DELFT UNIVERSITY OF TECHNOLOGY

MASTERS THESIS

ELLA: An Episodic Language Learning Assistant for Language-Anxious Learners

---

Author:
Karthik Prakash
Student number:
5746388

Thesis Supervisor:
Prof. Dr. M.M. SPECHT
Second Supervisor:
Prof. C.R.M.M. OERTEL

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Web Information Systems Group
Intelligent Systems

July 10, 2024

TU Delft

Declaration of Authorship

I, Karthik Prakash, declare that this thesis titled, “ELLA: An Episodic Language Learning Assistant for Language-Anxious Learners” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Karthik Prakash

Date: 2024-07-10

DELFT UNIVERSITY OF TECHNOLOGY

Abstract

Electrical Engineering, Mathematics and Computer Science
Intelligent Systems

Master of Science

ELLA: An Episodic Language Learning Assistant for Language-Anxious Learners

by Karthik Prakash

Learning a Foreign Language is crucial in today's interconnected world, as it enhances cross-cultural communication and fosters personal growth. Despite its many benefits, learners often face significant challenges, primarily due to Anxiety. To address these challenges, we introduce ELLA, a Dutch tutoring system powered by GPT-4. ELLA features an innovative Episodic Memory architecture, enabling learners to review and delete system-generated episodes from past interactions, thereby giving them a sense of control. Through this thesis, we explore the impact of Episodic Memory on Foreign Language Learners' Speaking Anxiety and the associated learning benefits. The investigation is divided into four sub-questions: the relationship between Foreign Language Anxiety and Language Learning, the optimization of the Episodic Memory framework for Learning, the influence of Episodic Memory on speaking-related State Anxiety, and the perceived benefits for learners with Foreign Language Anxiety. Our experimental study demonstrated that ELLA has the potential to reduce speaking-related State Anxiety in Dutch learners and offers significant learning benefits as perceived by users. However, areas for improvement are identified, such as the need for an Automatic Speech Recognition system that accurately recognizes foreign accents and a Large Language Model fine-tuned for the Language Learning task. Future research should focus on these enhancements, evaluate the system in different languages using standardized pre-and post-tests, and involve longer interactions with a larger and more diverse user base over time.

Acknowledgements

I would like to express my deepest appreciation to everyone who has supported me throughout the journey of my master's thesis.

I extend my heartfelt gratitude to Professors Dr.Marcus Specht and Catharine Oertel for their supervision and guidance throughout my research. Your expertise and unwavering support have been invaluable. I am also thankful to Dr.Yoon Lee, MSc.Xiaoling Zhang, MSc.Gabrielle Martins van Jaarsveld, and other members of the LDE-CEL group. Your patient explanations and shared experiences helped me immensely towards completing this thesis. I extend my gratitude to the participants of my experiment for their time and the valuable knowledge they contributed to this project.

I am eternally grateful to my family for their unwavering support and faith in me. Your boundless care and encouragement have always inspired me to push beyond my limits. I am also thankful to my peers Colin, Safouane, Rebecca, Casper, and Ilinca for exchanging ideas and making the process of working on my master's thesis enjoyable. Finally, I would like to thank my friends Akash, Lukas, Xavia and Baitian. Thank you for picking me up when I struggled, for listening to my incessant rants about how difficult my thesis was, and for reassuring me that everything would be okay in the end.

Contents

List of Figures

List of Tables

List of Abbreviations

FL	**F**oreign **L**anguage
CA	**C**onversational **A**gent
EM	**E**pisodic **M**emory
LLM	**L**arge **L**anguage **M**odel

Chapter 1

Introduction

Foreign language (FL) learning is a fundamental aspect of global education, fostering cross-cultural communication and personal growth. It equips individuals with the skills required to effectively engage in an increasingly interconnected world, thanks to the advent of social networks. According to a report by the European Commission, in 2021 approximately 61% of upper secondary and vocational education students learnt at least one FL during their compulsory education [1]. This widespread linguistic education reflects a global trend towards valuing multilingualism, which is seen as crucial in enhancing employment prospects and cultural empathy, especially among expatriates since their language learning is primarily motivated by survival, rather than curiosity. Various studies have linked FL learning to cognitive benefits, such as better memory and better problem-solving skills. [2] [3].

The benefits offered by FL learning are seen not only in personal contexts but also in professional settings. Employers and Academics alike are increasingly seeking candidates who can navigate multiple cultural nuances and communicate with diverse populations. As international collaborations expand, the demand for multilingual professionals increases, necessitating FL learning. Learning a FL can bring numerous benefits beyond the workplace. When one learns a new language, it opens up new doors to diverse cultures, literature, and history. It allows the learner to appreciate and understand different perspectives, ways of life, and traditions, which can ultimately enrich their view of the world [4]. Additionally, learning a FL can enhance cognitive abilities such as memory, problem-solving, and decision-making skills. It can also boost confidence and self-esteem, as it requires dedication, patience, and perseverance to become proficient in a new language [5]. However, despite these numerous benefits, FL learning comes with its own set of challenges, and one significant obstacle that many learners face is FL anxiety.

According to Horwitz et al., FL anxiety or FLA is defined as "a distinct complex of self-perceptions, beliefs, feedings, and behaviours related to classroom language learning arising from the uniqueness of the language learning process" [6]. They believe that FLA is caused by three main factors: Communication Apprehension, Fear of Negative Evaluation and Test Anxiety.

Communication Apprehension refers to the anxiety associated with the idea of communicating with others. In language learning, this often manifests as a fear of speaking or listening in a FL, particularly in front of native speakers or evaluative situations. Learners start to doubt their ability to express their thoughts accurately and fluently in FL, leading to significant stress and reluctance to participate in language activities.

Fear of Negative Evaluation refers to the anxiety and unease that arises from the possibility of being judged negatively by others. This fear is particularly intense in language learning and speaking, where mistakes are visible and can be subject to correction or criticism. Learners who experience high levels of FLA tend to be overly concerned about the opinions of those they are communicating with (such as teachers, peers, natives, etc.), which can hinder their willingness to engage in communication activities that are essential for language acquisition and understanding.

Test Anxiety in FL learning is a situation where learners become excessively anxious and worried about performing poorly on language learning tasks such as tests and exams. This anxiety can affect the learners' ability to demonstrate their true abilities in language assessments and harm their motivation and confidence in learning the language.

Given these significant challenges posed by FLA, as outlined by Horwitz et al., innovative methods have been identified to alleviate these anxieties associated with FL learning [6]. To reduce the effects of Communication Apprehension and Fear of Negative Evaluation, Conversational Agents (CAs) can be used. The creation of these CAs allows learners to practice speaking in a controlled environment where there is no risk of judgement from human listeners, which directly addresses the Fear of Negative Evaluation. Such technology fosters an environment where learners can engage in dialogue without the fear of making mistakes. This can significantly reduce the stress associated with speaking a new language and encourage more spontaneous language use.

Moreover, these CAs can be powered by Large Language Models (LLMs), which can simulate natural and diverse conversational interactions. These models are trained on vast datasets and can generate responses that mimic human speech patterns, making the practice sessions versatile, realistic and beneficial. The use of LLMs enables these CAs to provide feedback that is not only immediate but also highly personalized, based on the user's proficiency level and specific needs. This personalized feedback is crucial for effective language learning as it helps learners identify and focus on their weaknesses.

In this paper, we explore the combination of CAs and FL learning by creating ELLA, a Dutch tutoring system powered by GPT 4.0. For ELLA, we created an Episodic Memory (EM) architecture with a novel feature that allows learners to review and delete episodes before saving them in system memory. ELLA has 4 main goals:

1. To create a safe space where anxious learners can practice Dutch without being afraid to make mistakes.

2. To give feedback on mistakes and suggestions on how learners can improve.

3. To encourage learners to think back and reminisce by storing system-generated episodes from past conversations.

4. To give learners complete control over their learning by providing a user-friendly interface that allows them to explicitly rate and delete episodes before saving them in memory.

1.1 Research Questions

To achieve our goals through this thesis, we aim to answer the research question: **What is the impact of Episodic Memory of a Foreign Language teaching conversational agent on the Speaking Anxiety of Foreign Language learners, and what Learning Benefits does it offer?** This research question is further broken down into 4 sub-research questions:

1. RQ1: What are the current findings regarding Foreign Language Anxiety and its impact on learners' learning and speaking ability, as well as the role of Episodic Memory in Conversational Agents?

   This question aims to analyze existing research in two main areas: the relationship between anxiety and language learning, and the function and importance of Episodic Memory in Conversational Agents. By answering this research question, we establish a foundational understanding of how anxiety affects language learning and speaking in a Foreign Language, and how Conversational Agents can utilize Episodic Memories to enhance learning experiences.

2. RQ2: How can the design of an Episodic Memory framework be optimized to support learning for language-anxious learners?

   This question aims to investigate how an Episodic Memory framework can be augmented to enable learners to take control of their learning. We hypothesise that an augmented design that allows learners to review and delete system-generated episodes will significantly improve learners' sense of control over their Foreign Language learning process.

3. RQ3: How does the usage of episodes stored in Episodic Memory, and the option to delete them, influence Foreign Language learners' speaking-related State Anxiety?

   This research question seeks to explore how the usage of episodic memory, and the option to delete system-generated episodes before saving them, impact Foreign Language learners' State Anxiety levels towards speaking Dutch with our system. The objective is to determine whether these features are beneficial or detrimental to the anxiety levels of FL learners.

4. RQ4: What specific learning benefits do language-anxious users perceive from this system?

   This question aims to explore the subjective perception of users regarding the benefits they gain through using a language learning system that remembers episodes from previous interactions that can also be deleted if required. It seeks to identify which aspects of the system are most valued by users and how these perceived benefits encourage them to continue using the system.

1.2 Overview

Firstly, in Chapter 2, we explore the literature on FL learning anxiety, the role of EM in CAs, and the lack of research surrounding EM editing and deletion to address RQ1. Chapter 3 outlines the architecture of ELLA and the rationale behind the design decisions, aiming to answer RQ2. This chapter ensures the reproducibility of the thesis by providing justifications for the design choices. Chapter 4 details the experimental design, the evaluation of the system through various statistical analyses on the obtained state anxiety scores, and the perceived benefits and results of our experiments, addressing RQs 3 and 4. Chapter 5 discusses our findings and answers the RQs, along with the limitations and suggests further research directions related to the solutions presented. Finally, Chapter 6 presents the conclusion of the thesis.

Chapter 2

Related Work and Hypotheses

This chapter reviews relevant literature and research on foreign language (FL) anxiety and memory in conversational agents (CAs). Section 2.1 examines the causes of FL anxiety and explores mechanisms for mitigating it. Section 2.2 dives into computer-assisted language learning and identifies the components necessary for effective language acquisition. Section 2.3 provides an overview of the types of memory used in CAs, describes state-of-the-art memory architectures, and highlights the lack of frameworks aiding data deletion from memory. Finally, in section 2.4, we outline our hypotheses based on the literature research.

2.1 Foreign Language Learning Anxiety

To understand FL learning anxiety, we need to first understand the notion of anxiety clearly. Anxiety is a psychological state characterized by feelings of worry, nervousness, or fear that are strong enough to interfere with one's daily activities. Typically, it is a natural response to stress or perceived threats, serving as a mechanism to alert the body and prepare it for "fight or flight."

FL learning anxiety is a specific type of anxiety that affects an individual's ability to acquire and use a new language and has been a significant research topic in linguistics [7]. Horowitz et al. define FL learning anxiety as "a distinct complex of self-perceptions, beliefs, feelings, and behaviours related to classroom language learning arising from the uniqueness of the language learning process" [6]. In the past, this concept seemed to be a conundrum, according to Scovel et al and other researchers [8] [9]. In his research, he acknowledges the inherent complexity related to anxiety as a psychological construct and highlights that it cannot be simply classified as either wholly positive or negative in its effects on language acquisition. He highlights the differentiation between facilitating and debilitating anxiety, proposing that while the former can motivate and enhance learning by stimulating the learner's engagement and focus, the latter can hinder progress by overwhelming the learner and impeding their ability to absorb and use the new language effectively.

Scovel's insights are grounded in a broader discussion of affective variables in language learning, which he categorizes as intrinsic to the learner [8]. These affective variables, including anxiety, are positioned within a spectrum of factors that influence language learning outcomes. Fortunately, recent studies have extensively explored FL learning anxiety. Learning a FL is a complex and multifaceted process, influenced by affective, cognitive and social factors [10] Cognitive factors refer to the mental processes of perception, memory and problem-solving that are practised in the language acquisition process. It impacts the speed and proficiency of one's second language skills, which are dependent on his/her first language, age of acquisition and cognitive style. Social factors refer to the effects of the social environment surrounding the language learner. It includes classroom dynamics, cultural exposure and interaction with native speakers, wherein opportunities for meaningful communication in the target language facilitate language acquisition. Affective factors refer to the role of emotions, attitudes and motivation in learning a second language. Research in affective factors of language acquisition typically focuses on how positive affective factors can enhance language learning by fostering engagement and increasing motivation, while negative affective factors like anxiety can hinder progress by demotivating the learner. This observation is underscored by recent studies on scales and questionnaires that effectively measure FL anxiety have led to researchers systematically finding significant negative correlations between FL Performance and Anxiety [11].

2.1.1 Attribution Theory

Current research on the multifaceted ways in which anxiety impedes FL acquisition and performance has shed light on several critical outcomes. Krashen et al. highlight that anxiety serves as an affective filter that prevents a high level of proficiency in a FL [12]. Their research suggests that high levels of emotional distress can create a psychological barrier that hampers the acquisition and processing of new language information, thereby hindering learner performance. This affective filter mechanism is caused in part by attribution theory. In the realm of educational psychology, attribution theory examines how individuals' beliefs about the causes of their successes and failures affect their motivation and behaviour in learning environments [13]. For example, learners who attribute their difficulties in language learning to a lack of innate ability may experience a heightened fear of failure. This fear, deeply rooted in their perceptions of competence and control, can exacerbate the negative effects of anxiety. Individuals become overly concerned with avoiding negative emotions and the embarrassment of failure, a response that is particularly detrimental in the context of FL learning where mistakes are an integral part of the learning process.

Recent studies have explored the relationship between attributional styles, fear of failure, and perfectionism in FL learning. For instance, Hsieh et al. shed light on how fear of failure can hurt FL learning [14]. Their study highlights the importance of self-efficacy and attributions. They found that students who see their failures as controllable factors, such as a lack of effort, are more likely to maintain higher self-efficacy. Similarly, Peacock's research presented many statistically significant differences between attribution and English FL proficiency among students and teachers [15]. These findings are consistent with Gobel and Mori's claim that unsuccessful learners often attribute their failure to a lack of talent and effort, whereas successful students credit their success to professors and the learning environment [16].

2.1.2 Cognitive Suppression

Cognitive suppression refers to the deliberate effort to avoid certain thoughts or memories and is often employed as a defence mechanism towards anxiety-provoking or distressing thoughts. While it can be an effective short-term coping mechanism, according to Wegner and Wenzlaff, cognitive suppression in the long run can lead to various psychological issues including increased anxiety, intrusive thoughts and cognitive load [17]. When individuals continuously attempt to suppress certain thoughts, it can create a negative effect and lead to an increased preoccupation with those thoughts, aggravating anxiety and stress [18].

In the context of FL learning, cognitive suppression can negatively impact a learner's ability to acquire and use a new language effectively. According to MacIntyre and Gardner, learners who experience anxiety related to language learning experience increased cognitive burden due to their suppressive thoughts about their fears or perceived inadequacies [19]. This increased cognitive load can impair concentration, hinder the retention of information and negatively affect performance in language tasks since they get distracted during language learning activities, reducing their overall efficiency and effectiveness [20]. Cognitive suppression consumes a lot of memory resources in learners, leaving fewer available for language learning tasks which can result in difficulties with complex language processing and problem-solving in the new language [21].

Many language learners fear making mistakes or being negatively judged by others. This suppression of anxious thoughts, particularly related to speaking, can increase overall anxiety levels making it more difficult for learners to speak fluently and coherently [19]. This creates a vicious cycle where anxiety leads to poor performance, which in turn increases anxiety and can foster avoidant behaviours where learners try to avoid speaking situations altogether. According to Gregersen and Horowitz, this avoidance can hinder language development and reduce opportunities for improvement [22].

2.1.3 Cognitive Reappraisal

Looking at potential solutions to cognitive suppression in FL learning, the concept of cognitive reappraisal (CR) emerges as an important strategy for emotional regulation that can offer substantial benefits in this context. In essence, cognitive reappraisal refers to individuals changing their emotional response to a situation by reinterpreting the meaning of the emotional stimuli. Grounded in the broader framework of cognitive-behavioural theories, cognitive reappraisal posits that altering one's emotional reaction to stimuli through reassessment of situations can significantly impact learning outcomes [23].

Fallah et al. utilized CR within the framework of their study on mindfulness and anxiety among FL teachers to understand how this regulation strategy could affect the relationship between mindfulness, self-efficacy and FL teaching anxiety [24]. Their experimental results suggest that the ability to reframe one's perception of potentially anxiety-inducing situations in teaching could effectively reduce anxiety levels. Similarly, Catherine. A and her colleagues investigated how CR impacts learning outcomes in game-based environments. Their findings indicate that the use of CR has a positive effect on solving problems that cause frustration or confusion during gameplay. The study further suggests that the benefits of CR are more evident in participants who put in a high level of effort in the game [25].

2.2 Computer-assisted Language Learning (CALL)

Wei et al. stated that almost all language learners use tools to enhance their language learning [26]. In the rapidly changing technological landscape of the 21st century, advanced technologies such as online games, virtual and augmented reality, immersive classrooms, telepresence, and conversational AI have revolutionized the FL (L2) learning environment. Researchers in the linguistics domain have expressed excitement about the potential of these advanced technologies to improve FL acquisition [27] [28].

CALL refers to the use of technologies, specifically computers to aid in the teaching and learning of languages. Beatty et al. perfectly describe CALL as "any process in which a learner uses a computer and, as a result, improves his or her language" [29]. CALL covers a wide range of applications, from traditional rote learning programs to more interactive approaches that leverage multimedia, games and even AI. According to Levy and Hubbard, the primary goal of CALL is to use these interactive technologies to enhance the language learning process by providing learners with access to diverse resources and personalized feedback, leading to a more effective and engaging experience [30][31].

The evolution of CALL can be traced back to the 1960s, with its development being driven by advances in computing technologies. Early CALL applications, such as PLATO (Programmed Logic for Automatic Teaching Operations), focused on repetitive grammar and vocabulary drills [32]. Early systems like PLATO can be categorised as behaviourist CALL since they were grounded by principles of behaviourist psychology, emphasising rote learning and repetitive practice [33]. At that time, technologies similar to PLATO were groundbreaking in the field of language learning because students were allowed to directly engage with the material through exercises and received immediate feedback [34].

However, in this day and age, PLATO and its peers with their rudimentary programming and lack of advanced technologies, fall short when compared to modern CALL tools that leverage technologies such as AI, Machine Learning (ML) and Natural Language Processing (NLP) to provide more personalized and adaptive learning experiences. A hallmark of these modern CALL applications is their interactivity, adaptability and the ability to provide immediate feedback not only in terms of the content but also in terms of the delivery, facial expressions, spoken intonations and much more. These advanced systems are broadly termed Intelligent Tutoring Systems (ITS) [35]. They use AI and ML to personalize lessons by analyzing learners' responses, identifying areas of difficulty and accordingly adapting instructional content [36].

Examples of ITS applications like Babbel, Duolingo and AutoTutor employ sophisticated algorithms to cater language lessons according to the learner's proficiency level and learning pace. This is achieved by employing NLP techniques to validate lesson content, understand user input and provide accurate translations [37] [38]. In parallel, ML techniques, combined with gamification are used to keep learners engaged. A unique facet of Duolingo's AI is its ability to keep the relationship between learner engagement and learner difficulty stable. This is important since in a lot of CALL applications, students get frustrated and quit since the lesson material becomes too difficult to understand [39]. If executed properly, it provides an effective and immersive learning environment. Along with various multimedia elements and interactive exercises, studies have shown that such applications can greatly enhance vocabulary acquisition and improve language skills when used consistently [40].

2.2.1 Episodic Learning

A fundamental aspect of cognitive development, episodic learning emphasizes the significance of individual experiences or episodes in the acquisition and retention of new knowledge. It involves learners encoding, storing, and retrieving information about specific events or episodes, making it particularly effective for learning that requires contextual or situational understanding.

The influence of memory on lexical development has been motivated by a lot of research highlighting the integral role memory systems play in the way learners acquire, retain, and utilize language [41] [42]. Within this broad spectrum of memory impact, Episodic Memories (EMs) emerge as particularly significant due to their unique capacity to capture and store information about events and experiences. This specificity of EM paves the way for episodic learning, which occurs when students engage in learning activities that are memorable and distinct enough to be encoded as individual episodes. These could be through hands-on projects, immersive simulations, storytelling, or any engaging activity that creates a vivid memory of the learning experience.

Studies conducted by Hamrick et al. and Zhang et al. have established the effectiveness of EM in language learning [43] [44]. Hamrick and his team investigated the impact of differences in EM abilities on an individual's capacity to learn new words. They hypothesized that a person who performs well in a non-linguistic EM task is likely to perform well in word learning as well. This is because both tasks might share a common underlying non-linguistic, memory-based mechanism. The results of the experiment showed that individual differences in EM abilities could predict word learning abilities immediately following the learning process.

Zhang and her colleagues attempted to verify the Episodic second language (L2) hypothesis, which suggests that L2 lexical representations in more advanced L2 speakers are episodic [44]. Their research aimed to determine if individual differences in episodic memory also contribute to higher-level L2 lexical abilities. Experiments conducted with a group of higher proficiency L2 learners, who had years of exposure to their L2, showed that episodic memory abilities did predict L2 lexical knowledge. Their results were consistent with the Episodic L2 hypothesis.

2.2.2 Instructional Strategies and Feedback Mechanisms

Instructional strategies and feedback mechanisms are essential elements of an ideal language learning environment. In CALL applications, these elements are required to replicate the pedagogical techniques of human tutors and provide immediate, personalised feedback to learners. A popular, time-tested instructional strategy is scaffolding, which involves providing temporary support to learners to help them achieve a task they cannot accomplish independently [45]. Initially, the learner is provided with a lot of assistance and guidance. According to Walqui, this support can be provided by modelling behaviours, providing hints or prompts, breaking tasks into manageable parts, and offering periodic feedback and encouragement [46]. As the task progresses and as the learner learns the ways of the task at hand, this support provided is gradually reduced until the learner achieves mastery [47]. This process should be dynamic and responsive, adapting to the learner's progress and fostering competence in the learning process.

Scaffolding can be done in various ways, which are dependent on the task at hand. Liu et al., provide a framework for scaffolded language learning in virtual environments [48]. They say that the basic links of scaffolding the teaching process in virtual language learning can be divided into three aspects; Entering the situation, where instructors guide learners through problems by providing the necessary tools (videos, questions, hints, etc) required to build the knowledge for comprehension. Setting up support, where teachers guide students through the problem's increasing complexity by keeping in mind the "Zone of Proximal Development". This is achieved by personalising the questions and the hints provided according to the language proficiency of the learners. This step is very important and has been mentioned as such by various authors [49] [50], since achievable learning objectives are always within the zone of proximal development, as advocated by Vygotsky. In the final step, termed Exploring with independence, learners are given the freedom to decide their own problems and solving methods and are also encouraged to explore independently. This provides learners with autonomy over their learning, encouraging positive habits which eventually lead to mastery [51] [52].

Another popular instructional strategy is dialogic interaction, which refers to the use of conversational dialogue between the system and the learner to facilitate learning. According to Alexander, dialogic interactions refer to the exchanges where students ask questions, share their opinions and comment on each other's ideas [53]. A common form of dialogic interaction is the initiation-response-feedback (IRF) pattern, where the teacher initiates the dialogue, the learners respond and the teacher gives feedback [54]. This is a common scaffolding technique seen in many classroom interactions. Teachers can enforce a continuum of thought by asking follow-up questions, elaborations, reformulations and recapsulations. This scaffolds the learner's thought process by helping them connect past information and events to the current topic, which aids understanding behind their reasoning process [55]. Systems like AutoTutor achieve this by engaging learners in natural dialogue, asking questions, providing explanations and encouraging them to articulate their responses based on provided feedback [38].

The choice of feedback mechanisms depends on the type of learning to be encouraged and hence becomes an important design choice [56]. Immediate feedback, which is provided as soon as a learner responds to a task, helps learners correct their mistakes in real time and prevents the reinforcement of errors [57]. Elaborative feedback goes beyond indicating whether an answer is correct or incorrect; it provides detailed explanations as to why a response was incorrect and what the correct response should be for a specific context, helping to deepen learners' understanding. In language learning, this can be achieved by immediately correcting the learner's pronunciation mistake and providing the correct phrase, intonation, and the motivation behind the correction [58].

Recasts involve rephrasing a learner's incorrect response into the correct form without directly indicating the presence of an error, promoting implicit learning of correct language structures [59]. However, a significant drawback of recasting is that beginner language learners tend to forget the mistakes they made earlier, which prevents them from mindfully re-structuring their response to the correct form [60][61]. In contrast, corrective feedback explicitly indicates that a response is incorrect and provides the learner with the correct form. This allows learners to recognize and understand their mistakes and learn the corresponding correct language usage [62] [63]. Finally, positive reinforcement, through ecstatic responses such as "Well done!", increases motivation and encourages good learning habits. This is crucial in language acquisition, as lack of motivation is a key reason cited by language learners for abandoning their studies [64][65].

2.2.3 Aiding Reminiscence

In language learning, reminiscence involves recalling previously learned grammar, vocabulary and contexts to reinforce positive memory and facilitate language acquisition. When learners reminisce, they transfer knowledge from short-term memory to long-term memory [66]. This can improve retention and make it easier to recall concepts when needed. [67]. Moreover, according to Buxton et al., regularly recalling information helps learners use them effectively and eventually achieve mastery [68]. This is due to the fact that reminiscing allows learners to recognize their progress, which boosts their motivation and confidence, encouraging persistent learning [69].

Several techniques can be used to aid reminiscence in language learning, leveraging both traditional and modern technologies. Traditionally, reminiscence is achieved by engaging learners in story-telling and role-playing activities, which allows them to recall and use language in meaningful contexts [70]. Multimedia cues of lesson material such as videos, audios and interactive simulations can aid this recalling process. Another technique that can be used to aid reminiscence is spaced repetition, which involves revisiting learned material at increasing intervals to strengthen memory retention [71]. Popular language learning applications like Duolingo and Memrise use spaced repetition algorithms to ensure that learners regularly review their vocabulary and grammar [72]. In traditional classroom settings, spaced repetition can be achieved by encouraging learners to keep language journals in which they record new vocabulary, grammar corrections and reflections on their language use.

Modern technologies offer a more interactive medium for reminiscence in language learning, making the process more efficient and effective. VR environments can immerse learners in realistic language contexts, prompting them to really and use language skills on the spot in lifelike scenarios. According to Baker et al., this immersive experience can greatly enhance contextual understanding and memory retention [73]. Social forums like Reddit and language learning forums like Tandem allow learners to engage in constructive discussions, share their learning experience and recall learned material through interacting with online peers [74]. In addition to this peer-to-peer interaction, mobile language learning applications like Babbel and Teuida offer structured AI -driven review sessions that integrate spaced retention to reinforce areas of weakness. This is done at optimal intervals for effective retention [75].

2.2.4 Social Robots in Language Learning

Social robots are autonomous entities that interact with humans through social behaviours and communication. These robots can offer personalized and interactive learning experiences, mainly due to their ability to have human-like conversations and provide immediate feedback. This makes them a perfect medium to be integrated into educational settings, particularly for language learning.

One of the main advantages of using social robots in language learning is their ability to create a stress-free environment for learners. Unlike human teachers, robots do not judge or exhibit frustration whenever learners make mistakes. This can be especially beneficial for language learners who feel anxious or embarrassed about making mistakes [76]. According to Gregersen et al., language learners who are anxious about making mistakes are particularly harsh on their own performance and therefore spend all their energy avoiding making mistakes, which is not conducive to a positive learning environment [22]. Studies have shown that learners often feel more comfortable practicing a new language with robots rather than humans, leading to increased speaking time and improved confidence [77] [78].

Social robots can also serve as tireless teaching assistants who are available for a lesson at any time. This constant presence is particularly useful for language practice, as regular practice and revision are essential for effective language acquisition [79]. Moreover, social robots can be programmed to adapt to the individual needs of the learner by providing customized lessons that cater to different learning paces and styles. Kennedy et al.'s research has demonstrated that personalized interactions with social robots can significantly improve vocabulary retention and pronunciation skills [80]. This interactive approach encourages active participation, which according to Van Den Berghe et al. is the main reason why learners learning a second language with social robots showed greater engagement and retention compared to traditional learning methods [81].

In today's age of AI, social robots are integrated with advanced NLP and ML, leveraging complex text and visual perception to understand and respond to learners' inputs accurately. This interaction is more natural and contextually relevant, providing significant benefits for language learners, as training in a dynamic, real-time environment that closely resembles real-life scenarios can greatly enhance language retention [82].

2.2.5 Peer-like Interaction and Emotional Support

In the context of language learning, social robots have proven to be ideal peers due to their ability to simulate interactions akin to those with human peers, creating a more inclusive environment for language practice. Since social robots can be programmed to exhibit linguistic patterns similar to human peers learners' communicative competence grows as their interactions and familiarity with the robot progresses. This approach reduces the formality typically associated with teacher-student interactions. Studies have shown that learners are more willing to engage in spontaneous conversation with robots due to their non-judgemental nature [83]. Keeping this in mind, Belpaeme et al., framed the robot they used in their L2TOR framework as a peer but used adult-like teaching strategies to scaffold the lessons, which proved to be beneficial [84].

For anxious language learners, providing emotional support is imperative since anxiety and self-consciousness can severely impede progress. These negative emotions arise when language learners find traditional classroom interaction intimidating. Research conducted by Alemi et al., showed that the motivation of children practising English with a social robot increased as a result of interacting with the robot and their anxiety was reduced due to the non-judgemental friendly learning atmosphere [85]. This is because social robots can detect and respond to specific emotional cues that are considered to be signs of anxiety and nervousness by offering positive reinforcement, empathy and encouragement [78].

2.3 Memory in Conversational Agents

2.3.1 Types of Memory

According to Elvir et al., memory in Conversational Agents (CAs) has two aspects of interest, temporal and functional. They divide memory from the temporal standpoint into long-term and short-term memory [86]. Short-term Memory, also called working or belief memory, is used to temporarily store information/knowledge that is relevant to the current situation/task at hand. Information that is stored in short-term memory typically has a short life (minutes up to a few days at max), while long-term memory or (semi-)permanent memory is used to store information that typically lasts for a lifetime and is permanent.

From a functional standpoint, Elvir and his colleagues divide memory into procedural, declarative/semantic and episodic memory. According to them, procedural memory is used to store information that is necessary to perform a task. Information that is stored in Procedural Memory typically consists of rules and steps that need to be followed to succeed at a specific task at hand. Semantic memory is used to store facts/information about the world around the agent, while episodic or autobiographical Memory refers to a record of the agent's experiences as perceived by that agent [87] [88].

2.3.2 Short-Term Memory

In CAs, short-term memory is essential to handle multi-turn dialogues effectively [89] [90]. It enables the agent to remember the context of the ongoing conversation, ensuring that the responses it gives are context-aware. For instance, if a user asks a follow-up question or refers to something mentioned earlier, the agent can reference this prior context without requiring the user to restate information, creating a more natural conversational experience. This also means that it is easier to track conversation states that change as the conversation progresses, such as user intents and dialogue states. For example, in educational settings a tutoring bot needs to remember the exercises that are already completed by the learner and their corresponding responses to guide them accordingly.

In the context of language learning, an essential benefit that short-term memory provides is ambiguity resolution within conversations. It helps the agent to disambiguate pronouns and other referential expressions by recalling recent dialogue content. For instance, if a user mentions "he" or "she", the agent can use its short-term memory to infer what or whom the user is referring to, based on earlier parts of the conversation [91]. In language learning and teaching, this can be beneficial to verify if the pronouns and other referential expressions used for the current context are correct.

2.3.3 Long-Term Memory

Long-term memory, on the other hand, is used to support interactions that span multiple sessions where personalization of user experiences over time is paramount [92]. This is done by storing information across multiple interactions, enabling the agent to provide more tailored and user-relevant responses. By storing user preferences and habits the agent can tailor its responses to the individual users. This is done by storing tidbits from conversations, which might include remembering a user's favourite topic to talk about, preferred interaction style or specific preferences towards places, music, etc [93]. For example, in a language learning scenario, an agent with long-term memory can recall the user's previous performances including the mistakes they made and how they resolved it, allowing for a more personalized and effective lesson in future interactions.

Furthermore, long-term memory is essential for creating a sense of continuity and relationship with the user. This allows conversational agents to engage in more human-like interactions by building on past conversations, which is essential to develop their relationship with the user positively. For example, a virtual health assistant can track a user's progress over time, remember previous health concerns, and provide consistent advice based on historical data. As the agent builds its responses on historical data, it learns new information and adapts its behaviour accordingly. In educational settings, this can be seen in tutor chatbots which track student progress over time, adapting their teaching strategies based on what the student has mastered or struggled with previously [94][95].

2.3.4 Semantic Memory

In CAs, semantic memory is essential to generate meaningful responses based on stored knowledge. This includes understanding the meaning behind language and relationships between concepts stored in memory [96]. Generally, semantic memory in CAs is typically in the form of knowledge bases and ontologies. Knowledge bases are databases or knowledge graphs that store facts and relationships in a machine-readable format. These knowledge bases allow CAs to perform tasks such as answering factual questions with accurate explanations and engaging in context-aware conversations. Ontologies, on the other hand, are structured frameworks for representing knowledge domains, including entities, their attributes and relationships [97].

One of the main reasons for using semantic memory in CAs is personalization, as demonstrated by Mavropoulos et al. [98]. Certain facts and observations about the user that are temporally constant are usually stored in semantic memory. These can either be long-term or short-term. For example, universal facts like "The earth is the only planet in the solar system that supports life" or user-specific facts like "User X is Indian" are temporally constant concerning the lifespan of the agent, and are stored in long-term semantic memory. While other interaction-specific facts like lesson content or current tasks are stored in short-term semantic memory for a few hours up to a maximum of a couple of weeks.

Since semantic memory enhances an agent's understanding and context retention across multiple sessions, context-aware coherent responses are generated, improving user satisfaction and engagement. By accessing a rich repository of facts and knowledge, CAs with semantic memory can provide accurate and detailed answers to user queries. This can be particularly valuable in educational settings, as described by Garrido et al., who used a CA with semantic memory to teach content from the Great Aragonese Encyclopedia to students [99]. By making use of centrally stored user data in semantic memory, conversations held by CAs are consistent across different interactions and sessions. This is essential to sustain engagement with the user, which is necessary for effective learning and teaching [100].

2.3.5 Episodic Memory

According to Wood et al., episodic memory (EM) consists of personally experienced event information that is spatially and temporally organized [101]. In CAs, EM is employed to store and remember specific events, experiences, and their contexts as perceived by the agent. This includes user preferences, past queries, and specific events mentioned during conversations. Most events stored in EM are emotionally driven. For example, when a user expresses excitement towards a particular topic or mentions being sad because their dog died. This emotional intelligence is also why CAs with EM can exhibit a level of empathy that makes interactions more human-like and satisfying.

Similar to semantic memory, EM can be divided into long-term and short-term. Short-term EM allows CAs to temporarily retain information from recent interactions, enabling them to maintain context and coherence in the current interaction. In educational settings, this can involve remembering what mistakes the learner made during the previous lesson and referring to them in the current lesson. This capability is crucial for maintaining fluidity in learning. On the other hand, long-term EM involves the storage of significant events and interactions over extended periods. This can include remembering what topics were discussed and what mistakes were addressed in previous lessons, which is essential to facilitate reminiscence in learning (ref 2.2.3).

Events and knowledge stored in short-term EM transition to long-term memory when they are frequently referenced, akin to memory consolidation in humans [102]. This process ensures that CAs can handle immediate conversational needs while also building a rich, contextual understanding of user behaviour and preferences over time, thus enhancing the depth of user interactions.

2.3.6 Memory Architectures

Xu et al. use a long-term memory (LTM) architecture to dynamically and accurately manage user persona information over extended interaction periods [103]. Their PLATO-LTM system architecture is comprised of several key components.

The first main module is the persona extractor, which identifies and extracts user and chatbot persona-related information from the dialogue. It utilizes an ERNIE-CNN network architecture, combining the pre-trained ERNIE model for sentence representation from the conversations along with a CNN for classification. This classification process labels each input to determine if it contains persona information and stores it in the respective memory module.

The second main component of the architecture is the Long-Term Persona memory, which maintains and continuously updates the personas of both the user and the chatbot based on the extracted data. The memory is structured in a way that enables efficient querying operations for reading and writing, allowing the system to adapt to new information across multiple interactions. The memory to be retrieved is determined by using the current dialogue context as a query to search the LTM for relevant persona information. This information, combined with their context persona matching module, which ranks the relevance of persona details to the ongoing conversation, is fed through their state-of-the-art dialogue generation model, PLATO-2, to generate context-aware responses that align with the personas involved.

Campos and Paiva divided the EM architecture of their conversational agent, MAY (My Memories Are Yours), into three hierarchical levels: Memory Line, General Events, and Lifetime Periods [104]. At the Base level, the Memory Line captures event-specific knowledge, encompassing detailed memories of individual events. This includes sensory-perceptual details and contextual information such as dates, personal textual descriptions, images, and sounds. These detailed memories serve as the fundamental components of their EM architecture, allowing for highly specific retrieval.

The Middle level contains General Events, which are abstractions of the specific memories stored at the Base level. The information at this level is categorised into "What," "Where," "When," and "Who." General Events can either relate to overarching Lifetime Periods or specific detailed memories spanning various dates. At the Top level, Lifetime Periods organise memories into broader thematic periods in a person's life, such as "College Years" or "Professional Careers". This thematic organisation helps to manage the overlap of different memories by associating them with larger periods, facilitating easier navigation through the complex network of episodic memories.

Elvir et al.'s EM is structured to store rich, contextual information about each user interaction [86]. Each memory entry includes metadata such as the time and date of interaction, the users involved, the context and the content of the conversation. This provides their Embodied CA with a foundation to generate contextually appropriate and personalized responses in future conversations. For example, if a user previously mentioned a preference towards a certain entity, the ECA can recall this information and incorporate it into their dialogue, creating a more engaging and meaningful interaction. During the conversation, their system constantly parses it to identify significant pieces of information, which are then organized and stored with the appropriate metadata. This real-time encoding ensures that their ECA builds a memory database that reflects the user's history and preferences accurately. They use a database to store the memories of their ECA, and query it to retrieve entries that match the context during conversation, to generate a contextually accurate response.

Their memory architecture includes both consolidation and forgetting mechanisms. Consolidation involves periodically organizing and integrating stored memories to improve retrieval efficiency and coherence. This is done by merging related memory entries or indexing them in a way that makes future retrieval faster and more accurate. Forgetting involves discarding outdated or less relevant memories which is essential to maintain optimal retrieval performance and memory size.

2.3.7 Editing Memory

The topic of editing memory is very rarely explored in the domain of CAs due to a combination of ethical, technical and practical considerations. Liao and Sandberg highlight the potential for abusing memory editing technologies to manipulate individuals or edit their memories without their consent [105]. Implementing memory editing capabilities in these agents raises questions about user consent, the potential for manipulation, and the overall impact on users' mental and emotional well-being. Ensuring ethical compliance in memory editing practices is challenging and requires robust regulatory frameworks, which are currently lacking in the field.

Another important aspect of memory modification is obtaining explicit consent. In practice, users may be uncomfortable with a CA altering or deleting their memories. This discomfort stems from the fear of losing their control and the integrity of personal experiences. Also from a technical perspective, the process of modifying specific memories always warrants unintended side effects [106]. Translating these complexities into CAs involves creating nuanced algorithms that can identify which memories to edit, and what the appropriate modifications are and ensuring that these changes do not disrupt the reliability and coherence of the knowledge base.

In the era of AI and LLMs, many researchers are advocating for the idea of giving users explicit control over the memories retained by an AI agent [107]. Their main idea is that users should have the ability to delete memories for a variety of reasons, including safety-related concerns such as privacy or trade secrets, as well as more mundane issues. However, the addition, editing, and deletion of memories in AI agents can be important when the AI uses and recombines memories to create new plans for the future. For instance, negative memories can be added to the AI as a precaution against carrying out actions that the user is opposed to. Therefore, to facilitate safe editing, memories must be isolated in a format that allows for their addition or complete deletion by users.

In the past, memory editing, such as deletions, was typically done through natural forgetting processes. The primary goal was to eliminate irrelevant or outdated information that had lost its significance over time. This was done to free up memory space for storing new and relevant information and to enhance the retrieval process, as explained in Richard et al.'s paper [108]. However, Richard and his colleagues also advocate for user-initiated forgetting, which enables users to manually specify which information they want the system to forget. This is beneficial for empowering users during their interactions.

With this in mind, Huang et al. developed a "Memory Sandbox," an interactive framework that enables users to manage and manipulate the conversational memory of LLM-powered agents [109]. They consider past interactions as "memory objects," interactive components of conversational history that can be viewed, manipulated, recorded, summarized, deleted, and shared across conversations. This framework provides users with direct control to modify the agent's memory, aligning with our architectural objectives.

Despite the benefits, allowing users to directly edit EMs presents significant drawbacks. As the stored information grows, managing both deletion and storage processes in the same memory space becomes increasingly difficult, leading to potential consistency issues. For example, if only part of the EM is updated without ensuring coherence, the agent might provide irrelevant responses due to these inconsistencies. Additionally, user edits can be computationally costly, as the memory system must refresh and reload with each change.

2.4 Hypotheses

Based on extensive literature research on the causes and solutions of FL learning and speaking anxiety, and the role of EM in CAs, we propose the following hypotheses:

1. We hypothesize that a Dutch language tutor CA equipped with an EM, and a feature to delete system-generated episodes from past conversations, will result in lower levels of speaking-related State Anxiety in FL learners.

2. We hypothesize that a Dutch tutoring system consisting of two conversational lessons, a reminiscing session where learners can ask questions and discuss previously covered topics, and a review session after each of the first two lessons where learners can review system-generated episodes from lesson conversations, rate them based on their helpfulness towards foreign language learning, and delete them from memory if desired, leads to multiple perceived learning benefits.

Chapter 3

System Design

This chapter comprehensively explains the design of our system, ELLA. Section 3.1 outlines the technologies utilized in creating ELLA. Section 3.2 describes the prompts used to instruct the GPT-4 LLM, along with the corresponding prompt engineering techniques employed. Section 3.3 provides an overview of ELLA's components and the rationales behind their design choices. Finally, Section 3.4 offers a comprehensive description of ELLA's system and memory architectures.

3.1 Technologies Used

3.1.1 OpenAI GPT-4

OpenAI's GPT (Generative Pretrained transformer) is a state-of-the-art, closed-source LLM that is highly suitable for interactive learning environments [1]. It possesses advanced Natural Language Understanding (NLU) and Generation capabilities, which have been developed through training on a vast amount of data. GPT-4, the latest iteration, can interpret complex instructions and produce contextually appropriate responses, mimicking human-like conversations. As we discussed in 2, FL learning environments need to satisfy certain requirements such as peer-like language and actionable feedback, most of which are satisfied by GPT. Various authors have conducted extensive research on how GPT meets these requirements, as shown in studies such as [110], [111], [112], and [113].

Fuchs discussed the benefits of GPT in higher education [114]. He explained that GPT can provide tailored learning paths and personalized assessments, which can help students achieve academic success and build self-confidence. He also emphasizes the importance of on-demand support, particularly in language learning environments where learners often require immediate assistance. For instance, a learner might need to inquire how to ask for the price of eggs in Dutch while shopping at a Dutch supermarket. Since most language learners are self-taught, they need immediate assistance when they come across challenging situations. GPT-4's on-demand support can help learners overcome these obstacles and increase their self-efficacy.

In the paper by Markel et al., the benefits of GPT are identified in training novice learners, and an interactive teacher-training app for novice teachers is proposed, which enables them to practice with simulated students without affecting real students [115]. This app also allows for the migration of data across sessions and offers personas and learning goals to tailor GPT's responses to the teachers. Similarly, we aim to provide a personalized language learning environment for Dutch language learners. We achieve this by creating user personas and using data from previous

[1] https://platform.openai.com/docs/api-reference/chat/

lessons to personalize the conversation. This is done without a human in the loop between the learner and the system.

Considering these benefits, we employ the OpenAI GPT-4 LLM through their API for two main purposes: Response Generation and Episode Identification.

Response Generation

We use the GPT-4 LLM to generate responses based on user input and a set of instructions (or **prompts**) that guide the LLM on processing user input and generating responses. Each lesson in our system is governed by unique prompts that are fed into the LLM. An overview of the prompts used is given in 3.2.

When a learner's speech input is converted to text, the GPT-4 LLM processes this text input as described in 3.1. Based on the interpreted meaning, it then generates a contextually accurate response. Moreover, since we also feed certain learner data such as the learner's proficiency level and some of their Dutch knowledge through placeholders in the prompt, the LLM customises its responses by adapting to this fed data along with data from past interactions, increasing or decreasing the level of complexity of the language used. This tailored approach helps in conducting lessons and providing feedback that is appropriate to the learner's current skill level, making the learning experience more personalised and intuitive.

Given the input "*Ik hou van het centrum in Delft*", the LLM dynamically recognizes and responds according to the learner's Dutch language proficiency. For a beginner learner, the LLM might respond with "*Dat is leuk om te horen! 'Ik hou van' betekent 'I love' in het Engels. 'Het centrum' betekent 'the center'. Kun je proberen te zeggen 'I love the park' in het Nederlands?*". This output is deliberately simplistic, utilizing basic vocabulary and straightforward sentence structures to reinforce language fundamentals and enable the learner to replicate the syntactic pattern in other contexts.

For an intermediate learner, the LLM might respond with "*Mooi gezegd! Delft heeft een prachtig centrum. Weet je ook hoe je kunt zeggen waarom je het leuk vindt? Bijvoorbeeld, je kunt zeggen: 'Ik hou van het centrum in Delft omdat het historisch en levendig is.*" This response not only acknowledges the initial input but also prompts the learner to elaborate, thereby fostering the development of conversational skills through the addition of detailed context.

Whereas for a learner with advanced or near native proficiency, the LLM aims to enhance both linguistic precision and analytical thinking by encouraging the use of descriptive adjectives and critical reflection on the specifics of their affection for Delft's center. Therefore in this case the LLM might respond with "*Dat is een prachtige plek! Delft staat bekend om zijn rijke historie en cultuur. Kun je beschrijven welke specifieke aspecten van het centrum van Delft jou aanspreken? Probeer enkele bijvoeglijke naamwoorden te gebruiken die je gevoelens en observaties gedetailleerder weergeven.*" This approach not only aids in the expansion of vocabulary but also promotes a deeper cognitive engagement with the target language, aligning with educational objectives for advanced language acquisition [116].

Episode Identification

We follow a similar episode identification process as Elvir and his colleagues [86]. According to them, episodes are an agent's experiences/events as perceived by the agent. In our context, episodes are snippets from the conversation that our system perceives as either a 'mistake' or a 'success'.

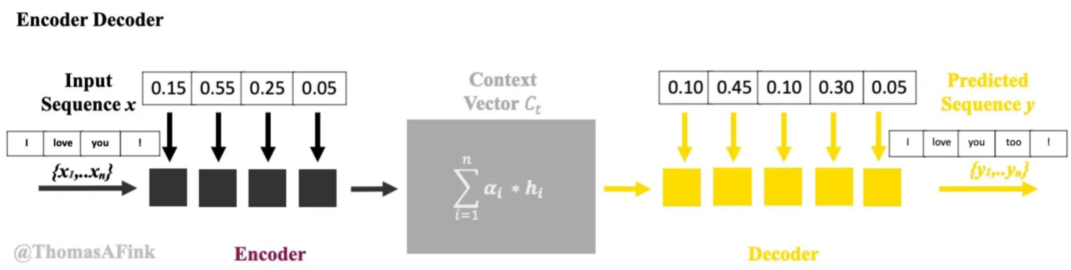

FIGURE 3.1: How GPT processes user input (as explained in [117])

To understand the methodology behind our system labels for identified episodes, we first need to know what is considered a 'mistake' and a 'success' episode. The differentiation is mainly due to specific linguistic criteria that measure the quality and accuracy of spoken language.

An episode is labelled as a "mistake' when it exhibits deficiencies or errors in the following areas:

- **Vocabulary**: Improper use or selection of words that may not fit the context or intended meaning.
- **Grammar**: Errors in verb tenses, noun-adjective agreement, or other grammatical rules that disrupt the clarity or standard structure of the Dutch language.
- **Typographical**: The overall organisation of sentences and paragraphs that may lack logical sequencing, making the text hard to follow.
- **Choice of Words**: Utilizing words that are either too complex, too simplistic, or inappropriate for the target audience or context.
- **Pronunciation**: In spoken language, incorrect pronunciation can lead to misunderstandings or the perception of a lack of fluency.
- **Syntax**: Faulty sentence construction that affects the readability and comprehension of the text.

Conversely, an episode is considered a 'success' when it displays proficiency and appropriateness for the given context. The criteria include, but are not limited to:

- **Structure**: Well-organized and logically constructed sentences that facilitate easy understanding and flow of ideas.
- **Vocabulary**: Appropriate and contextually relevant word choices that convey precise meanings and enhance communication.
- **Grammar**: Correct grammatical usage that adheres to the rules and conventions of the Dutch language.
- **Colloquial Expression**: Effective use of informal or conversational phrases that resonate with the interlocutor and add authenticity to the dialogue.
- **Pronunciation**: Accurate and clear pronunciation that promotes effective communication and demonstrates mastery over the Dutch language.

After a lesson is completed using our system, we send the conversational script along with specific instructions to the LLM through an API call. This helps us to identify the expert episode and align it with our educational objectives. We rely on the LLM for this task because it provides a nuanced evaluation that goes beyond simple error detection. It can evaluate the appropriateness of syntax and structure within the flow of conversation and can recognise the appropriate usage of colloquial expressions for the context. This ability is particularly valuable in educational settings where understanding the subtleties of language use is crucial for effective learning and improvement. Many researchers have explored the use of GPT as an expert annotator for various purposes [118][119][120], to mitigate various issues related to human annotation and evaluation.

For instance, manual assessment and annotation may be affected by biases, leading to what is known as the bias blind spot. This term refers to people being more conscious of biases in others than in themselves [121]. In educational settings, particularly in spoken language learning, this can cause evaluators to overlook their own biases when assessing learners who are not as proficient in the target language as they are. This may result in inaccurate and unfair assessments.

By utilizing LLMs like GPT-4, we mitigate these biases. The model applies a uniform set of criteria to analyze conversational scripts, reducing the subjective influence that individual educators might exert unintentionally.

Furthermore, LLMs offer a scalable and efficient feedback system that provides immediate, objective, and actionable feedback to learners, which is crucial for making rapid adjustments and improving their language skills.

3.1.2 MongoDB

MongoDB is a highly efficient NoSQL database that is document-oriented, which means it can easily manage and store different types of data. We chose MongoDB because of its flexibility and scalability, which are essential for handling the varied data generated by our system. Its robust querying capabilities enable us to quickly retrieve data, which is crucial for providing real-time feedback and personalized experiences in our chatbot. Because MongoDB is schema-less, we can store documents in a JSON-like format with varied structures, making it ideal for storing user profiles, interaction logs, and personalized learning paths without the constraints of a relational database schema.

Our system has five separate MongoDB databases, each with a specific purpose. Two databases store the episodes generated by the LLM for two lessons (which are saved in a JSON format), two databases store the metadata generated for the two lessons (topics and user preferences), and one database stores learner profiles containing language proficiency and knowledge about the Netherlands and the Dutch Language to personalise the conversations, along with a deletion log that contains a history of deleted episodes.

3.1.3 Furhat

Furhat Robotics' Furhat, is a social robot that facilitates advanced human-computer interaction through its sophisticated conversational AI system [2]. Unlike other robots, Furhat has an expressive, animatronic face that can display a wide range of emotions

[2]https://furhatrobotics.com/

and lip-synced speech, making user engagement more natural and intuitive compared to traditional screen-based language learning tools like Duolingo. Such realistic interactions are crucial to facilitate effective language learning, as they mimic real-life conversations and social contexts, which are essential for practising spoken language.

The expressive capabilities of Furhat also help increase learner engagement, as discussed by Du, Jessen and their respective colleagues [122] [123]. Seeing a robot responding with smiles or nods during a conversation makes the learner feel acknowledged, making the interaction far less intimidating. This can have a positive effect on the motivation of the learner, making the learning process more enjoyable and effective [124] [125].

Furhat also helps us make FL learning accessible to a broader range of learners, including those who might feel self-conscious in traditional classroom settings, which motivated our choice of the robot further. Its friendly and non-judgemental style can particularly benefit introverted learners, providing a safe space for conversation and encouraging them to practice more openly and frequently [126] [127].

Our system incorporates Furhat as an interlocutor who converses with the learner in Dutch. Furhat plays the role of a native Dutch teacher who converses with the learners, correcting them whenever they make a mistake and praising them whenever they respond correctly. These role-playing scenarios provide practical language practice and help learners understand cultural nuances, which can be challenging to grasp through traditional textbook-based learning [128]. Furhat's human-like features are central to creating a learning environment that mirrors real-life interactions. This makes the educational experience more intuitive, as learners are given the freedom to engage in a more natural and immersive dialogue, much like they would with a human teacher. This not only enhances the learner's ability to absorb linguistic nuances such as pronunciation, intonation and cultural expressions but also promotes emotional engagement, which is critical for memory retention [129].

We communicate with the learner through Furhat's built-in TTS which delivers the responses generated by the GPT-4 LLM. We use Furhat's built-in ASR system to convert the user's speech input to text and send it to the LLM for further communication.

3.1.4 Streamlit

Streamlit is an open-source framework tailored for app development in Python, facilitating the rapid creation and deployment of data applications with minimal coding. Additionally, some UI elements can be integrated using HTML and CSS scripts. We chose Streamlit for its stateful nature, which simplifies managing session variables that change with each interaction. Our system ELLA utilizes Streamlit to develop an interactive front-end interface. The Streamlit app is organized into several pages, detailed in A.

3.2 Prompt Engineering

3.2.1 Definition and Prompts Used

Prompt Engineering is the driving force of our system both in terms of conversing with the learner and generating episodes from the conversation. In essence, prompt engineering involves creating structured instructions that guide both the LLM and the learner through the learning process. In the context of creating a Dutch language learning assistant, prompt engineering helps to ensure that the lessons are personally scaffolded, engaging and relevant for the student while also being able to generate context-aware responses. Critically, prompt engineering also allows us to place some rules and restrictions that prevent GPT-4 from generating unwarranted or contextually incorrect responses. The prompts we have used for the two lessons are seen in 3.1 and 3.2, while the prompt for the reminiscing session is shown in 3.3.

LISTING 3.1: Prompt for Lesson 1

```
U bent een expert in het onderwijzen van de Nederlandse taal. U spreekt
     met een {lang_prof} student die Nederlands leert. De student heet
    "Learner", zij is bekend met steden {known_cities} en hun favoriete
     zinnen zijn {known_phrases}. Uw taak is om in eenvoudig Nederlands
     met de student te converseren, zonder het gebruik van emoji's of
    emoji-shortcodes.
Zeer belangrijk: Wanneer de student een fout maakt, bied dan een
    vriendelijke correctie met duidelijke markering. Bijvoorbeeld, als
    de student zegt "Ik loop in de park", formuleer het dan als volgt:
    "Om het nog beter te maken, zeggen we eigenlijk in het Nederlands '
    Ik loop in het park'. Zou je het met deze formulering kunnen
    proberen?" en moedig de student aan om de correcte zin te herhalen.
     Bied suggesties voor verbetering en bevestig dat de student de
    correctie heeft begrepen door hen te vragen de juiste zin te
    herhalen. Zorg voor een ondersteunende en educatieve interactie.
Wanneer de leerling de zinnen die zij kent correct gebruikt,
    complimenteer ze dan en bied aanmoediging. Bijvoorbeeld, als "Doei"
     een van de woorden in {known_phrases} is, en de leerling gebruikt
    het correct wanneer zij de les verlaat, complimenteer haar dan door
     te zeggen "Heel goed! Je hebt Doei correct gebruikt in deze
    context."
Begin met: "Goedendag Learner, leuk om met je te praten! Vandaag gaan
    we praten over waarom {city} jouw favoriete stad is. Kun je me
    vertellen wat je leuk vindt aan {city}?" Als er een fout wordt
    gemaakt, bijvoorbeeld "Ik houd van de meetings organized",
    corrigeer dit naar "In het Nederlands zouden we zeggen 'Ik houd van
     de georganiseerde bijeenkomsten'. Kun je deze zin herhalen?" of "
    Probeer het nog eens, maar nu met 'georganiseerde bijeenkomsten' in
     plaats van 'meetings organized'."
Ga pas verder nadat de student de correctie heeft herhaald, om hun
    begrip en gebruik van de Nederlandse taal te versterken. Gebruik
    alsjeblieft geen emoji's of emoji-shortcodes in je antwoorden om de
     focus op tekstuele communicatie te houden.
```

LISTING 3.2: Prompt for Lesson 2

```
U bent een expert in het onderwijzen van de Nederlandse taal. Vandaag
    gaat u verder met uw cursus met "Learner", een {lang_prof}
    Nederlandse student die genteresseerd is in Nederlandse gerechten {
    known_food} en culinaire tradities.
```

```
U dient gegevens die zijn verkregen uit eerdere gesprekken te gebruiken
     in uw antwoorden. Deze details bieden inzicht in zowel de
    struikelblokken als de succesmomenten van de student. {mistakes}
    bevatten de fout die de leerling in de vorige les heeft gemaakt,
    samen met een samenvatting die een idee geeft van de context waarin
     de fout is gemaakt.
{successes} bevatten de correcte zinnen die de leerling in de vorige
    les heeft gebruikt, samen met een samenvatting die een idee geeft
    van de context waarin de leerling deze correcte zin heeft gebruikt.
Gebruik deze samenvattingen actief om het huidige gesprek aan te passen
    . Bijvoorbeeld, als de student een soortgelijke fout maakt, verwijs
     dan naar de eerdere correctie met een opmerking zoals: "We hebben
    dit eerder bekeken, herinner je 'Ik vind het centrum mooi' na onze
    correctie? Laten we dat nog eens proberen."
Evenzo, als een student een zin correct herhaalt, bevestig dit dan met
    positieve feedback: "Uitstekend, net zoals je eerder correct zei '
    Ik vind het centrum mooi'!"
Wanneer de leerling de zinnen die zij kent correct gebruikt,
    complimenteer ze dan en bied aanmoediging.
Bijvoorbeeld, als Doei een van de woorden in {known_phrases} is, en de
    leerling gebruikt het correct wanneer zij de les verlaat,
    complimenteer haar dan door te zeggen "Heel goed! Je hebt Doei
    correct gebruikt in deze context."
Begin de sessie met een open vraag over Nederlandse gerechten om de
    dialoog te stimuleren: "Goedendag Learner, fijn je weer te zien.
    Laten we het vandaag hebben over jouw favoriete Nederlandse
    gerechten, zoals {known_food}. Wat waardeer je het meest aan deze
    gerechten, en kun je daar meer over vertellen?"
Moedig tijdens de sessie actieve betrokkenheid en herhaling aan voor
    zowel correcties als bevestigingen van correct gebruik. Dit helpt
    de student om vertrouwen te ontwikkelen in het gebruik van de
    Nederlandse taal en bevordert de taalretentie. Gebruik geen emoji's
     in je antwoorden om de focus op tekst te houden.
Bij fouten, bijvoorbeeld "Ik houd van de meetings organized", verfijn
    de correctie naar: "In het Nederlands zeggen we 'Ik houd van de
    georganiseerde bijeenkomsten'. Kun je dat eens proberen?" of "Laten
     we 'georganiseerde bijeenkomsten' proberen in plaats van 'meetings
     organized'." Bevestig de correctie voordat u verdergaat, om hun
    Nederlands te verbeteren.
Belangrijk: Pas uw feedback en correcties aan op basis van de voortgang
     van de student, gebruikmakend van aantekeningen uit de vorige les.
     Bijvoorbeeld, herinner de student aan vorige fouten of
    complimenteer correct gebruik van eerder geleerde uitdrukkingen met
     opmerkingen als 'Het lijkt erop dat je dezelfde fout als vorige
    keer hebt gemaakt', of 'Je hebt 'doei' weer correct gebruikt, goed
    gedaan!'.
```

LISTING 3.3: Prompt for Reminiscing Session

```
Je bent een uitstekende Nederlandse taaldocent. Vandaag haal je
    herinneringen op met je student 'Learner' over de onderwerpen die
    jullie eerder hebben besproken.
```

```
{lesson1_topics} bevat de onderwerpen van de eerste les en {
    lesson1_sentiments} de gevoelens van de leerling hierover. Zo ook
    bevatten {lesson2_topics} en {lesson2_sentiments} de onderwerpen en
     gevoelens van de tweede les.
Belangrijk: Je hebt een overzicht van de fouten {total_mistakes} en
    successen {total_successes} van de leerling uit de vorige lessen.
    Refereer aan deze gegevens als de leerling een soortgelijke fout
    maakt of succes herhaalt.
Zeg bijvoorbeeld: 'Je maakt weer dezelfde fout met "doei"' of 'Goed zo!
     Je hebt "lekker" weer correct gebruikt.' Deze feedback bevat
    gedetailleerde correcties en complimenten over uitspraak,
    grammatica, en woordgebruik. Gebruik deze informatie om de
    lesinhoud aan te passen en te personaliseren.
Bied vriendelijke correcties en laat de leerling de juiste formulering
    herhalen voor bevestiging, zoals in: 'In het Nederlands zeggen we "
    Ik loop in het park". Probeer je het zo te zeggen?' Zorg voor een
    ondersteunende en educatieve interactie en bevestig het begrip van
    de leerling voordat je verdergaat. Gebruik geen emoji's in je
    antwoorden om de focus op tekst te houden.
```

3.2.2 Abstract Rule Representation

When designing our prompts for the lessons, we follow a blueprint consisting of abstract rules that help structure the teaching process. This blueprint outlines the key components that guide the interactions between the GPT-4 LLM, which acts as a Dutch language teacher, and the learner, ensuring a consistent FL learning experience.

1. **Role Definition:** To set the stage, we first establish the LLM's identity as a Dutch language teacher. This gives the LLM an idea about how to structure its responses in a clear and informative manner. We also use Dutch for the system prompt to emphasize that the conversations should be in Dutch with the learner.

2. **Learner Customization:** We use data that is stored in user profiles, such as language proficiency, known Dutch cities, Dutch dishes and Dutch phrases, to personalize interactions. This is done by feeding the information into the lesson prompt through placeholders.

3. **Context Information:** We provide contextual information to create a theme for each lesson. In the first lesson, we instruct the LLM to engage the learner in conversation about their favourite cities. In the second lesson, the LLM is instructed to talk with the learner about their favourite Dutch dishes. Finally, in the reminiscing session, the LLM is instructed to help the learner look back on the previous lessons.

4. **Feedback Loop:** We direct the LLM to establish a feedback mechanism with the learner. This system will help in creating an environment that continuously guides the learner towards improvement. When the learner makes a mistake, the LLM should clearly explain the correction and encourage the learner to repeat the corrected form. This will help the learner better understand and

remember the material. Additionally, when the LLM identifies the correct usage of phrases or concepts, it should provide positive feedback to promote successful learning behaviours. This will encourage the repetition of successful behaviours, strengthening the learner's grasp of the material and boosting their confidence.

5. **Personalization Through Reference of Past Events:** We keep track of the learner's past mistakes and successes, and use this information to guide the prompts in subsequent lessons. This simulates EM retrieval and facilitates progressive learning.

6. **Examples:** We demonstrate how the LLM can process various user inputs and make sense of the provided data in the prompt.

7. **Repetition:** We repeat certain rules in the prompt to emphasize their importance.

3.2.3 Prompt Engineering Techniques

Directional Stimulus

Li et al. introduced a small and adjustable policy model to produce directional stimulus prompts [130]. These prompts act as nuanced and instance-specific hints that guide LLMs, such as GPT-4. The primary objective of these hints is to steer the model towards generating specific desired outcomes. For example, generating responses following a particular theme of conversation.

We don't use a second prompt generation model in ELLA. However, we utilize the prompt principles that have been introduced in this paper. We explicitly instruct the GPT-4 LLM to generate responses that follow a specific theme or purpose. For instance, in the prompt we employed for Lesson 2 (see 3.2), we asked the LLM to notify the learner whenever they made a mistake that was made previously in the first lesson. We also instructed the LLM on how to make sense of the mistakes and success events that we feed through the prompt.

Few-shot Prompting

Few-shot prompting involves providing demonstrations in the prompt to help the model learn in context. These demonstrations provide conditioning for subsequent examples, which helps the model generate better responses, as explained by [131].

In our lesson prompts, we provide examples of how the language model should respond to certain user inputs (such as when they make a mistake). This not only teaches the correct linguistic structures and rules for the LLM but also models the appropriate pedagogical tone and feedback style. We also provide examples of how the LLM can use the data we dynamically feed through the prompts to personalize its interactions based on the learner's proficiency level, known mistakes, and other contextual data. This adaptability makes the educational process more engaging and relevant for the learner.

For instance, we use placeholders for *mistakes* and *successes* that are filled with specific content related to the learner's previous lessons in 3.2 and 3.3. This allows the language model to reference past errors or successes, reinforcing learning through repetition and validation.

Role-Setting

Role setting is the process of defining tasks and responsibilities for language models to ensure they perform specific functions effectively. In a paper by Thomas and his colleagues, they explore how to set the roles of LLMs to act as a relevance assessor for search queries [132]. This involves evaluating and labelling search results based on their relevance to a given query. The role is defined by the instructions embedded within the prompts provided to the LLM. Specific phrases are introduced to make the LLM act as the desired role effectively.

We incorporate role setting into our lesson prompts (seen in 3.1, 3.2 & 3.3) by specifically instructing the LLM that it is an expert Dutch language teacher and that it is conversing with a Dutch language learner. In 3.2, we provide context indicating that the LLM is conducting a subsequent lesson with the learner, while in 3.3, we instruct the LLM to recall and discuss content from previous lessons.

Prompt Chaining

To effectively separate lessons and analyze conversations, we employed the technique of prompt chaining. As described by Wu et al., prompt chaining is a powerful method for customising human-AI interactions, enhancing both transparency and controllability [133]. This approach involves breaking down tasks into manageable sub-tasks, each designed to address specific aspects of the learning process. In our system, the following sub-tasks are defined:

1. **Lesson 1:** Engage in a conversation about the learner's favourite cities.
2. **Analyze Lesson 1:** Review the conversation from Lesson 1 to identify and catalogue the learner's mistakes and successes.
3. **Lesson 2:** Discuss the learner's favourite Dutch cuisine.
4. **Real-time Cross-reference:** During the conversation, continually compare with episodes from Lesson 1. If similar mistakes or correct usages occur, remind the learner of previous instances and provide appropriate corrections or praise.
5. **Analyze Lesson 2:** Examine the conversation from Lesson 2, compiling a list of the learner's mistakes and successes.
6. **Reminiscing Session:** Encourage the learner to reminisce and discuss topics from the first two lessons, while also answering any questions they might have related to the grammar and vocabulary used in those lessons.
7. **Ongoing Cross-reference:** During this conversation, refer back to episodes from Lessons 1 and 2. Highlight recurring mistakes or correct usages, providing reminders and feedback as necessary.

Once these sub-tasks are identified, the GPT-4 LLM is sequentially prompted with each sub-task. The response generated from one sub-task serves as the input for the subsequent prompt. This approach allows for easier debugging of model responses, facilitating detailed analysis and targeted improvements across various stages of the learning process.

3.3 Comprehensive System Overview

3.3.1 Room Setup

FIGURE 3.2: Set-Up of The Components

To understand how our system interacts with the physical components, we must first visualize the orientation of said components. The setup consists of a Furhat animatronic face, a PC with an attached keyboard and mouse, and a microphone. All these components are strategically placed in front of the user to ensure a seamless language-learning experience.

From Figure 3.2, we see that the Furhat animatronic face has been positioned on the left side of the table, slightly tilted towards the user seated in front of it. This left-sided placement ensures an unobstructed line of sight and interaction with the monitor displaying our system's web application. The slight tilt is intentional, prioritizing the user's focus on the Furhat face during conversation and ensuring that Furhat's built-in camera, responsible for facial recognition, remains unobstructed. This setup allows Furhat to maintain direct eye contact with the user throughout the conversation.

The monitor of the PC is positioned at the centre of the table as the primary display for running and visualizing the web application. This central placement ensures that it is easily viewable from all angles, providing a clear visual reference for the user during interactions. The keyboard and mouse are placed in front of the monitor to facilitate seamless interaction with the system.

The microphone is positioned on the right side of the monitor, away from the keyboard and mouse. This strategic placement ensures that it is close enough to effectively capture the user's voice without interference, facilitating accurate audio input acquisition.

By arranging the equipment in this manner, we create an environment conducive to natural, interactive language learning sessions, thereby enhancing the reliability and validity of the experimental results obtained from the same.

3.3.2 Application Design

Our web application is divided into 6 pages (refer A), and the list of pages is displayed on a sidebar for efficient navigation. An overview of the same is given in 3.3.

UserID Generation

When the application is launched, a random unique user ID is generated which is then used to store data from the interaction. We believe that using a randomly generated user ID is safer than using names or serial numbers to store session data. Random user IDs help to anonymize the data, making it difficult to trace the information back to the user through direct identifiers in datasets, and thus protect their identity. [134]. Studies have shown that names and serial numbers can often be linked back to the individual, increasing the risk of re-identification[135]. Using random user IDs also complies with EU GDPR, which requires researchers to minimize the amount of personally identifiable information (PII) collected and stored [136].

User Profile Creation

Once the user receives their unique ID, they are prompted to click the "About You" page. This page is used to create a personalized user profile for the lessons. On this page, the user is asked to select their language proficiency from a drop-down menu with 4 choices: Beginner, Intermediate, Advanced, and Native. These choices align with the standard Dutch language proficiency levels (A1 - Beginner, A2 - Elementary, B1 and B2 - Intermediate, C1 - Advanced, C2 - Native). Next, the user is asked to enter some of their favourite cities and foods in the Netherlands. Finally, the user is requested to input some Dutch phrases they have picked up from day-to-day conversations.

At the end of the page, the user is prompted to click a button to save their profile into the system's memory. A popup then directs the user to navigate to the next page, which contains the first lesson. The data from user profiles is used in the system prompts for the lessons, which are shown in 3.1, 3.2, and 3.3.

Lesson 1: Favourite Dutch Cities

In the first lesson, the user is required to have a conversation about their favourite cities in the Netherlands (which they mentioned on the previous page) with the LLM through the Furhat robot. The rationale behind this lesson topic is given in 3.3.5. During the conversation, whenever the user clicks the "Start Recording" button, Furhat listens to the user and their input is transcribed into the text after the user clicks "Stop Recording". This text is then transmitted to the LLM through an

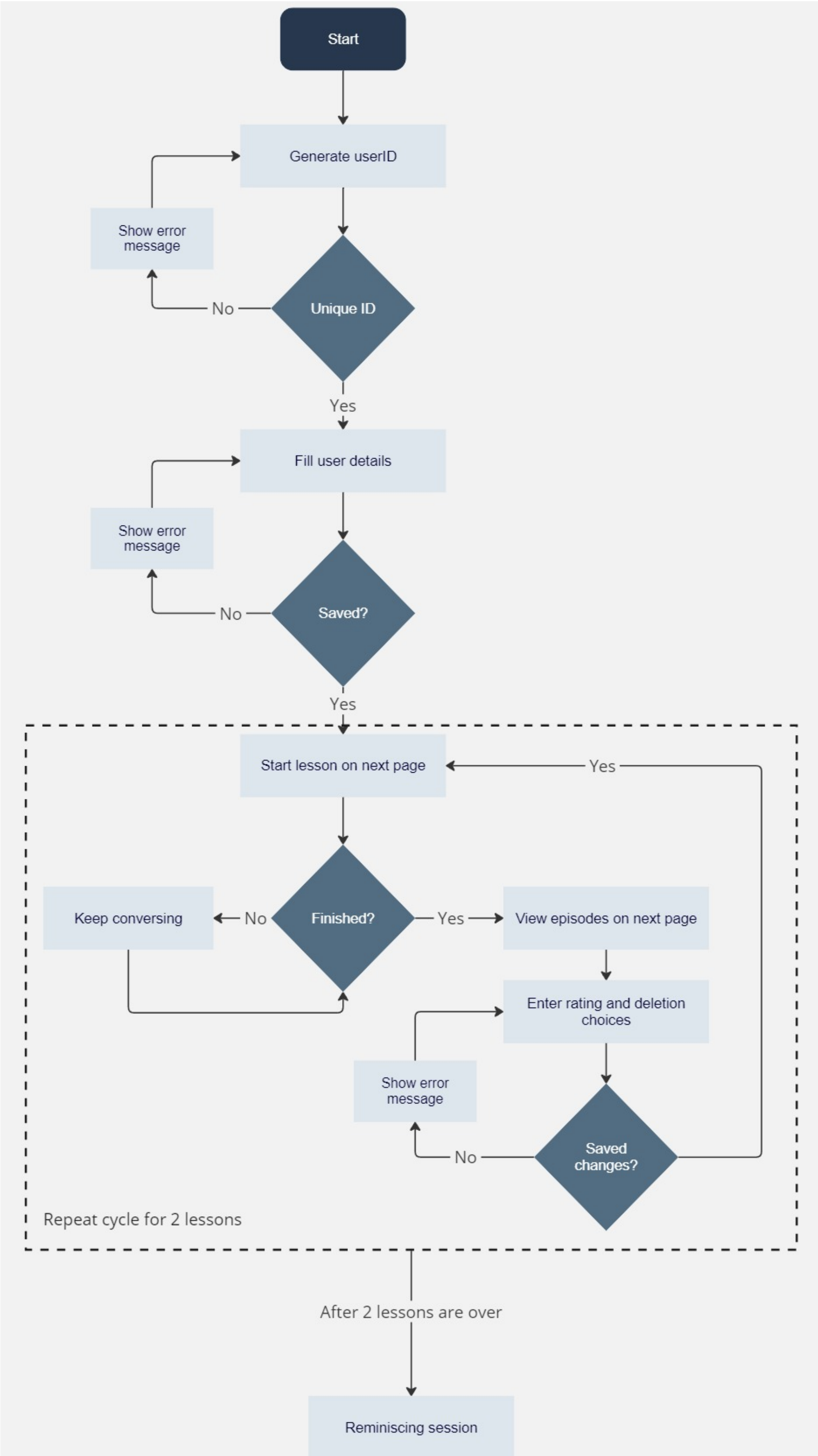

FIGURE 3.3: Flow Diagram of Our Web Application

API call, which processes the input and generates a response. The response is then sent to the Furhat robot to produce the speech output.

The lesson is structured as a natural conversation since we primarily focused on teaching spoken Dutch to the learner. During the lesson, whenever the user makes a mistake, the LLM politely corrects them and asks them to repeat the correct phrase. For example, if the user said "*Ik vind the people bijzonder.*", the LLM identifies the mistake (the user incorrectly used *people* which is English) and asks them to politely correct their response by saying "*Om het nog better te maken, zeggen we in the Nederlands 'Ik vind de mensen bijzonder'. Zou je het met deze formulering kunnen proberen?*". Here correcting the user politely is important since research conducted by Hattie and Timothy indicates that feedback is most effective when delivered in a manner that is respectful and supportive [137].

Reviewing System-Generated Episodes from Lesson 1

When the user wishes to end the lesson, they are instructed to click on the next page in the sidebar titled "Lesson 1 Performance." Upon opening this page, three models are executed. The first model is the expert episode identifier, which is detailed in 3.3.3. The second and third models extract metadata from the conversational script of the first lesson, and their workings are explained in 3.4.2.

After the models finish generating results, the user is presented with a list of episodes. Next to each episode, there is a rating dropdown, which the user can use to rate the episodes on a scale of 1 to 5. A rating of 1 indicates that the episode was not helpful at all for Dutch language learning, while a rating of 5 indicates that the episode was very helpful.

After the user has rated the episodes and indicated their preferences for deleting or saving certain episodes, they can save the changes by clicking the "Save Changes" button. Once clicked, the saved episodes (Episodes not indicated for deletion) are moved to the MongoDB database for Lesson 1 for permanent storage (long-term EM), while the deleted episodes (Episodes indicated for deletion) are transferred to the deletion log in the user's profile. We have separate databases for storing lesson episodes, lesson metadata, and user profiles as explained in 3.1.2.

Lesson 2: Favourite Dutch Food

Once the user rates each episode from Lesson 1 and saves their changes, they are prompted to go to the next page titled "Lesson 2" by clicking on the page listed in the sidebar. This is also a conversational lesson about the user's favourite Dutch dishes and cuisines (eg: Stroopwafels). The rationale behind this lesson topic for the second lesson is given in 3.3.5.

During Lesson 2, our system ELLA showcases signs of EM by referring to the data from the previous lesson. For example, in 3.4, our system remembers the user's favourite words that were mentioned earlier in the "About You" page and encourages the user whenever they use them correctly in the context. ELLA also suggests better words and asks the user to confirm their understanding by repeating the phrase, as seen in 3.5.

Through injecting the saved episodes from Lesson 1 in the prompt for Lesson 2 through placeholders, we are also able to simulate that the system remembers the user's past successes and indicates them in the conversation as seen in 3.6. A high-level overview of how this process occurs for subsequent lessons after the first lesson is shown in 3.7.

FIGURE 3.4: Remembering Favourite Words

FIGURE 3.5: Suggesting Additional Words

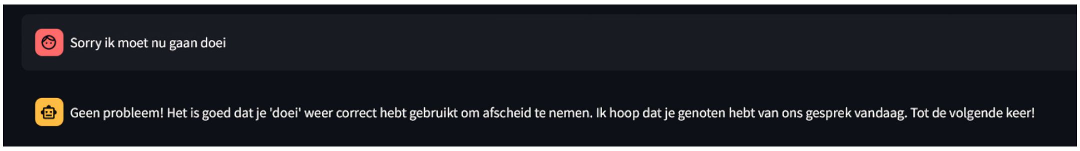

FIGURE 3.6: Remembering Past Phrases Used Correctly

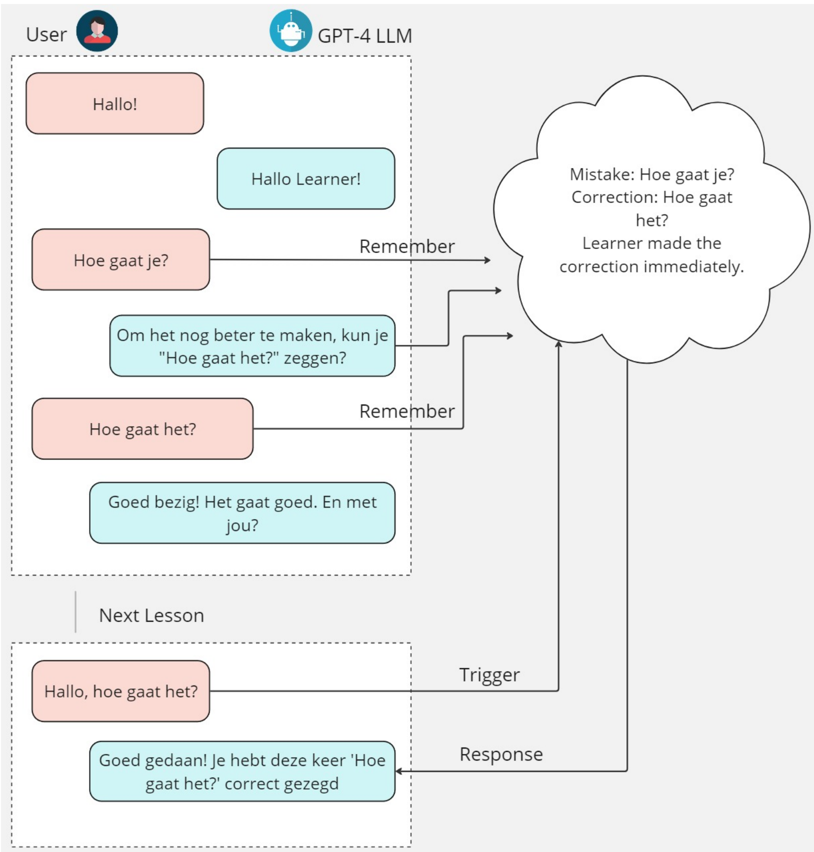

FIGURE 3.7: How ELLA's Memory Works

Rating System-Generated Episodes from Lesson 2

When the user wants to end the second lesson, they are instructed to click the next page on the sidebar titled "Lesson 2 Performance". Here, similar to the generation process after Lesson 1, we generate the lesson metadata along with a set of episodes using the models described in 3.4.2 and 3.3.3 respectively. The user is required to rate the episodes again on a scale of 1 to 5 and is asked to indicate their preference towards saving or deleting individual episodes from the system's long-term episodic memory. Once changes are saved by clicking the "Save Changes" button, the saved episodes are moved to the MongoDB database for Lesson 2 while the deleted episodes are again transferred to the deletion log in the user's profile.

Reminiscing Session

After saving changes, the user is directed to the final page, which is a reminiscing session. The rationale for this session is explained in 3.3.6. During the reminiscing session, the user can review the topics covered in previous lessons and ask questions related to the grammar and vocabulary used. The interaction ends when the user finishes reminiscing with the system.

3.3.3 Episode Generation

LISTING 3.4: Prompt that generates mistake and successful episodes from lesson conversations

```
Analyze the conversation from a text file between a Dutch language
    learner and a Dutch local, with a primary focus on meticulously
    identifying every single mistake made by the learner, including but
     not limited to grammar, vocabulary, and pronunciation errors.
    Capture corrections provided by the Dutch local for each mistake,
    along with explanations when available. Additionally, identify
    episodes of successful language usage and moments of encouragement,
     highlighting the learner's progress and positive reinforcement
    received.
Your response should structure the analysis as follows, with the
    username specified once for the entire conversation, rather than
    within each episode:
{{
 "username": "{userid}",
  "topic": (topic of the conversation),
  "episodes": [
    {{
      "episodeNo": ,
      "dialogNo": (the number of the dialog in which the episode is
          identified),
      If the episode is a mistake:
      - "episode_type": "Mistake",
      - "mistake_phrase": (the incorrect phrase the learner used),
      - "mistake": (a brief description of the mistake),
      - "mistake_type": (grammar, vocabulary, etc.),
      - "context": (in what context of the conversation is the phrase
          said),
```

```
      - "correction": (the correct form or usage),
      - "summary": (Concisely summarize the conversational episode by
          explicitly mentioning the phrase where the user made a mistake
          . Include the mistake, the correction, and the rule involved
          in no more than 100 words. Briefly elaborate on why the
          correction is necessary, aiming to clarify and reinforce the
          correct language usage. Ensure the summary encapsulates these
          elements clearly for educational reinforcement.)

      If the episode is a success:
      - "episode_type": "Success",
      - "correct_phrase": (the phrase used correctly by the learner),
      - "context": (in what context of the conversation is the phrase
          said),
      - "correct_type": (grammar, vocabulary, etc.),
      - "summary": (In no more than 100 words, describe the learner's
          correct language use, contextual relevance, and the success
          demonstrated along with explicitly mentioning the correct
          phrase being used. Offer concise praise and highlight the
          significance of the correct usage for ongoing language
          development.)
    }}
    // Additional episodes follow the same structure
  ]
}}
Your response should be in JSON format.
```

To generate the episodes after a lesson has ended, we feed the conversational script into the GPT-4 LLM. We treat the LLM as an expert episode identifier, based on previous research indicating GPT's capability as an expert annotator in various scenarios [138][120].

We design the system prompt (seen in 3.4) to obtain a comprehensive analysis of the conversation between a Dutch language learner and a native Dutch speaker (which is GPT-4 in this case). The primary objective is to meticulously identify and document every mistake made by the learner, including errors in grammar, vocabulary, and pronunciation, along with their context. Additionally, we ask the LLM to provide summaries of the episodes when available. The prompt also emphasizes recognizing successful language usage and moments of encouragement, highlighting the learner's progress and the positive reinforcement provided. A detailed explanation of the specific mistakes and successes considered is provided in 3.1.1.

Finally, we instruct the LLM to structure its response in a JSON format. For each conversational episode, GPT-4 categorizes it as either a "Mistake" or a "Success." For mistakes, the response includes details such as the incorrect phrase, mistake type, context, correction, and a summary explaining the episode. For successes, it includes the correct phrase, type, context, and a corresponding summary. This consistent structure for both mistake and success episodes facilitates uniformity, as we convert the JSON to BSON for easy storage in a MongoDB database.

To obtain the response from the LLM, we combine the conversational script with a query that briefly explains the task and then call the API. The combining process is illustrated in the code shown in 3.5. An example of the response generated is shown in D.

LISTING 3.5: Code to generate episodes through combining prompt with conversational script

```
user_query = "Can you identify the episodes between me(Learner) and the
     dutch teacher(Assistant) from this conversation? "
episode_query = user_query + "\n---\n" + script_content
combined_input = prompt + episode_query
messages = [
    {"role": "system", "content": prompt},
    {"role": "user", "content": episode_query}
]
response = client.chat.completions.create(
    model='gpt-4-0125-preview',
    response_format={"type": "json_object"},
    messages=messages
)
response_data = response.choices[0].message.content if response else {}
print(response_data)
```

3.3.4 Episode Rating and Deletion

After the episodes have been generated by the GPT-4 LLM, we present a summary of each episode to the user for evaluation. Users are asked to rate each episode summary on a helpfulness scale from 1 to 5, where 1 indicates not helpful at all and 5 signifies very helpful towards their FL learning (as seen in 3.8).

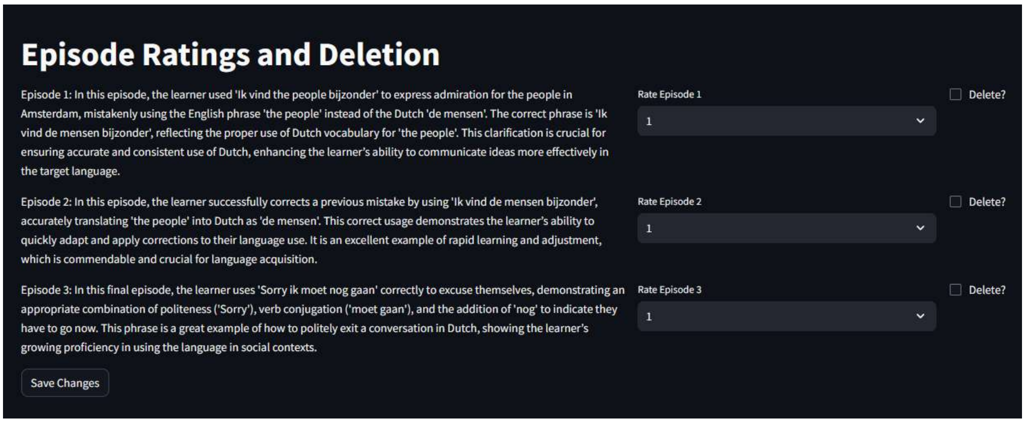

FIGURE 3.8: Reviewing Episodes

Additionally, users have the option to delete episodes from the system's episodic memory by clicking the **Delete?** checkmark if they do not agree with the summary for various reasons. We add a flag variable called *will_delete* to each episode before displaying the summaries for review. If an episode is marked for deletion, we set the value of *will_delete* to true; otherwise, we set it to false. After the user saves their changes, episodes marked for deletion (with *will_delete* set to true) are removed from the episode catalogue and transferred to the deletion log. The remaining episodes are permanently stored in the system memory. An example of a deleted episode is shown in 3.6.

LISTING 3.6: Example of an episode opted for deletion

```
"episodeNo": 2,
"dialogNo": 4,
"episode_type": "Success",
"correct_phrase": "doei",
"context": "Assistant recognizes learner's previous correct use of '
    doei' at the end of conversations.",
"correct_type": "vocabulary",
"summary": "Here, the assistant positively acknowledges the learner's
    correct previous use of 'doei' to conclude conversations. This
    success underscores the learner's ability to use specific
    vocabulary correctly according to social conventions, contributing
    positively to their language acquisition journey.",
"rating": 1,
"will_delete": true
```


This rating and deletion process is designed to work synergistically, providing insights into the user's perceptions of each episode. By collecting ratings and noting whether episodes are deleted, we can infer the reasons behind low ratings or deletions. Understanding these reasons is crucial for refining the system and improving the accuracy of future analyses. We speculate the following reasons for low ratings or deletions:

- **Technical Errors Caused by ASR:** Episodes may be rated poorly or deleted if there are inaccuracies in transcribing the spoken conversation, leading to misunderstandings in the analysis.
- **Misclassification of Success Episodes as Mistake Episodes:** Users may rate episodes low or delete them if the LLM incorrectly identify a successful language usage as a mistake.
- **Misidentification of System Responses as Mistake/Success Episodes:** Irrelevant episodes can be generated if the LLM mistakenly evaluates its responses during the conversation instead of the user's, leading to irrelevant episodes.
- **High Anxiety Associated with Specific Mistake Episodes:** If a user finds certain mistakes particularly stressful or discouraging, they might choose to rate them poorly or even delete them from memory.
- **Unsatisfactory explanations:** If the LLM gives explanations that are too short, too long or too vague, users may rate them low or delete them.

We believe that this rating and deletion process also reduces FL learning anxiety by enhancing users' sense of control over the learning process. Research suggests that when learners feel a greater sense of control over their educational activities, they experience lower levels of anxiety and higher levels of motivation and engagement [139] [140]. By enabling learners to curate their future lessons, we mitigate feelings of helplessness associated with FL learning, thereby reducing cognitive suppression. This participatory approach empowers learners and fosters a more personalized and supportive learning environment. Consequently, it leads to more effective and enjoyable foreign language acquisition through cognitive reappraisal[141].

3.3.5 Rationale Behind Topics for Lessons 1 and 2

We carefully selected the topics for Lesson 1 and Lesson 2 because we believe they provide the most effective learning opportunities for language learners. In Lesson

1, we used the GPT-4 LLM to initiate a conversation about the learner's favourite city. This approach is highly effective as it connects with personal relevance and motivation. Research by Schmitt and Carter has shown that learners are more engaged and motivated when the content is personally relevant and something they are passionate about [142]. Additionally, discussing cities allows for the introduction of contextual vocabulary, including geographical terms, descriptive adjectives, cultural references, and travel-related phrases. This is in line with Nation's principles of vocabulary acquisition, which highlight the importance of using language in context for effective vocabulary learning [143].

The process of learning a language is closely connected to understanding the culture associated with it. Discussing food in the second lesson offers learners valuable insights into Dutch culture and traditions. This cultural immersion aids learners in enhancing their cultural competence. Furthermore, food is a popular topic in everyday conversations, making it an ideal choice for language practice. By learning how to talk about their favourite dishes, learners can pick up useful phrases they will likely use in daily interactions. This approach aligns with the principles of communicative language theory, which emphasizes the use of functional language [144]. According to research by Calvert and his colleagues, this method also helps learners develop conversational skills that apply to real-world interactions [145]. Engaging in such meaningful conversations prepares learners to communicate more effectively in various social settings, thereby increasing their overall comfort with using the target language.

3.3.6 Rationale Behind Reminiscing Session

We designed a session towards the end of the interaction to allow learners to discuss the topics covered in previous lessons and address any grammatical mistakes that were previously identified and saved in the system. Scientifically, revisiting previously learned material at spaced intervals is a well-established method to improve language retention and counteract the natural tendency to forget information over time, as explained by Ebbinghaus' forgetting curve [146]. We believe that this review session also offers the opportunity for learners to examine and discuss the grammatical errors identified in earlier lessons. This allows learners to correct their mistakes and gain a clearer understanding of the underlying rules, which, according to Lyster and Ranta's research, is an important aspect of second language (L2) acquisition [147].

When learners are encouraged to ask questions and discuss their specific difficulties, they become more autonomous and actively engaged in the learning process. Research indicates that by involving learners actively in their learning, they develop a deeper understanding and a personal investment in the language learning process [148]. This approach provides a platform for addressing individual doubts and meeting specific needs, which, according to Tomlinson, is crucial for accommodating diverse learning styles for lexical development that can be applied to real-life scenarios [149]. Additionally, Schön's research suggests that reflection helps learners make sense of their learning and apply their insights to future learning situations [150].

3.4 System Architecture

3.4.1 Input-Output

Whenever the user speaks into the microphone, their speech input is converted into text. This text is then used for Response Generation and Natural Language Understanding (NLU) operations. This conversion is achieved by using a custom Streamlit ASR (Automatic Speech Recognition) plugin [3]. Ideally, with access to high-grade computational resources, we would utilize OpenAI's Whisper, renowned for its accurate context-aware speech-to-text transcriptions [4]. However, to minimize latency and reduce computational costs, we opted for the Streamlit plugin.

For generating the output, the user input is first processed using the GPT-4 LLM to create a response (the detailed generation process is explained in Section 3.1.1). Subsequently, we use Furhat's built-in Text-to-Speech (TTS) module to convert the text response into speech, which is then delivered to the user through Furhat's speakers. This streamlined process ensures efficient and effective interaction with minimal delay.

3.4.2 Natural Language Understanding

To extract meta-information from the conversations of the lessons, we perform the following NLU operations:

Preprocessing the Conversational Script

First, we pre-process the text by doing the following in order: tokenization (splitting the text into individual words), stop-word removal (removing common words that do not have a significant meaning), stemming and lemmatization (reducing words to their root forms). This preprocessing ensures that the text is clean and ready for NLP tasks.

Topic Modelling

To extract the topics from the text script of the conversation from each lesson, we utilize the power of Latent Dirichlet Allocation (LDA). LDA involves identifying abstract topics within a collection of documents. It works by assuming that each document in a collection is generated through a process where a distribution over topics is selected for the document, and for each word in the document, a topic is chosen from this distribution. Subsequently, a word is selected from the chosen topic's word distribution. The model includes parameters such as α (Dirichlet prior on per-document topic distributions) and β (Dirichlet prior on per-topic word distributions), which guide the probability distributions. This probabilistic model helps us reveal the hidden thematic structure in our conversational scripts.

We first convert the pre-processed text into a document-term matrix, where the rows represent documents and columns represent words. The matrix entries indicate the frequency of each word in each document.

This matrix is then passed as the input into the LDA model. During this process, we specified the number of topics to be extracted as 5, which we believe to be a good granularity of topics. Once the LDA model runs, it generates a set of topics, each represented by a group of words and their associated probabilities. We use these

[3] https://pypi.org/project/streamlit-mic-recorder/
[4] https://github.com/openai/whisper

keywords to understand the topics that are generated. For example, a topic containing words like "Delft," "City," and "canals," could be labelled as "Delft" or "Travel." An example of the topics extracted for a conversational script is given below:

["Topic 0: gaan, vandaag, heel, goed, nederland, hebt, correct, gesprek, leiden, context", 'Topic 1: hallo, gaan, the, people, vind, sorry, mensen, musea, bijzonder, amsterdam', "Topic 2: mensen, vind, bijzonder, we, maken, formulering, eigenlijk, 'ik, proberen, zeggen", 'Topic 3: leiden, leuk, cultuur, vind, bijzonder, vindt, praten, stad, goed, heel', 'Topic 4: amsterdam, musea, bijzonder, inderdaad, horen, bezoeken, bezocht, willen, mooi, museum']

Named Entity Recognition

After obtaining the topics from the conversation, we extract the entities the user talks about and their sentiment towards the same. To first extract the entities from the conversation, we use a pre-trained Dutch NER RobBERT model [5] that is specially designed for Named Entity Recognition (NER). This is based on the RobBERT model, which uses the well-optimized architecture of RobBERTa. It has been pretrained on a large Dutch corpus from the OSCAR dataset and fine-tuned to identify and categorize entities such as names of people, organizations, and locations in Dutch texts. Due to the 12 self-attention layers in the transformer architecture, RobBERT can capture complex linguistic nuances, making it an excellent choice for our task [151]. We input the preprocessed conversation script into the HuggingFace NER pipeline and obtain an entity and a label specifying the type of entity, as shown: **['name': 'Leiden', type: 'B-LOC']**

Sentiment Analysis

To determine the user's sentiments about the identified entities, we utilize a RobBERT model that has been fine-tuned specifically for sentiment analysis [6]. This particular RobBERT model was developed by the DTAI research group at KU Leuven through fine-tuning on the Dutch Book Reviews Dataset from *hebban.nl*. The dataset includes reviews initially rated on a five-star scale, which were then converted to positive, neutral, and negative sentiment classes. Despite the model creators' indication that it might not perform well without additional fine-tuning, we discovered that the model's output met our requirements when used in the Huggingface sentiment analysis pipeline.

The model generates a sentiment score and a label for every entity identified by the NER pipeline, as shown: **['name': 'Leiden', 'sentiment': ['label': 'Positive', 'score': 0.99949932098838867]]**

3.4.3 Our Memory Architecture

The memory architecture used in ELLA is heavily inspired by the architectures of [103] [152] and [153]. It is divided into four specific modules; a working conversation log, a long-term semantic memory, a short-term episodic cache and a long-term episodic memory. Figure 3.9 shows our architecture, whose individual modules are explained in detail below.

[5] https://huggingface.co/pdelobelle/robbert-v2-dutch-ner

[6] https://huggingface.co/DTAI-KULeuven/robbert-v2-dutch-sentiment

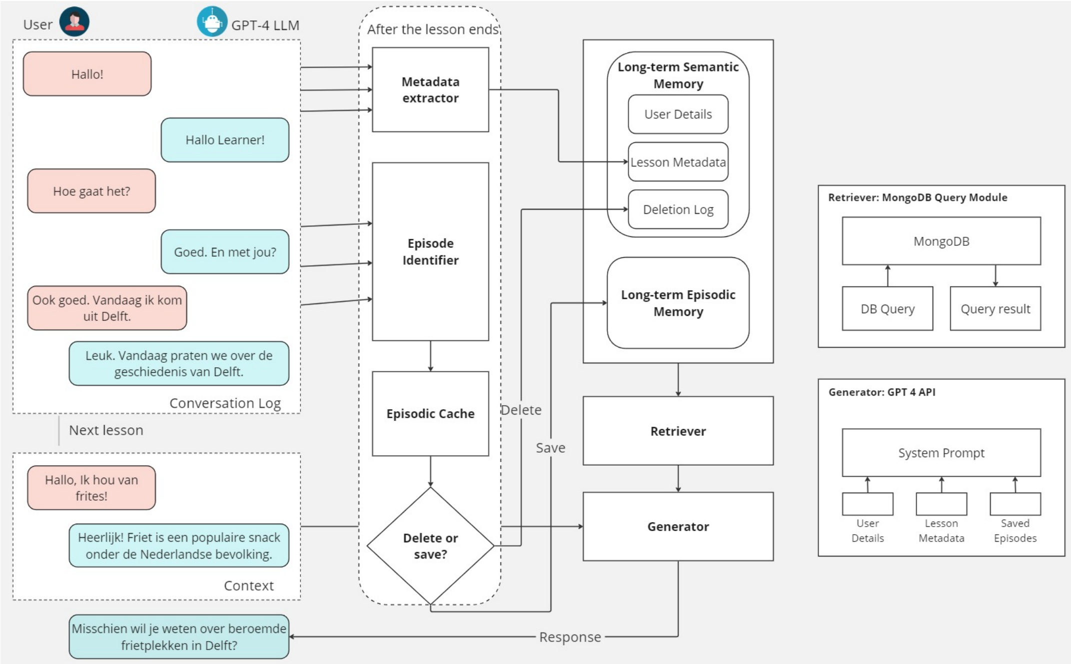

FIGURE 3.9: Memory Architecture

Working Conversational Log

We use a short-term conversational log, also known as working memory, to store the dialogues of both the user and the GPT-4 LLM. The use of short-term working memory to store conversational dialogues has been extensively researched previously. For instance, Henderson et al. employed a state-tracking long short-term memory (LSTM) model to store and update the dialogue state during conversations [154]. Similarly, Vinyals and Le explored the application of recurrent neural networks (RNNs) in a sequence-to-sequence conversational model to manage dialogues. Their model utilizes the hidden states of RNNs as a form of short-term working memory, enabling it to track the conversational history across turns [155].

Each time a dialogue occurs during a lesson, this log, stored as a .txt file, is appended with the latest exchange. Each time the user finishes their interaction with the system (ie. finishes two lessons, the episode rating process for the lessons and the reminiscing session), the file is wiped clean so no data leaks occur.

Long-Term Semantic memory

The long-term semantic memory architecture we use in our system is inspired by the work of Xu et al. [103]. Similar to their conversational agent, we utilize this memory primarily to store user details in the form of profiles. These profiles are created on the "About You" page of our web application (refer to 3.3.2). The details stored in user profiles include their language proficiency and specific knowledge related to the Dutch language and culture, such as cities, cuisines, and known Dutch phrases (example shown in 3.7).

LISTING 3.7: Data stored in User Profiles

```
"userID": "Z041",
"language_proficiency": "Beginner",
"known_cities": "Leiden",
```

```
"known_phrases": "lekker",
"known_food": "Stroopwafel"
```

In addition to user profiles, we use long-term semantic memory to store metadata from the lessons. This metadata is extracted using topic modelling and other NLU techniques, which are explained in detail in 3.4.2. Finally, we also store the deletion log from each lesson in long-term semantic memory. The deletion log contains records of episodes that users chose to delete during the episode review process after lessons 1 and 2 (seen in 3.6). Possible reasons for users deleting episodes are described in 3.3.4. For example, in 3.6 the user deleted the episode since the system identified its dialogue as a conversational episode instead of only generating conversational episodes from user dialogues.

Short-Term Episodic Cache

To address the limitations of Huang et al.'s paper mentioned in 2.3.7, we use a short-term episodic cache. The cache improves editing speed, enhances user control, and reduces computational costs. Crucially, allowing users to edit the cache means their actions do not affect the long-term EM until changes are saved. This approach prevents memory inconsistencies and facilitates scalability, as the short-term cache is cleared after each session, and only saved episodes are appended to long-term memory. This method is computationally efficient, as only the cache changes. Zhang et al. support this approach, noting that knowledge editing operations are better suited for small-scale memory adjustments to minimize computational costs [156].

In our episodic cache .json file, we store episodes generated by our episode identifier before they are chosen for deletion. After users rate the episodes and indicate preferences for deletion, the cache is cleared by transferring episodes not marked for deletion to the long-term episodic MongoDB database and deleting marked episodes from the cache file, storing a copy in the deletion log.

Long-Term Episodic Memory

We use the long-term EM to store the episodes that are saved by the user (ie. not indicated for deletion), as seen in 3.8. We have two MongoDB databases serving as our long-term episodic memory, one for the episodes that are saved from Lesson 1 and one for the episodes that are saved from Lesson 2. We use separate databases for the two lessons, to allow us to differentiate the episodes stored clearly between the two lessons. This also makes it easier for us to prevent any unwanted data leaks that might cause the LLM to generate inaccurate responses.

LISTING 3.8: An example of the data present in LT Episodic Memory

```
"episodeNo": 4,
"dialogNo": 7,
"episode_type": "Success",
"correct_phrase": "Ik heb nog niet naar een museum geweest maar ik ben
   wel naar de dierentuin geweest",
"context": "Correcting a previous mistake regarding a visit to the
   museum and the zoo.",
"correct_type": "grammar",
```

```
"summary": "The learner shows improvement by correctly reformulating
    their previous mistake, accurately using 'nog niet' for 'not yet'
    and 'ben... geweest' for 'have been'. However, there remains a
    slight mistake in the phrase 'Ik heb nog niet naar een museum
    geweest', it should be 'ik ben', showing a learning process. The
    successful repetition of the corrected structure 'ik ben wel naar
    de dierentuin geweest' demonstrates the learner's ability to adapt
    and correct errors based on feedback, a crucial skill in language
    learning.",
"rating": 4
"will_delete": "false"
```


Memory Retrieval

To access information from our long-term semantic and EMs, we employ MongoDB queries, a method utilised by many researchers in the past [157] [158]. To execute these queries in Python, as it is the primary language of our system, we make use of the PyMongo library. This library offers tools for establishing a connection to the database, carrying out CRUD operations, and efficiently managing data for Python applications.

For a particular user, we retrieve data such as user profile information, metadata of Lessons 1 and 2 (including topics, entities, and sentiment labels), and saved episodes from these lessons. The retrieved data is then incorporated into the system prompt for subsequent interactions (as shown in Figure 3.9).

Chapter 4

Experimentation

This chapter outlines the experimental design used to investigate the sub-research questions (3 and 4) central to our study. Section 4.1 provides details regarding the recruitment of participants. Section 4.2 gives an overview of the questionnaires we used in our experiments. Section 4.3 explains the division of participants for our experiments. Finally, section 4.5 presents the results of our study.

4.1 Participants

We recruited 28 participants aged between 20 and 30 who are either currently learning Dutch or have previously learned Dutch through the TU Delft Dutch language course. We aimed to include participants with varying levels of Dutch language proficiency to ensure a fair and unbiased evaluation of our system.

To ensure a random recruitment process, participants were asked to fill out an online survey indicating their willingness to participate. According to the survey results, 60.7% of the participants were male, 35.7% were female, and 3.6% preferred not to disclose their gender. Regarding language proficiency, 57.1% of the participants were at the Beginner level, 25.0% were at the Intermediate level, none were at the Advanced level, and 17.9% had near-native proficiency in Dutch.

4.1.1 Ethics

We made sure not to use any personally identifiable information (PII) such as participant names to store the content from the interactions. Instead, we assigned a random alphanumeric user ID to store the data generated from participants' interactions in our MongoDB. This was done to reduce the possibility of linking the database data to specific individuals.

All participants were informed of the risks associated with interacting with a closed-source model like the GPT-4 LLM through an informed consent form, and their data was used only after they provided their consent. Throughout the experiment, we ensured that our data storage and processing methods complied with GDPR. Data related to the experiments was stored on a personal laptop of the principal researchers, while all data from the interactions with our system was stored in a secure MongoDB repository.

4.2 Materials

4.2.1 Foreign Language Speaking Anxiety Scale (FLSAS)

We asked the participants to take the FLSAS Questionnaire before engaging with our system. This questionnaire was designed to measure their baseline anxiety levels related to speaking a FL. The questionnaire consisted of 17 questions that are asked on a 5-point Likert scale (1 - "Strongly Disagree", 2 - "Disagree", 3 - "Neutral", 4 - "Agree", 5 - "Strongly Agree), resulting in scores ranging from a minimum of 17 to a maximum of 85.

The questionnaire evaluated various dimensions of speaking anxiety, including discomfort in speaking, fear of making mistakes, and apprehension about communicating in a language-learning environment [159][160]. The responses received from this questionnaire were utilized to understand the overall emotional and psychological readiness of the learners before they engage in a language-learning speaking activity. A copy of this questionnaire can be found in A.

4.2.2 State Anxiety Questionnaire

Keeping our sub-research question 3 in mind, we requested participants to complete a questionnaire following their interaction with our system. This questionnaire, utilizing the State component of the State-Trait Anxiety Inventory (STAI) [161], measured their state anxiety levels related to speaking Dutch with our system. It comprised 20 questions, each rated on a 4-point Likert scale (1 - "Not at all", 2 - "Somewhat", 3 - "Moderately so", 4 - "Very Much so"), resulting in scores ranging from a minimum of 20 to a maximum of 80.

The STAI is a well-established tool for assessing State Anxiety, which refers to the temporary condition of feeling anxious or worried in response to specific situations or stimuli. The questions covered various aspects of State Anxiety, including nervousness, tension, apprehension and worry. Researchers in the past have used this questionnaire to determine the anxiety levels of individuals in response to a speaking task, like speaking in public or speaking a foreign language [162] [163]. A copy of this questionnaire can be found in B.

4.2.3 Perceived Learning Questionnaire

In addition to the State Anxiety Test, we asked participants to fill out a questionnaire that contains open-ended questions to answer our sub-research question 4. The purpose of this questionnaire was to gather feedback on the perceived learning benefits of our system. We aimed to use this information to qualitatively assess our system and determine if it has achieved our goals of helping learners reduce speaking anxiety and improve their language skills. The insights gained will inform future design improvements for language-learning applications to reduce anxiety associated with learning and speaking a FL.

The questionnaire includes the following questions:

1. Describe your overall experience using our system. What did you find most beneficial?

2. How has our system impacted your confidence in using the language you are learning?

3. Can you share an instance where you felt our system significantly helped you improve a specific language skill?

4. How well do you think the content provided by our system matches your learning goals and level?

5. What challenges did you face while using our system? How do you think these challenges could be addressed?

4.3 Procedure

Students who opted to participate in our experiments were asked to first complete the FLSAS Questionnaire, as detailed in 4.2.1. The results from this questionnaire were used to assess the levels of FL Speaking Anxiety among our participants concerning the Dutch language. Following the completion of the questionnaire, participants proceeded with the experiment.

Our study employed a 2x2 factorial design, with the independent variables being the presence or absence of EM usage and the presence or absence of the deletion feature. This experimental design is inspired by Berg et al., who examined the impact of two independent categorical variables (with/without learning support and/or chat sessions) on one dependent continuous variable (anxiety)[164].

In this design, if there is no EM usage, episodes from previous lessons will not be referenced in the current lesson. If there is no deletion feature, participants were unable to delete system-generated episodes. The dependent variable in our study was their State Anxiety.

The condition without EM usage and deletion served as our control condition to establish baseline scores for comparison. The other three conditions—EM usage with deletion, EM usage without deletion, and deletion without EM usage-served as our experimental conditions.

We randomly assigned participants to these four conditions, with 7 participants per condition. After participants had completed their interaction with our system,

Condition	Deletion	No Deletion
EM Usage	1	2
No EM Usage	3	4

TABLE 4.1: Split Up of Conditions

they were asked to fill out the State Anxiety Questionnaire from the STAI, as detailed in 4.2.2. The data collected from this questionnaire was statistically analyzed to identify significant differences between the experimental conditions. This analysis helped us understand how our system impacted the learners' State Anxiety levels towards speaking Dutch.

After completing the State Anxiety Questionnaire from the STAI, participants were asked to respond to five open-ended questions, as detailed in 4.2.3. These questions were designed to assess whether our system provides any Perceived Learning benefits.

Through this experimental setup, we aimed to answer the following sub-research questions:

1. How does the usage of episodes stored in Episodic Memory, and the option to delete them, influence Foreign Language learners' speaking-related State Anxiety?

2. What specific learning benefits do language-anxious users perceive from this system?

4.4 Data Analysis Strategy

4.4.1 FLSAS Scores

To analyze the responses to the FLSAS Questionnaire in 4.2.1, we employed descriptive statistics, including the mean, standard deviation, and variance. Additionally, we determined the minimum and maximum scores to understand the spread of FL speaking anxiety levels among participants. A Kernel-Density Estimation (KDE) curve was overlaid to visualize the distribution of FL Speaking Anxiety scores.

4.4.2 State Anxiety Scores

For the analysis of the State Anxiety scores obtained from the State Anxiety Questionnaire in 4.2.2, we first categorized the scores based on the experimental conditions described in 4.3. We then used descriptive statistics, such as the mean and variance of the State Anxiety scores per condition, to compare the distribution of State Anxiety scores with the distribution of FLSAS scores from 4.2.1.

Given our 2x2 factorial experimental design illustrated in 4.3, where the usage of EM in lessons and the presence of the deletion feature were the independent variables, and the State Anxiety score was the dependent variable, we determined that ANOVA was the most appropriate statistical test. Theoretically, ANOVA allowed us to assess both the main effects and the interactions between factors [165]. Due to our limited participant size, we chose a less stringent statistical significance threshold of 0.1, compared to the standard 0.05 [166].

Initially, we validated whether our data met the three primary assumptions of ANOVA: Independence, Normality, and Homogeneity of Variance. The assumption of Independence was satisfied, as each interaction with our system was independent of the others, ensuring that data collected from each participant was not influenced by data from other participants.

To test for Normality, we used the Shapiro-Wilk test. The Shapiro-Wilk test on the State Anxiety scores yielded a p-value of 0.017, which is less than the significance threshold of 0.1. Therefore, we rejected the null hypothesis (H_0) that the State Anxiety scores are normally distributed.

To address this issue, we performed a log transformation on the State Anxiety scores and conducted the Shapiro-Wilk test again. The p-value obtained was 0.038, which still required us to reject the null hypothesis. Since the assumptions of ANOVA were not met, we switched to a non-parametric equivalent of the ANOVA test.

We selected the Aligned Rank Transform ANOVA (ART ANOVA) due to its robustness to violations of ANOVA assumptions. This method is suitable for factorial designs and does not require the assumption of normality to be validated [167]. The results of our ART ANOVA statistical test are described in 4.5.3.

4.4.3 Post-Hoc Tests

To better understand how the main factors from our ART ANOVA test - EM Usage and Deletion - interacted with each other across their different layers, we performed Contrast Analysis [168] [169]. This method allowed us to explore specific comparisons between the groups defined by our main factors, providing a more nuanced understanding of how these factors might interact to influence State Anxiety scores. By looking at these interactions more closely, we found detailed patterns and effects that might not have been apparent from a simple Main Effect Analysis.

Next, we conducted a Correlation Analysis to examine the relationship between certain hidden variables in our setup and State Anxiety scores. Specifically, we calculated Spearman's Correlation coefficient (ρ) to assess the association between the number of generated episodes and the average rating of saved episodes with the State Anxiety scores. We chose Spearman's Correlation because it does not require the assumption of Normality, unlike Pearson's correlation. Our decision to avoid Pearson's correlation was based on the results of the Shapiro-Wilk test, which indicated that our data does not meet the Normality assumption. The results from the Correlation Analysis helped us determine whether these variables have a statistically significant effect, guiding us on whether to include them in our statistical model for further analyses.

4.4.4 Perceived Learning Questionnaire

We performed Thematic Analysis on the open-ended responses we obtained in the Perceived Learning Questionnaire, which is ideal for identifying recurring patterns and themes in qualitative data. Castleberry et al. describe Thematic Analysis as a prevalent method for analyzing open-ended data, often used in qualitative research [170]. We chose this approach to gain an understanding of participants' perceptions and experiences regarding FL learning with our system ELLA.

In theory, the Thematic Analysis process involves several steps:

1. Compiling: Organizing the gathered data into a format suitable for analysis.
2. Disassembling: Breaking down data into meaningful segments through coding.
3. Reassembling: Grouping codes into themes and identifying patterns.
4. Interpreting: Drawing analytical conclusions by understanding thematic patterns and relationships.
5. Concluding: Formulating conclusions based on the data to answer research questions.

In our study, we began by cleaning the data, removing punctuation and stop words. We manually identified themes in the responses for each question mentioned in 4.2.3. Subsequently, we identified common themes across all five questions to derive insights, leading to our conclusions. The results of the Thematic Analysis are detailed in 4.5.5.

4.5 Results

4.5.1 FLSAS Questionnaire

Looking at the obtained scores for the FLSAS questionnaire, we observed that the mean anxiety score was **46.82**, indicating that, on average, participants had a moderate anxiety score. The standard deviation was **8.06** and the variance was **64.97**, suggesting that the Speaking Anxiety scores are fairly spread out around the mean. The FLSAS anxiety scores obtained ranged from a minimum of **34** to a maximum of **64**. This indicates a relatively broad spectrum of FL Speaking Anxiety scores among the participants, with some experiencing low Speaking Anxiety and others experiencing higher Speaking Anxiety.

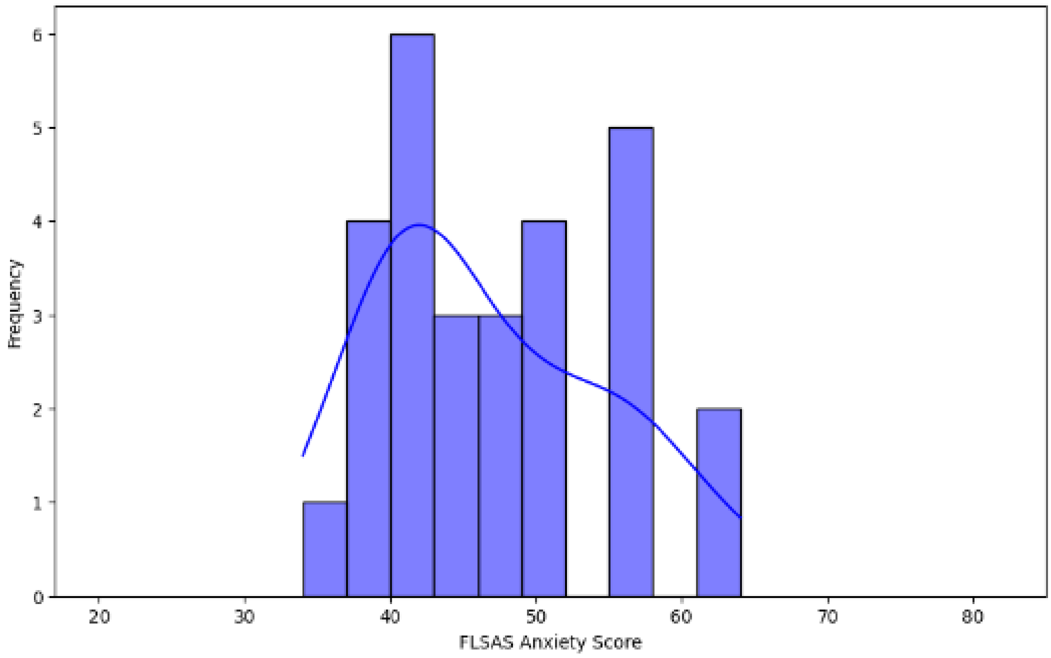

FIGURE 4.1: FLSAS Anxiety Scores

A histogram with an overlaid density curve was used to visualize the distribution of the anxiety scores obtained from the FLSAS Questionnaire. The histogram in 4.1 showed that the scores are roughly normally distributed, with a slight skew towards higher anxiety levels. The peak of the distribution was around the mean score of **46.82**, indicating that most participants had FL Speaking Anxiety scores in this range.

The distribution's shape suggests that while the majority of participants experienced moderate levels of FL Speaking Anxiety, there were a few outliers with significantly higher FL Speaking Anxiety scores. This slight skewness indicates that a subset of participants had notably higher anxiety levels compared to the rest.

From analyzing the FL Speaking Anxiety scores obtained from FLSAS and their corresponding distribution, we came to an understanding that our participants showed moderate levels of FL Speaking Anxiety for the Dutch language.

4.5.2 State Anxiety Scores

Condition	Mean	Variance
1	44.28	7.23
2	47	17.67
3	45.14	9.47
4	43.85	5.80

TABLE 4.2: State Anxiety Scores per Condition

Overall, the State Anxiety scores of the participants among the 4 conditions seemed to follow a similar trend to the scores of the FLSAS Questionnaire in 4.5.1. Looking at the mean and variance of the State Anxiety scores for each condition, we saw that Conditions 1 and 3 had similar mean anxiety scores with moderate variance, suggesting moderate and relatively consistent anxiety levels among participants in these conditions.

Condition 2 stood out due to its highest mean State Anxiety score (**47**) and greatest variance (**17.67**), suggesting that while participants generally experienced higher State Anxiety, their individual responses varied significantly.

Condition 4 was the most notable for its lowest mean State Anxiety score (**43.85**) and variance (**5.80**), indicating that participants in this condition tended to have lower and more consistent State Anxiety levels.

Initially, when we crafted our system, it was hypothesised (refer 2.4) that due to the presence of both EM and the deletion feature, learners would have a sense of control over their learning, resulting in the lowest State Anxiety levels among participants. However, the results indicated otherwise, since Condition 4, which did not have both EM Usage and the Deletion feature, exhibited the lowest State Anxiety scores and variance. We speculate the reason behind this result in 5.

4.5.3 Analysis Using ART ANOVA

Factors	F-value	p-value
EM Usage	1.20	0.28
Deletion	0.59	0.44
EM Usage:Deletion	2.92	0.09

TABLE 4.3: ART ANOVA results

From 4.3, we can see that neither of the factors (EM usage or Deletion) has a statistically significant main effect on the State Anxiety scores since their p-values are greater than the significance threshold of 0.1.

However, when examining the interaction effect between EM Usage and Deletion on State Anxiety scores, the results reveal a p-value of 0.09, which fell under our significance threshold of 0.1. This suggested that the interaction effect between EM Usage and Deletion was statistically significant, and was worth exploring further.

4.5.4 Post-Hoc Analysis

Given the results of ART ANOVA, we believed that the potential interaction effect warranted deeper investigation. Contrast Analysis on our first main factor EM Usage yielded an estimated difference of -3.5 between the conditions without EM usage and the conditions with EM usage. This suggested that the mean State Anxiety scores for the conditions without EM usage are 3.5 points lower than the conditions with EM usage. However, the p-value obtained was 0.283, which was greater than the significance threshold of 0.1.

Contrast analysis on our second main factor Deletion yielded an estimated difference of 2.5 between the conditions without Deletion and the conditions with Deletion. This suggested that the mean State Anxiety scores for the conditions with Deletion are 3.5 points lower than the conditions without the Deletion feature. However, the p-value obtained was 0.44, which was greater than the significance threshold of 0.1.

Performing Contrast Analysis on the interaction between our independent variables, we obtained the following results:

From 4.4, we observed that Condition 1 (EM Usage and Deletion) had a lower mean State Anxiety score than both Conditions 2 (EM usage without Deletion) and 3

Contrast between Conditions	Estimated Difference	p-value
1 - 2	-5.12	0.24
1 - 3	-1.14	0.79
1 - 4	0.92	0.83
2 - 3	4.07	0.36
2 - 4	6.14	0.17
3 - 4	2.07	0.64

TABLE 4.4: Contrast Scores of the Different Layers of Interaction Between Each Condition

(Deletion without EM Usage). However, Condition 1 had a higher mean State Anxiety score than Condition 4 (without both EM Usage and Deletion). From the results obtained, we inferred that Condition 1 performed relatively better when compared to the other conditions to obtain a lower State Anxiety score among its participants. However, all of the results in the Contrast Analysis of the interaction effects resulted in a p-value greater than the significance threshold of 0.1.

This meant that there was not sufficient statistical evidence for us to prove that both the usage of EM and the presence of the deletion feature resulted in lower State Anxiety scores. However, we wanted to determine if any other potential hidden variables influenced the State Anxiety scores.

Generated Episodes

For 28 participants, our system generated a total of 295 episodes, out of which 259 were saved and 36 were deleted. The number of episodes generated per condition is seen in 4.5.

Condition	Generated Episodes
1	78
2	70
3	67
4	80

TABLE 4.5: Number of Generated Episodes per Condition

Condition	Spearman's ρ	p-value
1	-0.15	0.74
2	-0.36	0.43
3	0.56	0.18
4	-0.25	0.58

TABLE 4.6: Spearman's ρ Between the Number of Generated Episodes and State Anxiety Scores per Condition

From 4.6, we observed that Spearman's correlation coefficient revealed small negative correlations between the number of generated episodes and State Anxiety scores across most conditions, with a notably moderate positive correlation in Condition 3. However, none of the relationships were statistically significant, since the p-values were greater than the significance threshold of 0.1.

Saved Episodes

Condition	Success	Mistake
1	18	41
2	23	52
3	18	29
4	30	48

TABLE 4.7: Number of Saved Mistake and Success Episodes per Condition

Attempting to understand the saved episodes in 4.7, we observed that most of them were categorized as mistake episodes. The definitions of mistake episodes and success episodes can be found in 3.1.1. Focusing on the specific types of mistake episodes, we observed the following patterns across our four conditions:

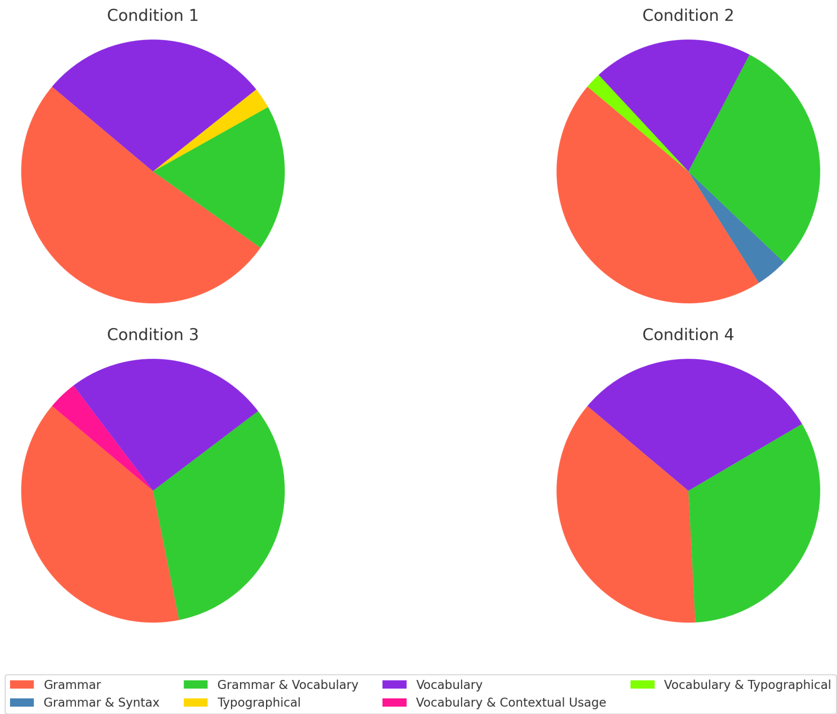

FIGURE 4.2: Types of Mistakes per Condition

From 4.2, we observed that across all four conditions, grammar mistakes (orange segment) were the most frequent type of error made by participants, followed closely by vocabulary mistakes (purple segment). This pattern highlighted a common difficulty with grammar among participants in all conditions. The consistent occurrence of combined grammar and vocabulary mistakes (green segment) indicated that these areas are particularly problematic when they occur together.

In Condition 2, there was a distinct presence of vocabulary and typographical mistakes (neon green segment), along with grammar and syntax mistakes (blue segment), which are not seen in the other conditions. This suggested a unique challenge faced by participants in Condition 2.

Similarly, Conditions 1 and 3 showed small proportions of typographical mistakes (yellow segment) and vocabulary and contextual usage mistakes(pink segment) respectively, which were absent in the other conditions. This could indicate specific difficulties encountered by participants in these conditions.

Condition	Grammar	Grammar & Syntax	Grammar & Vocabulary	Typographical	Vocabulary	Vocabulary & Contextual Usage	Vocabulary & Typographical
1	51.28	0.00	17.94	2.56	28.20	0.00	0.00
2	45.09	3.92	29.41	0.00	19.60	0.00	1.96
3	39.28	0.00	32.14	0.00	25.00	3.57	0.00
4	36.95	0.00	32.60	0.00	30.43	0.00	0.00

TABLE 4.8: Percentage of Mistake Types per Condition

The results in 4.8 were representative of our participant population, as approximately 57% were of beginner Dutch proficiency. Beginners are known to make a substantial number of basic errors as they are just starting to learn the language, while intermediate learners (about 25% of our participants) are also known to make the same errors but in a less frequent manner[171].

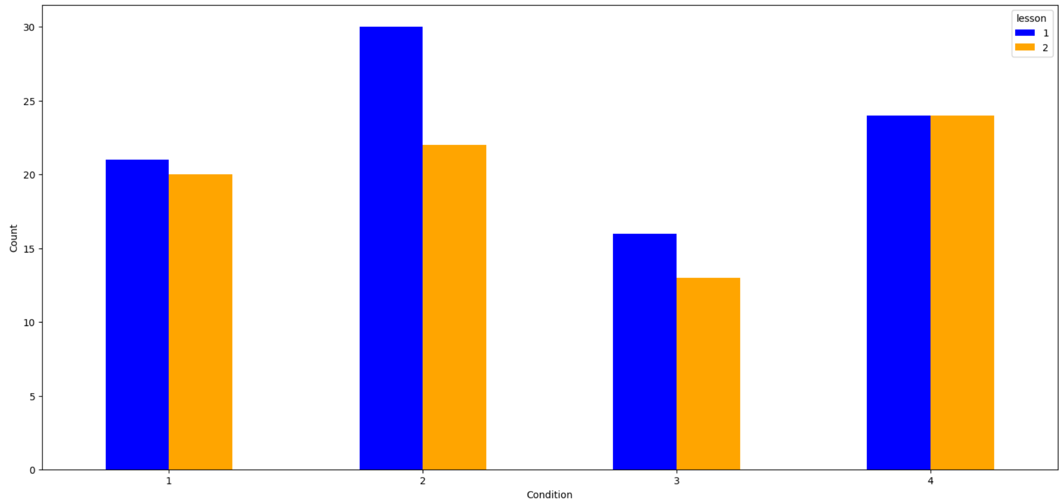

FIGURE 4.3: Mistake Episodes in Lessons 1 and 2 per Condition

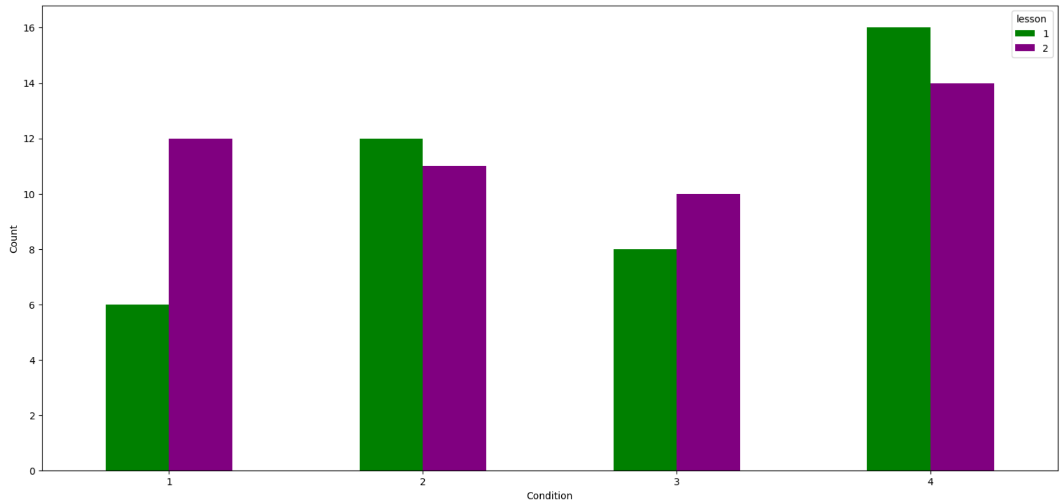

FIGURE 4.4: Success Episodes in Lessons 1 and 2 per Condition

From Figure 4.3, it was evident that participants consistently made fewer mistakes in Lesson 2 compared to Lesson 1. This suggested that as participants engaged with the system, they adapted and corrected the errors made in Lesson 1. Similarly, Figure 4.4 showed an increase in successful episodes in Lesson 2 compared to Lesson 1, indicating a positive learning direction where learners make fewer mistakes as they progress in FL learning with the system.

Examining the specific conditions, Condition 2 performed the worst, since it generated the highest number of mistakes. Contrary to our initial belief, Condition 1 was not the best-performing condition in terms of minimizing mistakes; instead, Condition 3 was the most successful. This observation was supported by our Contrast Analysis on the main factors 4.5.4, which suggested that conditions without EM usage and with Deletion are expected to achieve lower State Anxiety scores.

Average Rating of Saved Episodes

Condition	Average Rating of Episodes
1	3.74
2	3.36
3	3.38
4	3.28

TABLE 4.9: Average Rating of Saved Episodes per Condition

From 4.9, we saw that participants in Condition 1 (with EM Usage and Deletion) gave on average a higher helpfulness rating out of 5 to the saved system-generated episodes, while participants from the other 3 conditions gave a relatively similar low rating to the saved system-generated episodes. This meant that participants from Condition 1 thought that the system-generated episodes were more likely to help their FL acquisition, possibly explaining their lower State Anxiety scores.

After performing Spearman's correlation between the average rating of saved episodes and State Anxiety scores, we observed the following:

Condition	Spearman's ρ	p-value
1	0.49	0.26
2	-0.32	0.50
3	0.45	0.30
4	0.36	0.44

TABLE 4.10: Spearman's ρ Between Average Rating of Saved Episodes and State Anxiety Scores per Condition

From 4.10, we observed that Spearman's correlation analysis revealed varying relationships between average rating and State Anxiety scores across different conditions of EM usage and Deletion. Specifically, moderate positive correlations were observed in most conditions, with a notably smaller negative correlation in Condition 2. However, none of the relationships were statistically significant, since the p-values were greater than the significance threshold of 0.1.

Deleted Episodes

We offered the Deletion feature only to Conditions 1 and 3 as explained in 4.3. Out of the 295 episodes generated, only 36 episodes were deleted, or 12.95%. All deleted episodes were not rated, which meant that they had a default rating of 1. The type of episodes seen in deleted episodes resembled the ones described earlier under saved episodes. However, our speculations as to why the episodes were deleted, are mentioned in 3.3.4.

4.5.5 Percieved Learning Questionnaire

Usability and UI

Many participants praised our system ELLA's usability and UI. The adaptability and EM of the system were particularly appreciated, with participants noting that it enhanced their learning experience. Responses such as "I find the adaptability and memory of the system beneficial to the learning process", "It made the learning experience a lot better. The interactive methodology adds a natural element to conversing in Dutch." and "The overall experience of the system was quite enjoyable" reflect this sentiment. However, some participants did mention criticisms regarding the interactions, such as the robotic speech of furhat and the latency in responses.

Content Relevance and Learning Outcomes

Participants appreciated the system's ability to provide contextually appropriate and beneficial content. This sentiment is reflected in statements such as "It generates content that I would have never read particularly about," "I believe it adapted the complexity of its responses to my proficiency level," and "It matches my level and my ability, and attempts to push me forward." Additionally, the system's capacity to offer sound advice and real-time corrections, aligned with the participants' learning goals, was recognized as a major advantage. This is evidenced by responses like "It matches my learning goals and my level.". On the other hand, some participants felt the content was either too advanced or not tailored to their proficiency level, suggesting a need for better scaffolding to align with individual learning goals.

Feedback and Corrections

Feedback on grammatical errors and pronunciation was a prevalent theme among the responses. Participants valued the system's ability to provide immediate and detailed corrections, helping them understand and rectify their mistakes. Statements like "The general differentiation between using certain phrases as formal/informal was amazing.", "It corrected the small things, that the local dutch speakers wont correct me for" and "It showed me the difference between geen and niet. That helped." exemplify this appreciation. Some participants, however, felt the feedback could be overly judgemental, affecting their confidence negatively.

Safe Learning Environment

Participants felt more secure practising Dutch without the fear of judgement, which encouraged them to engage more freely. Statements like "I feel like my speaking level is not being judged while practising." and "Practising in an environment where mistakes are not negatively perceived by other humans is a very good opportunity for learners." reflect this sentiment.

Building Confidence

While responses varied, many participants indicated that our system positively impacted their confidence in using the language. The interactive nature and repeated practice were seen as beneficial for building confidence over time. Statements such as, "I feel more comfortable speaking to it and communicating with it, which helps me more confidently practice my Dutch.", "It would help a lot, especially with more time." and "I feel confident practicing Dutch with the system" reflect this sentiment. Conversely, some participants felt that the system's strict feedback on pronunciation along with its inability to recognise foreign accents could be discouraging.

Practice and Repetition

Improvements in pronunciation and speaking practice were also prevalent themes in the responses. Participants appreciated our system's corrections and feedback on their pronunciation, which helped them speak more accurately and confidently over time. Comments such as "I liked that it corrected my pronunciation", "I can repeat the same sentence over and over again until I say it correct" and "Yes it will encourage to speak better" highlight this positive impact. However, some participants faced challenges with the system's speech recognition, which created inescapable loops of repeated fault identification and correction, particularly with non-native accents.

Criticisms

While the overall feelings of the participants with our system were positive, several areas for improvement were identified. Participants cited issues such as the system's difficulty in recognizing foreign accents, overly fast responses that sometimes made the content difficult to understand, the need for more contextual and detailed instructions, and the occasional mismatch between the content generated and their learning level. Some participants also had issues with guiding the GPT-4 LLM back to "language-learning" mode, citing its tendency to go on tangential rants rather than focus on teaching Dutch.

Conclusion of Thematic Analysis

In summary, the thematic analysis revealed that the participants' experiences with our system were largely positive, highlighting the usability, relevant content and helpful grammatical and pronunciation corrections. Many appreciated the safe space provided by our system to make mistakes and get detailed explanations of the corrections provided. However, some criticisms of the robotic responses and occasional speech recognition errors were noted.

In terms of confidence in using the language, most participants had a neutral response, with some expressing increased confidence due to the interactive and non-judgmental nature of the system. Participants mentioned that repeated interaction and feedback on pronunciation were particularly beneficial.

Despite some challenges, such as response latency and difficulties with speech recognition, the system's ability to adapt to the participants' proficiency levels and provide precise feedback was well-received. The detailed feedback and corrections offered by the system were seen as valuable for achieving learning goals. Overall, the system appears to be a promising tool for language learning, verifying our hypothesis in 2.4, with room for further enhancements based on feedback.

Chapter 5

Discussion

In this chapter, we discuss the answers to our research questions, along with the limitations and future research possibilities. Section 5.1 explains how our main research question and its corresponding sub-research questions have been addressed. Section 5.2 outlines the technical and experimental limitations we encountered during this study. Finally, section 5.3 offers insights into potential improvements for our system and suggests different areas of research for a more holistic evaluation of our system.

5.1 Answers to Research Questions

Through this thesis, we aimed to answer the main research question, **What is the impact of Episodic Memory of a Foreign Language teaching Conversational Agent on the Speaking Anxiety of Foreign Language learners, and what Learning Benefits does it offer?** by answering its following sub-research questions.

5.1.1 What are the current findings regarding Foreign Language Anxiety and its impact on learners' learning and speaking ability, as well as the role of Episodic Memory in Conversational Agents?

Through our literature research, we found that FL learning anxiety is a significant factor that negatively impacts learners' learning and speaking abilities. Learners frequently worry about making mistakes and being negatively judged by their peers and other interlocutors, which can lead to avoidance behaviours and reduce participation in speaking tasks. Consequently, anxious learners are harsher on their performance, resulting in fewer opportunities and motivation to improve thus slowing their overall progress in FL acquisition. Cognitive overload caused by anxious thoughts also means that FL learners have fewer mental resources available for processing and producing language effectively.

Research conducted in CALL has shown that effective CALL systems integrate interactive exercises, multimodality, and real-time feedback to create an immersive learning experience. Key elements required for an effective language-learning environment include personalization, engagement, and context-immersive practice. Personalization can be achieved through adapting learning algorithms and user profiles. Engagement is fostered through interactive elements that make learning motivating and enjoyable. Additionally, providing context-immersive scenarios can help learners apply their skills in real-life situations, enhancing retention and fluency.

The role of EM in CAs is particularly relevant since these agents can play a significant role in reducing FL anxiety by offering a safe space for learners to converse without the fear of making mistakes and being negatively judged for it. EM enables CAs to remember specific instances of past interactions, which can be recalled to personalize and contextualize future conversations. This personal touch can make learners feel understood, increasing their engagement and reducing their communication apprehension. Moreover, the ability to provide contextually relevant feedback based on past interactions helps create a more natural and coherent learning process.

Current EM architectures include frameworks for storing, retrieving, and using past interactions to maintain continuity in conversations. However, there is a severe lack of research in optimising these frameworks for educational purposes, particularly in developing methods for explicit editing, deletion and updation of stored memories. Based on this research gap, we were motivated to create a CA system with a deletable EM architecture to understand how EM can be leveraged to enhance language learning and reduce anxiety related to speaking a FL more effectively.

5.1.2 How can the design of an Episodic Memory framework be optimized to support learning for language-anxious learners?

We designed an EM framework with a specific focus on optimizing it to support learning. Our innovative EM framework incorporates an explicit deletion mechanism, aimed at providing a sense of control for language-anxious learners. The key component of this framework is a short-term episodic cache, which temporarily stores the system-generated episodes during the rating process mentioned in 3.3.4.

When a learner interacts with these episodes, they have the opportunity to indicate their preference for either retaining or deleting specific episodes. This interaction is crucial for the learner to exercise control over their learning materials.

After the learner indicates their preference towards deleting episodes, the episodes indicated for deletion get sent to the deletion log. Conversely, episodes that the learner chooses to retain are moved to long-term episodic memory storage, where they are stored and retrieved for future reference.

The explicit rating and deletion process serves multiple pedagogical purposes. By granting the learners the ability to manage what the system saves in its memory, we align with educational theories that advocate for learner control. Research suggests that such frameworks can go a long way towards mitigating FL anxiety by fostering a sense of agency and personalized learning.

5.1.3 How does the usage of episodes stored in Episodic Memory, and the option to delete them, influence Foreign Language learners' speaking-related State Anxiety?

The results of our experiments suggest a possible correlation between the use of EM, the presence of the deletion feature, and learners' State Anxiety when speaking Dutch with our system. Due to the limited number of participants, we used a less strict statistical threshold of 0.1. Our ART ANOVA test revealed that the interaction of the main factors EM Usage and Deletion significantly influences State Anxiety scores, as their p-value was less than 0.1.

To validate this result, we conducted several post-hoc analyses. Unfortunately, the contrast analysis between the main factors of EM usage and deletion, as well as their interaction layers, showed no statistically significant effects (p-value>0.1) on State Anxiety scores. Upon examining hidden variables in our setup, such as the number of generated episodes and the average rating of saved episodes, we found a moderate negative Spearman's Correlation and a moderate positive Spearman's Correlation, respectively, with State Anxiety scores. However, none of these correlations were statistically significant, as their p-values were greater than 0.1.

When comparing the number of mistakes and successes between Lesson 1 and Lesson 2 for each condition, we noticed a general decrease in the number of mistakes from Lesson 1 to Lesson 2, indicating improvement and learning effectiveness. These results align with our previous findings from Contrast Analysis, where conditions without EM and with Deletion performed the best in terms of making the least number of mistakes in Lessons 1 and 2 across experimental conditions.

5.1.4 What specific Learning Benefits do language-anxious users perceive from this system?

The Thematic Analysis of responses from our participants revealed several key benefits and challenges associated with the system. Looking at the benefits, a significant theme that emerged is the system's usability. Participants frequently highlighted the quick, natural interactions, and pleasant user interface, which contributed to a positive user experience. Another major benefit that was highlighted is content relevance. Participants found the content engaging and useful, which maintained their interest and motivation.

The system's grammatical corrections were also complimented for being instant and accurate. Participants felt that this immediate feedback on their grammar, along with detailed explanations on correct usage, helped them correct their mistakes in real-time, reinforcing correct language use and reducing the likelihood of ingraining errors.

Participants felt that the system was a safe space to make mistakes, which allowed them to practice without the fear of judgement, essential for effective FL learning. Participants reported increased levels of confidence through repetition, along with improvements in articulation and a broader vocabulary, indicating that our system effectively supports language development and contributes to overall FL mastery.

Several challenges were identified alongside the positive benefits. Participants reported issues with ASR, including repeated misinterpretations of input. Some participants felt excessively monitored and judged by the system and found it to be fast-paced with its corrections. Others felt that the system's content was either too advanced or lacked direction. This mismatch with learners' proficiency and goals can sometimes create additional stress and hinder progress. Additionally, technical challenges such as a lack of natural flow, delayed responses, and issues with the LLM backend were mentioned. All of these issues can detract from the learning experience, causing learners to feel frustrated, stressed, or disengaged.

5.2 Limitations

5.2.1 Technical Limitations

ASR and Transcription

This was possibly the most discussed limitation among the participants. Because of time and computational resource constraints, we used an existing ASR for Streamlit that could recognize Dutch in speech input and transcribe it to text. However, this ASR module was unable to recognize foreign accents when Dutch was spoken. This resulted in numerous transcription errors, such as generating a correction even though the person responded correctly, producing an English word even though the person clearly used Dutch, and creating contextually inaccurate transcriptions.

These inaccuracies often led to unintentional error correction loops during interactions, causing frustration among some participants.

Using LLM for Conversations

Some participants found that the GPT-4 LLM's responses during conversational lessons didn't match their proficiency level, making it difficult for them to keep up with the conversation. Additionally, the LLM did not consistently follow the prompts we provided, sometimes generating responses with emojis, which caused issues in our backend storage. One participant had difficulty guiding the LLM to "Language-Learning" mode, as it would start discussing unrelated topics like beer making instead of focusing on teaching Dutch through conversation.

While the LLM performed well in generating episodes from conversations, there were some inaccuracies. For example, it would sometimes label its own responses during conversations as mistake/success episodes, causing confusion among participants. Furthermore, there were differences in the number of episodes generated for identical conversations due to the randomness pre-programmed in the LLM.

Furhat

Some participants thought that Furhat was too creepy and its voice sounded too robotic. While this didn't affect their interaction with our system, they believed that a more human-like voice and a friendlier face would encourage them to converse more freely with our system. Due to a slight delay between the response generated by the GPT-4 LLM and Furhat's TTS, some participants found it challenging to move their faces between the monitor and Furhat to read and then listen to the text.

Reviewing Episode Summaries, Not Episodes

Some participants suggested including a collapsible window of the episodes for each summary shown during the review process. This feature would allow them to rate the episodes themselves rather than the summaries, which can be influenced by GPT-4's unpredictable nature. Additionally, one participant suggested keeping the review process at the end of the interaction after finishing the lessons and the reminiscing session. This would enable participants to view the conversational snippets that generated the corresponding episodes and summaries, providing them with a clearer basis for their ratings.

5.2.2 Experimental Limitations

Limited Participants

Due to time and resource constraints, we had to limit our number of participants to 28. This limitation had two main implications in our study.

First, the small number of participants significantly affected the statistical power of our analyses. With only 28 participants divided into 4 conditions, there was high variability in our data which skewed the results, making it difficult to identify statistically significant differences or correlations. This means that even though our results showed interesting trends or patterns, they were not statistically significant. As a result, the generalizability of our findings is limited, and caution must be exercised when interpreting our results.

Second, the constraint on our sample size also limited the diversity in our participant pool. A larger sample space would have allowed for a more diverse range of participants in terms of demographics such as age, gender, linguistic background and proficiency. This diversity is crucial if we want our results to apply to a broader population of language learners.

Short Interactions

Since our participants spent between 30 to 60 minutes interacting with the system, this short interaction time was not enough to affect their State Anxiety levels drastically. Therefore caution must be exercised while building on the results obtained from our study. This is also mentioned in the research conducted by Myers et al., which suggests that brief interventions, or short interactions with interventions, may not significantly reduce learning anxiety levels despite improving performance [172]. Furthermore, research conducted on psychodynamic counselling interventions for university students has demonstrated that more extended and interactive interventions are required to have a meaningful influence on anxiety, as brief consultations or interactions might not provide sufficient information required to address anxiety problems [173].

Different Questionnaires Used

Unfortunately, due to using two different anxiety questionnaires that have varying Likert scales and score ranges, we were unable to accurately measure the system's effectiveness in reducing FL Speaking Anxiety among participants. A comparison using the same standardized questionnaire before and after the intervention (pre and post-tests) would have provided a more reliable evaluation of the system's impact on FL anxiety levels. This inconsistency in measurement tools introduced potential variability, which undermines the reliability of our findings.

5.3 Future Improvements

On the technical side, the first improvement would be to implement a custom-trained or fine-tuned ASR model capable of recognizing Dutch spoken with foreign accents. Given that our target users are foreign university students learning Dutch, this enhancement would significantly boost usability and help reduce FL-speaking anxiety. Additionally, using a local LLM specifically fine-tuned for teaching Dutch would alleviate privacy concerns and maintain focus on "Language-Learning" through proper scaffolding techniques. Employing a custom fine-tuned LLM, such as Mistral, known for excelling in reasoning tasks, can minimize the generation of inaccurate episodes. Furthermore, replacing Furhat with a more humanoid robot featuring human-like voice modulation could strengthen the personal connection learners form with the system, thereby enhancing learning efficiency.

On the experimental side, the initial change would be to implement a standardized questionnaire for both pre-test and post-test assessments. This would provide a clearer measure of our system's effectiveness in reducing FL speaking anxiety. Furthermore, increasing the number of participants would help achieve statistically significant results. Extending the duration of interaction through more lessons, longer sessions, or multiple sessions over several weeks would also contribute to obtaining results based on statistical reliability rather than ambiguity. Finally, conducting future experiments to investigate how different LLMs affect FL anxiety would be an intriguing approach to evaluating the effectiveness of GPT-4 in our system.

Chapter 6

Conclusion

Our thesis explored the integration of conversational agents (CAs) and foreign language (FL) learning through ELLA, a Dutch learning assistant with deletable episodic memory. Learners interacted with ELLA via a Furhat robot, discussing topics like cities and food, and could review and delete conversation episodes. This aimed to empower learners, reducing anxiety associated with FL learning and speaking. At the end of interactions, learners could revisit previous lessons to clarify grammar or vocabulary. Through our experimental design, our thesis demonstrated the potential of ELLA to reduce state anxiety related to speaking Dutch and offered numerous learning benefits as perceived by our users.

However, this study is limited by numerous factors, including a small participant pool, brief system interactions, and varied questionnaires, which affected the evaluation of our system's impact on FL speaking anxiety. Our system also faced challenges with its automatic speech recognition (ASR) system, which struggled with foreign accents, and its language model, which sometimes produced unhelpful content. Despite these limitations, this thesis provides a solid foundation for future work in using editable and modifiable memory modules in CAs for FL learning.

For future improvements, we recommend employing a custom fine-tuned language model and a specialized ASR to better accommodate Dutch spoken with foreign accents. Additionally, replacing Furhat with a more human-like interlocutor could enhance user experience. Further research should examine the long-term impact on FL speaking anxiety using standardized tests and explore our system's effectiveness with various languages and more complex lesson content to assess its broader applicability.

Appendix A

UI of Web Application

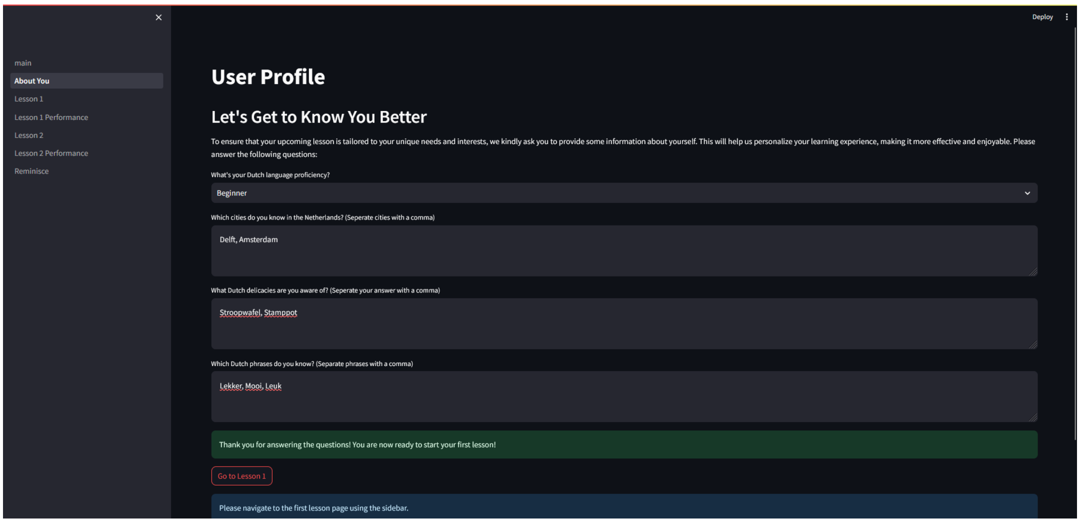

FIGURE A.1: About You

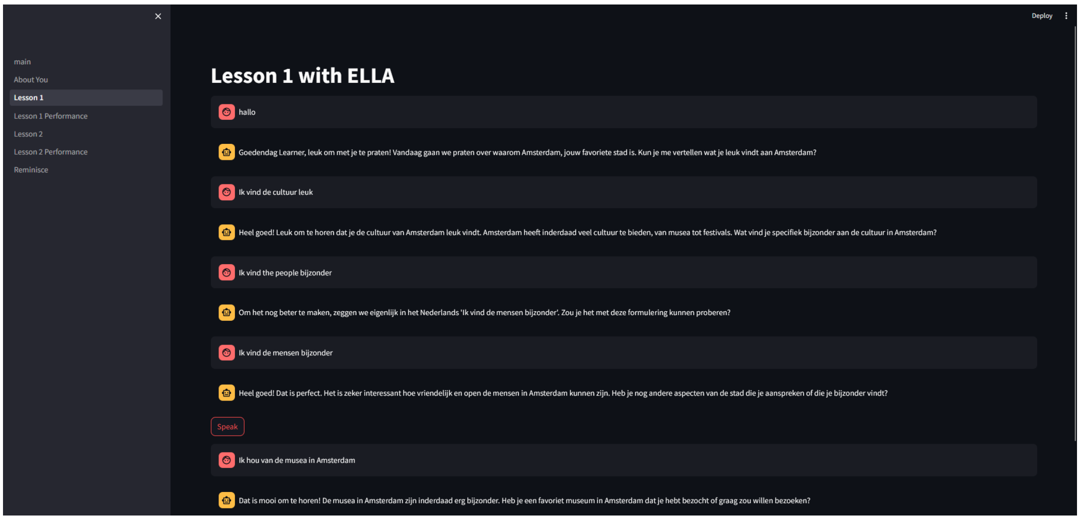

FIGURE A.2: Lesson 1

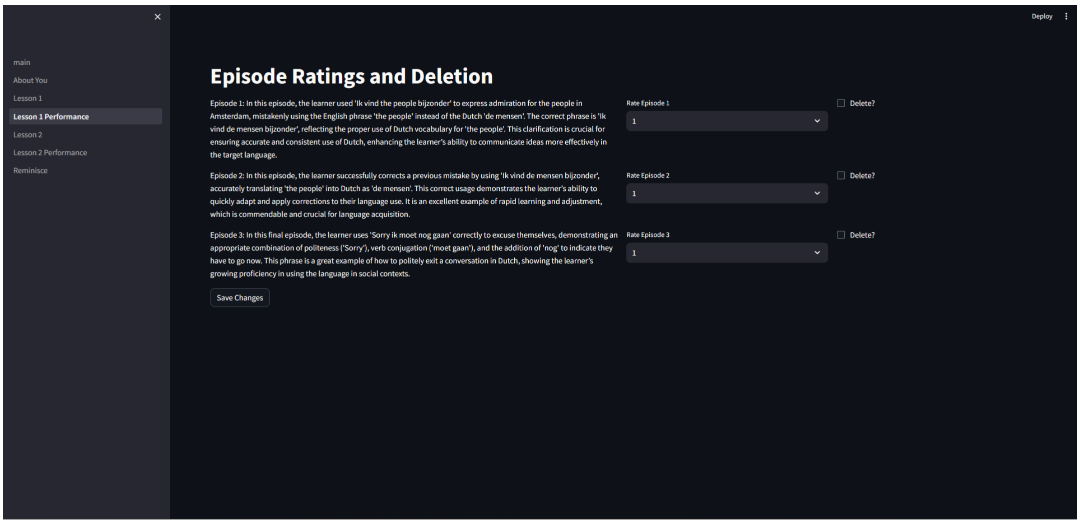

FIGURE A.3: Reviewing Episodes from Lesson 1

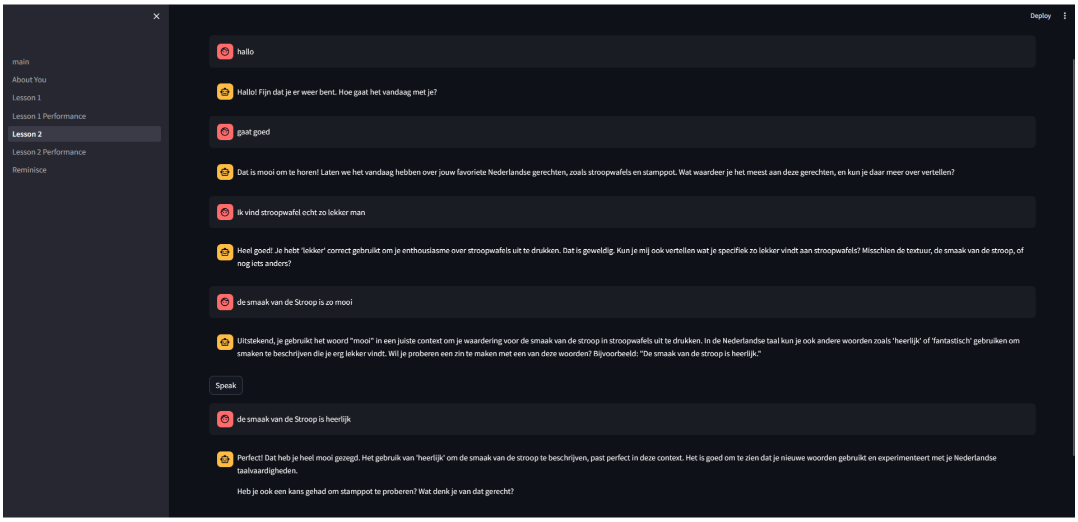

FIGURE A.4: Lesson 2

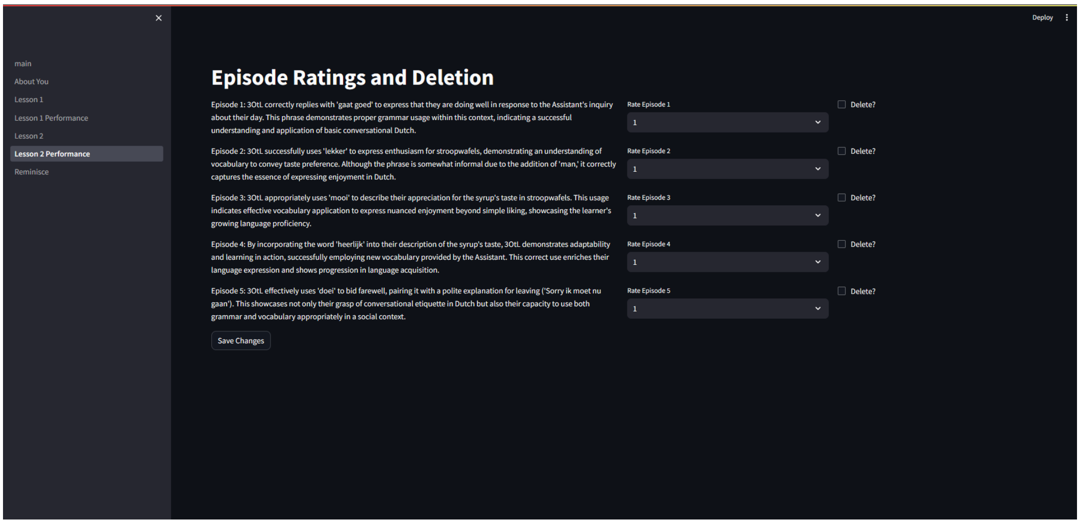

FIGURE A.5: Reviewing Episodes from Lesson 2

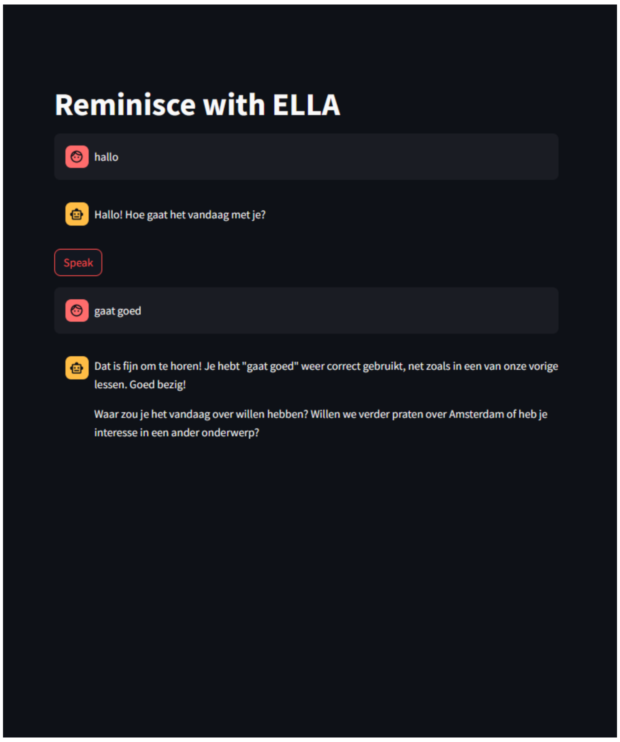

FIGURE A.6: Reminiscing Session

Appendix B

FLSAS Questionnaire

I never feel quite sure of myself when I am speaking Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I don't worry about making mistakes in Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I tremble when I know that I am going to have to speak Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

It frightens me when I don't understand what the other person is saying in Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

It wouldn't bother me at all to take more opportunities that allow me to speak Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

When I am speaking Dutch, I find myself thinking about things that have nothing to do with the conversation. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I keep thinking that other students are better at speaking Dutch than I am. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I start to panic when I have to speak in Dutch without preparation. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I worry about the consequences of my mistakes during the conversation. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I don't understand why some people get so upset over conversing in Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

When speaking Dutch, I can get so nervous that I forget things I know. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

It embarrasses me to volunteer to speak Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I would not be nervous speaking Dutch with native speakers. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I get upset when I don't understand what the other person is correcting. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Even if I am well prepared to speak Dutch, I feel anxious about it. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I often feel like escaping when I am required to speak Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I feel confident when I speak Dutch. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Appendix C

State Anxiety Questionnaire from STAI

I feel calm *

- ◯ Not at all
- ◯ Somewhat
- ◯ Moderately so
- ◯ Very much so

I feel secure *

- ◯ Not at all
- ◯ Somewhat
- ◯ Moderately so
- ◯ Very much so

I am tense *

- ◯ Not at all
- ◯ Somewhat
- ◯ Moderately so
- ◯ Very much so

I feel strained *

- Not at all
- Somewhat
- Moderately so
- Very much so

I feel at ease *

- Not at all
- Somewhat
- Moderately so
- Very much so

I feel upset *

- Not at all
- Somewhat
- Moderately so
- Very much so

I am presently worrying over possible mistakes *

◯ Not at all

◯ Somewhat

◯ Moderately so

◯ Very much so

I feel satisfied *

◯ Not at all

◯ Somewhat

◯ Moderately so

◯ Very much so

I feel frightened *

◯ Not at all

◯ Somewhat

◯ Moderately so

◯ Very much so

I feel comfortable *

- Not at all
- Somewhat
- Moderately so
- Very much so

I feel self-confident *

- Not at all
- Somewhat
- Moderately so
- Very much so

I feel nervous *

- Not at all
- Somewhat
- Moderately so
- Very much so

I am jittery *

- ◯ Not at all
- ◯ Somewhat
- ◯ Moderately so
- ◯ Very much so

I feel indecisive *

- ◯ Not at all
- ◯ Somewhat
- ◯ Moderately so
- ◯ Very much so

I am relaxed *

- ◯ Not at all
- ◯ Somewhat
- ◯ Moderately so
- ◯ Very much so

I feel content *

- Not at all
- Somewhat
- Moderately so
- Very much so

I am worried *

- Not at all
- Somewhat
- Moderately so
- Very much so

I feel confused *

- Not at all
- Somewhat
- Moderately so
- Very much so

I feel steady *

- Not at all
- Somewhat
- Moderately so
- Very much so

I feel pleasant *

- Not at all
- Somewhat
- Moderately so
- Very much so

Appendix D

Generated JSON file from a conversational lesson


```
{
   "username": "UWbG",
   "topic": "Discussing favorite city and activities in The
      Netherlands",
   "episodes": [
      {
         "episodeNo": 1,
         "dialogNo": 2,
         "episode_type": "Mistake",
         "mistake_phrase": "Den Haag en ik ben Happy leuk ik ben Elke
             keer daar geweest en",
         "mistake": "Improperly structured sentence.",
         "mistake_type": "grammar",
         "context": "Explaining what the learner likes about visiting
             Den Haag.",
         "correction": "Ik ben elke keer blij als ik in Den Haag ben
            geweest.",
         "summary": "The learner tried to express enjoyment in
            visiting Den Haag but struggled with sentence structure,
            saying 'Den Haag en ik ben Happy leuk ik ben Elke keer
            daar geweest en.' The correct formation is 'Ik ben elke
            keer blij als ik in Den Haag ben geweest,' streamlining
            the expression of happiness linked to their visits. This
            correction is important for clarity and adhering to
            proper Dutch sentence construction."
      },
      {
         "episodeNo": 2,
         "dialogNo": 4,
         "episode_type": "Success",
         "correct_phrase": "ik ben Elke keer blij als ik in Den Haag
            ben geweest",
         "context": "Repeating the corrected sentence structure
            provided by the Dutch local.",
         "correct_type": "grammar",
```

```
        "summary": "Learner successfully repeated the corrected
           sentence 'Ik ben elke keer blij als ik in Den Haag ben
           geweest,' demonstrating an ability to adopt corrections
           and improve sentence structure. This repetition
           underscores agile learning and correct application of
           grammatical rules in Dutch."
    },
    {
        "episodeNo": 3,
        "dialogNo": 5,
        "episode_type": "Mistake",
        "mistake_phrase": "Ik wil graag in de staat en soms loop ik
           naar de het strand en Scheveningen en als het goed is",
        "mistake": "Sentence structuring and article use error.",
        "mistake_type": "grammar & vocabulary",
        "context": "Describing activities in Den Haag, namely,
           walking in the city and going to the beach.",
        "correction": "Ik loop graag in de stad en soms ga ik naar
           het strand in Scheveningen, als het weer goed is.",
        "summary": "The learner intended to express enjoyment in
           walking in the city and visiting the beach, but misspoke,
            'Ik wil graag in de staat ... naar de het strand ... als
            het goed is.' The correction, 'Ik loop graag in de stad
           en soms ga ik naar het strand in Scheveningen, als het
           weer goed is,' rectifies the misuse of 'staat' for 'stad
           ', article redundancy in 'de het strand', and provides a
           clearer condition for beach visits. Such corrections are
           instrumental for accurate expression and comprehension in
            Dutch."
    },
    {
        "episodeNo": 4,
        "dialogNo": 7,
        "episode_type": "Mistake",
        "mistake_phrase": "Ik doe niks bijzonders zoals en",
        "mistake": "Incomplete thought and awkward phrase ending.",
        "mistake_type": "grammar",
        "context": "Attempting to explain usual activities done at
           the beach.",
        "correction": "Ik doe niets bijzonders, ik ontspan gewoon
           graag op het strand.",
        "summary": "When trying to convey a preference for simple
           relaxation at the beach, the learner mistakenly formed an
            incomplete thought, 'Ik doe niks bijzonders zoals en.'
           The improved sentence 'Ik doe niets bijzonders, ik
           ontspan gewoon graag op het strand' offers a complete and
            coherent expression of the learner's activities,
           highlighting the importance of clear and concise sentence
            structure in Dutch."
    }
]
```

```
}
```

Appendix E

Responses to Perceived Learning Questionnaire

Describe your overall experience using our system. What did you find most beneficial?

Looking at the answers to this question overall, we found out that most users expressed positive experiences, while a few expressed neutral reactions, indicating neither positive nor negative sentiments towards our system.

We can categorize individual responses to this question into different categories, and for each category we give some of the responses that the participants gave:

1. Usability: 'I find the adaptability and memory of the system beneficial to the learning process', 'The overall experience of the system was quite enjoyable.', 'Pleasant interface with moderate accuracy', 'it is quick which makes it feel more natural'
2. Content Relevance: 'It generates content that I would have never read particularly about it. So it is a nice way to get some reading or listening. ', 'the advice it generates is very sound.', 'I find the adaptability and memory of the system beneficial to the learning process'
3. Grammatical Corrections: 'What i liked most was how well the robot corrected the mistakes i made.', 'I can understand the feedback can be beneficial ', 'gave me the translation for the English words I used in the middle of the sentence', 'corrects the mistake instantly'
4. Safe space to make mistakes: 'how well the robot corrected the mistakes i made', 'the mistakes that I really made, the system caught it and suggested a correction that I really appreciate. Because these are the corrections that a local Dutch speaker wont help you with, they would just get along. But these small details help us improve dutch.', 'the system was beneficial when correcting my mistakes', 'Practicing in an environment where mistake are not negatively perceived by other humans is a very good opportunity for learners'
5. Detailed explanations: 'liked that it provided the corrected version of my answers and that i could see the explanation of my mistakes', 'The feedbacks are good', 'Attention to every aspect of speaking'
6. Pronounciation Corrections: 'I liked that it corrected my pronunciation', 'I can see the potential for usage in Dutch education for non native speakers', 'The feedback and suggestions how to phrase things', 'improve my speaking in repeating'

7. Criticisms: ' there are a few instances where it might not pick the exact wordings within the phrases', 'Sometimes the speech recognition wasnt working properly so even when I saying correctly, it wouldn't recognize my words properly and find mistakes. ', 'But not being able to recognizing my words correctly, made me upset.', ' The robot itself adds very little as most of the time I am focused on the monitor'

How has our system impacted your confidence in using the language you are learning?

Looking at the responses overall to this question, we found out that most users had a neutral response. However there are some responses that range from positive to overwhelmingly positive.

1. Interactivity:'It made the learning experience a lot better. The interactive methodology adds a natural element to conversing in Dutch.'
2. Safe space: 'Feels more secure'
3. Growing vocabulary: 'Sometimes it used difficult or unusual words and it influence my learning path and confidence '
4. Impact through pronunciation improvements: 'I had no problem in Speaking Dutch to the chat bot', 'Has contributed for online interaction', 'I can repeat the same sentence over and over again until I say it correct'
5. Repeated interaction: 'Not very much but I believe I can gain confidence after repeated interaction ', 'More confident as I realized I am pretty good at the language', 'In the long term it would enable me to have more confidence when speaking', ' using it on a longer term might help me improve my dutch and make me more confident since i have no conversation partner at the moment.'
6. Increased confidence: 'I feel like I dont have to perfectly know Dutch to make a conversation', 'I feel confident practicing Dutch with the system', 'I am more confident than before', 'It's inspiring', 'I feel more comfortable speaking to it and communicating with it, which helps me more confidently practice my Dutch'
7. Awareness of Mistakes: 'I feel like I am more aware of the minor mistake that I made. So it helped me'
8. Too judgemental: ' It made me worried about my accent and pronunciations', 'The robot is judging me too much'
9. No effect/negative effect: 'Not at all', 'I am already a native speaker, so it has not helped me a lot.', 'The voice is stern and it speaks too fast. I feel trouble catching up. Lowered my confidence a bit.', 'Did not really affect it, did not have any problems with it before.', 'It did not have a big impact at the moment. Probably because of a relatively short interaction', 'Not really', ' I noticed the incapability of the system in recognising my accent.'

"Can you share an instance where you felt our system significantly helped you improve a specific language skill?"

A lot of the responses to this question provided specific nuanced examples with less common themes. Most responses were positive about it, while some were neutral to slightly negative.

1. Grammar: 'The general differentiation between using certain phrases as formal/informal was amazing.', 'It told me how to use the words in the correct order and also gave the Dutch translation for words like research.', 'grammatical corections', 'The system fixed my usage of the word cuisine with keuken in the second lesson', 'Pointing out prepositions mistakes and suggesting superior phrasing seems to be helpful.', 'It showed me the difference between geen and niet. That helped.', 'To say welk eten instead of welke eten.', 'I made a sentence and it keeps correcting my mistake', 'Order of words, vocabulary', 'system corrected some of my grammatical errors that would have possibly gone undetected by an instructor in the moment', 'Sentence structure'

2. Pronounciation: 'Ja, ik moet beter articuleren en ik zei dorpse sfeer blijkbaar fout', 'it corrected the small things, that the local dutch speakers wont correct me for', 'helps me feel that there can be various ways of saying things and the system gives me the best possible one. And this is also how I learnt in my dutch class so I'll remember to practice Dutch this way from now on.'

3. Non-judgemental of speaking: 'I feel like my speaking level is not being judged while practicing'

4. Explanation of corrections: 'while reading the feedback, the system does not simply propose a correction but it explains why such correction is necessary and this makes it easier to remember in the future.'

5. Encouraging specificity: 'When I was prompted to use a broader range of vocabulary'

How well do you think the content provided by our system matches your learning goals and level?

A lot of the responses were overwhelmingly positive, while some were neutral.

1. Adaptation of responses: 'I believe it adapted the complexity of its responses to my proficiency level.', 'The level of dutch the system "spoke" to me matched my proficiency of the Dutch language.', ' it matches the my learning goals and my level', 'the system is flexible, being able to pick out mistakes at all levels of Speaking ability, making it worthwhile for learners at all levels', 'it matches my current level. there were a few words i did not know, but overall it matched my needs.', 'matches my level and my ability, and attempts to push me forward', '

2. Encouraging to speak more Dutch: 'It was a little more Dutch then expected but I also know very little so, it's fine as is.', 'Yes it will encourage to speak better', '

3. Too easy for native speakers: 'Too easy for me'

4. Good tool for speaking practice: 'I think it would be a wonderful tool to practice speaking', 'It would help a lot, especially with more time.'

5. Understandability of responses: 'Pretty well, I speak a little bit Dutch so I understood it for the most part'
6. Precision: 'the mistakes it pointed out in me were pretty precise.'
7. Criticisms: 'The level was ok, but the learning goal was not obvious to me ', 'the speech was a little fast.', 'It is a bit advance for my level', ' There were some words I did not understand', 'was a bit advanced for me. There were many words I did not know', 'might be better to provide different advise for people in different level of Dutch, so different sentences/words can be suggested', 'conversation it provide is a little bit too complicated ', 'content does not really have scaffolding in the material it gives to me for learning. Therefore, I had times that I could not understand lots of the terms in a sentence.'

What challenges did you face while using our system? How do you think these challenges could be addressed?

All of the responses provided by the participants provided us with valuable insights into how our system can be improved in the future. We attempted to categorize the responses to this question into the following categories:

1. Furhat took too long to respond: 'Furhat took so long to answer and it stretched out every topic', 'latency of responses is a little bit to long for fluent discussions', 'the system's speech was quite robotic and thus less easy to listen to for long sentences',
2. No conversation flow: ' I didn't know where the conversation was going'
3. Position of the screen and Furhat: 'Focusing on the robot rather than on a screen would make it a lot more natural', ' looking more into the screen and reading and not listening', 'read the text first and listen the response, so it kind of distract me from listening the response', '
4. Not being able to recognize foreign accents: 'Accents , wide and diverse datasets could be used', 'not being able to recognizing my words correctly was the biggest challenge and it can be improved by using more advance models', ' some phrases my pronunciation was not perfect, so the system misunderstood what i was saying and marking it as a mistake. that was a bit frustrating.', 'it should be less nitpicky', '
5. Being able to visualize conversation while rating episodes: 'When rating episodes I would like to see what I actually said and how it was corrected', 'number of interactions first to avoid giving feedback until the end of the conversation', 'When there is a pronunciation correction, I expect it provide the correct pronunciation again in episode ratings to help me practice and remember.
6. Long sentences: ' sentences said in Dutch by the robot are sometimes too long and difficult to understand for people who have just started off', '
7. Fast responses: 'Punctuation and it speaks a bit too fast and with a lot of complex words', 'was a little fast making it a little hard to understand longer sentences' ,'

8. Speech recognition error: 'it didn't understand me, and then asked me to repeat the same thing again and again even when I said right', 'for example it recognized some of the words I used as English (wat, what), (karamel, caramel)', 'had some faults, mainly at the start and end of sentences and when considering punctuation', 'sometimes derails conversation. It would be nice if the chat bot could maybe could check other interpretation of the the spoken language, to see if there might be a more sensible one.', '

9. More information required: 'more clarification or intro prior to starting the lessons/questionnaire', 'more instruction about how to use it', '

10. Not matching language proficiency: 'system is trying to teach me Dutch by talking to me in Dutch which is very challenging its more for Advanced Dutch speakers rather than entry-level speakers', 'system should be able to use a simpler Dutch for beginners like me', '

11. Error due to LLM backend: 'Its hard to navigate the assistant back to language learning mode.', 'Its slow and I encountered bugs during the experiment', 'I did not understand some words'

Bibliography

[1] E. Commision, "Foreign language learning increases among eu students," 2023. [Online]. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20230926-1.

[2] S. E. Pfenninger and S. Polz, "Foreign language learning in the third age: A pilot feasibility study on cognitive, socio-affective and linguistic drivers and benefits in relation to previous bilingualism of the learner," *Journal of the European Second Language Association*, vol. 2, no. 1, pp. 1–13, 2018.

[3] B. Klimova and M. Pikhart, "Current research on the impact of foreign language learning among healthy seniors on their cognitive functions from a positive psychology perspective—a systematic review," *Frontiers in psychology*, vol. 11, p. 522 211, 2020.

[4] E. P. Lazear, "Culture and language," *Journal of political Economy*, vol. 107, no. S6, S95–S126, 1999.

[5] F. Rubio, *Self-esteem and foreign language learning*. Cambridge Scholars Publishing, 2021.

[6] E. K. Horwitz, M. B. Horwitz, and J. Cope, "Foreign language classroom anxiety," *The Modern language journal*, vol. 70, no. 2, pp. 125–132, 1986.

[7] E. K. Horwitz, "Foreign and second language anxiety," *Language teaching*, vol. 43, no. 2, pp. 154–167, 2010.

[8] T. Scovel, "The effect of affect on foreign language learning: A review of the anxiety research," *Language learning*, vol. 28, no. 1, pp. 129–142, 1978.

[9] K. Chastain, "Affective and ability factors in second-language acquisition," *Language learning*, vol. 25, no. 1, pp. 153–161, 1975.

[10] G. R. Tucker, E. Hamayan, and F. H. Genesee, "Affective, cognitive and social factors in second-language acquisition," *Canadian modern language review*, vol. 32, no. 3, pp. 214–226, 1976.

[11] T.-Y. Chen and G. B. Chang, "The relationship between foreign language anxiety and learning difficulties," *Foreign language annals*, vol. 37, no. 2, pp. 279–289, 2004.

[12] S. Krashen, "Principles and practice in second language acquisition," 1982.

[13] B. Weiner, "Attribution theory, achievement motivation, and the educational process," *Review of educational research*, vol. 42, no. 2, pp. 203–215, 1972.

[14] P.-H. P. Hsieh and D. L. Schallert, "Implications from self-efficacy and attribution theories for an understanding of undergraduates' motivation in a foreign language course," *Contemporary Educational Psychology*, vol. 33, no. 4, pp. 513–532, 2008.

[15] M. Peacock, "Attribution and learning english as a foreign language," *ELT journal*, vol. 64, no. 2, pp. 184–193, 2010.

[16] P. Gobel and S. Mori, "Success and failure in the efl classroom: Exploring students' attributional beliefs in language learning," *EUROSLA yearbook*, vol. 7, no. 1, pp. 149–169, 2007.

[17] R. M. Wenzlaff and D. M. Wegner, "Thought suppression," *Annual review of psychology*, vol. 51, no. 1, pp. 59–91, 2000.

[18] D. M. Wegner, D. J. Schneider, S. R. Carter, and T. L. White, "Paradoxical effects of thought suppression.," *Journal of personality and social psychology*, vol. 53, no. 1, p. 5, 1987.

[19] P. D. MacIntyre and R. C. Gardner, "The subtle effects of language anxiety on cognitive processing in the second language," *Language learning*, vol. 44, no. 2, pp. 283–305, 1994.

[20] J. S. Abramowitz, D. F. Tolin, and G. P. Street, "Paradoxical effects of thought suppression: A meta-analysis of controlled studies," *Clinical psychology review*, vol. 21, no. 5, pp. 683–703, 2001.

[21] J. A. Linck, P. Osthus, J. T. Koeth, and M. F. Bunting, "Working memory and second language comprehension and production: A meta-analysis," *Psychonomic bulletin & review*, vol. 21, pp. 861–883, 2014.

[22] T. Gregersen and E. K. Horwitz, "Language learning and perfectionism: Anxious and non-anxious language learners' reactions to their own oral performance," *The Modern Language Journal*, vol. 86, no. 4, pp. 562–570, 2002.

[23] A. C. Strain and S. K. D'Mello, "Affect regulation during learning: The enhancing effect of cognitive reappraisal," *Applied Cognitive Psychology*, vol. 29, no. 1, pp. 1–19, 2015.

[24] N. Fallah, F. Abdolahzadeh, and M. Yaaghobi, "Mindfulness and anxiety among foreign language teachers: The role of cognitive reappraisal and self-efficacy," *Mindfulness*, vol. 14, no. 12, pp. 3020–3032, 2023.

[25] C. A. Spann, V. J. Shute, S. Rahimi, and S. K. D'Mello, "The productive role of cognitive reappraisal in regulating affect during game-based learning," *Computers in Human Behavior*, vol. 100, pp. 358–369, 2019.

[26] Y. Zhou and M. Wei, "Strategies in technology-enhanced language learning," *Studies in Second Language Learning and Teaching*, vol. 8, no. 2, pp. 471–495, 2018.

[27] D. Tafazoli, M. E. Gomez Parra, and C. A. Huertas-Abril, *Cross-cultural perspectives on technology-enhanced language learning*. IGI Global, 2018.

[28] R. Shadiev and M. Yang, "Review of studies on technology-enhanced language learning and teaching," *Sustainability*, vol. 12, no. 2, p. 524, 2020.

[29] K. Beatty, *Teaching & researching: Computer-assisted language learning*. Routledge, 2013.

[30] M. Levy, *Computer-assisted language learning: Context and conceptualization*. Oxford University Press, 1997.

[31] P. Hubbard and M. Levy, "The scope of call education," *Teacher education in CALL*, vol. 14, pp. 3–20, 2006.

[32] D. Alpert and D. L. Bitzer, "Advances in computer-based education: The plato program will provide a major test of the educational and economic feasibility of this medium.," *Science*, vol. 167, no. 3925, pp. 1582–1590, 1970.

[33] M. Warschauer and D. Healey, "Computers and language learning: An overview," *Language teaching*, vol. 31, no. 2, pp. 57–71, 1998.

[34] S. Jones, *The> friendly orange glow: The untold story of the plato system and the dawn of cyberculture*, 2018.

[35] H. S. Nwana, "Intelligent tutoring systems: An overview," *Artificial Intelligence Review*, vol. 4, no. 4, pp. 251–277, 1990.

[36] P. D. Reddy and M Sasikumar, "Student model for an intelligent language tutoring system," in *2014 IEEE 14th International Conference on Advanced Learning Technologies*, IEEE, 2014, pp. 441–443.

[37] B. S. Klinton Bicknell Claire Brust, "How duolingo's ai learns what you need to learn," *IEEE spectrum*, 2023. [Online]. Available: https://spectrum.ieee.org/duolingo.

[38] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "Autotutor: An intelligent tutoring system with mixed-initiative dialogue," *IEEE Transactions on Education*, vol. 48, no. 4, pp. 612–618, 2005.

[39] E. Novak, K. McDaniel, and J. Li, "Factors that impact student frustration in digital learning environments," *Computers and Education Open*, vol. 5, p. 100 153, 2023.

[40] E. M. Golonka, A. R. Bowles, V. M. Frank, D. L. Richardson, and S. Freynik, "Technologies for foreign language learning: A review of technology types and their effectiveness," *Computer assisted language learning*, vol. 27, no. 1, pp. 70–105, 2014.

[41] M. Van Koert, N. Leona, J. Rispens, *et al.*, "The role of memory in the acquisition of vocabulary and grammar in the first language and in english as a foreign," *Second Language Learning Before Adulthood: Individual Differences in Children and Adolescents*, vol. 65, p. 31, 2022.

[42] K. Morgan-Short, M. Faretta-Stutenberg, K. A. Brill-Schuetz, H. Carpenter, and P. C. Wong, "Declarative and procedural memory as individual differences in second language acquisition," *Bilingualism: Language and cognition*, vol. 17, no. 1, pp. 56–72, 2014.

[43] P. Hamrick, C. Graff, and B. Finch, "Contributions of episodic memory to novel word learning," *The Mental Lexicon*, vol. 14, no. 3, pp. 381–398, 2019.

[44] Y. Zhang, M. Ridchenko, A. Hayashi, and P. Hamrick, "Episodic memory contributions to second language lexical development persist at higher proficiencies," *Applied Cognitive Psychology*, vol. 35, no. 5, pp. 1356–1361, 2021.

[45] J. Hammond, *Scaffolding: Teaching and learning in language and literacy education.* ERIC, 2001.

[46] A. Walqui, "Scaffolding instruction for english language learners: A conceptual framework," *International journal of bilingual education and bilingualism*, vol. 9, no. 2, pp. 159–180, 2006.

[47] A. Abdurrahman, N. Nurulsari, H. Maulina, B. Rahman, R. Umam, and K. Jermsittiparsert, "Multi-level scaffolding: A novel approach of physics teacher development program for promoting content knowledge mastery," *International Journal of Innovation, Creativity and Change*, vol. 7, no. 8, 2019.

[48] Z. Liu, J. Hua, and Z. Zhang, "Scaffolding instruction in virtual language learning," *Journal of Language Teaching and Research*, vol. 13, no. 2, pp. 386–391, 2022.

[49] T. Fani and F. Ghaemi, "Implications of vygotsky's zone of proximal development (zpd) in teacher education: Zptd and self-scaffolding," *Procedia-Social and Behavioral Sciences*, vol. 29, pp. 1549–1554, 2011.

[50] M. Haider and A. Yasmin, "Significance of scaffolding and peer tutoring in the light of vygotsky's theory of zone of proximal development," *International Journal of Languages, Literature and Linguistics*, vol. 1, no. 3, p. 2015, 2015.

[51] R. Hasan and N. A. Ahmad, "Conceptual framework of scaffolding literacy module to help remedial students mastering reading skills," *International Journal of Academic Research in Business and Social Sciences*, vol. 8, no. 11, pp. 1031–1038, 2018.

[52] S. Madya, F. Hamied, W. A. Renandya, C. Coombe, and Y. Basthomi, "Improving learners' vocabulary mastery through the use of scaffolding strategies while storytelling in an efl multiethnic classroom," in *ELT in Asia in the Digital Era: Global Citizenship and Identity*, Routledge, 2018, pp. 151–158.

[53] R. J. Alexander, "Towards dialogic teaching: Rethinking classroom talk," 2008.

[54] J. M. Sinclair and M. Coulthard, "Towards an analysis of discourse: The english used by teachers and pupils," *(No Title)*, 1975.

[55] H. Muhonen, H. Rasku-Puttonen, E. Pakarinen, A.-M. Poikkeus, and M.-K. Lerkkanen, "Scaffolding through dialogic teaching in early school classrooms," *Teaching and teacher education*, vol. 55, pp. 143–154, 2016.

[56] A. B. Ateş-Şen and A. C. Küntay, "Children's sensitivity to caregiver cues and the role of adult feedback in the developmentof referential communication," in *The acquisition of reference*, John Benjamins, 2015, pp. 241–262.

[57] M. L. Epstein, A. D. Lazarus, T. B. Calvano, *et al.*, "Immediate feedback assessment technique promotes learning and corrects inaccurate first responses," *The Psychological Record*, vol. 52, pp. 187–201, 2002.

[58] J. Metcalfe, N. Kornell, and B. Finn, "Delayed versus immediate feedback in children's and adults' vocabulary learning," *Memory & cognition*, vol. 37, no. 8, pp. 1077–1087, 2009.

[59] H. Nicholas, P. M. Lightbown, and N. Spada, "Recasts as feedback to language learners," *Language learning*, vol. 51, no. 4, pp. 719–758, 2001.

[60] I. Panova and R. Lyster, "Patterns of corrective feedback and uptake in an adult esl classroom," *Tesol Quarterly*, vol. 36, no. 4, pp. 573–595, 2002.

[61] R. Ellis and Y. Sheen, "Reexamining the role of recasts in second language acquisition," *Studies in second language acquisition*, vol. 28, no. 4, pp. 575–600, 2006.

[62] S. Li, "The effectiveness of corrective feedback in sla: A meta-analysis," *Language learning*, vol. 60, no. 2, pp. 309–365, 2010.

[63] S. E. Mol, A. G. Bus, and M. T. De Jong, "Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language," *Review of Educational Research*, vol. 79, no. 2, pp. 979–1007, 2009.

[64] F. Manzoor, M. Ahmed, and B. R. Gill, "Use of motivational expressions as positive reinforcement in learning english at primary level in rural areas of pakistan," *British Journal of English Linguistics*, vol. 2, no. 3, pp. 30–42, 2014.

[65] L. Thohir, "Motivation in a foreign language teaching and learning," *Vision: Journal for language and foreign language learning*, vol. 6, no. 1, pp. 20–29, 2017.

[66] C. J. Brainerd, V. F. Reyna, M. L. Howe, J Kingma, and R. E. Guttentag, "The development of forgetting and reminiscence," *Monographs of the Society for Research in Child Development*, pp. i–109, 1990.

[67] D. P. McCabe and N. C. Soderstrom, "Recollection-based prospective metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (jorks).," *Journal of Experimental Psychology: General*, vol. 140, no. 4, p. 605, 2011.

[68] C. E. Buxton, "Level of mastery and reminiscence in pursuit learning.," *Journal of Experimental Psychology*, vol. 32, no. 2, p. 176, 1943.

[69] Z. Dörnyei and E. Ushioda, *Teaching and researching motivation*. Routledge, 2021.

[70] N. Fields, L. Xu, B. Troutman, K. Daniel, and A. Miles, "Intergenerational reminiscence and digital storytelling: Exploring the experience of college student volunteers," *Innovation in Aging*, vol. 7, no. Supplement_1, pp. 327–327, 2023.

[71] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis.," *Psychological bulletin*, vol. 132, no. 3, p. 354, 2006.

[72] C. H. Karjo and W. Andreani, "Learning foreign languages with duolingo and memrise," in *Proceedings of the 2018 international conference on distance education and learning*, 2018, pp. 109–112.

[73] S. Baker, R. M. Kelly, J. Waycott, *et al.*, "School's back: Scaffolding reminiscence in social virtual reality with older adults," *Proceedings of the ACM on human-computer interaction*, vol. 4, no. CSCW3, pp. 1–25, 2021.

[74] M. P. Stevenson and M. Liu, "Learning a language with web 2.0: Exploring the use of social networking features of foreign language learning websites," *CALICO journal*, vol. 27, no. 2, pp. 233–259, 2010.

[75] R. Gangaiamaran and M. Pasupathi, "Review on use of mobile apps for language learning," *International Journal of Applied Engineering Research*, vol. 12, no. 21, pp. 11 242–11 251, 2017.

[76] J. D. Ewald, "Foreign language learning anxiety in upper-level classes: Involving students as researchers," *Foreign Language Annals*, vol. 40, no. 1, pp. 122–142, 2007.

[77] T. Belpaeme, P. Vogt, R. Van den Berghe, *et al.*, "Guidelines for designing social robots as second language tutors," *International Journal of Social Robotics*, vol. 10, pp. 325–341, 2018.

[78] M. Alemi, A. Meghdari, and M. Ghazisaedy, "The impact of social robotics on l2 learners' anxiety and attitude in english vocabulary acquisition," *International Journal of Social Robotics*, vol. 7, pp. 523–535, 2015.

[79] M. F. S. John and M. A. Gernsbacher, "Learning and losing syntax: Practice makes perfect and frequency builds fortitude," in *Foreign language learning*, Psychology Press, 2013, pp. 231–255.

[80] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Social robot tutoring for child second language learning," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, IEEE, 2016, pp. 231–238.

[81] R. Van den Berghe, J. Verhagen, O. Oudgenoeg-Paz, S. Van der Ven, and P. Leseman, "Social robots for language learning: A review," *Review of Educational Research*, vol. 89, no. 2, pp. 259–295, 2019.

[82] S. A. Hockema and L. B. Smith, "Learning your language, outside-in and inside-out," 2009.

[83] F. Tanaka and S. Matsuzoe, "Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 78–95, 2012.

[84] T. Belpaeme, J. Kennedy, P. Baxter, *et al.*, "L2tor-second language tutoring using social robots," in *Proceedings of the ICSR 2015 WONDER Workshop*, 2015.

[85] M. Alemi, A. Meghdari, and N. S. Haeri, "Young efl learners' attitude towards rall: An observational study focusing on motivation, anxiety, and interaction," in *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, Springer, 2017, pp. 252–261.

[86] M. Elvir, A. J. Gonzalez, C. Walls, and B. Wilder, "Remembering a conversation–a conversational memory architecture for embodied conversational agents," *Journal of Intelligent Systems*, vol. 26, no. 1, pp. 1–21, 2017.

[87] B. A. Ober and G. K. Shenaut, "Semantic memory," in *Handbook of psycholinguistics*, Elsevier, 2006, pp. 403–453.

[88] M. Elvir, "Episodic memory model for embodied conversational agents," 2010.

[89] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[90] M. H. T. Bunga and S. Suyanto, "Developing a complete dialogue system using long short-term memory," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 2019, pp. 326–329.

[91] W. Zhang, Y. Cui, Y. Wang, *et al.*, "Context-sensitive generation of open-domain conversational responses," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 2437–2447.

[92] M. E. Ligthart, M. A. Neerincx, and K. V. Hindriks, "Memory-based personalization for fostering a long-term child-robot relationship," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2022, pp. 80–89.

[93] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "Memorybank: Enhancing large language models with long-term memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 19 724–19 731.

[94] T.-W. Chang, J. Kurcz, M. M. El-Bishouty, Kinshuk, and S. Graf, "Adaptive and personalized learning based on students' cognitive characteristics," *Ubiquitous learning environments and technologies*, pp. 77–97, 2015.

[95] L. Tetzlaff, F. Schmiedek, and G. Brod, "Developing personalized education: A dynamic framework," *Educational Psychology Review*, vol. 33, pp. 863–882, 2021.

[96] A.-e.-h. Munir and W. M. Qazi, "Artificial subjectivity: Personal semantic memory model for cognitive agents," *Applied Sciences*, vol. 12, no. 4, p. 1903, 2022.

[97] F. J. Serón and C. Bobed, "Vox system: A semantic embodied conversational agent exploiting linked data," *Multimedia Tools and Applications*, vol. 75, pp. 381–404, 2016.

[98] T. Mavropoulos, G. Meditskos, S. Symeonidis, *et al.*, "A context-aware conversational agent in the rehabilitation domain," *Future Internet*, vol. 11, no. 11, p. 231, 2019.

[99] P Garrido, F. Martínez, J. B. Peris, S Baldassarri, E Cerezo, and F. J. S. Arbeloa, "Learning through embodied conversational agents with semantic memory," *The International Journal of engineering education*, vol. 34, no. 2, pp. 442–457, 2018.

[100] J. LoCasale-Crouch, F. Jamil, R. C. Pianta, K. M. Rudasill, and J. DeCoster, "Observed quality and consistency of fifth graders' teacher–student interactions: Associations with feelings, engagement, and performance in school," *Sage Open*, vol. 8, no. 3, p. 2 158 244 018 794 774, 2018.

[101] R. Wood, P. Baxter, and T. Belpaeme, "A review of long-term memory in natural and synthetic systems," *Adaptive Behavior*, vol. 20, no. 2, pp. 81–103, 2012.

[102] R. P. Kesner and M. R. Hunsaker, "The temporal attributes of episodic memory," *Behavioural brain research*, vol. 215, no. 2, pp. 299–309, 2010.

[103] X. Xu, Z. Gou, W. Wu, *et al.*, "Long time no see! open-domain conversation with long-term persona memory," *arXiv preprint arXiv:2203.05797*, 2022.

[104] J. Campos and A. Paiva, "May: My memories are yours," in *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, Springer, 2010, pp. 406–412.

[105] S. M. Liao and A. Sandberg, "The normativity of memory modification," *Neuroethics*, vol. 1, pp. 85–99, 2008.

[106] E. A. Phelps and S. G. Hofmann, "Memory editing from science fiction to clinical practice," *Nature*, vol. 572, no. 7767, pp. 43–50, 2019.

[107] E. F. Villaronga, P. Kieseberg, and T. Li, "Humans forget, machines remember: Artificial intelligence and the right to be forgotten," *Computer Law & Security Review*, vol. 34, no. 2, pp. 304–313, 2018.

[108] D. Richards and K. Bransky, "Forgetmenot: What and how users expect intelligent virtual agents to recall and forget personal conversational content," *International Journal of Human-Computer Studies*, vol. 72, no. 5, pp. 460–476, 2014.

[109] Z. Huang, S. Gutierrez, H. Kamana, and S. MacNeil, "Memory sandbox: Transparent and interactive memory management for conversational agents," in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–3.

[110] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.

[111] B. Carr, "Revolutionizing education: Unleashing the power of chat gpt/ai to empower educators," *Technology and the Curriculum: Summer 2023*, 2023.

[112] P. Rospigliosi, *Artificial intelligence in teaching and learning: What questions should we ask of chatgpt?* 2023.

[113] M. Firat, "How chat gpt can transform autodidactic experiences and open education?," 2023.

[114] K. Fuchs, "Exploring the opportunities and challenges of nlp models in higher education: Is chat gpt a blessing or a curse?" In *Frontiers in Education*, Frontiers, vol. 8, 2023, p. 1 166 682.

[115] J. M. Markel, S. G. Opferman, J. A. Landay, and C. Piech, "Gpteach: Interactive ta training with gpt-based students," in *Proceedings of the tenth acm conference on learning@ scale*, 2023, pp. 226–236.

[116] J. A. Van Ek, *Objectives for foreign language learning*. Council of Europe, 1993, vol. 2.

[117] T. A. Fink, "Effectively querying chatgpt via the openai api," 2023. [Online]. Available: https://thomasafink.medium.com/effectively-querying-chatgpt-via-the-openai-api-fc1b24f61a04.

[118] A. Vujinović, N. Luburić, J. Slivka, and A. Kovačević, "Using chatgpt to annotate a dataset: A case study in intelligent tutoring systems," *Machine Learning with Applications*, p. 100 557, 2024, ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2024.100557. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827024000331.

[119] A. Toney-Wails, C. Schoeberl, and J. Dunham, "Ai on ai: Exploring the utility of gpt as an expert annotator of ai publications," *arXiv preprint arXiv:2403.09097*, 2024.

[120] L. Ostyakova, V. Smilga, K. Petukhova, M. Molchanova, and D. Kornev, "Chatgpt vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2023, pp. 242–254.

[121] I. Scopelliti, C. K. Morewedge, E. McCormick, H. L. Min, S. Lebrecht, and K. S. Kassam, "Bias blind spot: Structure, measurement, and consequences," *Management Science*, vol. 61, no. 10, pp. 2468–2486, 2015.

[122] J. Du, D. Blake, L. Wang, C. Conran, D. Mckibben, and A. Way, "Idea: An interactive dialogue translation demo system using furhat robots," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18*, Springer, 2019, pp. 645–648.

[123] M. Jessen, *Book-talks with furhat: How can interaction with conversational robots be used to motivate swedish middle schoolers to read?* 2022.

[124] R. Gardner, "Between speaking and listening: The vocalisation of understandings1," *Applied linguistics*, vol. 19, no. 2, pp. 204–224, 1998.

[125] E. Zachrisson, "Effects of the social robot furhat on vocabulary learning," 2021.

[126] S. Rasouli, G. Gupta, M. Ghafurian, and K. Dautenhahn, "Proposed applications of social robots in interventions for children and adolescents with social anxiety," in *Proceedings of the Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 2022, pp. 1–7.

[127] D. Utami and T. Bickmore, "Collaborative user responses in multiparty interaction with a couples counselor robot," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2019, pp. 294–303.

[128] Š Beňuš, R Sabo, and M Trnka, "Teaching l1 and l2 communication skills with a robotic head," in *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, IEEE, 2019, pp. 69–75.

[129] C. M. Tyng, H. U. Amin, M. N. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in psychology*, vol. 8, p. 235 933, 2017.

[130] Z. Li, B. Peng, P. He, M. Galley, J. Gao, and X. Yan, "Guiding large language models via directional stimulus prompting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[131] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[132] P. Thomas, S. Spielman, N. Craswell, and B. Mitra, "Large language models can accurately predict searcher preferences," *arXiv preprint arXiv:2309.10621*, 2023.

[133] T. Wu, M. Terry, and C. J. Cai, "Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–22.

[134] L. Sweeney, "K-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[135] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, IEEE, 2008, pp. 111–125.

[136] S. Stalla-Bourdillon and A. Knight, "Anonymous data v. personal data-false debate: An eu perspective on anonymization, pseudonymization and personal data," *Wis. Int'l LJ*, vol. 34, p. 284, 2016.

[137] J. Hattie and H. Timperley, "The power of feedback," *Review of educational research*, vol. 77, no. 1, pp. 81–112, 2007.

[138] C. Hou, G. Zhu, J. Zheng, *et al.*, "Prompt-based and fine-tuned gpt models for context-dependent and-independent deductive coding in social annotation," in *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 2024, pp. 518–528.

[139] B. F. Chorpita and D. H. Barlow, "The development of anxiety: The role of control in the early environment," *The Neurotic Paradox, Vol 2*, pp. 227–264, 2018.

[140] E. L. Deci and R. M. Ryan, "The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior," *Psychological inquiry*, vol. 11, no. 4, pp. 227–268, 2000.

[141] P. Hu and J. Zhang, "A pathway to learner autonomy: A self-determination theory perspective," *Asia Pacific Education Review*, vol. 18, pp. 147–157, 2017.

[142] N. Schmitt and R. Carter, "The lexical advantages of narrow reading for second language learners," *Tesol Journal*, vol. 9, no. 1, pp. 4–9, 2000.

[143] I. S. Nation and I. Nation, *Learning vocabulary in another language*. Cambridge university press Cambridge, 2001, vol. 10.

[144] S. J. Savignon, *Interpreting communicative language teaching: Contexts and concerns in teacher education*. Yale University Press, 2008.

[145] M. Calvert and Y. Sheen, “Task-based language learning and teaching: An action-research study,” *Language Teaching Research*, vol. 19, no. 2, pp. 226–244, 2015.

[146] J. M. Murre and J. Dros, “Replication and analysis of ebbinghaus’ forgetting curve,” *PloS one*, vol. 10, no. 7, e0120644, 2015.

[147] R. Lyster and L. Ranta, “Corrective feedback and learner uptake: Negotiation of form in communicative classrooms,” *Studies in second language acquisition*, vol. 19, no. 1, pp. 37–66, 1997.

[148] L Vygotsky and M Cole, “Lev vygotsky: Learning and social constructivism,” *Learning Theories for Early Years Practice. UK: SAGE Publications Inc*, pp. 68–73, 2018.

[149] C. A. Tomlinson, *How to differentiate instruction in mixed-ability classrooms*. Ascd, 2001.

[150] D. A. Schön, *The reflective practitioner: How professionals think in action*. Routledge, 2017.

[151] P. Delobelle, T. Winters, and B. Berendt, “Robbert: A dutch roberta-based language model,” *arXiv preprint arXiv:2001.06286*, 2020.

[152] A. I. Niculescu, R. Jiang, S. Kim, *et al.*, “Sara: Singapore’s automated responsive assistant, a multimodal dialogue system for touristic information,” in *Mobile Web Information Systems: 11th International Conference, MobiWIS 2014, Barcelona, Spain, August 27-29, 2014. Proceedings 11*, Springer, 2014, pp. 153–164.

[153] V. R. Martinez and J. Kennedy, “A multiparty chat-based dialogue system with concurrent conversation tracking and memory,” in *Proceedings of the 2nd Conference on Conversational User Interfaces*, 2020, pp. 1–9.

[154] M. Henderson, B. Thomson, and J. D. Williams, “The second dialog state tracking challenge,” in *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 2014, pp. 263–272.

[155] O. Vinyals and Q. Le, “A neural conversational model,” *arXiv preprint arXiv:1506.05869*, 2015.

[156] Z. Zhang, X. Bo, C. Ma, *et al.*, “A survey on the memory mechanism of large language model based agents,” *arXiv preprint arXiv:2404.13501*, 2024.

[157] J. Olson and E. Södergren, *Long term memory in conversational robots*, 2019.

[158] N. Godewithana and G. Kirindage, “Intelligent hybrid chatbot solution for archaeological sites tracking,” in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, IEEE, 2021, pp. 517–522.

[159] S. Çağatay, “Examining efl students’ foreign language speaking anxiety: The case at a turkish state university,” *Procedia-Social and Behavioral Sciences*, vol. 199, pp. 648–656, 2015.

[160] H. Karatas, B. Alci, M. Bademcioglu, and A. Ergin, “An investigation into university students foreign language speaking anxiety,” *Procedia-Social and Behavioral Sciences*, vol. 232, pp. 382–388, 2016.

[161] C. D. Spielberger, D. N. Hansen, *et al.*, “Effects of state anxiety and task difficulty on computer-assisted learning.,” *Journal of Educational Psychology*, vol. 60, no. 5, p. 343, 1969.

[162] E. Carrillo, L. Moya-Albiol, E. Gonzalez-Bono, A. Salvador, J. Ricarte, and J. Gomez-Amor, "Gender differences in cardiovascular and electrodermal responses to public speaking task: The role of anxiety and mood states," *International Journal of Psychophysiology*, vol. 42, no. 3, pp. 253–264, 2001.

[163] H. Luo, "Foreign language speaking anxiety: A study of chinese language learners," *Journal of the National Council of Less Commonly Taught Languages*, vol. 15, no. 1, pp. 99–117, 2014.

[164] M. Berg, A. Rozental, J. de Brun Mangs, *et al.*, "The role of learning support and chat-sessions in guided internet-based cognitive behavioral therapy for adolescents with anxiety: A factorial design study," *Frontiers in psychiatry*, vol. 11, p. 503, 2020.

[165] W Penny and R Henson, "Analysis of variance," *Statistical parametric mapping: The analysis of functional brain images*, pp. 166–177, 2006.

[166] V. Amrhein, F. Korner-Nievergelt, and T. Roth, "The earth is flat (p> 0.05): Significance thresholds and the crisis of unreplicable research," *PeerJ*, vol. 5, e3544, 2017.

[167] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 143–146.

[168] F. Konietschke, S. Bösiger, E. Brunner, and L. A. Hothorn, "Are multiple contrast tests superior to the anova?" *The International Journal of Biostatistics*, vol. 9, no. 1, pp. 63–73, 2013.

[169] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock, "An aligned rank transform procedure for multifactor contrast tests," in *The 34th annual ACM symposium on user interface software and technology*, 2021, pp. 754–768.

[170] A. Castleberry and A. Nolen, "Thematic analysis of qualitative research data: Is it as easy as it sounds?" *Currents in pharmacy teaching and learning*, vol. 10, no. 6, pp. 807–815, 2018.

[171] P. Feltsen, *Language acquisition and the errors we make: A comparison between beginners and intermediate learners*, 2009.

[172] S. J. Myers, S. D. Davis, and J. C. Chan, "Does expressive writing or an instructional intervention reduce the impacts of test anxiety in a college classroom?" *Cognitive Research: Principles and Implications*, vol. 6, no. 1, p. 44, 2021.

[173] E. Ierardi, M. Bottini, and C. Riva Crugnola, "Effectiveness of an online versus face-to-face psychodynamic counselling intervention for university students before and during the covid-19 period," *BMC psychology*, vol. 10, no. 1, p. 35, 2022.