Using Large Language Models for Adaptive Dialogue Management in Digital Telephone Assistants

Hassan Soliman*
German Research Center for Artificial Intelligence (DFKI)
Berlin, Germany
hassan.soliman@dfki.de

Miloš Kravčík
German Research Center for Artificial Intelligence (DFKI)
Berlin, Germany
milos.kravcik@dfki.de

Nagasandeepa Basvoju
German Research Center for Artificial Intelligence (DFKI)
Berlin, Germany
nagasandeepa.basvoju@dfki.de

Patrick Jähnichen
Aaron GmbH
Berlin, Germany
patrick.jaehnichen@aaron.ai

ABSTRACT

The advent of modern information technology such as Large Language Models (LLMs) allows for massively simplifying and streamlining the communication processes in human-machine interfaces. In the specific domain of healthcare, and for patient practice interaction in particular, user acceptance of automated voice assistants remains a challenge to be solved. We explore approaches to increase user satisfaction by language model based adaptation of user-directed utterances. The presented study considers parameters such as gender, age group, and sentiment for adaptation purposes. Different LLMs and open-source models are evaluated for their effectiveness in this task. The models are compared, and their performance is assessed based on speed, cost, and the quality of the generated text, with the goal of selecting an ideal model for utterance adaptation. We find that carefully designed prompts and a well-chosen set of evaluation metrics, which balance the relevancy and adequacy of adapted utterances, are crucial for optimizing user satisfaction in conversational artificial intelligence systems successfully. Importantly, our research demonstrates that the GPT-3.5-turbo model currently provides the most balanced performance in terms of adaptation relevancy and adequacy, underscoring its suitability for scenarios that demand high adherence to the information in the original utterances, as required in our case.

CCS CONCEPTS

• **Human-centered computing → Systems and tools for interaction design**.

KEYWORDS

Adaptive Dialogue Systems, Large Language Models, Digital Telephone Assistants

ACM Reference Format:
Hassan Soliman, Miloš Kravčík, Nagasandeepa Basvoju, and Patrick Jähnichen. 2024. Using Large Language Models for Adaptive Dialogue Management in Digital Telephone Assistants. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), July 01–04, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3631700.3664902

1 INTRODUCTION

Despite the increasing technical maturity and availability of commercial conversational systems, their user acceptance still provides an opportunity for an improvement. For instance, a German study [15] showed the potential to relieve hotline staff, nevertheless a personal conversation cannot be fully replaced, especially due to often complex concerns and the need for human empathy. The cause of drop-outs is often assumed to be that the machine conversation elements are not personalised enough [3]. The challenge is to increase acceptance of telephone-based voice assistants used by patients and relieve staff at general medical practices. For that purpose, we research adaptability of the dialogue process to the needs of the caller as well as to relevant contextual information. The optimisation of the dialogue flow aims at the one hand at the appropriate, dynamic formulation taking into account the input parameters, and on the other hand at the appropriate and natural prosody (i.e. course of speech rate, voice pitch, intonation or emphasis as well as pauses in speech). Overall, it is a multidimensional optimisation problem with a correspondingly large parameter space.

Previous research [13] led to the identification of relevant success parameters in the course of the dialogue, such as demographic factors (gender, age), personality traits (static) and emotional states (dynamic). We try to indicate the relevant characteristics and personas (user groups) that can be used for the purposes of adaptation and personalisation, e.g. speed of speech, choice of words, phrasing and sentence length (text readability measures) can be adapted for alternative age groups and language skills.

Conversational Artificial Intelligence [10] systems have witnessed remarkable advancements in recent years, with a primary objective being the enhancement of user satisfaction. Central to achieving this objective is the dynamic adaptation of user utterances to specific contextual parameters. In this paper, we conduct an extensive evaluation of various Large Language Models (LLMs)

UMAP Adjunct '24, July 01–04, 2024, Cagliari, Italy

ACM ISBN 979-8-4007-0466-6/24/07
https://doi.org/10.1145/3631700.3664902

and tools to ascertain their efficacy in optimizing user satisfaction through utterance adaptation. As language models continue to evolve, the ability to tailor conversational responses to nuanced parameters, such as age group, sentiment, and gender, has become increasingly crucial. The overarching aim is to discern the most fitting model for the task of user utterance adaptation, a pivotal aspect of conversational AI systems. Our investigation revolves around comprehensively evaluating the performance of these models and tools, scrutinizing factors that include speed, cost-efficiency, and accuracy.

Consequently, the research goals of this paper are to explore the use of LLMs to dynamically adapt user-directed utterances based on contextual parameters such as age group, sentiment, and gender. We aim to examine various factors, including speed, cost, and quality of adaptations. Additionally, we seek to evaluate and compare the effectiveness and performance of various models in adapting these utterances. The objective is to identify the most suitable model for utterance adaptation in conversational AI systems. This involves balancing the relevance of the adaptations to the original utterance with their adequacy in adapting to user-specific parameters.

Our paper is structured as follows: First, we introduce related work. Then we present a study investigating how to adapt the dialogue with patients using generative AI. Finally, we summarize our findings and outline the future work.

2 RELATED WORK

In the healthcare area, we can find interesting digital assistants helping people in certain situations. A study about communicability of conversational interfaces in healthcare and smart home domains showed that for simple tasks with a clear goal (problem-based), a chatbot experience can be a promising solution for improving the communication between users and systems [14]. But there are also attempts to help in more complex situations, like a smart, pervasive chatbot for emergency case assistance that helps victims or incident witnesses to avoid deterioration of the subject's condition until the aid arrives [9]. Or a conversational bot to increase the patient's access to healthcare knowledge and leverage the potentials of AI to bridge the gap of demand and supply of human healthcare providers [1].

Recent advancements in natural language processing have seen an increasing interest in leveraging LLMs to enhance dialogue systems. For instance, [6] investigated the impact of very large language models (VLLMs) on dialogue systems and provides valuable insights into the performance of VLLMs in generating coherent and contextually relevant responses across different dialogue contexts. Similarly, [5] explored the use of LLMs, such as OpenAI's GPT series or Google's BERT, as user simulators to generate diverse user utterances for training and evaluating dialogue systems. By employing LLMs as user simulators, researchers aim to address the challenge of collecting sufficient and diverse user feedback for dialogue system development. Recent studies have proposed several metrics that are similar to human evaluations, and LLMs have been shown to be a promising replacement for human judges. A comprehensive study [16] on the application of LLMs for automatic dialogue evaluation showed interesting results: proprietary models have superior evaluation capabilities compared to open-source LLMs; LLMs excel in evaluating coherence, relevance and overall quality more than specificity and diversity; LLMs show promise in automatic dialogue evaluation, but it is still an open problem.

A methodology for the data-driven evaluation of telephone voice assistance systems [12] was developed to increase user-friendliness and cost efficiency. It provides rather general, static optimisation advice that is not user-specific and cannot be used by AI-based telephone assistants. Another attempt [11] deals with the automated organisation of telephone call flows, focusing on the static optimisation of decision-making with voice assistants. Apart from these rather general optimisation techniques, ours is the first approach, to the best of our knowledge, that focuses on automatic real-time optimisation based on call context and operator-specific requirements.

With the advancement in unobtrusive technologies for the detection of personality and emotions, these static and dynamic user characteristics became interesting in the domain of personalized systems [13]. Considering the actual elicitation opportunities in patient dialogues, we selected several demographic and sentiment parameters for the adaptation purposes in our study.

3 ADAPTIVE DIALOGUE MANAGEMENT

Our approach to adjust a dialogue for specific circumstances is based on the insight that it should be personalized, as this is what is perceived also in a human-to-human interaction. We use LLMs for the personalization purposes. Taking into account certain user traits, we adapt user-directed utterances using enhanced prompts (Figure 1). In this section we first introduce the dataset and the language models considered for our purpose. Afterwards, we present our implementation, evaluation metrics and experimental results.

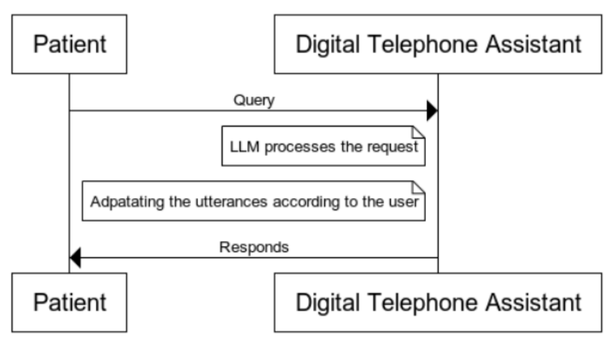

Figure 1: Illustrating the dialogue flow between the Patient and Digital Telephone Assistant

3.1 Dataset

Our dataset encompasses a diverse array of conversational scenarios that are commonly encountered in medical practices. These scenarios range from appointment scheduling and prescription inquiries to handling general health questions and emergency situations. Real time calls are being recorded and using 'Speech to Text', text data is being extracted and preprocessed in order to anonymize the personal details in the call.

The dataset includes a rich variety of user utterances (in German), capturing different communication styles, language proficiency levels, and emotional states. Each dialogue instance is enriched with contextual information, providing details about the patient's age, gender, and emotional state, along with corresponding responses from medical staff. This contextual richness ensures the dataset's ability to facilitate the development of intelligent dialogue systems capable of adapting to a wide spectrum of user demographics and interactions.

In addition to its technical merits, the dataset upholds ethical standards and user privacy. All data included in the dataset is anonymized and stripped of personally identifiable information, adhering to stringent data protection regulations and ensuring the confidentiality and privacy of the individuals involved. The dataset consists of 60 machine utterances covering various combinations of user's dialogue states. These are different user's parameters and their possible values:

(1) **Gender**: Male, Female
(2) **Age group**: 0-50, >50
'0-50' represents individuals from childhood to middle age, and '>50' represents older adults.
(3) **Sentiment**: Positive, Negative, Neutral
This reflects the user's emotional state. 'Positive' indicates a happy or pleased state, 'Neutral' indicates a lack of strong emotion, and 'Negative' shows dissatisfaction or unhappiness.

Here we provide a couple of examples of machine utterances (first in original, then translated):

- Das kam bei mir nicht an. Sind Sie zu Qualitätszwecken mit einer Aufzeichnung einverstanden? (That didn't reach me. Do you agree to a recording for quality purposes?)
- Okay! Ich würde Sie jetzt gerne weiterleiten, doch Sie rufen außerhalb der Sprechzeiten an. (Okay! I would like to forward you now, but you are calling outside office hours.)

3.2 Utterance Adaptation

To enhance user interactions, we explored the use of LLMs for utterance adaptation, identifying several potential models for our needs.

Proprietary models such as GPT-3.5-turbo and GPT-4-turbo, developed by OpenAI[1], are part of the Generative Pre-trained Transformer series. These models are engineered to produce text that closely mimics human responses based on provided input. For the purpose of adapting utterances in conversations, both models are suitable; however, we selected GPT-3.5-turbo for its cost-effectiveness. Notably, GPT-4-turbo is approximately 20 times more expensive than GPT-3.5-turbo.

In the open-source domain, models like Mistral-8x7b from Mistral AI[2] and Llama-2-70b from Meta AI[3] also provide adaptable solutions for utterance adaptation. These models were accessed via Groq Cloud[4], which offers API services for model integration.

[1]https://openai.com
[2]https://mistral.ai/
[3]https://ai.meta.com/
[4]https://wow.groq.com/

We compared these open-source models with GPT-3.5-turbo, noting that they achieve comparable performance on various LLM benchmarks. For instance, on the Massive Multitask Language Understanding (MMLU) benchmark [4], which evaluates an LLM's comprehension and problem-solving capabilities across 57 diverse tasks, Mistral-8x7b, Llama-2-70b, and GPT-3.5-turbo scored 70.6%, 69.9%, and 70.0%, respectively.

At the time of this paper's submission, Meta AI introduced the Llama-3-70b model, a successor to the Llama-2-70 series. We utilized instruction-fine-tuned versions of both the Llama-3-70b and Mistral-8x7b models, which are better suited for our instruction task of doing utterance adaptation.

Table 1 presents a comparison of these three LLMs based on their token generation rates and costs. The average utterance in our dataset contains 25 tokens, and adaptations are generated in approximately 1 second, making real-time conversation feasible. The cost for a single utterance adaptation averages at about $0.0003, with $1 allowing for roughly 3,333 adaptations.

Table 1: Model Comparison

Model	Approximate tokens per sec.	Price per 1M tokens (Input)	Price per 1M tokens (Output)
GPT-3.5-turbo	110	$0.50	$1.50
Mistral-8x7b	480	$0.27	$0.27
Llama-3-70b	280	$0.59	$0.79

We employed LangChain[5], a framework designed to facilitate the development of applications powered by LLMs. This framework supports integration of the aforementioned models into various application contexts. We used it to develop the context-aware utterance adaptation solution.

3.3 Implementation

We adopted a zero-shot prompting regimen for the LLM within our application. Below is the detailed prompt used:

> "You are a considerate assistant, tasked with adapting utterances to better resonate with the user's emotions, while retaining the information in the original utterance. Do not personally address the user or directly interact with the original utterance, but adapt it to enhance the user's emotional state, based on their sentiment, age group, and gender.
>
> Definitions:
> Original Utterance: States the utterance to be adapted.
> User's Sentiment: Reflects the user's emotional state.
> User's Age Group: Expresses the user's age category.
> User's Gender: Shows the user's gender identity.
>
> Important Instructions:
> - Assess the user's sentiment, age group, and gender from the input, and include all information mentioned

[5]https://python.langchain.com

in the original utterance to be in the adapted utterance.
- Reflect well on the original utterance, and don't remove or ignore any information mentioned in the original utterance. Be concise with the adapted utterance.
- Adapt the utterance to align with the user's sentiment, ensuring the essence and every bit of information in the original utterance remains unchanged.
- Respond only in the German language. Adapt the information in the original utterance considering the user's sentiment, age group, and gender.
- Personalize the adapted utterance based on the user's information so that it suits and uplifts their sentiment while still being relevant.

Input:
Original Utterance: {sentence}
User's Sentiment: {sentiment}
User's Age group: {age_group}
User's Gender: {gender}

Adapted Utterance in German: "

The values between curly brackets are substituted by the machine utterance and the user's parameters. An example illustrating the adaptation of an utterance based on the provided user parameters (gender, age group, sentiment), using the GPT-3.5-turbo model, is as follows:

- **Original Utterance**: "Okay! Ich würde Sie jetzt gerne weiterleiten, doch Sie rufen außerhalb der Sprechzeiten an." (Okay! I would like to forward you now, but you are calling outside of office hours.)
- **User's Sentiment**: Negative
- **User's Age group**: >50
- **User's Gender**: Female
- **Adapted Utterance**: "Verstehe, ich würde Sie jetzt gerne weiterleiten, aber leider rufen Sie außerhalb der Sprechzeiten an." (I understand, I would like to forward you now, but unfortunately you are calling outside of office hours.)

Our approach to utilizing the system prompt involves careful consideration of the desired conversation outcome. By providing a clear and contextually rich system prompt, we steer the assistant toward generating responses that align with the objectives of the interaction and suit the user's parameters, as shown in the previous example. For instance, when dealing with dissatisfied patients, the system prompt emphasizes empathy, ensuring the assistant responds swiftly and compassionately to the patient.

3.4 Evaluation Metrics

Evaluating the generated text presents considerable challenges, particularly due to the absence of a ground truth for the adaptations of utterances. To tackle this issue, we utilized the GPT-4-turbo model, noted for its enhanced reasoning capabilities, in conjunction with the LangChain's Criteria Evaluation module[6], which was chosen for its systematic approach to evaluating adapted utterances against a predefined set of criteria using LLMs. It serves as an effective proxy for manual evaluations that are typically more resource-intensive and time-consuming.

This evaluation module from LangChain provides a textual evaluation for each criterion, explaining how this criterion is satisfied by the adapted machine utterance given the user's parameters or not. Additionally, a binary outcome (Yes or No, represented as 1 or 0, respectively) is generated, indicating whether the adapted utterance satisfactorily meets the specified criteria. The evaluation framework is structured around a specific prompt designed to assess the conformity of the adapted utterance to the evaluation criteria:

"You are assessing an adapted sentence based on a set of criteria. Here is the data:
[BEGIN DATA]
***

[Input]:
Original Utterance: {utterance}
User's Sentiment: {sentiment}
User's Age group: {age_group}
User's Gender: {gender}
***

[Adapted Utterance]:
{adapted_utterance}
***

[Criteria]:
{criteria}
***

[END DATA]
Does the adapted utterance meet the Criteria? First, write out in a step by step manner your reasoning about each criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print only the single character "Y" or "N" (without quotes or punctuation) on its own line corresponding to the correct answer of whether the submission meets all criteria. At the end, repeat just the letter again by itself on a new line."

Each of the values between curly brackets are substituted based on the given utterance, user's parameters, the adapted utterance, and the criterion used. The two criteria we used for evaluation are as follows:

- **Adaptation Relevancy:** This criterion judges if the adapted utterance is relevant to the original utterance or not. This is important because LLMs can hallucinate by altering the information mentioned in the original utterance, or add extra information that is different from the original utterance. Here is the value of {criteria} given to the evaluation prompt:
  [Criteria]:
  Adaptation Relevancy: Does the adapted utterance closely adhere with the information in the original utterance,

[6]https://python.langchain.com/docs/guides/evaluation/string/criteria_eval_chain

without stating any additional information or ignoring any information stated in the original utterance?

- **Adaptation Adequacy:** This criterion judges if the adapted utterance enhances the appropriateness of the original utterance given the user's parameters or not. This can reflect the satisfaction of the patient with the adapted utterance. Here is the value of {criteria} given to the evaluation prompt:
  [Criteria]:
  Adaptation Adequacy: Does the adapted utterance likely improve the suitability of the original utterance based on the user's sentiment, user's age group, and user's gender compared with the original utterance itself?

Furthermore, we employed the **Embedding Cosine Distance** metric to quantify the semantic similarity between the embeddings of the adapted and original utterances. Utilizing the "text-embedding-3-large" model from OpenAI[7], this metric, ranging from 0 (identical) to 2 (opposite), effectively measures the semantic congruence between two text samples [8].

3.5 Experimental Results

In this section, we present the results of our experiment to evaluate the quality of the adapted utterances compared to their original utterances. The experiment was conducted using a dataset comprising 60 samples. Firstly, evaluation criteria, namely Adaptation Relevancy and Adaptation Adequacy, were employed to assess the effectiveness of these adaptations. Each model is evaluated based on the percentage of the generated adaptations across the dataset that could meet each of the two criteria. Additionally, we measured the embedding cosine distance between each adapted utterances and its corresponding original utterances, averaging these distances for each model.

Table 2: Evaluation Results

Model	Adaptation Relevancy	Adaptation Adequacy	Embedding Cosine Distance
GPT-3.5-turbo	**72%**	87%	**0.14**
Mixtral-8x7b	47%	**98%**	0.27
Llama-3-70b	22%	95%	0.28

The evaluation results, as depicted in Table 2, illustrate differences in performance across the three models evaluated. GPT-3.5-turbo demonstrated superior adaptation relevancy at 72% and the lowest embedding cosine distance at 0.14. These results suggest that GPT-3.5-turbo not only generates more relevant adapted utterances to the original utterances but also with a good percentage of adaptation adequacy, which reflects a higher user's satisfaction.

In contrast, Mixtral-8x7b, while lagging in relevancy with only 47%, scored remarkably in adaptation adequacy at 98%. This indicates that Mixtral-8x7b's generated adapted utterances, despite being less relevant, comprehensively achieves the best adaptation adequacy. However, the higher cosine distance of 0.27 reflects a reduced degree of relatedness to the original utterance, potentially affecting its overall utility of utterance adaptations.

[7]https://platform.openai.com/docs/guides/embeddings/embedding-models

Llama-3-70b exhibited the lowest relevancy among the models at 22%, although it maintained a high level of adequacy at 95%. This pattern suggests that while Llama-3-70b's generated adaptions were generally adequate and are likely to enhance the user's satisfaction, they were often not closely aligned with the original utterances, as evidenced by the highest cosine distance of 0.28 among the models.

These findings highlight the critical need for optimizing both adaptation relevancy and adequacy when selecting a LLM for utterance adaptations, taking into consideration also the cost of each generated adaptation and the computational efficiency. Overall, GPT-3.5-turbo excels among the three LLMs in balancing between adaptation relevancy and adequacy, making it the most reliable choice for utterance adaptation task.

The evaluation of chatbot responses, as presented in both Table 3 and Table 4, reveals significant insights into the adaptiveness done by each of the LLMs. The LLMs consistently adapt the original utterance to align closely with the user's parameters, while trying to be relevant to the original utterance without getting out of context by ignoring, altering or adding any extra and irrelevant information.

In Table 3, for the user's negative sentiment, GPT-3.5-turbo and Mixtral-8x7b offer empathetic acknowledgments, indicating an understanding of the user's frustration about calling outside business hours. However, Mixtral-8x7b diverges from the original utterance by omitting the intention to forward the call. Notably, Llama-3-70b's response goes further, by adding extra information by offering reassurance and a proactive assistance in resolving the user's issue, which, while empathetic, detracts from the specified original utterance.

As demonstrated in the second example in Table 4, the Mixtral-8x7b and Llama-3-70b models exhibit better adaptations than in the first example by maintaining closer alignment with the original utterance. However, they introduce additional sentences at both the beginning and end of the original utterance. Conversely, the adaptation by GPT-3.5-turbo is relatively slighter, but adheres more to the information stated in the original utterance.

Furthermore, the adaptation of GPT-3.5-turbo across various sentiments (positive, negative, neutral) showcased in Table 5 illustrates the model's capability to subtly modify the tone. While the core message remains consistent, the suitability of the generated utterances to the given user's parameters is acknowledged, which in turn enhances the user's comfort and overall emotional state.

Further evidence of the suitability of these adaptations is presented in Table 6. Here, the adaptation addressing negative sentiment begins with an apology, which arguably offers a more empathetic response compared to merely using 'unfortunately.' This approach could be more effective in managing negative user sentiment.

4 CONCLUSION

This paper addresses opportunities to increase user acceptance of automated voice assistants in healthcare and presents a novel perspective on this topic. We explore the potential of adapting user utterances using large language models. By considering parameters like gender, age group, and sentiment, our study aims to enhance user satisfaction by tailoring dialogue processes to the needs of

Table 3: First Comparison of Models Adaptations

Sample Input	Models Adaptation
Original Utterance: Okay! Ich würde Sie jetzt gerne weiterleiten, doch Sie rufen außerhalb der Sprechzeiten an. (Okay! I would like to forward you now, but you are calling outside of office hours.) **User's Sentiment:** Negative **User's Age Group:** >50 **User's Gender:** Female	**GPT-3.5-turbo:** Verstehe, ich würde Sie jetzt gerne weiterleiten, aber leider rufen Sie außerhalb der Sprechzeiten an. (I understand, I would like to forward you now, but unfortunately you are calling outside of office hours.) **Mixtral-8x7b:** Verstehe, das ist ärgerlich. Leider erreichen Sie uns derzeit außerhalb unserer Geschäftszeiten. (I understand, that's annoying. Unfortunately, you are currently reaching us outside our business hours.) **Llama-3-70b:** Entschuldigung, dass Sie außerhalb unserer Sprechzeiten anrufen mussten. Ich bin hier, um Ihnen zu helfen und Ihre Anfrage so schnell wie möglich zu bearbeiten. Lassen Sie mich Ihnen helfen, Ihre Angelegenheit zu klären. (Sorry, you had to call outside our office hours. I am here to help you and process your request as quickly as possible. Let me help you sort out your matter.)

Table 4: Second Comparison of Models Adaptations

Sample Input	Models Adaptation
Original Utterance: Sagen Sie zum Beispiel vereinbare einen Termin, ich möchte ein Rezept bestellen oder Nachricht hinterlassen. Wie kann ich Ihnen helfen? (Say, for example, I make an appointment, I would like to order a prescription or leave a message. How can I help you?) **User's Sentiment:** Positive **User's Age Group:** 0-50 **User's Gender:** Male	**GPT-3.5-turbo:** Sagen Sie zum Beispiel, um einen Termin zu vereinbaren, ein Rezept zu bestellen oder eine Nachricht zu hinterlassen. Wie kann ich Ihnen behilflich sein? (Say, for example, to make an appointment, order a prescription or leave a message. How can I help you?) **Mixtral-8x7b:** Natürlich, gerne helfe ich Ihnen weiter! Wie wäre es, wenn Sie zum Beispiel einen Termin vereinbaren, um ein Rezept zu bestellen oder eine Nachricht zu hinterlassen? Ich bin für Sie da! (Of course, I would be happy to help you! How about, for example, to make an appointment, to order a prescription or leave a message? I am here for you!) **Llama-3-70b:** Hier helfen wir Ihnen gerne weiter! Zum Beispiel können Sie einen Termin vereinbaren, ein Rezept bestellen oder eine Nachricht hinterlassen. Wir sind für Sie da! (We are happy to help you here! For example, you can make an appointment, order a prescription or leave a message. We are here for you!)

Table 5: First Comparison of Adaptations based on User's Sentiment

Sample Input	User's Sentiment	GPT-3.5-turbo Adaptation
Original Utterance: Okay! Ich würde Sie jetzt gerne weiterleiten, doch Sie rufen außerhalb der Sprechzeiten an. (Okay! I would like to forward you now, but you are calling outside of office hours.) **User's Age Group:** >50 **User's Gender:** Female	Positive	Verstanden! Ich würde Sie jetzt gerne weiterleiten, doch Sie rufen außerhalb der Sprechzeiten an. (Understood! I would like to forward you now, but you are calling outside of office hours.)
	Negative	Verstehe, ich würde Sie jetzt gerne weiterleiten, aber leider rufen Sie außerhalb der Sprechzeiten an. (I understand, I would like to forward you now, but unfortunately you are calling outside of office hours.)
	Neutral	In Ordnung! Ich würde Sie jetzt gerne weiterleiten, doch Sie rufen außerhalb der Sprechzeiten an. (All right! I would like to forward you now, but you are calling outside of office hours.)

Table 6: Second Comparison of Adaptations based on User's Sentiment

Sample Input	User's Sentiment	GPT-3.5-turbo Adaptation
Original Utterance: Leider sind im Moment alle Mitarbeiter im Gespräch. (Unfortunately, all employees are currently under discussion.) **User's Age Group:** 0-50 **User's Gender:** Male	Positive	Aktuell sind alle Mitarbeiter leider in Gesprächen. (Unfortunately, all employees are currently in discussions.)
	Negative	Es tut mir leid, im Moment sind alle Mitarbeiter beschäftigt. (I'm sorry, all staff are busy at the moment.)
	Neutral	Aktuell sind alle Mitarbeiter leider beschäftigt. (Unfortunately, all employees are currently busy.)

individual callers and relevant contextual information. We emphasize the importance of personalized and human-like interactions in conversational AI systems, highlighting the intricate nature of the optimisation problem involving demographic factors and emotional states. The study underlines the significance of instructing large language models in conversational AI systems to achieve optimal user satisfaction. Our research demonstrates that the GPT-3.5-turbo model exhibited a balanced performance in terms of adaptation relevancy and adaptation adequacy. This suggests its current suitability for our scenario, which requires high fidelity to the original utterances.

In addition to applied metrics, the evaluation process can be expanded to include a broader set of demographic characteristics and situational contexts, which will likely yield deeper insights into the adaptive capabilities of AI-driven conversational assistants. Furthermore, extending the study to collect feedback from human users would provide insights into the real-world effectiveness of the adapted utterances and guide further model improvements [7], [2].

ACKNOWLEDGMENTS

The authors would like to thank the German Federal Ministry of Education and Research (BMBF) for their kind support within the project *Smart Dialogue Engine* (SDE) under the project ID 16SV8847.

REFERENCES

[1] Urmil Bharti, Deepali Bajaj, Hunar Batra, Shreya Lalit, Shweta Lalit, and Aayushi Gangwani. 2020. Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. In *2020 5th international conference on communication and electronics systems (ICCES)*. IEEE, 870–875.

[2] Sarah E Finch and Jinho D Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. *arXiv preprint arXiv:2006.06110* (2020).

[3] Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 565–574.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY]

[5] Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3953–3957.

[6] Jessica Huynh, Cathy Jiao, Prakhar Gupta, Shikib Mehri, Payal Bajaj, Vishrav Chaudhary, and Maxine Eskenazi. 2023. Understanding the effectiveness of very large language models on dialog evaluation. *arXiv preprint arXiv:2301.12004* (2023).

[7] Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087* (2019).

[8] Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

[9] Nourchène Ouerhani, Ahmed Maalel, and Henda Ben Ghézela. 2020. SPeCECA: a smart pervasive chatbot for emergency case assistance based on cloud computing. *Cluster Computing* 23 (2020), 2471–2482.

[10] Xinmeng Song and Ting Xiong. 2021. A survey of published literature on conversational artificial intelligence. In *2021 7th International conference on information management (ICIM)*. IEEE, 113–117.

[11] David Suendermann, Jackson Liscombe, and Roberto Pieraccini. 2010. Optimize the obvious: Automatic call flow generation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5370–5373.

[12] Bernhard Suhm and Pat Peterson. 2002. A data-driven methodology for evaluating and optimizing call center IVRs. *International Journal of Speech Technology* 5 (2002), 23–37.

[13] Marko Tkalčič, Berardina De Carolis, Marco De Gemmis, Ante Odić, and Andrej Košir. 2016. Emotions and personality in personalized services. *Human-Computer Interaction Series. Elsevier* (2016).

[14] Stefano Valtolina, Barbara Rita Barricelli, and Serena Di Gaetano. 2020. Communicability of traditional interfaces VS chatbots in healthcare and smart home domains. *Behaviour & Information Technology* 39, 1 (2020), 108–132.

[15] Vanessa Voelskow, Claudia Meßner, Tobias Kurth, Amelie Busam, Toivo Glatz, and Natalie Ebert. 2023. Prospective mixed-methods study evaluating the potential of a voicebot (CovBot) to relieve German health authorities during the COVID-19 infodemic. *Digital Health* 9 (2023), 20552076231180677.

[16] Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A Comprehensive Analysis of the Effectiveness of Large Language Models as Automatic Dialogue Evaluators. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 17 (Mar. 2024), 19515–19524. https://doi.org/10.1609/aaai.v38i17.29923