Leveraging sentiment analysis and memory modules to increase a user's level of engagement with a conversational agent

Alexandra Neagu
Delft University of Technology

1 INTRODUCTION

In the landscape of user-agent interactions, memory serves as a critical component, enabling agents to store past user interactions and ultimately contributing to enhanced likability and the cultivation of enduring relationships. Nevertheless, personalization extends beyond the mere recollection of conversation episodes, encompassing the user's anticipation of the agent's behaviour evolving organically as shared experiences accumulate.

The impact of the evolution of an agent's behaviour on user engagement and sentiment is significant and multifaceted. Not only does it aid in establishing a stronger common ground [8, 12], it also enhances the level of personalization offered to the user [9]. Additionally, some users may prefer a less supportive agent in specific learning contexts, highlighting the importance of tailoring the agent's behaviour to individual preferences [8]. Combining these challenges with the memory of an agent is still being heavily explored and researched.

This paper seeks to address a fundamental research question in relation to the agent's memory and sentiment: "In what ways does tailoring a conversational agent's tone — by leveraging sentiment analysis on top of memory modules — to reflect its sentiment during a meal-prepping interaction influence the individual user's level of engagement?".

This is studied in the context of planning meals, as finding a diverse, yet nutritious food pattern is often perceived as difficult. The developed conversational agent will assist users by scheduling 4 meals for a day. Based on the user input, such as allergies and preferences, suggestions are made in order to aid the user in this process, recurring use of the agent helps generate more suitable suggestions since the agent incorporates memory of past conversations.

To model an appropriate memory architecture for this use case, both episodic and semantic memory is needed. Episodic memory, which recalls specific memories for individuals, is necessary for keeping track of user preferences related to recipes, taste, help from the agent, allergies etc., and semantic memory, long-term memory about general knowledge of the world, is needed for providing recipes and general cooking knowledge.

2 BACKGROUND

In recent years, the field of conversational agents has witnessed significant advancements, with memory playing a pivotal role in enhancing the capabilities of these agents. Understanding the nuances of memory in conversational agents is crucial for creating more natural and effective human-robot interactions. This background section delves into the most important findings from relevant memory-related papers, exploring key aspects such as content vs source, context, common ground, autobiography, belief-desire-intention, and the distinction between long-term and short-term memory.

Common Ground in Conversations , or the mutual knowledge and understanding developed during conversations, is crucial for reducing ambiguity and misunderstanding. It involves memory of conversation content (item memory) and context (who said what), and recognizes the role of participants as speakers or listeners [12].

Individual Preferences in Conversational Agents , the personality of the robot is important, with some users preferring unsupportive robots. This underscores the importance of customizing robot personalities to match specific user preferences, enhancing the human-robot interaction experience [8].

Autobiographic and Episodic Memory in Agents . Agents with autobiographic memory, which includes personal memories, appear more personal and believable [5]. Episodic memory, keeping track of experiences from the agent's perspective, uses a "gist" of the conversation for future reference [5]. Systems like Memory Graph Networks (MGN) learn natural paths among episodic memory nodes, improving engagement by allowing users to retrieve and browse through their stored memories [13]. A similar system using episodic memory is given in [16]. Here, the concept of 5W1H (When, Where, Who, Why, What and How) is mainly used to create and store the episodic memory, which can then be used later when appropriate.

Short-term and long-term memory . In addition to autobiographic and episodic memory, consolidating short-term memory into long-term memory is a costly process[15]. A study in [10] found that children preferred robots that remembered details about them. Simultaneously, keeping users engaged during prolonged interactions is crucial, in [8] a framework is proposed, which shows that memory-enabled robots excel at maintaining user interest, with some users preferring the challenging nature of unsupportive robots over supportive robots.

Understanding beliefs, desires and intentions of both the agent and the (human) user is essential. Leolani's robot employs a Belief, Desire, Intention (BDI) approach to model this in one-to-one conversations [18].

Multi-User Communication , existing frameworks rely on assumptions that make it hard to enable realistic group communication, thus possibly losing track of context. A solution is proposed [11] in which an additional sub-module keeps track of topics and participants in each (sub-)conversation. Expanding upon this, Leolani's implementation of a theory of mind emphasizes the importance of situational information about the world in combination

with the perspective of information sources to reason over knowledge, desires, and intentions during conversations [18].

Forgetfulness especially with minor details, has been shown to enhance the naturalness of a conversational agent [14]. By using Ebbinghaus' forgetting curve [9] a memory module can implement forgetfulness in a long-term memory system for personalized responses while maintaining appropriateness.

3 METHODOLOGY

Our experiment is primarily focused on the ability of a conversational agent to foster a social bond with users, rather than only attempting to understand and implement user preferences. The context in which the conversational agent operates is that of culinary preparation; specifically, it will help the user plan the meals for the next day. It adapts based on the history with the user by remembering important aspects such as past interactions, preferred dishes, known allergies, and the social context of their cooking. The history and relationship between the user and the agent will persist beyond individual sessions or isolated interactions, data will be stored in a database and organized for later retrieval in the same or in future conversations.

An issue which might arise is that of contrasting information. The user's current mood or circumstances might change and deviate from what the agent knows, which, if not accounted for, might result in sub-optimal interactions. To avoid over-relying on historical information, sentiment scores will be calculated on the user inputs such that a consistent history is kept and variations can be detected as the conversation progresses.

System Design: a schematic representation of our perception module can be found in Appendix A.2. The perception module will take as input audio speech. Speech input provides a more natural and intuitive means of communication, mimicking human conversation and catering to users who prefer or require hands-free interaction [4, 17]. During the conversation, the perception module will pass the parsed voice message on to the rest of the system through a JSON object.

The dialogue system consists of the User Input Processor, or the NLU module, which processes the user's input and extracts and structures the relevant information for the memory module to store or query. We have incorporated an Intent Recognition module to accurately recognise the user's intentions or requests from their input, with a particular focus on correctly identifying allergens, due to the need for higher accuracy in these aspects of the conversation.

The response generation will be done using Natural Language Generation (NLG) with the Cohere LLM [2]. By integrating the context into the prompt generation, we can generate appropriate responses that keep memory and context in mind. On the output side, the system includes a Response Generator module, the NLG module. This module constructs natural language responses to communicate effectively with the user. It incorporates information retrieved from the memory module by extracting to personalize these responses or to acknowledge the user's history, thereby creating a more engaging interaction.

The conversational agent's memory framework, as depicted in Appendix A.1, incorporates both factual and episodic elements. Within this framework, long-term memory is designated for storing factual memories. These memories consist of relatively static information, such as specific user details, notably including allergies. When a user starts the conversation, allergies are retrieved and kept in the short-term memory as Python objects so that they can be quickly retrieved during the interaction. Alongside, episodic memories are formed through ongoing user interactions. These will live in the short-term memory of the agent, and once the conversation is over, the planned meals are stored in an SQLite database intended for long-term memory. Inputs to the memory module consist of structured data processed by the NLU, while outputs are in a machine-readable format like JSON.

Throughout the interaction, the agent will recommend meals, and based on the user's response, either deem them satisfactory or process new requests. To contextualize these requests, the system tracks the last four made by the user, shifting earlier ones to long-term memory for consistent conversation management. Through this forgetfulness, previous satisfactory meals and requested modifications are taken into account when suggesting following meals.

Inputs and outputs of the system: in terms of processing techniques for this input, the agent will use Speech Recognition Systems to convert spoken words into text using Furhat's integrated speech recognition engine [7]. It will first generate transcribed speech from the audio it received, then tokenize this text, breaking it down into individual tokens (words or phrases). For the text processing, the TextBlob NLP library [3] will be used. TextBlob is an API-based text processor with a multitude of Natural Language Processing (NLP) methods that can help develop our agent such as sentiment analysis, tokenization, counting word and phrase frequencies and spelling correction. Lastly, we perform intent recognition on the user inputs in order to reliably detect when an allergen is mentioned in the conversation.

The output of the system will be produced by the NLG module which passes the generated text to Furhat. This is done through a JSON object which includes the message which has to be converted to speech and a status code which can be read to verify if the module was successful.

Data sources and error handling: the developed agent possesses a general knowledge base through the database that is being maintained, consisting of important information about users. In addition, the LLM plays an important role as a data source for the culinary responses. Another important element is Furhat, which can be programmed such that it recognizes responses such as names or allergies. Furhat is also given instructions on how to start and guide the conversation.

Whenever something unintended happens and Furhat does not understand, it will mention this to the user and ask the question again. Whenever the user input is understood, but the LLM does not understand, this will be expressed in the LLM response that is read to the user by Furhat. In no case will the system crash due to unintended behaviour.

Challenges encountered: as one could expect, the development of the conversational agent came with its challenges. The main challenges faced were those of designing a suitable, extensive, yet understandable system architecture. This required a good amount

of work and thinking. In addition, structuring the conversation between the agent and the user did not turn out to be an easy task, as the user should get some guidance in what it can say, but should also be left free, such that it does not feel restricted. To accommodate for this, the LLM receives some context with the user input, such that it gets a better understanding of what the user might mean. Finding the right format for prompting the LLM for suitable responses also required a good amount of testing and experimenting.

3.1 Experiment Procedure and Evaluation

In order to answer the proposed research question, an experiment was designed such that participants would interact with a formal or informal version of our agent. The goal of the experiment is to answer the question "How do users' levels of engagement differ during a daily meal planning conversation with a formal agent compared to an informal agent?".

The experiment has been conducted as a between-subject experiment, meaning that half of the group will test the formal agent, while the other half of the group will test the friendly agent. The main consideration for a between-subject experiment as opposed to a within-subject experiment is the reduction of possibly biased outcomes. Since participants will only have to test one agent, they will not suffer from fatigue and learning effects.

Procedure: the experiment was conducted physically by providing the participants with a consent form and explaining to them the experiment's purpose, procedures, potential risks and confidentiality, also emphasizing the experiment is on a voluntary basis and withdrawal is possible at any time. Afterwards, the users completed an interaction session with the agent, taking no more than 15 minutes. The agent briefly introduces itself and its intention, followed by requesting the allergies and dietary preferences of the participant. This information is subsequently used to plan a day's worth of meals. During this process, objective metrics have been recorded, such as the sentiment of the inputs and the frequency of questions asked. To conclude the experiment, participants were asked to fill out a questionnaire regarding the interaction [1]. The combined data collected during this experiment provide the possibility of drawing conclusions from the experiment.

Participants: the experiment has been executed by 30 participants, of which 19 are male and 11 are female, with their ages ranging from 20 to 55 and the mean age being 25. Most participants are TU Delft students, complemented with friends and family, all with technological proficiency, as this is an important inclusion criterion. The participants have evenly been divided between the agent applying the formal and the friendly tone and they apply an artificial name to conform to responsible data security and privacy measures and to ensure Furhat recognizes their input as a name.

Each participant in our experiment is assigned a fictitious allergy or none at all, ensuring equal distribution across the formal and friendly groups. This approach aims to evaluate the agent's ability to accurately recall and respond to these artificial allergy details, testing its performance and maintaining group balance. The primary focus is on the agent's effectiveness in handling different allergy scenarios and its impact on the user experience, rather than on actual allergy data from participants. Conversational history is collected solely for sentiment analysis and is promptly deleted post-analysis for privacy.

Evaluation Measures: in the experiment, the user's level of engagement and sentiment variation were measured during the interaction with the agent. The engagement has been measured with the short version of the Artificial-Social-Agent Questionnaire (ASAQ) [1], a questionnaire tailored for quick analysis of the interaction with an artificial social agent, which consists of 24 questions about the user's perception of the agent. The corresponding results are displayed in ASA charts [6], specifically tailored to report ASAQ results. The chart normalizes the scores from the questionnaire on a 7-point scale, ranging from -3 to 3. The arrangement of the metrics is based on factor analysis and theoretical similarities, meaning that neighbouring constructs can be compared, as they are related to user engagement. Answers to the questionnaires will be used to further evaluate the interactions in two ways:

- **T-tests** are performed on all categories to find any categories that show significant differences between the formal versus informal results, and the allergy versus no-allergy result.
- **Pearson correlation test** is performed to verify whether the correlation between engagement, trust, perceptions of attentiveness, and usability is statistically significant.

Sentiment values will be calculated through TextBlob [3]. The values will lie on a scale between -1 and 1 where 1 represents positive sentiment and -1 negative. The value will be calculated after each message is received. Once the experiment is concluded we average the sentiment score over the whole conversation.

4 RESULTS

ASA charts have been generated to make two comparisons, first of all, figure 1 compares the results obtained from participants using the formal with the friendly agent, where the score in the middle is the mean of the total score for both groups. Secondly, figure 2 compares the results obtained from participants who did and did not indicate an allergy.

Furthermore, the results of the t-tests can be found in Table 2 for the t-test between the formal and informal data, and Table 3 for the allergy and no-allergy data. For any category with a p-value < 0.05, the null hypothesis can be rejected (being that the category shows no significant difference in the mean of the scored number), which means that the difference in scores for that category is significant. This means that for the formal and informal data, Sociability and User's Trust show significant differences, and for the allergy and no-allergy data Emotional Intelligence Presence is deemed significantly different.

Analyzing the Pearson correlation coefficients (table 1) between user sentiment and various factors reveals that in both informal and formal contexts, user sentiment has a minimal impact on Usability (AU) and User Engagement (UE). However, there is a moderate positive correlation with Attentiveness (AA) and User Trust (UT) in informal settings and a lesser, but still positive correlation in formal settings. This suggests that, while user sentiment is not a strong predictor of usability or engagement, it has a more significant

relationship with attentiveness and trust, particularly in informal contexts.

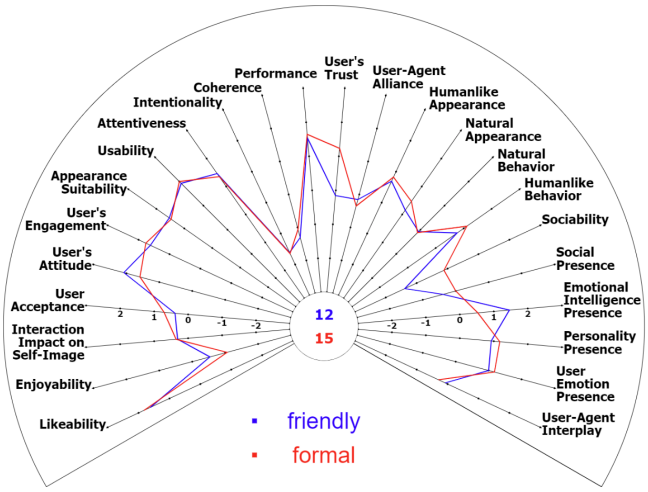


Figure 1: ASA chart for friendly and formal agent

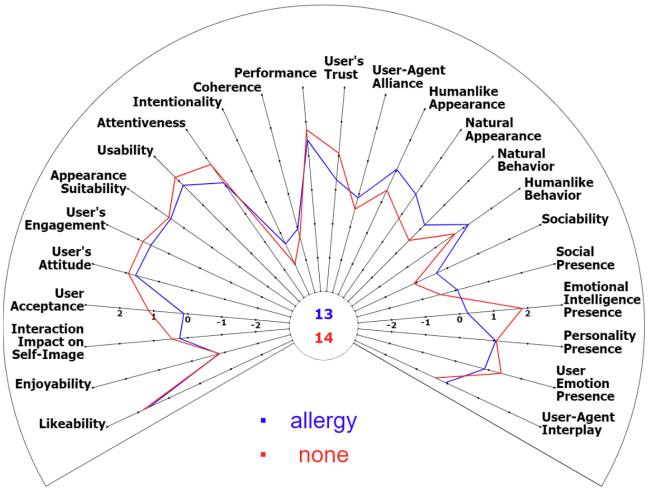


Figure 2: ASA chart for participants with allergy and no allergy

type	AU	UE	AA	UT
Informal	0.090	0.002	0.214	0.212
Formal	0.001	-0.068	0.131	0.165

Table 1: Pearson correlation coefficient between ASAQ constructs and user sentiment

Label	T-Statistic	P-Value
Likeability	-0.387	0.701
Enjoyability	0.88	0.387
Interaction Impact on Self-Image	-0.082	0.935
User Acceptance	-0.502	0.62
User's Attitude	1.53	0.137
User's Engagement	-0.367	0.716
Appearance Suitability	0.115	0.909
Usability	-0.161	0.873
Attentiveness	0.206	0.838
Intentionality	0.054	0.957
Coherence	-0.455	0.653
Performance	-0.238	0.814
User's Trust	-2.308	0.029
User-Agent Alliance	0.398	0.694
Humanlike Appearance	-0.217	0.83
Natural Behavior	-0.497	0.623
Natural Behavior	0.056	0.956
Humanlike Behavior	-0.646	0.524
Sociability	-2.372	0.025
Social Presence	-0.758	0.455
Emotional Intelligence Presence	1.385	0.177
Personality Presence	-0.437	0.666
User Emotion Presence	-0.306	0.762
User-Agent Interplay	0.401	0.691

Table 2: T-test results between formal and informal data

Label	T-Statistic	P-Value
Likeability	-0.37	0.714
Enjoyability	0.038	0.97
Interaction Impact on Self-Image	-0.412	0.684
User Acceptance	-1.691	0.102
User's Attitude	-0.658	0.516
User's Engagement	-0.637	0.529
Appearance Suitability	-0.115	0.909
Usability	-0.698	0.491
Attentiveness	-1.273	0.214
Intentionality	1.128	0.269
Coherence	0.455	0.653
Performance	-0.658	0.516
User's Trust	-1.223	0.231
User-Agent Alliance	0.692	0.495
Humanlike Appearance	1.141	0.264
Natural Behavior	1.387	0.176
Natural Behavior	0.883	0.385
Humanlike Behavior	0.903	0.374
Sociability	1.272	0.214
Social Presence	1.035	0.31
Emotional Intelligence Presence	-2.575	0.016
Personality Presence	-0.031	0.975
User Emotion Presence	-0.882	0.385
User-Agent Interplay	0.604	0.551

Table 3: T-test results between allergy and no allergy data

REFERENCES

[1] Artificial Social Agent Questionnaire - ASAQ | Artificial Social Agent Questionnaire - ASAQ. (????). https://ii.tudelft.nl/evalquest/

[2] Cohere | The leading AI platform for enterprise. (????). https://cohere.com/

[3] GitHub - sloria/TextBlob: Simple, Pythonic, text processing–Sentiment analysis, part-of-speech tagging, noun phrase extraction, translation, and more. — github.com. https://github.com/sloria/TextBlob. (????). [Accessed 07-12-2023].

[4] James Allen, George Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of the 6th international conference on Intelligent user interfaces*. 1–8.

[5] Miguel Elvir, Avelino J. Gonzalez, Christopher Walls, and Bryan Wilder. 2017. Remembering a Conversation – A Conversational Memory Architecture for Embodied Conversational Agents. *Journal of Intelligent Systems* 26, 1 (2017), 1–21. https://doi.org/doi:10.1515/jisys-2015-0094

[6] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. 2022. The Artificial-Social-Agent Questionnaire: Establishing the Long and Short Questionnaire Versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, Article 18, 8 pages. https://doi.org/10.1145/3514197.3549612

[7] David Griol, Jesús García-Herrero, and José M Molina. 2011. The educagent platform: Intelligent conversational agents for e-learning applications. In *Ambient Intelligence-Software and Applications: 2nd International Symposium on Ambient Intelligence (ISAmI 2011)*. Springer, 117–124.

[8] Zerrin Kasap and Nadia Magnenat-Thalmann. 2011. Building long-term relationships with virtual and robotic characters: the role of remembering. *The Visual Computer* 28, 1 (Sept. 2011), 87–97. https://doi.org/10.1007/s00371-011-0630-7

[9] Yonghee Kim, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2015. Acquisition and Use of Long-Term Memory for Personalized Dialog Systems. In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, Ronald Böck, Francesca Bonin, Nick Campbell, and Ronald Poppe (Eds.). Springer International Publishing, Cham, 78–87.

[10] Iolanda Leite, André Pereira, and Jill Fain Lehman. 2017. Persistent Memory in Repeated Child-Robot Conversations. In *Proceedings of the 2017 Conference on Interaction Design and Children (IDC '17)*. Association for Computing Machinery, New York, NY, USA, 238–247. https://doi.org/10.1145/3078072.3079728

[11] Victor R. Martinez and James Kennedy. 2020. A Multiparty Chat-Based Dialogue System with Concurrent Conversation Tracking and Memory. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages. https://doi.org/10.1145/3405755.3406121

[12] Geoffrey L. McKinley, Sarah Brown-Schmidt, and Aaron S. Benjamin. 2017. Memory for conversation and the development of common ground. *Memory & Cognition* 45, 8 (July 2017), 1281–1294. https://doi.org/10.3758/s13421-017-0730-3

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

[14] Deborah Richards and Karla Bransky. 2014. ForgetMeNot: What and how users expect intelligent virtual agents to recall and forget personal conversational content. *International Journal of Human-Computer Studies* 72, 5 (2014), 460–476. https://doi.org/10.1016/j.ijhcs.2014.01.005

[15] María-Loreto Sánchez, Mauricio Correa, Luz Martínez, and Javier Ruiz-del Solar. 2015. An Episodic Long-Term Memory for Robots: The Bender Case. In *RoboCup 2015: Robot World Cup XIX*, Luis Almeida, Jianmin Ji, Gerald Steinbauer, and Sean Luke (Eds.). Springer International Publishing, Cham, 264–275.

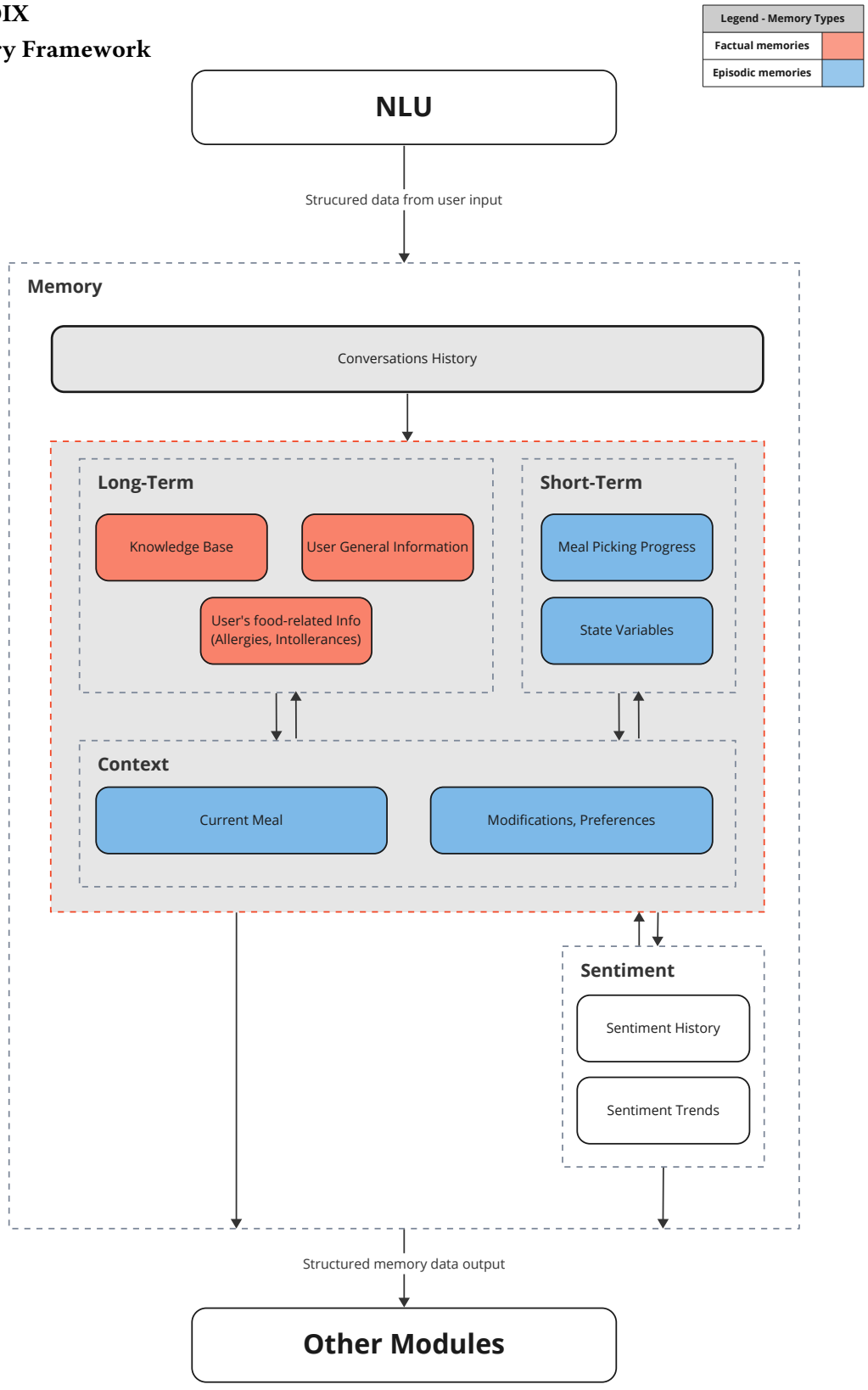
[16] Bart Schreuder Goedheijt. 2017. Recalling shared memories in an embodied conversational agent : personalized robot support for children with diabetes in the PAL project. (2017). http://essay.utwente.nl/80062/

[17] Vincenzo Scotti, Roberto Tedesco, and Licia Sbattella. 2021. A Modular Data-Driven Architecture for Empathetic Conversational Agents. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 365–368. https://doi.org/10.1109/BigComp51126.2021.00080

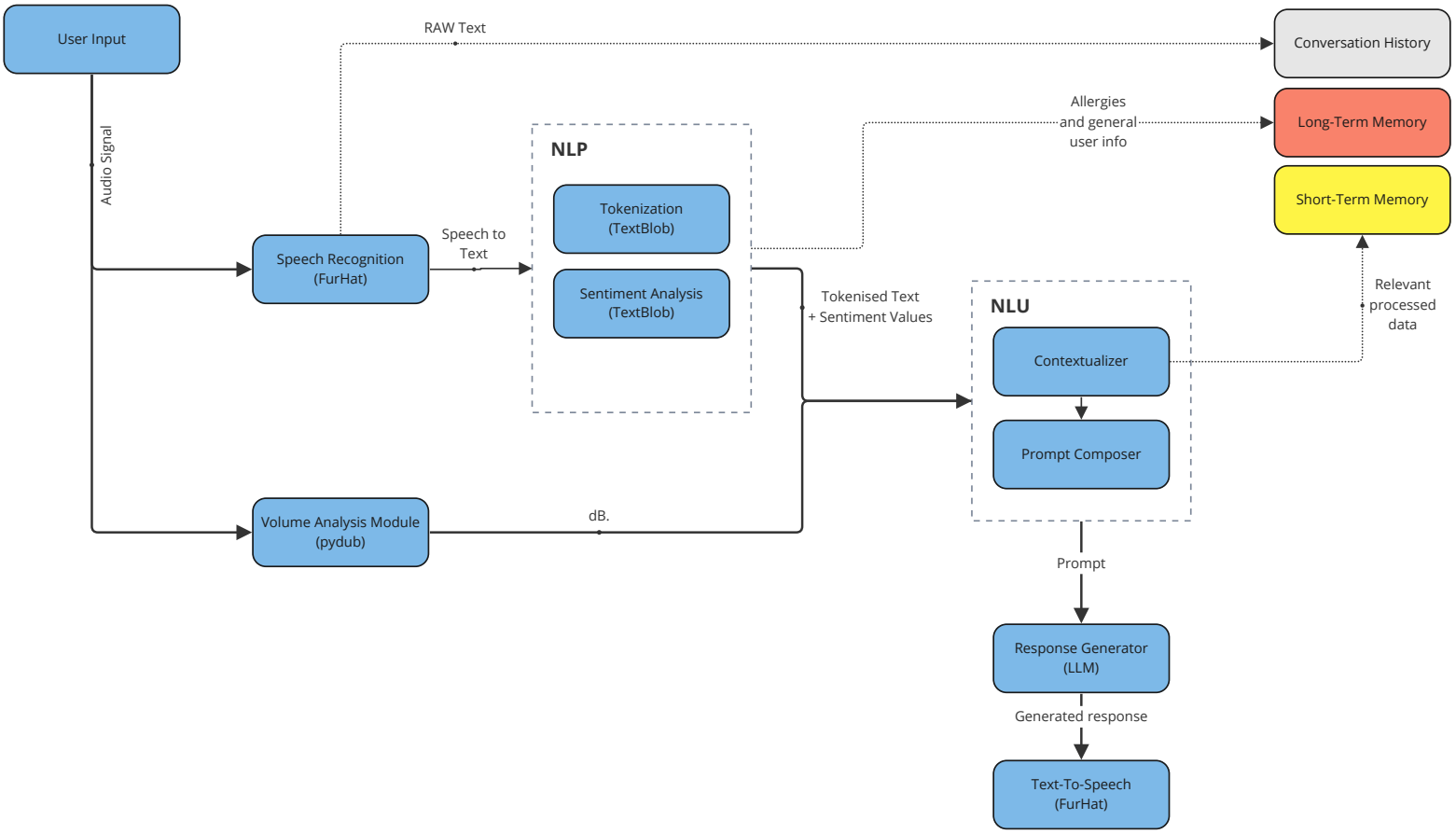
[18] Piek Vossen, Selene Baez, Lenka Bajcetić, and Bram Kraaijeveld. 2018. Leolani: A Reference Machine with a Theory of Mind for Social Communication. In *Text, Speech, and Dialogue*, Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (Eds.). Springer International Publishing, Cham, 15–25.

A APPENDIX

A.1 Memory Framework



A.2 Perception Framework



A.3 Human research ethics

In addition to the separate analyses for sentiment and ASAQ data, we will conduct a correlation analysis to explore the relationship between user engagement levels and sentiment scores. Scatter plots will allow us to plot user engagement scores from the ASAQ against sentiment scores obtained from TextBlob. By visualizing these data points, we can identify any patterns or trends that indicate a relationship between these two variables. This method is particularly useful for detecting whether higher levels of engagement correlate with specific sentiment trends or vice versa.

In general, when conducting experiments in this field of research, there are numerous risks for the participants involved. The magnitude of these risks differs per experiment. The most important risks participants might be exposed to are the following, where we have provided an explanation and risk assessment for each of them, as well as possible ways of dealing with them:

- Privacy: it should be ensured that as little data is collected about participants as possible, only information that might help in determining relations and connections is gathered. This information consists of age, gender and allergies. This data is kept anonymous and is only used for the experiment and its results. The data that is collected should be stored safely, such that it will only be accessible to the people needing it, which is our project group.
- Consent: the participants of the experiment should be informed about the exact intention and structure of the experiment before participating and should be left free to decide whether or not to participate. Therefore, we are providing possible participants with a clear description beforehand, as well as a consent form that should contain all the necessary information, enabling them to make an informed decision.
- Form of the address of the conversational agent towards the participants: since our experiment consists of different tones used by the agent to speak to users, we should prevent the tone from becoming undesired. Examples are abusive language or swearing and cursing, such behaviour by the agent should not occur in interaction with the user since this is not part of the experiment and is generally unwanted.
- Spread of false or misleading information: it might be possible that the agent comes up with suggestions about recipes or other matters related to the topic that are wrong, for example regarding recipes that are suitable for certain allergies. This is something to treat very carefully, since consuming ingredients that an individual is allergic to could lead to severe health issues.