Prompt Engineering Házi Feladat

Készítette: Czotter Benedek
Neptun: TFB4FY

1. Választott feladat leírása:

44. Egy adott ország társadalmi-gazdasági helyzetének gyors összefoglalása

- Az LLM generáljon egy rövid elemzést egy adott ország demográfiai, gazdasági és politikai helyzetéről.

2. Megvalósítás:

A feladat célja egy olyan LLM-alapú alkalmazás létrehozása volt, amely gazdasági és szociológiai elemzést készít egy adott országról. Ehhez a llama_cpp könyvtár segítségével egy GGUF formátumú nyelvi modellt töltöttünk be. A felhasználó egy ország nevét adja meg, amely alapján az alkalmazás egy részletes elemzést generál.

A fejlesztés során sikerült egy letisztult és esztétikus grafikus felületet (GUI) létrehozni a Tkinter segítségével. Az alkalmazás tartalmaz egy beviteli mezőt az ország nevének megadására, egy gombot az elemzés elindításához, valamint egy görgethető szövegmezőt az eredmény megjelenítéséhez. A vizuális élményt tovább javítottuk egy fejléc hozzáadásával és színes, formázott elemekkel.

A legnagyobb kihívás a válaszadási sebesség volt, mivel a modell mérete és a generált szöveg hossza befolyásolta a teljesítményt. Ezt a max_tokens paraméter finomhangolásával próbáltuk optimalizálni. Összességében a megoldás jól működik, de további fejlesztési lehetőségek vannak, például a válaszok streamelése vagy GPU-gyorsítás használata a teljesítmény növelése érdekében.

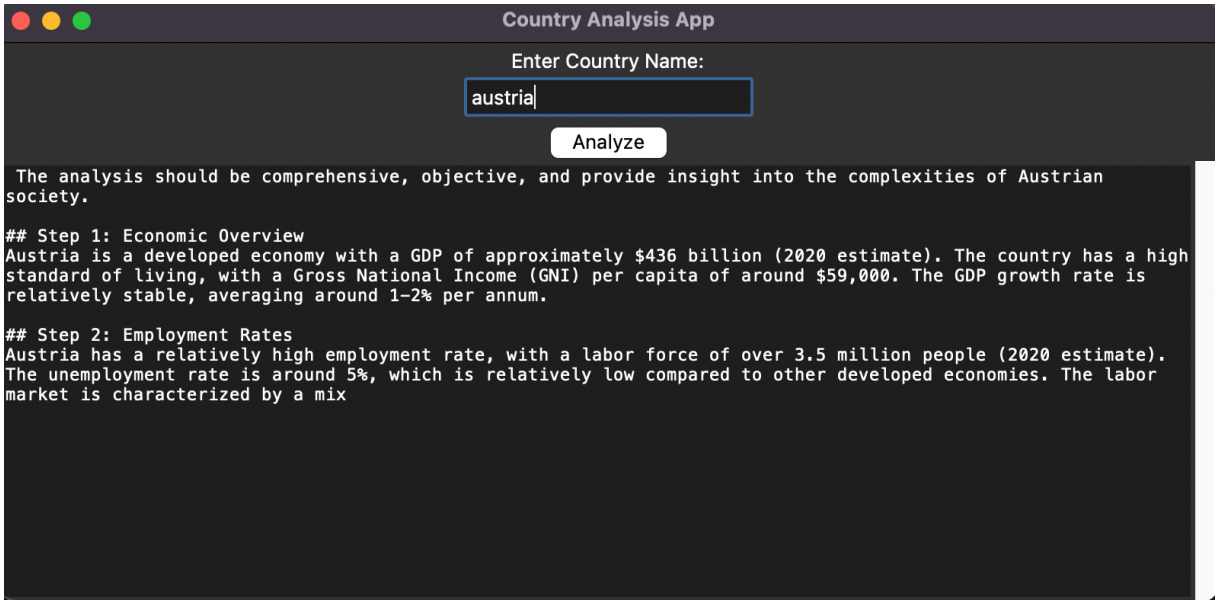
3. Kód elérhetősége:

Github: https://github.com/czotterbenedek/prompt-engineering-hf-czb

A használt .gguf fájlt a mérete miatta nem tudtam feltölteni a repository-ba. Az alábbi modellt használtam: Llama-3.2-1B-Instruct-Q5_K_M.gguf

Link a modellhez: https://huggingface.co/bartowski/Llama-3.2-1B-Instruct-GGUF/blob/main/Llama-3.2-1B-Instruct-Q5_K_M.gguf

4. Az alkalmazás futásáról kép:



5. Legfontosabb használt promptok:

- Create a small llm-based application in python which is prompted to give back a small economical and sociological analysis about the given country. Use a hugging face model and the gguf file which can be downloaded from hugginface. The app should have a small graphic interface.
- how to make the answer longer? and make the prompt more precise
- Why can it be that the response generation is very slow?

Illetve a kapott hibaüzenetek értelmezése is végig Chat-GPT és Gemini segítségével zajlott. A dokumentáció tartalmát is a lehető leginkább a nagy nyelvi modellek segítségével készítettem, kisebb javításoktól eltekintve.