Project Proposal *–*

LLM-Powered Natural Language Query System for Healthcare Analytics

(by – Keerthana Bhukya, Tran Thai Dang Khoa, Vraj Rajesheshkumar Soni)

Project Idea -

This project proposes the development of a Large Language Model (LLM)-powered natural language query system designed for healthcare analytics. The system will allow users to interact with healthcare databases using plain English queries rather than technical programming languages such as SQL.

Healthcare organizations generate vast amounts of structured data covering patient demographics, medical conditions, hospital admissions, and billing information. While this data is highly valuable for improving patient care and operational efficiency, it often remains inaccessible to non-technical users due to the complexity of database query languages. For example, retrieving insights such as *"How many diabetes patients were admitted under Medicare last year?"* usually requires IT staff or data analysts to write specialized queries.

Our system addresses this challenge by acting as a translator between natural language and database operations. Users will be able to type queries in plain English, which the LLM will translate into structured queries (SQL or Pandas). The system will then execute them securely and return results in both numerical and narrative form, enhanced with visualizations such as bar charts or histograms.

This approach makes healthcare data more accessible, interpretable, and actionable, enabling doctors, administrators, and insurance professionals to make faster and more informed decisions.

Objectives -

The objectives of this project are to:
Improve Accessibility – Allow healthcare professionals without technical expertise to query complex datasets using natural language.
Enhance Efficiency – Deliver real-time insights to reduce delays caused by reliance on IT staff and lengthy report-generation cycles.
Increase Clarity – Provide results with plain-English explanations and intuitive visualizations to make insights easier to understand.
Support Analytics Functions – Enable users to identify trends, analyze distributions, compare groups, and generate visual reports.
Ensure Reliability – Evaluate performance using at least 30 benchmark queries to measure accuracy and identify errors.
Promote Transparency – Display the generated SQL alongside outputs so users can validate and trust the system's results.

By meeting these objectives, the project will demonstrate how LLMs can effectively bridge the gap between human intent and technical database queries, advancing healthcare analytics toward greater inclusivity and trustworthiness.

Dataset -

The project will use an existing synthetic healthcare dataset containing 10,000 patient records. The dataset is designed to simulate real-world healthcare data while avoiding privacy issues.

Attributes:
Demographics: Name, Age, Gender, Blood Type
Medical Information: Medical Condition, Test Results, Medications
Hospitalization Details: Admission Type, Admission Date, Discharge Date, Hospital, Doctor
Financial Information: Billing Amount, Insurance Provider

Rationale for Dataset Selection:
Comprehensive Coverage – Includes demographics, conditions, admissions, and billing, enabling diverse queries across multiple aspects of patient care.
Structured Format – Ideal for testing the ability of LLMs to translate natural language into accurate SQL queries.
Synthetic and Ethical – Contains no real patient data, ensuring privacy and compliance with HIPAA while remaining realistic enough for testing.

This dataset provides a strong foundation for evaluating the system's ability to handle diverse healthcare queries and deliver reliable, user-friendly insights.