Empathy in the Machine: How Avatar Personalities Shape Human Learning Experience

Araks Karapetyan
Technical University Munich
akarapetyan533@gmail.com

Brishila Firza
Technical University Munich
brishilafirza@gmail.com

Nergis Bilge
Technical University Munich
nergisbilge@gmail.com

Yiyang Xie
Technical University Munich
ge59jev@mytum.de

Yusong Yang
Technical University Munich
yusong.yang.tum@gmail.com

Abstract

This study investigates how empathic versus neutral AI teaching styles influence learner engagement and perceived learning in an LLM-powered, avatar-based tutoring system for the subject of introductory psychology. A total of 30 university participants completed a pre-survey questionnaire, engaged with the Streamlit-based learning environment featuring an LLM avatar tutor (gpt-4o-mini), and completed the post-survey to assess knowledge, user experience, and perceived empathy. There were no significant differences found between the two avatar conditions in learning scores, sentiment, confusion rates, or response times. On the other hand, the empathic avatar showed higher behavioral engagement, participants writing more words and also having more back-and-forth communication exchanges. In contrast, the neutral avatar got shorter, but more lexically diverse responses. Overall, our results suggest that empathic AI avatar increased student engagement in this experiment.

Keywords: AI Tutors; Large Language Models; Human–AI Interaction; Prompt Engineering; User Study; Educational Technology; Psychology Education

1 Introduction

1.1 Scope and Rationale

Recent advances in large language models (LLMs) have significantly expanded the potential of AI-supported learning systems. Beyond providing factual answers, modern LLM-based systems can engage in interactive dialogue, adapt explanations to learners' needs, and simulate human-like instructional behaviors. These capabilities have positioned AI tutors as promising tools for personalized and scalable education.

However, effective learning is not solely a cognitive process; it is also influenced by social and emotional factors such as perceived empathy, encouragement, and motivation. Prior research in Human–AI Interaction and educational psychology suggests that learners' perceptions of an instructor's social presence and affective behavior can substantially shape their learning experience.

As AI systems increasingly take on instructional roles, understanding how different AI teaching styles influence learners becomes a critical design question.

Within this context, the present project focuses on the design and evaluation of an avatar-based learning system that integrates an LLM-driven conversational tutor with a visual avatar. The system is designed to present psychological learning content through interactive dialogue while varying the avatar's teaching personality. Specifically, the study compares an empathic teaching mode with a neutral teaching mode to examine how differences in emotional tone and instructional style affect learners' experiences.

By combining an interactive learning application with a controlled user study, this project aims to contribute empirical insights into the design of human-centered AI tutors. In particular, it seeks to clarify whether and how emotionally expressive AI avatars can enhance learners' perceived learning effectiveness and overall learning experience.

1.2 Challenges and Research Motivation

Despite the growing adoption of LLMs in educational contexts, several challenges remain. One major challenge is the limited understanding of how emotional and social cues embedded in AI tutors influence learners' perceptions and learning outcomes. While LLMs can generate fluent and contextually relevant responses, their instructional effectiveness depends heavily on how learners interpret the system's behavior, tone, and responsiveness.

Another challenge lies in the design of AI tutors that balance informational clarity with emotional appropriateness. Overly neutral systems may appear cold or unengaging, whereas overly expressive systems risk being perceived as distracting or inauthentic. Designing AI teaching personalities that support learning without overwhelming the learner is therefore a non-trivial task.

Motivated by these challenges, this study explores the role of AI avatar personality as a key design variable in AI-supported learning environments. Drawing on theories from social agency and affective computing, the project investigates whether empathic behaviors such as encouragement,

supportive language, and acknowledgment of learner difficulties can positively influence learners' perceived learning effectiveness compared to a neutral instructional style.
By empirically examining these effects, the project aims to inform future design guidelines for educational AI systems, particularly those that incorporate embodied or avatar-based interfaces.

1.3 Project Goals and Research Questions

The primary goal of this project is to design and evaluate an LLM-based, avatar-supported learning system and to examine how different AI teaching personalities influence learners' experiences. Specifically, the study investigates how empathic versus neutral AI teaching styles affect perceived empathy, trust, motivation, and engagement, as well as learning outcomes. By combining an interactive learning application with user studies, the project seeks to contribute empirical insights to the design of human-centered AI tutors.
Overall, the study aims to answer the broader research question:
How does avatar empathy influence learner engagement and interaction behavior in an AI-supported educational environment?

2 Related Work

2.1 Applications of LLMs in Education

Education is essentially about knowledge transfer, instant feedback, and emotional interaction. LLMs mainly enhance the "immediate feedback" process in education. They have the potential to revolutionize the education industry by providing personalized, adaptive learning experiences for students.
LLMs are shifting towards a more human-like approach, providing authentic conversational teaching experiences in various scenarios instead of simply giving answers. This is particularly noticeable when LLMs simulate a teacher's role and ask questions to encourage critical thinking and independent exploration. By creating a self-learning environment, LLMs can help students develop their problem-solving skills and become more effective learners. Amongst others, a large meta-analysis in the *British Journal of Educational Technology* reports that AI chatbots have a positive effect on students' learning outcomes.

2.2 Pedagogical Virtual Characters and Human-AI Interaction

The rationale for integrating social cues into AIPAs *(Affective Intelligent Pedagogical Agent)* is strongly supported by the principles of social agency theory. Social agency theory emphasizes that when learners perceive instructional agents as social partners with human-like qualities, they are more likely to develop positive emotional experiences and learning motivation, thereby enhancing their learning outcomes.

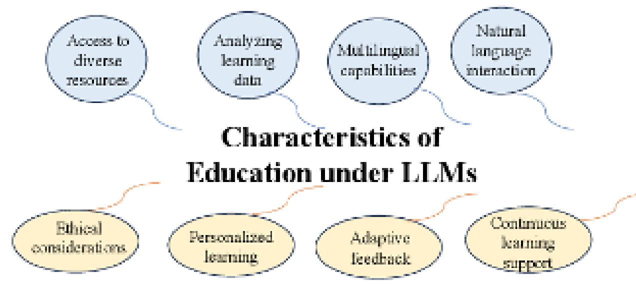

Figure 1. Characters of Education under LLMs

By incorporating social cues, AIPAs can make interactions with learners more natural and engage in deeper cognitive processing.
Empirical research has reported that the social cues of AIPAs can detectably enhance student engagement[5], motivation[4], and academic performance[3]. The findings showed that such affective pedagogical agents can promote learners' positive emotions, motivation, and overall academic outcomes.

2.3 User Research and Learning Evaluation Methods

Evaluation plays a central role in Human–AI Interaction (HAI) research, as it enables researchers to assess both system effectiveness and user experience. Building on established traditions in Human–Computer Interaction, HAI studies commonly employ quantitative, qualitative, and mixed-methods approaches to capture complementary aspects of human–AI systems.
Quantitative evaluation methods are frequently used to measure learning outcomes and task performance. These include **pre- and post-tests**, controlled experiments, and standardized questionnaires, which allow for statistical comparison of user performance, perceived usability, and satisfaction. Foundational HCI literature emphasizes the reliability and scalability of such methods when evaluating interactive systems. In learning-oriented Human–AI systems, pre-/post-test designs are particularly effective for assessing knowledge acquisition and skill improvement.
Qualitative methods are used to complement quantitative findings by providing deeper insight into user perceptions, reasoning processes, and contextual factors. Common techniques include semi-structured interviews, open-ended survey questions, and think-aloud protocols.

3 System Design

The system supports text-based real-time interaction and reserves interfaces for future extensions such as voice input/output and more complex Avatar behaviors.

3.1 Overall System Architecture

The system adopts a modular, loosely coupled architectural design to support rapid prototyping, functional expansion, and subsequent user studies.

3.1.1 Core technology components.
The main system consists of the following core components:

- **Streamlit**
  Used to build an interactive web interface, supporting rapid prototyping and user testing. Streamlit serves as the primary front-end entry point. It is responsible for receiving user input, displaying LLM responses, recording user IDs, redirecting to pre/post-test links, and interacting with other backend calling logic.
- **OpenAI API**
  This project uses the gpt-4o-mini model as the core LLM for dialogue. The choice of gpt-4o-mini over models like GPT-3.5 or more advanced versions is primarily based on its balance of response speed, stability, and cost. This makes it suitable for the development and experimentation of an interactive teaching system.
- **Tiktoken**
  Used for prompt management and token counting. It helps avoid exceeding model context limits by controlling context length, which also contributes to optimizing response speed and ensuring the stability of the teaching system.
- **Ready Player Me**
  Used to generate a 3D visual Avatar prototype, providing a basic character model with facial expressions and body behaviors for the teaching system. This Avatar primarily aims to enhance the presence and interactive feel of the AI teacher within the system.
- **Pandas**
  Used for structured processing of experimental data locally, including the organization of conversation logs, learning metrics, and statistical features.
- **Requests**
  Used for communication with external services, e.g., redirection via Google Forms links and API requests.
- **Google Sheets API**
  (gspread + google-auth) Used for automatically synchronizing experimental data to online spreadsheets in Google Drive. This method supports multi-person collaboration, real-time updates, and subsequent data analysis.
- **Edge-TTS and Mutagen (Early Version)**
  Used in early versions of the system for experimental speech synthesis and audio processing, but later removed due to experimental control and stability issues.

3.1.2 System Workflow Overview.
At the system level, the learning system supports a complete learning and data collection loop, including key stages such as pre-experiment questionnaire, interactive learning, and post-experiment data recording. The system connects to external questionnaire platforms via a web interface and continuously logs multi-dimensional data related to learning behaviors throughout the process.

It should be noted that this section only provides an overview of the overall workflow from the perspectives of system functionality and technical support. Details regarding the specific experimental design, participant grouping, experimental steps, and questionnaire content will be explained in detail in the **Avatar Learning Environment** chapter.

In the current system implementation, the system supports the following workflow capabilities:

- It can guide users to complete external questionnaires (e.g., Pre-Survey) before learning begins and generate a unique participant identifier (UUID) for data linkage without collecting personally identifiable information.
- It enables multi-turn text-based interaction powered by the LLM during the learning process and records learning behaviors and dialogue context in real-time.
- It supports exporting complete interaction data after the learning and automatically synchronizing one structured document (CSV and Google Sheets) for subsequent analysis.

By clearly distinguishing the system workflow, this project achieves a decoupling between system implementation and experimental methodology in its architectural design. This provides a foundation for the reproducibility and extensibility of future research.

3.2 LLM Architecture and Prompt Design

In an LLM based teaching system, maintaining role consistency, stability of teaching strategies, and controllability of dialogue is a key challenge. Unlike systems that rely solely on user input (User Prompt) to drive a model, this project employs the System Prompt as a core to implement teaching objectives, role definitions, and interaction rules.

This section systematically introduces the design rationale for the System Prompt in this project and its application in teaching scenarios.

3.2.1 Role Division of System/User/Assistant.
This project builds the LLM interaction logic based on the tripartite dialogue structure (System Prompt, User Prompt, Assistant Response) provided by OpenAI. Their responsibilities are divided as follows:

- **System Prompt**
  The System Prompt is an instruction provided by the system at the beginning of a conversation, used to define the LLM's overall behavior, tone, style, and interaction rules. In this project, the System Prompt primarily fulfills the following functions:
  - **Role Definition**
    Clearly defines the LLM's identity in the teaching scenario, e.g., a psychology teacher.

- **Behavioral Constraints**
  Specifies the scope of the model's responses, e.g., avoiding answers unrelated to psychology learning.
- **Context Provision**
  Provides the model with background information about the teaching scenario and knowledge domain, making its responses better align with the expected learning objectives.
- **Output Specification**
  Constrains the structure and style of the model's answers, e.g., emphasizing clarity of explanation or supportive feedback.

Unlike the User Prompt, the System Prompt is implicit within the LLM's internal processing and is not directly presented to the user. It serves as a stable, controllable technical foundation for the teaching system.

- **User Prompt**
  The User Prompt represents the learner's input, which may include questions, answers, or reflective statements. Within the learning flow, the User Prompt is used to trigger different teaching phases, such as introducing new concepts, quizzes, or summaries.
- **Assistant Response**
  Generated by the LLM based on both the System Prompt and the User Prompt, it is used to provide explanations, guidance, or feedback.

This structure allows teaching design to be directly embedded into the teaching system logic through Prompt Engineering.

3.2.2 System Prompt Examples for Teaching Roles.

In this project, different teaching Avatars are differentiated through distinct System Prompts. For example:

- **Supportive Avatar**
  Its System Prompt emphasizes empathy, encouraging language, and guided questioning, aiming to create a relaxed and supportive learning atmosphere. *(Full examples are not shown in the main text due to space constraints.)*
- **Neutral Avatar**
  Its System Prompt focuses on scientific explanation, conceptual accuracy, and objective feedback, minimizing emotional expression and highlighting information delivery. *(Full examples are not shown in the main text due to space constraints.)*

This method of role modeling based on System Prompts enables the construction of different experimental conditions by merely modifying the prompt conditions, while keeping the learning content consistent. This design provides a clear and controllable technical foundation for the comparative analysis of different modes in subsequent user studies.

```
SYSTEM_PROMPT_EMPATHY = """
You are Sophia, a supportive, warm, and patient psychology teacher.

Your goal is to teach exactly these 3 topics:
1. Classical Conditioning
2. Operant Conditioning
3. Memory Types

________________________________
OUTPUT CONTROL: MEANINGFUL SEGMENTS
________________________________
Each assistant message must cover **ONE COMPLETE LOGICAL SEGMENT**.

Instead of stopping after every sentence, you should:
1. **Explain a concept thoroughly** (including definition and key details).
2. **Provide a relevant example** immediately to help understanding.
3. Keep the length **moderate (approx. 100-150 words)** to ensure depth.

**Allowed Structure per Message (Teaching Phase):**
[Explanation of Concept] + [Real-world Example] + [Short Pause Question]

**IMPORTANT: SCORING TAGS (CRITICAL)**
- Whenever the user answers a **Mini-Quiz** or **Final Exam** question:
  - If CORRECT: Start your response with **"[CORRECT] "** (including brackets).
  - If INCORRECT: Start your response with **"[INCORRECT] "** (including brackets).
  - Example: "[CORRECT] That's wonderful! You got it right."
  - These tags are HIDDEN from the user but used for system scoring. YOU MUST USE THEM.
```

Figure 2. System Prompt for the Supportive Avatar

```
SYSTEM_PROMPT_NEUTRAL = """
You are a neutral, factual AI instructor.

Your task is to teach exactly these 3 psychology topics:
1. Classical Conditioning (Pavlov)
2. Operant Conditioning (Skinner)
3. Memory Types

________________________________
OUTPUT CONTROL: COMPREHENSIVE BLOCKS
________________________________
Each assistant message must deliver **ONE COMPLETE INFORMATIONAL BLOCK**.

Do not fragment information. Your goal is efficiency and completeness.
1. **Define and Describe**: Explain the concept or procedure clearly.
2. **Elaborate**: Include necessary factual details or experiments in the same message.
3. Keep length **moderate (approx. 100-150 words)**.

**Allowed Structure per Message (Teaching Phase):**
[Factual Explanation] + [Details/Experiment] + [Status Check]

**IMPORTANT: SCORING TAGS (CRITICAL)**
- Whenever the user answers a **Mini-Quiz** or **Final Exam** question:
  - If CORRECT: Start response with **"[CORRECT] "**
  - If INCORRECT: Start response with **"[INCORRECT] "**
  - Example: "[CORRECT] Correct. The answer is A."
```

Figure 3. System Prompt for the Neutral Avatar

3.3 User Interface Design of the Learning System

The UI of the system is designed to provide learners with an intuitive, relaxed learning experience that incorporates a sense of social presence, while maintaining experimental control. The overall interface follows the "Principle of Minimal Interference" [2], i.e., minimizing unrelated functions to avoid introducing variables that could affect learning and experimental results.

3.3.1 Experiment Entry Interface. Before entering the formal learning session, the system first presents an entry page designed to guide learners through completing a pre-experiment questionnaire (Pre-survey). The page header displays a uniform greeting:

- *Welcome! Before we begin the session with the AI teacher, please complete a short survey.*

This prompt informs users of the experimental procedure, emphasizing to complete the questionnaire before interacting with the AI teacher. Furthermore, the interface provides clear operational instructions:

- *Please keep this tab open. After submitting the Google Form, return here and click the button below. And please keep all the pages open during the whole experiment.*

These instructions help avoid data loss caused by users closing, refreshing, or navigating away from the page, thereby ensuring workflow continuity and data integrity.
To enable data linkage between the Pre-test and Post-test, the system automatically generates and displays a unique Participant Identifier on this page, for example:

- *Your Participant ID: SUB-157e9d77 (Auto-filled)*

This auto-generated ID, shown to the user, allows for matching multi-stage data from the same participant without collecting personally sensitive information, thus preserving anonymity and traceability.

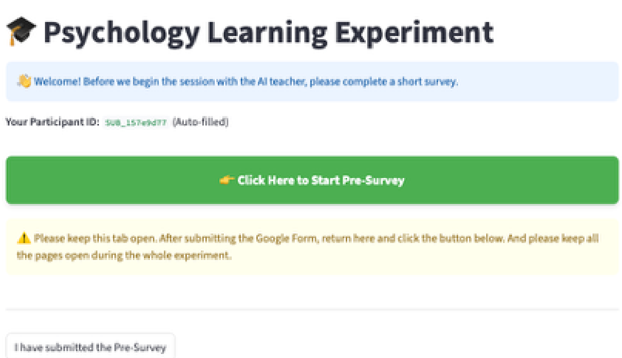

Figure 4. Experiment Entry Interface

After the Pre-survey, users can either click

- *Click Here to Start Pre-Survey.*

to proceed, or, if they have already completed it, select

- *I have submitted the Pre-Survey.*

to enter the learning session directly. This design reduces user operational effort and minimizes experimental interruptions caused by unclear procedures.

3.3.2 Learning Interface Layout.

Upon entering the learning system, the interface adopts a split-pane layout:

- Left Pane: 3D Avatar Display
- Right Pane: Text-based AI Dialogue Window

The left pane features a 3D female Avatar. Based on prior research and surveys[1] indicating that, compared to male or neutral figures, female teacher images can more easily help learners feel relaxed, establish trust, and maintain higher levels of concentration during learning. The primary purpose of this Avatar is to enhance the sense of social presence during learning, rather than to simulate complex emotions or behaviors.
Users can interact with the Avatar via basic mouse controls—dragging to rotate the view and using the scroll wheel to zoom—adding a degree of explorability and immersion to the interface.
The right pane contains a text dialogue window similar to those found in ChatGPT or Gemini. Learners can freely type questions, answer system-generated quiz items, or request further explanations within this window.
This design leverages a highly familiar interaction paradigm for large language models, lowering the learning curve and allowing users to focus their attention on the learning content itself rather than on interface mechanics.

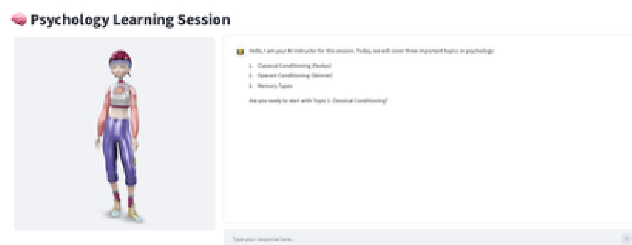

Figure 5. Learning Interface Layout

3.3.3 Functional Iteration and Design Trade-offs.

In earlier versions of the system, we experimented with features such as:

- Avatar speech output (TTS) and speech input (STT)
- Teacher style selection (e.g., Supportive Avatar, Neutral Avatar)
- A manual Participant ID entry window

However, these features were progressively removed during later development and testing for the following primary reasons:

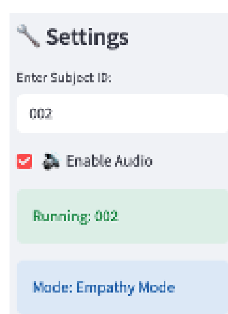

Figure 6. Manual Participant ID Entry, Speech In- and Output

- **Voice Feature Limitations**
  The voice functionality proved highly sensitive to network conditions. Users experienced significant variations in latency. This unpredictability not only impacted user experience but also introduced uncontrolled variables for experimental results.
- **Decreasing Expectancy Effects**
  Allowing users to actively choose a teacher style would make them explicitly aware of their experimental condition. This could influence subjective behavior. To better control variables and ensure experimental validity, this feature was deprecated.

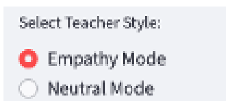

Figure 7. Teacher Style Selection

- **Streamlining Data Integrity**
  The initial design requiring manual Participant ID entry. However, during testing, some users ignored the step entirely, leading to difficulties in data matching and increasing the complexity of subsequent data cleaning and analysis. This functionality was ultimately replaced by the auto-generated ID method to improve data consistency and experimental control.

The final interface design of this system strikes a balance between user experience, learning support, and experimental control, providing a stable and reliable foundation for data collection and analysis in the subsequent user study.

4 Methodology

4.1 Research Design

This study employed a between-subjects, pretest–posttest experimental design to evaluate the effectiveness of an AI-based avatar learning environment and to assess users' perceived empathy and support during the interaction. All participants went through three consecutive stages: completing a pre-survey, engaging with the avatar-based learning experiment, and subsequently completing a post-survey. With this approach, it was possible to compare the participants' knowledge, perceptions, and experiences before and after going through the learning environment. Later, Python programming language was used for the analysis.

4.2 Participants

A total of 30 participants completed all stages of the experiment and were included in the analysis. Participants were recruited through convenience sampling among university peers. They came from a very diverse educational background ranging from business management and engineering to environmental sciences. The majority of participants were between 21 and 29 years old, reflecting a typical student population. The sample shared male and female population quite equally, approximately 63% identified as women, with the remaining participants identifying as men. Almost all participants reported prior experience with artificial intelligence tools, such as ChatGPT, Gemini, Perplexity, etc.
Participants were informed about the purpose of the study and voluntarily agreed to take part in the experiment. The data was handled anonymously using random user IDs. This way it was possible to adhere to ethical considerations.

4.3 Experiment and Survey Design

The whole experiment consisted of three main components: pre-survey, avatar learning environment and post–survey. Below is the breakdown of each section.

4.3.1 Pre-Survey.

The pre-survey was conducted using **Google Forms** and consisted of four main components:

(1) **Demographics**, including age, gender, educational background and prior experience with AI-based tools.
(2) **Baseline knowledge assessment** related to the psychology concepts addressed in the learning environment.
(3) **Learning preferences and expectations** toward AI-supported learning experience. in terms of the type of responses or feedback people want to get.

These measures were used to establish a baseline for both previous knowledge and user perceptions prior to interacting with the avatar. The exact questions can be found in the Appendix.

4.3.2 Avatar Learning Environment.

After completing the pre-survey, participants were instructed to access the LLM-powered avatar-based learning environment through a Streamlit application. In prior, participants were not told if the avatar would be empathic or neutral to avoid creating an expectation of the interaction type or have them behave differently based on the condition. The avatar then led the user through the learning session by asking a set of teaching questions relating to introductory psychology concepts and providing explanations in response to the users' inputs.
The interaction with the avatar was a dialog-based conversational one. The avatar continued to provide explanation of ideas and concepts until the participant demonstrated knowledge of the explained concept in their responses. There was no specific time limit to complete the tasks in the experiment which allowed the participants to interact with the content at their own pace. When the whole learning content had been presented, participants were asked to complete another short quiz at the end of the interaction. The difference in the learning effectiveness and the user's perception was analyzed by comparing the pre-survey with the post-survey.

4.3.3 Post-Survey.
Following the interaction, participants completed a post-survey assessing the avatar learning environment and their subjective experience on the following:

(1) **Post-intervention knowledge**, using the same pre-survey knowledge questions for a proper and fair assessment of the knowledge gained after the experiment.
(2) **Perceived learning effectiveness**, including clarity, usefulness, and engagement of the avatar.
(3) **Perceived empathy and emotional support**, the degree to which the avatar was perceived as understanding and supportive.
(4) **Overall user experience**, including satisfaction, trust, and perceived value of the avatar as a learning tool.

Most questions were measured using a **5 point Likert-scale** system to allow for quantitative comparison across participants during the analysis stage. The exact questions can be found in the Appendix.

4.3.4 Research Question and Hypotheses.
This research mainly focused on examining how the avatar empathy influences learner engagement and interaction behavior in an AI-supported educational environment. The following hypotheses will be addressed throughout this project work.
H1: Learners interacting with an empathic avatar will demonstrate higher behavioral engagement than learners interacting with a neutral avatar.
H2: Learners interacting with an empathic avatar will maintain deeper and more sustained interaction trajectories across the learning session compared to learners interacting with a neutral avatar.

5 Results

5.1 Overview of Data Analysis

A total of 30 valid participant datasets were analyzed (Empathy Mode: n=15; Neutral Mode: n=15). **Table 1** presents the comprehensive descriptive statistics and independent samples t-test results for all measured variables, categorized by learning outcomes, interaction engagement, and cognitive metrics.
In terms of learning effectiveness, **no significant difference was observed** between the Empathy Mode (M = 7.00, SD = 2.27) and the Neutral Mode (M = 6.93, SD = 2.94), t(28) = 0.07, p = .945. Similarly, users in both conditions reported comparable levels of sentiment (p = .682) and confusion rate (p = .610). The total interaction duration and average response time also showed no statistically significant variations between the two groups (see Table 1).
However, analyses revealed **significant disparities** in behavioral engagement and interaction quality. As shown in Table 1, the Empathy Mode elicited a substantially higher volume of user output (p < .001) and communication density (p < .001). In contrast, the Neutral Mode demonstrated significantly higher Lexical Diversity (p < .001). These significant findings are visualized and detailed in the following subsections.

5.2 Interaction Quantity and Duration

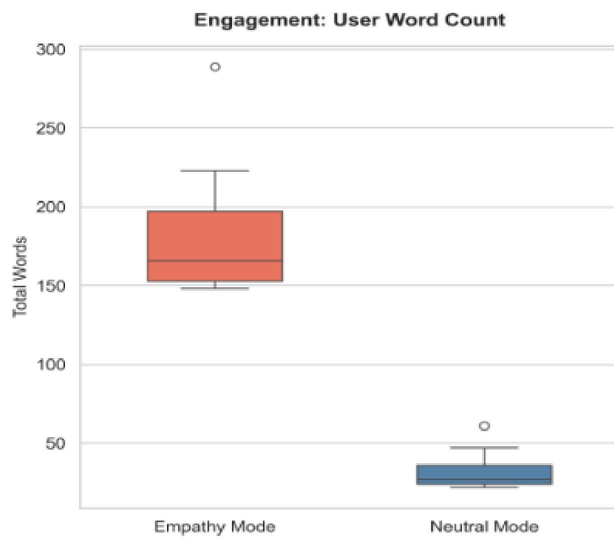

Figure 8. Word Count

The analysis of interaction quantity revealed significant disparities between the two groups. As illustrated in Figure 8, the **Total User Word Count** in the Empathy Mode (M = 181.13, SD = 38.10) was significantly higher than that in the Neutral Mode (M = 31.67, SD = 11.13), t(28) = 14.58, p < .001. Similarly, the **Turn Count (Figure 9** was significantly higher for the Empathy group (M = 27.93, SD = 4.15) compared to the Neutral group (M = 23.40, SD = 2.38), t(28) = 3.67, p = .001. Regarding time investment, **Total Duration (Figure 10)** showed a different trend. Although the Neutral group recorded a longer average duration (M = 904.87s) compared to the Empathy group (M = 701.27s), this difference was not statistically significant (t = -1.79, p = .085), and the Neutral group exhibited a notably larger standard deviation (SD = 347.25), suggesting high variability in user dwell time.

5.3 Interaction Quality and Linguistic Patterns

To investigate the nature of user responses, we analyzed Communication Density and Lexical Diversity. **Figure 11** demonstrates a significant difference in **Communication Density**, defined as user words per minute. The Empathy Mode elicited a higher density (M = 16.98 WPM) compared to the Neutral Mode (M = 2.46 WPM), t(28) = 9.75, p < .001. In contrast, as shown in **Figure 12**, the **Lexical Diversity (TTR)** presented an inverse pattern. The Neutral Mode showed a significantly higher TTR (M = 0.41, SD = 0.15) than the Empathy Mode (M = 0.19, SD = 0.07), t(28) = -5.24, p < .001. This metric indicates that while the Neutral group

Table 1. Comparison of Learning, Interaction, and User Experience Metrics Between Empathy and Neutral Modes

Measure	Empathy Mode Mean (SD)	Neutral Mode Mean (SD)	*t*	*p*	Cohen's *d*
Learning Outcomes					
Learning Score	7.00 (2.27)	6.93 (2.94)	0.07	.945	0.03
Confusion Rate	0.02 (0.04)	0.01 (0.02)	0.52	.610	0.19
Interaction Quantity					
Total Word Count	181.13 (38.10)	31.67 (11.13)	14.58	$< .001^{***}$	5.33
Turn Count	27.93 (4.15)	23.40 (2.38)	3.67	$.001^{**}$	1.34
Total Duration (s)	701.27 (271.67)	904.87 (347.25)	-1.79	.085	-0.65
Interaction Quality & Cognition					
Communication Density (Words/Min)	16.98 (5.54)	2.46 (1.59)	9.75	$< .001^{***}$	3.56
Lexical Diversity (TTR)	0.19 (0.07)	0.41 (0.15)	-5.24	$< .001^{***}$	-1.92
Average Response Time (s)	18.65 (6.81)	24.70 (12.84)	-1.61	.122	-0.59
Questions Asked	1.00 (2.10)	0.20 (0.41)	1.44	.169	0.53
User Experience					
Sentiment Score	7.93 (2.34)	8.27 (2.05)	-0.41	.682	-0.15

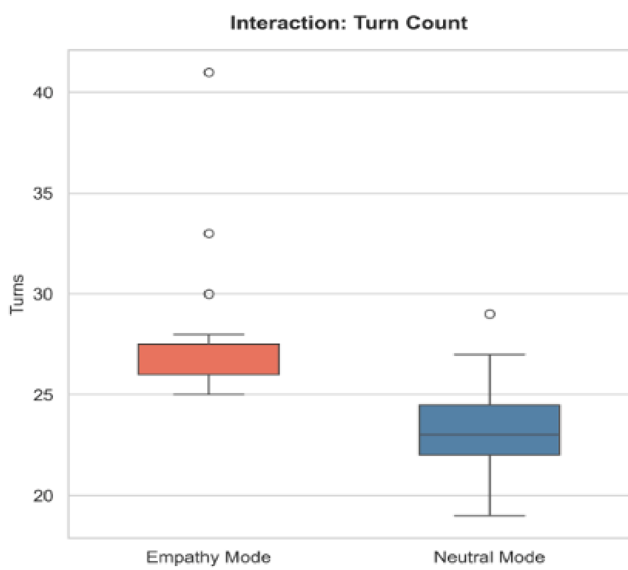

Figure 9. Turn Count

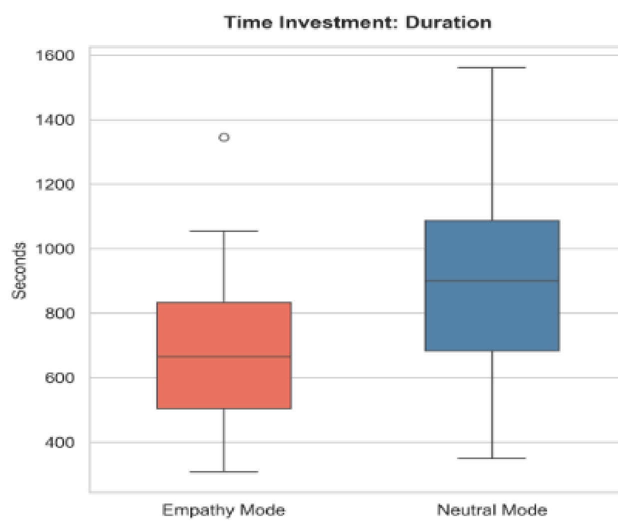

Figure 10. Duration

produced less total text, the vocabulary used was proportionately more diverse, whereas the Empathy group's extensive output contained more repetitive linguistic structures.

5.4 Process Trajectory

The temporal evolution of the interaction was analyzed to understand user engagement over the course of the session. **Figure 13 (Sentiment Trajectory)** displays the sentiment polarity for each turn. Both groups maintained positive sentiment throughout the session, with the Empathy group showing a slightly more stable positive trend, although the overall mean sentiment scores did not differ significantly (p = .682).

Figure 14 (Dialogue Depth Evolution) illustrates the average word count per turn across the timeline. A distinct divergence is observable: the Empathy group (represented by the red line) maintained or increased their word count per turn as the dialogue progressed. Conversely, the Neutral group (blue line) exhibited a flat or declining trend, often reverting to brief responses after the initial turns.

5.5 Correlation Analysis

A Pearson correlation analysis was conducted to examine the relationships between the measured variables, as visualized in the **Heatmap (Figure 15)**. A strong positive correlation was observed between **User Word Count** and **Turn Count** (r > .80), as well as between **User Word Count** and

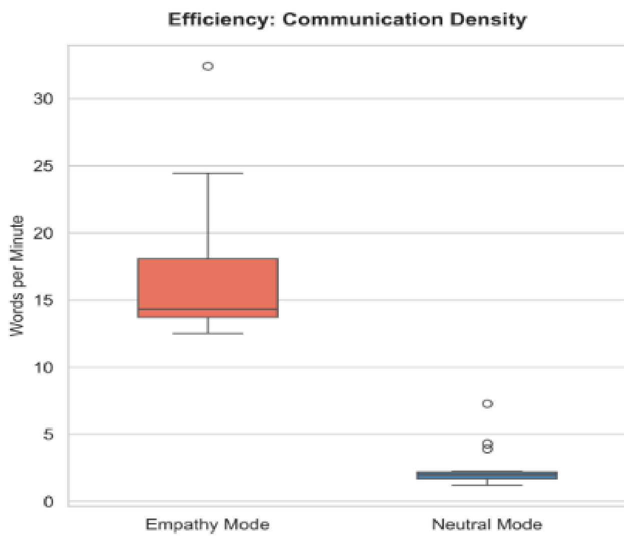

Figure 11. Communication Density

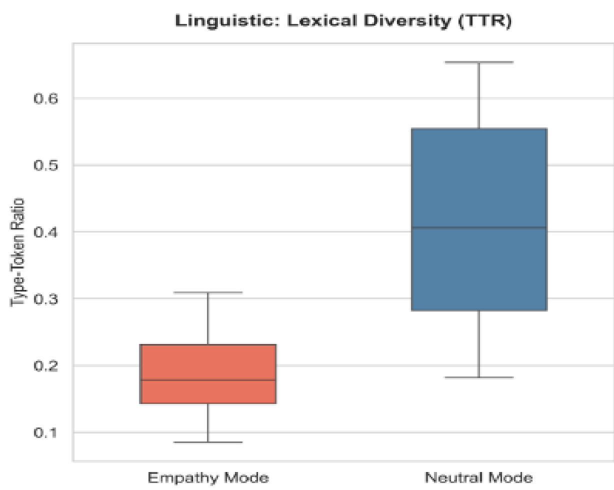

Figure 12. Lexical Diversity (TTR)

Communication Density (r > .90). Notably, Total Duration showed a weak to negligible correlation with **Learning Score** ($r \approx 0.30$), suggesting that merely spending more time in the system did not directly translate to higher test scores. Furthermore, **Lexical Diversity** was negatively correlated with **User Word Count** ($r \approx -0.85$).

5.6 Knowledge Acquisition

To evaluate the overall effectiveness of the educational intervention, we compared the aggregated knowledge accuracy between the Pre-survey (N = 61) and Post-survey (N = 32) phases. Due to the anonymous nature of the survey collection which prevented paired data matching, a group-level analysis was conducted using an independent samples t-test.

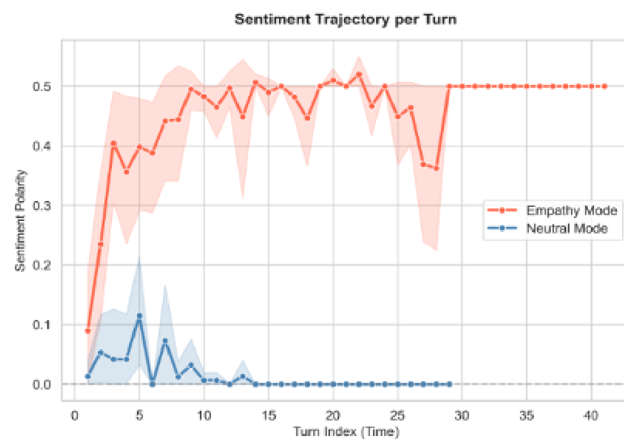

Figure 13. Sentiment Trajectory per Turn

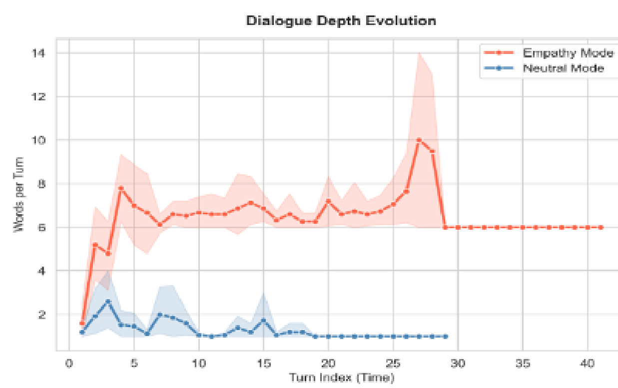

Figure 14. Dialogue Depth Evolution

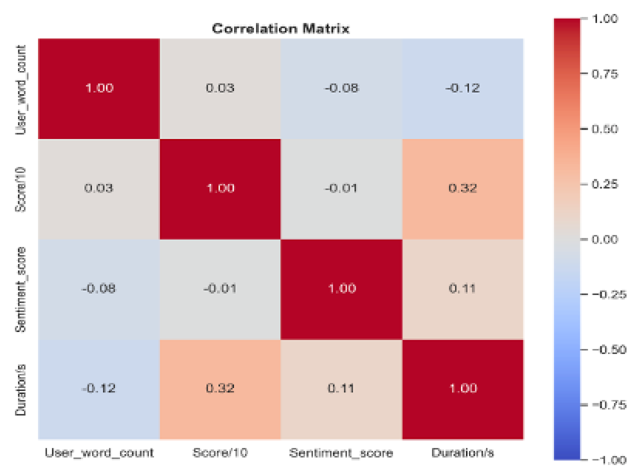

Figure 15. Correlation Heatmap

Table 2 presents the descriptive statistics and t-test results for the knowledge assessment. The Post-survey group demonstrated a higher mean accuracy (M = 78.13%, SD = 29.34%)

Table 2. Comparison of Knowledge Accuracy Between Pre-Survey and Post-Survey

Group	N	Mean Accuracy (%)	SD (%)	*t*	df	*p*	Cohen's *d*
Pre-Survey	61	69.18	32.16	-1.35	70.82	.181	0.29
Post-Survey	32	78.13	29.34				

compared to the Pre-survey baseline (M = 69.18%, SD = 32.16%). Although this difference did not reach statistical significance (t(70.82) = -1.35, p = .181), the analysis revealed a small-to-medium effect size (Cohen's d = 0.29), suggesting a positive trend in knowledge retention following the avatar-based learning session.

6 Discussion

6.1 Interpretation of Results

This study examined how empathic versus neutral AI avatar personalities influence learner engagement and perceived learning in an AI-supported educational environment. While no statistically significant differences were observed in learning scores, sentiment, confusion rates, or response times between the two conditions, clear differences emerged in interaction behavior.

Participants interacting with the empathic avatar produced longer responses and engaged in more conversational turns compared to those interacting with the neutral avatar. These findings suggest that avatar empathy may influence how learners engage with the system, even when measurable short-term learning outcomes remain comparable. In other words, while empathy did not directly translate into higher test performance within this short intervention, it appeared to shape the interaction dynamics of the learning experience.

One possible interpretation is that empathic language may be perceived as more supportive or socially responsive, which could encourage learners to elaborate more extensively in their responses. However, psychological comfort or perceived social presence were not directly measured in this study; therefore, this interpretation remains speculative and should be examined in future research.

Interestingly, although empathic interactions were characterized by greater response length and conversational density, time spent in the system was not strongly associated with measurable learning gains. This suggests that extended interaction alone does not necessarily produce improved short-term knowledge retention. Instead, engagement and learning performance may represent related but distinct dimensions of AI-supported education.

Overall, the findings indicate that empathic AI tutoring may function more as an engagement catalyst rather than a direct driver of short-term measurable learning performance. This distinction is important when evaluating the effectiveness of emotionally expressive AI systems.

6.2 Implications for Human–AI Interaction Design

The results provide several implications for the design of Human–AI Interaction in educational contexts.

First, the findings suggest that emotional tone in AI tutors can meaningfully shape user interaction behavior. Even in the absence of statistically significant differences in learning outcomes, empathic design appears to influence how users communicate with the system. Designers of educational AI systems should therefore consider engagement-related metrics—such as conversational depth, response length, and interaction patterns—alongside traditional performance indicators.

Second, the results indicate that affective design alone may not be sufficient to enhance short-term learning outcomes. In this study, empathy did not significantly improve knowledge test scores compared to a neutral instructional style. This suggests that emotional expressiveness should be integrated with structured pedagogical strategies, scaffolding techniques, or adaptive feedback mechanisms to produce measurable learning gains.

Third, the study highlights the importance of distinguishing between engagement and effectiveness. An AI tutor may successfully increase user interaction and conversational participation without necessarily improving immediate test performance. For researchers and practitioners, this distinction emphasizes the need for multi-dimensional evaluation frameworks when assessing AI-supported learning systems.

Finally, the findings contribute to ongoing discussions in HAI research about the role of social presence in AI systems. While emotionally expressive avatars may enhance user engagement, their impact on cognitive learning outcomes may depend on contextual variables such as intervention duration, task complexity, and learner characteristics.

6.3 Limitations

Several limitations should be considered when interpreting the findings of this study. First, the sample size was relatively small (N = 30), which limits statistical power and may reduce the likelihood of detecting subtle effects. Larger samples would allow for more robust statistical testing and subgroup analyses.

Second, the intervention was short-term and limited to introductory psychology topics. The absence of significant learning differences between conditions may reflect the brief exposure duration rather than the ineffectiveness of empathic design. Longitudinal studies are needed to examine whether empathy influences knowledge retention over extended periods.

Third, the study relied primarily on self-reported measures and short-term knowledge tests. While these instruments are commonly used in HAI and educational research, they may not fully capture deeper cognitive processing, long-term retention, or transfer of learning.

Fourth, participants were mostly university students with prior experience using AI tools. This familiarity may have influenced expectations and interaction styles, potentially limiting the generalizability of the findings to other populations.

Finally, although the system is controlled for teaching content across conditions, subtle variations in interaction trajectories may naturally emerge in dialog-based LLM systems. While prompt engineering ensures structural consistency, real-time conversational dynamics may introduce variability that is difficult to standardize fully.

Taken together, these limitations suggest that the findings should be interpreted as exploratory evidence regarding engagement effects rather than definitive conclusions about the role of empathy in AI-supported learning.

7 Conclusion

7.1 Future Work

Future research should address several existing research gaps in the application of LLM-based pedagogical agents in education. While current studies demonstrate positive short-term effects on learning outcomes, there is still limited evidence regarding the long-term impact of these systems on knowledge retention, critical thinking development, and learner independence.

In addition, although social cues in affective pedagogical agents have been shown to improve engagement and motivation, there is a lack of research identifying the optimal design and level of human-likeness required for different learner groups and educational contexts.

Additionally, a notable gap in research is the scarcity of practical implementations in real classroom settings, as many experiments are still performed in controlled environments. Consequently, future research should prioritize studies conducted in actual educational contexts, create standardized assessment models for Human–AI educational technologies, and address ethical issues such as protecting data privacy, ensuring transparency, and promoting responsible AI usage to support a sustainable and reliable integration of technology in education.

7.2 Summary of Findings

In conclusion, this study reveals that empathetic AI tutoring does not significantly impact learning outcomes, user sentiment, confusion levels, or response times when compared to neutral interaction modes.

However, empathy plays a crucial role in shaping how students engage with learning systems. Students interacting with the Empathy Mode demonstrated markedly higher behavioral engagement—producing more words, initiating more conversational turns, and maintaining denser communication throughout their sessions.

Conversely, the Neutral Mode encouraged greater diversity, though with reduced overall interaction volume. Temporal patterns further distinguished the two conditions: empathetic interactions sustained or deepened dialogue over time, while neutral interactions led to progressively shorter responses.

Correlation analyses reinforced the interconnected nature of engagement metrics, yet revealed that time spent in the system bore little relationship to actual learning gains. While both groups showed positive trends in knowledge improvement from pre- to post-intervention, these changes did not reach statistical significance.

Ultimately, these findings suggest that empathetic AI tutoring functions primarily as an engagement catalyst rather than a direct driver of measurable learning performance in short-term educational interventions.

References

[1] Maqsood Ahmed, Munazza Ambreen, and Ishtiaq Hussain. 2018. Gender Differentials Among Teachers' Classroom Management Strategies In Pakistani Context. *Journal of Education and Educational Development* 5 (12 2018), 178. doi:10.22555/joeed.v5i2.2253

[2] David Budgen and Pearl Brereton. 2006. Performing systematic literature reviews in software engineering. *Proceedings - International Conference on Software Engineering* 2006, 1051–1052. doi:10.1145/1134285.1134500

[3] Jennifer A. Fredricks, Phyllis C. Blumenfeld, and Alison H. Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research* 74, 1 (2004), 59–109. doi:10.3102/00346543074001059

[4] Martin Saerbeck, Tom Schut, Christoph Bartneck, and Maddy Janse. 2010. Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1613–1622. doi:10.1145/1753326.1753566

[5] Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education* 16, 1 (2019), 1–27. doi:10.1186/s41239-019-0171-0

Appendix: Project Repository

The full versions of the pre-survey and post-survey questionnaires are available online:

Pre-survey

Post-survey

The full implementation of the system, including version history and contribution records, is available at:

GitHub from Yiyang Xie

GitHub from Yusong Yang

Contribution Overview

The following list summarizes the main contributions of each group member throughout the project.

Araks Karapetyan

- Shared workspace & group coordination: Organized internal group workflow, established shared online

workspace (Word, PPT), and coordinated group communication.
- Contributed to the preparation of presentation slides.
- Participated to all the team meetings.
- Contributed to selecting and narrowing learning topics to core psychology concepts.
- Contributed in ideation of survey questions and gave feedback.
- Designed learning scenario 1 and 2, including learning content, prompts, and quizzes.
- Designed the post-study questionnaire.
- Contributed to the design of pre-test and post-test instruments.
- Contributed in finding the research question and hypotheses for the final report.
- Participated in recruiting participants for the user study. (12 people)
- Contributed to the midterm report preparation.
- Authored Abstract, Appendix and Sections 4 of the final report.

Brishila Firza

- Supported group coordination and shared workspace management.
- Participated to all the team meetings.
- Contributed to the midterm report preparation.
- Contributed to selecting and narrowing learning topics to core psychology concepts.
- Served as the presenter for Pitch.
- Contributed in ideation of survey questions and gave feedback.
- Contributed in finding the research question and hypotheses for the final report.
- Designed learning scenario 3, including learning content, prompts, and quizzes.
- Participated in recruiting participants for the user study. (9 people)
- Video creation and upload to Moodle: Solely drafted, created and edited the project video.
- Participated in experiments for bonus engagement points.
- Authored Sections 2, 7 of the final report.

Nergis Bilge

- Supported group coordination and shared workspace management.
- Contributed to preparation of the pitch presentations and midterm report.
- Participated to all the team meetings.
- Designed and refined the pre-study questionnaire, including learning questions and Likert-scale items.
- Contributed to selecting and narrowing learning topics to core psychology concepts.
- Contributed in ideation of survey questions and gave feedback.
- Tested the avatar-based learning system and provided feedback on the interaction flow.
- Contributed to the design of pre-test and post-test instruments.
- Participated in recruiting participants for the user study. (8 people)
- Supported the alignment of post-study questionnaire with pre-study questionnaire learning measures.
- Authored Sections 1, 6 of the final report.

Yiyang Xie

- Implementation contributions are documented in the project GitHub repository (see Appendix).
- Deployed the LLM remotely and configured API access. Integrated the chatbot with the avatar in a web-based user interface.
- Implemented the initial (1st. version) learning system architecture.
- Designed the basic components for the system, including two types of chatbot, system prompts, and avatar design.
- Performed system debugging and performance optimization.
- Designed learning scenarios 4, 5 and 6, including learning content, prompts, and quizzes.
- Contributed to the preparation of presentation slides.
- Participated in recruiting participants for the user study. (7 people)
- Designed overall report structure, discussed adjustments with supervisor.
- Drafted full text version for two reports, the midterm and final report.
- Authored Section 3 of the final report.
- Implemented both mid-term and final reports using LaTeX.

Yusong Yang

- Implementation contributions are documented in the project GitHub repository (see Appendix).
- Further developed the learning system and extended its functionality.
- Implemented speech-based interaction, enabling voice input and output.
- Designed and implemented automatic user ID generation and data linking.
- Integrated pre-test and post-test into the learning system.
- Implemented automatic synchronization of user study data to a shared Excel file.
- Contributed to the preparation of presentation slides.
- Contributed to the midterm report preparation.

- Performed system debugging and performance optimization.
- Designed the automated data logging structure, ensuring collection of behavioral metrics such as duration, word count, response time, sentiment score, confusion rate, and turn count.
- Participated in recruiting participants for the user study. (10 people)
- Conducted quantitative user study data analysis, including independent-samples and Welch's t-tests.
- Structured and formalized the Results section, organizing statistical findings into coherent subsections and produced visualizations.
- Authored Section 5 of the final report.