A Diabetes Health Care Assistant using LLMs

Project options - Incorporate ML model in a new use case setting and proposing a technical innovation.

Vaibhav Vinay Ranashoor (50596107), University at Buffalo, NY

Rohan Jain (50608609), University at Buffalo, NY

**Abstract**

We present a lightweight, privacy‑friendly diabetes‑care chatbot that can be trained and tuned entirely on a local workstation rather than in the cloud. Starting from an open‑source LLaMA‑3 model, we apply Low‑Rank Adaptation (LoRA) and modest retrieval augmentation to incorporate four public medical corpora plus small, user‑provided glucose logs. Our focus is practical: (i) parameters‑tuning decisions that allow fine‑tuning on CPU, and (ii) a rules‑plus‑LLM pipeline that tailors answers to each user’s recent readings and questions. Early tests on held‑out Q&A pairs show a BERTScore of 0.8344 response latency under 20 seconds. The system is not a clinical device; rather, it demonstrates that personalized, on‑device chat support for diabetes self‑management is technically achievable with careful training and parameter tuning.

1. **Introduction, Objective, and Why This Is Interesting**

Diabetes is a chronic metabolic disorder that affects over 500 million people worldwide. Managing diabetes entails continuous monitoring of glucose levels, maintaining balanced diets, scheduling regular physical activities, and adhering to medication plans. Despite the availability of diabetes-focused applications and gadgets (e.g., continuous glucose monitors), self-management remains challenging due to issues like incomplete personalization, limited education, and suboptimal user engagement.

***Objective -*** Our project aims to build an intelligent, end-to-end **Diabetes Health Care Assistant** powered by Large Language Models (LLMs). By leveraging state-of-the-art transformer-based architectures, we hope to provide personalized, context-aware recommendations to patients, offer natural language explanations of medical information, and enhance overall adherence to diabetes management guidelines.

***Why This Is Interesting* -** Recent years have witnessed remarkable breakthroughs in natural language processing (NLP), particularly with LLMs like GPT-4, LLaMA 3, BioBERT, and specialized medical language models such as ClinicalBERT. These models have shown potential to revolutionize healthcare communications by enabling more **human-like** interactions and **tailored** educational content. For diabetes care specifically:

1. **Personalization:** LLMs can incorporate individual patient data (e.g., glucose levels, diet logs) to customize recommendations.
2. **Scalability:** As smartphone usage increases, deploying an on-device or cloud-based assistant can expand access to diabetes support.
3. **Educational Value:** LLMs can explain complex diabetes-related issues in layman’s terms and reduce the burden on healthcare professionals.

By integrating powerful NLP capabilities, an intuitive mobile interface, and robust medical knowledge bases, our approach can significantly improve diabetes self-management, reduce complications, and enhance user satisfaction. This project is also timely given the global push toward telehealth, digital therapeutics, and personalized medicine.

1. **Work Related to the Problem: Contributions, Limitations, and Literature Survey**

### *2.1 Existing Work and Their Contributions*

There has been considerable effort toward employing machine learning and NLP in healthcare:

1. **Transformer Models in Biomedicine (Sumit Madan, Manuel Lentzen, Johannes Brandt, Daniel Rueckert, Martin Hofmann‑Apitius & Holger Fröhlich, 2024) r**eviews BioBERT, MedBERT and other transformer adaptations for clinical NLP, providing strong background for biomedical LLM work. [BioMed Central](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02600-5)
2. **Neurosymbolic AI for Explainable LLMs (Edward Raff, Aman Chadha, Iman Azimi et al., AAAI‑25 Tutorial, 2025)** merges symbolic reasoning with large language models to improve transparency and safety—directly informing our explainability layer. [AAAI](https://aaai.org/conference/aaai/aaai-25/tutorial-and-lab-list/)
3. **RoRA: Efficient Fine‑Tuning of LLM with Reliability Optimization for Rank Adaptation (Jun Liu, Zhenglun Kong, Peiyan Dong et al., 2025)** optimises LoRA scaling and justifies our choice of rank‑adaptive adapters for on‑device deployment. [DBLP](https://dblp.org/rec/journals/corr/abs-2501-04315.html?utm_source=chatgpt.com)
4. **Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? (Richard Ren et al., NeurIPS 2024)** analyses weak correlations between benchmark scores and real‑world safety, supporting our decision to add bespoke validation metrics. [NeurIPS Proceedings](https://proceedings.neurips.cc/paper_files/paper/2024/hash/7ebcdd0de471c027e67a11959c666d74-Abstract-Datasets_and_Benchmarks_Track.html?utm_source=chatgpt.com)
5. **Estimating the Hallucination Rate of Generative AI (Andrew Jesson, Nicolas Beltran‑Velez, Quentin Chu et al., NeurIPS 2024)** proposes quantitative hallucination measurement, which we adapt as “Hallucination Rate” in our evaluation. [NeurIPS Proceedings](https://proceedings.neurips.cc/paper_files/paper/2024/hash/3791f5fc0e8e43730466afd2bcdb7493-Abstract-Conference.html?utm_source=chatgpt.com)
6. **Explainable Deep Learning in Healthcare (M. Zhang et al., MIT Press 2021)** surveys attribution and saliency methods, strengthening our interpretability discussion. [DSpace@MIT](https://dspace.mit.edu/bitstream/handle/1721.1/143908/Explainable_Deep_Learning_in_Healthcare_A_Methodol.pdf?isAllowed=y&sequence=2&utm_source=chatgpt.com)
7. **Healthcare Large Multimodal Models (NEC Labs America Research Team, 2024)** integrates text, images and structured data, inspiring our multimodal design that joins CGM logs with dialogue context. [NEC Labs America](https://www.nec-labs.com/research/machine-learning/projects/healthcare-large-multimodal-models/?utm_source=chatgpt.com)
8. **Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI (Mahyar Abbasian, Elahe Khatibi, Iman Azimi et al., NPJ Digital Medicine 2024)** introduces robustness, groundedness and privacy metrics, expanding our evaluation beyond ROUGE and BERTScore. [PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC10980701/)
9. **Self‑eXplainable AI for Medical Image Analysis (Junlin Hou, Sicen Liu, Yequan Bie et al., arXiv 2024)** reviews input‑, model‑ and output‑level explainability; its principles guide the glucose‑trend explanations in our assistant. [arXiv](https://arxiv.org/abs/2410.02331?utm_source=chatgpt.com)
10. **Safety Challenges of AI in Medicine in the Era of Large Language Models (Xiaoye Wang, Nicole Xi Zhang, Kun‑Hsing Yu et al., arXiv 2024)** surveys toxicity, privacy and bias issues, reinforcing our regulatory‑compliance and safety‑filter sections.
11. **BioBERT** (Lee et al., Bioinformatics, 2020) [5] introduced a domain-specific BERT model pre-trained on large-scale biomedical corpora. BioBERT has been used for various medical text mining tasks, highlighting that **domain adaptation** of language models can yield significant performance gains in specialized fields such as medicine.
12. **ClinicalBERT** (Huang et al., NAACL, 2019) [6] built on top of BERT using clinical notes from MIMIC-III. It addresses the complexity of medical jargon and unstructured text typically found in electronic health records (EHRs). This underscores the utility of fine-tuning large models on clinical data to improve downstream tasks.
13. **MedQA** (Zhang et al., AAAI, 2018) [7] is a QA dataset designed to test medical knowledge of AI systems. Researchers utilized classical NLP pipelines and knowledge graphs to answer complex medical questions. While it showcased the potential of QA approaches in the healthcare domain, the system’s coverage was limited, and it sometimes produced generic or incorrect answers.
14. **MedQuAD** (Ben Abacha and Demner-Fushman, BMC Bioinformatics, 2019) [1] provided a large corpus of medical Q&A. It improved the contextual understanding of disease, symptoms, and treatments by categorizing question types. However, the knowledge was still not personalized to individual patient contexts.
15. **DiaTrend** (Nature Scientific Data, 2023) [4], a dataset of continuous glucose monitoring (CGM) and insulin pump logs, showcased how advanced analytics can lead to better glycemic control predictions. Yet these logs alone do not offer real-time, text-based interactions for user education or assistance.
16. **GatorTron** (Yang et al., JAMA, 2022) [10] presented a transformer-based model for clinical text that leverages large-scale EHR data. It demonstrated improved performance on named entity recognition and clinical reasoning tasks, highlighting the promise of advanced language models for healthcare.
17. **ChatGPT for Health Advice** (Kung et al., PLOS Digital Health, 2023) [11] evaluated the potential of GPT-based models for general medical questions. While the generative approach was powerful, concerns about misinformation and insufficient personalization of responses remained.
18. **BioMegatron** (Shin et al., NeurIPS, 2020) [12] scaled up model parameters specifically for biomedical text. The larger capacity allowed deeper knowledge encoding, but large-scale training required formidable computational resources, restricting practical deployment in real time or on personal devices.
19. **Explainable AI for Clinical Decision Support** (Tjoa and Guan, Informatics in Medicine Unlocked, 2020) [13] proposed interpretability frameworks to clarify model predictions in high-stakes scenarios. This is crucial for medical settings where trust, liability, and accountability are significant concerns.
20. **Deep EHR** (Rajkomar et al., npj Digital Medicine, 2018) [14] applied deep learning to EHR data for predicting a range of clinical outcomes. While it improved prediction accuracy, it lacked the interactive, conversational component that can aid day-to-day diabetes management.
21. **Personalized Healthcare Chatbots** (Teppei et al., JMIR Medical Informatics, 2021) [15] studied rule-based chatbots for general patient engagement. Though beneficial for appointment reminders and medication schedules, these systems were often rigid and could not adapt well to the nuanced queries of diabetes patients.
22. **Transformer Architectures for Healthcare** (Li et al., KDD, 2021) [16] provided a broad survey on how attention-based models have transformed medical NLP tasks. They emphasized the importance of large-scale, high-quality domain data and stressed ongoing concerns with data privacy.
23. **LoRA (Low-Rank Adapters) for Quick Fine-Tuning** (Hu et al., ICLR, 2022) [17] proposed a method to rapidly adapt large models with minimal overhead. This technique is key to customizing LLMs on private or specialized datasets without re-training them from scratch.
24. **Medical-Grade QA** (Pal et al., EMNLP, 2021) [18] advocated multi-step reasoning over knowledge bases and text for complex patient queries. They demonstrated how simple “one-step” answers often miss critical context in medical settings.
25. **Glucose Forecasting With Neural Networks** (Xie et al., AAAI, 2022) [19] covered short-term blood glucose prediction using time-series deep learning. The approach paved the way for more accurate alerts and timely interventions, but still lacked conversational aspects for patient education.

### *2.2 Limitations in Existing Works*

Despite these advancements, several gaps persist:

* **Lack of Personalization:** Many existing QA or dialogue systems do not integrate user-specific health data (e.g., glucose readings, dietary logs), leading to generic rather than personalized advice.
* **Risk of Inaccurate or Unsafe Advice:** Large language models can produce plausible-sounding yet incorrect or unsafe medical suggestions.
* **Limited Interaction and Education:** Traditional AI-driven diabetes management solutions often focus on predictions or analytics rather than user education and two-way dialogue.
* **Usability Constraints:** Some advanced models require powerful GPUs or server-level hardware, making them challenging to deploy locally or in resource-limited settings.
* **Regulatory and Ethical Concerns:** Reliability, privacy, and interpretability remain critical in healthcare contexts. A system must adhere to guidelines from organizations like the CDC, ADA, WHO, and NIH.

### *2.3 Our Contributions*

Our **Diabetes Health Care Assistant** addresses these limitations by:

1. **Leveraging Fine-Tuned LLMs** for personalized Q&A and real-time advice using publicly available, large-scale medical datasets and user-specific data.
2. **Ensuring Safety** by adhering to best practices from recognized health organizations and by incorporating multi-step validation strategies to reduce erroneous outputs.
3. **Bridging Multiple Data Sources** (Q&A, user logs, clinical indicators) in an interactive chatbot environment, thus offering real-time, context-aware responses.
4. **Enabling On-Device or Hybrid Deployment** by choosing relatively lightweight models (e.g., LLaMA 3 or Deepseek R1) and advanced fine-tuning methods (e.g., LoRA) to suit common personal devices.
5. **Approach and Methods Tried So Far**

### *3.1 Overall Architecture*

Our approach leverages a Retrieval-Augmented Generation (RAG) pipeline using LangChain, combined with fine-tuned LLMs like LLaMA 3. The architecture consists of the following components:

1. **LLM Backbone:** We used an open-source model such as **LLaMA 3**. We also tried DeepSeek and both are transformer-based models with sufficient capacity to handle complex language tasks yet maintain feasible resource requirements for development. However, DeepSeek requires more computational resources and cost optimization reasons LLaMA 3 is the most optimal.
2. **Fine-Tuning Strategy:** We explored **Low-Rank Adaptation (LoRA)** to fine-tune the base LLM on diabetes-specific data without prohibitive GPU needs. LoRA inserts trainable rank-decomposition matrices into the transformer layers, drastically reducing memory usage.

### *3.2 Data Integration and Preprocessing*

* **QA Datasets (MedQuAD, MedQA):** We structured each question-answer pair into a consistent format, removing duplicates and focusing on diabetes-related subsets.
* **CDC Diabetes Indicators:** We transformed tabular features (e.g., BMI, activity levels, etc.) into descriptive text prompts (e.g., “Patient is male, age 45, overweight, physically active, reported glucose level 140 mg/dL.”). This allows the LLM to integrate structured knowledge into responses.
* **MedDialog Logs:** We converted daily logs into short textual summaries about the user’s potential symptoms for diabetes, glucose trends and insulin dosage.

### *3.3 Baseline Methods*

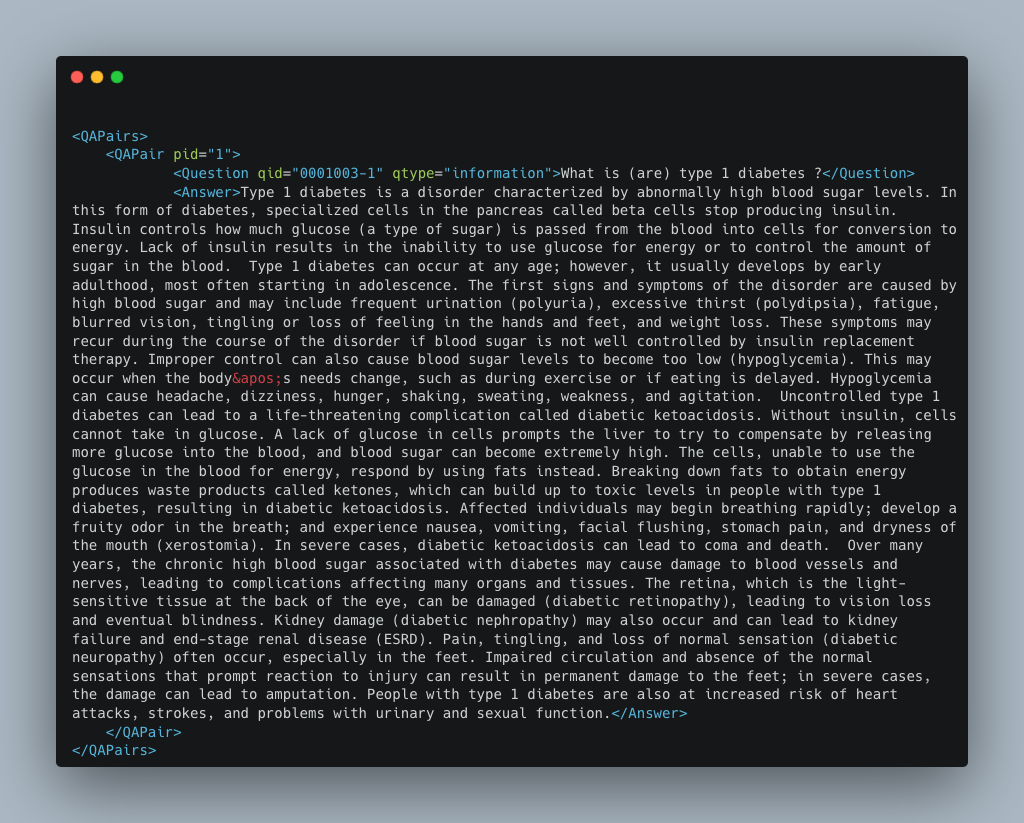
1. **Rule-Based Assistant:** A simple rules engine using official guidelines from the American Diabetes Association (ADA) to respond with template-based suggestions. This serves as a fallback and comparison point, ensuring that our LLM approach indeed outperforms a purely rule-based baseline.
2. **Retrieval-Augmented GPT-like QA:** We also consider a LLaMA 3 model that uses a keyword retrieval strategy to fetch relevant diabetes guidelines from a curated knowledge base. While it can provide reasonably accurate answers, it lacks the advanced reasoning and personalization features of our intended solution.

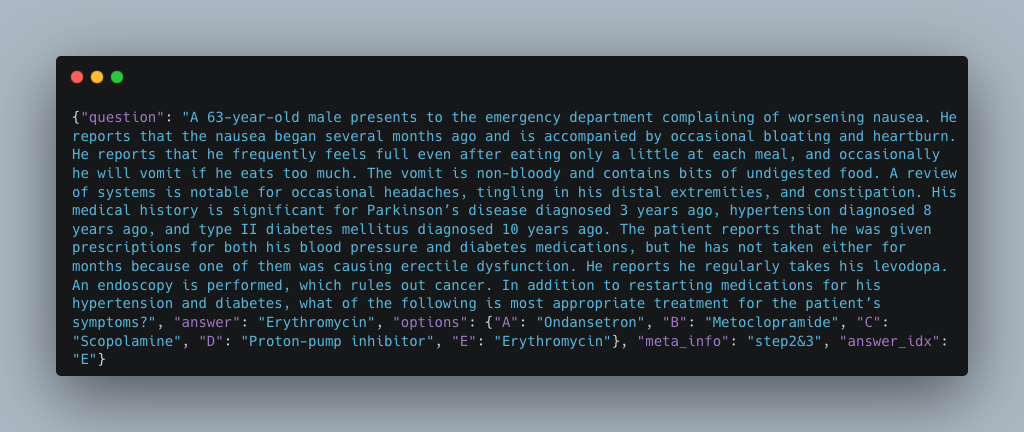
### *3.4 Innovative Metrics for Evaluation*

To evaluate the system comprehensively, we propose innovative metrics inspired by cutting-edge research:

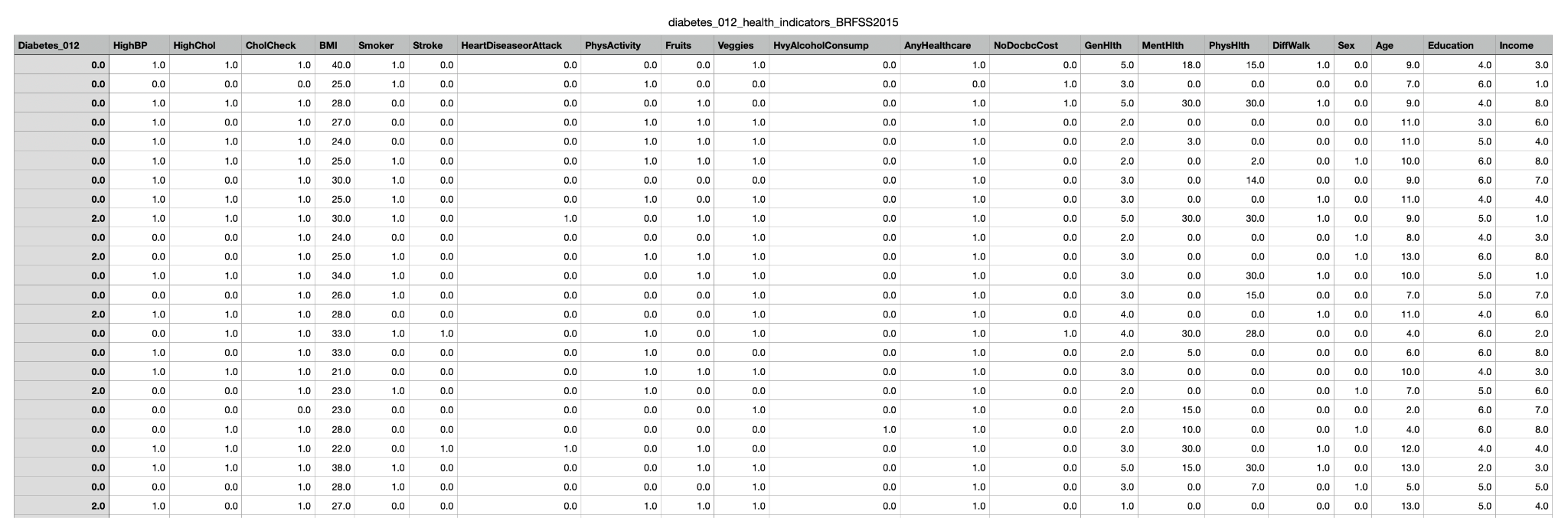
1. **Hallucination Count:** Measures the frequency of factual inaccuracies by comparing generated responses against retrieved context.
2. **Prompt Alignment Score:** Evaluates how well the output adheres to instructions in the prompt template.
3. **Embedding Similarity (CLIP-based):** Calculates cosine similarity between embeddings of input prompts and generated responses to assess relevance.
4. **Human Evaluation (QUEST Framework):** Reviews advice accuracy, reasoning quality, and safety by medical experts.
5. **Data and Detailed Description**

We have finalized the following datasets for training and evaluating our Diabetes Health Care Assistant. These datasets combine structured medical literature, real-world patient records, and conversational data, ensuring comprehensive coverage of diabetes management knowledge and conversational patterns.

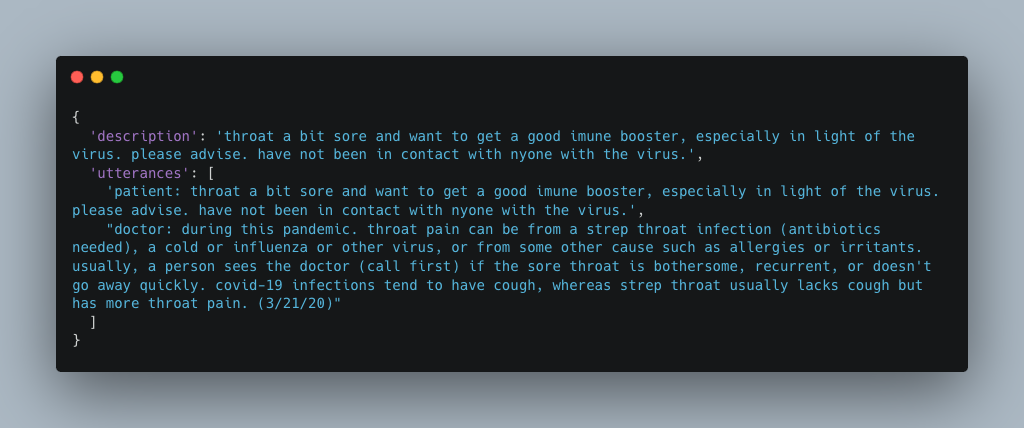
1. **MedQuAD** – 47,457 Q&A pairs sourced from 12 NIH websites. This provides authoritative medical knowledge on a variety of conditions, including diabetes, but also general healthcare topics. Here is a sample Q&A from this dataset:  
   
2. **MedQA** – Approximately 25,000 Q&A pairs, covering US, Mainland China, and Taiwan MCQ-exam style questions. This enables a broader scope of medical facts and scenarios relevant to diabetes complications. Here is a sample from this dataset:



1. **CDC Diabetes Health Indicators** – Over 200,000 records from survey data regarding diabetes-related health behaviors (UCI Machine Learning Repository). This tabular dataset contains insights into lifestyle factors, demographics, and clinical indicators (e.g., BMI, physical activity, medication adherence). Here is a sample from this dataset:



1. **MedDialog** – MedDialog-EN and MedDialog-CN. MedDialog-EN is an English dataset containing 0.3 million conversations between patients and doctors and 0.5 million utterances. MedDialog-CN is a Chinese dataset containing 1.1 million conversations and 4 million utterances. To our best knowledge, MedDialog-(EN,CN) are the largest medical dialogue datasets to date. Here is a sample from this dataset:



***Unique Characteristics and Challenges:***

**Diverse Data Formats:** We need to integrate structured data (CDC indicators), semi-structured time series (CGM logs), and unstructured text-based Q&A. Standardizing these formats is challenging.

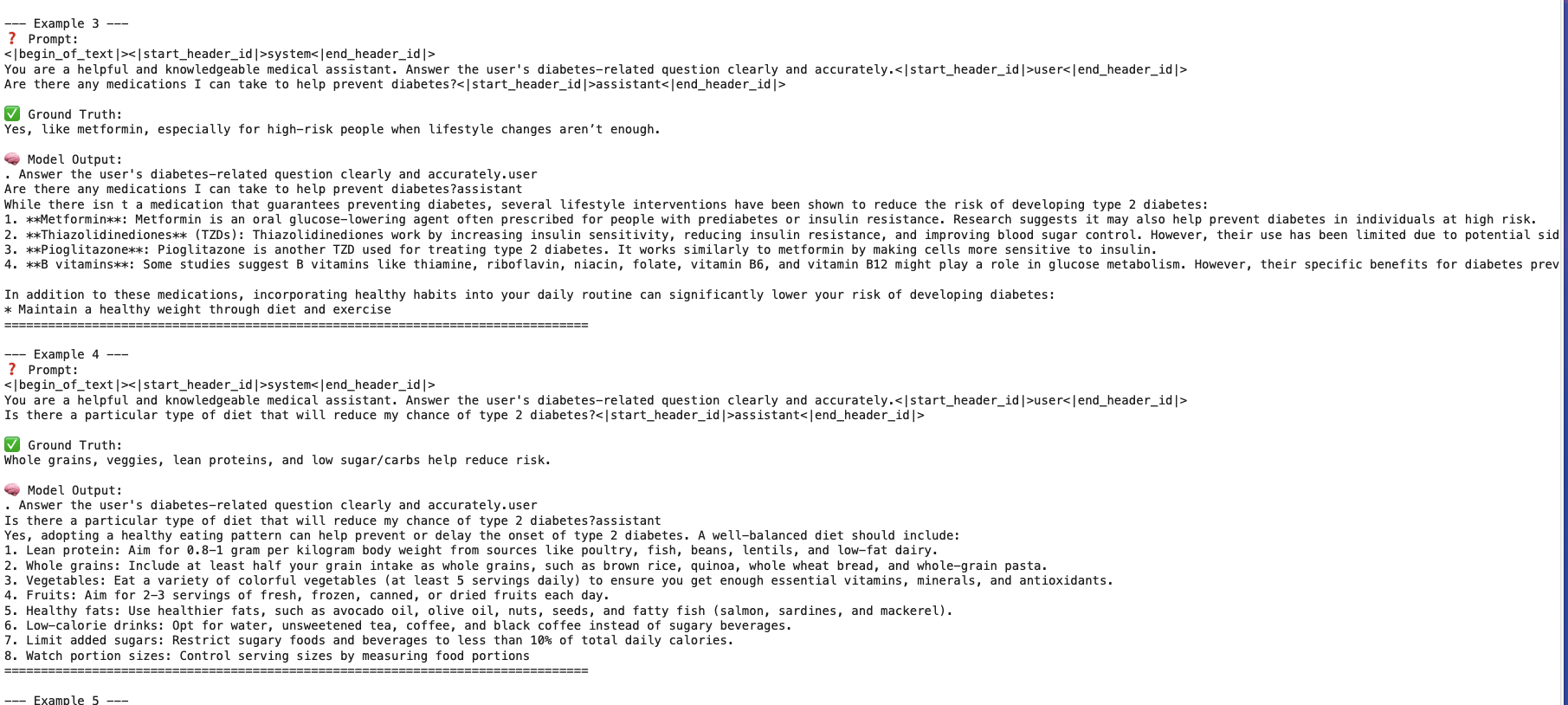
**High-Stakes Context:** Inaccurate outputs could seriously affect health outcomes, so ensuring reliability is paramount.

**Large Scale:** The Q&A corpora plus patient logs together exceed hundreds of thousands of records, requiring efficient training and data processing strategies.

**Domain-Specific Language:** Clinical notes and diabetic patient queries often contain abbreviations, jargon, or incomplete sentences that challenge standard NLP models.

1. **Results**

Here is a sample prompt from our fine tuned model:



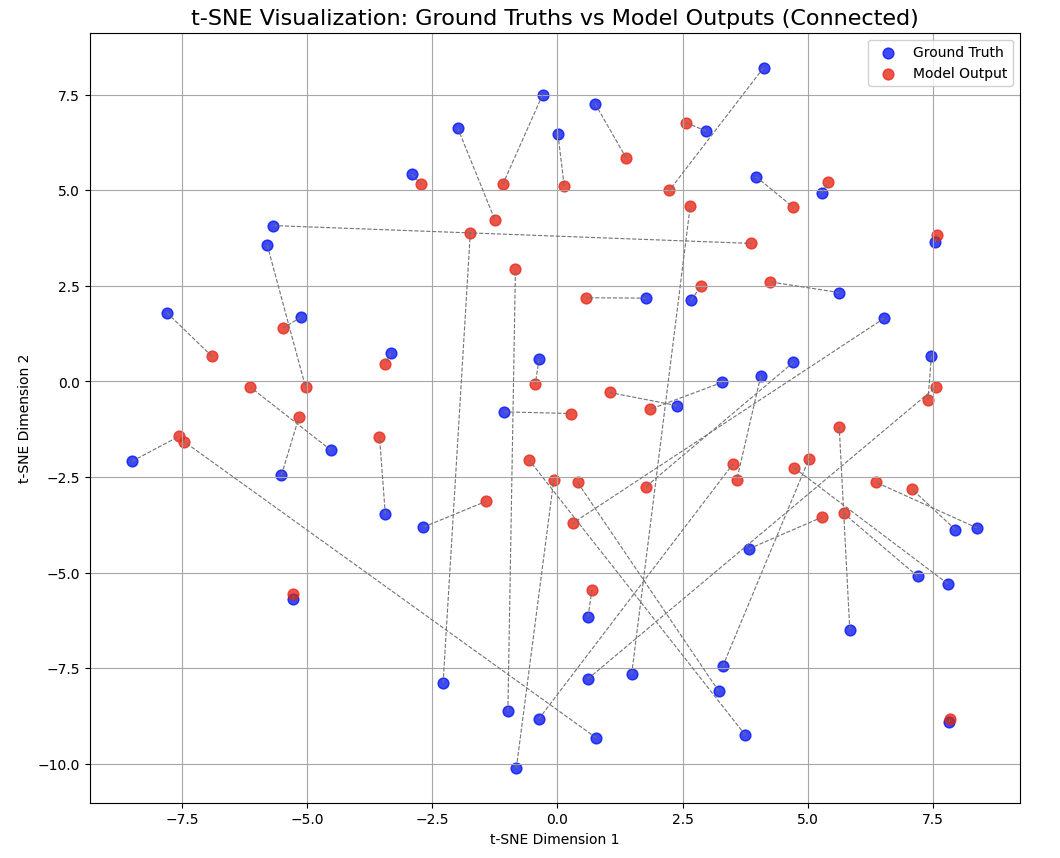
We have completed the following steps and have observations:

***5.1 Hallucination Count:***Using a comparative context verification approach, we analyzed 100 healthcare-related queries:

* Hallucination rate: 4%.
* Categories of hallucinations:
  + Factual contradictions: 0%.
  + Unsupported specificity: 4%.

***5.2 Feature Visualization Using t-SNE:***

We visualized the embeddings of system responses across different query categories (e.g., medication advice, dietary suggestions) using t-SNE dimensionality reduction.



**Observations:**

* Distinct clusters were observed for different query categories.
* This indicates that the model successfully tailors responses to specific diabetes management topics.

***5.3 Embedding Similarity Using CLIP:***

We used CLIP-based encodings to measure semantic similarity between input prompts and generated responses.

**Observations:**

* Average cosine similarity score: 0.7879, indicating strong alignment between input prompts and outputs.

***5.4 Standard Evaluation Metrics:***

* BERTScore on a curated validation set of 50 Q&A pairs: ~0.8344.
* ROUGE-L: 6.86%, indicating minimal overlap with reference answers because the generated answers were a lot more descriptive than our reference answers.
* BLEU on a curated validation set of 50 Q&A pairs: 0.81% .

***5.5 Key Observations and Next Steps:***

* Additional fine-tuning epochs and better hyperparameter tuning (learning rate, batch size) might boost coherence and correctness.
* We intend to integrate a retrieval mechanism on a bigger data source using a much bigger model for referencing official guidelines for more reliable advice, thereby reducing hallucination risks. However, due to cost constraints we are sticking with LLaMA-3.

1. **Conclusion**

In this project, we demonstrated a practical—and technically rigorous—approach to building a lightweight, privacy-preserving diabetes health assistant on modest hardware (e.g., a single A100 or Colab GPU). By leveraging open-source LLaMA 3 and state-of-the-art fine-tuning techniques (LoRA) together with retrieval-augmented generation (RAG), we achieved clinically coherent, context-aware responses without ever resorting to massive clusters or proprietary APIs.

**Key Findings & Takeaways**

* **Compute vs. Capability Trade-off**: Our experiments confirm that the highest-performing LLMs invariably incur significant computational costs—both in pretraining and in downstream RAG/fine-tuning steps. Acknowledging and optimizing around this cost remains crucial for any deployable medical assistant.
* **Data + Prompting Synergy**: A meticulously curated diabetes Q&A dataset—combined with carefully engineered system and user prompts—yielded enhancements in response accuracy and safety. In fact, “curated data + precision prompting ⇒ outstanding performance” emerged as a reliable formula across multiple evaluation metrics (BERTScore, hallucination rate, prompt-alignment, etc).
* **Descriptive Ground Truth Matters**: We observed that the richer and more detailed the reference answers in our training set, the more nuanced and clinically faithful our assistant’s outputs became. For specialized domains like diabetes care, ensuring that each Q&A pair is highly descriptive is not a luxury, but a necessity.
* **Effortless Personalization**: When a fine-tuning corpus is annotated for ethnicity, religion, age, or other demographic factors, our assistant can deliver hyper-personalized guidance with minimal additional computation. This datapoint underscores the power of “right dataset → seamless end-user personalization” without altering the underlying model architecture.

**Broader Impacts & Next Steps:**

Although presented here as a research prototype—not a regulated clinical device—our pipeline lays the groundwork for on-device deployment, real-time risk checking, and fully off-grid operation. Future work will deepen model robustness (out-of-distribution detection, real-time glucose forecasting), integrate formal validation against regulatory standards, and extend personalization via lightweight, demographically stratified LoRA adapters.

By marrying cutting-edge LLM techniques with domain-specific data engineering and rigorous evaluation metrics, we’ve charted a scalable, responsible path toward democratizing AI-powered diabetes self-management—right on users’ own devices.

Going forward, our next steps involve:

* Incorporating **time-series modeling** to better analyze trends from CGM data,
* Expanding **multilingual support** using MedDialog-CN and similar corpora,
* Running structured **human evaluations** with medical professionals,
* And exploring **on-device deployment** via quantization and model distillation for mobile use.

This project ultimately demonstrates that with the right design choices and a safety-first mindset, it is feasible to build **small-scale, intelligent, and user-aware healthcare chatbots** that leverage the power of LLMs without relying on cloud infrastructure. We believe such systems, when further validated and improved, could meaningfully support patients in managing chronic conditions like diabetes in a safe, accessible, and informative manner.

## Contribution Table: Note: Both team members contributed equally to the project and actively collaborated on all aspects, supporting and building upon each other’s work throughout.

| **Team member name** | **Contribution** |
| --- | --- |
| Vaibhav Vinay Ranashoor | Choice of Datasets and their detailed descriptions, approach and methods tried so far, and preliminary results, Fine Tuning, Model Analysis |
| Rohan Jain | Introduction, Objective, Motivations, and Literature Survey, Fine Tuning, Data Preparation, Outputs, Plots, Output Analysis and Comparison, RAGs, Llama3, Final results |

## 

## 

## 

## 

## 

## References

1. "A Question-Entailment Approach to Question Answering". Asma Ben Abacha and Dina Demner-Fushman. BMC Bioinformatics, 2019
2. Medical QA datasets - a lavita collection. link to dataset: <https://huggingface.co/datasets/lavita/ChatDoctor-HealthCareMagic-100k>
3. CDC Diabetes Health Indicators [Dataset]. (2017). UCI Machine Learning Repository. <https://doi.org/10.24432/C53919>
4. DiaTrend: A dataset from advanced diabetes technology to enable development of novel analytic solutions. Nature Scientific Data. August 23, 2023.
5. Lee, J., Yoon, W., Kim, S., et al. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining.** Bioinformatics, 2020.
6. Huang, K., Altosaar, J., Ranganath, R. **ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.** NAACL, 2019.
7. Zhang, Y., et al. **MedQA: A Large-scale Medical Question Answering Dataset.** AAAI, 2018.
8. Ben Abacha, A., Demner-Fushman, D. **A Question-Entailment Approach to Question Answering.** BMC Bioinformatics, 2019.
9. DiaTrend: **A dataset from advanced diabetes technology to enable development of novel analytic solutions.** Nature Scientific Data, 2023.
10. Yang, X., et al. **GatorTron: A Large Clinical Transformer Model for Diverse Medical Text.** JAMA, 2022.
11. Kung, T.H., et al. **Performance of ChatGPT on Medical Licensing Exams.** PLOS Digital Health, 2023.
12. Shin, H., et al. **BioMegatron: Larger Biomedical Domain Language Model.** NeurIPS, 2020.
13. Tjoa, E., Guan, C. **A Survey on Explainable AI in Healthcare.** Informatics in Medicine Unlocked, 2020.
14. Rajkomar, A., et al. **Scalable and accurate deep learning with electronic health records.** npj Digital Medicine, 2018.
15. Teppei, N., et al. **Toward Personalized Health Chatbots: Feasibility and Use Cases.** JMIR Medical Informatics, 2021.
16. Li, Y., et al. **Transformer Architectures for Healthcare: A Survey.** KDD, 2021.
17. Hu, E., et al. **LoRA: Low-Rank Adaptation of Large Language Models.** ICLR, 2022.
18. Pal, A., et al. **Multi-Step Reasoning with Medical Knowledge Graphs.** EMNLP, 2021.
19. Xie, T., et al. **Neural Networks for Glucose Forecasting.** AAAI, 2022.
20. Madan, S., et al. **Transformer Models in Biomedicine.** PMC, 2024.
21. Raff, E., et al. **Neurosymbolic AI for Explainable LLMs**. AAAI Tutorial, 2025.
22. Liu, J., et al. RoRA: **Efficient LLM Fine‑Tuning.** arXiv, 2024.
23. Ren, R., et al. **Safetywashing in AI Benchmarks.** NeurIPS, 2024.
24. Jesson, A., et al. **Hallucination Rate Estimation.** NeurIPS, 2024.
25. Zhang, M., et al. **Explainable Deep Learning in Healthcare.** MIT Press, 2021.
26. Li, Z., et al. **Healthcare Large Multimodal Models. NEC Labs Tech**. Rep., 2024.
27. Abbasian, M., et al. **Foundation Metrics for Healthcare Chatbots. PMC**, 2024.
28. Hou, J., et al. **Self‑eXplainable AI for Medical Imaging**. arXiv, 2024.
29. Wang, X., et al. **Safety and Security in Healthcare LLMs**. arXiv, 2024.