A LARGE LANGUAGE MODEL IN THE MEDICAL DOMAIN WITH A FOCUS ON DIABETES

Ijlal Sellak Ibrahim Bouaicha Badr Joulali

ABSRACT

Medical Large Language Models (LLMs) have demonstrated impressive performance on a wide variety of medical NLP tasks; however, there still lacks a LLM specifically designed for phenotyping identification and diagnosis in diabetes domain. Moreover, these LLMs typically have several billions of parameters, making them computationally expensive for healthcare systems. Thus, in this study, we propose DiabetesLLM, a model with 7 billion parameters built on a Mistral architecture, fine-tuned on more than 14000 clinical Q&A and over 5000 diabetes specific Q&A pairs covering the 2 diabetes types.

1. Introduction

Non-communicable diseases (NCDs)—including diabetes, cancers, chronic respiratory diseases, and cardiovascular diseases (CVDs) —are now the leading cause of death in Morocco, accounting for approximately 80–84% of all deaths in the country. This burden is significantly higher than the global average and is rising, with NCDs impacting all age groups but especially older adults.

As of 2021, diabetes affects approximately 9.1% to 9.7% of Moroccan adults aged 20-79, translating to over 2.3 million adults living with the disease. The country is notable for its high prevalence of type 1 diabetes among children, with more than 43,000 cases under the age of 19, placing it among the top ten globally for this age group. Diabetes imposes a significant financial strain on individuals, families, and the healthcare system.

2. Methodology

- **Language model choice :**

Mistral 7B is an efficient, scalable language model with 7 billion parameters, built on a decoder-only transformer architecture. It uses advanced attention mechanisms like Sliding Window Attention (SWA) and Grouped-Query Attention (GQA) to boost speed and memory efficiency. Mistral performs well on the Massive Multitask Language Understanding (MMLU) benchmark, excelling in areas like math, code generation, and reasoning, comparable to larger models. It is especially effective in specialized fields like medical diagnostics. Models like CancerLLM, trained with Mistral, show strong performance in healthcare, particularly in diabetes management.

- **Data collection and preparation:**

In this project, a dataset comprising over 15.000 Q&A pairs was generated using medical papers and a Kaggle dataset in addition we collected over 1.500 diabetes related question, simulating queries from diabetes patients. These questions were collected from various reliable sources to ensure a diverse and representative dataset.

- **Fine tuning**

The model undergoes fine-tuning using this dataset that focuses specifically on medical Q&As. This dataset includes detailed information about diabetes types, management strategies, medications, lifestyle recommendations, and complications. The fine-tuning process helps the model become proficient in understanding and responding to both general medical queries and diabetes-specific queries, ensuring that it can provide accurate, relevant answers to patients or healthcare providers.

LoRA was used to fine-tune Mistral-7B efficiently by updating only a small subset of parameters. By adding low-rank matrices to attention layers, it reduced memory and computational costs, enabling faster, more resource-efficient training while maintaining strong performance.

3. System Design and Architecture

Component	Description
Base Model	Mistral-7B (open-source LLM)
Fine-tuning	LoRA (Low-Rank Adaptation) for efficient parameter tuning
Frameworks	Hugging Face Transformers, PEFT (for LoRA)
Compute	Google Colab with A100 GPU (low-cost, accessible hardware)
Training Data	Custom medical and diabetes Q&A dataset (15,500 + 1500 examples) curated and cleaned
Interface	Web app
Deployment	TBD

4. Objectives & Impact Analysis:

Accuracy: We are hoping to achieve over 90% relevance in patient Q&A sessions compared to expert-verified answers.

Speed: Provide real-time (<5 second) answers on consumer-grade devices.

Accessibility: Make diabetic education available to underserved communities.

Immediate benefit: Empower diabetic patients with accessible, accurate, and friendly support — reducing misinformation and improving self-care.

Community focus: Initially target Moroccan regions with limited access to medical care.

Scalability: The system can easily expand to cover other diseases (e.g., hypertension, cardiovascular health) by fine-tuning on new datasets.

Long-term improvements:

- Early diagnosis awareness
- Reduction in diabetes-related complications
- Decrease in public healthcare system overload

5. Conclusion

DiabetesLLM, a specialized, efficient language model based on **Mistral-7B**, fine-tuned to improve diabetes diagnosis and education in Morocco. Using a focused dataset of over **17,000 Q&A pairs**, the model delivers fast, accurate answers while remaining lightweight and accessible.
By addressing gaps in patient education and healthcare support, DiabetesLLM aims to reduce complications, promote early diagnosis, and strengthen community health. Future work will expand its scope to other non-communicable diseases.