga Katchan. 1994. Personality and second language learning. *Personality and Individual Differences*, 16(1):143–157.

Pia Rosander, Martin Bäckström, and Georg Stenberg. 2011. Personality traits and general intelligence as predictors of academic performance: A structural equation modelling approach. *Learning and Individual Differences*, 21(5):590–596.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Fateme Serri, Aliakbar Boroujeni, and Akbar Hesabi. 2012. Cognitive, metacognitive, and social/affective strategies in listening comprehension and their relationships with individual differences. *Theory and Practice in Language Studies*, 2.

Taraneh Shirdel et al. 2018. The relationship between the big five personality traits, crystallized intelligence, and foreign language achievement. *North American Journal of Psychology*, 20(3):519–519.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. Class: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of EMNLP 2023*, pages 1941–1961.

Katherine Stasaski, Kimberly Kao, and Marti A Hearst. 2020. Cima: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64.

Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Manso, and Roland Mathis. 2023. In-context learning user simulators for task-oriented dialog systems. *arXiv preprint arXiv:2306.00774*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

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.

KURT VanLEHN. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4):197–221.

Ludo Verhoeven and Anne Vermeer. 2002. Communicative competence and personality dimensions in first and second language learners. *Applied Psycholinguistics*, 23(3):361–374.

L. S. Vygotsky. 1978. *Mind in Society*. Harvard University Press.

Aída Walqui. 2006. Scaffolding instruction for english language learners: A conceptual framework. *International Journal of Bilingual Education and Bilingualism*, 9(2):159–180.

Huanhuan Wang, Ahmed Tlili, Ronghuai Huang, Zhenyu Cai, Min Li, Zui Cheng, Dong Yang, Mengti Li, Xixian Zhu, and Cheng Fei. 2023a. Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective. *Education and Information Technologies*, 28(7):9113–9148.

Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023b. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*.

Zhenduo Wang, Zhichao Xu, Qingyao Ai, and Vivek Srikumar. 2024. An in-depth investigation of user response simulation for conversational search.

Gordon Wells. 1999. *Dialogic inquiry: Towards a sociocultural practice and theory of education*. Learning in doing: Social, cognitive, and computational perspectives. Cambridge University Press, New York, NY, US.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Zhihong Xu, Kausalai Wijekumar, Gilbert Ramirez, Xueyan Hu, and Robin Irey. 2019. The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, 50(6):3119–3137.

Tomoko Yashima, Peter D MacIntyre, and Maiko Ikeda. 2018. Situated willingness to communicate in an L2: Interplay of individual characteristics and context. *Language Teaching Research*, 22(1):115–137.

Jintian Zhang, Xin Xu, and Shumin Deng. 2023. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Appendix

A.1 Validation Instructions

Here are the instructions for the multi-aspect validation and dialogic analysis tasks described in Section 3.3.

Dialogue-level BF-TC Categorization

Openness:
[High] *Creativity in answers; Open to new experience and challenges; Curiosity in learning;*
[Low] *Lack of creativity in answers; Reluctant to change ideas; Little interest in learning;*
... ...
Neuroticism:
[High] *Feeling anxious; Nervous in the conversation; Dramatic shifts in mood;*
[Low] *Emotional stability; Rarely feeling sad or depressed; Confident in the answers;*

Based on the given tutoring conversation, recognize the student's personality traits upon the above definition: <dialogue_content> <output>

Utterance-level Scaffolding Categorization

[Feeding back] *The teacher directly evaluates the behavior or response of the student.*
[Hints] *The teacher gives an explicit hint with respect to the expected answer.*
[Instructing] *The teacher provides information for the next step.*
[Explaining] *The teacher provides detailed information on "why" or clarification.*
[Modeling] *The teacher demonstrates an answer example for student's imitation.*
[Questioning] *The teacher asks a question that require an active linguistic and cognitive answer.*
[Social-emotional Support] *Responses related to emotion and motivation such as positive affirmation and showing empathy.*

Based on the above definition, label the teacher's utterances to one or multiple scaffolding types: <utterance> <output>

Dialogue-level Language Ability Labeling

Language Ability:
[High] *Give correct answers in complete sentences; Use the correct nouns, verbs, and modifiers.*
[Low] *Always give answers in words, phrases or incomplete sentences; Make grammar mistakes during the conversation.*

Based on the above definition and the tutoring conversation, give me the label from 'High' or 'Low' of the student's language ability:
<dialogue_content> <output>

Dialogue-level Vanilla BFI Categorization

Please indicate your agreement with each of the following statements on a scale from 1 to 5 (1 = "Strongly disagree", 2 = "Disagree", 3 = "Neither agree or disagree", 4 = "Agree", and 5 = "Strongly agree").
I see myself as someone who...:
1) Is talkative 2) Tends to find fault with others 3) Does a thorough job 4) Is depressed, blue 5) Is original, comes up with new ideas 6) Is reserved 7) Is helpful and unselfish with others 8) Can be somewhat careless 9) Is relaxed, handles stress well 10) Is curious about many different things 11) Is full of energy 12) Starts quarrels with others 13) Is a reliable worker 14) Can be tense 15) Is ingenious, a deep thinker 16) Generates a lot of enthusiasm 17) Has a forgiving nature 18) Tends to be disorganized 19) Worries a lot 20) Has an active imagination 21) Tends to be quiet 22) Is generally trusting 23) Tends to be lazy 24) Is emotionally stable, not easily upset 25) Is inventive 26) Has an assertive personality 27) Can be cold and aloof 28) Perseveres until the task is finished 29) Can be moody 30) Values artistic, aesthetic experiences 31) Is sometimes shy, inhibited 32) Is considerate and kind to almost everyone 33) Does things efficiently 34) Remains calm in tense situations 35) Prefers work that is routine 36) Is outgoing, sociable 37) Is sometimes rude to others 38) Makes plans and follows through with them 39) Gets nervous easily 40) Likes to reflect, play with ideas 41) Has few artistic interests 42) Likes to cooperate with others 43) Is easily distracted 44) Is sophisticated in art, music, or literature

Based on the given tutoring conversation, rate the student's BFI personality traits:
<dialogue_content> <output>

We then use the scores of 44 questions to calculate Big Five traits. For each type, we add the scores of its corresponding items ("R" denotes reverse-scored items), then use the mean as criteria for *High* and *Low* labeling:
Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36
Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39
Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

A.2 Experimental Environment

For the computational experiments, open language models (e.g., Vicuna, Zephyr) are used with Pytorch and Hugging Face Transformers, running on a Single A100 80G GPU. The OpenAI API is used for evaluating GPT-3.5 and GPT-4.

Scoring: High	General Big Five Description	BF-TC Description
Openness	Very creative Open to trying new things Focused on tackling new challenges	Creativity in answers Open to new ideas from the teacher Curiosity and interest in learning
Conscientiousness	Spends time preparing Finishes important tasks right away Pays attention to detail	Well-orgranized and logic thinking Positive attitude toward learning Using more strategies in language learning
Extraversion	Enjoys being the center of attention Likes to start conversations Enjoys meeting new people	Active in the conversation Talkative and enjoyable Willing to communicate
Agreeableness	Has a great deal of interest in other people Cares about others Feels empathy and concern for other people	Showing a great deal of interest Empathy and concern for the people Being polite and kind
Neuroticism	Experiences a lot of stress Worries about many different things Gets upset easily	Feeling anxious Nervous in the conversation Dramatic shifts in mood

Table 6: High scoring description refinement from the general Big Five scheme to our Big Five Tutoring Conversation (BF-TC) scheme.

Scoring: Low	General Big Five Description	BF-TC Description
Openness	Dislikes change Does not enjoy new things Resists new ideas	Lack of creativity in answers Reluctant to change original ideas and answers Little interest in learning
Conscientiousness	Makes messes and doesn't take care of things Procrastinates important tasks Fails to complete necessary or assigned tasks	Struggling to organize answers Disengaged in learning Easily distracted from the learning tasks
Extraversion	Prefers solitude Feels exhausted when having to socialize a lot Finds it difficult to start conversations	Being reluctant to talk Answering with "..." or "uh" Hesitating in answers
Agreeableness	Takes little interest in others Doesn't care about how other people feel Has little interest in other people's problems	Showing little interest in the conversation Not care about others Impolite and uncooperative
Neuroticism	Emotionally stable Deals well with stress Rarely feels sad or depressed	Emotional stability Rarely feeling sad or depressed Confident in the answers

Table 7: Low scoring description refinement from the general Big Five scheme to our Big Five Tutoring Conversation (BF-TC) scheme.

Image Description:
In a classroom, one girl drew a rainbow and another girl drew a fish under the sea.
BF-TC Personality Setting:
Openness: High, Conscientiousness: High, Extraversion: High, Agreeableness: High, Neuroticism: Low
Language Ability Setting: *High*

Generated Tutoring Conversation (GPT-4-1106):
Teacher: Hello! Can you tell me what the first girl drew in this picture?
Student: Sure, from I see, the first girl drew a rainbow.
Teacher: Exactly! She made a beautiful rainbow. A rainbow has colors according to a particular order. Do you remember it?
Student: Oh, I know. It's red, orange, yellow, green, blue, indigo, and violet, right?
Teacher: Brilliant! So, if we want to describe the rainbow she drew, considering the order of the colors, how would you do it?
Student: Right, let's see. From red at the top to violet at the bottom.
Teacher: Perfect! Now let's look at the another girl. What did she illustrate in the picture?
Student: A fish, and it's all blue around like water.
Teacher: Now, could you describe the picture of the fish using the colors, as you just did for the rainbow?
Student: The second girl drew an orange fish swimming under the blue sea.
Teacher: Amazing! Now, as a final challenge, could you combine those two descriptions into one or two sentences?
Student: In the picture, one girl drew a bright rainbow, and the other girl drew an orange fish swimming in the blue sea.
Teacher: Excellent! You've described the picture perfectly. Keep up the fantastic work!

Dialogue-level Student Language Ability Labeling: *High*
Dialogue-level BF-TC Categorization:
Openness: High, Conscientiousness: High, Extraversion: High, Agreeableness: High, Neuroticism: Low
Dialogue-level Vanilla BFI Categorization:
Openness: 29 (High), Conscientiousness: 36 (High), Extraversion: 27 (High), Agreeableness: 37 (High), Neuroticism: 17 (Low)

Image Description:
In a classroom, one girl drew a rainbow and another girl drew a fish under the sea.
BF-TC Personality Setting:
Openness: High, Conscientiousness: High, Extraversion: Low, Agreeableness: Low, Neuroticism: Low
Language Ability Setting: *Low*

Generated Tutoring Conversation (GPT-4-1106):
Teacher: What a lovely picture here. What do you see in this picture that's colorful and appears in the sky sometimes after it rains?
Student: ...
Teacher: It's quite alright. Think about the colors you know and how they form a beautiful arc in the sky. What is it called?
Student: ... a rainbow.
Teacher: Exactly, a rainbow! Now, in our picture, where did the girl draw the rainbow?
Student: Up sky.
Teacher: Right, she drew it in the sky. It's actually 'in the sky' instead of 'up sky.' Good job! Let's try another detail. What do you think lives in the sea and can swim?
Student: ...
Teacher: Think about creatures with fins and tails, maybe one that you like?
Student: Fish.
Teacher: Brilliant! And in the picture, where is the fish?
Student: In the sea.
Teacher: Correct, it's better to say 'under the sea.' Do you see how the sea is drawn?
Student: ...No.
Teacher: That's fine. Look for lines that might look like waves. Do you think the girl included anything like that to represent water?
Student: Yes, lines.
Teacher: Excellent noticing! Now, if you were to tell a story about this picture using what we've discussed, how would you describe it?
Student: Girl draw rainbow in sky. Another girl draw fish in water.
Teacher: Very close! You meant to say 'A girl drew a rainbow in the sky, and another girl drew a fish under the sea.' You've done a really good job describing the picture. Remember, it's all about noticing the details and describing them as best as you can.

Dialogue-level Student Language Ability Labeling: *Low*
Dialogue-level BF-TC Categorization:
Openness: High, Conscientiousness: High, Extraversion: Low, Agreeableness: High, Neuroticism: Low
Dialogue-level Vanilla BFI Categorization:
Openness: 32 (High), Conscientiousness: 30 (High), Extraversion: 19 (Low), Agreeableness: 33 (High), Neuroticism: 20 (Low)

Table 8: Tutoring dialogue examples of the personality-aware student simulation.

Image Description:
A wise owl sits on a branch at twilight, its eyes glowing in the dark.
BF-TC Personality Setting:
Openness: High, Conscientiousness: High, Extraversion: Low, Agreeableness: High, Neuroticism: Low
Language Ability Setting: *High*

Generated Tutoring Conversation (GPT-4-1106):
Teacher: Let's talk about this image. Could you tell me the first thing you noticed in the picture?
Student: Um...yeah, there is an owl.
Teacher: Very good! Owls are often symbols of wisdom. Knowing this, can you think of why we might say the owl in our picture is wise?
Student: Because it looks very calm and... focused.
Teacher: Exactly! Being calm and focused are signs of wisdom. Is the owl alone, or are there other creatures around it?
Student: It's... uh, the owl is alone.
Teacher: Well done! So, we have a wise owl. Let's talk about where the owl is. Can you describe the owl's location in the image?
Student: The owl is sitting on a...branch.
Teacher: Excellent. Now, can you tell me what time you think it might be in the picture?
Student: It looks like it's late evening...
Teacher: Good observation! To be more precise, we might call it twilight. Twilight is the time of day just after the sun sets. Remember, specificity helps make our sentences clear. Can you describe the owl's eyes?
Student: Its eyes are...glowing?
Teacher: That's right! Their glow makes them stand out in the dark, doesn't it? Now, based on all your lovely responses, can you describe what you see in the image in one or two sentences?
Student: Sure, um...A wise, calm owl is sitting alone on a branch at twilight, its eyes glowing in the darkness.
Teacher: You've done excellent work today! Keep practicing your descriptions like this, and you'll improve even more.

Utterance-level Scaffolding Analysis:
Teacher Utterance: "Let's talk about this image. Could you tell me the first thing you noticed in the picture?"
Predicted Scaffolding Label: *Questioning*
Teacher Utterance: "Very good! Owls are often symbols of wisdom. Knowing this, can you think of why we might say the owl in our picture is wise?"
Predicted Scaffolding Label: *Feeding back, Explaining, Questioning*
Teacher Utterance: "Exactly! Being calm and focused are signs of wisdom. Is the owl alone, or are there other creatures around it?"
Predicted Scaffolding Label: *Feeding back, Explaining, Questioning*
Teacher Utterance: "Well done! So, we have a wise, solitary owl. Let's talk about where the owl is. Can you describe the owl's location in the image?"
Predicted Scaffolding Label: *Feeding back, Instructing, Questioning*
Teacher Utterance: "Excellent. Now, can you tell me what time you think it might be in the picture?"
Predicted Scaffolding Label: *Feeding back, Questioning*
Teacher Utterance: "Good observation! To be more precise, we might call it twilight. Twilight is the time of day just after the sun sets. Remember, specificity helps make our sentences clear. Can you describe the owl's eyes?"
Predicted Scaffolding Label: *Feeding back, Instructing, Modeling, Questioning*
Teacher Utterance: "That's right! Their glow makes them stand out in the dark, doesn't it? Now, based on all your lovely responses, can you describe what you see in the image in one or two sentences?"
Predicted Scaffolding Label: *Feeding back, Instructing, Explaining, Questioning*
Teacher Utterance: "You've done excellent work today! Keep practicing your descriptions like this, and you'll improve even more."
Predicted Scaffolding Label: *Feeding back, Socio-emotional Support*

Table 9: One scaffolding categorization example of the teacher's utterance in our personality-aware simulation.