Illy the IELTS CA: The Significance of Episodic Memory in Embodied Conversational Agent Tutors for Language Learning
Conversational Agents (CS4270), Delft University of Technology

Zohar Cochavi (4962281) **Irtaza Hashmi** (4829360) **Radek Kargul** (4770625) **Niyousha Najmaei** (5739276) **Ardy Zwanenburg** (4856848)

Abstract

Embodied Conversational Agents (ECA) have been shown to be more effective in working with users than their non-embodied counterparts. Especially ECAs which introduce an episodic memory could potentially further improve the engagement of a user in conversation and the building of long-term relationships. This study aims to answer if ECAs in an educational context improve user engagement by implementing episodic memory. Our agent, Illy, practices with users according to the IELTS English speaking exam format, where the capability of memory is presented in the form of short- and long-term feedback by giving a small report to the user in question. From an experiment including a diverse set of participants ($N = 30$), we found that there was a strong preference for an agent that gave feedback over one that did not when considering the likability ($F(6, 56) = 2.97$, $p = 0.014$). The usability of the agent did not show a significant change when participants were presented with either of the two versions.

1 Introduction

In this work, we investigate the effectiveness of using an Embodied Conversational Agent (ECA) with episodic memory for improving spoken language skills. Wik highlights the obstacles of non-natives when improving their spoken English, and demonstrates how people's tendency to anthropomorphize could be a potentially powerful asset in teaching a language [53]. Moreover, pedagogical conversational agents have been shown to be effective in several domains [40], [52].

As language learning is an intrinsically long-term process, the agent needs to be able to build a long-lasting relationship with the user. Numerous sources have shown the impact of using episodic memory in ECAs developing a sense of empathy and intimacy [8], [27] and building long-term relationships with the user [24], [28], [45]. Furthermore, [24] highlights the importance of the affective aspects of user-agent interactions in pedagogical conversational agents. They show that an agent that is able to create deep relationships with the learners and respond to their individual needs and emotions, leads to higher user satisfaction [24]. Adaptation to the student's affective state, and taking into account past interaction, is possible by implementing an episodic memory in the agent. Therefore, we aim at verifying a similar hypothesis to that of [24] in the domain of ECAs for language learning, by investigating the importance of recognizing affect and giving feedback based on previous interactions in a conversational agent with episodic memory for teaching language.

To this end, we have developed an ECA for training users for the speaking test of IELTS, a Standardized English language test. Illy, the IELTS CA, simulates the speaking exam of IELTS for the user, gives feedback on various measures of fluency and speech quality, provides the user with a progress report of the previous practice sessions, and points out weaknesses using motivating speech. Our goal is to investigate the following research question:

> *How much does the existence of episodic memory in a conversational agent for language learning affect the tendency of the users to keep interacting with the agent over a long period of time?*

To investigate this matter, the following sub-questions were addressed.

1. Will an agent with memory create a higher level of user satisfaction?
2. Will participants prefer the agent with memory that recalls previous interactions and gives feedback on progress?
3. Will participants interacting with the agent with memory have a higher tendency to have more practice sessions with the agent and continue their relationship with the agent?
4. Will participants feel more confident about their language skills after using the agent with memory and receiving their progress report over previous sessions?
5. Will participants with more language learning experience have more tendency towards the agent with memory?
6. Will participants more comfortable using the agent (and were more tech-savvy) have a higher tendency towards the agent with memory?

This report is organised as follows. In the next section, we present a number of related works. After that, we describe the architecture of Illy and the different modules. We then present the design, procedure, and results of the conducted experiment.

2 Related Works

There is a great amount of research done in the domain of ECAs. Some works are quite closely related to our investigation when it comes to memory architecture, and affect models. Campos et al. [7] create an agent that uses an episodic memory for maintaining a coherent social relationship over time. Their research emphasizes the importance of topic modeling and signaling explicit recall of previous episodes. Elvir et al. [15] investigate the role of memory in embodied conversational agents. They do that by proposing and implementing a unified episodic memory architecture. Their work is closely related to the design of Illy, as the agent is an empathetic companion that remembers the gist of a conversation, knows what to remember and how to store and retrieve it efficiently. Similar to our case, they assume the robot performs all communications verbally, rather than through non-verbal sounds, and does not include a forgetting mechanism. Commented the following sentences as they were irrelevant and not even analogue.
Illy's goal is to prolong the relationship between the agent and the user by conveying empathy. Leite et al. [27] use an empathetic model for social robots that can interact with children. After conducting a 5-week experiment, the results showed that the long-term interaction was positively perceived. Children that interacted with the robot felt supported by it, just like they would by their peers.
Ahuja et al. [1] investigate spoken language and co-speech gestures, perform a multi-modal fusion of speech and acoustic cues to make a robot seem like it is speaking, and effectively show that their method of gesture generation outperforms the state-of-the-art solution.
Schreuder [45] created a companion for diabetes: PAL (a physical robot). The challenge was to have a long-term relationship with the user, more specifically - children with diabetes. A module helping in capturing and referring to shared experiences between children and the PAL. However, the interaction did not help with the children's episodic memory, it did not increase affection, motivation, and diabetes self-management. Similarly to Illy, the bot uses memory to give motivational feedback to the user.
Kasap et al. [24] designed a system, Eva, that keeps the user's attention even after the first interaction. An important factor in the study was building long-lasting relationships, by keeping the novelty effect by remembering and referring to the engagement level of the user. The agent was tested in four distinct scenarios, with supportive/unsupportive personality and with/without memory. The study concluded that the existence of memory in a long-term interaction system can help keep the users' attention as time passes.
Moreover, Numerous works have been done on turn-taking management [4], [25], [38], the modalities that can be used [12], [13], [37], and back-channels [34], [36] and how they can be used to increase satisfaction and engagement in conversation.

3 Illy the IELTS CA

Numerous sources show that people respond more positively to embodied conversational agents than to their non-embodied counterparts [9], [16], [46]. Illy has a feminine voice and embodiment, as female embodied agents are associated with conveying trustworthiness and empathy [29], and taking a caring role [10] which leads to better user experience.
Illy's architecture consists of four modules: Memory Management module, Natural Language Processing (NLP) module, Turn-taking and Interruption Management module. In the following, we discuss the main components of our proposed agent for validating our hypothesis about the research question.

3.1 Memory Management Module

Memory is an important part of the dialogue and human behavior in general, not only because it is necessary for directing the conversation and responding intelligently [17], but also for building a long-term relationship with the agent and keeping users engaged [23]. The latter is especially important as Illy is designed to be a tool for long-term improvement of English speech fluency (within the context of the IELTS exam). These short-term and long-term goals are also reflected in the type of memory used. Namely short-term and long-term memory, which roughly corresponds to the 'practice' and 'evaluation' stages of dialogue respectively. Tying together the different types of memory and extracting relevant information from each of these sources is the responsibility of the Memory Management Module.

Architecture

McKinley et al. [31] associate common ground or mutual knowledge with an efficient conversation. They conclude that both item and context memories are positively correlated with common ground, of which the former is used in our agent. Our short-term memory is not a context-memory as we only assume a single user per session. Similar to Elvir et al. [15], a database contains the utterances in raw-text format with metadata that is attached after processing (which also includes a part-of-speech (POS) tagger stage [15]). Besides POS tags, this metadata contains the 'fluency' of this particular utterance as well as the probability of being part of different topics. Names, and facial features (as extracted by our facial recognition module) are either learned from user input or retrieved from long-term memory and then stored as attributes in the Memory Manager. Session data such as a count of off-topic utterances and a user talking over-time are stored similarly.
Long-term memory is an episodic context-memory that is used to adapt to the learning needs of the user by selecting cue cards and follow-up questions from topics that were more difficult for the user during the previous interactions. In similar work, Kasap et al. [23]'s approach, the episodic memory is stored in an architecture based on the Belief-Desire-Intention (BDI) model. They represent the episodes in terms of context, content, and outcome [50]. In our architecture, context, content, and outcome correspond to the "topics" (Metadata) and "cue_card_id" (Session), "tokens" (Utterance), and "fluency" (MetaData) fields respectively (see Figure 2). It contains all the session- and

user-data as stored in the short-term memory. The metadata paired is condensed with their utterances into an over-all set of scores to better reflect human memory in terms of short- versus long-term memory function. This concretely means that the scores of the utterances are averaged out, and the number of times a user has deviated from the topic of the cue (i.e. when the utterance does not share the topic with the highest probability with that of the cue) is normalized over the number of submitted utterances. Since long-term memory needs to be persistent over multiple sessions, it will be stored in a MongoDB database along with short-term memory. It also includes a semantic memory in the form of a CSV file which contains the 'cue cards'.
The long- and short-term memory schemas are portrayed in Figure 2. These only contain "who" spoke "when" explicitly since "where" a dialogue took place does not matter for determining the fluency of a user. "What" a user has said is important, but only so far as how it relates to the relevance with regard to the cue given by Illy (as mentioned previously).

Forgetting and Remembering

Short-term memory will be flushed to long-term as soon as a session has ended, meaning that, while some information is lost after a period of time, we do preserve some information to determine cues in future sessions. As mentioned in Olson & Sodergren, the use of a decay function such as the Ebbinghaus (which accurately describes human forgetfulness) curve could be useful for reusing cues that have already been used and probably forgotten by the user [39] akin to existing spaced repetition models [56]. Perhaps forgetfulness could also improve the engagement of the users with the agent, but decay functions in human-to-agent interactions are currently an open field of research [43] and not necessary for the purposes of this study.
While persistence is not needed, short-term memory entries are also stored in a MongoDB collection to maintain consistency in memory retrieval. This memory retrieval is done at three stages: user identification, cue choice, and producing session/progress feedback. As the IELTS exam has preset follow-up questions and cues, there is no need for, what Kasap & Magenat-Thalmann refer to as, interactive memory-retrieval when choosing a cue for the user to respond to [55]. User identification and producing reports (which in turn requires memory retrieval) are similarly deliberate actions and thus require no systems for dynamic memory retrieval.

3.2 Turn-taking and Interruption Management Module

The ability to take turns in a fluent way (i.e., without long response delays or frequent interruptions) is a fundamental aspect of any spoken dialog system [14]. Turn-taking is natural to humans and starts to be learned at an early age by infants [35]. The correct and appropriate timing of turn shifts is crucial to the conversation's flow. In the context of human-machine interaction, user satisfaction is directly impacted by this issue because inappropriate interruptions or premature turn-taking can drastically degrade the user experience, and cause frustration. Nonetheless, turn-taking remains a challenging area of research in human-machine interactions [49].
Turn-taking and interruption manager is an essential perception module in the case of our agent that ensures a smooth and correct flow of the dialog FSM, which is illustrated in Figure 3. The turn-taking and interruption manager module should decide whether a pause is turn-yielding or not. During the *practice session* part of the dialog, see Figure 4, the agent should be able to identify whether a pause indicates the user thinking about what should be said next or whether the speech is over. Moreover, in case the user exceeds the limit of two minutes, the agent performs a Floor Taking Interrupt [6], and asks the follow-up question. It should be noted that we do not take the "rude examinee" scenario into account. In other words, we assume that the user will not interrupt the agent.
Traditionally, distinguishing between turn-holding and turn-yielding pauses is done by segmenting the speech into Inter-Pausal Units, which are stretches of audio from one speaker without any silence exceeding a certain amount [48]. Despite the limitations of this approach, like being purely reactive [51], and potentially unnatural, we use this method, as it is suitable for our use-case. During the simulation of the exam, the examiner should give the test taker enough time to think between sentences when below the time limit. Therefore, slightly longer delays in identifying the need for turn-taking, will not sound unnatural.

3.3 NLP Module

Topic Modeling

Prediction of conversation topics is crucial for coherent and engaging dialogue systems [26]. This can be achieved with topic modeling[20]. We use it to process and cluster user utterances to see if the user stays on a topic related to asked questions. Topic modeling is done with Latent Dirichlet Allocation (LDA), which clusters topics according to relevant keywords [54] (for architecture, see Figure 5). The training data for the model was downloaded from IELTS Speaking Corpus. It contains 826 examples (topic question + sample answer). Data with missing parts were ignored. Next, the text is tokenized with NLTK library, where stop words are removed. Finally, the text was lemmatized, and put in a bag of words using Gensim library. LDA model is trained on 32 general topics [19] provided by the IELTS speaking test. The coherence score was used to evaluate the quality of data and how many topics the data represents. Two coherence scores were used, CV (computes the co-occurrences of words and creates content vectors, and calculates the score by using normalized pointwise mutual information (NPMI) and the cosine similarity [44]) and UMass (computes how often two words i and j appear together in the corpus [32]). Finally, perplexity was used as a metric to see how the model performs on unseen data. According to the metrics analysis, 28-32 topics yielded the best coherence score. Using these results, the final model was trained using scikit-learn with 32 topics. The clusters of the 32 topics are visualized using pyLDAvis in Figure 6. Based on observations, the model did fairly well in clustering

topics and predicting new text documents, given that they are of similar length to the training data documents.

User Intent

Knowing the user intent allows us to maintain the correct flow of the dialog FSM, therefore it is a critical part of the system. The architecture of user intent classification is shown in Figure 5. Snips NLU library was used to train the user intent classification model. It is a Natural Language Understanding (NLU) that parses sentences in natural language and extracts structure information. The library was used to classify user intent based on our data. We created training data and trained a model for our custom user intents. Our user intents included "decline", "confirm", "greeting", "introduction", "practice", "clarification", "feedback", "speech", and "silence". For each of the intents, sample text was created to represent the intent. In each of the sample texts, entities were included. The entities allow resolving the user intents. For example, "sayingHello" entity was created to represent whenever the user is saying "hello". Synonyms for an entity were also defined. For example, the synonyms of "hello" were defined as "hey", "good morning", etc. This was done for all the user intents. Different types of sample text were created containing different entities, such that the model is able to understand a wide range of text and classify them into one of the user intents.

Language Fluency

The IELTS Speaking exam consists of different criteria that can get a band between 0 and 9 (lowest to highest) [1]. The "Fluency and coherence" criterion indicates word frequency and self-corrects in a sentence. As the bot uses speech-to-text conversion, fluency has a rule-based scoring system, hence the score is directly related to the number of repetitions made by a user in a sentence. A model-based approach was considered, utilizing user speech (audio features) [41]. The language fluency architecture can be seen in Figure 5.
The user's performance is always measured with respect to a certain topic, and we consider the use of connectives and range of vocabulary to attribute most to "fluency" (as per the IELTS rubric mentioned in the first section, [18]). Since the main difference between topics is vocabulary, we can conclude that varying the topic could lead to different levels of fluency. For this reason, we determine the appropriate topic for the session and user based on the principle of spaced repetition [22].

3.4 Affect Module

Building a long-lasting relationship with the user is a crucial necessity for Illy to have a significant impact on the user's speaking abilities and result in progress. The agent provides feedback on fluency, staying on topic, anxiety during the speech, and the duration of the speech, and points out the user's progress over multiple practice sessions. For this progress report to have a positive impact on the user's speaking skill, the interactions between the user and the agent need to be sustained over time.

[1]https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx

Empathy-enabled agents are essential for sustained interaction between the agent and the user, and to behave empathically, social robots need to understand some of the user's affective states and respond appropriately [27]. Therefore, affect recognition is used in dialog management to make the user more empathetic and increase user satisfaction. Moreover, we employ affect recognition for identifying any probable anxiety in the user's speech and addressing this problem in the feedback given to the user, since the test taker has a better chance of impressing the examiner if he/she stays calm and confident.
Our agent uses a multi-modal statistical approach for detecting emotion, by assessing linguistics with EmoRoBERTa mode and vocal cues with MLP and RAVDESS dataset. The output of these models is then fused by late output fusion.

Linguistics model

For the affect model, an enhanced emotion detection model EmoRoBERTa[21] is used. EmoRoBERTa is a robust model trained on a GoEmotions dataset[2]. The dataset consists of over 58000 data samples and includes 28 emotions in total (including a neutral one). We group "anger", "disgust", "annoyance", "confusion" and "curiosity" into "frustration emotions" and "confusion emotions" categories, respectively. Upon perceiving confusion emotions, the agent tries to clarify by repeating the question. Moreover, when frustration emotions are perceived, the agent tries to calm the user down with empathetic language. Other models that make use of BERT, such as Valence Arousal Dominance Lexicon [33] and SpanEmo [2], were outperformed by GoEmotion (also employs BERT) [11]. This includes the state-of-the-art model biLSTM [42]. RoBERTa, is a more advanced version of BERT. It is a more trained model, exceeding the benchmark scores of BERT and being the current state-of-the-art [30].

Vocal cues model

To detect the user's emotion through speech, we used a pre-trained Multi-layer Perceptron (MLP) model. The model was trained on the RAVDESS dataset. The dataset had 24 different voices (60 samples each). Each contains a different modality: vocal channel, emotion, emotional intensity, statement, repetition, and actor. The model is trained using sklearn MLP Classifier. Vocal cues of the user's speech are then classified into eight different emotions, namely neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The "angry" and "surprised" classes are then considered to be "frustration" and "confusion" emotions respectively. Moreover, the "fearful" class is used as an indication of anxiety in the user's speech.

Multi-modal fusion

Finally, both linguistics and vocal cues are fused. The multimodality complements Illy's potential to detect user emotions. Uniform weighting of both modalities by averaging [47] ensures a more consistent emotion detection. [3] mentions the challenges of multi-modal fusion, like

[2]https://huggingface.co/arpanghoshal/EmoRoBERTa

fusion and alignment of inputs, however, these challenges do not apply to our case, since we fuse the outputs in the end.

4 Methodology

We carried out an experiment in order to measure the effect of episodic memory on user experience and satisfaction. We expected an agent with memory that adapts to the user's individual emotional and learning needs to create higher user satisfaction and lead to a longer-lasting relationship with the user. Our general hypotheses are listed below.

1. An agent with memory will create a higher level of user satisfaction.

2. Participants will have a tendency toward the agent with memory that recalls previous interactions and gives feedback on progress.

3. Participants interacting with the agent with memory will have a higher tendency to have more practice sessions with the agent and continue their relationship with the agent.

4. Participants will feel more confident about their language skills after using the agent with memory and receiving their progress reports over previous sessions.

5. Participants with more language learning experience will have more tendency towards the agent with memory.

6. Participants that were more comfortable using the agent (and were more tech-savvy) will have a higher tendency towards the agent with memory.

In this section, the design of the experiment, participants, measures of evaluation, and the procedure is briefly discussed.

4.1 Design

The experiment was designed as a between-subject study. Two versions of the agent were given to the control and experimental group. The control group interacted with a version of the agent that gave no feedback or progress report, whereas the experimental group interacted with a version that gave feedback and progress report.

4.2 Participants

The experiment had 30 participants, 17 of which identified as female, 11 as male, and 2 as non-binary. Participants were chosen from non-native English speakers who either were learning English or had learned English previously. The mean and variance of the number of languages spoken by each participant from the control group were $\mu = 3.07, \sigma^2 = 0.60$, whereas those of the experimental group were $\mu = 2.67, \sigma^2 = 0.89$.

4.3 Measures

Qualitative measures were gathered from the questionnaires to evaluate the two versions of the system in terms of "likability" and "usability". A scale with five discrete options ranging from "Strongly Disagree" to "strongly Agree" was used. The questionnaires provide a qualitative measure of the subject's satisfaction with the agent, tendency to train with the agent again, satisfaction with the given feedback, and idea about the lack of feedback. Moreover, gender, age, and the number of languages the user speaks are other qualitative and quantitative measures that are collected and used in the analysis.

4.4 Procedure

The procedure of the experiment, collected data, and the risks associated were communicated to the participants before the start of the experiment in an informed consent form. However, the participants were not informed that there are two versions of the agent and that they are interacting with one of the two versions. During the experiment, users were permitted to interact freely with the agent and attempt multiple practice sessions. Each experiment took about 20-30 minutes, depending on whether it was for the control or the experimental group.
After concluding the practice, the introductory questionnaire, Appendix E, was given to both groups. After that, the questionnaires in Appendix F and G, were given to the experimental group and the control group respectively. The questionnaires are based on the System Usability Scale [5].

5 Results

When analyzing the data, we used the Multivariate Analysis of Variance (MANOVA) to compare different variables, in this case questions in the survey. These questions can be split into the categories 'likeability' and 'usability', which will answer sub-questions 1,2 and 3,4,6 respectively.
Sub-question 5 can be answered by the presence of a correlation between language experience and the tendency towards the agent with memory.
Usability is analyzed as described by System Usability Scale (SUS) [5], but did not show any significant difference between the test and control groups as the Pillai's trace is $F(20, 42) = 0.77$, $p < 0.732$. Moving to likeability, MANOVA compared the question *"I would like to use the bot in the future to further improve my English speaking skills"* (which was present in both questionnaires) with *"The feedback I got from the agent was helpful in improving my speaking ability"* (which was only present in the questionnaire for the group with feedback), and *"It would be useful to have some feedback from the agent."* (which was only present in the questionnaire for the group without feedback). This analysis showed a significant difference with $F(6, 56) = 2.97$, $p = 0.014$. As for users who speak more languages (sub-question 5), no significant difference was found in the preference this group had to the agent with feedback than other groups, where $F(6, 56) = 0.55$, $p = 0.771$.

Taking Pillai's trace difference between the control and test group, $F(6, 56) = 0.55$, $p = 0.771$, shows that there is no significant difference in terms of age, gender, or the number of languages between the two. For further details, the mean values of users' responses to individual questions can be found in H.

References

[1] C. Ahuja, D. W. Lee, R. Ishii, and L.-P. Morency, "No gestures left behind: Learning relationships between spoken language and freeform gestures," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1884–1895. DOI: 10.18653/v1/2020.findings-emnlp.170. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.170.

[2] H. Alhuzali and S. Ananiadou, "Spanemo: Casting multi-label emotion classification as span-prediction," *arXiv preprint arXiv:2101.10038*, 2021.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[4] J. B. Bavelas, L. Coates, and T. Johnson, "Listener responses as a collaborative process: The role of gaze," *Journal of Communication*, vol. 52, no. 3, pp. 566–580, Sep. 2002. DOI: 10.1111/j.1460-2466.2002.tb02562.x. [Online]. Available: https://doi.org/10.1111/j.1460-2466.2002.tb02562.x.

[5] J. Brooke, "Sus: A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, Nov. 1995.

[6] A. Cafaro, N. Glas, and C. Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *Proceedings of the 2016 International Conference on Autonomous Agents amp; Multiagent Systems*, ser. AAMAS '16, Singapore, Singapore: International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 911–920, ISBN: 9781450342391.

[7] J. Campos, J. Kennedy, and J. F. Lehman, "Challenges in Exploiting Conversational Memory in Human-Agent Interaction," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '18, Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, Jul. 2018, pp. 1649–1657. (visited on 09/15/2022).

[8] J. Campos and A. Paiva, "MAY: My memories are yours," in *Intelligent Virtual Agents*, Springer Berlin Heidelberg, 2010, pp. 406–412. DOI: 10.1007/978-3-642-15892-6_44. [Online]. Available: https://doi.org/10.1007/978-3-642-15892-6_44.

[9] J. Cassell, T. Bickmore, M. Billinghurst, *et al.*, "Embodiment in conversational interfaces: Rea," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '99, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1999, pp. 520–527, ISBN: 0201485591. DOI: 10.1145/302979.303150. [Online]. Available: https://doi.org/10.1145/302979.303150.

[10] A. Danielescu, "Eschewing gender stereotypes in voice assistants to promote inclusion," in *Proceedings of the 2nd Conference on Conversational User Interfaces*, ser. CUI '20, Bilbao, Spain: Association for Computing Machinery, 2020, ISBN: 9781450375443. DOI: 10.1145/3405755.3406151. [Online]. Available: https://doi.org/10.1145/3405755.3406151.

[11] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[12] N. Dethlefs, H. Hastie, H. Cuayahuitl, Y. Yu, V. Rieser, and O. Lemon, "Information density and overlap in spoken dialogue," English, *Computer Speech and Language*, vol. 37, pp. 82–97, May 2016, ISSN: 0885-2308. DOI: 10.1016/j.csl.2015.11.001.

[13] J. Edlund and M. Heldner, "Exploring prosody in interaction control," *Phonetica*, vol. 62, no. 2-4, pp. 215–226, Dec. 2005. DOI: 10.1159/000090099. [Online]. Available: https://doi.org/10.1159/000090099.

[14] E. Ekstedt and G. Skantze, "Projection of turn completion in incremental spoken dialogue systems," in *SIGDIAL*, 2021.

[15] M. Elvir, A. J. Gonzalez, C. Walls, and B. Wilder, "Remembering a Conversation – A Conversational Memory Architecture for Embodied Conversational Agents," en, *Journal of Intelligent Systems*, vol. 26, no. 1, pp. 1–21, Jan. 2017, Publisher: De Gruyter, ISSN: 2191-026X. DOI: 10.1515/jisys-2015-0094. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/jisys-2015-0094/html (visited on 09/15/2022).

[16] M. E. Foster, "Face-to-face conversation: Why embodiment matters for conversational user interfaces," in *Proceedings of the 1st International Conference on Conversational User Interfaces*, ser. CUI '19, Dublin, Ireland: Association for Computing Machinery, 2019, ISBN: 9781450371872. DOI: 10.1145/3342775.3342810. [Online]. Available: https://doi.org/10.1145/3342775.3342810.

[17] W. S. Horton and R. J. Gerrig, "The impact of memory demands on audience design during language production," *Cognition*, vol. 96, no. 2, pp. 127–142, 2005, ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2004.07.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027704001416.

[18] *IELTS scoring in detail*. [Online]. Available: https://www.ielts.org/for-organisations/ielts-scoring-in-detail (visited on 09/30/2022).

[19] *IELTS Speaking Part 1 Topics & Questions*, [Online; accessed 4. Nov. 2022], Nov. 2022. [Online]. Available: https://ieltsliz.com/ielts-speaking-part-1-topics.

[20] H. Jelodar, Y. Wang, C. Yuan, and X. Feng, "Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *CoRR*, vol. abs/1711.04305, 2017. arXiv: 1711.04305. [Online]. Available: http://arxiv.org/abs/1711.04305.

[21] R. Kamath, A. Ghoshal, S. Eswaran, and P. B. Honnavalli, "Emoroberta: An enhanced emotion detection model using roberta," in *IEEE International Conference on Electronics, Computing and Communication Technologies*, 2022.

[22] S. H. K. Kang, "Spaced repetition promotes efficient and effective learning: Policy implications for instruction," *Policy Insights from the Behavioral and Brain Sciences*, vol. 3, no. 1, pp. 12–19, 2016. DOI: 10.1177/2372732215624708. eprint: https://doi.org/10.1177/2372732215624708. [Online]. Available: https://doi.org/10.1177/2372732215624708.

[23] Z. Kasap and N. Magnenat-Thalmann, "Building long-term relationships with virtual and robotic characters: The role of remembering," *The Visual Computer*, vol. 28, no. 1, pp. 87–97, Sep. 2011. DOI: 10.1007/s00371-011-0630-7. [Online]. Available: https://doi.org/10.1007/s00371-011-0630-7.

[24] Z. Kasap and N. Magnenat-Thalmann, "Building long-term relationships with virtual and robotic characters: The role of remembering," en, *The Visual Computer*, vol. 28, no. 1, pp. 87–97, Jan. 2012, ISSN: 1432-2315. DOI: 10.1007/s00371-011-0630-7. [Online]. Available: https://doi.org/10.1007/s00371-011-0630-7 (visited on 09/15/2022).

[25] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967, ISSN: 0001-6918. DOI: https://doi.org/10.1016/0001-6918(67)90005-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0001691867900054.

[26] C. Khatri, R. Goel, B. Hedayatnia, *et al.*, "Contextual topic modeling for dialog systems," *CoRR*, vol. abs/1810.08135, 2018. arXiv: 1810.08135. [Online]. Available: http://arxiv.org/abs/1810.08135.

[27] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 329–341, Mar. 2014. DOI: 10.1007/s12369-014-0227-1. [Online]. Available: https://doi.org/10.1007/s12369-014-0227-1.

[28] I. Leite, A. Pereira, and J. F. Lehman, "Persistent Memory in Repeated Child-Robot Conversations," in *Proceedings of the 2017 Conference on Interaction Design and Children*, ser. IDC '17, New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 238–247, ISBN: 978-1-4503-4921-5. DOI: 10.1145/3078072.3079728. [Online]. Available: http://doi.org/10.1145/3078072.3079728 (visited on 09/15/2022).

[29] N. Liu, R. Liu, and W. J. Li, "Identifying design feature factors critical to acceptance of smart voice assistant," in *HCI*, 2021.

[30] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[31] G. L. McKinley, S. Brown-Schmidt, and A. S. Benjamin, "Memory for conversation and the development of common ground," en, *Memory & Cognition*, vol. 45, no. 8, pp. 1281–1294, Nov. 2017, ISSN: 1532-5946. DOI: 10.3758/s13421-017-0730-3. [Online]. Available: https://doi.org/10.3758/s13421-017-0730-3 (visited on 09/15/2022).

[32] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. Mccallum, "Optimizing semantic coherence in topic models," Jan. 2011, pp. 262–272.

[33] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2018, pp. 174–184.

[34] L.-P. Morency, I. Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, pp. 70–84, Jan. 2010. DOI: 10.1007/s10458-009-9092-y.

[35] V. Nguyen, O. Versyp, C. Cox, and R. Fusaroli, "A systematic review and bayesian meta-analysis of the development of turn taking in adult–child vocal interactions," *Child Development*, vol. 93, no. 4, pp. 1181–1200, Mar. 2022. DOI: 10.1111/cdev.13754. [Online]. Available: https://doi.org/10.1111/cdev.13754.

[36] C. Oertel, P. Jonell, D. Kontogiorgos, J. Mendelson, J. Beskow, and J. Gustafson, "Crowd-sourced design of artificial attentive listeners," in *Interspeech 2017*, ISCA, Aug. 2017. DOI: 10.21437/interspeech.2017-926. [Online]. Available: https://doi.org/10.21437/interspeech.2017-926.

[37] C. Oertel and G. Salvi, "A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue," Dec. 2013, pp. 99–106. DOI: 10.1145/2522848.2522865.

[38] C. Oertel, M. Wlodarczak, J. Edlund, P. Wagner, and J. Gustafson, "Gaze patterns in turn-taking," vol. 3, Nov. 2012.

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

[40] D. Pérez-Marın and I. Pascual-Nieto, "An exploratory study on how children interact with pedagogic conversational agents," *Behaviour & Information Technology*, vol. 32, no. 9, pp. 955–964, 2013.

[41] A. Preciado-Grijalva and R. F. Brena, *Speaker fluency level classification using machine learning techniques*, 2018. DOI: 10.48550/ARXIV.1808.10556. [Online]. Available: https://arxiv.org/abs/1808.10556.

[42] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A cnn-bilstm model for document-level sentiment analysis," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 832–847, 2019.

[43] 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, ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2014.01.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581914000147.

[44] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15, Shanghai, China: Association for Computing Machinery, 2015, pp. 399–408, ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. [Online]. Available: https://doi.org/10.1145/2684822.2685324.

[45] B. Schreuder Goedheijt, *Recalling shared memories in an embodied conversational agent : Personalized robot support for children with diabetes in the pal project*, 2017. [Online]. Available: http://essay.utwente.nl/80062/.

[46] A. Shamekhi, Q. V. Liao, D. Wang, R. K. E. Bellamy, and T. Erickson, "Face value? exploring the effects of embodiment for a group facilitation agent," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18, Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–13, ISBN: 9781450356206. DOI: 10.1145/3173574.3173965. [Online]. Available: https://doi.org/10.1145/3173574.3173965.

[47] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2016, pp. 160–170.

[48] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Saarbrücken, Germany: Association for Computational Linguistics, Aug. 2017, pp. 220–230. DOI: 10.18653/v1/W17-5527. [Online]. Available: https://aclanthology.org/W17-5527.

[49] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech &amp Language*, vol. 67, p. 101 178, May 2021. DOI: 10.1016/j.csl.2020.101178. [Online]. Available: https://doi.org/10.1016%2Fj.csl.2020.101178.

[50] D. G. Tecuci and B. W. Porter, "A generic memory module for events," in *FLAIRS Conference*, 2007.

[51] M. Tice and T. Henetz, "Reading between the turns: Social perceptions of turn-taking cues in conversation," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2443–2443, Oct. 2011. DOI: 10.1121/1.3654799. [Online]. Available: https://doi.org/10.1121/1.3654799.

[52] N. Wellnhammer, M. Dolata, S. Steigler, and G. Schwabe, "Studying with the help of digital tutors: Design aspects of conversational agents that influence the learning process," 2020.

[53] P. Wik, "The virtual language teacher : Models and applications for language learning using embodied conversational agents," QC 20110511, Ph.D. dissertation, KTH, Speech Communication and Technology, 2011, pp. 45–51, ISBN: 978-91-7415-990-5.

[54] J.-F. Yeh, C.-H. Lee, Y.-S. Tan, and L.-C. Yu, "Topic model allocation of conversational dialogue records by latent dirichlet allocation," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, 2014, pp. 1–4. DOI: 10.1109/APSIPA.2014.7041546.

[55] Z. Yumak and N. Thalmann, "Towards episodic memory-based long-term affective interaction with a human-like robot," Oct. 2010, pp. 452–457. DOI: 10.1109/ROMAN.2010.5598644.

[56] A. Zaidi, A. Caines, R. Moore, P. Buttery, and A. Rice, "Adaptive Forgetting Curves for Spaced Repetition Language Learning," in *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 358–363,

ISBN: 978-3-030-52240-7. DOI:
10.1007/978-3-030-52240-7_65.

Appendix

A Memory Architecture

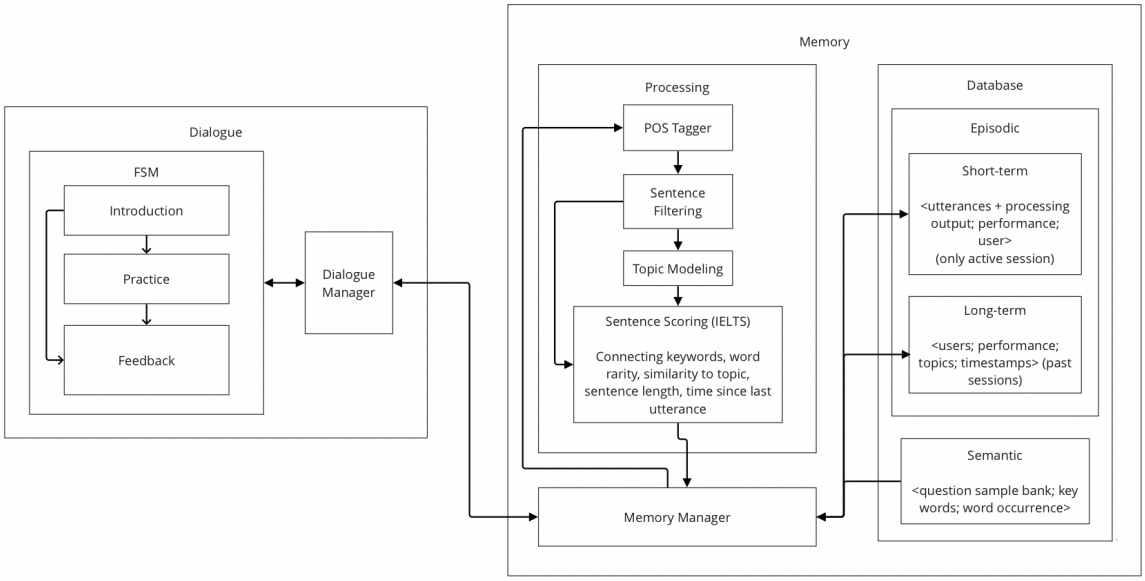

Figure 1: The schema of the memory architecture

Field	Type
user	User
_id	int
start_time	float
end_time	float
cue_card_id	int
follow_ups_idx	list[int]
average_score	float
on_topic	float
over_time	bool

(a) Session

Field	Type
name	str
encodings	NDArray
_id	int

(b) User

Field	Type
tokens	list[tuple[str, str]]
timestamp	float
speech_state	bool
metadata	MetaData
_id	int

(c) Utterance

Field	Type
topics	dict[int, float]
fluency_score	int

(d) MetaData

Figure 2: Different models used in the short- and long-term memory. (a), (b) are found in short- as well as long-term memory, while (c) and (d) (which are exclusively paired) are only found in short-term memory. When short-term memory is flushed, the MetaData objects paired with the Utterance objects are condensed and turned into average_score and on_topic.

B Dialogue Finite State Machine

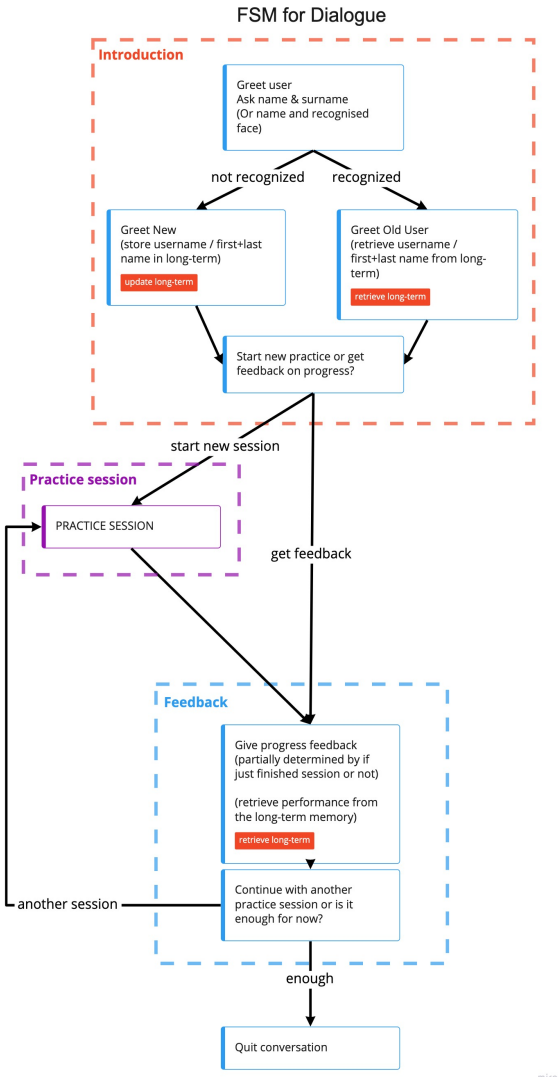

Figure 3: Finite State Machine (FSM) for Dialogue. The tags in color boxes indicate memory actions The text in parentheses indicates the action that happens with data and memory.

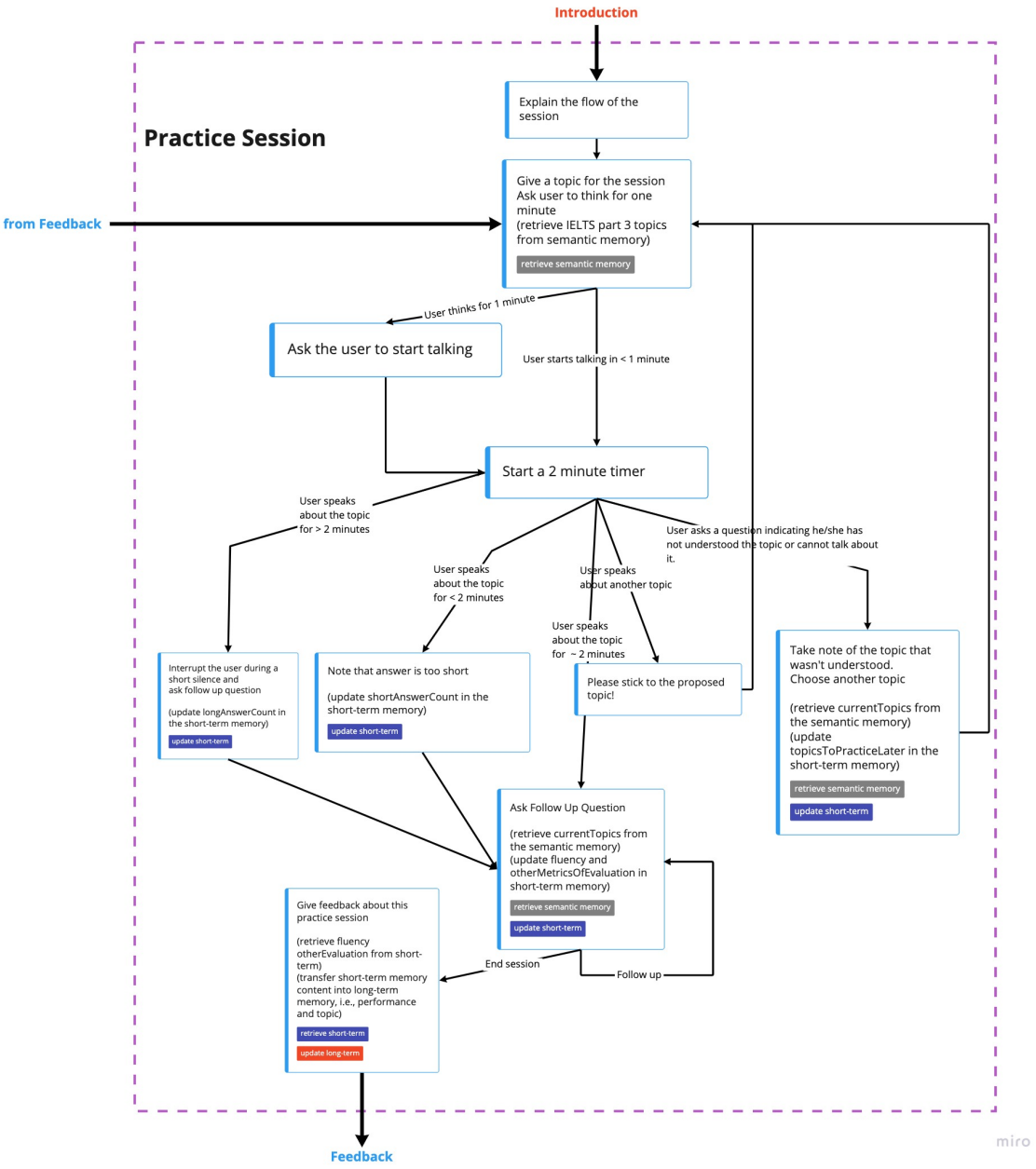

Figure 4: Finite State Machine (FSM) for Dialogue: Practice Session component. Note that the tags are for clarification purposes only and do not indicate how and when the memory manager updates and retrieves data from the memory.

C Perception

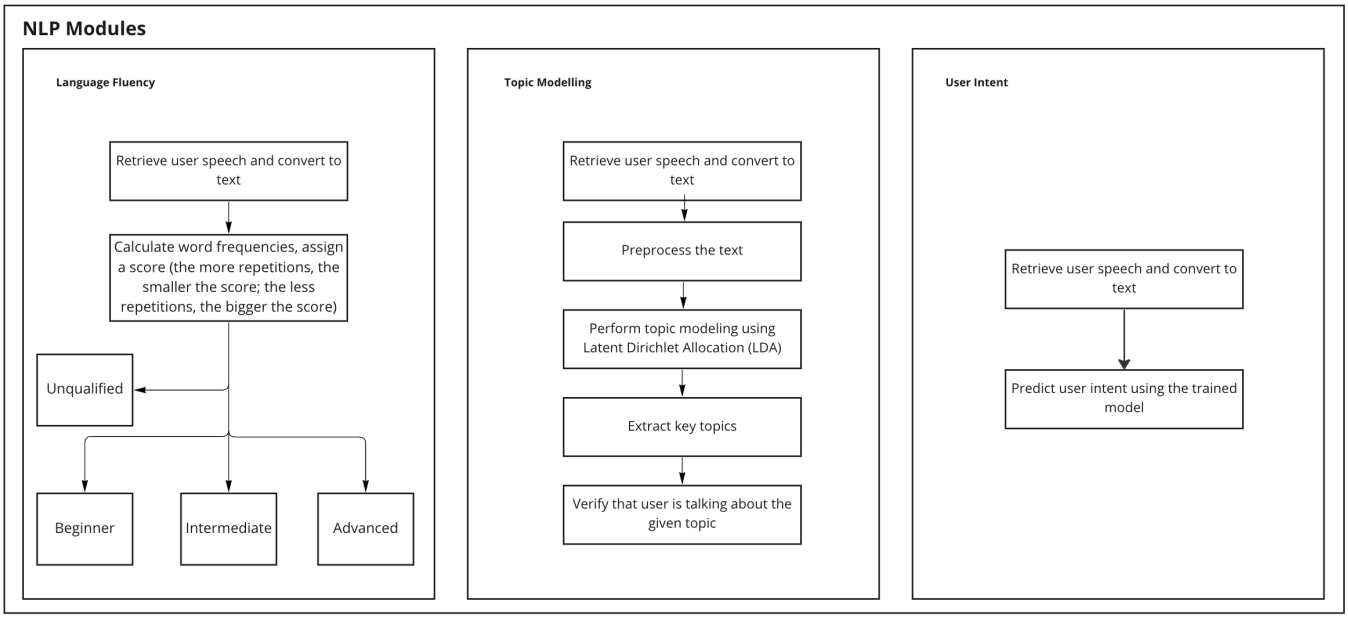

Figure 5: The NLP module and its corresponding sub-modules

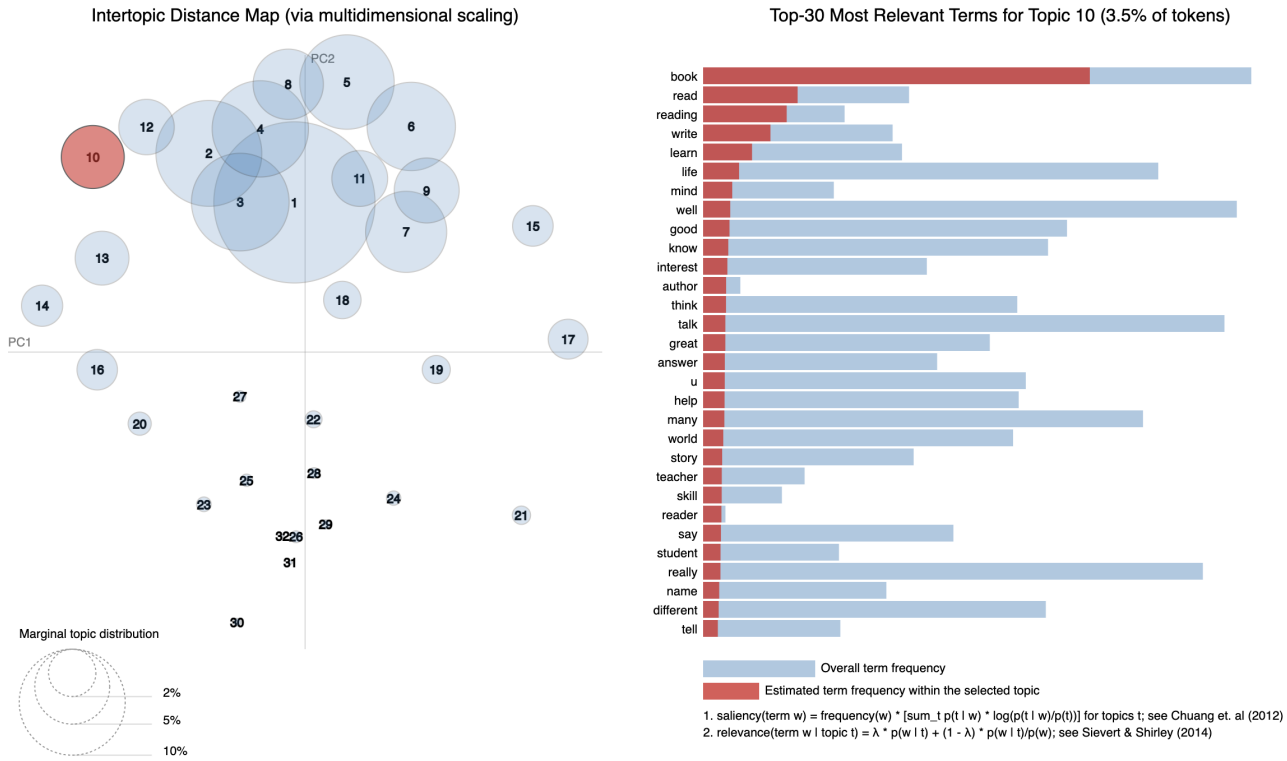

Figure 6: Clustered topics using LDA topic modeling

D Consent form

Please tick the appropriate boxes	Yes	No
Taking part in the study		
I have read and understood the study information dated [DD/MM/YYYY], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction	☐	☐
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
I understand that taking part in the study involves my name being used by the agent during the experiment and filling in questionnaires about the experience with the agents.	☐	☐
Risks associated with participating in the study		
I understand that taking part in the study involves the following risk when having fear of failure: stress beyond normal experience.	☐	☐
Use of the information in the study		
I understand that information I provide will be used for a report for the TU Delft course Conversational Agents (CS4370).	☐	☐
I understand that only the answers to my questionnaires will be stored and used for this research and kept until the end of the research/course. Any personal data will not be stored after the experiment	☐	☐

Participant

_____

Signature

_____

Date

Researcher

_____

Signature

_____

Date

E Intro Questionnaire

The following questions are about users, to determine their age, gender, and the number of languages they speak. The questionnaire is necessary to use the statistics to perform our experiment. Please enter the information in the boxes. Mark the box that resonates most strongly with your experience.

	Male	Female	Other
Please specify your gender.			

	18-21	21-30	¿30
Please specify your age range.			

Please specify the number of languages you speak. [ ]

F Questionnaire 1 - a group with feedback version

The following questions are about the sessions you had with the agent. Please rate each question on a scale of 1 to 5. Where the first box is "Strongly Disagree", the second box is "Neutral" and the last box is "Strongly Agree". Mark the box that resonates most strongly with your experience.

	Strongly Disagree				Strongly Agree
I think that I would like to use this system frequently					
I found the system unnecessarily complex					
I thought the system was easy to use					
I think I would need the support of a technical person to be able to use this system					
I found various functions in this system were well integrated					
I thought there was too much inconsistency in this system					
I would imagine that most people would learn to use this system very quickly					
I felt confident about my language skills after using the agent					
I need to learn a lot of things before I could get going with this system					
I would like to use the bot in the future to further improve my English speaking skills.					
The evaluation results I got from the agent match my opinion about my own skills.					
The feedback I got from the agent was helpful in improving my speaking ability.					

Any other comments on the session you could not express in the questions above?

_____

_____

_____

G Questionnaire 2 - the control group with no feedback

The following questions are about the sessions you had with and without the agent. Please rate each question on a scale of 1 to 5. Where the first box is "Strongly Disagree", the second box is "Neutral" and the last box is "Strongly Agree". Mark the box that applies

	Strongly Disagree				Strongly Agree
I think that I would like to use this system frequently					
I found the system unnecessarily complex					
I thought the system was easy to use					
I think I would need the support of a technical person to be able to use this system					
I found various functions in this system were well integrated					
I thought there was too much inconsistency in this system					
I would imagine that most people would learn to use this system very quickly					
I felt confident about my language skills after using the agent					
I need to learn a lot of things before I could get going with this system					
I would like to use the bot in the future to further improve my English speaking skills.					
It would be useful to have some feedback from the agent.					

Any other comments on the session you could not express in the questions above?

_____

_____

_____

H Questionnaire results

In the following table the Averages and Variances where f+ is with feedback and f- is without feedback.

	Avg F+	Avg F-	Var F+	Var F-
I think that I would like to use this system frequently	3.13	3.40	1.45	0.77
I found the system unnecessarily complex	1.40	1.27	0.77	0.20
I thought the system was easy to use	4.60	4.60	0.64	0.64
I think I would need the support of a technical person to be able to use this system	1.27	1.07	0.60	0.06
I found various functions in this system were well integrated	3.47	3.80	0.52	0.29
I thought there was too much inconsistency in this system	2.27	1.93	1.13	1.13
I would imagine that most people would learn to use this system very quickly	4.20	4.60	0.69	0.64
I felt confident about my language skills after using the agent	4.20	3.07	1.23	1.40
I need to learn a lot of things before I could get going with this system	2.20	1.93	1.36	1.66
I would like to use the bot in the future to further improve my English speaking skills.	3.40	3.67	1.17	1.02
It would be useful to have some feedback from the agent	N.A.	4.87	N.A.	0.25
The evaluation results I got from the agent match my opinion about my own skills.	3.73	N.A.	1.13	N.A.
The feedback I got from the agent was helpful in improving my speaking ability.	3.27	N.A.	1.26	N.A.