AI-LLP: A Memory-Augmented Conversational Agent for AI-Assisted Language Education

Jordy del Castilho (5550394) Liwia Padowska (5523192)
Yizhen Zang (6140130) Zeryab Alam (5486548)
Delft University of Technology

1 INTRODUCTION

Language learning has advanced significantly with the integration of artificial intelligence. Traditional language learning applications have provided structured educational resources, but they often fall short in providing personalized and adaptive learning experiences [1]. Although several studies have demonstrated AI's effectiveness in language education [15, 20], critical gaps remain in ensuring engagement over the long term.

One such underexplored area is the role of memory-enhanced systems in shaping user experience and sustaining engagement over time. Learning is a continuous process, and a good learning environment should adjust dynamically to students' emotional and cognitive states while considering past interactions [7]. Memory, particularly episodic memory, plays a crucial role in human learning by allowing individuals to recall past experiences and apply that knowledge in future interactions [17]. Moreover, semantic memory, which refers to generalized knowledge and facts (e.g., vocabulary and grammar rules), is equally essential in educational contexts [2]. Together, these memory systems mirror essential aspects of human cognition: episodic memory fosters relational continuity and motivation, while semantic memory supports conceptual understanding and skill development. An effective AI tutor should incorporate both memory types to support dynamic dialogue and knowledge transfer.

However, existing AI-based language learning platforms such as Duolingo mainly focus on predefined lesson plans and generic instructional models that lack personalization. These applications often rely on repetitive exercises, slow progression, and rigid task structures, failing to personalize learning paths or leverage memory for user proficiency.

To address these limitations, we introduce AI-based Language Learning Partner (AI-LLP), a memory-augmented AI tutor for personalized language learning. AI-LLP incorporates a dual memory architecture, short-term memory (STM) to maintain coherence within sessions and long-term memory (LTM) to personalize learning across sessions based on prior user interactions. This dialogue system can store and retrieve past conversations, track learner progress, and adapt lesson plans based on prior user behavior.

Building on this architecture, this study investigates the research question: **Does a memory-augmented conversational agent enhance user experience (UX) in AI-based language learning systems?** We address this by evaluating AI-LLP through a within-subject user study, assessing how memory influences perceived usability, relational quality, and overall engagement. We aim to demonstrate the value of conversational memory in moving AI tutors beyond static instruction toward dynamic and learner-centered support.

2 BACKGROUND

Numerous studies have explored how memory can enhance human-agent interaction, particularly in language learning contexts. This section highlights key findings and research in these areas.

2.1 Memory in Conversational Agents

Memory enables conversational agents to recall user preferences and past interactions. Researchers often distinguish between short-term memory (STM) and long-term memory (LTM) in these systems. STM typically stores information about the current session while LTM maintains information across sessions.

Unlike static databases, memory must be dynamic and adaptable to evolving interactions. Nuxoll and Laird [14] advocated for episodic memory in intelligent agents. Campos et al. [3] observed that it is difficult for agents to maintain socially coherent relationships if memory is not structured to adapt to past conversations.

While both are critical, LTM is especially important for building user trust and boosting engagement. Bang et al. [8] presented a framework with a personal knowledge database (PKB) and a forgetting model for personalized interactions. Similarly, Kasap and Magnenat-Thalmann [7] emphasized the importance of episodic memory in sustaining long-term engagement, demonstrating its effectiveness through the photography tutor Eva.

From a cognitive perspective, memory helps with building shared understanding over time. McKinley et al. [12] found that both item and context memory improve mutual understanding between dialogue partners. This aligns with the Belief-Desire-Intention theory, where agents update their knowledge based on shared beliefs [18]. Le Bigot et al. [9] further showed that collaborative experiences shape conversational memory. Richards and Bransky [16] highlighted the importance of memory management. Their study showed that while accurate recall enhances trust and engagement, memory failures frustrate users and reduce believability.

2.2 Proactiveness and Engagement

Proactiveness is another key capability of engaging agents. Moon et al. [13] argued that effective conversational agents should utilize memory to maintain context and guide users proactively through their learning journey, rather than merely reacting to queries. Deng et al. [4] proposed a human-centered framework enabling AI to simulate internal thought processes, allowing for proactive interventions. However, excessive proactivity may be counterproductive. Liu et al. [11] cautioned that proactive agents must balance initiative-taking with user expectations to avoid being perceived as intrusive.

Age also affects how proactivity and memory are perceived. A study by Leite et al. [10] found that older children responded positively to agents with persistent memory, while younger children preferred agents with less memory-based personalization. This suggests a need to adapt proactive strategies based on user demographics.

2.3 Language Learning and AI Integration

AI has been widely applied to language learning. Studies by Qiao & Zhao [15] and Wei [20] highlighted improvements in learners' speaking skills and motivation through AI-driven platforms like Duolingo. However, these systems often lack memory mechanisms that could personalize instruction and reinforce learning through contextual recall. A recent review by Alhusaiyan [1] identified persistent challenges in dialogic competence and teacher intervention.

Despite AI's effectiveness in language learning, most of the research on language learning was done in the English as a Foreign Language (EFL) setting, leaving a gap in assessing its effectiveness for other languages. Building on this, we aim to explore whether memory-augmented conversational agents can improve user experience in the context of AI-assisted Mandarin learning.

2.4 Frameworks and Tools

Memory-augmented conversational agents commonly rely on toolkits and frameworks such as FAISS[1] for vector database, NLTK[2] for natural language processing (NLP), LangChain[3] for LLMs arrangement, and PyTorch[4] for GPU-accelerated processing.

3 METHODOLOGY

To explore the impact of memory on user experience in language learning, we developed AI-LLP. This section outlines the system architecture, memory design, and user study setup used to evaluate AI-LLP's effectiveness.

3.1 System Architecture

The overall system architecture is illustrated in Figure 1.

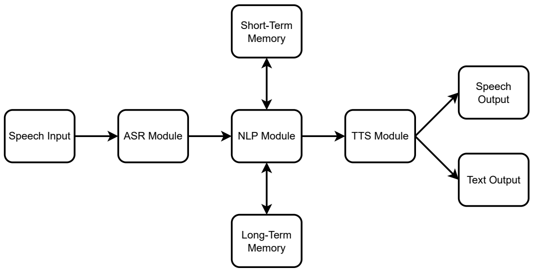

Figure 1: AI-LLP System Architecture.

The system begins by capturing speech input, which is then transcribed into text through the Automatic Speech Recognition (ASR) module, powered by Whisper[5]. This transcribed text is forwarded to the NLP module, utilizing Llama 3.1[6] through Ollama[7] for easy integration and quantization, where the system analyzes the text to extract intent and entities.

The NLP module interacts with both STM and LTM. Each user has a private vector database for confidentiality and security. STM is kept as a context window to the LLM to provide contextual awareness within a session. LTM is a vector database used to store and retrieve information. The NLP module queries it semantically through similarity search or saves user-specific data like preferences through function calls. This enables coherence across sessions. Each session is stored in LTM once it ends, allowing the agent to pick up where it left off. The NLP module generates a response that is passed to the text-to-speech (TTS) module for voice and text output.

3.1.1 *Memory Module.*

Components AI-LLP's memory module consists of STM and LTM, enabling the agent to retain context across sessions and personalize interactions [10]. STM is structured around a context window with a fixed size that is passed to the LLM at every invocation before the user message. LTM is structured around three key building blocks, as shown in Figure 2. The first component is the vector database, responsible for storing episodic interactions, user preferences, learning history, and personalized instructions. The second building block is the tokenizer, which converts text into unique tokens, representing words or word fragments as unique numbers. Lastly, the embedding module maps tokens to high-dimensional vectors, where the vector represents the semantic meaning of the token. This module will also add context between tokens within a sentence or passage.

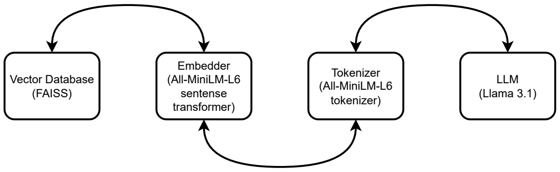
Vector Database (FAISS)
Embedder (All-MiniLM-L6 sentense transformer)
Tokenizer (All-MiniLM-L6 tokenizer)
LLM (Llama 3.1)

Figure 2: Building blocks of long-term memory.

Memory Implementation To implement the vector database for LTM, we used the FAISS framework, which supports high-performance similarity-based retrieval of stored embeddings [5]. For the embedding model, we selected All-MiniLM-L6[8], a compact and efficient transformer that offers fast inference. While it doesn't provide the highest accuracy compared to larger models, its performance is sufficient for our purposes. Since our use case focuses on semantic similarity rather than precise semantic understanding, All-MiniLM-L6 yields reliable results.

STM is implemented using LangChains' Conversation Buffer Memory[9] module, which maintains a fixed-length sliding window of recent conversational turns. This ensures the agent retains context across short interactions while keeping the prompt size manageable.

[1] https://github.com/facebookresearch/faiss
[2] https://www.nltk.org/
[3] https://www.langchain.com/
[4] https://pytorch.org/
[5] https://github.com/openai/whisper
[6] https://huggingface.co/meta-llama/Llama-3.1-8B
[7] https://ollama.com/
[8] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[9] https://python.langchain.com/v0.1/docs/modules/memory/types/buffer/

Memory Usage The agent has bidirectional access to LTM, enabling both retrieval and storage of user-related information. This is done through tool functions that accept queries to either insert or search content in the vector store. If the underlying LLM is trained to perform function calls, it can invoke these tools autonomously. We use LangChain to create a tool-calling agent, which listens for structured (JSON) function calls in the model's output, dispatches the appropriate memory function, and returns the results to the LLM for incorporation into its next response.

After each session, the entire conversation history is automatically archived into LTM for the agent to utilize in future sessions. This process is system-driven rather than agent-initiated. Through empirical testing, we found that storing each message from STM as a separate entry in LTM yielded the best performance. Longer memory chunks tended to confuse the LLM, sometimes causing it to lose track of the memory context entirely. Breaking memory into message-level units helped fix this problem and made retrieval more relevant.

3.1.2 Perception Module.

Perception plays an important role in facilitating natural and context-aware interactions [19]. The core component is the Automatic Speech Recognition (ASR) module. We used several state-of-the-art tools and frameworks, including Whisper, sounddevice[10], librosa[11], and noisereduce[12]. Instead of manually extracting features like Mel-frequency cepstral coefficients (MFCCs), the Whisper model directly processes the audio and extracts representations. It's chosen for its robustness to variations in speech (e.g., different accents).

Before audio transcription, we handle noise cancellation using the librosa and noisereduce libraries, which enhance speech clarity for better transcription accuracy. Once the speech is transcribed, it's directly passed to the NLP module.

3.1.3 Dialogue Flow.

To show how AI-LLP interacts with the user, we present basic elements of a possible dialogue flow in Figure 3.

The system starts by greeting the user and asking if they're ready to begin. It then checks whether the user is new or returning. New users receive an introduction to the agent's capabilities, while returning users are offered the option to resume or start fresh. During the session, the agent presents level-appropriate prompts or exercises. It evaluates the user's response and gives positive feedback if correct, or hints and explanations if incorrect. The session concludes with a summary and review of the topics covered.

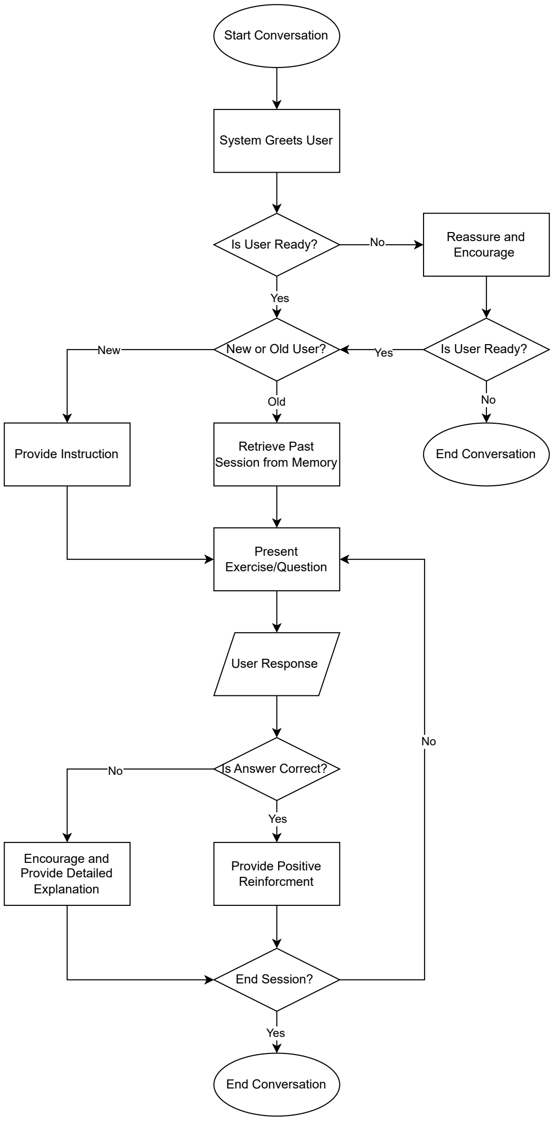

Figure 3: Simplified Dialogue Flow of AI-LLP.

3.2 User Study

To assess the impact of memory augmentation in AI-LLP on user experience, we conducted a comparative study, comparing participant responses across two conditions: **Version A (non-memory)** and **Version B (memory-augmented)**. Given the time constraints of the study, participants only learned Mandarin.

3.2.1 Study Design.

A within-subject design was used, where each participant interacted with both versions. This allows for direct comparison to minimize the impact of individual differences in Mandarin proficiency, AI familiarity, and learning aptitude. Moreover, the sequence in which participants interacted with the two versions was randomized to control for potential order effects.

The independent variable is the presence of memory capacity in the agent. The dependent variables are the self-reported measures, collected through post-interaction questionnaires. Confounding variables include prior Mandarin proficiency and experience with AI language tools.

Besides, to ensure feasibility, we also conducted a pilot study with a small group (3 people) to get feedback on our procedures, questionnaires, and overall testing setups.

[10] https://python-sounddevice.readthedocs.io/en/0.5.1/
[11] https://librosa.org/doc/latest/index.htm
[12] https://pypi.org/project/noisereduce/

3.2.2 Participants.

We recruited 30 participants, 5 female and 25 male (ages 18-32), primarily CS and DSAIT Master's students at TU Delft. Their background in conversational agents reduced confounding factors related to unfamiliarity. To provide comparative insights, we recruited participants with varying levels of Mandarin proficiency and prior experience with language learning tools.

3.2.3 Measures.

AttrakDiff Questionnaires To assess the impact of memory augmentation on user experience, we looked into the responses from AttrakDiff questionnaires. AttrakDiff [6] is a well-established and validated questionnaire, suitable for evaluating user experience in interactive systems like AI-LLP. It contains 28 bipolar adjective pairs (e.g., technical – human), with each rated on a 7-point semantic differential scale. It measures four core subscales:

- Pragmatic Quality (PQ): usability and efficiency
- Hedonic Quality - Stimulation (HQS): novelty and stimulation
- Hedonic Quality - Identity (HQI): self-identification and relatedness
- Attractiveness (ATT): overall impression

We developed digital versions of these questionnaires using Microsoft Forms, as recommended by TU Delft [13].

Open-Ended Questions In addition to the standard questionnaires on user experience, participants filled in a final survey with open-ended questions (details in Appendix A.1).

3.2.4 Procedure.

The study took 22–30 minutes per participant with the following steps.

(1) Participants read and sign the consent form (2 minutes).
(2) A survey on participants' age, language efficiency, etc. (2 minutes).
(3) Participants interact with either Version A or B for 2-3 sessions (6-10 minutes).
(4) Participants complete the AttrakDiff questionnaire (2 minutes).
(5) Participants interact with either Version B or A (switched) for 2-3 sessions (6-10 minutes).
(6) Participants complete the AttrakDiff questionnaire again (2 minutes).
(7) A final survey to gather overall feedback (2 minutes).

3.2.5 Analysis.

Paired Statistical Tests We first performed statistical tests on the four dimensions of AttrakDiff. The following hypotheses were formulated based on the assumption that the memory-augmented agent offers a better user experience:

- H_0 (Null Hypotheses): There is no significant difference in perceived user experience between the non-memory agent (Version A) and the memory-augmented agent (Version B) for each AttrakDiff dimension.
  - $H_{01} : PQ_A = PQ_B$
  - $H_{02} : HQS_A = HQS_B$
  - $H_{03} : HQI_A = HQI_B$
  - $H_{04} : ATT_A = ATT_B$
- H_1 (Alternative Hypotheses): Participants will rate Version B significantly higher than Version A on the AttrakDiff dimensions.
  - $H_{11} : PQ_B > PQ_A$
  - $H_{12} : HQS_B > HQS_A$
  - $H_{13} : HQI_B > HQI_A$
  - $H_{14} : ATT_B > ATT_A$

For each dimension, we used the Shapiro-Wilk test [14] to assess whether the difference scores were approximately normally distributed, which guided the choice of statistical test. If the normality assumption was met ($p > 0.05$), we used paired t-tests. Otherwise, we used the Wilcoxon signed-rank test [15], a non-parametric alternative suitable for within-subject comparisons.

For all comparisons, we calculated Cohen's d [16] as a measure of effect size, defined as the mean difference divided by the standard deviation of the difference scores. Since four comparisons were performed, we applied the Holm correction [17] to adjust p-values for multiple testing and reduce the risk of Type I errors. Each test result was reported with the mean, standard deviations, test statistic (t or W), raw and corrected p-values, Cohen's d, and the result of the normality test.

Visualizations To complement the statistical analysis and facilitate interpretation, we generated a bar chart showing the mean score and standard error of the mean for each dimension. The bar chart made it easy to compare overall differences between conditions. We also created boxplots to show the full distribution of participant scores, including medians, interquartile ranges, and potential outliers. With both conditions plotted side-by-side, we could visually assess the extent of overlap and distributional shifts across conditions. Moreover, we included the following question in the final survey: "Did you notice if one of the agents seemed to remember more from your previous conversations?" The distribution of answers was visualized using a pie chart to quantify the perceived effectiveness of the memory-augmented agent.

Qualitative Analysis of Open-Ended Responses Some open-ended questions were included in the final survey to invite participants to reflect freely on their experience with both versions of the agents. We manually reviewed the responses to cluster thematically similar feedback and draw out recurring insights related to agent memory behavior, intelligence, and consistency.

4 RESULTS

In this section, we present and analyze the results from the AttrakDiff measures and the final survey.

4.1 Quantitative Comparison: Paired Tests

The results of paired statistical tests are summarized in Table 1.

We found statistically significant improvements for the memory-augmented agent on both Hedonic Quality - Identity (HQI) ($p = 0.0266$) and Attractiveness (ATT) ($p = 0.0265$), even after correction.

[13] https://teaching-support.tudelft.nl/typo3-educational-tools-overview-of-tools-used-in-education/

[14] https://builtin.com/data-science/shapiro-wilk-test

[15] https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/how-to-conduct-the-wilcox-sign-test/

[16] https://statisticsbyjim.com/basics/cohens-d/

[17] https://www.statisticshowto.com/holm-bonferroni-method/

Table 1: Paired test results across AttrakDiff dimensions. Significant corrected p-values are marked with *.

	PQ	HQS	HQI	ATT
Test Type	t-test	t-test	Wilcoxon	Wilcoxon
Mean (Version A)	4.20	4.02	3.99	4.21
Mean (Version B)	4.45	4.16	4.62	4.80
SD (Version A)	0.71	0.67	0.97	0.88
SD (Version B)	0.77	0.60	1.01	0.74
Cohen's d	0.32	0.18	0.50	0.54
Corrected p-value	0.1815	0.3409	0.0266	0.0265
Significant?	No	No	**Yes***	**Yes***

However, no significant differences were observed for Pragmatic Quality (PQ) or Hedonic Quality - Stimulation (HQS). These findings partially support our hypotheses in section 3.2.5. We cannot reject H_{01} and H_{02}, but we can reject H_{03} and H_{04}. While we did not observe significant gains in perceived functionality or stimulation, the memory-augmented agent did enhance participants' perception of relational quality (HQI) and overall attractiveness (ATT).

4.2 Visual Exploration of Objective Measures

Figure 4 compares mean scores for the four subscales of AttrakDiff. We can see that Version B scored higher than Version A across all four dimensions. The largest improvements were observed in HQI and ATT, where the differences were also found to be statistically significant in our hypothesis tests. This supports the quantitative findings and reinforces the conclusion that memory augmentation contributes positively to the relational and affective qualities of the conversational agent.

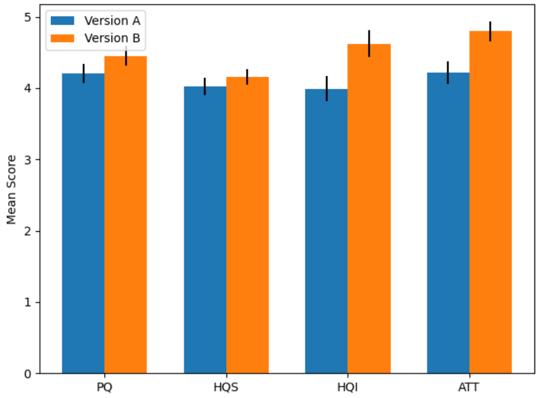

Figure 4: Comparison of mean scores for AttrakDiff dimensions between Version A and Version B.

Figure 5 shows the boxplots for each of the four dimensions, highlighting the distribution, central tendency, and spread of participant scores across both versions. These boxplots reinforce the pattern of improvements observed in the statistical analysis: while functional and stimulating aspects of the agents (PQ and HQS) remained similar, affective and relational qualities (HQI and ATT) showed meaningful enhancement with memory augmentation (Version B).

Figure 6 illustrates participants' responses regarding the perceived memory capabilities of the agents. The majority of participants (76.7%) reported that they noticed a difference in memory behavior between the agents, suggesting that the memory augmentation was both functionally and perceptually salient. A smaller group (16.7%) was uncertain, while only a minimal fraction (6.7%) explicitly stated they did not notice any difference. This implies that the memory-enhanced agent exhibited behaviors that were distinguishable and meaningful to most participants.

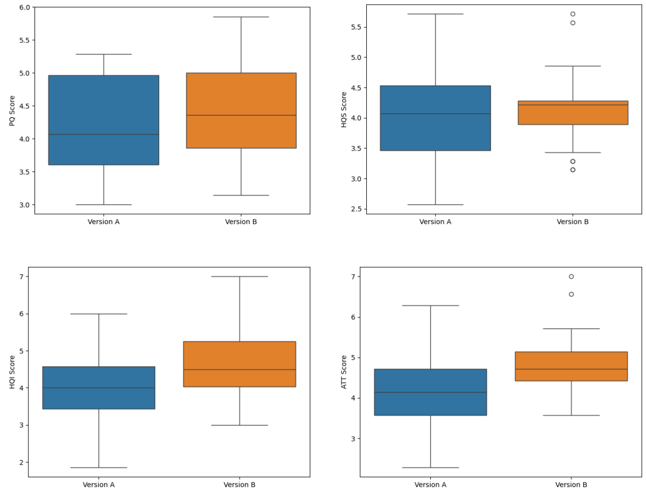

Figure 5: Boxplots of AttrakDiff dimensions comparing Version A and Version B.

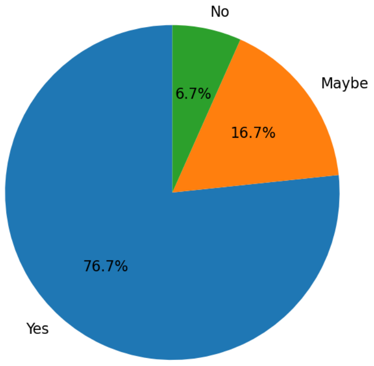

Figure 6: Perception of Agent Memory Recall.

4.3 Qualitative Insights from Open-Ended Feedback

Participants' responses to the open-ended question "What stood out to you the most when comparing the two agents?" revealed several recurring themes that help understand how the memory-augmented agent was perceived against the non-memory version.

Some participants highlighted that one agent (often the memory-augmented one) provided more informative or culturally rich responses. Terms like "actually helpful" and "different" were associated with Version B. This suggests that memory may have contributed to a more knowledgeable and varied experience. Comments referencing the agent “remembering” prior topics or “going off the rails” highlight how consistency (or its absence) shaped user impressions.

Our results show that memory augmentation improves user experience, especially in relational and affective dimensions. Improvements in HQI and ATT suggest users noticed and valued memory behaviors. While functional aspects remained largely unchanged, the results highlight memory's role in building coherent and engaging interactions.

REFERENCES

[1] Eman Alhusaiyan. 2025. A Systematic Review of Current Trends in Artificial Intelligence in Foreign Language Learning. *SJLS* 5, 1 (Jan. 2025), 1–16. https://doi.org/10.1108/SJLS-07-2024-0039

[2] Jeffrey R. Binder and Rutvik H. Desai. 2011. The Neurobiology of Semantic Memory. *Trends in Cognitive Sciences* 15, 11 (Nov. 2011), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

[3] Joana Campos, James Kennedy, and Jill F Lehman. 2018. Challenges in Exploiting Conversational Memory in Human-Agent Interaction. (2018).

[4] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards Human-centered Proactive Conversational Agents. (2024). arXiv:cs.IR/2404.12670 https://arxiv.org/abs/2404.12670

[5] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss Library. (Feb. 2025). https://doi.org/10.48550/arXiv.2401.08281 arXiv:cs/2401.08281

[6] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003*. Vieweg+Teubner Verlag, 187–196.

[7] Zerrin Kasap and Nadia Magnenat-Thalmann. 2012. Building Long-Term Relationships with Virtual and Robotic Characters: The Role of Remembering. *Vis Comput* 28, 1 (Jan. 2012), 87–97. https://doi.org/10.1007/s00371-011-0630-7

[8] Yonghee Kim, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2015. Acquisition and Use of Long-Term Memory for Personalized Dialog Systems. In *Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, Ronald Böck, Francesca Bonin, Nick Campbell, and Ronald Poppe (Eds.). Vol. 8757. Springer International Publishing, Cham, 78–87. https://doi.org/10.1007/978-3-319-15557-9_8

[9] Ludovic Le Bigot, Cléo Bangoura, Dominique Knutsen, and Sandrine Gil. 2022. When Non-Salient Information Becomes Salient in Conversational Memory: Collaboration Shapes the Effects of Emotion and Self-Production. *Quarterly Journal of Experimental Psychology* 75, 7 (July 2022), 1330–1342. https://doi.org/10.1177/17470218211055005

[10] Iolanda Leite, André Pereira, and Jill Fain Lehman. 2017. Persistent Memory in Repeated Child-Robot Conversations. In *Proceedings of the 2017 Conference on Interaction Design and Children*. ACM, Stanford California USA, 238–247. https://doi.org/10.1145/3078072.3079728

[11] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang Anthony Chen. 2025. Proactive Conversational Agents with Inner Thoughts. (2025). arXiv:cs.HC/2501.00383 https://arxiv.org/abs/2501.00383

[12] Geoffrey L. McKinley, Sarah Brown-Schmidt, and Aaron S. Benjamin. 2017. Memory for Conversation and the Development of Common Ground. *Mem Cogn* 45, 8 (Nov. 2017), 1281–1294. https://doi.org/10.3758/s13421-017-0730-3

[13] Seungwhan Moon, Pararth Shah, Rajen Subba, and Anuj Kumar. 2019. Memory Grounded Conversational Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, 145–150. https://doi.org/10.18653/v1/D19-3025

[14] Andrew M. Nuxoll and John E. Laird. 2012. Enhancing Intelligent Agents with Episodic Memory. *Cognitive Systems Research* 17–18 (July 2012), 34–48. https://doi.org/10.1016/j.cogsys.2011.10.002

[15] Hongliang Qiao and Aruna Zhao. 2023. Artificial Intelligence-Based Language Learning: Illuminating the Impact on Speaking Skills and Self-Regulation in Chinese EFL Context. *Front. Psychol.* 14 (Nov. 2023), 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

[16] Deborah Richards and Karla Bransky. 2014. ForgetMeNot: What and How Users Expect Intelligent Virtual Agents to Recall and Forget Personal Conversational Content. *International Journal of Human-Computer Studies* 72, 5 (May 2014), 460–476. https://doi.org/10.1016/j.ijhcs.2014.01.005

[17] María-Loreto Sánchez, Mauricio Correa, Luz Martínez, and Javier Ruiz-del-Solar. 2015. An Episodic Long-Term Memory for Robots: The Bender Case. In *RoboCup 2015: Robot World Cup XIX*, Luis Almeida, Jianmin Ji, Gerald Steinbauer, and Sean Luke (Eds.). Vol. 9513. Springer International Publishing, Cham, 264–275. https://doi.org/10.1007/978-3-319-29339-4_22

[18] Antonia Tolzin and Andreas Janson. 2023. Mechanisms of Common Ground in Human-Agent Interaction: A Systematic Review of Conversational Agent Research.. In *HICSS*. 342–351.

[19] Piek Vossen, Selene Baez, Lenka Bajčetić, and Bram Kraaijeveld. 2018. Leolani: A Reference Machine with a Theory of Mind for Social Communication. (June 2018). https://doi.org/10.48550/arXiv.1806.01526 arXiv:cs/1806.01526

[20] Ling Wei. 2023. Artificial Intelligence in Language Instruction: Impact on English Learning Achievement, L2 Motivation, and Self-Regulated Learning. *Front. Psychol.* 14 (Nov. 2023), 1261955. https://doi.org/10.3389/fpsyg.2023.1261955

A APPENDIX

A.1 User Study: Final Survey

Thank you for completing both sessions with the agent! In this final survey, we would like to gather your overall impressions of the two versions of the agent you interacted with.

If you are unsure about any question, please select the option that best represents your experience. For open-ended questions, if you would not like to answer, you can leave them blank.

If you have any questions or concerns regarding this study, please feel free to contact:

- Jordy del Castilho - jordydelcastilho@gmail.com
- Liwia Padowska - liwia.padowska@gmail.com
- Yizhen Zang - yizhenzang@tudelft.nl
- Zeryab Alam - zeryabalam272@icloud.com

To help us analyze your responses, please indicate which version of the agent you interacted with first:

- Version A
- Version B

Did you notice if one of the agents seemed to remember more from your previous conversations?

- Yes
- No
- Maybe

If yes, which one?

- Version A
- Version B

What stood out to you the most when comparing the two agents?

You can mention anything that surprised you, felt different, or was memorable.

Do you have any suggestions for improving either version of the agent?

Feel free to focus on language learning, interaction style, or anything else.