POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Benchmarking Techniques for Evaluation of Large Language Models**
Jméno autora:	**Adam Jirkovský**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	Ing. Luboš Král, PhD
Pracoviště oponenta práce:	Fakulta elektrotechnická (FEL)

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **náročnější**

Hodnocení náročnosti zadání závěrečné práce.

Zadání má tři hlavní části:

- Testování základních a instrukčních LLM: Vytvoření Czech Bench, který umožňuje hodnotit kvalitu jazykové velkých jazykových modelů (LLM), jako je gramatická správnost, stylistika a porozumění sémantice. Dále zahrnuje testy instrukčních LLM, které vyžadují pochopení a plnění zadaných úkolů.
- Hodnocení koherence a fakticity: Czech Bench odhaduje schopnost LLM generovat koherentní a fakticky správný text
- Pokrytí různých kategorií: Czech Bench testuje LLM v rozmanitých oblastech, jako jsou systémy pro otázky a odpovědi, shrnutí textu, generování textu v různých formátech, překlad a další.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Diplomant ve své práci prokázal, že je schopen k dané problematice vybrat relevantní informační zdroje, literaturu. Tu následně srozumitelně a logicky shrnuje v úvodních částech diplomové práce. V celku je předložená práce dobře zpracovanou, promyšlenou a zejména komplexní studií dané problematiky, která se v současnosti vyvíjí neuvěřitelně rychlým tempem.

Diplomant dále pokračoval aktivním přístupem a realizoval praktickou implementaci úlohy, ke si prokazatelně osvojil dovednost vyhodnocovat modely na různých anglických platformách LLM. Na základě analýzy byly následně vybrány nejslibnější modely, které diplomant přeložil do češtiny. Je nutno podoktnout, že překlad nebyl pouhou mechanickou transformací textu, ale zahrnoval i důkladnou tvorbu a úpravu testovacích skriptů pro zajištění správného fungování v morfologicky bohaté češtině. Kromě toho diplomant samostatně trénoval několik testovacích českých LLM, čímž prokázal pokročilé znalosti a dovednosti v oblasti strojového učení. Veškeré aktivity, včetně trénování a experimentů, jsou podrobně a srozumitelně dokumentovány v sekci výsledků, čímž diplomant demonstruje svůj systematický přístup a precizní práci.

Zvolený postup řešení **vynikající**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Diplomant zvolil vhodný a adekvátní přístup k úloze, odpovídající současnému stavu v problematice evaluace jazykových LLM. Jeho postup byl systematický a je podrobně popsán v diplomové práci. Kladně hodnotím výběr odpovídající množiny testů, následnou modifikaci metrik a navržení testovacích skriptů. Součástí je také očekávaný přehled s přesností vybraných modelů. Celý postup řešení je na odpovídající úrovni a obsahuje všechny relevantní fáze vyhodnocování modelů.

Odborná úroveň **A - výborně**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Kvalita LLM modelů a jejich systematické vyhodnocování je jednou ze současných velkých výzev v oblasti umělé inteligence. Zejména s velkým množstvím možného nasazení do různých profesních aplikací je vyhodnocování a benchmarking modelů naprosto zásadní. Nemluvě o provedení této úlohy v českém jazyce, který má chronicky malé množství dat a zastoupení v

modelech.
Řešení této problematiky tedy vyžaduje porozumění široké škále inovativních technologií, které běžně nejsou součástí standardních studijních programů. Diplomant prokázal na vysoké odborné úrovni schopnost jít do hloubky problému a řešil odpovídající detaily tak, jak jsou popsány v odborné literatuře. Nezůstával tedy pouze na povrchu řešení, což je vždy pro praktické nasazení zásadní.
To samozřejmě vyžaduje odpovídající pracovní zapojení. Hledání a stahování relevantních testovacích českých textů a jejich překlad z rozmanitých zdrojů je časově velmi náročné. Vyzdvihl bych také prokázané dobré programátorské schopnosti v návrhu testovacích procedur.

Formální a jazyková úroveň, rozsah práce **A - výborně**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Diplomová práce je napsána jasnou a srozumitelnou angličtinou s důrazem na detail. Formálně tj. typografickou kvalitou, řazením obrázků, grafů atd. odpovídá standardům technických dokumentů.

Výběr zdrojů, korektnost citací **A - výborně**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Diplomová práce je postavena na rozsáhlém množství internetových a literárních zdrojů. Zahrnuje relevantní nejnovější a nejpokročilejší práce z oblasti testování modelů. Diplomant věnoval pozornost
Správné citaci odkazovaných zdrojů v souladu s citačními praktikami. Uvedené citace dodržují standardní citační formát a veškeré převzaté prvky jsou řádně odlišeny od vlastních výsledků a úvah.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Není.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Diplomová práce poskytuje komplexní přehled o současném stavu v oblasti testování a hodnocení LLM, což je v současnosti jedno z řešených a důležitých témat v odborné komunitě. Diplomant při své práci postupoval systematicky a bral v potaz veškeré aspekty příslušné technologie. Pro porozumění současnému stavu využil relevantní odborné literatury, kterou korektně citoval a v práci jsou jednoznačně odlišeny části převzaté a části vlastní práce. Ve výsledku byly tyto znalosti využity k vytvoření vlastní hodnoticí platformy a pro návrh prvního českého testovacího souboru. Výstup je má vysoké kvalitativní úrovni a má předpoklady pro využití dalšími výzkumnými skupinami, které se zabývají vývojem českých modelů. Celkově hodnotím práci na výborné úrovni.

Dotaz k obhajobě: Testování LLM je poměrně nákladné. Jak by se daly náklady na testování minimalizovat?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně.**

Datum: 7.6.2024 Podpis: