Large Language Models for identification of medical data in unstructured records

The purpose of this thesis was to test and evaluate how Large Language Models extract key points of information in medical data, have them further analyze the input and expand on it by providing predictions on its contents, and finally, we measure their comparative performance.

This repository contains scripts for making requests to the respective API endpoints, which were used to generate the output text, manually classified input data from a sample of unstructured medical records, files with the predictions the LLMs made, and a script that calculates the F1-Score and visualizes a confusion matrix for each given model, based on the output data.

Below are the questions we aimed to answer.

1. *Does the text contain information about the patient's bad habits? – Yes or No.*
2. *Is the patient a smoker? – Yes or No.*
3. *Does the text contain information about the duration of diabetes in the patient? – Yes or No.*
4. *What is the duration of diabetes in the patient? – Number of years.*

After months of reviewing the latest models, figuring out how to create a scalable and automated approach to generate the answers, and coming up with a way to evaluate the results, this repository contains everything used in our solution, outside of the models themselves.

Many of the models are free to use, and we took advantage of that by downloading them from the platform Hugging Face. The models are quantized in the "GGUF" file format. While this hinders performance, the quantization options vary and more precise versions or the raw models themselves, if hardware capabilities allow it, can be used. However, some of the models have charges associated with them per number of input/output tokens. This would have to be considered for any further testing or implementation of our approach. The final lineup of models we ended up using is listed below.

- *llama-2-13b.q8_0.gguf*
- *hermes-2-pro-mistral-7b.q8_0.gguf*
- *vicuna-13b-v1.5.q8_0.gguf*
- *wizard-vicuna-13b.q8_0.gguf*
- *bggpt-7b-instruct-v0.2.q8_0.gguf*
- *mistral-7b-instruct-v0.2.q8_0.gguf*
- *gemma-7b.q8_0.gguf*
- *claude-3-opus-20240229*
- *gpt-3.5-turbo*
- *gpt-4*
- *gpt-4-0125-preview*

We saw a lot of promising results with the free models, which the community continually improves upon, by fine-tuning and quantizing in ways that allow for exploitation by a larger subset of people. In the end, the best performing models were the paid ones, and particularly OpenAI's GPT models, GPT-4 having achieved a 96% F1-Score accuracy and a small percentage of incorrect values in the confusion matrix.

https://www.researchgate.net/publication/379754048_Large_Language_Models_for_identification_of_medical_data_in_unstructured_records