Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

POSUDEK OPONENTA DIPLOMOVÉ PRÁCE

Student: **BC. LUKÁŠ RICHTER** **Oponent:** **Ing. Jiří Pálka, Ph.D.**

Studijní program: **Informační technologie**
Studijní obor/Specializace: **Softwarové inženýrství**
Akademický rok: **2023/2024**

Téma diplomové práce: **Implementace LLM jako chatbot pro e-shop**

Hodnocení práce:

Předložená diplomová práce se věnuje implementaci LLM jako chatbota, konkrétně na e-shopu. Požadovanou funkcionalitou chatbota je, aby disponoval vlastní knowledge base. Téma práce je velmi aktuální a lze ji určitě považovat za solidní základ pro využití v praxi.

Práce má bez příloh 66 stran. Z formálního hlediska je dobře členěná, kapitoly na sebe logicky navazují, použitá literatura je v práci řádně citována a celkem autor vychází z 52 zdrojů.

V teoretické části autor popisuje chatboty, jejich typy a technologie. Ve druhé kapitole teoretické práce se věnuje velkým jazykovým modelům (LLM) a stručně popisuje jejich princip a architekturu. Vedle vlastního přínosu LLM neopomněl popsat i možná rizika, která nese nasazení chatbota založeného na velkém jazykovém modelu. Ve třetí kapitole teoretické části je popsán přístup RAG (Retrieval Augmented Generation), pomocí kterého se dá LLM rozšířit o znalostní bázi a kterou diplomant následně využil v praktické části. Závěr teoretické části je věnován požadavkům na funkčnost a návrhu vlastního řešení.

V praktické části diplomant vhodně vybral a použil nástroje Ollama a LangChain, které výrazně usnadňují vývoj a experimentování s LLM. Oba nástroje patří mezi špičku v tomto oboru. Ollama poskytuje snadný způsob lokálního spuštění open-source LLM a framework LangChain nabízí mnoho tříd, které výrazně usnadňují integraci LLM do koncových aplikací.

Vytvořené vlastní řešení sestávající se serverové části napsané jazyku Python a klientské části postavené na HTML a JavaScriptu využívá zmíněné nástroje Ollama a LangChain. Aplikace obsahuje vlastní knowledge base vyřešenou pomocí metody RAG (Retrieval Augmented Generation) a využívající open-source vektorovou databázi ChromaDB.

Nad rámec zadání diplomant provedl vlastní porovnání open-source modelů Llama, Gemma a Mistral a proprietárních modelů ChatGPT od společnosti OpenAI.

Práce je zpracována na vysoké úrovni

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Otázky k obhajobě:

1. Měl jste omezení 16 GB RAM na grafické kartě. Pokud byste chtěl vyzkoušet open-source LLM s více parametry jak byste postupoval?

2. Jak náročné by bylo Vaši aplikaci upravit tak, aby knowledge-base nebyla načtena z vyextrahovaných textů z webu Alensa, ale např. z PDF souboru?

Celkové hodnocení práce:
Známku uvede oponent dle svého uvážení dle klasifikační stupnice ECTS:
A – výborně, B – velmi dobře, C – dobře, D – uspokojivě, E – dostatečně, F – nedostatečně.
Stupeň F znamená též „nedoporučuji práci k obhajobě“.

Předloženou diplomovou práci doporučuji k obhajobě a navrhuji hodnocení

A - výborně.

V případě hodnocení stupněm „F – nedostatečně“ uveďte do připomínek a slovního vyjádření hlavní nedostatky práce a důvody tohoto hodnocení.

Datum 23.5.2024 Podpis oponenta diplomové práce