KATEDRA INFORMATIKY A VÝPOČETNÍ TECHNIKY

Bc. Milan Horínek

Generování jednotkových testů s využitím LLM

Diplomová práce

Softwarové a informační systémy
2023/2024

Vedoucí práce:
Ing. Richard Lipka, Ph.D

Abstrakt

Tato diplomová práce zkoumá použití velkých jazykových modelů (LLM) pro automatizované generování softwarových testů, konkrétně GUI webových aplikací. Zvolený přístup zkoumá, zda vygenerované testy jsou nejen sémanticky smyslupné, aby odpovídali funkčním požadavkům softwaru, ale také zda jsou schopné odhalit i vložené chyby do softwaru. Výzkum zahrnuje návrh a vývoj automatizovaného nástroje, který využívá zpracování přirozeného jazyka k interpretaci popisů testů a generování odpovídajících testových skriptů. Tento nástroj byl extenzivně vyhodnocen na vzorové webové aplikaci za účelem posouzení přesnosti, úplnosti a spolehlivosti jednotlivých výstupních testů. Výsledky naznačují, že nástroje založené na LLM dokáží efektivně generovat softwarové testy a úspěšně identifikovat vložené chyby do aplikace

Úvod

Cílem práce bylo seznámit se s existujícími technologiemi velkých jazykových modelů (LLM) a jejich aplikacemi, zejména v oblasti generování jednotkových testů a poté navrhnout a implementovat automatizovaný nástroj, který využívá LLM k tvorbě testů na základě přirozeného jazyka a dalších informací. Nakonec také zhodnotit kvalitu a možnosti současných LLM při generování těchto testů.

Východiska, analytická část

Základ informací, na kterých je práce postavena tvoří:

- **LLM Modely:** Nastudování současně dostupných LLM modelů a jejich vhodnosti pro tento projekt.
- **Generování testů:** Studium literatury o metodách a přístupech k automatizovanému generování softwarových testů s LLM (state of the art).
- **Existující nástroje:** Prozkoumání nástrojů pro detekci chyb a generování testů, vhodných pro tento výzkum.
- **Návrh a implementace nástroje:** Tvorba návrhu pro nástroj a jeho pipeline využívající LLM pro generování testových skriptů. Detailní popis kroků a použitých technologií pro automatizaci testů a jejich tvorby.

Ověření výsledků

Pro účely ověření schopnosti jednotlivých vygenerovaných testů detekovat jak správné tak poruchové chování softwaru, byl zvolen projekt TbUIS, který představuje zjendodušený univerzitní IS a poskytuje 28 poruchových klonů, díky kterým se dají testy spustit na variantě se správným chováním tak i s vloženou chybou. Zároveň obsahuje sadu předdefinovaných testových případů, které lze využít jako vstup pro tvorbu testů. Z nich bylo zvoleno 10 případů a 14 variant aplikace, dle kterých se generovaly ukázkové testy. Pro každý testový případ (scénář) bylo vygenrováno 10 variant testu.

Dosažené výsledky

Z dostupných LLM modelů bylo zvoleno 10 zástupců a to jak proprietárních tak volně dostupných, kdy každý model vygeneroval sadu testů pro veškeré scénáře a ve zvoleném počtu variant. Ty poté byli spuštěny na všech zvolených kombinacích scénář-varianta aplikace. Výsledky tohoto spuštění poté byli vyhodnoceny dle toho, jaký očekávaný výsledek v každém případě měl být (úspěch/selhání testu).

Nejlépe zde vycházejí proprietární modely jako *Claude 3*, *GPT-4* či *Gemini 1.5 Pro*, které dokázaly vygenerovat mezi **35** až **50%** plně validních testů, které byly shopny odhalit veškeré vložené chyby. Zároveň pouze jeden z použitých modelů byl schopen vygenerovat alespoň jeden platný test pro všechny z testových scénářů. Více jak polovina všech vygenerovaných testů nešla spustit nebo nebyla schopna detekovat korektní chování testovaného softwaru. Lokální LLM modely až na *CodeLlama* nebyly schopné vygenerovat větší množsví platných testů a objevovaly se v jejich výstupu nedostatky již na syntaktické úrovni, ne pouze na logické.

Závěr

Výsledky naznačují, že navrhovaný postup tvorby testů s využítím LLM na bázi popisu přirozeným jazykem je možný a lze skrze něj úspěšně generovat jednotkové testy. Samotná kvalita testů nemusí být pro všechny případy dostatečná a zásadně se liší mezi jednotlivými modely a pro případný navazující výzkum je doporučeno zaměřit se na jejich zkvalitnění.

Srovnání použitých LLM modelů dle zavedených metrik

Model	Úspěšnost pro scénáře	Celková úspěšnost případů	Validita	Úspěšnost z validních	Celková úspěšnost testů
Claude 3 Opus	**100,00%**	**43,71%**	47,00%	**95,74%**	**45,00%**
Gemini 1.5 Pro	90,00%	39,43%	**48,00%**	87,50%	42,00%
GPT-4	90,00%	36,71%	39,00%	92,31%	36,00%
GPT-4-Turbo	70,00%	37,29%	39,00%	61,54%	24,00%
Mistral Large	40,00%	17,79%	18,00%	83,33%	15,00%
GPT-3.5 Turbo	20,00%	3,00%	3,00%	100,00%	3,00%
CodeLlama	50,00%	7,07%	9,00%	22,22%	2,00%
Llama 3	10,00%	2,00%	2,00%	100,00%	2,00%
WizzardCoder	10,00%	0,86%	1,00%	0,00%	0,00%
Mistral 7B	0,00%	0,00%	0,00%	0,00%	0,00%

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI