

POSUDEK OPONENTA DIPLOMOVÉ PRÁCE

Student: BC. RUBÁŠ JAN

Oponent: Ing. Jiří Pálka, Ph.D.

Studijní program: Informační technologie
Specializace: Softwarové inženýrství
Akademický rok: 2024/2025

Téma diplomové práce: Benchmarking LLM modelů

Hodnocení práce:

Diplomová práce představuje návrh, implementaci a evaluaci vlastního open-source nástroje pro benchmarking výstupů velkých jazykových modelů (LLM) a systémů typu Retrieval-Augmented Generation (RAG). Autor porovnává kvalitu odpovědí, latenci a rozsah výstupu modelů LLaMA, Mistral a DeepSeek.

Praktická část je založena na čtyřiceti technicky orientovaných otázkách na které každý model generoval odpovědi. DeepSeek byl navíc testován v režimu RAG s využitím vektorového indexu dokumentů.

Přínosy

- Velice oceňuji aktuálnost tématu – práce reaguje na rychlý rozvoj LLM
- Modularita nástroje – lokální provoz bez odesílání dat na externí servery je vhodný pro citlivé prostředí a splňuje požadavky GDPR a firemní bezpečnostní politiky.
- Přehledné vizualizace – autor nabízí tabulky, grafy a exporty, které umožňují rychlou interpretaci výsledků.

Struktura a jazyk

Text je logicky členěn na teoretickou a praktickou část; úvodní kapitoly poskytují solidní přehled o vývoji LLM a metrikách benchmarkingu. Psaní je srozumitelné, odborné termíny jsou vysvětleny.

Přínos pro praxi

Navržený nástroj je díky jednoduché architektuře a nízkým nárokům na infrastrukturu vhodný pro další rozšiřování a může sloužit firmám i akademické sféře při lokální evaluaci modelů.

Jako **zásadní nedostatek** vnímám absenci popisu o jaké konkrétní modely se jedná. Jen jednou je v práci zmíněna generace LLaMA 2, která v době psaní práce byla asi již překonána novější verzí. Nenašel jsem nikde verzi ostatních modelů a chybí zcela zásadní informace o počtu parametrů nebo kvantizaci. Bez verze modelu a bez počtu parametrů je srovnávání modelů zcela zavádějící a prosím o vysvětlení v rámci dotazů k obhajobě.

Dotazy k obhajobě:

1. Jaké generace modelů byly použity? Kolik parametrů měly modely?
2. Proč nebyla použita novější generace modelu Llama?
3. Použité zařízení Apple Macbook M3 Pro je určitým omezením k provozování LLM modelů. Které LLM jste na něm provozoval?
4. Jak jinak jste mohl testování LLM provést, pokud byl na některé Macbook nedostatečný?

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

C - dobře.

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

Datum

8.6.2025

Podpis oponenta diplomové práce