Text Classification for Medical Data using Text Embedding models and Vector Databases

***Abstract*—This paper discusses and measures the robustness of combining Large Language Models (LLM), Text embedding Models and Vector databases in measuring the classification accuracy of medical queries to be classified into the correct ailment.**

Index Terms*—*Vector, Database, LLM, Text Embedding Models, Text Classification

I. Introduction

In this paper the Author will use a specific LLM such as **GPT-3.5 turbo**, **LLaMA 2** and **Google-flan** for medical query production. Then use Text embedding models such as **text-embedding-ada-002** and **textembedding-gecko@001** for text embedding the queries, then uses the same LLM models to produce ground truth data in the vector Databases. To have a Total of 18 permutations and analyse the classification results.

II. Methods

A. Vector Database

A vector database is a method of storing many vectors or lists of numbers. In this specific Research the author used **Pinecone** Vector Database.

B. Vector Embeddings

A vector embedding is a numerical representation of any type of data, in this context is converting text into numerical data using **OpenAI's** text-embedding-ada-002 model and **Google's** textembedding-gecko@001

C. Framework

"Ground Truths" values is Going to be generated using one of the LLM such as **GPT-3.5 turbo**, **LLaMA 2** and **Google-flan** then stored in The Pinecone Vector Database and then be Tested using the following process, a medical query is going to be generated using one of the LLM used in the research then the query will be text embedded , then the output of text embedding will be used to query the type of ailment in the vector database to be able to measure the classification accuracy.

D. Dataset

The research conducts the test on 8 ailments Glaucoma, Jaundice, Cyanosis, Psoriasis, Conjunctivitis, Scoliosis, Skin cancer, Gingivitis. Each ailment has one in the vector database acting as the ground Truth, not to mention that the data brought for these ailments are produced and gathered from the LLM.

III. Results

A. Robustness Test

To address the issue of the dataset being generated by LLMs instead of medical professionals, the author thought about if the embedding model can handle inputs from different LLMs, the outputs of which consist of varying levels of length and specificity, then the embedding model is able to find similarities in data even though the ideas presented are represented in different ways.

B. Performance with Different Models as Query and Knowledge Base

It is not surprising if a text embedding model is able to more clearly detect similarities of text generated from the same model, but how these models perform when queried with data from other models.

Using **LLMA 2** for query generation and **ada-002** for embedding and **GPT-3.5** for ground truth dataset. This did very quit well with very few misclassification noticing that LLMA and GPT have comprehensive data representation however they are different in data representation.

Using **Google-flan** for query generation and **ada-002** for embedding and **GPT-3.5** for ground truth dataset. This did Exceptionally well with only 58 misclassifications out of 1600 queries. This appeared because **Google-flan** generates exceptionally short queries, so it is easier to use in finding similarities in vector database especially that **GPT** produces comprehensive data for ground truth.

Using **GPT-3.5** for query generation and **ada-002** for embedding and **Google-flan** for ground truth dataset. This Did the worst with a lot of misclassifications specially in skin cancer only correctly classified 35 cases of skin cancer and misclassified 110. This is because flan generates exceptionally short sentences and GPT produces comprehensive amounts of data.

C. Vertex AI embeddings vs OpenAI embeddings

The same thing was repeated except that the text embedding model has been changed to **textembedding-gecko@001.** This made less accurate classification on smaller scale because the **ada-002** has double dimensionality results than **gecko so ada-002** can encode more information into its vectors.

IV. Future work

The purpose of this paper was to study and measure the viability of text embedding models and vector databases to classify medical text and present a new use case for these emerging new technologies in the field of AI.

A possible continuation of this study could be the exploration of multimodal embeddings in the field of medicine. However, the author chosen not to incorporate multimodal embedding in the process due to lack of state-of-the-art models to use however it is recommended to use these models to see how they improve/hurt the use of Vector embedding and vector databases in field of medicine.