Posudek oponenta bakalářské práce

Autor/autorka práce: **Andrei Akhramchuk**

Název práce: **Artificial intelligence for facilitating software development**

Práce obsahuje přehled potenciálně použitelných AI technologií a vyhodnocení možností jejich použití. Jde rozhodně o velmi aktuální a užitečné téma, jakékoliv objektivní hodnocení výsledků současných AI modelů umožňuje udělat si lepší představu o možnostech jejich uplatnění. V práci jsou porovnávány 4 modely, zároveň obsahuje tvorbu jednoho na míru vytvořeného modelu. Bohužel jsou experimenty prováděné na jednotlivých modelech vzájemně neporovnatelné (pokaždé jde o zásadně jinou sadu pokusů, navíc vždy s poměrně malým vzorkem dat) a jejich vyhodnocení se často omezuje na subjektivní ústní popis, který je někdy i v rozporu se získanými výsledky. Řada textů působí velmi genericky, jako informace získané z marketingových materiálů nebo generované LLM – nenesou prakticky žádnou užitečnou informaci, kromě obecného komentáře o univerzální přínosnosti AI.

V popisu existujících technologií (kapitola 4) chybí odkazy na příslušné modely, očekával bych zde také informace o možnostech použití, licenčních podmínkách nebo ceně.

Řada tvrzení o AI je nepodložena odkazy na příklady modelů nebo technik které by je demonstrovaly – v textu se mluví o optimalizacích DB, optimalizaci lambda výrazů, AI syntaktických analyzátorech a transpilerech nebo o vysvětlení legacy kódu (např. „AI … improves code quality“ – p. 21, „AI can analyze database structure … and offer optimized version“, „AI can … simplify lambda expression“ – proč ne, ale chybí odkaz na AI která by to dělala). Netvrdím zde že to není možné nebo že by se podobné techniky a publikace neobjevovaly, a výrazně mi chybí odkazy na příklady demonstrující jejich vlastnosti a použití. V odborném textu je žádoucí se soustředit na tvrzení podložená fakty. Zejména v tomto typu práce, kde by měly být vyhodnoceno jaký skutečný přínos existující technologie mají. Použitá literatura je relevantní a aktuální, ale řada tvrzení neodkazuje na žádné publikované výsledky.

Řada generovaných odstavců (např. 4.1.3.) neobsahuje žádnou skutečnou informaci ani analýzu ze strany autora práce, jen obecná tvrzení o možných přínosech, bez jakéhokoliv vyhodnocení.

Nejzajímavější částí je kapitola 5, která popisuje praktické experimenty. Z textu práce je ale obtížené pochopit jakých výsledků bylo dosaženo, je nutné se zároveň podívat na výsledky v příloze. Vyhodnocení typicky probíhalo na menších datových sadách. V experimentu s vysvětlením SQL (5.1.1) není příliš rozebráno nakolik vysvětlení pomůže pochopit SQL dotazy – výsledky jsou označeny za úspěch, ale osobně se domnívám že k pochopení předložených dotazů vysvětlení příliš nepomáhá, jen přepisuje SQL dotaz do podoby, kde jsou vysvětleny základní konstrukce SQL jazyka (pomůže tedy člověku který nezná SQL, ale ne člověku, který nezná strukturu dané DB a význam jednotlivých identifikátorů). Podobné je i vysvětlení lambda výrazů (5.1.3), opět soustředěné především na syntaxi – zde se ale objevují i náznaky správně zpracované sémantiky. Obecně u hodnocení vysvětlení chybí metoda, podle které se hodnotilo, zda je vysvětlení srozumitelné / srozumitelnější než původní zdrojový text. Rozbor optimalizací SQL (5.1.2) je výrazně lepší, ale nevidím v něm vyhodnocení správnosti nových skriptů, navíc množina dotazů, na které se ověření provádělo je poměrně malá. U optimalizací lambda výrazů (5.1.4) je hodnocena jen čitelnost (opět bez jasného kritéria co je čitelnější), není hodnocena změna ve výkonu. S ohledem na prezentované výsledky považuji závěr v 5.1.6 za příliš optimistický.

Sekce 5.2 se věnuje ChatGPT 4, ale jsou v ní prováděné zcela jiné experimenty, modely proti sobě tedy není možné porovnat. I zde je vyhodnocení řady experimentů nejasné (zejména pokud jde o interpretaci kódu). Při vyhodnocování schopnosti vysvětlit kód jsou v uvedeny příklady ve kterých je zdrojový text editován a není jasné, jestli jsou výsledky generovány z těchto příkladů nebo z celých zdrojových textů. Výsledky se mi nepodařilo při použití příkladů opakovat a kvalita vysvětlení kódu v přiložených příkladech je nesrovnatelně větší než v pokusech, které jsem udělal. Na druhou stranu, v příkladech nejsou ve zdrojových textech žádné komentáře, po doplnění komentářů se mé výsledky blíží tomu, co je v příkladech, ale zároveň je patrné že LLM z komentářů silně čerpá.

V sekci 5.3 a 5.4 je ukázaná řada příkladů, ale opět bez měřitelného vyhodnocení. Základem pro vyhodnocení zde jsou záznamy práce s použitým AI asistentem, popis výsledků ale obsahuje především obecné fráze na místech, kde bych čekal konkrétní detaily kde AI funguje dobře a kde ne. Tvrzení jako „TabNine excelled at providing sophisticated code completions that were well aligned with developer intent … “ nejsou nijak podložena hodnocením generovaného kódu. Ačkoliv jde v zásadě o podobné technologie, příklady zvolené pro 5.3 a 5.4 jsou opět odlišné a není tak možné porovnat modely mezi sebou.

Vyhodnocení v příkladech (sekce Review v souborech příloh) navíc často působí také jako generované texty, navíc obsahují zavádějící tvrzení (např je zmíněna „zkušenost z předchozích úloh“, což je v nejlepším případě nepřesnost).

Poslední částí práce je vlastní AI model pro potřeby Eurosoftware, založený na interních datech firmy, jestli chápu text správně jako nadstavba nad existujícími modely LLaMA a Megatron. Nakolik chápu text práce, data byla připravena ručně v podobě sady dvojic otázka – odpověď. Editor těchto dat je poměrně jednoduchá aplikace, která se příliš nezabývá uživatelskou použitelností (jde jen o editor dvojic, ale nezohledňuje např. že pro editaci odpovědi je potřeba víc místa než na otázku, neumožňuje žádné vyhledávání a data ani nejsou nijak řazena, takže např. hodnotit, jestli do dat nejsou vkládány duplicitní položky je téměř nemožné). Na s. 58 je zmíněno že k vyhodnocení je použita jiná metrika než přímé porovnání odpovědí s očekávaným řetězcem (což je určitě vhodné), ale není zmíněno, jak je to nakonec řešeno. V textu je také zmíněno že probíhalo manuální hodnocení výsledků, ale není jasné, jak bylo prováděno.

Dotazy k práci

Proč jste volil tak širokou škálu experimentů, místo opakování stejných pokusů s různými modely, aby bylo možné je porovnat?

Jak velká část textu je generovaná LLM?

Jak byla vyhodnocována kvalita vašeho vlastního modelu – podle čeho byly určovány správné odpovědi, co bylo základem pro výpočet F-míry?

Zadání považuji za splněné, s tím že řada vyhodnocení nepřináší konkrétní poznatky. Navrhuji hodnocení známkou **dobře** a práci doporučuji k obhajobě.

V Plzni 22.5.2024 Ing. Richard Lipka, Ph.D.