Кратки теоретични бележки, които би могъл да включиш в презентацията или в самата теоретична част.

Large Language Models (LLM) или на български Големи Езикови Модели са големият пробив, който се направи наскоро в полето на ИИ. Това е езиков модел, състоящ се от невронна мрежа с много параметри, тренирани върху големи обеми от текст с надзираващо учение или полу-такова. ChatGPT e най-известия в момента такъв.

Големите му предимства са, че може да обработи голям обем от текст, да добие разбиране за този текст и да върне релевантната информация. В тази връзка може да се използва за създаване на рапорти от най-различни документи. Може да се използва за изясняването на правно-юридически документи, които често съдържат сложен и неразбираем изказ, които за непрофесионалисти е труден за разбиране. Всъщност всякакъв по-технически изказ може да бъде сведен до едно по-разбираемо ниво. Търсенето в множество документи на определено нещо може значително да се улесни. Създаването на резюмета и рапорти би могло да става много по-бързо.

Потенциалните ползи са огромни.

Но трябва да се имат предвид лимитациите. OpenAI имат лимит колко символа могат да се изпратят. В trial версията на модела ChatGTP 3.5, лимита е 4096 символа на един промпт. Това означава, че много големи текстове трябва да се раздробят на по-малки, като не се губи смисъла. Това и е целта на Langchain и Cohere. Текстът се раздробява на малки части, като всяка част съдържа по-малко и от другите, след това се превръщат думите в числа, защото операциите с числа са значително по-бързи.

A picture containing diagram, text, font, plan

Description automatically generated

Процесът е илюстриран на тази картинка. От документа се изтегля съдържанието, разделя се на по-малки части, защото OpenAI имат лимит на символите. Всяка част се превръща във вектор от числа. Тези числа се групират на едно място, което се нарича Knowledge Base. Когато се зададе въпрос от потребителя, въпросът се превръща отново във вектор от числа, изпраща се към Knowledge Base, прави се сравнителен анализ на търсенето и частта, която е най-близка до въпроса се връща като отговор. Това което прави OpenAI много успешно е да превърне думите в такива вектори от числа, чиито стойност много добре отрязва връзката им с останалите думи.

В нашият случай е използвана българската Конституция, като тя е разделена на членовете си. Тези които са прекалено дълги, са допълнително раздробени. Могат да се задават въпроси свързани със законите на България. Трябва да се има предвид, че на български не работят толкова добре нещата, понеже OpenAI не е имал бил трениран достатъчно на български. Другият голям недостатък са така наречените халюцинации. Това се отговори, които моделът сам си измисля, давайки уверен вид, че знае отговор. За да се избегне отчасти това, се добавя към въпросът да посочи източник и също когато не знае да каже, че не знае. Фалшивата увереност може да е голям проблем.

Langchain е библиотека за python, която се стреми да използва API на OpenAI и да реализира по-голям потенциал на модела. Тук може да се контролира какъв да бъде отговорът, стилът, да има запомняне между отговорите и тн.

Cohere се използва за създаването на тези Embeddings или превръщането на думите във вектор от числа. За създаването на резюме се използва K-mean clustering. Това включва клъстерирането на текста на групи по еднозначност на значениено.