Pilot Study Report — TEACHme

Andrea Federici (andrea3.federici@mail.polimi.it)
Alireza Yahyanejad (alireza.yahyanejad@mail.polimi.it)
Mahdi Valadan (mohammadmahdi.valadan@mail.polimi.it)
Paolo Pertino (paolo.pertino@mail.polimi.it)
Pietro Moroni (pietroguglielmo.moroni@mail.polimi.it)

July 2024

1 Introduction

The purpose of this pilot study was to evaluate the TEACHme application in terms of usability, conversational feedback relevance, and overall user experience. The study aimed to identify strengths and areas for improvement in order to refine the application.

2 Methodology

2.1 Participants

The pilot study included four participants selected to represent a diverse range of demographics and technical proficiencies. The participants varied in age and included both male and female genders. Occupations among the participants included a bank intern, students, and a physical education teacher in a primary school. Technical proficiency levels ranged from moderately proficient to very proficient, ensuring a broad spectrum of user experiences with the TEACHme application. A detailed description of each participant, including specific characteristics, will be illustrated in a later section.

2.2 Procedure

Participants signed up, logged in, and were assigned to three different conversation scenarios. Feedback was collected after each conversation, focusing on conversation feedback relevance, pronunciation and synonym suggestions, difficulty level, and general usability.

Each participant had the following conversations with the related difficulty levels:

- **Participant 1:**
  - Conversation 1: Last summer holidays (Difficulty: easy)
  - Conversation 2: Hobbies (Difficulty: medium)
  - Conversation 3: Favorite food (Difficulty: hard)
- **Participant 2:**
  - Conversation 1: Last summer holidays (Difficulty: hard)
  - Conversation 2: Hobbies (Difficulty: easy)
  - Conversation 3: Favorite food (Difficulty: medium)
- **Participant 3:**
  - Conversation 1: Last summer holidays (Difficulty: hard)
  - Conversation 2: Hobbies (Difficulty: easy)
  - Conversation 3: Favorite food (Difficulty: medium)
- **Participant 4:**
  - Conversation 1: Last summer holidays (Difficulty: medium)
  - Conversation 2: Hobbies (Difficulty: hard)
  - Conversation 3: Favorite food (Difficulty: easy)

The duration was set to 5 minutes for each conversation.

3 Data Collection

3.1 Demographic Data

The participants' demographic information was collected to provide context for their feedback.

3.2 Usability Feedback

Participants were asked to rate the ease of navigation, report any issues or bugs, and provide overall usability feedback.

3.3 Conversational Feedback

Participants' responses were collected on the relevance of the conversation feedback and the accuracy of pronunciation/synonym suggestions.

3.4 Difficulty Levels

Participants were asked to assign difficulty levels to each conversation (easy, medium, hard).

3.5 Naturalness of Conversations

Participants provided feedback on whether conversations felt natural and human-like.

4 Results

4.1 Participant Demographics

- **Participant 1:**
  - Age: 23 years
  - Gender: Female
  - Occupation: Intern in a bank in Switzerland
  - Technical Proficiency: 3
- **Participant 2:**
  - Age: 25 years
  - Gender: Male
  - Occupation: Student
  - Technical Proficiency: 3/4
- **Participant 3:**
  - Age: 23 years
  - Gender: Male
  - Occupation: Student
  - Technical Proficiency: 4
- **Participant 4:**
  - Age: 25 years
  - Gender: Male
  - Occupation: Student and physical education teacher in a primary school
  - Technical Proficiency: 3/4

4.2 Usability Feedback Analysis

- **Participant 1:**
  - **Ease of navigation:** Rated 3-4. The process to sign up, log in, and navigate the home page was clear. However, there were issues understanding some buttons on the conversation page. Adding instructions could improve clarity.

  - **Issues reported:** The participant experienced interruptions when thinking about what to say. Sometimes, the overall feedback was not immediately visible on the feedback page and required reloading.

- **Participant 2:**
  - **Ease of navigation:** Rated 4. Fairly easy to navigate, with no major issues.
  - **Issues reported:** The microphone would occasionally cut off, urging the participant to think quickly. Pronunciation of Italian location names was problematic.

- **Participant 3:**
  - **Ease of navigation:** Rated 5. Easy to navigate, but some buttons were in unintuitive places.
  - **Issues reported:** The participant experienced significant microphone issues, with frequent interruptions.

- **Participant 4:**
  - **Ease of navigation:** Rated 4-5. The application was easy to navigate but there was some confusion about how to use the buttons on the conversation screen.
  - **Issues reported:** While thinking, the application would sometimes send the message before the participant finished speaking.

4.3 Conversational Feedback Analysis

- **Participant 1:**
  - **Relevance of feedback:** The feedback was informative and relevant, helping the participant discover new ways to say things.
  - **Pronunciation and synonym suggestions:** Synonym suggestions were accurate and helpful. Pronunciation challenges were sometimes triggered by non-English words, reducing their informativeness.

- **Participant 2:**
  - **Relevance of feedback:** The feedback was somewhat generic but useful overall.
  - **Pronunciation and synonym suggestions:** Some suggestions were accurate and helpful, while others were less so.

- **Participant 3:**
  - **Relevance of feedback:** Feedback was skewed by microphone issues, leading to frustration.

    - **Pronunciation and synonym suggestions:** Generally accurate, with some suggestions more interesting than others.

- **Participant 4:**
    - **Relevance of feedback:** The feedback was relevant and useful for improving English speaking skills.
    - **Pronunciation and synonym suggestions:** Synonym suggestions were accurate, but some were difficult to understand and it was not clear how to use them in a sentence.

4.4 Difficulty Level Analysis

- **Participant 1:**
    - Easy conversation: Rated as Easy
    - Medium conversation: Rated as Hard
    - Hard conversation: Rated as Medium
- **Participant 2:**
    - Difficulty levels not explicitly rated.
- **Participant 3:**
    - Difficulty levels not explicitly rated.
- **Participant 4:**
    - Easy conversation: Rated as Easy
    - Medium conversation: Rated as Hard
    - Hard conversation: Rated as Medium

4.5 Naturalness and Human-Like Quality

- **Participant 1:**
    - The conversation was engaging and stimulated continuous discussion. However, the participant noted that the agent did not naturally integrate personal experiences or feelings into the conversation unless explicitly prompted.
- **Participant 2:**
    - The conversation felt natural overall.
- **Participant 3:**
    - The conversation was mostly human-like, but safety principles were too strict, and some answers felt empty.

- **Participant 4:**
  - The conversation felt quite natural. However, the participant sometimes preferred shorter replies from the agent.

4.6 Additional Comments and Suggestions for Improvement

- **Participant 1:**
  - Overall, the participant liked the application and suggested adding functionalities for self-study, improving UI clarity, and making the agent's responses more natural.
- **Participant 2:**
  - Fix bugs and improve the interface.
- **Participant 3:**
  - Fix bugs, improve the interface, and make the conversation feel more natural by addressing the strictness of safety principles.
- **Participant 4:**
  - Improve the time the agent waits while the user is thinking and consider allowing longer thinking times.

5 Discussion

5.1 Key Findings

- Positive feedback on usability and conversational relevance.
- Some UI elements were unclear, and minor bugs were reported.

5.2 Strengths and Weaknesses

- **Strengths:** Ease of use, engaging conversations, accurate synonym suggestions.
- **Weaknesses:** UI clarity issues, occasional bugs, and the need for more natural conversational integration.

5.3 Participant Feedback

The participants suggested improvements in UI clarity, more natural conversation integration, and functionalities for self-study.

6 Conclusion

6.1 Summary of Findings

The TEACHme application showed promise in usability and engaging conversations but required minor improvements in UI clarity, natural conversation integration, and functionalities for self-study.

6.2 Recommendations

- Improve UI clarity by adding instructions or on-hover descriptions.
- Address minor bugs, such as interruptions during conversation and delayed feedback visibility.
- Enhance the naturalness of conversations by integrating personal thoughts and experiences into the agent's responses.
- Add functionalities for self-study to allow students to practice independently.
- Adjust the time the agent waits while the user is thinking to allow for longer thinking times.

7 Appendices

7.1 Appendix A: Participant Responses

7.1.1 Participant 1

- Age: 23
- Gender: Female
- Occupation: Intern in a bank in Switzerland
- Technical proficiency: 3 (Moderately Proficient)
- Was the conversation feedback relevant in identifying your shortcomings?: Informative and relevant.
- Were pronunciation/synonym suggestions accurate?: Synonym suggestions accurate; pronunciation challenges sometimes triggered by non-English words.
- How easy was it to navigate the application (1 - Very difficult, 5 - Very easy): 3-4
- Did you encounter any issues or bugs?: Interruptions when thinking, delayed feedback visibility.

- Did the conversation feel natural and human-like to you?: Engaging but lacked integration of personal experiences.
- Any additional comments or suggestions for improvement?: Add functionalities for self-study, improve UI clarity, and enhance conversation naturalness.
- Difficulty matching (ground truth — predicted by the user):
  - Last summer holidays: Easy — Easy
  - Hobbies: Medium — Hard
  - Favorite food: Hard — Medium

7.1.2 Participant 2

- Age: 25
- Gender: Male
- Occupation: Student
- Technical proficiency: 3/4 (Moderately to Very Proficient)
- Was the conversation feedback relevant in identifying your shortcomings?: Somewhat generic but useful overall.
- Were pronunciation/synonym suggestions accurate?: Mixed accuracy, generally aware of suggestions.
- How easy was it to navigate the application (1 - Very difficult, 5 - Very easy): 4
- Did you encounter any issues or bugs?: Microphone issues, pronunciation problems with Italian names.
- Did the conversation feel natural and human-like to you?: Felt natural overall.
- Any additional comments or suggestions for improvement?: Fix bugs, improve interface.

7.1.3 Participant 3

- Age: 23
- Gender: Male
- Occupation: Student
- Technical proficiency: 4 (Very Proficient)

- Was the conversation feedback relevant in identifying your shortcomings?: Feedback skewed by microphone issues.
- Were pronunciation/synonym suggestions accurate?: Generally accurate, some more interesting than others.
- How easy was it to navigate the application (1 - Very difficult, 5 - Very easy): 5
- Did you encounter any issues or bugs?: Significant microphone issues, frequent interruptions.
- Did the conversation feel natural and human-like to you?: Mostly human-like but with strict safety principles and some empty responses.
- Any additional comments or suggestions for improvement?: Fix bugs, improve interface, make conversations more natural.

7.1.4 Participant 4

- Age: 25
- Gender: Male
- Occupation: Student and physical education teacher in a primary school
- Technical proficiency: 3/4 (Moderately to Very Proficient)
- Was the conversation feedback relevant in identifying your shortcomings?: Relevant and useful for improving English speaking.
- Were pronunciation/synonym suggestions accurate?: Synonym suggestions were accurate, but some were difficult to understand and use in a sentence.
- How easy was it to navigate the application (1 - Very difficult, 5 - Very easy): 4-5
- Did you encounter any issues or bugs?: While thinking, the application would sometimes send the message before the participant finished speaking.
- Did the conversation feel natural and human-like to you?: The conversation felt quite natural. However, the participant sometimes preferred shorter replies from the agent.
- Any additional comments or suggestions for improvement?: Improve the time the agent waits while the user is thinking and consider allowing longer thinking times.
- Difficulty matching (Conversation scenario: ground truth — predicted by the user):

- Favorite food: Easy — Easy
- Last summer holidays: Medium — Hard
- Hobbies: Hard — Medium