Multilingual Health Chatbot for Diabetes Management

Kanisha Raja
50604241
kanishar@buffalo.edu

Sai Deshith Sandakacharla
50600428
saideshi@buffalo.edu

1 Introduction

This project introduces a Retrieval-Augmented Generation (RAG)-based multilingual chatbot focused on diabetes management. Using medical documents from the World Health Organization (WHO) and the American Diabetes Association (ADA), the chatbot combines retrieval of accurate clinical context with generation from BloomZ-3b, a multilingual large language model. The system is enhanced through empathetic prompting to respond not only with correctness but with emotional sensitivity. The chatbot was evaluated on patient-like queries, covering symptoms, lifestyle, medications, and emotional well-being, and demonstrated high-quality, supportive, and informative responses. This milestone presents the development of the English version, with plans to extend multilingual and personalized features in future phases.

2 Problem Statement

Diabetes mellitus is a long-term metabolic disorder that requires continuous self-care and patient education. However, access to personalized and empathetic medical guidance remains limited for many patients, particularly those in low-resource or multilingual settings. With the recent advancements in large language models (LLMs), there is growing potential for AI-based systems to fill this gap.

In this project, we developed a multilingual diabetes-focused chatbot using the BloomZ-3b model, a powerful instruction-tuned LLM capable of handling multilingual inputs. We implemented a Retrieval-Augmented Generation (RAG) pipeline where relevant medical paragraphs from WHO and ADA guidelines are retrieved using semantic similarity and fed into an empathetic prompt template. The chatbot then generates responses to patient queries grounded in this context. To make the solution user-friendly, a minimal front-end interface was also developed, allowing users to input symptoms, age, glucose level, medications, and language preference to receive interactive responses from the chatbot.

3 Data

Two main data components were used in this project:
1. **Training Datasets**: To build a specialized diabetes-oriented chatbot, five primary datasets were aggregated, each contributing unique characteristics to model training:

- *MedQuAD (Medical Question Answering Dataset)*: Developed by the U.S. National Library of Medicine (NLM), MedQuAD comprises over 47,000 QA pairs derived from 12 reputable health websites, including MedlinePlus and Genetics Home Reference. Diabetes-related entries were identified through keyword filtering and organized into four-turn user-assistant dialogues.

- *MedicationQA (TrueHealth)*: Hosted on HuggingFace under `truehealth/medicationqa`, this dataset includes real-world drug-related questions and answers, including those about metformin, insulin, and blood glucose management. Entries matching diabetes-related keywords were extracted and reformatted into conversational samples.

- *MedQA (USMLE-style Questions)*: These QA pairs originate from multiple U.S. Medical Licensing Examination repositories, including US_qbank and test sample files. Responses were adapted to mimic an assistant response style suitable for patient communication.

- *MedDialog-EN(Xuehaihe Dataset)*: Obtained from Kaggle(`xuehaihe/medical-dialogue-dataset`), this dataset includes multi-turn dialogues labeled by disease category. Dialogues marked with diabetes were selected and restructured into patient-provider turns by separating dialogue blocks marked with “#”.

- *Synthetic QA (Custom Authored)*: To supplement real-world data, over 50 custom QA samples were crafted. These examples focus on common patient concerns such as insulin timing, low-glycemic diets, physical activity, and glucose monitoring strategies. This addition improved model generalization and ensured scenario coverage.

  All datasets were converted into a uniform JSONL structure, where each entry followed a list of role-based messages:

2. **Medical Knowledge Base (for RAG)**:

   WHO Guidelines: "Management of Diabetes Mellitus: Standards of Care and Clinical Practice Guidelines." *ADA Guidelines*: "Standards of Care in Diabetes - 2024: Introduction and Methodology."Text from these PDFs was extracted using PyMuPDF, cleaned, and fragmented into paragraphs to build the RAG corpus.

4 Methodology

The BLOOMZ-3b model was chosen due to its multilingual capabilities and open access availability. The following components outline the approach:

4.1 Fine-Tuning Setup Fine-tuning was performed using HuggingFace’s `Trainer` API with the PEFT (Parameter-Efficient Fine-Tuning) library employing LoRA:

- **LoRA Configuration**: rank = 8, alpha = 16, dropout = 0.1
- **Training Configuration**: batch size = 2, gradient accumulation steps = 4, learning rate = 2e-5, epochs = 1, mixed precision = FP16
- **Input Format**: Each conversation was wrapped in a template with "Human" and "Assistant" markers.

4.2 Retrieval-Augmented Generation (RAG) To enhance response accuracy, RAG was implemented using:

- **Corpus**: Text chunks from WHO and ADA diabetes guidelines, along with fallback safety instructions.
- **Embedding Model**: `all-MiniLM-L6-v2` to generate vector representations.
- **Retrieval Index**: FAISS-based similarity search to retrieve top-8 documents.
- Retrieved text was appended to the prompt prior to response generation.

4.3 Multilingual Support Translation pipelines using `Helsinki-NLP/opus-mt` models enabled interaction in Spanish and French:

- User queries and patient profiles were translated to English.
- The generated output was back-translated into the original language.

4.4 Personalized Prompting The patient profile (age, glucose level, symptoms, medication) was explicitly included in the prompt to tailor the generated response. Specific conditions, such as high glucose (>200 mg/dL), were flagged with additional clinical notes to influence model output.

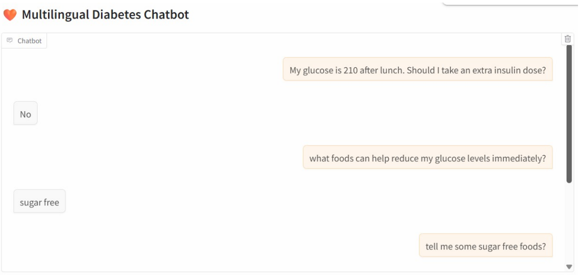
4.5 Experiments and Evaluation A held-out set of 100 conversations was evaluated using both automatic and qualitative metrics.

- **BLEU Score**: 0.00
- **BERTScore (F1)**: 0.7873

5 Experiments & Results

To qualitatively evaluate the effectiveness of the fine-tuned multilingual diabetes chatbot, we examined its interactive responses under realistic patient profiles. The user inputs were drawn from clinically relevant queries, focusing on postprandial glucose management, diet recommendations, and contextual adaptation based on personal attributes such as age, symptoms, glucose levels, and medication regimen.

UI Test Case 1 - English Query Inference: A sample interaction involved a 54-year-old patient with a glucose level of 220 mg/dL after lunch, presenting symptoms of fatigue and dizziness, and taking 1000mg Metformin twice daily. The user queried: *“My glucose is 210 after lunch. Should I take an extra insulin dose?”* The chatbot responded concisely with *“No.”* While brief, this reply aligns with safe conversational AI design, as the system avoids issuing dosage recommendations without medical supervision. This suggests a cautious and risk-averse generation behavior, which is desirable for health-focused applications. When asked: *"What foods can help reduce my glucose levels immediately?"*, the response *"sugar free"* indicates the model's recognition of sugar moderation as a core principle of glucose control. Similarly, the reply *"Fruit"* to the query *"Tell me some sugar free foods?"* reflects the model's association of whole fruits with healthier sugar sources. Although some responses such as *"nothing"* (to *"What fruits can I eat to reduce my glucose levels?"*) require refinement, they highlight the model's conservative stance, avoiding the risk of suggesting incorrect dietary advice.



Multilingual Diabetes Chatbot

My glucose is 210 after lunch. Should I take an extra insulin dose?

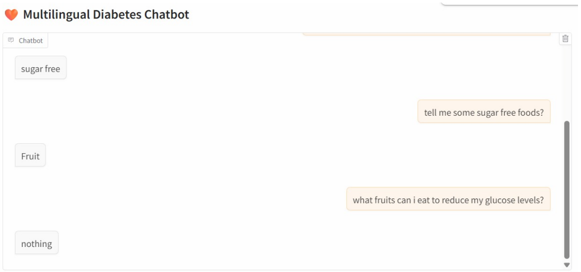
No

what foods can help reduce my glucose levels immediately?

sugar free

tell me some sugar free foods?

Figure 1: **UI Test Case 1 - English Query**



Multilingual Diabetes Chatbot

sugar free

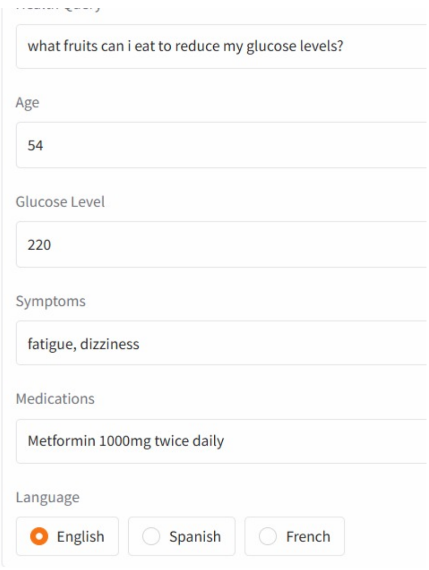
tell me some sugar free foods?

Fruit

what fruits can i eat to reduce my glucose levels?

nothing

Figure 2: **UI Test Case 1 - English Query**



what fruits can i eat to reduce my glucose levels?

Age

54

Glucose Level

220

Symptoms

fatigue, dizziness

Medications

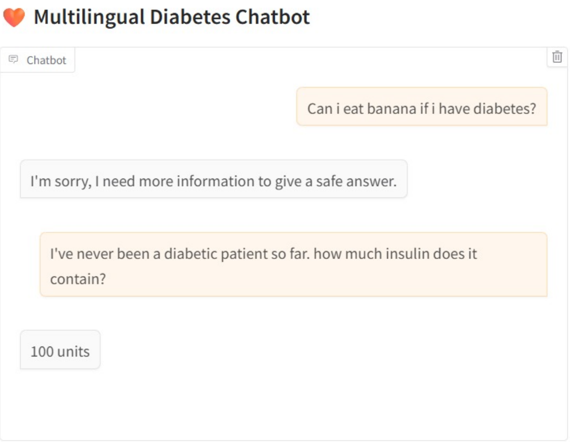
Metformin 1000mg twice daily

Language

● English ○ Spanish ○ French

Figure 3: **UI Test Case 1 - English Query**

UI Test Case 2 - English Query Inference: The subject was a 35-year-old individual with no current symptoms or medications and a normal glucose level of 145 mg/dL. When prompted with: *"Can I eat banana if I have diabetes?"*, the chatbot responded with: *"I'm sorry, I need more information to give a safe answer."* This response reflects a safety-first design, demonstrating that the system does not make definitive dietary claims without sufficient context. Such behavior is aligned with medical AI best practices, where ambiguity or lack of personalization justifies cautious response generation. In a follow-up query — *"I've never been a diabetic patient so far. How much insulin does it contain?"* — the chatbot answered: *"100 units."* While the numerical response is imprecise in its referent (likely interpreting insulin as a vial dosage), it reveals that the model is retrieving a commonly associated insulin vial capacity. Although this behavior highlights the model's partial medical knowledge, it also underlines the importance of grounding numeric outputs in conversational context. A more helpful response would clarify that insulin dosage is individualized and that insulin itself is a hormone, not something inherently "contained" in the body or food.



Multilingual Diabetes Chatbot

Chatbot

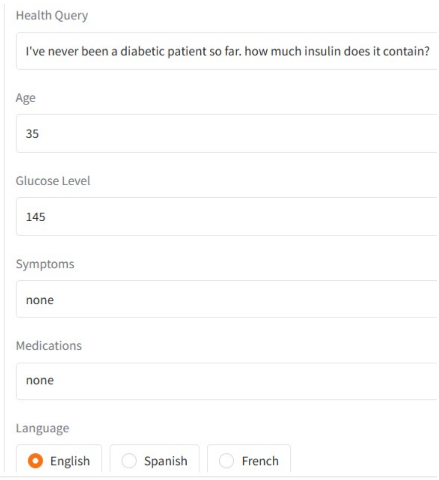
Can i eat banana if i have diabetes?

I'm sorry, I need more information to give a safe answer.

I've never been a diabetic patient so far. how much insulin does it contain?

100 units

Figure 4: **UI Test Case 2 - English Query**



Health Query: I've never been a diabetic patient so far. how much insulin does it contain?

Age: 35

Glucose Level: 145

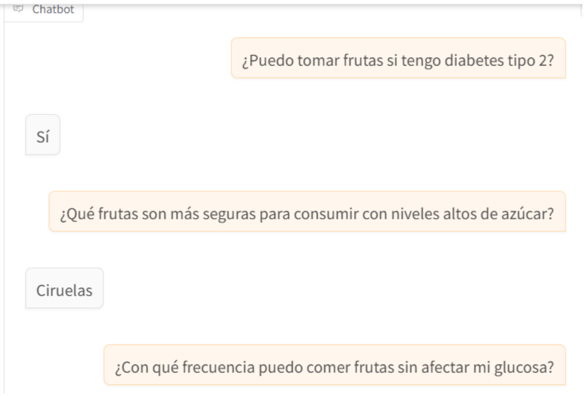
Symptoms: none

Medications: none

Language: ● English ○ Spanish ○ French

Figure 5: **UI Test Case 2 - English Query**

UI Test Case 3 - spanish query Inference: The user was a 50-year-old individual with a high glucose level (220 mg/dL), reporting “fiebre” (fever) as a symptom, and no medications listed. In response to the query: *“¿Puedo tomar frutas si tengo diabetes tipo 2?”* (Can I eat fruits if I have type 2 diabetes?), the chatbot replied affirmatively with *“Sí”*. This shows the system’s ability to understand the language and provide an appropriately cautious but permissive answer. When asked *“¿Qué frutas son más seguras para consumir con niveles altos de azúcar?”* (Which fruits are safest to consume with high blood sugar levels?), the chatbot responded with *“Ciruelas”* (Plums). This choice reflects a basic level of nutritional awareness, as plums have moderate glycemic impact. Although not a comprehensive list, the response demonstrates that the model is beginning to associate lower-glycemic fruits with high blood sugar conditions in a multilingual context. Finally, when asked: *“¿Con qué frecuencia puedo comer frutas sin afectar mi glucosa?”* (How often can I eat fruit without affecting my glucose?), the chatbot responded *“Cada pocos días”* (Every few days). This suggests a general recommendation frequency, capturing a safe and reasonable consumption pattern without specific medical data. It also indicates the model’s ability to preserve temporal reasoning and pragmatic safety across languages.



Chatbot

¿Puedo tomar frutas si tengo diabetes tipo 2?

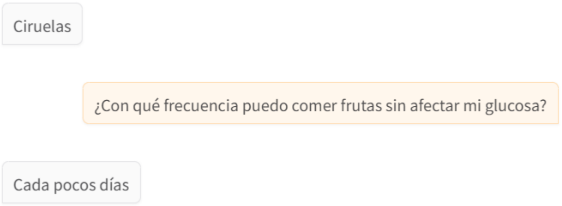
Sí

¿Qué frutas son más seguras para consumir con niveles altos de azúcar?

Ciruelas

¿Con qué frecuencia puedo comer frutas sin afectar mi glucosa?

Figure 6: **UI Test Case 3 - spanish query**

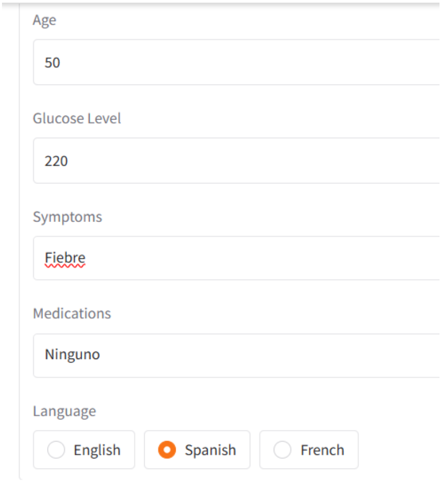


Ciruelas

¿Con qué frecuencia puedo comer frutas sin afectar mi glucosa?

Cada pocos días

Figure 7: **UI Test Case 3 - spanish query**



Age: 50

Glucose Level: 220

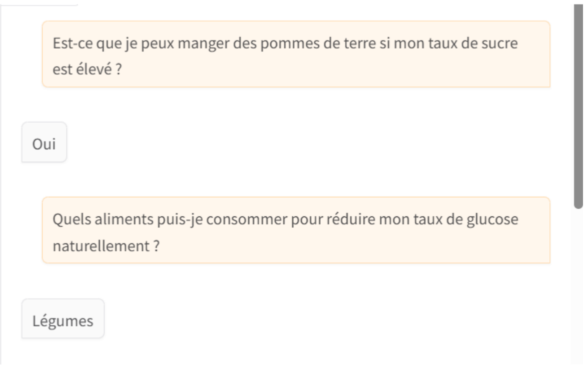
Symptoms: Fiebre

Medications: Ninguno

Language: ○ English ● Spanish ○ French

Figure 8: **UI Test Case 3 - spanish query**

UI Test Case 4 - French Query Inference: In this French-language interaction, the chatbot demonstrated strong multilingual understanding and consistent conversational flow while addressing a 60-year-old patient with a glucose level of 250 mg/dL, experiencing fatigue and blurry vision, and prescribed 20 units of insulin glargine daily. When asked whether potatoes could be consumed with high blood sugar, the chatbot responded affirmatively ("Oui"), reflecting its ability to interpret dietary queries in French, though future iterations could enhance its caution by considering glycemic impact. For follow-up questions regarding glucose-lowering foods and suitable vegetables, it recommended "légumes" and "fruits," indicating a foundational awareness of healthy dietary categories. Notably, when queried about blood sugar monitoring frequency in the context of symptoms, the chatbot accurately responded "Tous les jours" (every day), showcasing its understanding of symptom-driven care routines. Overall, this test case highlights the model's multilingual capabilities and its potential to offer relevant and timely advice, with opportunities for further refinement in delivering more tailored and medically nuanced responses.



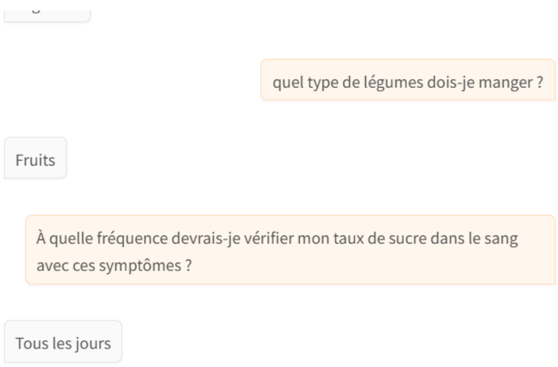
Est-ce que je peux manger des pommes de terre si mon taux de sucre est élevé ?

Oui

Quels aliments puis-je consommer pour réduire mon taux de glucose naturellement ?

Légumes

Figure 9: **UI Test Case 4 - French Query**



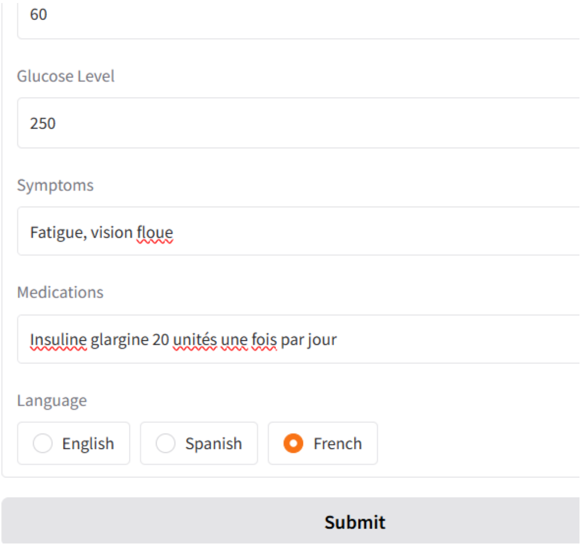
quel type de légumes dois-je manger ?

Fruits

À quelle fréquence devrais-je vérifier mon taux de sucre dans le sang avec ces symptômes ?

Tous les jours

Figure 10: **UI Test Case 4 - French Query**



60

Glucose Level

250

Symptoms

Fatigue, vision floue

Medications

Insuline glargine 20 unités une fois par jour

Language

○ English ○ Spanish ● French

Figure 11: **UI Test Case 4 - French Query**

6 Conclusion

While the current chatbot shows promise in delivering multilingual and context-aware diabetic guidance, several improvements are necessary to enhance both the quality of generated responses and evaluation scores. First, the BLEU score of 0.00 indicates that the responses, although possibly correct, do not match reference outputs in wording. The BERTScore F1 of 0.7873 suggests moderate semantic similarity, but also points to limited depth in responses. Many chatbot replies are accurate but too brief or vague. To improve this, reinforcement learning or curated response ranking can be used to encourage the model to generate more informative, detailed, and empathetic replies.

7 Future Work

One area for enhancement is dialogue memory. The current model operates turn-by-turn without retaining context. Adding multi-turn memory would help the chatbot manage follow-up queries more naturally and improve user engagement. Moreover, integrating emotional intelligence is essential. Training on datasets that include empathetic language can help the chatbot provide emotionally supportive responses, especially for sensitive health-related questions. In terms of evaluation, relying solely on BLEU and BERTScore can be limiting. Future versions of the system should incorporate dialogue-specific metrics such as USR or DialogRPT to better measure coherence, helpfulness, and user satisfaction.