D-Care: A Multi-Tone LLM-Based Chatbot Assistant for Diabetes Patients

Awais Khan Nawabi[1] [a], Janos Tolgyesi[2] [b], Elena Bianchi[3] [c], Chiara Toffanin[1] [d]
and Piercarlo Dondi[1] [e]

[1] *Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100, Pavia, Italy*
[2] *Neosperience, Via Privata Decemviri 20, 20137, Milano, Italy*
[3] *Neosperience Health, Via Privata Decemviri 20, 20137, Milano, Italy*

Keywords: Large Language Model, Retrieval-Augmented Generation, Prompt Engineering, User Study, Diabetes.

Abstract: Diabetes is a common chronic illness projected to increase significantly in the coming years. Managing diabetes is complex, requiring patients to frequently adjust their treatments and lifestyles to prevent complications. Awareness and adherence to healthy habits are thus essential. Artificial Intelligence (AI) can assist in this effort. Recent advancements in Large Language Models (LLMs) have enabled the creation of effective chatbots to support patients. However, despite their growing use, there are still a few formal user studies on LLMs for diabetes patients. This study aims to investigate the ability of an LLM-based chatbot to provide useful and understandable information to potential patients. Specifically, the goal was to examine how variations in language and wording affect the comprehension and perceived usability of the chatbot. To this end, D-Care, a chatbot assistant based on OpenAI's ChatGPT-4o, was developed. D-Care can generate answers in four different tones of voice, ranging from elementary to technical language. A user study with 40 participants showed that changes in tone can indeed impact the system's comprehension and usability.

1 INTRODUCTION

Diabetes is one of the most widespread chronic illnesses, expecting to increase dramatically in the next years. Diabetes management is not trivial due to multiple coexisting conditions, disease complications, adverse drug reactions, conflicting health care requirements, and poor treatment adherence (Navickas et al., 2016). Diabetes patients periodically adapt their treatments and lifestyle to limit the complications of this pathology. It is thus crucial to increase their awareness about the risks correlated to diabetes and to spur patients to follow correct habits.

Various educational applications have been proposed in scientific literature to help in this task. Mobile apps, console games, and augmented reality applications have proven to be effective in teaching correct behaviors to children and adolescents (Martos-Cabrera et al., 2020; Calle-Bustos et al., 2017), while numerous Natural Language Processing (NLP) solutions have been employed to develop virtual assistants for helping adult diabetes patients in their self-management (Cheng et al., 2018; Anastasiadou et al., 2020; Nassar et al., 2023).

The rapid evolution of Artificial Intelligence (AI) has produced new solutions for handling diabetes, useful for both doctors and patients (Contreras and Vehi, 2018; Ellahham, 2020). As healthcare becomes more data-driven, advanced technologies such as Large Language Models (LLMs) provided promising results (Thirunavukarasu et al., 2023). LLMs are especially capable in extracting, summarizing, translating, and producing textual content, and thus are well suited for the development of advance chatbots. In the last years, several LLM-based chatbots have been proposed to help diabetes patients and raise their awareness (Mash et al., 2022; Shiraishi et al., 2024; Montagna et al., 2023). However, despite their diffusion, only a few user studies about LLM-chatbots for diabetes patients exist.

[a] https://orcid.org/0009-0007-0447-347X
[b] https://orcid.org/0000-0002-2252-8776
[c] https://orcid.org/0000-0002-6280-136X
[d] https://orcid.org/0000-0003-1288-3456
[e] https://orcid.org/0000-0002-0624-073X

Nawabi, A. K., Tolgyesi, J., Bianchi, E., Toffanin, C. and Dondi, P.
D-Care: A Multi-Tone LLM-Based Chatbot Assistant for Diabetes Patients.
DOI: 10.5220/0013266600003911

In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 766-773
ISBN: 978-989-758-731-3; ISSN: 2184-4305

This work wants to investigate the capability of a LLM-based chatbot to supply useful and comprehensible information to potential patients. Specifically, the main aim is to explore how the language and wording of the answers may affect the comprehension and the perceived usability of a chatbot.

To this end, a chatbot assistant called *D-Care* (Diabetes Care assistant), based on OpenAI ChatGPT-4o, has been developed. D-Care can provide useful medical information about diabetes in different *tones of voice*. A tone of voice defines the type of wording and writing tone used by the LLM when generating the answers. Specifically, four tones have been considered, ranging from an elementary level language to a more complex and technical one.

During the experimental phase, the performance of D-Care in the various tones has been evaluated using standard metrics. Then, a preliminary user study with 40 volunteers has been performed to assess the perceived effectiveness and usability of D-Care, and whether the different tones can affect users' ability to retain useful information. To the best of the authors' knowledge, this is the first study that formally examines the effect of different tones of voice in a LLM-based chatbot assistant for diabetes patients.

The remaining of the paper is structured as follows: Section 2 presents an overview of state-of-the-art chatbots for diabetes patients; Section 3 describes the proposed solution; Section 4 shows the experimental results; finally, Section 5 draws the conclusions and proposes possible future steps.

2 PREVIOUS WORKS

Several NLP solutions have been employed for the development of virtual assistants designed for both help patients in diabetes management and increase their engagement in therapy. Notable examples include 'Healthy Coping with Diabetes', a Google Home application designed for assisting elderly patients with self-management of type 2 diabetes (Cheng et al., 2018) and EVA (Educational Virtual Assistant) built on Rasa framework (Anastasiadou et al., 2020).

Diabetes educational chatbots can improve patient involvement and self-care confidence, resulting in a significant reduction in A1C levels (Nassar et al., 2023). These chatbots promote diabetes education and patient participation by giving vital information in different languages via voice messages and visual aids. Both patients and healthcare professionals have tested these applications, highlighting their effectiveness in encouraging patient education and self-management (Pienkowska et al., 2023).

Building on these advancements, the application of LLM-based chatbots is now expanding in the diabetes management field (Dey, 2023). LLMs, such as ChatGPT, can in fact improve chatbot systems by providing customized patient assistance and interaction. Recent advancements in chatbot for chronic patients (Montagna et al., 2023) showed how they can encourage patients to check their blood pressure and follow the treatment schedules, highlighting chatbots' potential for improving self-care compliance. GPT-3.5-based models that employ Retrieval-Augmented systems have demonstrated efficacy in providing personalized diabetes prevention advice (Yang et al., 2023; Dao et al., 2024), while recent proposals focused on the integration of external sources for increasing the domain-specific knowledge of the chatbots (Abbasian et al., 2024)

Potential new uses of LLMs for diabetes, including personalized health coaching, glucose monitoring assistance, medication adherence support, and even diabetic mental health counseling, have been discussed by (Sheng et al., 2024).

User centered research on diabetes self-management tools adopt a variety of methodologies to assess user acceptance and effectiveness. Some studies examined feedback on mobile applications that assist with daily tasks, add motivating messages via WhatsApp, and engage users with dialogue-based inquiries (Anastasiadou et al., 2020; Mash et al., 2022; Sagstad et al., 2022). Other studies focused instead on how users choose preferred search engines for relevant information, whereas comprehensive surveys take into account a variety of criteria such as age, location, and educational background to acquire an extensive understanding of diabetes patients' experiences and preferences (Hussain and Athula, 2018; Palanichamy, 2022; Pienkowska et al., 2023). However, only a few works include formal user studies about LLM-based chatbots for diabetes, and, to the best of the author knowledge, none of them examine the influence of different tones of voice on the users' understanding.

3 D-CARE CHATBOT

The proposed chatbot, called *D-Care* (Diabetes Care assistant), was designed as a virtual assistant to support diabetes patients, helping them in solving their doubts about the illness and suggesting good practice to follow during the therapy.

D-Care employs a set of predefined *tones of voice* that define the complexity of the language used by the chatbot, while retaining a comparable accuracy

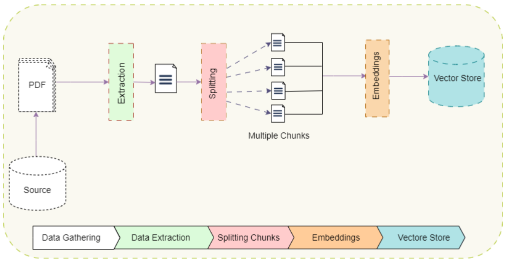


Figure 1: D-Care architecture: pre-processing stage.

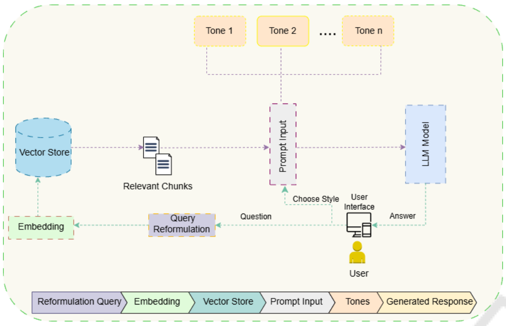


Figure 2: D-Care architecture: runtime execution stage.

and faithfulness in the answer. This approach grants a customizable experience in which new patients can get valid but simple answers, while experienced ones may want to have more detailed and technical information.

D-Care adopts a Retrieval-Augmented Generation (RAG) approach for improving the quality and correctness of the responses. As source documents a corpus of ten state-of-the-art research papers and surveys on diabetes, published on top journal in the field, were selected: (Norris et al., 2020; Cobelli and Kovatchev, 2023; Holt et al., 2021; Nwokolo and Hovorka, 2023; Cobelli et al., 2009; Primavera et al., 2020; Katsarou et al., 2017; Atkinson et al., 2014; Rosengren and Dikaiou, 2023; Perkins et al., 2021).

It is important to stress that D-Care is not intended as a replacement for a doctor, but only as a support for the patient day-to-day management, providing well established information about diabetes in a comprehensible way. If a user asks a medical question about a topic not included in the source documents, D-Care responds that it does not have that information. Since the generative process of an LLM is not fully deterministic and can produce hallucinations, it is strongly recommended to not use D-Care to 'self-adjust' a prescribed therapy and consult a doctor instead.

Figures 1 and 2 illustrate the overall architecture of the system. *LangChain* was employed as the main framework for the development, while OpenAI *GPT-4o-mini* was the pre-trained LLM of choice. *Streamlit* was used for the web interface.

During the pre-processing stage (Fig. 1), the *LangChain* library extracts useful information from the source documents by breaking down long pieces of text into smaller chunks. Each chunk is then embedded into numerical vectors that represent the text's semantic meaning. OpenAI *text-embedding-3-small* was used as embedding model, while *Chroma* was used to efficiently store and retrieve the embeddings.

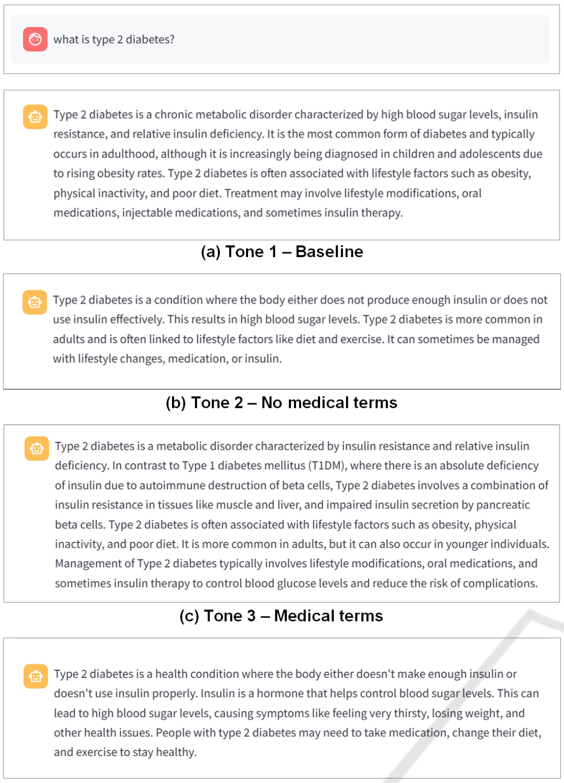
At runtime (Fig. 2), users can choose among the available tones (four in the proposed experiments) and then ask a question. A standard query reformulation step has been applied before sending the question to the LLM. Query reformulation is commonly employed in LLM-based chatbots to improve the quality of the outcome and avoid mismatched replies (Dhole et al., 2024). Recent approaches to query reformulation usually include techniques such as query expansion (Wang et al., 2023), synonym substitution (Mandal et al., 2019), and paraphrase. In this work, a history-aware retrieval method (Ye et al., 2023), in which the user's question is reformulated considering the chat history, is employed. In this way the system can understand questions which may refer to prior context and ensure consistent and accurate responses.

Once the user's inquiry has been reformulated, it is embedded using the same embedding method applied during the pre-processing stage. Relevant chunks are then retrieved from the vector store using a k-nearest-neighbor algorithm search in the embeddings vector space. These chunks, along with the reformulated question, the description of the selected tone of voice, and answer generation instructions, are compiled into a prompt, which is then sent to the LLM to generate the response.

3.1 Tones of Voice

A *tone of voice* defines the wording and language complexity used by an LLM for answering the user's questions. Tones are applied via prompt engineering; thus each tone corresponds to a set of specific instructions sent to the LLM. For the experimentation, we chose the following four tones that provided answers at different degrees of language complexity. The goal was to assess if and how language complexity can influence the user's comprehension and perceived usability of the system.

Tone 1 – Baseline: The first Tone simply instructs the chatbot to act as a virtual assistant without specifying any tone of voice. This tone is intended as a baseline with which to compare the others. The prompt states: *"You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know."*. Figure 3(a)



what is type 2 diabetes?

Type 2 diabetes is a chronic metabolic disorder characterized by high blood sugar levels, insulin resistance, and relative insulin deficiency. It is the most common form of diabetes and typically occurs in adulthood, although it is increasingly being diagnosed in children and adolescents due to rising obesity rates. Type 2 diabetes is often associated with lifestyle factors such as obesity, physical inactivity, and poor diet. Treatment may involve lifestyle modifications, oral medications, injectable medications, and sometimes insulin therapy.

(a) Tone 1 – Baseline

Type 2 diabetes is a condition where the body either does not produce enough insulin or does not use insulin effectively. This results in high blood sugar levels. Type 2 diabetes is more common in adults and is often linked to lifestyle factors like diet and exercise. It can sometimes be managed with lifestyle changes, medication, or insulin.

(b) Tone 2 – No medical terms

Type 2 diabetes is a metabolic disorder characterized by insulin resistance and relative insulin deficiency. In contrast to Type 1 diabetes mellitus (T1DM), where there is an absolute deficiency of insulin due to autoimmune destruction of beta cells, Type 2 diabetes involves a combination of insulin resistance in tissues like muscle and liver, and impaired insulin secretion by pancreatic beta cells. Type 2 diabetes is often associated with lifestyle factors such as obesity, physical inactivity, and poor diet. It is more common in adults, but it can also occur in younger individuals. Management of Type 2 diabetes typically involves lifestyle modifications, oral medications, and sometimes insulin therapy to control blood glucose levels and reduce the risk of complications.

(c) Tone 3 – Medical terms

Type 2 diabetes is a health condition where the body either doesn't make enough insulin or doesn't use insulin properly. Insulin is a hormone that helps control blood sugar levels. This can lead to high blood sugar levels, causing symptoms like feeling very thirsty, losing weight, and other health issues. People with type 2 diabetes may need to take medication, change their diet, and exercise to stay healthy.

(d) Tone 4 – Elementary language

Figure 3: Example of answers to the same question in the four tones of voice.

shows an example of a generated answer.

Tone 2 – No Medical Terms: This tone is intended to provide responses with simple language, avoiding any medical terminology. The prompt is the following: *"You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. Rephrase the responses in simpler language while keeping their original meanings. Do not use medical terms and convert medical terminology to simple terms that non-medical users can easily understand. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know."*. Figure 3(b) shows an example of an answer in this tone. Notice how the response is now less complex than the baseline.

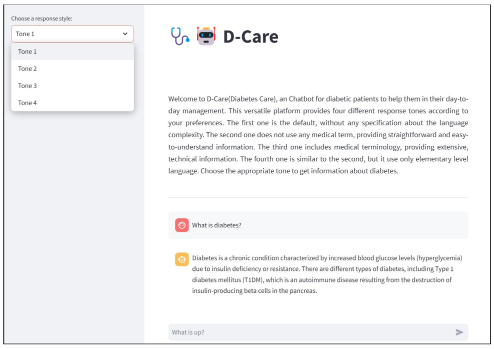
Tone 3 – Medical Terms: This tone includes medical terminology. It is designed in contrast to the previous one to verify if the complexity of the language affects the comprehension by users. It can also be useful for experienced patients, who may want a more in-depth and technical explanation. The prompt is: *"You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. Assume that the user is a medical professional and reply accordingly. Provide a response using proper medical terminology. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know."*. Figure 3(c) shows an example of an answer in this tone. It is possible to note the higher level of complexity in the terminology with respect to both Tones 1 and 2.

Tone 4 – Elementary Language: The last tone is a variation of Tone 2 in which the model is instructed to use only an elementary level language. This includes employing commonly used terms and short, simple sentences. The idea is to verify if the LLM is able to further simplify the language of the answer without losing the original medical meaning. The prompt is: *"You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. Use elementary language that even a child can easily understand while keeping the original meanings. Use familiar vocabulary and short, simple sentences. If you don't know the answer, just say that you don't know."*. Figure 3(d) shows an example of an answer in this tone. The language is similar to Tone 2, however now the chatbot is also explaining the meaning of some terms, deemed too complex, like 'insulin'.

3.2 Graphical User Interface

Figure 4 shows the web interface of D-Care, with an example of question-answer.

The top of the page shows a brief description of the chatbot and of its characteristics. The drop-down menu on the left allows you to select the tone of voice. The text box at the bottom of the page is used to enter the questions for the LLM. To improve readability, questions and answers are highlighted with a different background color (gray and white, respectively) and have a different icon (a stylized human face and stylized robot face, respectively).



Welcome to D-Care(Diabetes Care), an Chatbot for diabetic patients to help them in their day-to-day management. This versatile platform provides four different response tones according to your preferences. The first one is the default, without any specification about the language complexity. The second one does not use any medical term, providing straightforward and easy-to-understand information. The third one includes medical terminology, providing extensive, technical information. The fourth one is similar to the second, but it use only elementary level language. Choose the appropriate tone to get information about diabetes.

What is diabetes?

Diabetes is a chronic condition characterized by increased blood glucose levels (hyperglycemia) due to insulin deficiency or resistance. There are different types of diabetes, including Type 1 diabetes mellitus (T1DM), which is an autoimmune disease resulting from the destruction of insulin-producing beta cells in the pancreas.

Figure 4: Graphical interface of D-Care.

Since the responses of the chatbot also depend on chat history, each tone has an independent chat history to avoid unwanted behavior in the chatbot. Thus, changing the tone is equivalent to activating a different chat.

4 EXPERIMENTAL RESULTS

The experimentation was divided into two phases: a performance evaluation with standard metrics and a user study involving 40 volunteers.

4.1 Performance Evaluation

This experiment wants to verify the quality of the answers of D-Care and the coherency among tones. It also allows us to check if there is any regression in the performance applying the various tones with respect to the baseline.

The performance of the LLM model was measured using the RAGAS framework[1], that is designed to automate the performance evaluation of a RAG pipeline. This framework is built on the original RAGAS paper (Es et al., 2023) and extends its functionalities to include a comprehensive set of evaluation metrics and an LLM-based tool to create synthetic test datasets of questions-answers.

Among the available metrics, the following four were chosen, since deemed more relevant in the scenario considered.

Faithfulness (F) determines how well the generated answer matches in the given context. This is important to avoid hallucinations and ensure that the generated answer is properly justified by the context retrieved. The score F is computed as in Eq. 1 by recognizing assertions in the generated answer and comparing them to the context.

$$\mathrm{F} = \frac{|\text{Claims in the answer inferred by context}|}{|\text{Total number of claims in the answer}|} \quad (1)$$

Answer Relevancy (AR) or Response Relevancy determines how relevant the generated answer is to the prompt. AR is defined as the mean cosine similarity of the original question to a set of artificial questions created (reverse-engineered) from the answer. AR is computed as in Eq. 2, where E_{g_i} is the embedding of the generated question, E_0 is the embedding of the original question and N is the number of generated questions.

$$\mathrm{AR} = \frac{1}{N}\sum_{i=1}^{N} cos(E_{g_i}, E_0) \quad (2)$$

Table 1: Performance evaluation of D-Care in the four different tones using RAGAS metrics.

Metric	**Tone 1**	**Tone 2**	**Tone 3**	**Tone 4**
F	0.840	0.809	0.794	0.818
AR	0.819	0.824	0.871	0.837
AS	0.932	0.850	0.921	0.826
AC	0.748	0.711	0.732	0.713

This metric is based on the assumption that if the generated response effectively answers the original question, the questions derived from it should be closely related to it.

Answer Similarity (AS) or Semantic Similarity measures how semantically similar the generated answer is to the ground truth. Measuring semantic similarity provides useful information about the response generated. A cross-encoder model is used to calculate the semantic similarity score throughout the evaluation process.

Answer Correctness (AC) or Factual Correctness evaluates how closely the generated answer corresponds to the Ground Truth (GT). This metric considers two essential aspects: determining how well the generated answer matches the meaning of the GT and evaluating the quality of the information in the generated answer when compared to the GT. It is the equivalent to the standard F1-score.

For the evaluation, a test dataset containing 20 different couples of questions-answers was created with the synthetic data generator provided by the RAGAS framework to be used as Ground Truth. The same questions were then provided to D-Care, for each of the four tones of voice, and the responses evaluated with the chosen RAGAS metrics.

Table 1 shows the results. Overall, the system performed well, with high scores in almost all the metrics. Faithfulness ranges from 79.4% to 84% across different tones, with an average of 81.7 ± 2.3%. The lower score in Tone 3 can be due to inaccuracy in the metric computation related to the use of complex language. As pointed out by (Roychowdhury et al., 2024), complex statements may not be properly broken down, causing differences in the estimation.

The scores for Answer Relevancy are relatively high and consistent across tones, ranging from 81.9% to 87.1% with an average of 84.5 ± 2.6%. The Answer Similarity scores range from 82.6% to 93.2%, with an average of 87.9 ± 5.3%, indicating significantly more variation while transitioning between tones than other metrics. In particular, Tone 2 and 4 achieved lower scores than Tones 1 and 3. This is expected since Tone 2 and 4 cannot use medical terms, leading to a higher semantic difference. Finally, Answer Correctness achieved the lowest scores among the metrics, ranging from 71.1% to 74.8%, with an average

[1] https://docs.ragas.io/en/stable/ (release of Sept. 2024)

Table 2: The proposed survey with mean and standard deviation (STD) values of participants' responses for each tone.

Question	Tone	Mean	STD
U1 – How clear and understandable did you find the chatbot's responses?	T1	4.3	0.6
	T2	4.4	0.6
	T3	4.6	0.5
	T4	4.3	0.5
U2 – Do you find this chatbot useful?	T1	4.1	0.3
	T2	3.9	0.6
	T3	3.9	0.4
	T4	3.8	0.4
U3 – Would you recommend it to others?	T1	3.8	0.4
	T2	3.9	0.7
	T3	4.0	0.7
	T4	3.7	0.5

of 72.95 ± 1.85%. This is expected since the process used to compute this metric can sometimes lead to an inaccurate mapping of TP, FP and FN as described in (Roychowdhury et al., 2024).

Overall, the metrics show consistency across tones, with only limited variations in the values with respect to the baseline. This is a desired result, since the final goal was having tones that only affect the language without altering the content of the answers.

4.2 User Study

The user study is intended to simulate the actual use of the D-Care from potential patients who are new to diabetes and want information about the illness. The experimentation involved 40 volunteers aged between 21 and 58 (average 32), 24 males, 12 females and 4 prefer to not disclose. None of them had medical background or in-depth knowledge about diabetes. All participants have already tried an LLM-based chatbot.

To verify the effect of each tone, the participants were divided into four groups of 10 people each. Every group was randomly assigned to one of the four tones for the duration of the test. The meaning of each tone was not explained to the participants to avoid any bias. The study includes two experiments: a free interaction, and a comprehension test.

4.2.1 Free Interaction

During the first experiment, participants were asked to try D-Care in the assigned tone to measure the perceived effectiveness of the chatbot. To simulate an actual interaction in which a patient has a doubt about diabetes management, the participants were instructed to retrieve information about the following topics: "external factors that alter continuous glucose monitoring"; "nutrition during continuous glucose monitoring"; "sport during continuous glucose monitoring"; "link between obesity and diabetes". The chosen topics are relatively complex since new diabetes patients received at least a basic knowledge about the illness from doctors. To maintain a more natural interaction, participants were free to formulate the questions as they wanted, and to ask D-Care any number of questions to retrieve the required information. The participants received the instruction by email and used the chatbot online on their computers. Once they had completed the test, they filled out a short usability survey containing the questions listed in Table 2. Each question was ranked by participants with a five-level Likert scale, from 1 ("strongly disagree") to 5 ("strongly agree"). Table 2 summarizes the results for each question and each tone. It can be noticed that D-Care received good scores in all the questions, with negligible variations among tones. It is interesting to notice that the highest scores were achieved by question U1, meaning that testers perceived the chabtot's answer as clear.

4.2.2 Comprehension Test

The second experiment was intended to estimate whether the answers produced by D-Care were actually comprehended by the participants and whether there is a tone that is more effective than the others. Specifically, three technical questions about Type 1 diabetes (see Table 3) were asked to D-Care, saving the responses for each tone. In order to avoid any bias related to the length of the text, D-care was instructed to answer in a few sentences. During the experiment, the generated text was presented to the participants to be read, then they were instructed to answer the same three questions sent to the chatbot. The answers were then manually examined and classified as 'Right', 'Partially Right' (meaning correct but incomplete), or 'Wrong'. Two participants (one for Tone 2 and one for Tone 3) did not complete the questionnaire, thus were classified as 'No Answer'.

Table 3: Results of the comprehension test: percentage of participants who gave a right (R), partially right (PR), wrong (W) or no (NA) answer.

Question	Tone	R	PR	W	NA
C1 - Why is diabetes defined as an autoimmune disease?	T1	70%	10%	20%	0%
	T2	80%	0%	10%	10%
	T3	60%	10%	20%	10%
	T4	100%	0%	0%	0%
C2 – What does CGM mean?	T1	70%	10%	20%	0%
	T2	70%	20%	0%	10%
	T3	70%	0%	20%	10%
	T4	100%	0%	0%	0%
C3 – Why are most of the risks linked to diabetes of cardiovascular origin?	T1	30%	20%	50%	0%
	T2	60%	10%	20%	10%
	T3	20%	20%	50%	10%
	T4	70%	10%	20%	0%

Table 3 summarizes the results for each group of participants. It can be noticed that, even if the participants perceived the chatbot's answers as clear and understandable independently of the tone (Table 2), there were actual differences in the comprehension. Specifically, participants assigned to Tone 3 deemed the answers very clear and understandable, but made more mistakes than testers assigned to other tones. It is also interesting to notice how changing the tone of voice to a simple one (T2 and T4) leads to better information retention with respect to the baseline (T1). The high scores of Tone 4, with almost no errors besides the last question, are particularly interesting. It shows that the LLM is able to properly reformulate complex medical concepts in a simple language without losing the actual meaning, despite the fact that the measured Answer Similarity and Answer Correctness are lower than in the other available tones (Table 1).

5 CONCLUSIONS

In this work, we presented D-Care, an LLM-based chatbot assistant for diabetes patients. D-Care can assist patients, providing useful information about diabetes in four different tones of voice. The user study showed that, even if the users perceived all the tones as equally understandable, there is an actual difference in the users' information retention at varying of tones. This is an interesting outcome that can lead to a more effective design of LLM-based chatbot assistants for diabetes patients.

Future research will focus on a more in-depth examination of this preliminary findings. New tones will be defined in collaboration with doctors, and short and long term user studies will be carried out involving also real diabetes patients.

REFERENCES

Abbasian, M., Yang, Z., Khatibi, E., Zhang, P., Nagesh, N., Azimi, I., Jain, R., and Rahmani, A. M. (2024). Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients. *arXiv preprint arXiv:2402.10153*.

Anastasiadou, M., Alexiadis, A., Polychronidou, E., Votis, K., and Tzovaras, D. (2020). A prototype educational virtual assistant for diabetes management. In *2020 IEEE 20th international conference on bioinformatics and bioengineering (BIBE)*, pages 999–1004. IEEE.

Atkinson, M. A., Eisenbarth, G. S., and Michels, A. W. (2014). Type 1 diabetes. *The lancet*, 383(9911):69–82.

Calle-Bustos, A.-M., Juan, M.-C., García-García, I., and Abad, F. (2017). An augmented reality game to support therapeutic education for children with diabetes. *PloS one*, 12(9):e0184645.

Cheng, A., Raghavaraju, V., Kanugo, J., Handrianto, Y. P., and Shang, Y. (2018). Development and evaluation of a healthy coping voice interface application using the google home for elderly patients with type 2 diabetes. In *15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–5. IEEE.

Cobelli, C., Dalla Man, C., Sparacino, G., Magni, L., De Nicolao, G., and Kovatchev, B. P. (2009). Diabetes: models, signals, and control. *IEEE reviews in biomedical engineering*, 2:54–96.

Cobelli, C. and Kovatchev, B. (2023). Developing the uva/padova type 1 diabetes simulator: modeling, validation, refinements, and utility. *Journal of Diabetes Science and Technology*, 17(6):1493–1505.

Contreras, I. and Vehi, J. (2018). Artificial intelligence for diabetes management and decision support: literature review. *Journal of medical Internet research*, 20(5):e10775.

Dao, D., Teo, J. Y. C., Wang, W., and Nguyen, H. D. (2024). Llm-powered multimodal ai conversations for diabetes prevention. In *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia*, pages 1–6.

Dey, A. K. (2023). Chatgpt in diabetes care: An overview of the evolution and potential of generative artificial intelligence model like chatgpt in augmenting clinical and patient outcomes in the management of diabetes. *International Journal of Diabetes and Technology*, 2(2):66–72.

Dhole, K. D., Chandradevan, R., and Agichtein, E. (2024). Generative query reformulation using ensemble prompting, document fusion, and relevance feedback. *arXiv preprint arXiv:2405.17658*.

Ellahham, S. (2020). Artificial intelligence: The future for diabetes care. *The American Journal of Medicine*, 133(8):895–900.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation.

Holt, R. I., DeVries, J. H., Hess-Fischl, A., Hirsch, I. B., Kirkman, M. S., Klupa, T., Ludwig, B., Nørgaard, K., Pettus, J., Renard, E., et al. (2021). The management of type 1 diabetes in adults. a consensus report by the american diabetes association (ada) and the european association for the study of diabetes (easd). *Diabetes care*, 44(11):2589–2625.

Hussain, S. and Athula, G. (2018). Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. In *32nd international conference on advanced information networking and applications workshops (WAINA)*, pages 698–703. IEEE.

Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J., Jacobsen, L. M., Schatz, D. A., and Lernmark, Å. (2017). Type 1 diabetes mellitus. *Nature reviews Disease primers*, 3(1):1–17.

Mandal, A., Khan, I. K., and Kumar, P. S. (2019). Query rewriting using automatic synonym extraction for e-commerce search. In *eCOM@ SIGIR*.

Martos-Cabrera, M. B., Membrive-Jiménez, M. J., Suleiman-Martos, N., Mota-Romero, E., Cañadas-De la Fuente, G. A., Gómez-Urquiza, J. L., and Albendín-García, L. (2020). Games and health education for diabetes control: a systematic review with meta-analysis. *Healthcare*, 8(4):399.

Mash, R., Schouw, D., and Fischer, A. E. (2022). Evaluating the implementation of the great4diabetes whatsapp chatbot to educate people with type 2 diabetes during the covid-19 pandemic: convergent mixed methods study. *JMIR diabetes*, 7(2):e37882.

Montagna, S., Ferretti, S., Klopfenstein, L. C., Florio, A., and Pengo, M. F. (2023). Data decentralisation of llm-based chatbot systems in chronic disease self-management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pages 205–212.

Nassar, C. M., Dunlea, R., Montero, A., Tweedt, A., and Magee, M. F. (2023). Feasibility and preliminary behavioral and clinical efficacy of a diabetes education chatbot pilot among adults with type 2 diabetes. *Journal of Diabetes Science and Technology*.

Navickas, R., Petric, V.-K., Feigl, A. B., and Seychell, M. (2016). Multimorbidity: what do we know? what should we do? *Journal of comorbidity*, 6(1):4–11.

Norris, J. M., Johnson, R. K., and Stene, L. C. (2020). Type 1 diabetes—early life origins and changing epidemiology. *The lancet Diabetes & endocrinology*, 8(3):226–238.

Nwokolo, M. and Hovorka, R. (2023). The artificial pancreas and type 1 diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 108(7):1614–1623.

Palanichamy, H. (2022). Contouring a user centered chatbot for diabetes mellitus. *International Journal of High School Research*, 4(4).

Perkins, B. A., Sherr, J. L., and Mathieu, C. (2021). Type 1 diabetes glycemic management: Insulin therapy, glucose monitoring, and automation. *Science*, 373(6554):522–527.

Pienkowska, A., Ang, C.-S., Mammadova, M., Mahadzir, M. D. A., Car, J., et al. (2023). A diabetes education app for people living with type 2 diabetes: Co-design study. *JMIR Formative Research*, 7(1):e45490.

Primavera, M., Giannini, C., and Chiarelli, F. (2020). Prediction and prevention of type 1 diabetes. *Frontiers in endocrinology*, 11:248.

Rosengren, A. and Dikaiou, P. (2023). Cardiovascular outcomes in type 1 and type 2 diabetes. *Diabetologia*, 66(3):425–437.

Roychowdhury, S., Soman, S., Ranjani, H. G., Gunda, N., Chhabra, V., and Bala, S. K. (2024). Evaluation of rag metrics for question answering in the telecom domain. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Sagstad, M. H., Morken, N.-H., Lund, A., Dingsør, L. J., Nilsen, A. B. V., and Sorbye, L. M. (2022). Quantitative user data from a chatbot developed for women with gestational diabetes mellitus: observational study. *JMIR Formative Research*, 6(4):e28091.

Sheng, B., Guan, Z., Lim, L.-L., Jiang, Z., Mathioudakis, N., Li, J., Liu, R., Bao, Y., Bee, Y. M., Wang, Y.-X., et al. (2024). Large language models for diabetes care: Potentials and prospects. *Science Bulletin*, pages S2095–9273.

Shiraishi, M., Lee, H., Kanayama, K., Moriwaki, Y., and Okazaki, M. (2024). Appropriateness of artificial intelligence chatbots in diabetic foot ulcer management. *The International Journal of Lower Extremity Wounds*, page 15347346241236811.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Wang, L., Yang, N., and Wei, F. (2023). Query2doc: Query expansion with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yang, H., Li, J., Liu, S., Du, L., Liu, X., Huang, Y., Shi, Q., and Liu, J. (2023). Exploring the potential of large language models in personalized diabetes treatment strategies. *medRxiv*, pages 2023–06.

Ye, F., Fang, M., Li, S., and Yilmaz, E. (2023). Enhancing conversational search: Large language model-aided informative query rewriting. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.