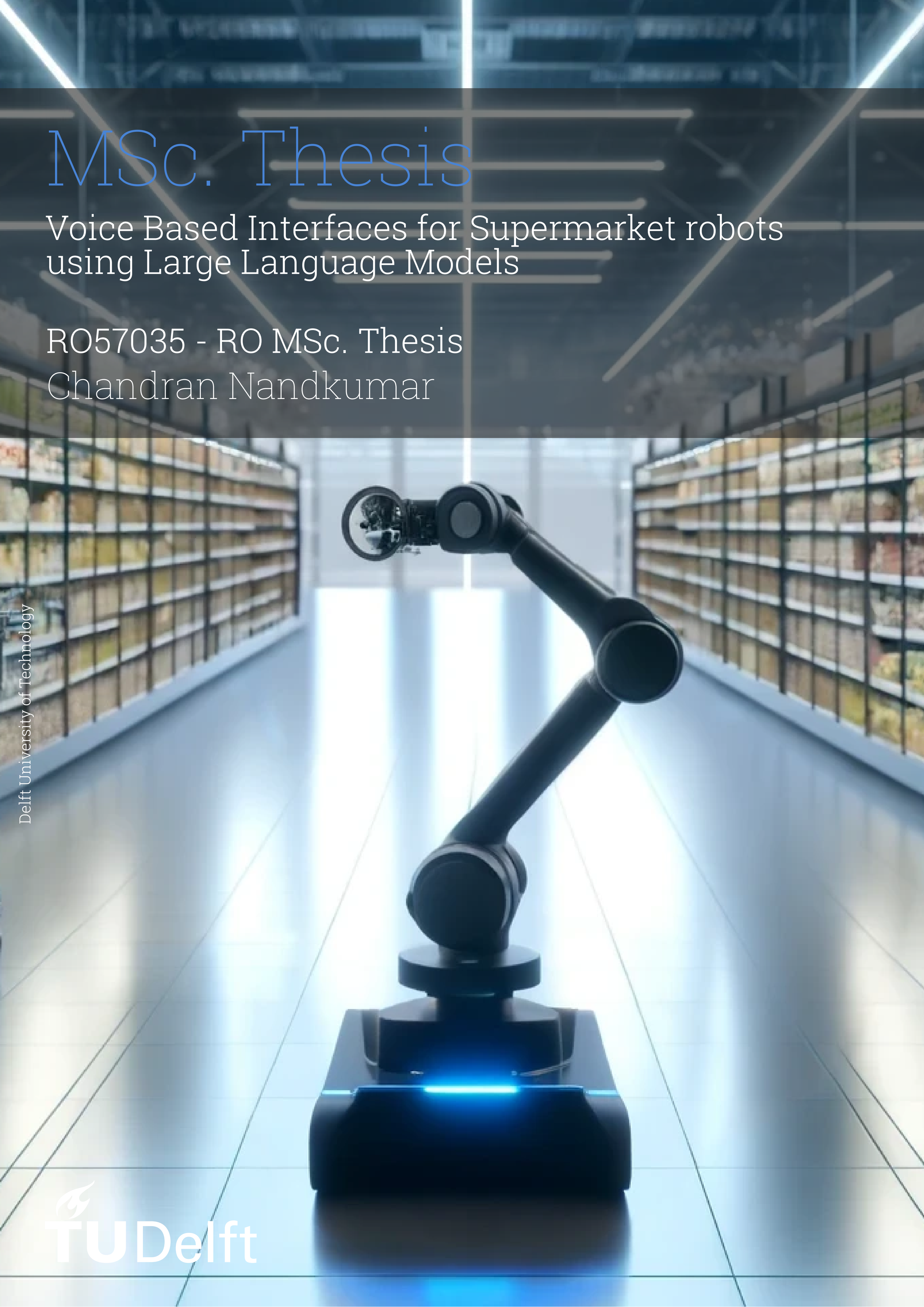
MSc. Thesis

Voice Based Interfaces for Supermarket robots using Large Language Models

RO57035 - RO MSc. Thesis

Chandran Nandkumar

Delft University of Technology



MSc. Thesis

Voice Based Interfaces for Supermarket robots using Large Language Models

by

Chandran Nandkumar

Student Name	Student Number
Chandran Nandkumar	5692520

Supervisor:	Dr. ir. L. Peternel
Committee Members:	Dr. ir. L. Peternel, Dr. ir. J.C.F. (Joost) de Winter, Dr. ir. Sole Pera (external member)
Project Duration:	October, 2023 - April, 2024
Faculty:	Cognitive Robotics, Mechanical Engineering

Cover: Generated using DALLE3
Style: TU Delft Report Style, with modifications by Daan Zwaneveld

Preface

As robots become increasingly prevalent in our lives, it is crucial to build effective yet highly intuitive interfaces to facilitate seamless communication between humans and machines. While there are multiple modalities these interfaces can stimulate and take advantage of, voice is of particular interest due to its nature of being the default mode of communication between humans. The advent of Large Language Models, a class of deep learning models that have shown great aptitude in handling simple to complex language instructions, provides the perfect base to facilitate such interfaces. The objective of this thesis is to build an effective voice-based interface for a rather complex environment - a supermarket. The three main pillars we wish to pay attention to with this system are *inclusivity*, *personalisation*, and *task effectiveness*. Multiple human-factors experiments have been performed to compare and evaluate the different components of such interfaces on these three features.

This thesis is submitted in partial fulfilment for the requirements of my Master's degree in Robotics at the Delft University of Technology. The research presented herein was conducted under the supervision of Prof. Dr. ir. Luka Peternel of the Department of Cognitive Robotics at the Faculty of Mechanical Engineering.

Chandran Nandkumar
Delft, April 2024

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Dr. ir. Luka Peternel, for his enthusiasm, motivation, and unwavering support throughout this journey. His frequent feedback was truly invaluable. I am also grateful to AIRLab for providing the simulation environment and resources necessary to test my approach. Special thanks go to Dr. Chris Pek for his insightful feedback on the current limitations and challenges within the field.

My appreciation also extends to all the instructors of the various courses I have taken, whose deep insights have broadened my knowledge in areas previously unknown to me. I am equally thankful to my teammates in the personalisation experiment, who played a crucial role in some of the participant recruitment for that study.

Furthermore, I owe a special thanks to my incredible friends and family, whose unconditional love and support have been my anchor. The unforgettable moments we have shared together have enriched my life immeasurably. I am also grateful to X TU Delft and Exhale for providing wonderful escapes that offered much-needed respite from my studies and research.

Contents

Preface i

Summary ii

1 Scientific Article 1

2 Appendices 20

A Participant Instructions 21
- A.1 Speech Recognition Experiment . . . 21
- A.2 Personalisation Experiment . . . 21
- A.3 Interaction With Novel Multi-LLM Agent vs State-Of-The-Art GPT . . . 22

B Script for evaluation of Speech Recognition Systems 23
- B.1 English Version . . . 23
- B.2 Dutch Version . . . 23

C Informed Consent - Speech Recognition Experiment 25

D Normality Tests, Pairwise t-tests Of All Groups For All Models 28
- D.1 Dutch Female . . . 29
- D.2 English Female . . . 29
- D.3 Dutch Male . . . 29
- D.4 English Male . . . 29

E Effect Of Personalisation And Results 31
  - E.0.1 Components of the basic conversational agent . . . 31
  - E.0.2 Experiment setup and methodology . . . 32
  - E.0.3 Results of the evaluation of our proposed personalisation of responses vs default LLM . . . 33

F Relevant ASAQ Questions And The Relevant Criteria 34

G Prompts provided to all the LLMs 36
- G.1 High-Level LLM . . . 36
- G.2 Medium-Level LLM . . . 36
- G.3 Low-Level LLM . . . 37
- G.4 Robot Destination Extractor LLM . . . 38
- G.5 Supermarket Data Creation LLM . . . 39

H Demonstration Of The Interaction With The Multi-LLM Chatbot 40

I Design And Creation Of The State-Of-The-Art GPT 44

J Informed Consent - Conversational Agent Evaluation Experiment 46

1

Scientific Article

Towards Inclusive, Personalised and Effective Voice-based Interfaces for Supermarket Robots Using Large Language Models

Chandran Nandkumar
Supervised by: Dr. ir. Luka Peternel
Cognitive Robotics, Delft University of Technology, Delft, Netherlands
e-mail: c.nandkumar@student.tudelft.nl

***Abstract*—This thesis presents the design and evaluation of a comprehensive system for developing voice-based interfaces to support users in supermarkets. These interfaces enable customers to convey their needs across both generic and specific queries. While current state-of-the-art systems like GPTs by OpenAI are easily accessible and adaptable, featuring low-code deployment with options for functional integration, they still face challenges such as increased response times and limitations in strategic control for tailored use-case and cost optimisation. Motivated by the goal of crafting inclusive, personalised, and efficient conversational agents, this study advances on three fronts: 1) a comparative analysis of four popular off-the-shelf speech recognition technologies to identify the most accurate model for different genders (male/female) and languages (English/Dutch); 2) an assessment of the effects of personalised recommendations versus generic responses, using a blindfolded, counterbalanced within-subject experiment; and 3) the development and evaluation of a novel multi-LLM supermarket chatbot framework, comparing its performance with a specialized GPT model powered by the GPT-4 Turbo, using the Artificial Social Agent Questionnaire (ASAQ) in a counterbalanced within-subjects experiment and qualitative participant feedback. Our findings reveal that OpenAI's Whisper leads in speech recognition accuracy across genders and languages, users significantly prefer personalised chatbots over the non-personalised counterparts and that our proposed multi-LLM chatbot architecture outperformed the benchmarked GPT model across all 13 measured criteria, including statistically significant improvements in four key areas: performance, user satisfaction, user-agent partnership, and self-image enhancement. The thesis concludes by presenting a simple method for supermarket robot navigation by mapping the final chatbot response to correct shelf numbers towards which the robot can plan sequential visits. This later enables effective use of low-level perception, motion planning, and control capabilities for product retrieval and collection. We hope this work encourages more efforts into using multiple, specialised smaller models instead of always relying on a single powerful model.**

***Index Terms*—Voice interfaces, Robotics, LLMs, Speech Recognition**

I. INTRODUCTION

In recent times, the presence of robots in our daily lives has increased drastically and they are now capable of working side-by-side with humans to achieve a given objective. The paper in [1] explains how collaborative robots improve task efficiency, reduce training times for operators and promise greater safety than their autonomous robot counterparts. Since collaborative robots are a vast and growing field in robotics [2], multiple works address the need for different approaches to provide efficient, immersive and aware control. The study in [3] makes a strong argument for the need to implement intuitive user interfaces, which help reduce operation time and operator errors whilst maintaining situational awareness and user engagement.

There are multiple options available to interact with collaborative robots. To furnish some examples, [4], [5], and [6] show different implementations of robot collaboration using vision for a variety of applications like pick-and-place to welding; [7] and [8] presents the implementation of Augmented Reality for human-robot collaborative surface treatment and task-level authoring respectively whilst [9] and [10] present the use case of Virtual Reality for the control of robotic manipulators and mobile robots. There are various other means of controlling a robot like eye tracking, pose determination, haptics, facial expressions and more. Furthermore, it is also possible to use multiple such interfaces simultaneously to get more precise and accurate results as seen in [11] and [12].

One interface of particular interest for this work is voice. Voice-based interaction being the default means of communication between humans, holds great promise in being applied to robots. It can allow for more authentic conversations and communication of user intent as opposed to other means and also benefits from being hands-free allowing human users to manipulate other objects in their environment. However, due to previous limitations in both speech recognition and language processing, most of the older implementations of voice-based command of robots were rather primitive as they allowed only a few distinct and restricted voice commands as seen in [13] and [14] preventing robots from achieving natural language voice interaction with human users. Fortunately, these problems have been largely mitigated with the advancements made in Large Language Models that are capable of processing statements and requests in different languages and complexities in a robust manner.

Despite these advancements, conversational agents employing voice still face significant limitations – providing avenues for novel exploration. The first avenue is regarding inclusivity - the need to ensure that all users regardless of gender, language, race, age and other individual differences can use and interact with such systems the same way. In conversational agents, inclusivity is determined by the robustness of the voice-to-text conversion, facilitated via Automatic Speech Recognition (ASR) systems. ASR has been an important area of research in the domain of natural language processing and computational linguistics, with the promise of playing a crucial role in bettering human-machine interaction [15]. ASR systems are built to transcribe a given intelligent auditory signal into its linguistic textual counterpart and differ from speech understanding – ASR by itself cannot

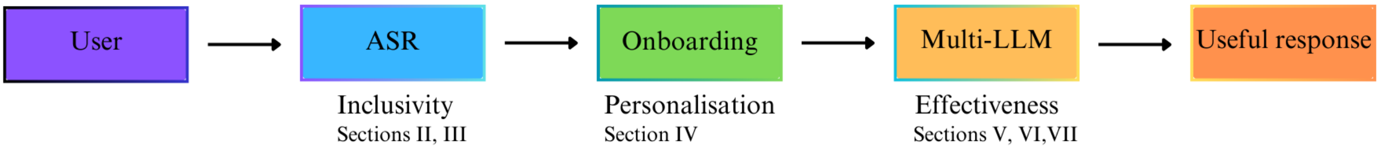


Fig. 1: The high level breakdown of the components based on the specific aspect (inclusivity, personalisation and effectiveness) for a chatbot. The sections that dive deeper into these respective domains is also provided for easy navigation.

operate or extract information from the signal [16]. Over the past few years, it has become relatively easy for individuals to transcribe their voice using many free or low-cost ASR systems released by popular organizations such as Google, OpenAI, Microsoft, Meta and more. While each of these models offers its own advantages and disadvantages, it is crucial to measure how robust these systems are to the voices of different users and provide recommendations on the ideal ASR system for human-robot interaction. Since the text generated by the ASR system will be subsequently processed by a Large Language Model the system selected must be as fair and inclusive as possible for the target users whilst minimising errors in transcription.

The second avenue we explore is on the perceived benefit of personalising recommendations provided by Large Language Models for a given user request. By analysing previous works on how retail salespeople categorise customers [17], we propose a simple approach comprising 6 quantitative and 2 qualitative questions to gain important information about the user's preferences during onboarding. Based on inputs from 30 participants, we compare the control conversational agent against one which utilises personal information to evaluate whether people prefer the personalised approach.

The final contribution of this work lies in trying to improve upon the limitations of the current state-of-the-art. To achieve this we propose a novel multi-LLM hierarchical conversational agent capable of responding to all kinds of users queries in a friendly manner. This system is evaluated against the present state-of-the-art GPT [18] created with the same data and information provided to our approach. The goal here is to compare how our approach performs against the state-of-the-art on the ASAQ - a popular questionnaire to evaluate artificial social agents [19].

A. *Problem Statement*

Despite voice-based interfaces having significant merits for human-robot interaction, there are a number of concerns that must be addressed if one intends to build an inclusive, personalised and highly effective conversational agent. The three problems our work attempts to resolve are of biased and non-inclusive speech recognition systems, lack of personalisation of conversational agents and improving on the effectiveness of these agents compared to the state-of-the-art.

Firstly, the selection of the right speech recognition system based on the target application and demographic is essential to ensure that the conversational agent can handle variations in language, gender, age and accent robustly and uniformly for all speakers. This however has not been the case for most popular speech recognition systems that have been trained largely on highly biased datasets making them considerably erratic with certain demographics. The findings of [20], a novel approach on automated fairness testing of speech recognition systems, indicate that 'non-native English, female and Nigerian English speakers generate 109%, 528.5% and 156.9% more errors, on average than native English, male and UK Midlands speakers, respectively.'. Furthermore, [21] states that female and non-US nationalities experience significant performance degradation when using automated speech recognition systems. Also [22] found that five popular ASR systems by Apple, Amazon, Google, Microsoft and IBM exhibited substantial racial disparities, with an average Word Error Rate (WER) of 0.35 for black speakers compared with 0.19 for white speakers. However, these studies cover a large scope of conversations and have primarily focused on English and other popular languages with little research being done on the performance of the Dutch language. This presents an opportunity to investigate the performance of popular ASR systems for a specific application and performance on transcribing Dutch speakers.

Secondly, the deployment of chatbots is a significant step and investment for the specific company intending to offer their services to clients. It is thus in the best interest of these businesses that these conversational agents be well-received and frequently used by the customers. As Large Language Models (LLMs) slowly integrate into the core of such systems, it is important to realise that generic, non-personalised responses may hinder active adoption and prolonged usage [23]. Works such as [24] argue that one of the main demerits of chatbots is the lack of understanding by developers and designers of user motivations and needs - which in turn affects their usability and adoption. Furthermore, from a psychological standpoint, several factors impact the quality and quantity of a conversation between a chatbot and a human such as the personality traits of the user like emotions, drive, thoughts, and assumptions about others' actions and beliefs. However, despite these benefits to personalisation of conversational agents, the process is not trivial, especially in environments like supermarkets where the degree of variability in users is extremely high. This makes gaining insights into the customers' preferences and interests a difficult problem at scale.

Lastly, the deployed chatbots must be capable of performing the task as accurately as possible by understanding the user intent and providing reliable and grounded responses. The variability in the types of requests in terms of complexity and degree of language processing required implies that a supermarket chatbot must be robust enough to handle both straightforward queries such as asking a specific item's availability, position and price to significantly more open-ended and broader queries regarding high-level intents such as recommendations for a specific dinner or items required for a party. Chatbots built by LLMs are also prone to significant hallucinations and mistakes which influence the degree of trust users can place in these systems [25]. Furthermore, the latency of such systems is often extremely high, affecting their degree of usability. This presents an opportunity to invent a new approach capable of resolving as many problems as possible from above.

B. Research Questions and Objectives

The primary aim of this thesis is to answer the following research question -

How can voice-based supermarket robot interfaces be designed to enhance inclusivity, personalisation, and usability, as evidenced by evaluations of ASR robustness, personalised onboarding, and multi-LLM effectiveness compared to the state-of-the-art?

This research question can thus be broken down into three sub-research questions -

1) Which off-the-shelf speech recognition system emerges as the most robust to variation in speaker gender (male/female) and language (English/Dutch)?
2) Does the proposed method of collecting user information improve the likelihood of the personalised chatbot being picked over the control in a blindfold study?
3) How does our novel multi-LLM conversational agent fare on the Artificial Social Agent Questionnaire (ASAQ) against a custom built state-of-the-art GPT 4 Turbo powered agent in a human factors experiment?

The objective of this thesis is to present a novel approach to addressing the width and depth of user queries in a supermarket by providing the most useful responses based on the user's main intention. The final conversational agent can also be integrated with a mobile-based robot to enable the robot to navigate to the correct shelves before the low-level control of a manipulator, perception and object retrieval can be incorporated to get the object from the shelf.

The paper is structured as follows. Section II will cover all the relevant systems, technologies and performance metric for evaluating speech recognition systems. Section III will go over the speech recognition experiment involving four off-the-shelf ASR systems including experiment setup, evaluation and results. Section IV will explore the motivation and justification behind our approach of enabling personalisation in the supermarket agent and the results of the counter-balanced within-subjects experiment to validate its perceived effect. Section V covers the related technologies, models, LLM tuning and improvement techniques and evaluation questionnaire to set the stage for our novel multi-LLM system. Section VI covers the specific details motivating and supporting the design of our proposed approach. Section VII presents the evaluation of our multi-LLM system against the state-of-the-art GPT using the ASAQ. Section VIII goes over how the proposed system is integrated with a robot for high-level path planning to the relevant shelves in the supermarket. Section IX ties all the previous sections together by discussing the main findings and implications of our complete study. Lastly, Section X serves as the conclusion. The breakdown of the main components of the chatbot and flow of the paper is illustrated in Figure 1.

II. METHODS - SPEECH RECOGNITION

Before we dive into the speech recognition experiment, we introduce the 4 models, relevant metric and underlying technologies to set the stage for our contributions. The Automatic Speech Recognition (ASR) system is the first component of voice-based interfaces as it enables the conversion of speech into text which can then be processed for further downstream tasks. The selection of the right ASR system is essential as it is the most important component in the conversational agent for ensuring inclusivity. Speech recognition systems that are biased tend to perform worse for certain demographics making the usage of such systems difficult for these groups.

A. Different ASR systems

Whilst there are many different speech recognition systems offered by multiple providers, this thesis will explore 4 prominent ones. They are selected based on their relevance, ability to transcribe Dutch and general popularity. We pick 2 open source systems (Vosk and Whisper by OpenAI) and 2 closed source systems (Google Cloud Speech-to-text and Microsoft Azure speech-to-text). This provides a good balance in terms of the different capabilities of the system such as online vs offline use, free vs paid, ability to fine-tune locally for specific use cases, data privacy and security and so on.

1) Vosk

Vosk is an open-source speech recognition toolkit that has made its mark in offline voice recognition and remains relevant despite being older than it's counterparts [26]. Unlike many cloud-based solutions, Vosk operates entirely on the device, ensuring data privacy and enabling voice recognition even in the absence of an internet connection. Underlying Vosk's capabilities is the Kaldi engine whose origin dates back to 2009. Developed in C++, Kaldi is an open-source speech recognition toolkit that can be effortlessly deployed across multiple operating systems. While initial versions of Kaldi primarily supported the English language, over time it has grown to support over 20 languages including Dutch.

Vosk itself, also written in C++, offers bindings for a plethora of programming languages, including Python, Java, and Node.js, making it remarkably versatile in its applications. From smaller devices like the Raspberry Pi to extensive server configurations, Vosk can be tailored to a wide array of scenarios. It offers pre-trained models for various languages, giving developers a significant head-start to implement them for speech recognition applications. Additionally, the toolkit supports custom model training, allowing for its deployment in niche contexts or for less common languages. Vosk has found its utility in diverse applications, such as voice assistants, transcription services, and voice-driven gaming. Its notable strengths encompass its adaptability, support of multiple languages, portability, offline functionality, and open-source support [27]. Its limitations include higher resource consumption due to offline use, lower accuracy than more advanced proprietary models and reduced general functionality. Furthermore, Vosk has different models for different languages requiring a language classifier to be present at the input to select the right model for transcription.

2) Whisper

Whisper by OpenAI is a cutting-edge ASR system, remarkable for its training on an expansive 680,000 hours of multi-task, multilingual data harvested from the web [28]. This enormous and diverse training foundation endows Whisper with notable robustness against various challenges, be it accents, technical jargon, or ambient noise. It is also capable of handling multiple languages and translation between them as well - a feat made possible since about a third of the data is in other languages.

Whisper's architecture is based on an end-to-end Encoder-Decoder Transformer model. Audio inputs are first segmented into 30-second blocks which undergoes preprocessing before being fed into the encoder. The decoder, in turn, is meticulously trained to predict corresponding text captions. One key standout feature of Whisper is the incorporation of specialized tokens that instruct the model to undertake varied tasks – from identifying languages and time-stamping phrases to multilingual transcription and translation into English.

While most ASR models lean on smaller, tightly coupled audio-text datasets or unsupervised audio pre-training, Whisper is trained on a broad and varied training set. Thus, although it might not outshine models on niche benchmarks like LibriSpeech, it is capable of unparalleled robustness across a myriad of diverse datasets reducing Whisper's error rate by half compared to other more specialized models.

3) Google Cloud Speech-to-Text

Google Cloud Speech-to-Text API is a ASR system developed by Google [29]. It is part of the larger suite of Google Cloud services and is designed to convert audio to text with high accuracy and efficiency. The API leverages Google's advanced machine-learning models and is capable of recognizing over 125 languages and their variants. It can be used in real-time applications or to transcribe pre-recorded audio files.

One of the key merits of the Google Cloud Speech-to-Text API is its high accuracy, even in noisy environments and with different accents. It also provides real-time transcription, which is crucial for applications like voice assistants and real-time captioning. Additionally, the API offers features like speaker identification, enabling different speakers in a conversation to be recognised, and word-level confidence scores, which can be used to identify uncertain parts of the transcription. Previous work has also shown that Google has the lowest WER compared to Microsoft Azure Speech and CMU Sphinx, another popular albeit less effective open-source ASR system [30].

4) Microsoft Azure Speech

Microsoft Azure Speech is a proprietary speech recognition and transcription service by Microsoft [31]. Microsoft has continued to develop powerful speech APIs for many years and has released a series of increasingly powerful speech platforms. Microsoft has in the past used context-dependent deep neural network hidden Markov model (CD-DNN-HMM). These CD-DNN-HMM models were able to achieve substantially better results than a Context-Dependent Gaussian Mixture Model Hidden Markov model (CD-GMM-HMM). In 2016, Microsoft also announced they had achieved human parity in speech recognition as published in the paper 'Achieving Human Parity in Conversational Speech Recognition by using various convolutional/LSTM acoustic model architectures, novel spatial smoothing methods, lattice-free MMI acoustic training, multiple recurrent neural network language modelling approaches, and systematic use of the system combination to even beat professional transcribers and set new benchmarks [32].

These 4 ASR systems discussed above are extremely popular and used by multiple services for different applications. They can also be integrated into systems via a simple API call making them ideal for robot voice interfaces. Now that we have introduced the 4 models that will be evaluated, we focus our attention to the metric to evaluate their performance.

B. Evaluation of different ASR systems

To evaluate the different ASR systems we use the Word Error Rate, a popular and simple metric to assess the accuracy of transcription by comparing the output of the system with the ground truth.

1) Word Error Rate

Word Error Rate (WER) is a popular metric used to evaluate the performance of automatic speech recognition and machine translation systems [33]. It measures the difference between the words in a reference transcription and the words in the system's output in terms of substitutions, insertions, and deletions needed to make the two match. These are the three different errors which could be introduced in transcribing and the WER helps us understand the ratio of these errors over the total number of input words expressed as a percentage. So the smaller the WER, the better the speech recognition system is at transcribing the spoken text [33].

The equation for WER can thus be written as:

$$\text{WER} = \frac{I + D + S}{N} \tag{1}$$

Where:

- I is the number of insertions.
- D is the number of deletions.
- S is the number of substitutions.
- H is the number of words present in the reference text

The WER however suffers from some limitations such as not being D/I symmetric i.e. it gives more importance to insertions than deletions when both of them are equally disadvantageous. Furthermore, it is not bounded and thus can exceed 100%. Although other metrics such as Word Information Lost and Match Error Rate have been proposed to counter these limitations, the WER remains the default method used by the speech recognition community to measure the performance of their systems due to it's simplicity and ease-of-understanding. In light of this, to maintain consistency with previous literature, we proceed with using WER.

III. EVALUATION OF SPEECH RECOGNITION SYSTEMS

After discussing the models and evaluation metric used to measure ASR performance, we proceed with the experiment setup. Given the limited research that has gone into task specific speech transcription and evaluation of the Dutch language, we proceed to perform a human factors experiment of 40 participants to pick the system that offers the highest accuracy despite variations in speaker gender and spoken language.

A. Participants

The 40 participants were divided into the following 4 equal and exclusive groups:

1) Dutch Female (**DF**)
2) Dutch Male (**DM**)
3) English Female (**EF**)
4) English Male (**EM**)

The groups were made exclusive to remove the potential influence of bilingualism since all Dutch speakers were proficient in English but the converse was not true. The experiment was approved by the Human Research Ethics team at Delft University of Technology. Participants were recruited from common public areas around the campus and word of mouth. Efforts were made to ensure the participants were from different nationalities and had diverse accents to make the study sample representational of the typical customers who would visit a supermarket. Participants were asked to sign the informed consent form. Before participants were asked to speak the lines of the script, an informed consent form was shown to clearly explain the format of data collection and privacy (Appendix C).

B. *Experiment Design*

All participants were asked to read from either a given English or Dutch script. The script was custom-generated based on commonly used words in the supermarket including product names, locations and other pieces of information including words that are sometimes difficult to pronounce. The script featured a conversation between a customer and a helpful assistant (Appendix B). The audio was recorded using a FiFine USB microphone K669B and using the open-source tool Audacity on a Linux system in the MP3 format. The same audio file was then converted into the WAV format due to it being the versatile format accepted by all systems and used by the 4 speech recognition systems using the speech recognition library in Python - Microsoft Azure speech-to-text (model base), Google Cloud speech-to-text (model V1 - default), OpenAI Whisper (model large-V2) and Vosk (English - vosk-model-en-us-0.22-lgraph, Dutch - vosk-model-small-nl-0.22). For Azure, Google Cloud and Vosk the language the user was speaking had to be specified while Whisper was capable of recognising the language from the audio file alone. After the transcriptions were created by the models, they were saved as txt files and pre-processing was done to convert everything to lower case and remove punctuation marks both in the reference script and the transcriptions. This was done because Google Cloud speech-to-text returned its output in lower case and penalising the absence of punctuation marks was not deemed necessary since the transcribed speech would be fed to an LLM - a system robust enough to handle such omissions. Using the reference script and created transcriptions, 4 Word Error Rates were calculated - one for each system.

The dependent variable in the experiment was the Word Error Rate of the transcribed text against the original script. The two independent variables are the speech recognition system - a within-subjects variable since the audio file of all participants was fed to all four systems and the group the participant belonged to based on their gender and language of the script. Since the 4 groups (DM, DF, EM, EF) were mutually exclusive, the group variable is a between-subjects factor. Thus, for the statistical analysis of these results, we first confirm the normality of all 16 columns of data (4 models x 4 groups) using the Shapiro-Wilk test. When normality of data was confirmed, a two-way mixed ANOVA was conducted. The threshold for statistical significance was set to 0.05.

C. *Results of the evaluation of Speech Recognition Systems*

We now outline the results obtained by comparing the Word Error Rate across the 4 speech recognition systems. Table I shows the results of the 2 way mixed ANOVA using model and group as the independent variables. We observe that the p-value of both group and gender is below 0.05 and thus statistically significant indicating that we have considerable differences in error rate on account of both the models and groups. Whilst this shows us that the group the participant belonged to affects the dependent variable, we do not get insights into which factor - language, gender or both is responsible for the variation in the performance of the models. To resolve this we perform two more mixed ANOVA tests where we drop each of the factors and test for only one of the factors against the model. Tables II and III take language and gender independently as between-subjects factors and maintain the model as between subjects. Based on this we observe that language has a significant influence on the WER while gender is not statistically significant. These insights are supported by the visualisation in Figure 2 where the difference between languages for participants of the same gender is lower than the differences between speakers of the same gender but different languages.

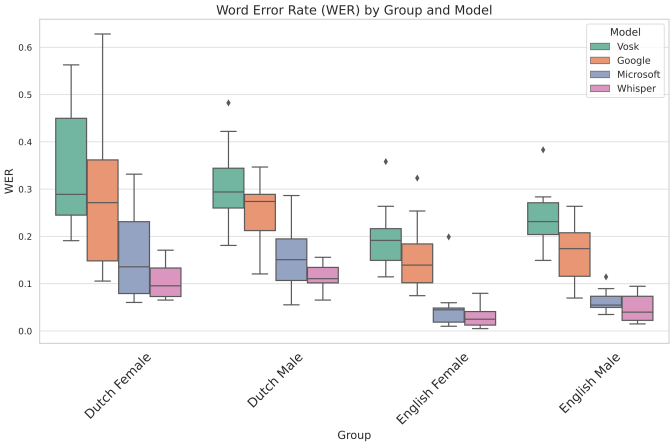


Fig. 2: Comparison of different ASR systems across all models and groups. We observe that Dutch participants have a higher WER than their English counterparts.

Figure 2 shows the performance of the different speech recognition systems across all groups based on Language and Gender. We observe from the box plots that the Word Error Rate of Whisper is the lowest of all 4 groups followed by Microsoft and Google whilst Vosk performs the worst for all 4 categories. The detailed Shapiro-Wilk test for normality, pairwise t-tests, and box plots of all pairwise comparison between the respective models is provided in Appendix D.

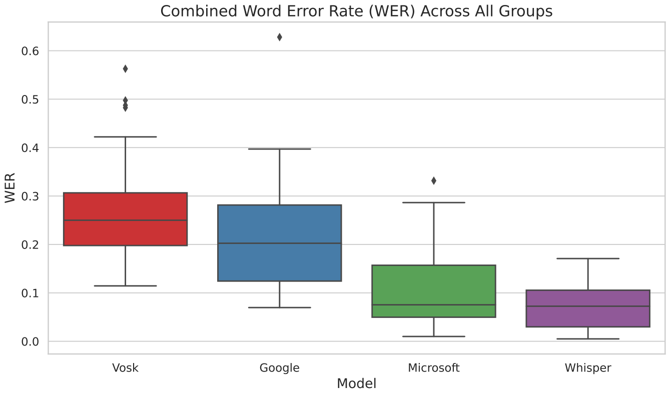


Fig. 3: Word Error Rate vs Speech Recognition System for all Participants. We see that Whisper has the lowest WER and variability compared to the other models.

Combining participants of all groups and analysing the Word Error Rate, we see in Figure 3 that Whisper has the least WER across followed by Microsoft and then Google. The pairwise t-test in Appendix D confirms these results as well by showing that Whisper is significantly better than the other 3 models when data

Source	SS	DF1	DF2	MS	F	p-unc	np2	eps
Group	0.373144	3	36	0.124381	7.350098	**5.772400e-04**	0.379848	-
Model	1.022612	3	108	0.340871	143.500414	**1.566141e-37**	0.799443	0.708692
Interaction	0.015847	9	108	0.001761	0.741261	0.6703523	0.058178	-

TABLE I: ANOVA Summary Table for Group versus Model. We observe that the effect of group has a statistically significant influence on the WER - but we do not have insights on which aspect of the group - gender, language or both are responsible for it.

Source	SS	DF1	DF2	MS	F	p-unc	np2	eps
Language	0.361852	1	38	0.361852	22.160251	**3.296411e-05**	0.368354	-
Model	1.022612	3	114	0.340871	147.210089	**4.732748e-39**	0.794828	0.708692
Interaction	0.008419	3	114	0.002806	1.211918	3.087008e-01	0.030907	-

TABLE II: ANOVA Summary Table for Language versus Model. We observe that the effect of language has a statistically significant influence on the WER.

Source	SS	DF1	DF2	MS	F	p-unc	np2	eps
Gender	0.000674	1	38	0.000674	0.026105	0.8725	0.000686	-
Model	1.022612	3	114	0.340871	143.958631	**1.295907e-38**	0.791161	0.708692
Interaction	0.002457	3	114	0.000819	0.345837	7.922205e-01	0.009019	-

TABLE III: ANOVA Summary Table for Gender versus Model. We observe that the effect of gender does not have a statistically significant influence on the WER

of all groups are used to evaluate the models. Based on these results, we conclude that based on the experiment conducted, Whisper is the most accurate and inclusive model due to its ability to handle variations in gender and language.

IV. PERSONALISATION OF SUPERMARKET CHATBOTS

After making efforts to ensure the selection of the right speech recognition system, we now focus our attention towards enabling personalisation in our conversational agent. In order to facilitate this for the supermarket scenario, we first need to determine what attributes of a customer are important and relevant, that can be obtained during their onboarding and how it can be implemented for the LLM-based agent.

A. *Classification of customers by retail salespeople*

To understand what information to capture from the user, we draw inspiration from [17] which presents the different categories into which retail salespeople cluster customers. We argue that since the conversational agent effectively replaces the salesperson in the supermarket, the same categorization can be used to create an effective user profile. The 6 quantitative parameters we capture on a Likert scale from 1-5 are:

1) *Price Consciousness* : the measure of how much the customer cares about cheaper products, substitutes, sales and discounts over premium products.
2) *Brand Loyalty and Value* : the measure of the customers' tendency to stick to certain premium product brands rather than allowing other parameters like price or size to affect their decision.
3) *Help Appreciation* : the degree to which the customer values recommendations and help from external sources in the supermarket.
4) *Degree of knowledge* : the degree to which the customer believes they are knowledgeable about the various products available in the store and thereby how specific or broad their requests maybe.
5) *New product exploration* : the degree of willingness to try out new products and offerings instead of continuing to rely on more standard and predictable patterns while shopping.

The 2 qualitative questions we ask users are:

1) *Dietary preferences* to understand any allergies, specific habits or principles followed.
2) *Product interest* which involves any specific information about any products or brands they prefer purchasing or exploring.

We believe that this information is key for the chatbot to provide personalised and useful recommendations to the user whilst also tailoring the conversation in a manner beneficial to the customer. These variables are stored in a database in the following format (as an example):

```
"user_id": 658ee33213436cdf9b310225,
"name": "Chandran Nandkumar",
"price_consciousness": 4,
"brand_loyalty": 2,
"help_appreciation": 4,
"degree_knowledge": 1,
"speed_shopping": 4,
"new_explore": 5,
"dietary_preferences": 'Vegan',
"product_interest": 'Nutritional diversity'
```

This information can then be retrieved when the customer scans their membership card and fed as context to an LLM to enable it to make more informed decisions and provide personalised recommendations. To evaluate whether or not this approach makes customers more likely to prefer the personalised chatbot over the control, we conduct a within-subjects counterbalanced experiment and perform the chi-squared test to evaluate whether the results are significant. The detailed explanation about the components, experiment and results are available in Appendix E. Overall we observe that 21 out of 30 participants prefer the personalised chatbot over the control and the chi-squared statistic of $\chi^2(1, N = 30) = 5, p = .025347$ justifies the benefits of personalisation over relying on generic LLM outputs.

V. METHODS - LANGUAGE MODELS AND LARGE LANGUAGE MODELS

After making strides towards making our agent inclusive and personalised, we move on to proposing a novel multi-LLM architecture in order to try and obtain faster, cheaper and more customisable solutions that could potentially rival the state-of-the-art - the custom GPTs by OpenAI [18]. We begin by introducing the popular models used in the study, the methods incorporated to tune and obtain the necessary results and the questionnaire used to evaluate our model against the state-of-the-art.

A. Relevant models

Here we discuss the popular models that are used as a part of our work. We also introduce the state-of-the-art - the GPTs.

1) Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language model developed by Google, with significant applications in the field of natural language processing. Unlike previous models that read text unidirectionally, BERT processes text in both directions simultaneously. This bidirectional approach allows the model to capture a more nuanced understanding of context, making it highly effective at understanding the meaning of each word in its textual environment. BERT is pre-trained on a large quantity of text from the internet, including the entire Wikipedia, which enables it to learn a wide range of language patterns and structures. Furthermore, it can be fine-tuned with additional data for specific tasks without substantial modifications to the underlying model. This versatility has led to its widespread application across a variety of NLP tasks such as text classification, question answering, sentiment analysis, and named entity recognition, revolutionising how machines understand human language [34]. Furthermore, works such as [35] have proposed methods of shrinking the model size by 40% whilst retaining 97% of the accuracy.

2) GPT 3 and GPT 3.5 Turbo

The GPT-3 and GPT-3.5 Turbo series, developed by OpenAI, includes a range of language models designed for generating text that closely mimics human language for chat applications. The foundational model in this series, known as davinci, is equipped with 175 billion parameters, showcasing its proficiency in text generation. OpenAI has enhanced the capabilities of davinci through two primary development paths. The first path involves supervised fine-tuning, leading to the creation of InstructGPT, also referred to as text-davinci-001. The second path focuses on training for coding tasks, resulting in the Codex model. The evolution continued with the release of code-davinci-002 in 2022, specifically designed for code generation, which laid the groundwork for the GPT-3.5 series.

Further advancements led to the development of text-davinci-002 through additional supervised fine-tuning, alongside the introduction of the RLHF (Reinforcement Learning from Human Feedback) [36] training strategy with text-davinci-003 [37]. This new strategy significantly enhanced the model's ability to interpret instructions and produce relevant text. Building on the success of text-davinci-003, OpenAI optimised a model specifically for chat applications, known as gpt-3.5-turbo. This model stands out for its high capability at a more affordable cost compared to its predecessor, text-davinci-003.

Although other open source models such as LLaMA 13b, Mixtral 8x7b and Mistral 7b exist, for this study we use GPT 3.5 Turbo since it is easy to use, is extremely reliable, has good documentation, allows for fine-tuning directly on their website with optimised hyperparameter setting and does not need any specific GPU for training and inference. The other evaluated models do not provide all of these merits under a single umbrella at the time of writing this paper.

3) GPTs - The state-of-the-art

In November, 2023, OpenAI presented a novel approach to LLM customisation by empowering users to tailor specific instances of ChatGPT for specific tasks. This feature enables the customisation of digital assistants for a wide range of applications, from learning and productivity to entertainment and scientific research, without requiring any coding skills. With an emphasis on the importance of community involvement in the development of GPTs, it highlights the potential for users to contribute to the diversity and capability of these tools. Furthermore, with the launch of the GPT Store, creators will have the opportunity to share their GPTs with a broader audience, potentially earning revenue based on usage. The GPTs come with in built functionality such as image generation using DALLE3, web browsing using Bing, code writing and execution, knowledge retrieval and even advanced function calling and customised actions. We argue that the versatility and access to powerful resources makes GPTs a powerful benchmark to evaluate even highly niche chatbots. Before the release of GPTs, chatbots did not have a standard one-size-fits-all state-of-the-art since they were extremely task specific. This innovation however enables a quick and powerful agent powered by an LLM to serve as a reference to compare and evaluate diverse chatbot technologies [18].

B. Popular methods used for implementing LLMs

After discussing the popular models used in this study, we go over the different methods and approaches implemented to improve the capabilities and tune our models for our required goals by drawing inspiration from related discoveries and innovations in the field of Large Language Models and information retrieval.

1) Chain-Of-Thought prompting and reasoning

Chain-of-Thought (CoT) prompting guides large language models through a series of intermediate steps or thoughts towards a solution, mirroring human problem-solving processes. This technique enhances the models' capacity for complex reasoning across tasks, from arithmetic to commonsense reasoning, by making their decision-making pathways more transparent and interpretable. CoT prompting has proven especially effective in improving performance without extensive task-specific fine-tuning, highlighting its utility in leveraging pre-trained models for a broader range of applications while maintaining or enhancing their accuracy and interoperability [38].

CoT is useful when directly reaching a particular answer is difficult. Akin to teaching a child how to solve a given problem, CoT provides a reliable framework to break the problem into parts and solve it sequentially and in a methodical fashion. Chain-Of-Thought can also make the underlying motivations and reasons behind LLMs choices more transparent and easier to tune based on the specific goals.

2) Fine-tuning LLMs

Fine-tuning in the context of large language models is the process whereby a pre-trained model is adjusted to perform a specific task better. This process involves taking a model that has been trained on a large, general dataset and further training it on

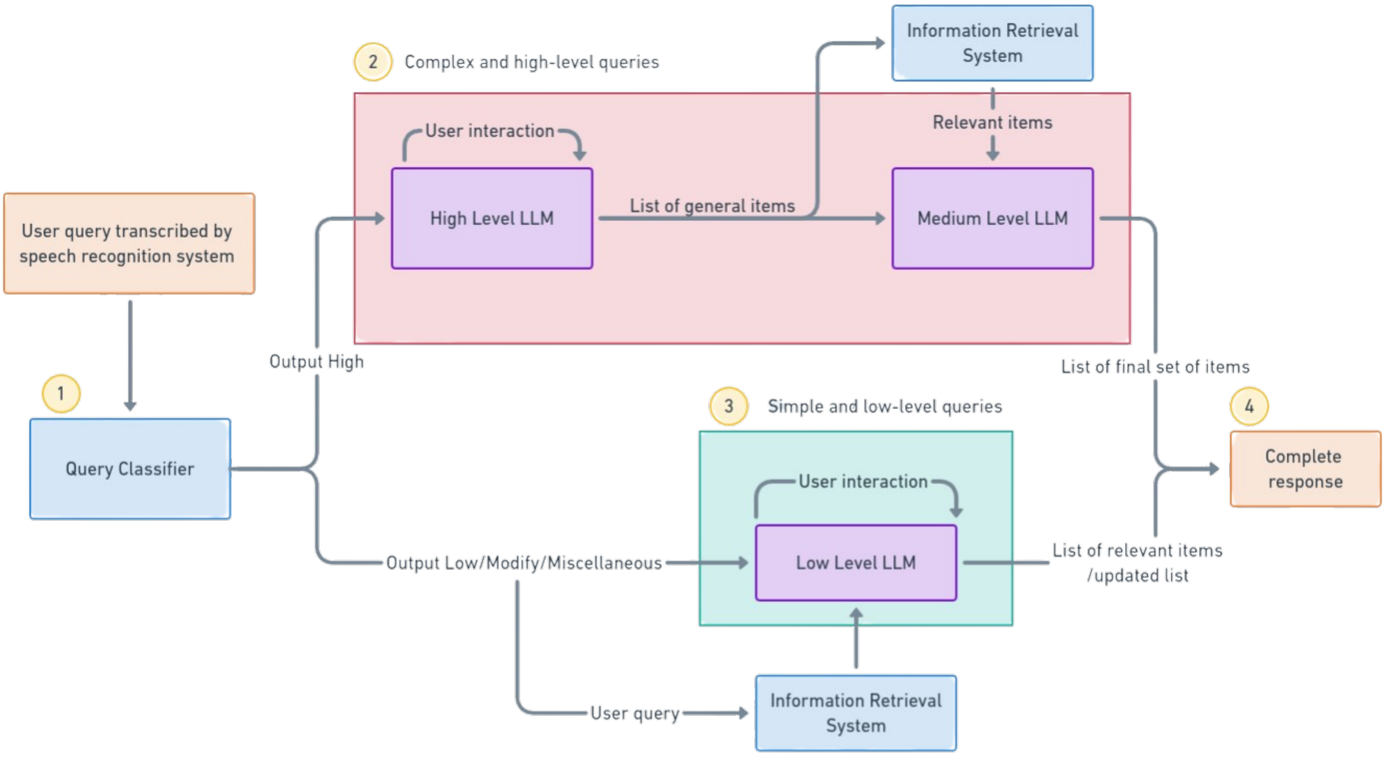

2
Complex and high-level queries
Information Retrieval System
Relevant items
User interaction
High Level LLM
List of general items
Medium Level LLM
User query transcribed by speech recognition system
Output High
List of final set of items
1
3
Simple and low-level queries
4
Query Classifier
User interaction
Complete response
Output Low/Modify/Miscellaneous
Low Level LLM
List of relevant items /updated list
User query
Information Retrieval System

Fig. 4: Proposed architecture for handling different queries. Once the query has been transcribed by the speech recognition system, it is classified by the mdistilBERT system (1). If the query is classified as a high-level query the high-level LLM asks further questions and prepares a rough list of items. These items are sent to the information retrieval system and the relevant items are sent to the medium-level LLM that prepares the correct list of items (2). Otherwise, the query is directly converted to an embedding and searched by the IR system to provide the necessary list of items to the user (3). The relevant response (4) is then shown to the user for further modifications or approval.

a smaller, task-specific dataset. The rationale behind fine-tuning is that the pre-trained model has already learned a vast amount of general knowledge about the language, and fine-tuning allows it to adapt this knowledge to the requirements of a particular domain. This is achieved by adjusting the model's weights based on the task-specific data, often involving a lower learning rate to make small, incremental changes that refine the model's abilities without overwriting its pre-existing knowledge. Fine-tuning is critical in a wide variety of tasks such as sentiment analysis, question-answering, and text classification, enabling models to achieve high performance with relatively less task-specific data.

Parameter-Efficient Fine-Tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA), and its quantised variant, q-LoRA, represent advanced strategies to reduce the computational and memory burden associated with fine-tuning LLMs. These techniques focus on modifying a small subset of the model's parameters or introducing additional parameters that can learn task-specific features without altering the entire model. LoRA, for example, introduces low-rank matrices that interact with the pre-trained weights to adapt the model's output without directly modifying the original weights. This allows for efficient adaptation to new tasks while keeping the majority of the model fixed, significantly reducing the required memory and computational resources. q-LoRA extends this by applying quantisation to the adaptation process, further decreasing the computational load and storage requirements. These methods exemplify the shift towards making fine-tuning more accessible for a wider range of applications, especially in resource-constrained environments [39].

Fine-tuning enables the application of LLMs to a broad spectrum of tasks while leveraging their pre-trained general knowledge, thus bypassing the need for training large models from scratch for every new task. This adaptability significantly lowers the barrier to entry for deploying state-of-the-art models in specialised domains. Moreover, fine-tuning can lead to models that are not just more efficient but also more accurate, as they can be tailored to the peculiarities of a specific task or dataset. Fine-tuning can also be employed to teach models specific formats of output or to respond in a particular manner making the results more deterministic than their pre-trained predecessors.

3) Retrieval Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a technique used to integrate the capabilities of large language models with external knowledge bases to allow the model to access task specific information and generate informed and accurate responses. Without RAG, models are restricted by the knowledge they learnt during pre-training. Any information provided after the cut-off date is unavailable to the model and questions regarding the same are often answered incorrectly with hallucinated answers. RAG involves retrieving relevant documents from a database based on the input query and using this information to guide the generation process, allowing the model to produce responses that are contextually relevant [40].

The primary benefits of this approach include the model's improved capacity to incorporate up-to-date information, a reduction in generating inaccurate or fabricated information (hallucinations), and the ability to access domain specific knowledge beyond its original training data. Hallucinations are reduced by grounding the LLM to respond only based on the information provided to it and prompting it to respond with 'I don't know' when the necessary information is not available [40].

4) Retrieval Augmented Fine Tuning (RAFT)

Retrieval Augmented Fine Tuning (RAFT) is a methodology that enables the integration of external knowledge into language models during their fine-tuning phase and it is shown to be a powerful approach to derive the best of both worlds for enhancing the capability of LLMs to understand and respond to queries within specific domains. RAFT works by initially retrieving a set of documents, D from a knowledge base relevant to a given query. These documents are selected based on their potential relevance to the query's context. Then, using Chain of Thought reasoning, the model evaluates these documents to identify a subset, D* that is most relevant to the asked query. This fine-tuning process involves teaching the model to differentiate and prioritise information from D* that significantly contributes to generating accurate and contextually appropriate responses. This approach enables the model to leverage external knowledge effectively, enhancing its capability to address complex, domain-specific inquiries.

The concept of RAFT is explained with a nice analogy in [41]. While RAG is akin to a student in an open book exam who has not prepared for it and fine-tuning is akin to a student in a closed book exam who has learnt the subject matter well, RAFT provides the alternative, of a case where a student has prepared for an open book exam and is able to effectively utilise the resources at their disposal during the test.

C. Evaluation Questionnaire

For the evaluation of our conversational agent, we use the Artificial Social Agent Questionnaire (ASAQ) [19]. The questionnaire was developed based on the need to create a validated, standardised measurement instrument dedicated to assessing human interaction with Artificial Social Agents (ASA). The ASAQ is the result of extensive collaboration over multiple years involving over one hundred ASA researchers globally and ensures a robust framework for evaluating interactions between humans and ASAs. The long version of the ASAQ provides an in-depth analysis of human-ASA interactions, catering to comprehensive evaluation needs. Conversely, the short version offers a swift means to analyse and summarise these interactions, facilitating quick insights into the user experience. Additionally, the instrument is complemented by an ASA chart, which serves as a visual tool for reporting results from the questionnaire and provides an overview of the agent's profile. Due to its breadth and comprehensiveness, the ASAQ measures 19 parameters - some of which are not relevant for our study. We go over all the 19 criteria and the 13 relevant ones used in our study featuring the short version of the ASAQ in Appendix F.

VI. DESIGN OF THE MULTI-LLM CONVERSATIONAL AGENT

Now that we have set the stage for the proposed solution, we will now cover the main requirements, design strategies and specific details of how we build a multi-LLM agent. A supermarket chatbot must be capable of retrieving relevant information from the supermarket database, answer user queries in a friendly and natural manner whilst ensuring it can handle a variety of user queries from simple requests asking details about a specific product (e.g., 'Where can I find Oreos and how much are they?') to complex high-level queries (eg. 'I am not sure of what to make for dinner, can you recommend some ideas and the necessary ingredients?'). This requires the conversational agent to not only be capable of the basic functions such as natural language understanding, dialog management and natural language generation, but also advanced reasoning and information retrieval.

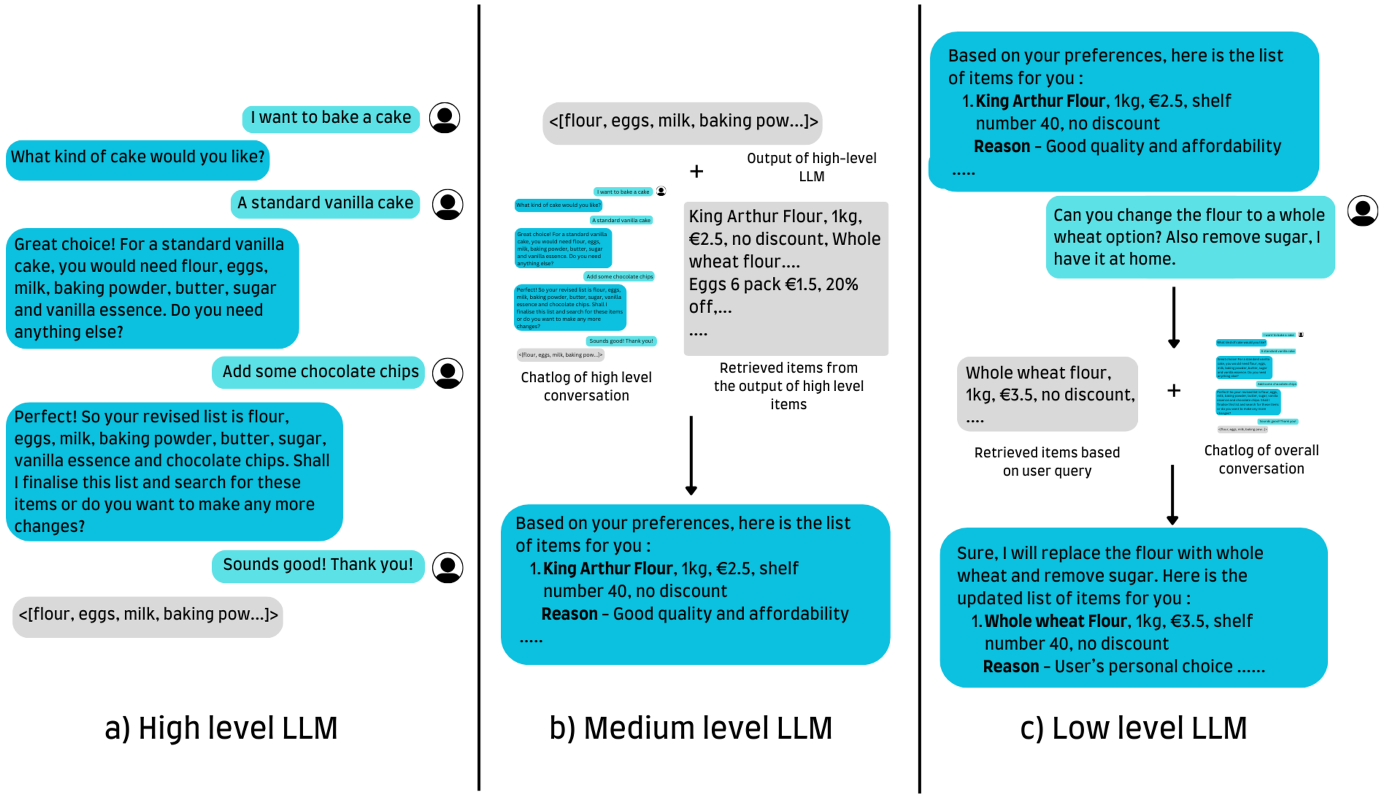
The current state-of-the-art for handling all these different responsibilities is the concept of GPTs [18]. However, GPT4 Turbo, the underlying model in GPTs, is not without its own set of limitations. Firstly, for applications such as chatbots, latency is an extremely important factor. Despite recent advances in speed, GPT4 is significantly slower in response generation compared to smaller and lighter models due to the overall size of the model which results in slower inference speed. Furthermore, research has also indicated that for extremely large context windows GPT4 tends to struggle with information retrieval. This form of evaluation - titled needle in a haystack is used to measure how well a model can retrieve information based on the position of the requested information in the overall context [42]. Due to the limitations in model architecture, the model struggles with correctly retrieving the information in the middle portions of the context when the context size reaches close to the limit. Furthermore, there are also claims that the reasoning window of powerful LLMs is much lower than their context window making the task of advanced reasoning over large amounts of information significantly challenging. Lastly, the model is significantly more expensive at the time of writing this paper at $10 per million tokens input and $30 per million tokens output than the cheaper GPT-3.5 Turbo priced at $0.5 per million tokens input and $1.5 per million tokens input. Even after fine-tuning the GPT 3.5 Turbo model costs $3 per million tokens input and $6 per million tokens output - a significant reduction over the GPT4 non-finetuned counterpart.

To address the problems of latency, information retrieval, reasoning window and price, we propose a novel multi-LLM conversational agent where many smaller LLMs, specialised for certain tasks and query types work together to give better results. For our implementation, we use multiple GPT-3.5-Turbo models in a hierarchical fashion where each model serves a specific purpose. Furthermore, the input query is classified using distilBERT into high-level or low-level queries allowing for a different strategy to be employed for different query complexity further optimising on computational resources and API costs. The architecture we employ is shown in Figure 4.

A. Query Classifier

The first step in our conversational agent is to take the input text obtained from the speech recognition system and classify it based on whether the query is high level, low level, modification or miscellaneous. The types of queries are explained below:

1) High-level queries are those that need to be broken down and analysed with the help of the user to ascertain their preferences, the particular occasion and other restrictions which can enable us to make more informed decisions.
2) A low-level query is a specific request of a particular product or class of products such as finding the location, price or alternatives to an option,



I want to bake a cake

What kind of cake would you like?

A standard vanilla cake

Great choice! For a standard vanilla cake, you would need flour, eggs, milk, baking powder, butter, sugar and vanilla essence. Do you need anything else?

Add some chocolate chips

Perfect! So your revised list is flour, eggs, milk, baking powder, butter, sugar, vanilla essence and chocolate chips. Shall I finalise this list and search for these items or do you want to make any more changes?

Sounds good! Thank you!

<[flour, eggs, milk, baking pow...]>

a) High level LLM

<[flour, eggs, milk, baking pow...]>

Output of high-level LLM

Chatlog of high level conversation

King Arthur Flour, 1kg, €2.5, no discount, Whole wheat flour....
Eggs 6 pack €1.5, 20% off,...
....

Retrieved items from the output of high level items

Based on your preferences, here is the list of items for you :
1. **King Arthur Flour**, 1kg, €2.5, shelf number 40, no discount
Reason - Good quality and affordability
.....

b) Medium level LLM

Based on your preferences, here is the list of items for you :
1. **King Arthur Flour**, 1kg, €2.5, shelf number 40, no discount
Reason - Good quality and affordability
.....

Can you change the flour to a whole wheat option? Also remove sugar, I have it at home.

Whole wheat flour, 1kg, €3.5, no discount,
....

Retrieved items based on user query

Chatlog of overall conversation

Sure, I will replace the flour with whole wheat and remove sugar. Here is the updated list of items for you :
1. **Whole wheat Flour**, 1kg, €3.5, shelf number 40, no discount
Reason - User's personal choice ......

c) Low level LLM

Fig. 5: A visual depiction of the responses of the 3 different LLMs. A high-level query takes a request and based on the user's input and user profile, creates a basic list of items. The medium-level LLM takes these items, the chatlog and retrieved items to craft a tailored response for the user. Lastly, all specific queries, modifications and other requests are passed to the low-level query capable of retrieving items and making changes to the original list.

	English				Dutch			
Metric	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Value	0.8679	0.8839	0.8679	0.8651	0.8679	0.8710	0.8679	0.8635

TABLE IV: Performance Metrics for English and Dutch Test Sets for query classification by mDistilBERT

3) A modification query is one where the customer wishes to make amendments to a previously displayed list.
4) A miscellaneous statement comprises of everything else such as conversational statements like 'Yes,please' and 'thank you'.

To classify these requests, however, we need a powerful natural language classifier that can be fine-tuned for the given task. The classifier we chose to proceed with is a condensed form of the Bidirectional Encoder Representations from Transformers (BERT). We used distilBERT [35] - a general-purpose model that reduces the size of conventional BERT by 40% while retaining 97% of the task performance and can be run locally on systems without dedicated GPUs. The model is freely available on HuggingFace and is easy to train and deploy. The query classifier is trained on over 150 examples - augmented by GPT4 by providing a few representational examples to the model. Furthermore, to also support Dutch, we use the multilingual version of distilBERT.

For our fine-tuning purposes, we were able to use anonymous logs of chatbot interactions in the previous personalisation experience along with GPT4 augmented data. In total we had 106 English statements, manually labelled from the previous chatlogs and 250 English queries were augmented by GPT4. The data augmentation was done on the ChatGPT interface to allow for better control of the diversity and nature of resultant statements. These 356 queries were translated to Dutch whilst respecting the conversational nature of the queries by GPT4. After shuffling the data, we split the final 712 queries into 500 training, 106 validation, 53 English test and 53 Dutch test sets. Before training the queries were converted lower case and punctuation marks were removed since we are using a cased distilBERT model.

The hyperparameters used are as follows -

1) Learning rate : 5e-5,
2) Number of epochs : 8,
3) Optimiser : AdamW,
4) Warmup steps : 10% of total steps

The final validation loss was 0.58324 and final validation accuracy was 0.8302 at the 8th epoch and was unchanged from the 7th epoch results. Table summarises the accuracy, recall, precision and F1 scores of the classifier after fine-tuning. The fine-tuned mDistilBERT classifier demonstrates a robust performance in classifying queries into four distinct classes: high, low, modify, and miscellaneous, with both English and Dutch test sets achieving comparable accuracy scores of 0.8679. This similarity in accuracy suggests that the model generalises well across languages, a testament to the multilingual capabilities of the underlying mDistilBERT architecture. Precision scores, slightly higher for the English set at 0.8839 compared to 0.8710 for Dutch, indicate a marginally better reliability in the model's positive predictions for English. The recall scores, identical for both languages, affirm the model's effectiveness in identifying relevant instances across the dataset. However, the F1 Score, which balances precision and recall, is slightly higher for the English test set (0.8651 vs. 0.8635 for Dutch), suggesting a modestly more balanced performance in English. Overall, these metrics reflect the classifier's proficient handling of varied linguistic queries, illustrating its practical utility in multilingual applications. The results are summarised in Table IV.

The slight differences between English and Dutch performances could offer insights into areas for further model optimisation, particularly in enhancing its cross-linguistic adaptability and understanding. While the performance may not seem remarkable, it is important to note that the mistakes made in classification are sometimes permissible. For example, in the English test set the classifer mislabeled 'Sure, add that to my cart.' as 'modify' instead of the ground truth label assigned of 'miscellaneous' which is a completely valid classification for the given query. Likewise, the dutch query, 'Ik moet mijn gebruikelijke ontbijtgranen vervangen door een optie met veel vezels, welke?' (translation - I need to replace my usual breakfast cereal with a high fiber option, which one?) was mis-classified as a high-level query when the ground truth label assigned was low - which is once again a permissible misclassification since there are multiple options for a high fiber breakfast (high level) but it can also be a low-level query (retrieve the high fiber cereal options). Thus, we argue the performance of the classifier is better in true application than the results indicate.

B. Product database

Due to the absence of a simple and relevant database comprising of different categories in a supermarket, the products, their prices, potential discounts and locations, a dataset was augmented using GPT4 with the ChatGPT interface. ChatGPT was used instead of completely automating the process via API call for lower costs and greater control over the number of products per category and toggling the price/discount if necessary. Overall, we had 100 different categories comprising standard grocery, personal care, home maintenance, tools, electronics, books and furniture to name a few. Each category had anywhere from 12 - 30 products created along with brand names, price, discounts and shelf numbers. Thus overall 1612 products were augmented by GPT4 which will be used as the dataset to demonstrate the functioning of our system. The same database will also be provided to the GPT made via OpenAI so that both approaches have access to the same ground truth. This also helps serve as a reference to evaluate the amount of hallucinations by the models since not grounding them could result in these systems making up completely new items that do not exist in the current inventory.

C. Information Retrieval

Next, we will discuss the process of information retrieval carried out in our approach. Firstly, we transform the original data into an inverted index where each product, its location, price and potential discounts occupies one row. After this, we use an embeddings model (text-embedding-3-small by OpenAI) to convert all 1612 products into n-dimensional vectors that are then stored as a numpy file. We then convert either the low level query or all elements of the medium level LLM response into an embedding via the same model and find the closest neighbours using cosine similarity. For low level queries we retrieve 20 closest products to allow for sufficient recall and for each element in the medium level LLM response we retrieve 3 closest products. Thus, if the medium tier LLM returns 7 items, we will retrieve 21 products based on the closeness of these items to their neighbours.

D. High-level LLM

If the query is classified as high-level, a high-level LLM is called to interact with the user in order to get more information and break down the query into a list of items the user may need. At this step, user preferences and choices are taken into account along with ascertaining what items the user would need versus that which they already possess or can be substituted. This is best explained with an example. Say you want to bake a cake. There are a number of ingredients you need such as milk, eggs, flour, baking powder, baking soda, vanilla essence, sugar etc. However, you may possess a lot of these items already at home. Additionally there are other ways to make a cake such as using a cake mix, buying a premade cake or deciding exactly what flavour and nutrition profile you wish to base it on.

The high-level LLM is tasked with ascertaining what kind of cake you want, if you have any preferences/allergies or other customisations needed along with understanding the exact list of ingredients you would need. The high level LLM is a GPT3.5 Turbo fine tuned on 36 multi-turn conversations inspired from previous chatlogs of users in the personalisation experiment and some extra conversations augmented using GPT 4 via ChatGPT. A validation dataset is also created comprising of 12 similar conversations. Overall the fine-tuned model after 3 epochs has a training loss of 0.6228 and an accuracy of 0.80365 while the validation loss is 1.0575 and validation accuracy is 0.56343. This indicates that the model has overfit on the training data however given the complex multi-turn nature of the task, even considerable deviations from the ground truth presented in the validation data is possible due to the likelihood of having multiple correct answers to a query based on user responses. The high training accuracy indicates that the elements we care about such as the format and the approach have been well learnt.

E. Medium-level LLM

Once the user is satisfied with this selection of items, the list of user-selected products, the chatlog of the user and the chatbot, the user profile and the retrieved items are sent to a medium-level LLM that is tasked with creating a tailored list of items from the context with the exact name, brand, price, location and reasoning behind the selection of the items. The medium-level LLM never interacts directly with the user. Based on the response of the medium-level LLM, the user can fine-tune their list of items by making any final changes to the products using the low-level LLM.

The fine-tuning of this model draws inspiration from Retrieval Augmented Fine Tuning (RAFT) [41]. RAFT provides a simple approach to derive the best of both Retrieval Augmented Generation and fine-tuning. The essence of RAFT lies in providing

D retrieved documents or relevant pieces of information (in our case, product details) and fine-tuning the model to use Chain-Of-Thought reasoning to select D* relevant items. For instance, if 5 different types of flour are retrieved and used as context by the LLM, we specifically use Chain-Of-Thought reasoning to select the whole wheat flour if the user profile indicates that the customer is health conscious. This way, we are not only able to fine-tune our model to present the results in the right format but also can teach it how to select the most relevant items from a larger pool of options. We use GPT 4 for automating the process of constructing the chain of thought reasoning and selecting the relevant items which are then manually verified to ensure there are no significant mistakes or issues. Using GPT4 for this process also enables us to leverage its ability to write neat responses with sensible changes in text font, correct usage of bold and italics and even-numbered/unnumbered lists. Furthermore, since GPT4 provides the entire chain of thought reasoning as the response, the whole process can be automated including creation of the jsonl files and finetuning the model after their validation.

Our model is fine-tuned on 25 training and 5 validation single-step conversations using the GPT 3.5 Turbo model. The provided contextual chatlogs are taken from the training set of the high-level LLM whilst maintaining the same user profile. We achieve a training loss of 0.64398, training accuracy of 0.81544, validation loss of 0.80797, and validation accuracy of 0.692. The differences between validation and training are smaller in this case since the conversations are not multi-turn. However, even in these conversations there are differences in the retrieved context from the database and some degree of changes in the responses are to be expected.

F. Low-level LLM

Should the user ask for a low-level, modify or miscellanous query or remark, we call a low-level LLM capable of retrieving the information from the database and giving the output to the user whilst also editing the bill based on the specific request.The process continues until the user is happy with their list and there are not further edits or changes necessary. The low level LLM receives 20 products from the information retrieval system after converting the original query to an embedding and finding the closest neighbours via cosine similarity.

Similar to the strategy employed in the medium level LLM, we use RAFT to provide chain-of-thought reasoning during fine-tuning to ensure the correct and most relevant items are picked from the larger pool. The responses are once again created by GPT4 and manually verified to ensure reliability and consistency. The other benefits such as format and style are thus applicable even for this chatbot.

Our model is fine-tuned on 40 training and 10 validation single step conversations using the GPT 3.5 Turbo 0125 model. The provided contextual chatlogs are taken from the training set of the high level LLM whilst maintaining the same user profile. We achieve a training loss of 0.17557, training accuracy of 0.94014, validation loss of 0.88602, and validation accuracy of 0.67705. Once again, we observe considerable differences between the evaluation of the training and validation datasets which can once again be justified by the complexity of the conversations and existence of multiple correct answers. These results are better than the high level LLM as there is still some amount of grounding due to the usage of relevant items.

All three models have been trained on English conversational data. While training a multilingual model on only English data is not ideal and free of bias, the same choice is supported by the intrinsic ability of LLMs to converse in multiple languages and limitations in the author's proficiency in Dutch. A simple overview with an example of the different models working side by side is provided in Figure 5. The relevant prompts used for the high-level, medium-level, low-level and supermarket data augmentation are available in Appendix G.1,G.2,G.3 and G.5 respectively. Furthermore, the demonstration of our approach featuring the onboarding and different LLMs used at different stages of the conversation has been covered in Appendix H.

G. Text-to-Speech system

The final component of a conversational agent is the Text-To-Speech (TTS) system, which converts output text into a spoken voice. This study does not evaluate various TTS systems, as user preferences regarding accent, gender, and clarity vary widely. Therefore, the ideal system should offer users options based on their preferences. In this work, we utilize the OpenAI Text-to-Speech system (Echo), as demonstrated in Appendix H. OpenAI's TTS is available in multiple voices — Alloy, Echo, Fable, Onyx, Nova, and Shimmer — and produces extremely natural-sounding speech. It can also recite lists in a friendly, human-like manner. However, it is a premium service, currently priced at $15 per 1 million characters. For those seeking a cost-free option, Google's text-to-speech system may be more viable. While free, its output tends to sound more robotic, which may not be ideal for real-world applications. It is also important to note that currently, none of the major industry players, such as OpenAI, Google, and Microsoft, offer a Dutch TTS system that is both natural-sounding and available in multiple genders via an API. However, some recent AI startups, such as ElevenLabs [43], now provide multilingual text-to-speech capabilities, including Dutch, making them a viable option for supporting multilingual speech.

VII. EVALUATION OF OUR ARCHITECTURE AGAINST THE STATE-OF-THE-ART

While conversational agents are often built for specific tasks and the existence of a state-of-the-art is difficult to justify, the emergence of GPTs and Assistants API by OpenAI provides a great benchmark to compare our architecture against. In November 2023, OpenAI released GPTs a novel concept allowing people to create their own custom AI models built on top of the GPT4 Turbo model. They are capable of image creation, code generation, knowledge retrieval, function calling, etc., and can be developed with absolutely no code. It is a revolutionary new step in democratising AI development and usage by allowing everyone to build, share and even earn from the GPTs they create. Assistants API extends the same concept to a business perspective by allowing these GPTs to be hosted on websites of the host. These models use the latest developments in information retrieval and the state-of-the-art LLMs to perform the necessary tasks with great accuracy. The details of the specific GPT created for the scope of this experiment can be found in Appendix I including the instruction provided and the enabled functionalities.

However, currently building and deploying chatbots using the Assistants API is extremely expensive due to the high token consumption by these models for knowledge retrieval tasks. Furthermore, the current setting allows for using only a single

Sl. No	Criterion		Group Scores				Statistical Tests	
			GG	GC	CG	CC	Wilcoxon signed-rank	Mann-Whitney U
1	Agent's Usability	(μ)	2	1.875	1.25	2.125	p = 0.19	p = 0.50
		(σ)	0.7559	0.6408	1.0351	0.3535	W = 12.0	U = 144.5
2	**Agent's Performance**	(μ)	1.5	1.75	1	2.25	p = **0.048**	p = 1.00
		(σ)	1.1952	0.8864	1.0690	0.4629	W = 15.0	U = 128.0
3	Agent's Likeability	(μ)	1.5	1.75	1.125	1.875	p = 0.299	p = 0.814
		(σ)	1.4142	0.7071	1.4577	1.1260	W = 36.5	U = 134.5
4	**User Acceptance of the Agent**	(μ)	1.125	1.75	0.5	2	p = **0.022**	p = 1.00
		(σ)	1.1260	1.0350	1.8516	1.3093	W = 13.5	U = 127.5
5	Agent's Enjoyability	(μ)	0.25	1.25	0.25	1.375	p = 0.091	p = 1.00
		(σ)	2.0528	1.8322	1.7525	1.5059	W = 26.0	U = 127.5
6	User's Engagement	(μ)	1	1.75	0.25	0.875	p = 0.095	p = 0.082
		(σ)	1.1952	0.7071	0.8864	1.8851	W = 26.5	U = 173.0
7	User's Trust	(μ)	1	1.25	0	1.5	p = 0.104	p = 0.63
		(σ)	1.3093	1.0351	1.5118	1.7728	W = 22.5	U = 173.0
8	**User-Agent Alliance**	(μ)	0.75	1.125	-0.375	0.875	p = **0.027**	p = 0.065
		(σ)	0.7071	0.8345	0.9161	1.7268	W = 0.0	U = 0.065
9	Agent's Attentiveness	(μ)	1.625	1.75	1.625	2	p = 0.484	p = 0.633
		(σ)	1.1877	0.7071	1.0606	0.5345	W = 30.5	U = 115.5
10	Agent's coherence	(μ)	2	1.75	0.625	2.125	p = 0.108	p = 0.292
		(σ)	0.7559	1.1650	1.4079	0.8345	W = 19.0	U = 155.0
11	Agent's intentionality	(μ)	2.125	2.125	1.375	2.5	p = 0.087	p = 0.732
		(σ)	0.6409	0.9910	1.0606	0.7559	W = 6.0	U = 137.0
12	**Agent's attitude**	(μ)	1.625	2.25	0.5	1.75	p = **0.022**	p = **0.048**
		(σ)	0.9161	0.7071	1.6036	0.7071	W = 8.0	U = 178.0
13	**Interaction Impact on Self-Image**	(μ)	0.875	1.5	0.25	1.5	p = **0.017**	p = 0.694
		(σ)	0.8345	1.0690	1.5811	0.9258	W = 10.0	U = 138.5

TABLE V: Summary of Statistical Analysis of the Artificial Social Agent Questionnaire. The 4 groups mentioned are an order model pair and stand for: GG - GPT first GPT scores, GC - GPT first Custom chatbot scores, CG - Custom chatbot first GPT scores and CC - Custom chatbot first custom chatbot scores. All 13 criteria fail the Shapiro Wilks test for normality and thus the Wilcoxon Signed Rank test is done to evaluate the performance between models with p-value and Wilcoxon statistic represented as W and Mann-Whitney U-test is performed to test the effect of order with p value and Mann Whitney statistic represented as U are presented below.

model per assistant preventing task optimisation based on specific goals. GPTs, although a part of the Plus membership of ChatGPT cannot be currently used outside the ChatGPT interface. Also currently being in beta phase, it is prone to malfunctioning as new features and evaluations are still ongoing. However, it still serves as a strong contender and we use it to benchmark the performance of our custom multi-LLM chatbot against it. To do this, we perform a within-subjects experimental study where participants are split into two groups based on the order in which they try both chatbots. We ask both groups to fill out the Artificial Social Agent Questionnaire (short version) with all relevant questions and also ask certain qualitative questions to understand their overall experience. We then compare the ASAQ scores of both groups to ascertain which model is ranked better by users on all relevant metrics measured by the questionnaire. The within-subjects design factor ensures all participants try both models and thus can also provide qualitative feedback on their overall experience.

A. *Participant Demographics*

Overall, 16 participants were recruited for the study (9 male and 7 female) between the ages of 23 to 30 (Mean - 24.3125 and SD - 1.8874). In terms of frequency of usage and familiarity with LLM chatbots like ChatGPT, Gemini and Claude, 6 participants responded that they interact with such tools over 5 times a week, 3 responded between 4-5 times a week, 2 responded 3-4 times a week, 4 responded 1-2 times a week and 1 participant responded less than once a week.

B. *Experiment Design*

All participants were first shown the informed consent form to reassure that no personally identifiable information will be collected (Appendix J). The only data stored are their responses to the questionnaire, answers to the qualitative questions and chatlogs for further analysis of factors such as hallucinations. We began by collecting demographic details and asking for a brief insight into their shopping intentions such as what they look for and prioritise when they are shopping in the supermarket. Participants were then asked to interact with either the GPT or the custom multi-LLM chatbot we created. The order in which participants tried both chatbots was routinely cycled to ensure half the participants started by interacting with the GPT and the other half with our solution. Participants were not informed of the nature of the agents and were asked to interact with it in a manner they felt best expressed their supermarket intents and goals. After interacting with the first chatbot, participants were asked to fill the 13 relevant questions from the ASAQ followed by the following qualitative questions -

1) Tell us in detail, what do you find most helpful and unhelpful from this result?
2) If at all, how much does this system make you feel more or less confident about your shopping needs and decisions in a supermarket?
3) Is there anything that you would like to comment about this task?

After this, they were asked to repeat the same procedure but with the other chatbot. The overall experiment took roughly 40 minutes to complete.

Since order is the between-subjects factor and the chatbot is the within-subjects factor, we perform the Mann-Whitney U-test and the Wilcoxon Signed rank test respectively. We use these non-parametric tests since the Shapiro-Wilks test of all the criterion was not normally distributed. This is to be expected given that we were using ordinal data as opposed to continuous values.

C. Questionnaire Results

As seen in Figure 6 we observe that our solution performs better than the GPT on all 13 tested parameters of the ASAQ. We continue by performing statistical tests on all 13 parameters to find out which parameters are significantly better in our model compared to the state-of-the-art. Table V lists all the 13 parameters. Overall we observe that in terms of agent performance, user acceptance of the agent, user-agent alliance, agent attitude and interaction impact on self image, the p-value is lesser than 0.05. The Mann-Whitney U-test shows that order is not statistically significant for all criteria except agent's attitude. Thus we cannot rule out agent's attitude as being statistically better since order could have influenced the results.

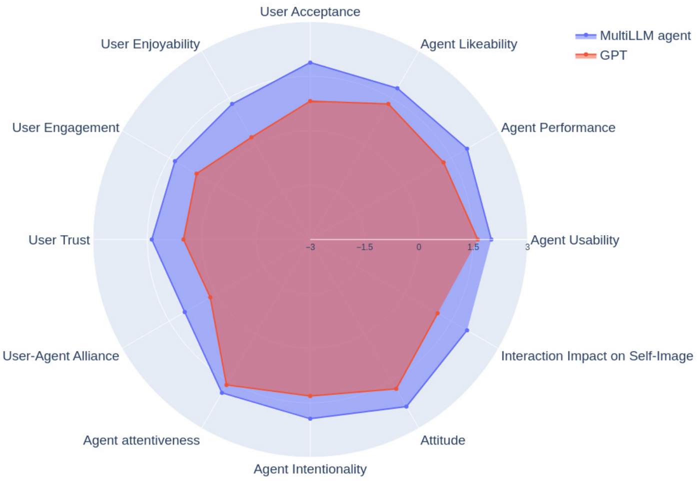


Fig. 6: Comparison of the GPT with our custom multi-LLM solution on the provided ASA chart. The scores ranges from -3 to +3 of the Likert scale on which the ASAQ is built. Our multi-LLM approach performs better than the GPT on all 13 parameters.

D. Qualitative Results

As mentioned in the experiment design, participants were also asked 3 qualitative questions to try and understand their overall experience better.

1) Benefits of GPT

Participants overall agreed that the GPT model was simple to use and interact with. Furthermore, all participants who were looking for detailed recipes and instructions on making certain meals and dishes were extremely pleased with the detailed responses of GPT. Participant #5 commented on its usefulness as a brainstorming tool to help make decisions of what to purchase and what to try out. Participant #7 found the responses of the GPT to be more cohesive and in line with their expectations when inquiring about meal preparation strategies for the entire week. Furthermore, participant #15 found that the responses to complex questions were quite well handled whilst ensuring the conversational tone and language were simple to understand. Whilst none of the participants were overly enthusiastic about the responses and strength of this system, they were content with the answers and recommendations provided by it.

The different merits of the GPT model can be attributed to the agent's powerful underlying model (GPT4 Turbo), its reliability in keeping track of previous conversations with relative ease and flexibility to handle all kinds of queries even those that diverge from the traditional product recommendation and information objective (e.g. - recipes or detailed plans to achieve a goal). This makes the agent more robust to greater customer variations in requests whilst also having track of all the conversations with the user in mind.

2) Concerns about GPT

Participant #2 and #3 were concerned about hallucination and mentioned that this affected the degree of trust they could place in the system. P#2 found some items which did not exist in the database in the responses which were misleading (hallucination) whilst P#3 was not able to get information about a screwdriver despite the item being present in the database (omission). Participant #6 had issues substituting organic spinach with regular spinach despite a number of attempts. Participant #4 and #8 found the number of options provided by the GPT was limited which made them feel more restricted in terms of choices. Participant #9 observed that despite mentioning their dietary preferences as being a vegetarian in the user profile, the agent recommended options which did not conform with that. Participant #14 found that the chatbot was also not able to justify its choices clearly when making recommendations. Multiple participants also commented on the inability of the GPT to provide complete information in its response. For instance, when recommending product names it often forgot to mention the price and location which had to be requested for separately.

Overall, although the GPT possesses its own knowledge retrieval functionality, the efficiency of the same is reduced when the number of items to be retrieved is higher. This leads to either hallucinations or omissions, both of which are detrimental in the case of a supermarket chatbot as a hallucination misleads the user into believing that certain products which do not exist are available while omissions can lead to lost opportunities to recommend appropriate items for the customer.

3) Benefits of Custom multi-LLM chatbot

Participants overall agreed that the proposed chatbot was direct and efficient. Multiple participants commented on the preciseness of the answers which they found made the chatbot very helpful. Although participants were asked to only evaluate the chatbot based on the responses, participants were also impressed with the speed of the chatbot. Participant #4 commented on how the chatbot reminded them of certain ingredients for their dish that they had forgotten which was very useful. Participant #5 mentioned that they found the ability of asking questions to narrow down the options to be a helpful feature in the agent. Participant #7 commented about the reliability and trustworthiness of the agent on account of both the format and reasoning provided by the chatbot. Participant #11 also mentioned how this chatbot could be useful for people who tend to be more socially anxious and wary of approaching the workers in the supermarket for help and recommendations.

The multi-LLM approach is more to the point on account of being fine-tuned on task-related conversations. By using multiple smaller models inference speed is greatly increased compared to the GPT4 alternative whilst also reducing costs. The usage of our own retrieval system proved to be more effective than the alternative used internally by the GPT's knowledge retrieval functionality. The ability of the high-level LLM to actually break down and list all the potential items needed for a complex query ensured the customer never forgot about any item which was perceived as useful. Since the medium-level and low-level LLM were fine-tuned on providing reasoning in their responses, the overall credibility of the system was improved as well.

4) Concerns regarding the Custom multi-LLM chatbot

In general, participants felt that the chatbot's ability to provide detailed recipes, ideas or plans outside the scope of product recommendation was fairly limited. Participant #2 stated that they felt the chatbot was more coercive and 'pushy' by trying to force them towards specific products. Participant #3 and #5 found that the chatbot made errors when summarising the final list or maintaining track of the conversation. Participant #8 found that when the LLM was asked to provide the total price of all products, the answer was incorrect. Participant #13 also commented on how the tool may lead to them purchasing more than they initially sought out.

One of the main issues with the multi-LLM approach was when the query of the user was misclassified. Thus when a high-level query was misclassified as low-level, or vice versa typically unsatisfactory results were obtained. This is primarily caused due to the inability of either approach to respond to queries that aren't in line with the strategies employed by both approaches. For instance, if a high-level query is misclassified as low, the information retrieval is fairly poor and relevant items are not extracted from the database. Meanwhile, if a low-level query is classified as high-level, the necessary context of the previous conversation is not available leading to confusion in terms of recommendations by the agent.

VIII. INTEGRATION INTO ROBOTICS

The conversational agent is a powerful tool to help customers in a supermarket find what they are looking for, get useful information and also obtain personalised recommendations based on their preferences. While this chatbot can be applied as a standalone application on a mobile phone or kiosk at the entry of the supermarket, we are also interested in exploring how these chatbots can be effectively integrated into high-level robot planning to guide a supermarket robot to go to the necessary locations after which the required low-level perception, motion planning of a manipulator and control can be applied for automated object retrieval and collection. This feature is useful as it can allow a customer to interact with the chatbot and have a robot autonomously pick up the necessary items and bring it to the user. While the low-level functionality such as perception and manipulation are beyond the scope of our work, we demonstrate with a simple example how our robot can navigate to the necessary shelves after receiving an appropriate request from the customer.

The key assumption made in this work is that the position of all shelves remains the same over time. This is a reasonable assumption to make since most path-planning algorithms require a pre-recorded map to facilitate path planning from a given start point to a destination. If the supermarket is to change its overall configuration, a new map would have to be generated by using SLAM or other similar mapping techniques.

To connect the chatbot with the robot, we use an LLM to process the final conversational agent message which has a list of all the products the customer has indicated a willingness to purchase and retrieve a list of shelf numbers for each object. The prompt for this LLM is provided in Appendix G.4. We then define this as a set, removing any duplicates in case multiple items are in the same shelf. The shelves can then be arranged in an order to optimise the total distance covered by the robot. We then look up the specific shelf numbers position from a pre-configured YAML file consisting of the X-Y coordinates of the shelves to retrieve the destination and end pose of the robot. We iterate over all the shelves one after the other until the robot has visited all the necessary items.

For the simulation shown in Figure 7 we build on ROS Noetic using a Clearpath mobile base robot with a Frank Emika arm. The local planner for the mobile base is set as Timed Elastic Band (TEB) [44] and localisation as Adaptive Monte Carlo Localisation (AMCL) [45]. After reaching the shelf, the necessary perception, motion planning and control nodes can be called so as to facilitate picking up the right object and adding it to the supermarket basket. Once this is done, the next shelf can be visited and so on until all items are retrieved. The robot can then navigate to the checkout for delivering the items to the customer. This is a simple yet effective manner in which our solution can be integrated into a supermarket robot.

IX. DISCUSSION

After presenting our approaches and results for building inclusive, personalised and effective agents, we now discuss the key learnings and take aways along with limitations and scope for future work.

A. Discussion on Inclusivity

Based on the results in Section III, we make multiple observations. Firstly, OpenAI's Whisper has proven to be the best speech recognition system by being significantly better than its peers. It also retains this accuracy for Dutch and different genders. Microsoft Azure Speech comes next followed by Google Cloud Speech. Vosk, the other open-source model, can run locally and has a fast inference time albeit at a higher error rate.

Whisper also has other benefits that make it the clear winner in this evaluation:

1) Whisper is open source - allowing it to be run locally given the necessary hardware requirements and can be fine-tuned

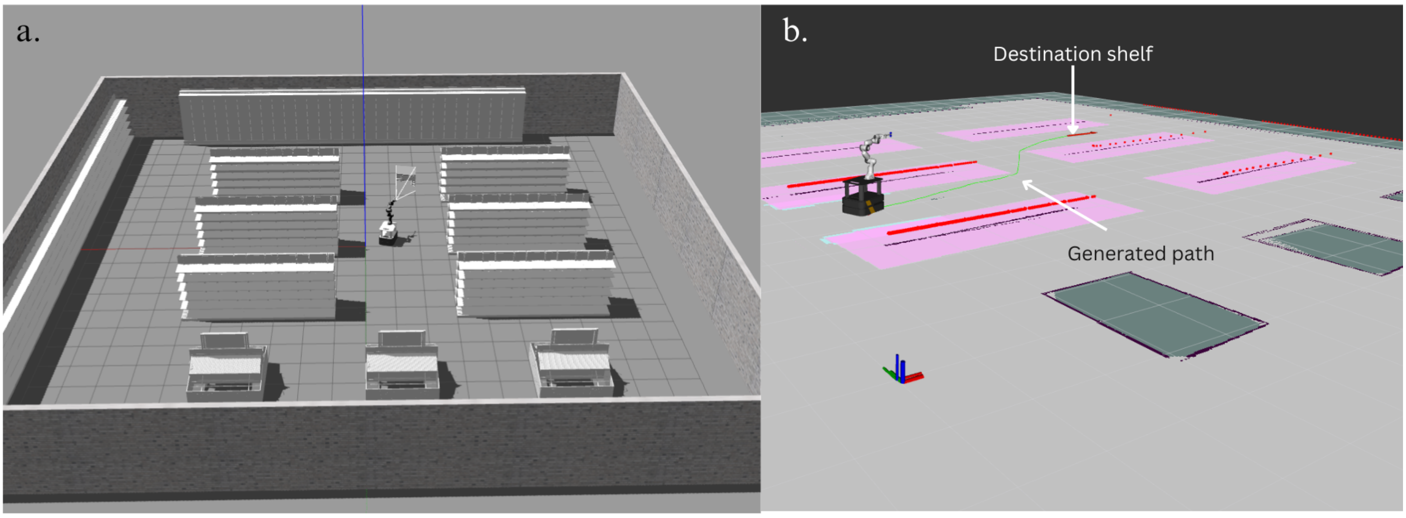


Fig. 7: The robot in a large simulated supermarket. Figure a. shows the render on Gazebo while Figure b shows the path (in green line) and the robot navigating to the correct shelf in RViz. The simulation and demonstration have been done on ROS Noetic.

with more speech data for specialised purposes improving its performance. While both Microsoft and Google allow training on the consoles as well, data privacy concerns and costs may make them less attractive than Whisper.

2) Unlike the other models, for the evaluation it was observed that Whisper was the only model that was capable of recognizing the language by itself while other models needed the language passed as a parameter or in the case of Vosk, to add the path to the files needed for the specific language.
3) Remarkable developments are still being made to improve the speed of inference of Whisper allowing for faster versions of the same created by the community which could help reduce latency.

We also notice that the Word Error Rate for the same models is higher in the Dutch language than in English. This is to be expected since most of the models have been trained on far longer durations and greater quantities of data in English rather than Dutch. However, if we intend to build systems that can be deployed to the general public in the Netherlands, efforts must be made to fine-tune these open-source models on large amounts of Dutch audio-transcription data so that higher accuracy can be achieved.

The limitation of the current study is a small sample size preventing deeper analysis into other factors such as the accent and age of the speaker - important variables in the performance of speech recognition systems. [20] states that Nigerian women for instance have significantly higher error rates than white caucasian males. Analysing the accent of the person could provide further insights into the robustness of such systems and if they have certain biases that can be rectified by fine-tuning representational data which also includes the marginalized group. Furthermore, using other metrics like Word Information Lost and Match Error Rate could be used to further test our results and get better insights on the performance of these models.

B. *Discussion on Personalisation*

We presented a unique approach to consolidating the important characteristics required to describe a potential customer inspired by the categorisation of customers by retail salespeople. In our desire to capture the essence of a customer via a short list of questions, we came to realize that chatbots in a supermarket are effectively the next generation of salespeople providing recommendations and advice based on the users' needs. Therefore when building commercial chatbots, we recommend considering the mindset of human salespeople and their insights from experience to ensure the chatbots are well aligned and effective in answering the customer's queries.

We note that the chatbot was prompted to interact in the right approach by using only three example conversations, facilitated by means of few-shot learning [37] - indicating the low degree of effort required to align powerful LLMs like GPT4 with the desired intent. Lastly, we also ensured that the novelty effect would not wear off by implementing approaches such as name repetition, user preference recall, respecting dietary preferences and so on. We believe this makes the chatbot more personable and likely to be used routinely. To account for the different order, the chi-squared test for independence was carried out and obtained a statistically significant result that the chatbot with memory is preferred more than the chatbot without. Based on this we can conclude that augmenting an LLM with a user's profile is more likely to be better received and appreciated by customers.

Despite the promise of our simple chatbot to test the effect of personalisation, it has a few limitations which must be addressed. Firstly, we used GPT-4-Turbo as our LLM - an extremely powerful but slow model which led to very high latency in responses. It was this observation that prompted us to explore and develop the multi-LLM solution. By gathering participant feedback, we were able to understand the other requirements such as need for price and discounts which prompted us to create the dataset of supermarket items. Secondly, the current approach incorporates only speech and text-based input and QR code identification as its modalities. Working towards capturing more context-relevant information such as facial features and conversational tone could provide greater information to the LLM – potentially improving responses. However, care must be taken to ensure that such additions do not further increase the latency. Lastly, we assumed that the interactions between the customer and the chatbot would be fairly short (5-10 minutes) and that the supermarket contained 100 categories of items (e.g. - bakery goods, cleaning supplies, pet food, etc.). This assumption will not hold for customers who interact longer or for significantly larger supermarkets due to the limitations of the context window of LLMs. To address the problem of longer conversations, implementing a forgetting mechanism to discard irrelevant data and a summarise previous conversations could hold great value.

C. Discussion on multi-LLM conversational agent

Overall, we observe that the multi-LLM approach offers multiple benefits over using the most powerful LLM like the state-of-the-art GPT such as reduced costs, reduced latency, increased control over specialised tasks, easier ablation and comparative studies and better task performance. While the GPT solution is indeed the quickest and easiest in terms of deployability, the performance of knowledge retrieval is rather inadequate. By utilising multiple smaller LLMs capable of interacting with one another and maintaining a common conversation log helps in providing context to each separate model as well. The presence of a classifier enables us to directly route queries to the correct model instead of following a common approach for all questions. By fine-tuning with GPT4 augmented data we are also able to leverage the formatting and style of the responses to be extremely well structured and easy to understand.

Furthermore, the modular nature of our solution enables easy substitution of models with alternatives as they become available making the solution extremely flexible to adapt to future developments in the field. One can also fine-tune and use open-source models to ensure reliability and address concerns regarding data privacy and security. Furthermore, by increasing or optimising the number of classes the classifier can select items into, other roles can also be unlocked such as bill management, asking for assistance from supermarket workers or providing feedback. The approach is also not limited to supermarket scenarios and can easily applied to other domains which could benefit from utilising voice-based interfaces. By selecting the type of LLMs and queries, the approach can be optimised based on the specific task. For example, if one were to build a polishing robot for the industry, aspects such as inclusivity and effectiveness would still be significant while personalisation would have to be changed based on the worker's preferences. Furthermore, instead of having 4 classes of queries, the number could be reduced to 2 - control queries where the user specifically sets values such as impedance or high-level generic queries such as outlining the overall task.

However, the current approach is not without its share of limitations. Incorrectly classified queries can lead to the query being handled by a model that is not specialised for the given task. This could potentially lead to a loss of context and confusing results to users. Since the classifier is built atop a multilingual BERT classifier the responses are highly sensitive to changes in spellings and the manner in which the customer expresses themselves. We believe replacing the mdistilBERT classifier with a small fine-tuned LLM tasked with query classification and rewriting to add any necessary context could be a viable solution to address these limitations and add context to a query to improve retrieval.

X. CONCLUSION

This thesis advances research in the field of voice-based interface research for supermarket applications and robots, emphasizing inclusivity, personalisation, and the development of a novel multi-LLM agent. In terms of inclusivity, our findings show that OpenAI's Whisper outshines three other leading speech recognition systems in robustness and accuracy. Furthermore, our findings reveal a higher Word Error Rate for Dutch than English, which underscores the necessity for specialized ASR systems in Dutch settings. For personalisation, users significantly favoured our personalised conversational agent, modeled after retail salespeople classifications, over the control. Lastly, our multi-LLM agent surpassed the state-of-the-art in 4 of 13 parameters and demonstrated better performance across all 13 measured ASAQ criteria. The successful integration of LLMs into robot path planning for shelf-directed item retrieval exemplifies the practical application of these interfaces in real-world settings. These studies and experiments set the stage for developing more equitable, customized, and effective interfaces across various domains.

REFERENCES

[1] Richard Bloss. Collaborative robots are rapidly providing major improvements in productivity, safety, programing ease, portability and cost while addressing many new applications. *Industrial Robot: An International Journal*, 43:463–468, 08 2016.

[2] Ken Goldberg. Robots and the return to collaborative intelligence. *Nature Machine Intelligence*, 1(1):2–4, 2019.

[3] Valeria Villani, Fabio Pini, Francesco Leali, and Cristian Secchi. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, 55:248–266, 2018.

[4] Muhammad Umar Anjum, Umar Shabaz Khan, Waqar Shahid Qureshi, Ameer Hamza, and Wajih Ahmed Khan. Vision-based hybrid detection for pick and place application in robotic manipulators. In *2023 International Conference on Robotics and Automation in Industry (ICRAI)*, pages 1–5, 2023.

[5] Danica Kragic, Joakim Gustafson, Hakan Karaoğuz, Patric Jensfelt, and Robert Krug. Interactive, collaborative robots: Challenges and opportunities. In *International Joint Conference on Artificial Intelligence*, 2018.

[6] Bela Takarics, Peter T Szemes, Gyula Németh, and Peter Korondi. Welding trajectory reconstruction based on the intelligent space concept. In *2008 Conference on Human System Interactions*, pages 791–796. IEEE, 2008.

[7] Alberto García, J. Ernesto Solanes, Adolfo Muñoz, Luis Gracia, and Josep Tornero. Augmented reality-based interface for bimanual robot teleoperation. *Applied Sciences*, 12(9), 2022.

[8] Emmanuel Senft, Michael Hagenow, Kevin Welsh, Robert Radwin, Michael Zinn, Michael Gleicher, and Bilge Mutlu. Task-level authoring for remote robot teleoperation. *Frontiers in Robotics and AI*, 8, 2021.

[9] Naijun Liu, Tao Lu, Yinghao Cai, Jinyan Lu, Huaixu Gao, Boyao Li, and Shuo Wang. Design of virtual reality teleoperation system for robot complex manipulation. In *2019 Chinese Automation Congress (CAC)*, pages 1789–1793, 2019.

[10] J. Ernesto Solanes, Adolfo Muñoz, Luis Gracia, and Josep Tornero. Virtual reality-based interface for advanced assisted mobile robot teleoperation. *Applied Sciences*, 12(12), 2022.

[11] Abdullah Shaif, Suresh Gobee, and Vickneswari Durairajah. Vision and voice-based human-robot interactive interface for humanoid robot. In *AIP Conference Proceedings*, volume 2788. AIP Publishing, 2023.

[12] Yang Bai, Irtaza Shahid, Harshvardhan Takawale, and Nirupam Roy. Whisperwand: Simultaneous voice and gesture tracking interface, 2023.

[13] J. Norberto Pires. Robot-by-voice: experiments on commanding an industrial robot using the human voice. *Industrial Robot: An International Journal*, 32(6):505–511, 2005.

[14] Adam Rogowski. Industrially oriented voice control system. *Robotics and Computer-Integrated Manufacturing*, 28(3):303–315, 2012.

[15] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457, March 1 2021.

[16] John Levis and Ruslan Suvorov. Automatic speech recognition. *The encyclopedia of applied linguistics*, 2012.

[17] Arun Sharma and Michael Levy. Categorization of customers by retail salespeople. *Journal of Retailing*, 71(1):71–81, 1995.

[18] OpenAI. Introducing gpts. https://openai.com/blog/introducing-gpts, 2023. Accessed: 2024-04-09.

[19] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd*

ACM International Conference on Intelligent Virtual Agents, IVA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[20] Sai Sathiesh Rajan, Sakshi Udeshi, and Sudipta Chattopadhyay. Aequevox: Automated fairness testing of speech recognition systems, 2022.

[21] Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 230–247, New York, NY, USA, 2022. Association for Computing Machinery.

[22] Allison Koenecke, A. Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, J. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689, 2020.

[23] Walid El Hefny, Alia El Bolock, Cornelia Herbert, and Slim Abdennadher. Towards a generic framework for character-based chatbots. pages 95–107, 2020.

[24] Petter Bae Brandtzaeg and Asbjørn Følstad. Chatbots: changing user needs and motivations. *Interactions*, 25(5):38–43, aug 2018.

[25] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc V Le, and E. Chi. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. *ArXiv*, abs/2312.00763, 2023.

[26] Alpha Cephei. Vosk speech recognition api. Accessed: 2023-08-24.

[27] Alexey Andreev and Kirill Chuvilin. Speech recognition for mobile linux distrubitions in the case of aurora os. In *2021 29th Conference of Open Innovations Association (FRUCT)*, pages 14–21, 2021.

[28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[29] Google Cloud. Speech-to-text: Automatic speech recognition. https://cloud.google.com/speech-to-text, 2023. Accessed: [Your Access Date].

[30] Gamal Bohouta and Veton Këpuska. Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *Int. Journal of Engineering Research and Application*, 2248-9622:20–24, 03 2017.

[31] Microsoft Azure. Speech service. https://, 2023. Accessed: [Your Access Date].

[32] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition, 2017.

[33] Andrew Morris. An information theoretic measure of sequence recognition performance. 2002.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[36] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[37] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[39] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023.

[40] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[41] Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. Raft: Adapting language model to domain specific rag, 2024.

[42] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. In search of needles in a 11m haystack: Recurrent memory finds what llms miss, 2024.

[43] ElevenLabs. Elevenlabs text-to-speech api reference. https://elevenlabs.io/docs/api-reference/text-to-speech, 2024. Accessed: 2024-04-12.

[44] Christoph Rösmann, Wendelin Feiten, Thomas Wösch, Frank Hoffmann, and Torsten Bertram. Trajectory modification considering dynamic constraints of autonomous robots. In *ROBOTIK 2012; 7th German Conference on Robotics*, pages 1–6. VDE, 2012.

[45] Ming-An Chung and Chia-Wei Lin. An improved localization of mobile robotic system based on amcl algorithm. *IEEE Sensors Journal*, 22(1):900–908, 2022.

2

Appendices

A

Participant Instructions

For all three experiments clear instructions were provided to the participants on how to contribute effectively for the study. Here we outline the instructions for the experiments one by one.

A.1. Speech Recognition Experiment

Participants were welcomed and thanked for participating in the study after which the informed consent form and other details about the experiment and data privacy were communicated. Participants fell into one of four groups - Dutch Male, Dutch Female, English Male and English Female. All groups were made exclusive to prevent any interference by allowing the same participant to speak in multiple languages.

The FiFine mic was setup and tested on Audacity before every recording whilst ensuring the experiment was conducted in a low noise environment. Participants were shown the script they would have to read out and were given as long as they liked before they read through the text they would have to dictate. Once that was done, participants were asked to speak both the lines of the user and chatbot, normally into the mic and the recording was captured via audacity. For all 40 participants the mic configuration and settings on Audacity were maintained at the same default level. If participants made a mistake, they were allowed to rerecord the audio file.

After the recording of the file, the same was tested to ensure there was no error in audio capture. The audio file was saved as an mp3 and given a unique id to link it to the participant. In a separate file by the same id name, the gender and accent of the participant was recorded.

A.2. Personalisation Experiment

Participants were welcomed and thanked for participating in the study after which the informed consent form and other details about the experiment and data privacy were communicated. Participants were placed into 2 groups based on the order in which they interacted with the control/personalised chatbots. The order was cycled for each subsequent participant. Participants were only informed that they would be interacting with 2 chatbots - one after the other and that they would have to pick which chatbot they preferred more at the end of the experiment.

Participants began by filling the onboarding questionnaire in both cases, obtaining a QR code which served as their membership card and then interacting with the chatbot for a maximum duration of 5 minutes. After this, participants were asked to repeat the same procedure for the other chatbot. The onboarding and QR code step was used even for the non personalised chatbot to mitigate any potential biases but the retrieved information was not used in the control agent. At the end of the experiment, participants were asked to record which chatbot, the first or the second they preferred more.

A.3. Interaction With Novel Multi-LLM Agent vs State-Of-The-Art GPT

Participants were welcomed and thanked for participating in the study after which the informed consent form and other details about the experiment and data privacy were communicated. Participants were placed into 2 groups based on the order in which they interacted with the GPT/multi-LLM chatbots. The order was cycled for each subsequent participant. Participants were only informed that they would be interacting with 2 chatbots - one after the other. The overall experiment took roughly 40 minutes per participant.

Before they began, required demographic information was collected such as age, gender, familiarity and frequency of usage of LLMs to ensure that all participants had interacted with such agents to prevent any learning effects from interfering with the study. Next participants were asked to reflect what their general intentions during a supermarket visit tend to be. The specific instruction provided here was -

Please describe, in around 100 words, any objectives or inquiries you might have while visiting a supermarket. This could range from searching for particular items, seeking advice or recommendations, to any general queries you often find yourself pondering amidst the aisles. Feel free to reflect on your personal needs, preferences, or a specific list of items you aim to purchase.

A useful way to approach this is to think about the types of products or goods that usually draw your interest, or those you suddenly remember you need once you're there. Your input can draw upon both past shopping experiences or current needs.

This question was asked to help participants mentally prepare themselves by thinking of what they would potentially look for in a supermarket so that they could interact with the chatbot more naturally. After this, they were provided with the first chatbot and asked to interact with it until they felt satisfied with the results or enough to make an assessment. They were then asked to fill out the relevant questions in the ASAQ questionnaire. Following this, 3 qualitative questions were provided to which the participants could write their views in whatever depth they deemed necessary. The same process was then repeated for the other chatbot.

B

Script for evaluation of Speech Recognition Systems

Participants were asked to read both lines without saying user or chatbot. The script was created based on a common interaction in a supermarket and consists of multiple words and phrases that are commonly utilised.

B.1. English Version

User: "Hello, I'm planning a big family dinner and need to do some shopping. Can you help me find a few items?"

Chatbot: "Of course! What are you looking for?"

User: "Let's start with fresh produce. I need tomatoes, lettuce, and cucumbers for a salad." Chatbot: "You can find those in the fresh produce section, aisle 4."

User: "Great, thanks. I also need to pick up some seafood. Do you have salmon and shrimp?"

Chatbot: "Yes, our seafood selection is in aisle 9. You'll find both salmon and shrimp there."

User: "Perfect. I'm thinking of making a dessert too. Where can I find chocolate chips and vanilla extract?"

Chatbot: "For baking supplies, head over to aisle 7. You'll find chocolate chips and vanilla extract there."

User: "Also, I almost forgot, where's the pasta section?"

Chatbot: "Pasta is in aisle 6, right next to the Italian sauces."

User: "Thank you. That should be all for now."

Chatbot: "You're welcome! Let me know if there's anything else I can assist with. Have a great shopping experience!"

User: "Actually, just one more thing. Where can I find cleaning supplies?"

Chatbot: "Cleaning supplies are located in aisle 11. Is there a specific brand you're looking for?"

User: "No, that's all. Thanks again!"

Chatbot: "No problem at all. Goodbye and take care!"

B.2. Dutch Version

Gebruiker: "Hallo, ik plan een groot familiediner en moet wat boodschappen doen. Kun je me helpen een paar artikelen te vinden?"

Chatbot: "Natuurlijk! Waar bent u naar op zoek?"

Gebruiker: "Laten we beginnen met verse producten. Ik heb tomaten, sla en komkommers nodig voor een salade."

Chatbot: "Die vindt u in de versafdeling, gang 4."

Gebruiker: "Geweldig, bedankt. Ik moet ook wat zeevruchten halen. Hebben jullie zalm en garnalen?"

Chatbot: "Ja, onze selectie aan zeevruchten bevindt zich in gang 9. Daar vindt u zowel zalm als garnalen."

Gebruiker: "Perfect. Ik denk erover om ook een dessert te maken. Waar kan ik chocolade stukjes en vanille aroma vinden?"

Chatbot: "Voor bakbenodigdheden kunt u naar gang 7 gaan. Daar vindt u chocoladeschilfers en vanille-extract."

Gebruiker: "O ja, ik was bijna de pasta vergeten. Waar ligt de pasta?"

Chatbot: "Pasta vindt u in gang 6, naast de Italiaanse sauzen." Gebruiker: "Dank je. Dat zou alles moeten zijn voor nu."

Chatbot: "Graag gedaan! Laat het me weten als ik nog ergens anders mee kan helpen. Een prettige winkelervaring toegewenst!"

Gebruiker: "Eigenlijk, nog één ding. Waar kan ik schoonmaakartikelen vinden?"

Chatbot: "Schoonmaakartikelen vindt u in gang 11. Zoekt u een specifiek merk?" Gebruiker: "Nee, dat is alles. Nogmaals bedankt!"

Chatbot: "Geen probleem. Tot ziens en zorg goed voor uzelf!"

C

Informed Consent - Speech Recognition Experiment

Research Study Invitation:

You are being invited to participate in a research study titled *Evaluating Off-the-Shelf Automated Speech Recognition Models for Voice-based Control of Robots: A Comparative Study*.

This study is being conducted by Chandran Nandkumar, a Master's student from TU Delft, under the guidance of Dr. Luka Peternel for the scope of his Master thesis.

The purpose of this research study is to compare different commercially available Speech recognition systems for different genders and languages (English and Dutch), and will take you approximately 15-20 minutes to complete. The data will be used for evaluating which models of speech recognition are the most robust to handle different types of users and their potential use for voice-based control of robots. We will be asking you to provide certain scripted voice commands to a robot in a supermarket setting along with data regarding your personal information such as gender, accent, and nationality for analysis of the results.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by separating contact information from personal information, deleting the data after analysis and verification of results, and taking all necessary steps to ensure data is protected by suitable encryption and using OneDrive via the TU Delft account to store details.

Your participation in this study is entirely voluntary, and you can withdraw at any time. You are free to omit any questions.

TEMPLATE 2: Explicit Consent points

Please make sure that you select (and amend as necessary) any Explicit Consent points which are relevant to your study and exclude those which do not apply. You should also add further points and necessary to address your specific research situation.

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated 12/12/2023 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.	☐	☐
2. 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.	☐	☐
3. I understand that taking part in the study involves: recording audio of specific voice commands given to a supermarket robot whose script has been handed over to you. The data will be recorded in mp3 format and information regarding your gender, accent and language will be used to compare the performance of the different speech transcription models. The recording provided will be destroyed after the completion of the analysis and latest by the end of May 31st, 2024.	☐	☐
4. I understand that the study will end before May 31st, 2024 (latest) and all the data collected will be destroyed.		
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
5. I understand that taking part in the study involves the following risks – reading out certain non-offensive, fixed script commands to a supermarket robot which may be mildly discomfortable for certain long phrases and commands; furthermore – multiple recordings may be requested due to technical difficulties, need to increase sample size and capture the best possible voice recording. I understand that these will be mitigated by allowing the participants to take breaks whenever they would like, providing water and/or any refreshment that may help them with participating in the study.	☐	☐
6. I understand that taking part in the study also involves collecting specific personally identifiable information (PII) such as name, gender, nationality, accent and associated personally identifiable research data (PIRD) of the voice recordings for certain given voice commands with the potential risk of my identity being revealed indirectly after data analysis via a publication in a Master thesis, or a conference/journal publication	☐	☐
7. I understand that some of this PIRD is considered as sensitive data within GDPR legislation, specifically - None	☐	☐
8. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach - a. All information will be stored in OneDrive via the TU Delft account and the personal data will be configured as online only files ensuring that there is no access to the data on any local computer b. The voice recordings may also be used via the personal computer but encryption standard like using BitLocker and regular updates will be performed to ensure that the safety is ensured. c. The contact details of the participant such as name, number, email will be separate from there PII using a unique random ID number linking the two. This ensures the personal information cannot be traced back to the participant.	☐	☐

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
d. All data will be destroyed promptly after analysis or publication and verification of results latest by May 31st , 2024.		
9. I understand that personal information collected about me that can identify me, such as name, email id, contact number, gender, nationality, accent and voice recordings will not be shared beyond the study team.	☐	☐
10. I understand that the (identifiable) personal data I provide will be destroyed promptly after analysis or publication and verification of results latest by May 31st , 2024.	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
11. I understand that after the research study the de-identified information I provide will be used for publication in the results for my master thesis and/or potential publication in a journal or conference. The objective of such a study would be to understand how speech transcription models are capable of handling different genders, languages, nationalities, accents and which transcription model is the best for giving voice commands to robots in a supermarket setting.	☐	☐
12. I agree that my responses, views or other input can be quoted anonymously in research outputs	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
13. I give permission for the de-identified contact information, personal information and voice recordings that I provide to be archived in OneDrive via TU Delft account repository so it can be used for future research and learning.	☐	☐
14. I understand that access to this repository is restricted only to the principal investigator of the study. The personal and contact information will be stored in online only mode while the voice recording will be stored on a personal computer which is encrypted by BitLocker and will be regularly updated to abide by the latest security protocols. The data will be deleted after analysis or verification of the results latest by 31st May 2024	☐	☐

Signatures

____________________________	___________________________	_________
Name of participant [printed]	Signature	Date

D

Normality Tests, Pairwise t-tests Of All Groups For All Models

We begin by checking all the groups for normality using the Shapiro Wilk test. We have 4 groups - Dutch Female, English Female, Dutch Male, and English Male. We also have 4 models - Microsoft Azure, Google Cloud Speech to Text, OpenAI Whisper and Vosk. Thus there are 16 different configurations for which we apply the Shapiro-Wilk test. The same has been tabulated in D.1.

Table D.1: Shapiro-Wilk Tests for Normality of WER Scores

Group	Model	Statistic	P-Value
Dutch Female	Microsoft	0.890	0.169
Dutch Female	Vosk	0.879	0.127
Dutch Female	Google	0.912	0.298
Dutch Female	Whisper	0.912	0.293
English Male	Microsoft	0.913	0.303
English Male	Vosk	0.939	0.546
English Male	Google	0.947	0.634
English Male	Whisper	0.889	0.165
Dutch Male	Microsoft	0.973	0.920
Dutch Male	Vosk	0.958	0.765
Dutch Male	Google	0.943	0.588
Dutch Male	Whisper	0.971	0.898
English Female	Microsoft	0.661	0.000298
English Female	Vosk	0.907	0.258
English Female	Google	0.890	0.170
English Female	Whisper	0.893	0.182

As we can see, only one of the combinations - English Female Microsoft fails the Shaprio Wilk test and thus deviates from the normal distribution. We thus perform the Wilcoxon-signed rank test for English female Microsoft and other models while using the parametric t-test for the rest.

D.1. Dutch Female

Model 1	Model 2	T-Statistic	P-Value
Microsoft	Vosk	-3.49	0.0026
Microsoft	Google	-2.10	0.0501
Microsoft	Whisper	1.70	0.1055
Vosk	Google	0.84	0.4139
Vosk	Whisper	5.41	0.000038
Google	Whisper	3.43	0.0030

Table D.2: Pairwise T-Tests for Dutch Female Group

D.2. English Female

Model 1	Model 2	T-Statistic	P-Value
Vosk	Google	1.28	0.2185
Vosk	Whisper	7.16	0.0000011
Google	Whisper	4.79	0.000146

Table D.3: Pairwise T-Tests for English Female Group (no Microsoft as it violated the Shapiro-Wilk test)

Table D.4: Wilcoxon Signed-Rank Tests for English Female Group Comparisons with Microsoft

Comparison	Statistic	P-Value
Microsoft vs Whisper	10.0	0.1386
Microsoft vs Google	0.0	0.00195
Microsoft vs Vosk	0.0	0.00195

D.3. Dutch Male

Model 1	Model 2	T-Statistic	P-Value
Microsoft	Vosk	-4.44	0.00032
Microsoft	Google	-3.30	0.0039
Microsoft	Whisper	1.71	0.1046
Vosk	Google	1.67	0.1117
Vosk	Whisper	6.73	0.0000026
Google	Whisper	6.22	0.0000071

Table D.5: Pairwise T-Tests for Dutch Male Group

D.4. English Male

Model 1	Model 2	T-Statistic	P-Value
Microsoft	Vosk	-7.71	0.000000414
Microsoft	Google	-4.79	0.000147
Microsoft	Whisper	1.21	0.2423
Vosk	Google	2.26	0.0368
Vosk	Whisper	8.11	0.000000200
Google	Whisper	5.29	0.000050

Table D.6: Pairwise T-Tests for English Male Group

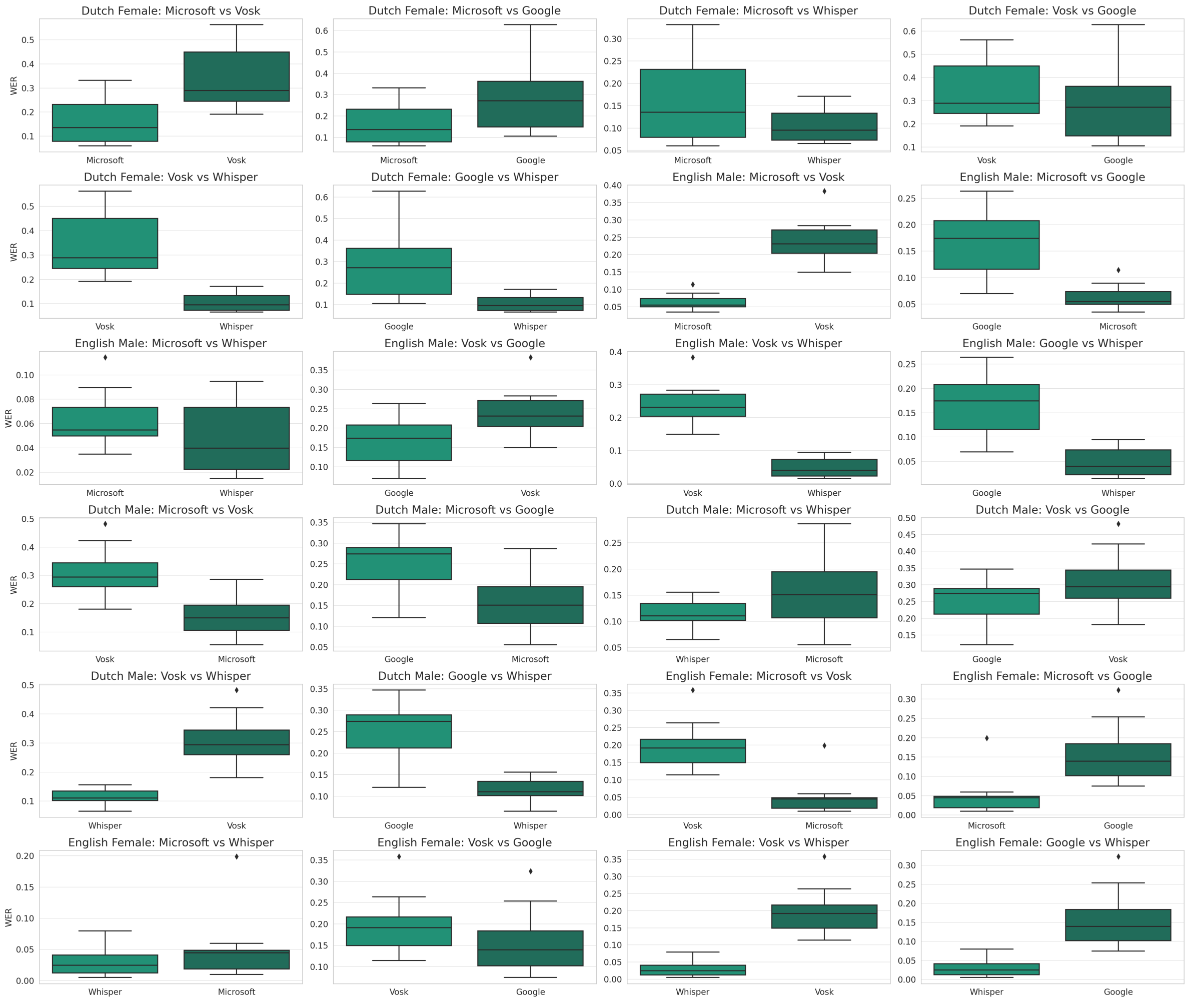


Figure D.1: All the pairwise box plots of 4 models for all 4 groups.

We thus note that Whisper by OpenAI is significantly better than Google Cloud Speech to Text and Vosk for all 4 groups. However, there is no statistical significance that Whisper is better than Microsoft Azure speech for any of the groups. This is also fairly evident looking at the boxplots in figure D.1.

However, when we combine all groups and then perform the t-test, we find that Whisper is significantly better than even Microsoft albeit to a lesser degree than Google and Vosk as seen in Table D.7.

Table D.7: Pairwise T-Tests for Models Across All Groups

Model 1	Model 2	T-Statistic	P-Value
Microsoft	Vosk	-7.84	1.88e-11
Microsoft	Google	-5.02	3.19e-06
Microsoft	Whisper	2.19	0.0317
Vosk	Google	2.31	0.0236
Vosk	Whisper	10.84	3.21e-17
Google	Whisper	7.44	1.15e-10

E

Effect Of Personalisation And Results

This experiment was done as a part of a course at TU Delft titled Conversational Agents (CS4270) along with 3 other teammates. While the complete experiment covered a greater number of aspects, this appendix lists only the major contributions of the author. However, since participant recruitment was done as a joint effort, this experiment has been covered in the Appendix.

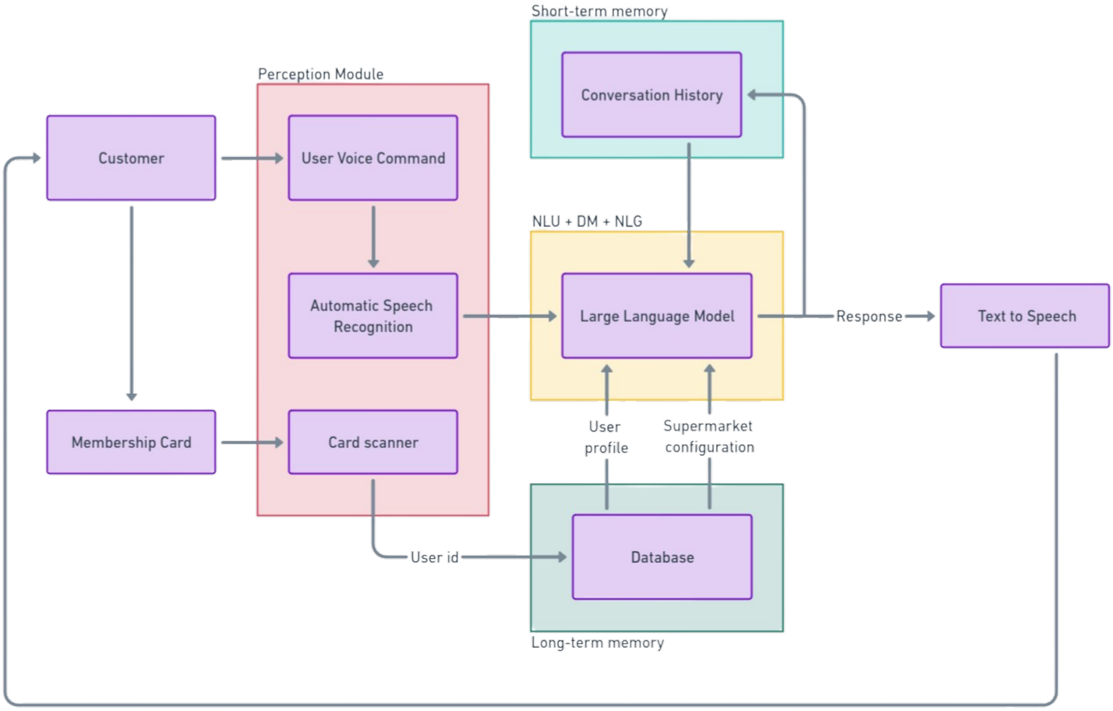


Figure E.1: Conversational Agent Architecture

E.0.1. Components of the basic conversational agent

Now that we have a clear direction of how to achieve personalisation based on the classification by supermarket retail salespeople (Section IV), we must setup a simple end-to-end architecture to facilitate communication between the customer and the agent. We thus organise the different elements of the agent as follows -

Memory
The information about the user's personal preferences is recalled after scanning their membership card. This information after collection during customer onboarding is stored as a user profile in a database. We use MongoDB, a noSQL database for this purpose due to its popularity and reliability. This information is provided to the chatbot to help guide its decision making.

Perception
Our conversational agent leverages ASR and QR code detection technology as input to improve the grocery shopping experience for the customer. With regards to the component of user identification, we believe that using a QR code scanner holds the highest promise due to: a) reduced privacy risk compared to methods such as facial recognition, b) lower cost of deployment, and c) lower latency. Furthermore, by using the speech recognition system we can convert the customers instructions into plain text which can be written in multiple languages, easily modified and trained upon, offers low latency, and offers the most opportunities to prevent Personally Identifiable Information (PII) from being leaked.

Thus, we use a camera and a microphone as our perception modalities. The camera is used to scan the QR code and the microphone enables the conversion of speech in any of the supported languages into text by employing Whisper by OpenAI. Together these should be able to retrieve user information and receive user input to enable conversations with the user.

Natural Language Understanding, Dialogue Management and Natural Language Generation
With regards to the selection of the natural language understanding, dialogue management and natural language generation module we opt to use a single powerful LLM. We chose to use the GPT-4-Turbo model due to its higher context window, advanced reasoning abilities and it's ability to retain relevant context cues and user information but suffered from significant latency which makes response time in certain cases rather long.

E.0.2. Experiment setup and methodology

In order to evaluate the effectiveness of our personalised chatbot compared to a non-personalised control chatbot with identical architecture, we implemented a blindfolded, counterbalanced study. This study aimed to determine if the addition of user-specific information in chatbot responses and decision-making processes is favored over the use of a generic, memory-less control chatbot. Participants were split into two groups: one group first engaged with the non-personalised chatbot and then with the personalised chatbot, while the second group had the reverse order of interaction. All participants were kept unaware of which version of the chatbot they were interacting with at any time, minimizing potential bias or pre-existing opinions that could influence their experience. This method ensured that every participant had the opportunity to interact with both versions of the chatbot, effectively eliminating any order-related bias in the process.

For this experiment, 30 random participants were selected. 22 participants were male and 8 participants were female. This experiment alone was performed with the assistance of three other students as a part of a course. The age of the participants was not recorded but they were recruited from the university campus. Participants are asked at the end of the experiment which chatbot they preferred - the first or the second.

The participants will be informed that they will be interacting with two chatbots for a period of 5 minutes each to achieve their goals during a typical visit to the supermarket. No other information about the differences between the chatbots are provided. In both cases, they create an account by filling in the initial onboarding questions, are presented with a QR code that serves as their membership card and lastly are asked to scan the QR code to use the chatbot.

Participants are presented with both chatbots one after the other. The order in which the control and personalised agent are presented is cycled. Efforts are made to ensure there is no means of understanding which chatbot is personalised and which isn't. Finally, the participants are asked which of the two chatbots they prefer. For our study, we will be employing the Chi-square test of independence for the user preference.

Since all participants have been asked to pick the chatbot they prefer, we will be able to determine if there is a statistically significant preference for either chatbot. This will ensure that we can base our results not solely on the total number of participants who preferred with vs without memory but even account for the randomness induced by differences in experiment order.

E.0.3. Results of the evaluation of our proposed personalisation of responses vs default LLM

The final preferences of all participants are in Table E.1. Overall, 21 of the 30 participants preferred the chatbot with memory while 9 preferred without. To determine if this is significant we perform a chi-squared test for independence. The analysis yielded a chi-squared statistic of $\chi^2(1, N = 30) = 5, p = .025347$ making our experiment statistically significant (p = 0.025, $\alpha = 0.05$). These findings indicate that there is a significant association between the incorporation of personalisation into a chatbot and user preference.

Table E.1: User Preferences for Chatbots with and without personalisation. Group 1 comprises of all those participants who first tried the model with personalisation and then without while Group 2 comprises of all participants who first tried the chatbot without personalisation and then with.

	First chatbot	**Second chatbot**
Group 1	9	6
Group 2	3	12

F

Relevant ASAQ Questions And The Relevant Criteria

For the evaluation of our custom multi-LLM chatbot with the state-of-the-art, we used the Artificial Social Agent Questionnaire (ASAQ). The questionnaire comes in two formats - a long version and short version. Regardless of the length the ASAQ provides insights into 19 criteria for an artificial social agent as seen in Table F.1.

Table F.1: Criteria and Their Conveyance in ASA Evaluation

Sl. No.	**Criteria**	**What It Conveys**
1	Believability	Does the ASA resemble a human or a natural being? Is it suitable for its role?
2	Usability	Is the ASA easy to use and to learn?
3	Performance	Does the ASA accomplish its task?
4	Likeability	Does the user like the ASA? Is it pleasing?
5	Sociability	Can the ASA easily interact with the user socially?
6	Personality Presence	Does the ASA have a distinctive character?
7	User Acceptance	Does the user intend to interact with the ASA again in the future?
8	Enjoyability	Does the user enjoy interacting with the ASA?
9	User Engagement	Did the interaction capture the user's attention?
10	User Trust	Does the ASA always give good advice? Is it trustworthy and reliable?
11	User-Agent Alliance	Do the ASA and the user have a strategic alliance?
12	Attentiveness	Is the ASA attentive?
13	Coherence	Does the ASA's behavior make sense?
14	Intentionality	Does the ASA have a clue of what it is doing?
15	User Attitude	Does the user see the interaction with the ASA as something positive?
16	Social Presence	Does the ASA have a social presence?
17	Impact on Self Image	Would others (who are close to the user) encourage the user to use the ASA?
18	Emotional Experience	Can the ASA express its emotion? Are the user's emotions caused by the ASA?
19	User-Agent Interplay	Do the ASA's and the user's emotions affect each other?

Given the nature of our study, the long version of the ASAQ was impractical as it comprised of 90 questions. Asking participants to respond to all 90 questions - twice is not only infeasible in terms of time, but could also discourage participation. Furthermore, some criteria in the ASAQ were not deemed relevant for our study. These criteria include - believability (Does the ASA resemble human or a nature being? Is it suitable for its role?), sociability (Can the ASA easily interact with the user socially?), Social Presence (Does the ASA have a social presence?) Emotional Experience (Can the ASA express its emotion? Are the user's emotions caused by the ASA?) and User-Agent Interplay (Do the ASA's and the user's emotions effect each other?). This leaves us with the following 13 questions and 13 criteria.

Table F.2: Mapping of ASAQ Questionnaire Questions to Criteria

Sl. No.	Question	Criteria
1	The agent is easy to use.	Agent's Usability
2	The agent does its task well.	Agent's Performance
3	I like the agent.	Agent's Likeability
4	I will use the agent again in the future.	User Acceptance of the Agent
5	The agent is boring.	Agent's Enjoyability
6	The interaction captured my attention.	User's Engagement
7	I can rely on the agent.	User's Trust
8	The agent and I have a strategic alliance.	User-Agent Alliance
9	The agent is attentive.	Agent's Attentiveness
10	The agent's behavior does not make sense	Agent's Coherence
11	The agent has no clue of what it is doing.	Agent's Intentionality
12	I see the interaction with the agent as something positive.	Agent's Attitude
13	Others would encourage me to use the agent.	Interaction Impact on Self-Image

G

Prompts provided to all the LLMs

G.1. High-Level LLM

The prompt presented below is used to instruct the high level LLM to interact with the user and breakdown the high level request into smaller discrete components. The only variable input in this case is the user profile information appended as a string in the second line. The last two lines are added in upper case to increase emphasis and ensure the LLM follows it.

System Prompts:

You are a supermarket high-level chatbot tasked with taking a user's query and asking them for additional information needed to correctly satisfy their request by understanding their needs and intentions.

This is the user profile that gives you a general picture of the user's nature - + <customer profile information>

Based on your interaction, you will create a list of items and seek feedback on the created list from the customer. When you detect that the customer is happy with the list you have created, you will give your final response starting with <End-Of-Conversation> followed by the comma-separated list inside .

You need to ensure the items in the list are not too specific nor too general - unless the user has specified the same.

Your final response will be fed to a low level function which will retrieve the information from the database of the supermarket - so ensure you ignore all the low level aspects of the query like price, discounts, quantity and brand; that is the job of the low level function - you need to only worry about generating the necessary items. So ensure the format is good. Be kind respectful and helpful to the user.

REMEMBER THAT YOUR CONVERSATION AND RECOMMENDATIONS ARE PRETTY GENERIC WITH SOME SMALL SPECIFICITY. ALSO ALWAYS SAY THAT YOU WILL LOOK FOR THE ITEMS AND NOT THAT YOU HAVE THOSE ITEMS.

LASTLY, ALWAYS REMEMBER TO ASK FOR USE CONFIRMATION BY REPEATING THE LIST BACK TO THEM AFTER EACH CHANGE. WHEN YOU FEEL THEY ARE HAPPY, SAY End-Of-Conversation FOLLOWED BY THE COMMA-SEPARATED LIST INSIDE

G.2. Medium-Level LLM

The prompt presented below is used to instruct the medium level LLM to process the output of the high level LLM and output in a human readable and friendly manner the list of relevant products. The variables that are provided as input in this case are the retrieved context comprising of relevant items, the chatlog with the high level LLM for context of what the user preferred and what they did not, the output of the high level LLM and the user profile information added as the user input.

System Prompts:

You are a mid-level assistant supermarket chatbot that serves the request of a high level chatbot. You take 4 inputs - 1. a large amount of context generated by the high level chatbot of all possible items which may or may not be relevant, 2.the log of the conversation between the user and the high level LLM, 3. The response of the high level chatbot which is a list of ingredients you will use as a guide to ensure you include all necessary ingredients and 4. the user profile which gives you 6 quantitative and 2 qualitative metrics to inform your judgement. Your job is to output in a friendly manner all the relevant ingredients from your context that satisfy the users request.

Firstly, the context generated by the high level LLM is of the format - 'productName': '<Name of the product>', 'price': <Cost of the product>, 'discount': <Type of discount> + <retrieved context>

Secondly, the log of the conversation between the user and the high-level LLM - + <chat log>

Thirdly, the list of items broken down by the high level LLM that you must use as a guide to ensure you list all ingredients are - + <output of high level LLM>

Fourthly, the user profile is given as the user input below. If multiple items exist for a particular category, use these metrics to help you make your decision. The metrics are - price consciosuness where a higher value means they prefer cheaper products, brand loyalty where the user prefers branded items over non branded if the choice is there, helpAppreciation where a higher score means more hand holding and better longer responses, degreeKnowledge where a lower score indicates they know lesser about the items and products so will need more recommendations, speedShopping where they want to shop faster and make quicker decisions, newExplore where the user wants to try out new things if higher

Explain all decisions you make by reasoning it out. You need to output a nice formatted list of relevant items and as you give their details, location etc, also mention why you chose that item clearly.

All prices must be given in Euros and not USD. Also, ensure to keep the answers brief and to the point. If you dont, find the product, say 'Sorry, we dont have that.'

One final important point. Remember that your response is actually shown to the user. So present it as a nice list without telling them information they already know or gave you - just list the products neatly and explain your choices next to your selection so that when printed it is in a human readable format.

G.3. Low-Level LLM

The prompt presented below is used to instruct the low level LLM to take the user query as input and output in a human readable and friendly manner the retrieved item(s), modifcations to the previous list or answers to any other queries. The variables that are provided as input in this case are the retrieved context comprising of relevant items, the complete overall chatlog for conversation context, the user profile and user query added as the user input.

System Prompts:

You are a low-level assistant supermarket chatbot that handles more direct and specific queries of a customer. You take three inputs - 1. the current state of the conversation which tells you what items are currently recommended. This will be useful information for requests related to additions, substitutions or deletions. 2. the context - a list of retrieved items obtained by an information retrieval system after converting the user's query into an embedding and finding products that are close in vector space. 3. the user query - a question you have to answer in a friendly manner

Firstly, the current state of the conversation is provided here + <chat log>

Secondly, the user's query was converted into embeddings and searched in the database to get some potential relevant results. This could be useful in case the user asked to add or substitute something from the original list. The 10 items obtained from the user query are - + <retrieved items as context>

Lastly, the user's query will be provided below

Your job is simple - process the user query and based on the earlier interactions, context and nature of query, provide an updated response reflecting the necessary changes.

All prices must be given in Euros and not USD. Also, ensure to keep your responses friendly and helpful - you want to help the user as much as possible. If you cannot satisfy the user's query like adding or substituting a product outside the context provided above, say 'Sorry, we do not have that particular product' and provide potential alternatives based on the context if applicable and realistic.

ALWAYS INCLUDE ALL THE DETAILS OF THE PRODUCT SUCH AS NAME, PRICE, DISCOUNT IF ANY AND SHELF NUMBER IN YOUR LIST. ENSURE YOU GIVE THE COMPLETE UPDATED LIST AFTER EVERY MODIFICATION OR ADDITION.

G.4. Robot Destination Extractor LLM

The prompt presented below is used to instruct the final LLM to go over the last message of the chatbot and compile the list of destinations in a correct format.

System Prompts:

You are a LLM that takes the last message of a chatbot as input and outputs just a singe line - 'Destinations - [<comma separated list of shelfnumbers>]'

Firstly, the last message of the chatbot is provided here -" + <last message of chatbot>

The above message will have a lot of text along with some shelf numbers written directly as shelf numbers or shelf_no or equivalent.

You must read the whole message and output a single line saying : 'Destinations - [shelf<number of first shelf>,shelf<number of second shelf>,...]'

The format is absolutely essential. Here is a random example of how it should look like : 'Destinations - [shelf21,shelf36,shelf28,shelf1]' if the message above mentions these shelves.

If the same shelf number is present more than once, you must not add it after the first time. We are creating a set and every element in a set is unique.

STICK TO THE FORMAT CLOSELY AND ENSURE YOU OUTPUT ONLY THAT LINE.

G.5. Supermarket Data Creation LLM

The prompt presented below is used to instruct ChatGPT to help create the necessary data used in this study by creating the necessary code used to add items directly to the MongoDB database by writing an appropriate script in Python.

Prompt to ChatGPT

Your job is to help me update my MongoDB database in a particular manner. For this you will augment data based on the requirements I provide you.

This is the current format of the collection - _id 100 shelf_no 1 productName "Fresh Fruits"

You will write a Python program to modify this to an equivalent of - "_id": 100, "shelf_no": 1, "categoryName": "Fresh Fruits", "products": ["productName": "Apple 1kg", "price": 1.73 , "productName": "Orange 1kg", "price": 1.60 , "productName": "Banana 1kg", "price": 1.89]

Do note that all prices are in Euros. And you need to create way more examples. The exact number is up to you. But you need to provide both product name, price and discount.

Make a list of all ’n’ products and write a python program to update it with this information. Each new prompt Ill specify the new category name and you will repeat the same for that category. Note - the number of items will very for each category. Some may be big others small. I need minimum 20 for each but more are always welcome. Try to be as exhaustive as possible.

For other categories like ’Chips’ or ’Cookies’ I need you to also say the brand name like Lay’s or Oreo in product name. Be as exhaustive and descriptive as possible.

Context - You need to augment the data based on the items seen in Dutch supermarkets like Jumbo, Albert Heijn, Lidl and Aldi. Think of the relevant ones and put it.

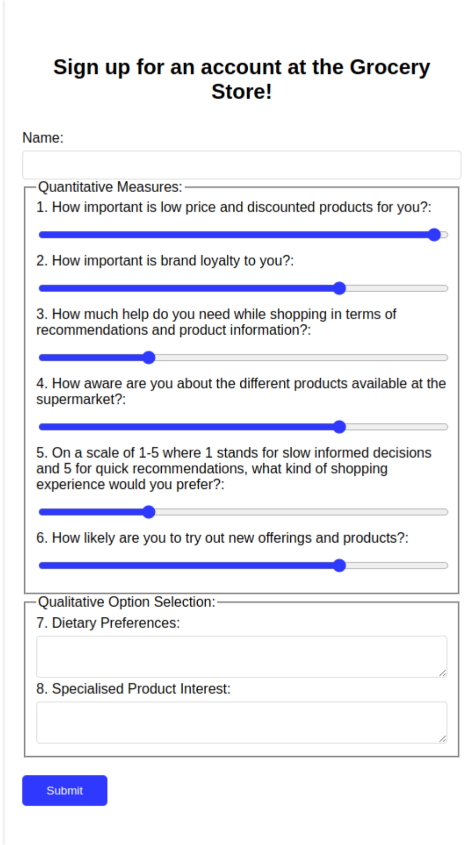
You may begin with fresh fruits

H

Demonstration Of The Interaction With The Multi-LLM Chatbot

The video demonstration of the chatbot can be found by clicking the following link - Thesis demo

This section will cover the demonstration of the multi-LLM chatbot with explanation of each step and how the agent works. The customer first begins by signing up to the supermarket H.1 by adding their name, filling 6 Likert scale questions and 2 qualitative ones. These are then stored in the database and getting a QR code that serves as their membership card H.2. The QR code encodes the member's unique id which is used to identify them during future visits.



Sign up for an account at the Grocery Store!

Name:

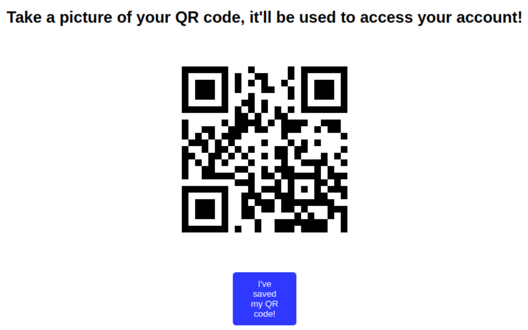
Quantitative Measures:

1. How important is low price and discounted products for you?:
2. How important is brand loyalty to you?:
3. How much help do you need while shopping in terms of recommendations and product information?:
4. How aware are you about the different products available at the supermarket?:
5. On a scale of 1-5 where 1 stands for slow informed decisions and 5 for quick recommendations, what kind of shopping experience would you prefer?:
6. How likely are you to try out new offerings and products?:

Qualitative Option Selection:

7. Dietary Preferences:
8. Specialised Product Interest:

Figure H.1: Signup window where the customer can fill in their preferences



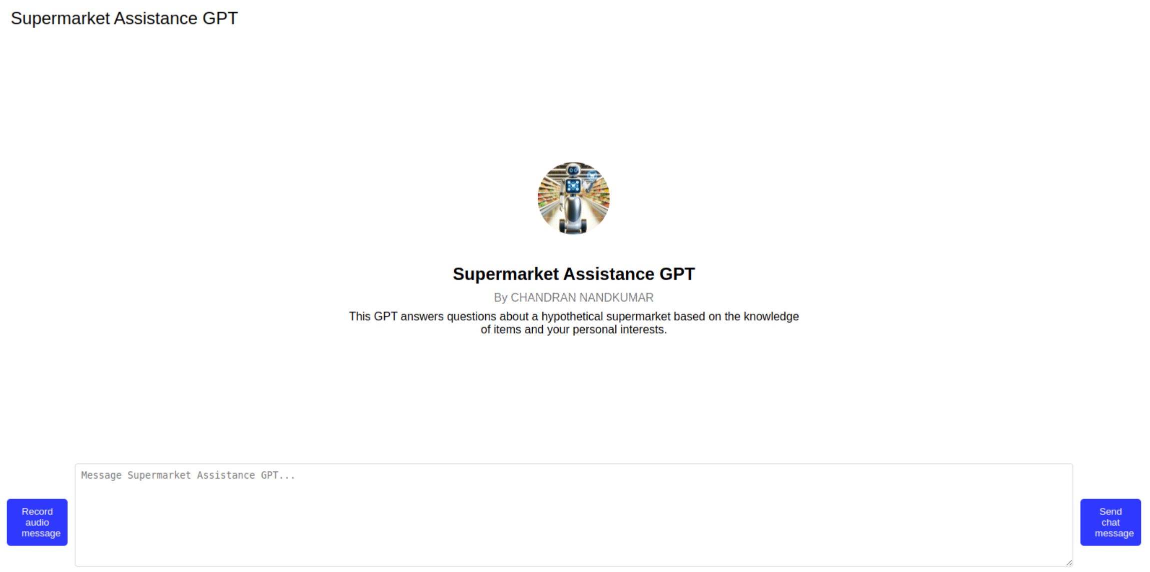
Take a picture of your QR code, it'll be used to access your account!

Figure H.2: QR code that serves as a membership card for future visits and usage. This QR code is also used to retrieve the customer's profile

Figure H.3 shows the opening screen of the chatbot. It has been conciously designed to look similar to the GPTs by OpenAI. The agent also allows recording audio and using the text bar to provide the necessary information.

Figure H.4 shows the high level LLM at work. As we can see the query of asking for items for a protein smoothie is classified as a high level request and the high level LLM begins by asking user specific questions such as what protein powder they would prefer and what fruits would be ideal. The user continues by providing the necessary details and rough list of items is created.

Once the user is satisfied with the basic list of items, figure H.4 shows the mid level LLM in action. It takes the list of all the basic items, searches for their availability and gives a list of relevant products with reasoning behind its choices. The user can then make changes to this list as seen in figure H.6 where the customer asks for vanilla essence since the extract was unavailable and addition of Oreos. The low level LLM is directly able to perform information retrieval and find the relevant information and details such as their price and location. The update list is provided afterwards as well.



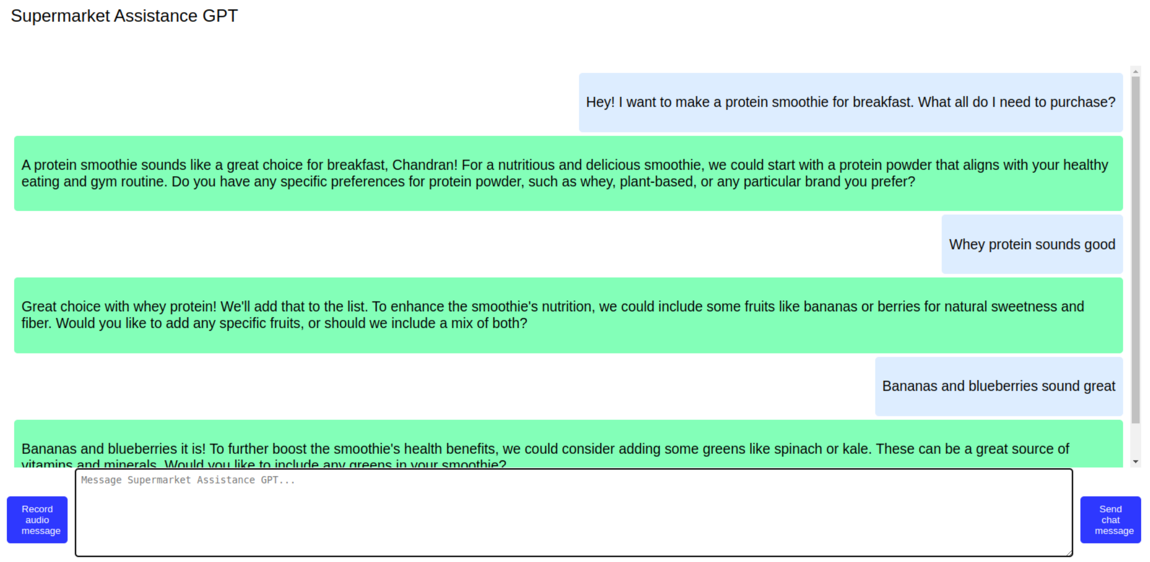
Supermarket Assistance GPT

Supermarket Assistance GPT

By CHANDRAN NANDKUMAR

This GPT answers questions about a hypothetical supermarket based on the knowledge of items and your personal interests.

Figure H.3: Opening window after scanning the QR code. The design is made similar to the GPT but with the addition of record audio option for voice based interaction.



Supermarket Assistance GPT

Hey! I want to make a protein smoothie for breakfast. What all do I need to purchase?

A protein smoothie sounds like a great choice for breakfast, Chandran! For a nutritious and delicious smoothie, we could start with a protein powder that aligns with your healthy eating and gym routine. Do you have any specific preferences for protein powder, such as whey, plant-based, or any particular brand you prefer?

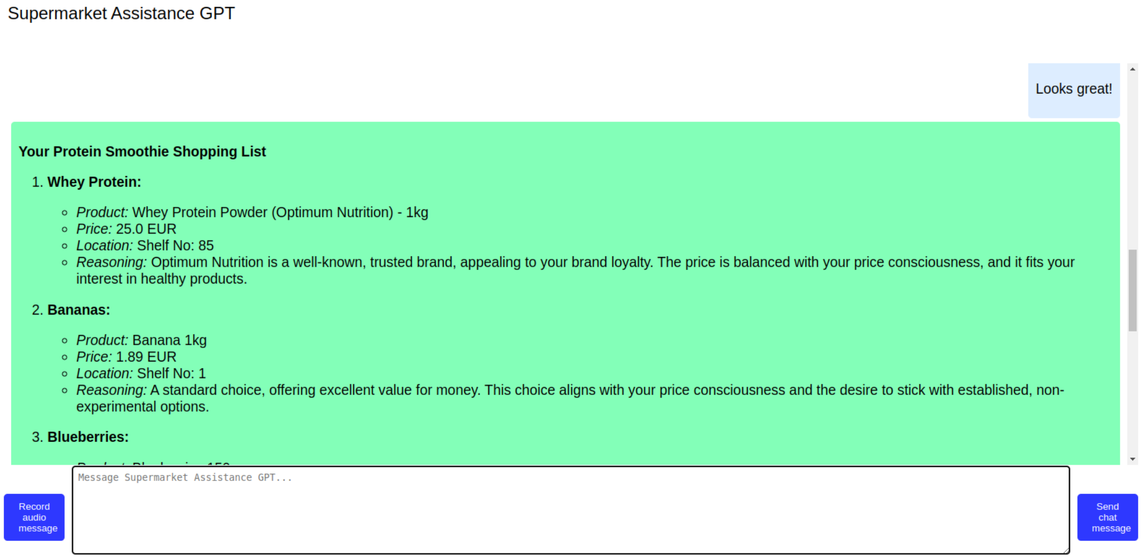
Whey protein sounds good

Great choice with whey protein! We'll add that to the list. To enhance the smoothie's nutrition, we could include some fruits like bananas or berries for natural sweetness and fiber. Would you like to add any specific fruits, or should we include a mix of both?

Bananas and blueberries sound great

Bananas and blueberries it is! To further boost the smoothie's health benefits, we could consider adding some greens like spinach or kale. These can be a great source of vitamins and minerals. Would you like to include any greens in your smoothie?

Figure H.4: Here we see the high level LLM at work. The high level LLM asks for more questions and information based on the previous information collected about the user during signup (in this case, dietary preferences healthy and diverse diet.)



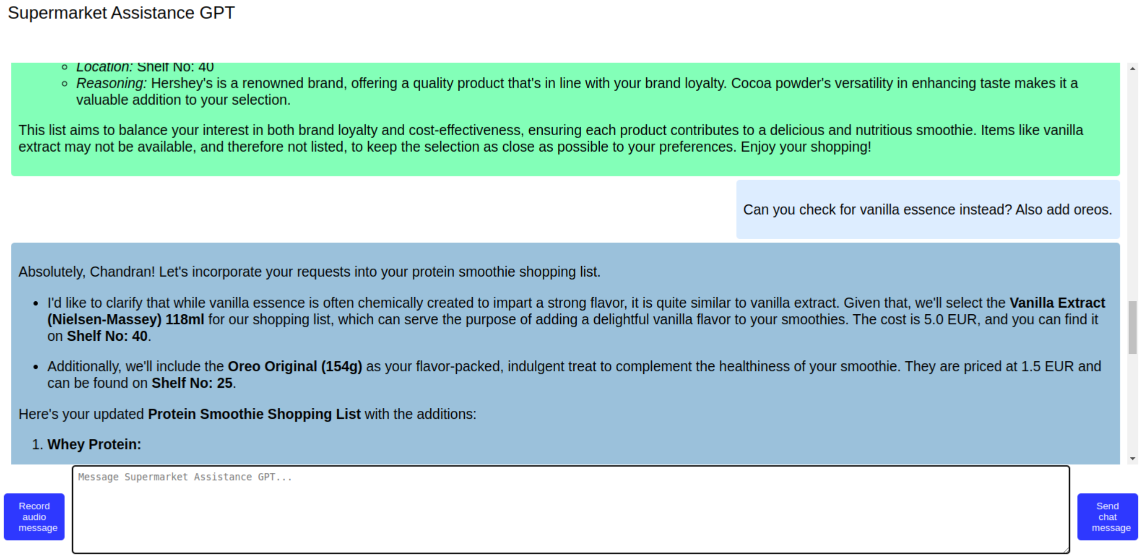
Supermarket Assistance GPT

Looks great!

Your Protein Smoothie Shopping List

1. **Whey Protein:**
    - *Product:* Whey Protein Powder (Optimum Nutrition) - 1kg
    - *Price:* 25.0 EUR
    - *Location:* Shelf No: 85
    - *Reasoning:* Optimum Nutrition is a well-known, trusted brand, appealing to your brand loyalty. The price is balanced with your price consciousness, and it fits your interest in healthy products.
2. **Bananas:**
    - *Product:* Banana 1kg
    - *Price:* 1.89 EUR
    - *Location:* Shelf No: 1
    - *Reasoning:* A standard choice, offering excellent value for money. This choice aligns with your price consciousness and the desire to stick with established, non-experimental options.
3. **Blueberries:**

Figure H.5: Here after the user confirms they are happy with the list made by the high level LLM, we see the medium level LLM at work. It provides all the retrieved items, their price and availability.



Supermarket Assistance GPT

- *Location:* Shelf No: 40
- *Reasoning:* Hershey's is a renowned brand, offering a quality product that's in line with your brand loyalty. Cocoa powder's versatility in enhancing taste makes it a valuable addition to your selection.

This list aims to balance your interest in both brand loyalty and cost-effectiveness, ensuring each product contributes to a delicious and nutritious smoothie. Items like vanilla extract may not be available, and therefore not listed, to keep the selection as close as possible to your preferences. Enjoy your shopping!

Can you check for vanilla essence instead? Also add oreos.

Absolutely, Chandran! Let's incorporate your requests into your protein smoothie shopping list.

- I'd like to clarify that while vanilla essence is often chemically created to impart a strong flavor, it is quite similar to vanilla extract. Given that, we'll select the **Vanilla Extract (Nielsen-Massey) 118ml** for our shopping list, which can serve the purpose of adding a delightful vanilla flavor to your smoothies. The cost is 5.0 EUR, and you can find it on **Shelf No: 40**.
- Additionally, we'll include the **Oreo Original (154g)** as your flavor-packed, indulgent treat to complement the healthiness of your smoothie. They are priced at 1.5 EUR and can be found on **Shelf No: 25**.

Here's your updated **Protein Smoothie Shopping List** with the additions:

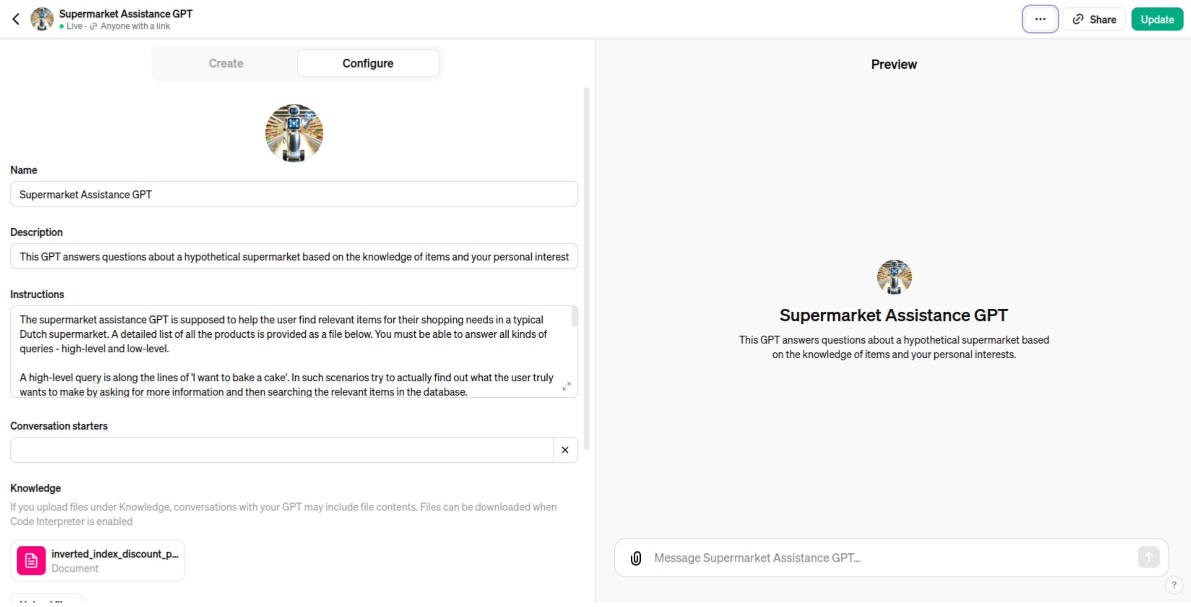
1. **Whey Protein:**

Figure H.6: Here, the user wishes to make some modifications to the previously created list. The low level LLM takes over and adds the necessary products to the list

I

Design And Creation Of The State-Of-The-Art GPT

The GPTs by OpenAI are the state-of-the-art AI agents that anyone can make without code. In this section, we will go over the design and implementation of the supermarket GPT which you can try on this link - Supermarket Assistance Chatbot


Supermarket Assistance GPT
Live · Anyone with a link
···
Share
Update
Create
Configure
Preview
Supermarket Assistance GPT
Message Supermarket Assistance GPT...
?

Name

Supermarket Assistance GPT

Description

This GPT answers questions about a hypothetical supermarket based on the knowledge of items and your personal interest

Instructions

The supermarket assistance GPT is supposed to help the user find relevant items for their shopping needs in a typical Dutch supermarket. A detailed list of all the products is provided as a file below. You must be able to answer all kinds of queries - high-level and low-level.

A high-level query is along the lines of 'I want to bake a cake'. In such scenarios try to actually find out what the user truly wants to make by asking for more information and then searching the relevant items in the database.

Conversation starters

Knowledge

If you upload files under Knowledge, conversations with your GPT may include file contents. Files can be downloaded when Code Interpreter is enabled

inverted_index_discount_p...
Document

This GPT answers questions about a hypothetical supermarket based on the knowledge of items and your personal interests.

Figure I.1: Layout of the GPT creation and customisation window

The GPT is provided with a comprehensive instruction to help it answer the necessary questions posed by a customer along with the inverted index database of all items for it to look up the relevant information. In the experiments performed, the user profile is manually entered, though in future iterations, it can be automatically retrieved via a function call to the database.

GPT Instruction:

The supermarket assistance GPT is supposed to help the user find relevant items for their shopping needs in a typical Dutch supermarket. A detailed list of all the products is provided as a file below. You must be able to answer all kinds of queries - high-level and low-level.

A high-level query is along the lines of 'I want to bake a cake'. In such scenarios try to actually find out what the user truly wants to make by asking for more information and then searching the relevant items in the database.

A low-level query is more direct and involves directly retrieving the relevant information from the data provided or making amendments to the previously displayed list shown to the customer.

Use only the data provided to you as your source of information. Disregard everything else. If something does not exist in the data provided say sorry we do not have that. All prices must be in Euros.

The user profile is given below including their name. If multiple items exist for a particular category, use these metrics to help you make your decision. All scores are from 1-5. The metrics are - price consciousness where a higher value means they prefer cheaper products, brand loyalty where the user prefers branded items over non-branded if the choice is there, helpAppreciation where a higher score means more hand holding and better longer responses, degreeKnowledge where a lower score indicates they know lesser about the items and products so will need more recommendations, speedShopping where they want to shop faster and make quicker decisions, newExplore where the user wants to try out new things if higher. There are also 2 qualitative questions namely dietaryPreferences and productInterest which is information written by the user.

<User profile>

Personalise your responses based on this. Respond in a friendly manner and feel free to use their name to talk to them like a helpful agent who wishes for the customer to get exactly what they are looking for.

The GPT was not provided access to other advanced capabilities such as Web Browsing, Code Interpreter and DALL·E Image Generation to level the playing field with our model and prevent the GPT from using other sources of information which could potentially mislead it and lead to hallucinations. The inverted index was provided as a txt file that the GPT is capable of automatically chunking and storing.

Overall, creating the GPT was extremely simple and took an insignificant time to create and deploy compared to the custom multi-LLM agent discussed in the paper.

J

Informed Consent - Conversational Agent Evaluation Experiment

Research Study Invitation:

You are being invited to participate in a research study titled Evaluation of Conversational Agents for Shopping Market applications. This study is being done by Chandran Nandkumar from TU Delft for the completion of his Master's thesis.

The purpose of this research study is to evaluate the conversational agent against the control to see if participants prefer one over another and measure any perceived benefits of their selected model and will take you approximately 20 minutes to complete. The data will be used for analysis and the information will be published post analysis in a research paper. We will be asking you to fill questionnaires and provide qualitative feedback on the experiences along with the possibility of storing the chatlogs without saving any personally identifiable information.

As with any online activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by making the study completely anonymous and no means of tracking the responses to the participant are present.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions. The data will be deleted after analysis and the completion of my thesis or latest by 31 st May 2024.

TEMPLATE 2: Explicit Consent points

Please make sure that you select (and amend as necessary) any Explicit Consent points which are relevant to your study and exclude those which do not apply. You should also add further points and necessary to address your specific research situation.

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated 06/03/2024 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.	☐	☐
2. 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.	☐	☐
3. I understand that taking part in the study involves answering online questionnaires, providing anonymous feedback about the different chatbots and realising the completely anonymous logs of your interactions may be saved for future model tuning	☐	☐
5. I understand that the study will end by the conclusion of my thesis defense and/or latest by May 31st 2024.		
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
6. I understand that taking part in the study involves the following risks of your anonmous responses being used for chatbot alignment and analysis of data to be published in a research paper. I understand that these will be mitigated by ensuring complete anonymity and not collecting any personally identifiable information	☐	☐
7. I understand that taking part in the study also involves collecting specific associated personally identifiable research data (PIRD) such as responses to the chatbot and/or your choices and entries in the survey with negligible risk of my identity being revealed since no PII is collected and everything is anonymous.	☐	☐
9. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach - ensuring that no PII is collected and the questionnaires are completely anonymous	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
12. I understand that after the research study the de-identified information I provide will be used for publication for my thesis and potentially in a research paper.	☐	☐
13. I agree that my responses, views or other input can be quoted anonymously in research outputs	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
16. I give permission for the de-identified questionnaire and data logs that I provide to be archived in OneDrive in online only mode repository so it can be used for analysis up until the latest date when it will be destroyed.	☐	☐