



FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY



Diplomová práce

Generování jednotkových testů s využitím LLM

Milan Horínek





FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Diplomová práce

Generování jednotkových testů s využitím LLM

Bc. Milan Horínek

Vedoucí práce

Ing. Richard Lipka, Ph.D.

© Milan Horínek, 2024.

Všechna práva vyhrazena. Žádná část tohoto dokumentu nesmí být reprodukována ani rozšiřována jakoukoli formou, elektronicky či mechanicky, fotokopírováním, nahráváním nebo jiným způsobem, nebo uložena v systému pro ukládání a vyhledávání informací bez písemného souhlasu držitelů autorských práv.

Citace v seznamu literatury:

HORÍNEK, Milan. *Generování jednotkových testů s využitím LLM*. Plzeň, 2024. Diplomová práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky. Vedoucí práce Ing. Richard Lipka, Ph.D.

Podklad pro zadání DIPLOMOVÉ práce studenta

Jméno a příjmení: **Bc. Milan HORÍNEK**
Osobní číslo: **A23N0089P**
Adresa: **Česká 1024/6, Most, 43401 Most 1, Česká republika**
Téma práce: **Generování jednotkových testů s využitím LLM**
Téma práce anglicky: **LLM based unit test generator**
Jazyk práce: **Čeština**
Vedoucí práce: **Ing. Richard Lipka, Ph.D.**
Katedra informatiky a výpočetní techniky

Zásady pro vypracování:

1. Seznamte se s existujícími LLM, jejich technologií a možnostmi použití.
2. Prostudujte existující literaturu ohledně generování jednotkových testů, zejména s ohledem na využití neuronových sítí a LLM.
3. Seznamte se s existujícími ukázkovými programy s možností injekce chyb a vyberte vhodný testovací program.
4. Navrhněte a implementujte automatizovaný nástroj využívající popis testu v přirozené řeči, LLM a případně další informace, který vygeneruje sadu jednotkových testů.
5. Ověřte kvalitu automaticky vytvořených testů zejména s ohledem na přesnost a úplnost.
6. Zhodnoťte možnosti současných LLM pro generování testů.

Seznam doporučené literatury:

Dodá vedoucí práce.

Podpis studenta:

Datum:

Podpis vedoucího práce:

Datum:

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného akademického titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Západočeská univerzita v Plzni má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

Plzeň dne 14. května 2024

.....

Milan Horínek

V textu jsou použity názvy produktů, technologií, služeb, aplikací, společností apod., které mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

Abstrakt

TBD

Abstract

TBD

Klíčová slova

LLM • testing • unit • Robot Framework

Poděkování

TBD

Zkratky

GGML GPT-Generated Model Language. 21

GGUF GPT-Generated Unified Format. 20, 21

GUI Graphical User Interface. 11

ML Machine Learning - strojové učení. 21

PUT Parameterized Unit Test. 5

Slovník pojmů

filtr Ve vývoji softwaru a testování se odkazuje na mechanismy nebo kritéria používaná k selektivnímu výběru testovacích vstupů nebo k rozhodování, které výstupy testů jsou relevantní pro další analýzu. Filtry mohou být použity k odstranění redundantních, nelegálních nebo jinak nežádoucích vstupů z procesu generování testů, což zvyšuje efektivitu testování tím, že se zaměřuje pouze na vstupy, které mohou odhalit chyby nebo porušení kontraktů. 5

kontext Rozsah textu nebo dat, který může model zpracovat při generování odpovědí. Funguje jako krátkodobá paměť modelu a určuje, kolik informací může model vzít v úvahu při tvorbě odpovědi.. 21

kontrakt V kontextu jednotkových testů odkazuje na formálně definované podmínky nebo pravidla, které musí být dodrženy během vykonávání kódu. Kontrakty mohou specifikovat, co funkce očekává od svých vstupů a jaké výstupy nebo stavy by měla produkce kódu způsobit. Použití kontraktů pomáhá zajistit, že kód se chová podle očekávání a usnadňuje identifikaci chyb, když tyto podmínky nejsou splněny. 5

LLM Velký jazykový model (LLM - large language model) je typ modelu strojového učení, který je navržen tak, aby rozuměl a generoval text v přirozeném jazyce. Tyto modely jsou trénovány na obrovských datasetech a jsou schopny provádět různé úkoly, jako je textová klasifikace, generování textu, strojový překlad a další. 7–9, 20, 21, 23, 24

teplota V terminologii strojového učení se jedná o hyperparametr, který ovládá relativitu a náhodnost vstupu modelu. Vyšší teplota vede k méně předvídatelným a kreativnějším výsledkům, kdežto nižší teplota produkuje více konzervativnější výstup.. 21, 23

testové pachy Unit Test Smells jsou špatné návyky nebo postupy, které se mohou objevit při psaní jednotkových testů. Tyto "smelly" často vedou k neefektivním nebo nespolehlivým testům a mohou ztížit údržbu testovacího kódu. Příklady zahrnují Duplicated Asserts, Empty Tests a General Fixture. 8

token V NLP oblasti se jedná o rozdělení sekvence textu na menší části (například slova, části slov, atd.).. 21–23

Obsah

1	Úvod	3
1.1	Testy	3
1.2	Jazykové modely	3
1.3	Motivace	3
2	Rešerše	4
2.1	Provedená práce v problematice	4
2.1.1	Předchozí automatizovaná řešení	4
2.1.2	Vydané publikace	7
2.1.3	Modely	8
2.2	Testovací program	11
3	Návrh	13
3.1	Co a jak testujeme	13
3.2	Navrhované řešení	13
4	Generování testů	14
4.1	Výběr scénářů	14
4.2	Nahrávání scénářů	17
4.3	Výběr požadavků	18
4.3.1	Vytvoření nového testu	18
4.3.2	Vyplnění požadavků	20
4.4	Dotazování LLM	20
4.4.1	Modely a jejich spuštění	20
4.4.2	Dotazy	23
5	Spuštění testů	28
5.1	Spuštění testovaného programu a jeho orchestrace	28
6	Vyhodnocení výsledků	29

7	Budoucí vylepšení	30
8	Závěr	31
	Seznam obrázků	32
	Seznam tabulek	33
	Seznam výpisů	34

Úvod

1

TDB - Co jsou unit testy, LLM, a jejich předpoklady, GUI testování

Něco na úvod

1.1 Testy

1.2 Jazykové modely

1.3 Motivace

2.1 Provedená práce v problematice

2.1.1 Předchozí automatizovaná řešení

Jazykové modely nebyli prvními pokusy o automatizované generování jednotkových testů. Ještě před nimi existovala spousta metod zahrnující příklady jako *fuzzing*, *generování náhodných testů řízených zpětnou vazbou*, *dynamické symbolické exekuce*, *vyhledávací a evoluční techniky*, *parametrické testování*. Zároveň také již na počátku století byli pokusy o vytvoření vlastní neuronové sítě sloužící právě čistě k úkolu testování softwaru. V této sekci je ukázka několika z nich.

2.1.1.1 Programatická řešení

Jedna z používaných programatických automatizovaných metod pro tvorbu jednotkových testů je tzv. *fuzzing*. V rámci těchto testů musí uživatel stále definovat jeho kódovou strukturu, resp. akce, které test bude provádět a jaký výstup očekávat. Automaticky generovaný je pouze vstup tohoto testu. Výhodou zde tedy je, že uživatel nemusí vytvářet maketu vstupních dat testu, která se zde vytvoří automatizovaně. Zůstává zde však problematika, že pro uživatele není kód *black-box*, ale celou jeho strukturu včetně požadovaného výstupu musí sám definovat. [**fuzzing**]

Pouze vstupy dokáže také generovat metoda *symbolické exekuce*, která postupně analyzuje chování větvení programu. Začíná bez předchozích znalostí a používá řešitel omezení k nalezení vstupů, které prozkoumají nové exekuční cesty. Jakmile jsou testy spuštěny s těmito vstupy, nástroj sleduje cestu, kterou se program ubírá, a aktualizuje svou znalostní bázi (q) s novými podmínkami cesty (p). Tento itera-

tivní proces se opakuje a nadále zpřesňuje sadu známých chování a snaží se maximalizovat pokrytí kódu. Nástroje běžně zvládají různé datové typy a respektují pravidla viditelnosti objektů. Snaží se také používat mock objekty a parametrizované makety k simulaci různých chování vstupů, čímž zlepšuje proces generování testů, aby odhalila potenciální chyby a zajistila komplexní pokrytí kódu testy. **[parizek_symbolic_execution]** Tato metoda je implementována například v nástroji *IntelliTest* v rámci IDE *Visual Studio*. Je používána v kombinaci s *parametrickými testy*, také označovanými jako PUT. Ty na rozdíl od tradičních jednotkových testů, které jsou obvykle uzavřené metody, mohou přijímat libovolnou sadu parametrů. Nástroje se pak snaží automaticky generovat (minimální) sadu vstupů, které plně pokryjí kód dosažitelný z testu. Nástroje jako např. *IntelliTest* automaticky generují vstupy pro put, které pokrývají mnoho exekučních cest testovaného kódu. Každý vstup, který pokrývá jinou cestu, je "serializován" jako jednotkový test. Parametrické testy mohou být také generické metody, v tom případě musí uživatel specifikovat typy použité k instanci metody. Testy také mohou obsahovat atributy pro očekávané a neočekávané vyjímky. Neočekávané vyjímky vedou k selhání testu. put tedy do velké míry redukuje potřebu uživatelského vstupu pro tvorbu jednotkových testů. **[IntelliTestInputGeneration2023]** **[microsoft2023testgen]**

Pokud zvolíme symbolické řešení vstupu společně s determinovanými vstupy a testovací cestou, vzniká tak hybridní řešení zvané jako *konkolické testování* nebo *dynamická symbolická exekuce*. Tento druh testů dokáží tvořit nástroje jako *SAGE*, *KLEE* nebo *SzE*. Problémem tohoto přístupu však je, když program vykazuje *ne-deterministické* chování, kdy tyto metody nebudou schopny určit správnou cestu a zároveň tak ani zaručit dobré pokrytí kódu/větví. Velká míra používání stavových proměnných může vést k vysoké výpočetní náročnosti těchto metod a nenalezení praktického řešení. **[engler2006exe]** **[sen2005cute]** **[zhou2006safedrive]**

Další metodou je *náhodné generování testů řízené zpětnou vazbou*, která je vylepšením pro generování náhodných testů tím, že zahrnuje zpětnou vazbu získanou z provádění testovacích vstupů v průběhu jejich vytváření. Tato technika postupně buduje vstupy tím, že náhodně vybírá metodu volání a hledá argumenty mezi dříve vytvořenými vstupy. Jakmile dojde k sestavení vstupu, je provedena jeho exekuce a výsledek ověřen proti sadě kontraktů a filtry. Výsledek exekuce určuje, zda je vstup redundantní, proti pravidlům, porušující kontrakt nebo užitečný pro generování dalších vstupů. Technika vytváří sadu testů, které se skládají z jednotkových testů pro testované třídy. Úspěšné testy mohou být použity k zajištění faktu, že kódu jsou zachovány napříč změnami programu; selhávající testy (porušující jeden nebo více kontraktů) ukazují na potenciální chyby, které by měly být opraveny. Tato metoda dokáže vytvořit nejen vstup pro test, ale i tělo (kód) testu. Ovšem pro uživatele je

stále vhodné znát strukturu kódu. [FeedbackDirectedRT]

Z programatických metod se zdají být nejpokročilejší *evoluční algoritmy* pro generování sad jednotkových testů, využívající přístup založený na vhodnosti, aby vyvíjely testovací případy, které mají za cíl maximalizovat pokrytí kódu a detekci chyb. Tyto algoritmy mohou autonomně generovat testovací vstupy, které jsou navrženy tak, aby prozkoumávaly různé exekuční cesty v aplikaci. Uživatelé mohou interagovat s vygenerovanými testy jakožto s *black-boxem*. Testy se zaměřují na vstupy a výstupy, aniž by potřebovali rozumět vnitřní logice testovaného systému. Tento aspekt evolučního testování je zvláště výhodný při práci se složitými systémy nebo když zdrojový kód není snadno dostupný. Proces iterativně upravuje testovací případy na základě pozorovaných chování, upravuje vstupy pro efektivnější prozkoumání systému a identifikaci potenciálních defektů. Tato metoda podporuje vysokou úroveň automatizace při generování testů, snižuje potřebu manuálního vstupu a umožňuje komplexní pokrytí testů s menším úsilím. [CAMPOS2018207] [abs-2111-05003]

2.1.1.2 Neuronové sítě

Ke generování jednotkových testů lze využít i vlastní neuronové sítě. Takové se pokoušeli vytvářet například v práci "Unit test generation using machine learning"[Saes2018UnitTestGenera kde byly testovány primárně RNN a experimentálně CNN sítě (ty však měli problém s větším množstvím tokenů). Modely byli testovány na jazyce Java. Metoda přistupovala k programům jakožto white box, tedy měli k dispozici celý zdrojový kód včetně zkompilevaného bytecodu. Při nejlepší konfiguraci dosáhl výsledek modelu 70.5% parsovatelného kódu (tedy takového bez chyb) natrénovaný z bezmála 10000 příkladů *zdrojový kód - test*. Výsledek práce je však stále jakýsi "proof of concept", protože zatímco vygenerují částečně použitelný výsledek, je vždy nutný zásah experta, aby mohlo dojít k vytvoření celého testovacího souboru. Takovéto sítě se však mohou silně hodit jako výpomoc programátorovi při psaní testů.

2.1.1.3 Nevýhody současných metod

Současné metody generování jednotkových testů, jako je *fuzzing* a *symbolická exekuce*, často vyžadují podrobnou znalost struktury kódu a očekávaných výstupů, což omezuje jejich efektivitu a zvyšuje složitost tvorby testů. Jen některé z těchto metod jsou schopné vygenerovat testy pouze na bázi specifikace (z black box pohledu) bez vnitřní znalosti kódu. Většina z klasických metod je zároveň schopna generovat

pouze vstupy jednotkových testů, ale už ne samotné tělo (kód) testu nebo očekávané výstupy, a tedy pouze složí jako jakási konstra pro programátora, který musí test doimplementovat.

Velké jazykové modely (LLM) mohou být atraktivní alternativou, protože pomocí nich lze potenciálně automatizovat generování jak vstupů pro testy, tak přidruženého testovacího kódu, čímž se snižuje potřeba hlubokého porozumění struktuře kódu. S takovým nástrojem není potřeba programátora, ale může s ním pracovat i méně zkušený uživatel (např. *tester*). Dále je zde možnost otestovat kód za pomoci slovní specifikace pro funkci nebo vlastnosti. Takové specifikace se používají například v *aero-space* nebo *automotive* sektoru. LLM také mohou objevit všechny možné stavy, které mohou u vstupu nebo výstupu nastat a pokusit se pro ně navrhnout test.

2.1.2 Vydané publikace

Jeden z poměrně nedávno vydaných článků (září 2023) nazvaný "An Empirical Evaluation of Using Large Language Models for Automated Unit Test Generation"[[schafer2023empirical](#)] se zabývá využitím velkých jazykových modelů (LLM) pro automatizované generování jednotkových testů v jazyce JavaScript. Implementovali nástroj s názvem **TEST-PILOT**, který využívá LLM *gpt3.5-turbo*, *code-cushman-002* od společnosti OpenAI a také model StarCoder, který vznikl jako komunitní projekt [[StarCoder2023](#)]. Vstupní sada pro LLM obsahovala signatury funkcí, komentáře k dokumentaci a příklady použití. Nástroj byl vyhodnocen na 25 balíčcích npm obsahujících celkem 1684 funkcí API. Vygenerované testy dosáhly pomocí *gpt3.5-turbo* mediánu pokrytí příkazů 70,2% a pokrytí větví 52,8%, čímž překonaly nejmodernější techniku generování testů v jazyce JavaScript zaměřenou na zpětnou vazbu, Nessie.

Zmíněný model *StarCoder* byl představen v článku "StarCoder: may the source be with you!"[[StarCoder2023](#)] z května 2023. Vytvořeny byly konkrétně 2 verze, *StarCoder* a *StarCoderBase*, s 15,5 miliardami parametrů a délkou kontextu 8K. Tyto modely jsou natrénovány na datové sadě nazvané *The Stack*, která obsahuje 1 bilion tokenů z permissivně licencovaných repozitářů GitHub. *StarCoderBase* je vícejazyčný model, který překonává ostatní modely open-source LLM modely, zatímco *StarCoder* je vyladěná verze speciálně pro Python, která se vyrovná nebo překoná stávající modely zaměřené čistě na Python. Článek poskytuje komplexní hodnocení, které ukazuje, že tyto modely jsou vysoce efektivní v různých úlohách souvisejících s kódem.

Článek "Exploring the Effectiveness of Large Language Models in Generating

Unit Tests"[siddiq2023exploring] z dubna 2023 hodnotí výkonnost tří LLM - *Codex*, *CodeGen* a *GPT-3.5* - při generování jednotkových testů pro třídy jazyka Java. Studie používá jako vstupní sady dva benchmarky, *HumanEval* a *EvoSuite SF110*. Klíčová zjištění ukazují, že *Codex* dosáhl více než 80% pokrytí v datové sadě *HumanEval*, ale žádný z modelů nedosáhl více než 2% pokrytí v benchmarku *SF110*. Kromě toho se ve vygenerovaných testech často objevovaly tzv. testové pachy, jako jsou *duplicitní tvrzení* a *prázdné testy* [testsmells].

2.1.2.1 Srovnání výsledků

Výsledky diskutovaných studií v předchozím bodě jsme srovnali v tabulce 2.1. V rámci první práce dosahuje nejlepších výsledků model *gpt-3.5-turbo*, který dosáhl 70% pokrytí kódu testy a 48% úspěšnosti testů. U druhé studie má tento model na testovací sadě *HumanEval* velice podobný výsledek, ovšem model *Codex* dosáhl lepších výsledků. Může však také jít o rozdíl způsobený programovacím jazykem. Zatímco v práci [schafer2023empirical] se využívá jako benchmark sada balíčků jazyka *JavaScript*, který kvůli absenci explicitního typování, může být obtížnější pro strojové testování oproti jazyku Java, který je využit ve zbylých 2 pracích. [jutai] také využívá Javu a s modelem *gpt-3.5-turbo* dosahuje podobného pokrytí kódu a úspěšnosti jako *Codex* v práci [siddiq2023exploring].

2.1.3 Modely

Na LLM modelech nás konkrétně zajímá schopnost pozorovat programovací jazykům a ty poté také generovat na výstupu. Důležité pro nás také je, zda daný model je proprietární či otevřený a pod jakou licenci, tedy zda by byl vhodný pro naši práci. V případě analýzy zdrojového kódu může být také klíčovou vlastností délka kontextu daného modelu. Tyto vlastnosti jsou také zaneseny do tabulky 2.2.

StarCoder. Jedním z často používaných modelů v předchozích pracích je *StarCoder* a *StarCoderBase*, diskutovaný již v sekci 2.1. *Base* verze je schopna generovat kód pro více jak 80 programovacích jazyků. Model je navržen pro širokou škálu aplikací obsahující generování, modifikaci, doplňování a vysvětlování kódu. Jeho distribuce je volná a licence **CodeML OpenRAIL-M 0.1** [BigCode2023] umožňuje ho využívat pro množství aplikací včetně komerčních nebo edukačních. Jeho uživatel však má povinnost uvádět, že výsledný kód byl vygenerován modelem. Licence má své

restrikce z obavy tvůrců, protože by model mohl někoho při nesprávném použití ohrozit. Tyto restriktce se aplikují na všechny derivace projektů pod touto licencí. Zároveň není kompatibilní s Open-Source licencí právě kvůli těmto restrikcím.

Model StarCoder se také dočkal novější verze *StarCoder2*, který nepřímo navazuje na model *StarCoderBase*. Je naučen na archivu GitHub repoziářů archivovaných v rámci organizace *Software Heritage*, čítajících přes 600 programovacích jazyků a dalších pečlivě vybraných dat jako například *pull requesty*. Mimo toho také trénovací data obsahují staženou dokumentaci k vybraným projektům. Model se vyjímá tím, že se snaží udržet malou velikost. Je nabízen ve verzích s 3, 7 a 15 miliardy parametrů, i přesto však dle jejich úvodní studie [lozhkov2024starcoder] nabízí na sadě populárních programovacích jazyků shodné či lepší výsledky oproti modelu *CodeLlama* s 34 miliardy parametrů. Výhodou tohoto modelu nepochybně je, že ho lze spustit na spoustě dnešních počítačů s plným offloaded na grafickou kartu.

CodeLlama. Nedávno vydaným modelem je *Code Llama* od společnosti Meta. Jedná se o evoluci jejich jazykového modelu *Llama* specializovaný však čistě na úlohy kódování. Je postaven na platformě *Llama 2* a existuje ve třech variantách: *základní Code Llama*, *Code Llama - Python* a *Code Llama - Instruct*. Model podporuje více programovacích jazyků, včetně jazyků jako Python, C++, Java, PHP, Typescript, C nebo Bash. Je určen pro úlohy, jako je generování kódu, doplňování kódu a ladění. *Code Llama* je zdarma pro výzkumné i komerční použití a je uvolněn pod licencí MIT. Uživatelé však musí dodržovat zásady přijatelného použití, ve kterých je uvedeno, že model nelze použít k vytvoření služby, která by konkurovala vlastním službám společnosti Meta.

Copilot. Velmi populárním nástrojem pro generování kódu za pomoci LLM je GitHub Copilot, který je postaven na modelu *codex* od OpenAI. Původní model však byl přestal být zákazníkům nabízen a namísto něj OpenAI doporučuje ke generování kódu využívat chat verze modelů GPT-3.5 a GPT-4. Na architektuře GPT-4 je také postavený nástupce služby Copilot, Copilot X. Zmíněné modely chat GPT-3.5 a GPT-4 jsou primárně určeny pro generování textu formou chatu. Zvládají však zároveň i dobře generovat kód a jsou vhodné i úlohu generování jednotkových testů. Narozdíl od předchozích modelů však nejsou volně distribuovány a jsou poskytovány pouze jako služba společností OpenAI skrze API nebo je lze hostovat v rámci služby Azure společnosti Microsoft, která zajišťuje větší integritu dat. Jedná se tedy o uzavřený model a jeho uživatelé musí souhlasit s jeho podmínkami použití.

Práce	Model	Benchmark	Pokrytí testy	Úspěšnost
An Empirical Evaluation of Using Large Language Models for Automated Unit Test Generation	<i>gpt-3.5-turbo</i> <i>code-cushman-002</i> <i>StarCoder</i>	Sada NPM balíčků	70.2% 68.2% 54%	48% 47.1% 31.5%
Exploring the Effectiveness of Large Language Models in Generating Unit Tests	<i>gpt-3.5-turbo</i> <i>CodeGen</i> <i>Codex (4k)</i>	HumanEval SF110 HumanEval SF110 HumanEval SF110	69.1% 0.1% 58.2% 0.5% 87.7% 1.8%	52.3% 6.9% 23.9% 30.2% 76.7% 41.1%
Java Unit Testing with AI: An AI-Driven Prototype for Unit Test Generation	<i>gpt-3.5-turbo</i>	JUTAI - Zero-shot, temperature: 0	84.7%	71%

Tabulka 2.1: Přehled a srovnání studií

Model	Otevřenost	Licence	Počet parametrů	Počet programovacích jazyků	Datum vydání
<i>StarCoderBase</i>	Volně dostupný	CodeML OpenRAIL-M 0.1	15,5 miliardy	80+	5/2023
<i>Code Llama</i>	Volně dostupný	Llama 2 Licence	34 miliard	8+	8/2023
<i>gpt-3.5-turbo</i>	Uzavřený	Proprietární	175 miliard?	?	5/2023
<i>gpt-4</i>	Uzavřený	Proprietární	1.76 bilionu	?	3/2023

Tabulka 2.2: Přehled a srovnání modelů generujících kód

2.2 Testovací program

Pro účely této práce jsme se rozhodli vydat cestou GUI testů, konkrétně webových stránek / webové aplikace. Mezi požadavky na testovanou aplikaci a její výběr bylo mimo možnosti vytvořit pro ní automaticky sadu testů, také možnost zavedení (*injekce*) chyb do aplikace, díky kterým bude možno ověřit nejen fakt, že vygenerované testy jsou schopné detekovat *korektní* chování softwaru, ale také úspěšně detekovat *chyby*. Volitelnou podmínkou také byla existence již existujících testů vytvořených lidským programátorem, se kterými by bylo možné strojově generované testy porovnat.

Těmto požadavkům vyhovuje jeden z předešlých univerzitních projektů nazvaný **TbUIS** (*Testbed University Information System*), na který vytvořili Matyáš a Šmaus pod vedením doc. Herouta. [Matyas2018] [Smaus2019] Aplikace představuje *marketu* univerzitního informačního systému, avšak i tak implementuje většinu funkcionality, kterou bysme od podobného systému čekali a nabízí pohled jak ze strany *studenta* tak ze strany *vyučujícího* (viz obr. 2.1). Funkce tohoto systému vycházejí ze sady *use casů*, které popisuje web ¹ aplikace. Výhodou tohoto systému primárně je, že nabízí nejen plně funkční a otestovanou variantu, ale také 27 dalších poruchových klonů, vždy porušující alespoň jeden *use case* a s ním potenciálně spojený test.

Projekt dále také obsahuje i sadu jak *funkčních* testů primárně pro backend vytvořených Poubovou tak také *akceptačních* testů držejících se pevně specifikací, vytvořených Vaisem. [Poubova2019] [Vais2020] Tyto testy využívají nejen pevné znalosti jednotlivých uživatelských scénářů, ale také jsou parametrizovány pomocí všech dostupných dat k aplikaci (*např. uživatelská jména, předměty, zkoušky, atd.*) k vyhodnocení chování softwaru v co nejvíce případech. Právě *akceptační* testy využívají námi zvolený **RobotFramework** pro testování a tedy vytvořené scénáře bude možné do určité míry přirovnat a případně zhodnotit jejich kvalitu.

¹Aplikaci lze nalézt na adrese: <https://projects.kiv.zcu.cz/tbuis/web/>

University Information System Home Mia Orange Logout

Student's View

- Overview
- My Subjects
- Other Subjects
- My Exam Dates
- Other Exam Dates

First name:

Last name:

Email:

Obrázek 2.1: Prostředí systému TbUIS z pohledu studenta.

Návrh

3

3.1 Co a jak testujeme

3.2 Navrhované řešení

Generování testů

4

4.1 Výběr scénářů

Pro vygenerování testů za pomoci LLM bylo nejdříve nutné vybrat ze sady *use case*¹ scénáře vhodné pro demonstraci nejen správné funkčnosti, ale také s možností ověření nesprávného chování na poruchových klonech (diskutováno v sekci 2.2). Požadavek tedy byl, aby většina z vybraných scénářů vycházející z *use casu* měl alespoň jeden poruchový klon, případně více. Vybráno bylo 10 scénářů, testovaných celkově na 14 variantách programu. Seznam vytvořených specifikací pro automatické generování testů se nachází v tabulkách 4.1 a 4.1. Scénáře budou dále také referovány jako *specifikace*. Každá specifikace má číslo, které odpovídá číslu *use casu*, ze kterého vychází (např. specifikace 04 vychází z *use casu UC.04*). Zde je nutné poznamenat, že ne všechny specifikace vychází pevně z jejich *use casů*, ale byli upravené tak, aby šli provést v jednom chodu bez nutných závislostí (jako například namísto podmínky přihlášeného uživatele se uživatel vždy musí na začátku či během testu přihlásit do systému). Popis specifikací v tabulce je pouze orientační a pro bližší upřesnění jednotlivých kroků, které se mají provést referujte web projektu. V tomto seznamu se také u každé specifikace nachází výčet klonů, na kterých lze očekávat poruchu. Celkový seznam klonů, a kterých se budou všechny vytvořené testy spouštět lze nalézt v tabulce 4.1. Podobně jako v případě specifikací, čísla klonů odpovídají číslům, jak jsou uvedena na oficiálním webu² projektu společně s vysvětlením jednotlivých chyb varianty. Pro přehlednost jsme bezchybnou variantu označili číslem 00.

¹Seznam dostupný na adrese: <https://projects.kiv.zcu.cz/tbuis/web/page/uis#use-cases>

²Seznam poruchových klonů společně s možností stažení: <https://projects.kiv.zcu.cz/tbuis/web/page/download>

Specifikace	Popis	Porucha na klonech
1	<i>Přihlášení do aplikace</i> Student i učitel se přihlásí do aplikace přihlašovacími údaji, dostupnými v databázi systému. Dále se zkontroluje, zda systém pro účet s neexistujícím uživatelským jménem nebo neplatným heslem vypíše chybovou hlášku.	02
4	<i>Odepsání předmětu</i> Student se přihlásí, odepíše předmět v patřičné sekci systému. Předmět by následně měl zmizet v ostatních sekcích a to jak v pohledu studenta tak učitele.	04, 19, 25, 26, 28
6	<i>Zapsání předmětu</i> Student se přihlásí, zapíše předmět v patřičné sekci, který se následně zobrazí i v ostatních částech systému. Z učitelského pohledu by se student měl zobrazit na seznamu studentů daného předmětu.	25, 26, 28
8	<i>Registrace na zkoušku</i> Student se přihlásí do systémů a zapíše se na jeden z možných zkouškových termínů. Tento termín by se měl přesunout mezi již zapsané termíny. V učitelském pohledu bude student na seznamu zapsaných na konkrétní zkoušku a také by mělo být možno studenta ohodnotit.	22, 25, 26, 28
9	<i>Zobrazení spolužáků u zkoušky</i> Student se přihlásí a u zkoušky si může rozkliknout seznam všech účastníků.	
10	<i>Zrušení předmětu</i> Učitel po přihlášení klikne na tlačítko <i>Remove</i> u předmětu, od jehož výuky se chce odhlásit. Ten se přestane zobrazovat ve všech ostatních sekcích systému z jeho pohledu, až na jeho znovuzapsání. Student by u předmětu neměl najít jméno daného učitele, který se odhlásil.	26, 28
11	<i>Zobrazení studentů u předmětu</i> Učitel se přihlásí do systému a u předmětu je schopen si zobrazit seznam zapsaných studentů.	26, 28

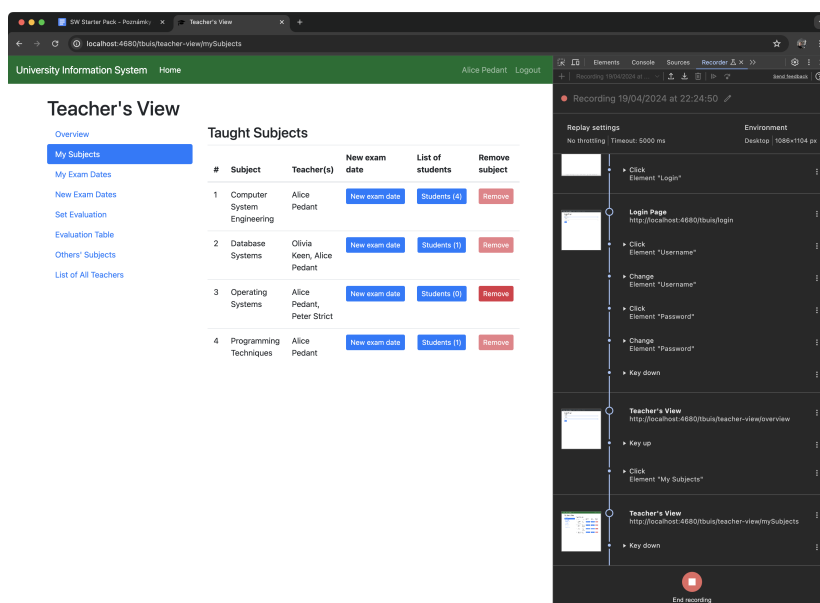
Tabulka 4.1: Specifikace pro generované testy - část 1

Specifikace	Popis	Porucha na klonech
12	<i>Zrušení zkoušky</i> Učitel se přihlásí a v sekci jeho přidělených zkoušek odstraní konkrétní termín. Tento termín by nyní neměl být vidět jak v učitel-ském tak studentském pohledu do systému.	20, 21, 23, 26, 28
17	<i>Přihlášení se k výuce předmětu</i> Učitel se přihlásí do systému a v seznamu předmětů se přihlásí k výuce daného předmětu. Ten by se poté měl zobrazit ve zbytku systému v rámci patřičných sekcí. Zároveň studenti by nyní měli u předmětu vidět jméno tohoto vyučujícího.	18, 24, 25, 26, 27, 28
18	<i>Zobrazení seznamu učitelů a předmětů, které vyučují</i> Přihlášený učitel je schopen si zobrazit seznam všech učitelů.	27, 28

Tabulka 4.2: Specifikace pro generované testy - část 2

Číslo klonu	Porucha
00	Bez defektu
02	Překlep v nadpisu
04	Návrat na špatnou stránkau
18	Chybějící sloupec v tabulce
19	Náhodně chybějící tlačítko
20	Nefunkční tlačítko
21	Změna se nepropíše do UI
22	Nefunkční tlačítko
23	Smazání se nepropíše do DB
24	Přidání se nepropíše do DB
25	Tabulka studentů prázdná
26	Tabulka učitelů prázdná
27	Nesprávný výběr z DB
28	Mix chyb (včetně interní chyby systému)

Tabulka 4.3: Seznam poruchových klonů využitých pro testování.



Obrázek 4.1: Nahrávání scénáře za pomoci nástroje v Google Chrome

4.2 Nahrávání scénářů

Dle *specifikace* z předchozího bodu, vytvoří uživatel LLM generačního nástroje nahrávku jednotlivých kroků, které má test provést. V případě, že specifikace udává více uživatelských pohledů (*student / učitel*), nahraje uživatel tyto pohledy jednotlivě. Současný projekt pracuje s nahrávacím nástrojem v rámci vývojářských nástrojů webového prohlížeče *Google Chrome* (pro vývoj byla využita verze 124). Tento nástroj je stále experimentální funkce, takže nelze vyloučit možnost, že v následujících verzích již nemusí být dostupný. Díky této funkci lze nahrávat uživatelské vstupy a interakce s jednotlivými prvky v rámci webových stránek. Mimo jiné také umožňuje přidávat *asserty*, avšak tyto možnosti jsou nedostatečné a tedy v rámci této práce se omezíme pouze na údaje o interakci s prvky. Nahrávání zobrazeno na obr. 4.1. Nástroj je schopen vyexportovat výstup nahrávání v řadě formátů. Zde jsme zvolily JSON formát, který popisuje jednotlivé akce dle objektů. Nahrané scénáře uživatel vhodně pojmenuje a uloží do složky input programu, který je součástí tohoto projektu. Vytvořený program pro generování testů za pomoci LLM se stará o veškerou orchestraci generování i spouštění testů a také vytovření *šablon* pro testy, ze kterých se generují. Součástí těchto *šablon* je právě i uživatelská nahrávka. Ukázkovou strukturu input složky lze vidět ve výpisu 4.1, kde se nachází *šablona* pro *specifikace* 1 a 4. Význam *šablon* je diskutován v následující sekci.

Výpis 4.1: Ukázková struktura input složky

```

1 milan:dp\program\input$ ls -l
2 milan      448 Apr 21 20:52  ..
3 milan      3815 Mar 27 20:24  rec-spec-1-student.json
4 milan      3820 Mar 27 20:28  rec-spec-1-teacher.json
5 milan      6043 Mar 30 19:42  recording-spec-4-student.json
6 milan     10635 Mar 31 09:18  recording-spec-4-teacher.json
7 milan      1370 Apr 21 20:52  spec-1.txt
8 milan      1328 Mar 31 09:28  spec-4.txt
9

```

4.3 Výběr požadavků

4.3.1 Vytvoření nového testu

Primárním programem celé práce je soubor `test.py`, který se nalézá v kořenové složce programu (`/program`) a má konkrétně 3 pracovní režimy:

1. Vytvoření šablony pro test
2. Vygenerování testu dle šablony
3. Spuštění sady testů

Právě první režim je v současném kroce důležitý. Jednotlivé režimy programu jsou spuštěny za pomoci argumentů (argumenty jednotlivých režimů ukázané ve výpisu 4.2). V případě *vytvoření nového testu* se za argument přidává název šablony testu. Tento název musí být unikátní a využívá se později ve zbytku programu jakožto identifikátor daného testu. Po potvrzení příkazu se ve vstupní složce vytvoří nová šablona pro test se zvoleným názvem a příponou `.txt`. Tuto šablonu může uživatel může dále upravit. Celý příkaz je ukázaný ve výpisu 4.3. Vytvořenou šablonu uživatel zobrazí a upraví v *textovém editoru*.

Výpis 4.2: Hlavní režimy programu

```

1 milan:dp\program\input$ python3 test.py -n #Nový test
2 milan:dp\program\input$ python3 test.py -i #Generování
3 milan:dp\program\input$ python3 test.py -r #Spuštění testů
4

```

Výpis 4.3: Vytvoření nového testu (šablony)

```
1 milan:dp\program\input$ python3 test.py -n specification-1
```

```
2
```

Zdrojový kód 4.4: Vzor pro vyplnění šablony testu

```
1 Write Robot Framework scanario. Open page like in this JSON
  recording and then when you execute all the steps in the
  recording, do this:
```

```
3 - //TODO
```

```
5 {% include 'recording.json' %}
```

```
6
```

Zdrojový kód 4.5: Vyplněná šablona testu pro specifikaci 18

```
1 Write Robot Framework scanario. Open page like in this JSON
  recording and then when you execute all the steps in the
  recording, do this:
```

```
3 - Check if there are these names present on the page: Julia
  Easyrider, Olivia Keen, John Lazy, Alice Pedant, Thomas
  Scatterbrained, Peter Strict
```

```
4 - Check if element with path /*[@id="tea.listOfAllTeachers.
  table.teacherRow-0"]/td[3] has text matching "Numerical
  Methods"
```

```
5 - Check if element with path /*[@id="tea.listOfAllTeachers.
  table.teacherRow-1"]/td[3] has text matching "Database
  Systems, Fundamentals of Computer Networks, Introduction
  to Algorithms, Mobile Applications, Web Programming"
```

```
6 - Check if element with path /*[@id="tea.listOfAllTeachers.
  table.teacherRow-2"]/td[3] should not contain text
```

```
7 - Check if element with path /*[@id="tea.listOfAllTeachers.
  table.teacherRow-3"]/td[3] has text matching "Computer
  System Engineering, Database Systems, Operating Systems,
  Programming Techniques"
```

```
8 - Check if element with path /*[@id="tea.listOfAllTeachers.
  table.teacherRow-4"]/td[3] has text matching "Computation
  Structures"
```

```
9 - Check if element with path /*[@id="tea.listOfAllTeachers.
  table.teacherRow-5"]/td[3] has text matching "Operating
  Systems, Programming in Java, Software Engineering,
  Software Quality Assurance"
```

```
11 {% include 'recording-spec-18.json' %}
```

```
12
```

4.3.2 Vyplnění požadavků

Do šablony je vložen základní prompt společně s požadavky, které uživatel může vyplnit (viz ukázka 4.4). Pod požadavky je defaultně vložena nahrávka `recording.json`. Tuto hodnotu nahradí uživatel za název souboru vložené nahrávky. Formát šablony jakožto vstupu pro prompt testu byl zvolen pro jednodušší parametrizaci. Šablona využívá jazyk *Jinja2*. V rámci vytvořených šablon v této práci byly tyto parametrizační vlastnosti využity například pro vytvoření pohledu *studenta* a *učitele*. Místo, které je v šabloně označeno jako `\\TODO` slouží pro napsání požadavků testu. Od uživatele se čeká, že tyto požadavky vypíše v odrážkách, čemuž odpovídá i formát *promptu*. Ukázka vyplněných požadavků pro konkrétní test je součástí výpisu 4.5, který ukazuje vypsání vlastností pro specifikaci 18, jak bylo popsáno v sekci 4.1.

4.4 Dotazování LLM

4.4.1 Modely a jejich spuštění

Jak již bylo řečeno v kapitole 2, velké jazykové modely existují jak v proprietární formě, které jsou dostupné pouze přes API poskytovatele nebo webové rozhraní, ale také se dají najít modely dostupné v *public* či *open source* formě dále referovány jako *otevřené* modely). V rámci tohoto projektu byly využity obě varianty, tedy *proprietární* modely srkze API poskytovatele a *otevřené* lokálně nebo také srze API u některého z poskytovatelů a to z důvodu vysokých hardwarových nároků některých modelů. Konkrétní seznam použitých modelů v práci včetně typu přístupu k nim lze nalézt v tabulce 4.4. Pro dotazování vzdálených modelů využívá většina koncových bodů API od OpenAI, tedy jsou kompatibilní s jejich knihovnou pro různé jazyky. Některé společnosti však pro své modely využívají vlastní definici API, mezi ně se řadí například *Anthropic* nebo *Google*. V projektu tedy využíváme pro veškeré kompatibilní modely (resp. jejich běhová prostředí) knihovnu pro OpenAI API, se kterým jsou kompatibilní, a pro zbytek jejich vlastní knihovny.

4.4.1.1 Prostředí pro lokální modely

Pro *lokální* modely a jejich spuštění byl využit software *LM Studio*, který je postaven nad *C++* knihovnou *llama.cpp*. Ta umožňuje jednoduché spuštění LLM modelů ve formátu GPT-Generated Unified Format (GGUF). Jedná se o *binární* formát souborů určený pro rychlé načítání a ukládání modelů využívaných pro účely

Tvůrce	Model	Použitá verze	Runtime
OpenAI	GPT-4	gpt-4-32k	API
	GPT-4 Turbo	gpt-4-turbo-2024-04-09	
	GPT-3 Turbo	gpt-3-turbo-0125	
Mistral	Mistral vo.2 7B	TBD	Lokální
	Mistral Large	mistral-large-2402	API <i>La Plateforme</i>
Anthropic	Claude 3 Opus	claude-3-opus-20240229	API
Meta	Codellama	TBD	Lokální
Google	Gemini 1.5 Pro	gemini-1.5-pro-preview-0409	API <i>Google Cloud</i>

Tabulka 4.4: Použité LLM modely

interferenčních úloh. Jeho návrh umožňuje snadnou úpravu při zachování zpětné kompatibility. [ggerganov_gguf] [huggingface_gguf] Modely mohou být vyvíjené za pomoci různých frameworků (např. *PyTorch*) a poté převedeny do GGUF formátu pro použití v rámci GGML knihovny, která umožňuje efektivní spuštění modelů na CPU a GPU. Knihovna také umožňuje *kvantizaci*, tedy techniku využívanou ke snížení paměťových nároků ML úloh. Zahrnuje reprezentaci vah a aktivací za pomoci datakových typů s nižší přesností (např. *int8*) oproti obvyklým 32 bitovým číslům (např. *float32*), ve kterých bývají reprezentována originální data modelu. Jejím cílem je snížit počet bitů, což vede k nižší velikosti modelu a to k nižším HW nárokům a s tím spojených vlastností (*snížená energetická spotřeba, nižší latence, ...*). Kvantizace má však i nežádoucí dopady a to *ztrátu přesnosti* nebo případně *výkonu* modelu. [huggingface_quantization] GGUF tedy tvoří jednotný formát, ve kterém se spousta otevřených LLM modelů (případně jejich konverzí) distribuuje. Jednou z platforem pro jejich distribuci je například HuggingFace.

LM Studio umožňuje spuštění právě modelů ve formátu GGUF, včetně jejich stažení z repozitářů a funkcemi s tím spojených (například *vypsání seznamu kvantizací, spojení rozdělených modelů, atd.*). Jeho primární funkcí je spouštět lokální LLM modely jako *chat* nebo *lokální server* nabízející API kompatibilní s definicí dle OpenAI. Pro každý druh modelů (jako *Llama, Mistral, Command R, ...*) také obsahuje předdefinované nastavení, které si uživatel může upravit. Mezi tímto nastavením jsou hyperparametry jako *teplota, Min P, Top P, délka kontextu, maximálního počtu vygenerovaných tokenů a dalších*. Mimo toho obsahuje i nastavení pro hardware jako *počet vláken CPU, počet GPU vrstev, GPU framework* a podobně. Vrstvou je zde myšlena vrstva neutronů, která provádí určit výpočty a zpracování vstupních dat. Vrstvy mohou být ku příkladu *vstupní, skryté, výstupní*. Klíčovou vlastností nastavitelnou v rámci runtime tohoto softwaru je možnost nastavit chování kontextového okna. Mezi možnostmi se řadí:

1. *Klouzavé okno* - Model si pamatuje posledních x tokenů z konverzace, které využívá ke generování výstupu a predikci.
2. *Začátek a konec* - Model si pamatuje první prompt (případně i systémový) a zbylé tokeny na konci konverzace.
3. *Zastavení* - Při dosažení počtu tokenů v rámci konverzace rovnající se x , je generování odpovědi zastaveno.

Proměnnou x je zde myšlena celková délka kontextu, kterou *model* nebo *runtime* podporuje. V případě *runtime* v rámci *LM Studio* (resp. `llama.cpp`) může uživatel délku kontextu zvolit až do délky 16 384 tokenů. Pokud však tato délka bude větší, jak podporovaný kontext samotným modelem, délka kontextu, který model má v paměti, bude dán hodnotou z modelu. Například modely vycházející z architektury *Llama-2* disponují kontextem 4096 tokenů, jak bylo diskutováno v sekci 2.1.3. Pokud tedy *runtime* bude nastaven s větším kontextovým oknem, bude rozhodující právě tato hodnota modelu.

Pro spuštěnou instanci modelu lze také nastavit *systémový prompt*. Jedná se o první zprávu, která se modelu předává a měla by obsahovat pokyny pro model definující jeho chování, roli a další relevantní informace, které mohou zlepšit přesnost jeho výsledků nebo mu umožnit lépe pochopit kontext. Tento prompt je běžně vložen pouze na začátku konverzace a poté uložen do paměti modelu. Ukázku takového systémového promptu lze vidět ve výpisu 4.6. Tento konkrétní příklad pochází právě ze softwaru *LM Studio*.

Zdrojový kód 4.6: Příklad systémového promptu

```

1         You're an intelligent programming assistant.
        Output code only, no text around.
2
```

4.4.1.2 Testovací sestava a nastavení

Pro spuštění lokálních modelů byla využita PC sestava, jejíž komponenty jsou popsány v tabulce 4.5. Runtime programu *LM Studio*, popsaného v minulé sekci, byl nastaven, aby využíval 12 z 16 dostupných CPU vláken. GPU akcelerace fungovala v režimu *OpenCL*. Počet GPU vrstev se však liší model od modelu. Pro správnou funkčnost je potřeba udržet celý model v paměti RAM, tedy v případě naší sestavy jsme mohli pracovat s modely pouze cca do 60GB, s tím, že do VRAM grafické karty lze nahrát pouze vrstvy o celkové velikosti přibližně 20GB. Na GPU tedy byl

Typ	Komponenta
CPU	AMD Ryzen 8700F
Základní deska	Asus B650-PLUS TUF
CPU Chlazení	BOX
RAM	Kingston FURY Renegade DDR5 64GB 6400 MHz (4x16GB)
VGA	Radeon RX 7900 XT 20GB Asus TUF
PSU	ADATA XPG CYBERCORE 1300W
Case	BeQuiet! Dark Base Pro 901

Tabulka 4.5: Seznam komponent testovací PC sestavy

akcelerován jen takový počet vrtev, který z jejich celkového počtu odpovídal této velikosti. Délka *kontextového okna* byla nastavena na 8 192 tokenů, avšak pro některé modely je tato délka zkrácena samotným modelem (viz sekce 4.4.1.1). Jednotlivé hyperparametry byly poté nastaveny jako:

1. **teplota:** 0.7
2. **Top P:** 1
3. **Maximum vygenerovaných tokenů:** 3000

Zbytek *hyperparametrů* zůstal na defaultních hodnotách daného modelu, který byl testovaný. Toto nastavení bylo využito jak pro *lokální* modely tak pro *API* dotazy na poskytovatele proprietárních či jiných vzdáleně spuštěných modelů. Použité hodnoty těchto parametrů vycházejí z prvotních testovacích experimentů, kdy se osvědčili jako vhodné pro generování jednotkových testů, protože by měli zaručovat *nižší determiničnost* výsledku a jeho *vyšší přesnost*.

4.4.2 Dotazy

Samotné vygenerování testů za pomoci dotazování LLM je řešeno skrze *generující* režim programu, jako je ukázáno ve výpisu 4.2. Tento režim jako argument vyžaduje *název šablony*, která funguje jako *prompt* pro model a byla vytvořena v předchozím kroku (viz sekce 4.3.1). Tento režim doporučuje i více nepovinných argumentů jako:

- **count** - Počet variací testů, které se mají vygenerovat. Celé číslo. Defaultní hodnota 1.

- `manual` - Zkopírovat prompt do schránky pro manuální vložení do rozhraní modelu. Vlajka. Vhodné pro debug.
- `cmd` - Vypsát prompt do standardního výstupu. Vlajka.
- `compress` - Vlajka vyjadřující použití promptové komprese.

Ukázkové volání lze vidět v ukázce 4.7. Program je schopen vygenerovat pro jeden test více variant na bázi stejného promptu (vytváření nových dotazů na LLM). Detailní popis algoritmu je nastíněn v obrázku 4.2. Šablona, kterou uživatel vytvořil je použita pro *render* finálního promptu za pomoci šablonovacího jazyka *Jinja*. Tato metoda byla zvolena z důvodu *jednoduché a předhledné modifikace* vstupů (zde nazvaných jako *šablon* nebo *uživatelských specifikací*), *možnosti nezávislého vložení nahrávky* a také především případné *parametrizace* vstupu, díky které by mohla jít měnit *uživatelská jména, názvy prvků* a další možné údaje v šabloně vhodné pro parametrizaci.

Výpis 4.7: Ukázka volání generace testu dle šablony

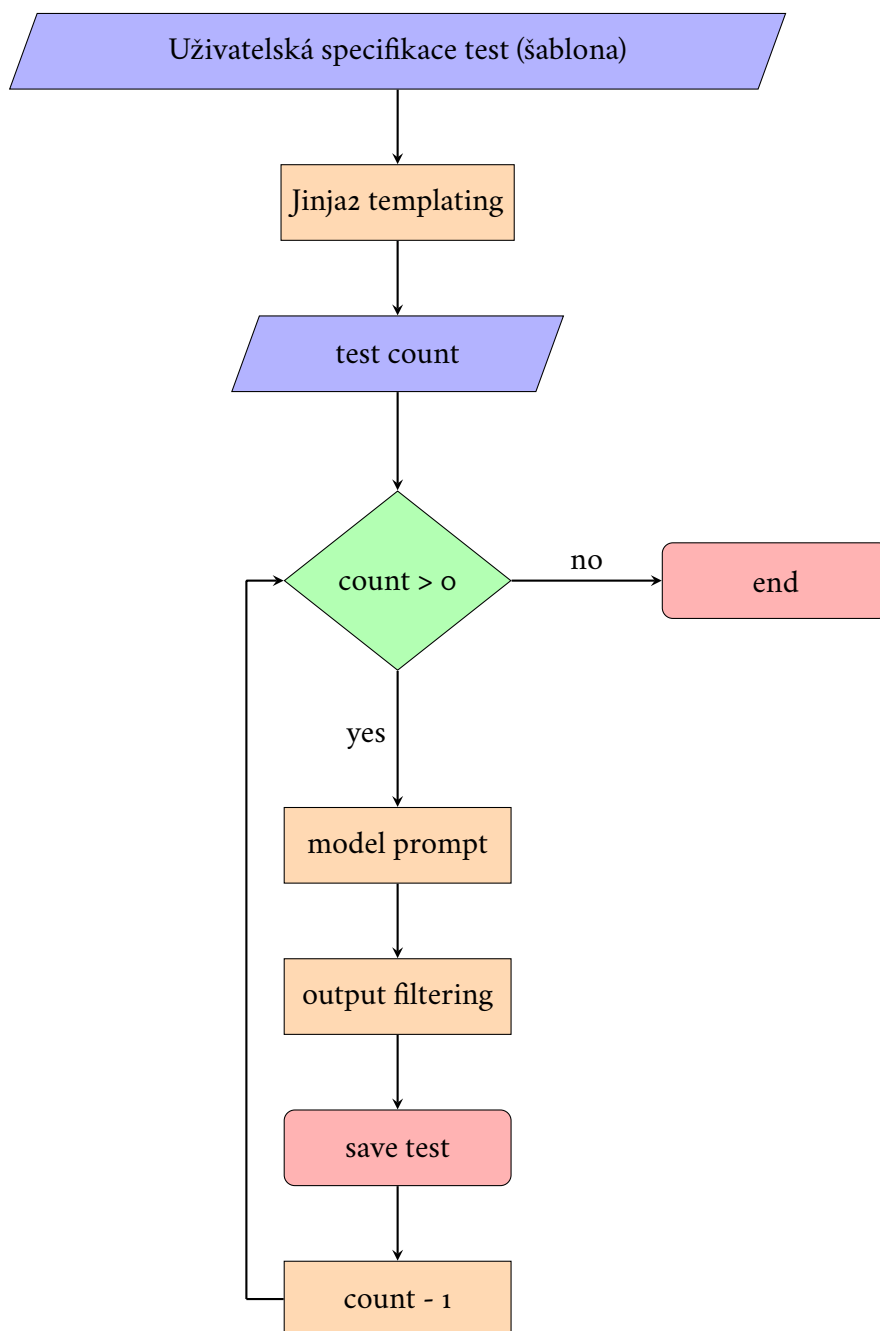
```
1 milan:dp\program$ python3 test.py -i spec-4 --count 10
```

Finální podoba promptu je poté skrze *konektor* pro příslušné LLM (například zmíněná *OpenAI* knihovna, *HTTP* dotaz či jiný druh konektoru) předána jako dotaz modelu. Společně s ním je i modelům předán v rámci těchto konektorů i *systémový prompt*. V případě, že se je aktivní vstupní příznak programu `--manual`, systémový prompt se přidá před vyrenderovaný prompt a tento spojený text je zkopírován do *schránky*. Program načítá systémový prompt ze souboru `templates/system.txt`. Jeho znění (viz výpis 4.8) vychází požadavků a nedostatků, které byli při testovacím generování zpozorovány.

Zdrojový kód 4.8: Použitý systémový prompt

```
1 You're an intelligent programming assistant. You write Robot
  Framework scenarios and scripts using the Selenium Library
  . Insert delays between steps. Output code only, no text
  around. Use Chrome as browser. Locate elements using XPath
  . Close browser between scenarios.
```

I přes důraz na nutnost pouze *programového* výstupu v rámci systémového promptu, mají modely tendenci tuto podmínku ignorovat a generovat text okolo kódu. Protože jsou modely naučené tak, aby jejich výstup byl text v **Markdown** formátu, obsahují i značky se začátkem kódového bloku a značením jazyka, ale také



Obrázek 4.2: Zjednodušené schema algoritmu pro dotazování jazykového modelu při generování testu.

prvky jako odrážky, apod. Tyto prvky se pak objevují i ve výstupech, kde je požadován pouze *čistý text* (jako například v ukázce 4.9). V našem případě je toto chování nevhodné, protože vyžadujeme na výstupu pouze *kód testu*. Díky struktuře *Markdown* formátování však lze kód z tohoto výstupu jednoduše vyparsovat. Toto parsování má právě na starosti funkce *filtrování výstupu* ve schématu 4.2. V rámci programu byla implementována tak, že ze vstupního textu vyjme kód v prvním nalezeném kódovém bloku, označeného pomocí znaků ‘‘‘. Pokud blok nenajde, vrátí původní text bez změny (předpokládá, že model se držel systémového promptu). V případě více kódových bloků přítomných ve výstupu, kdy požadovaný kód nebude v prvním z bloků, dojde k nesprávnému parsování, avšak jedná se o očekávaný případ, protože detekce správného kódu je netriviální a zároveň *subjektivní* problematika.

Zdrojový kód 4.9: Výstup modelu s nadbytečným textem.

```

1 Here's a Robot Framework scenario that follows the recorded
  JSON steps for interacting with the web page, and then
  performs the checks as specified:

3 '''robotframework
4 *** Settings ***
5 Library                SeleniumLibrary

7 *** Variables ***
8 ${URL}                 http://localhost:4680/tbuis/index.jsp
9 ${USERNAME}            lazy
10 ${PASSWORD}           pass
11 ${BROWSER}            Chrome

13 *** Test Cases ***
14 Open and Verify Webpage Content
15 ...
16 '''

18 In this Robot Framework script, I've incorporated commands
  that:
19 — Open the specified page and set the viewport size.
20 — Perform the login process using the provided username and
  password.
21 — Navigate to the "List of All Teachers" section.
22 — Check for the presence of specified names on the page.
23 — Validate the text content of specific table cells related
  to the teachers' courses.
24 Adjust the locators (xpath) as necessary to match the exact
  structure and identifiers used in your application's HTML.
25

```

Vyfiltrovaný výstupní text modelu je poté zapsán do souboru ve výstupní složce nazvané `generated`, nacházející se v kořenové složce programu. Test je uložen s názvem ve formátu `{nazev-vstupu}-{cislo-varianty}.robot` jakožto *RobotFramework* soubor. První část tohoto názvu souhlasí s názvem *promptové šablony*, pro kterou uživatel vytváří testy. Druhá část poté čísluje vygenerované varianty. Pokud již ve výstupní složce existuje vygenerovaný test pro danou specifikaci, program nové variantě přiřadí číslo o 1 vyšší, tedy pokračuje v číselné řadě. Názvy vygenerovaných testů jsou postupně vypisovány do konzole. Od uživatele je očekáváno, že vygenerované testy ve výstupní složce zkontroluje a ověří, že obsahují validní (např. *obsahuje text, připomínající RobotFramework test*). Je zde totiž stále možnost, že model nevygenerovat žádný kód a vypsat pouze text nebo *filtrace* neproběhla validně. Tento krok by v ideálním případě bylo možné automatizovat za pomoci *analyzátoru* pro *RobotFramework*, který se nachází v rámci LSP pro *RobotFramework* `robotframework-lsp` nachází. Bohužel však neumí validovat zápis pro knihovnu *SeleniumLibrary*, která je v rámci systémového promptu vyžadována a očekávána, že bude přítomná ve vygenerovaném testu.

Spuštění testů

5

5.1 Spuštění testovaného programu a jeho orchestrace

Testovaný program *TbUIS* (popsaný v sekci 2.2) je distribuován jakožto `.WAR` soubory, tedy komprimovaný webový archiv jazyka Java, spustitelný jakožto servlet. Z dokumentace projektu a předchozích diplomových prací vyplývá, že pro nasazení je určený aplikační server *Tomcat* verze 7 až 9. S vyššími verzemi není aplikace kompatibilní. Každá varianta aplikace (*poruchové klony*) je distribuována jako samostatný `WAR` soubor. Pro spuštění konkrétní varinaty a spuštění vygenerovaných testů na této variantě bylo nutné vytvořit *automatizované nasazení* aplikace *TbUIS*. Pro nasazení je nutné brát v úvahu nejen samotnou webovou aplikaci, ale i její závislost, kterou je *databáze MySQL*, která musí být inicializována dodaným skriptem.

Pro účely *automatizovaného nasazení* byla zvolena technologie *Docker*, tedy *kontejnerové* řešení s možností vytvářet obrazy a kompozice. Právě kompozice je vhodným nástrojem pro jednoduché *lokální* nasazení aplikace *TbUIS*, protože potřebujeme 2 samostatné kontejnery; první s *aplikačním serverem* a druhý s *databází*.

Vyhodnocení výsledků



Budoucí vylepšení

7

Závěr

8

Seznam obrázků

2.1	Prostředí systému TbUIS z pohledu studenta.	12
4.1	Nahrávání scénáře za pomoci nástroje v Google Chrome	17
4.2	Zjednodušené schema algoritmu pro dotazování jazykového modelu při generování testu.	25

Seznam tabulek

2.1	Přehled a srovnání studií	10
2.2	Přehled a srovnání modelů generujících kód	10
4.1	Specifikace pro generované testy - část 1	15
4.2	Specifikace pro generované testy - část 2	16
4.3	Seznam poruchových klonů využitých pro testování.	16
4.4	Použité LLM modely	21
4.5	Seznam komponent testovací PC sestavy	23

Seznam výpisů

4.1	Ukázková struktura input složky	18
4.2	Hlavní režimy programu	18
4.3	Vytvoření nového testu (šablony)	19
4.4	Vzor pro vyplnění šablony testu	19
4.5	Vyplněná šablona testu pro specifikaci 18	19
4.6	Příklad systémového promptu	22
4.7	Ukáзка volání generace testu dle šablony	24
4.8	Použitý systémový prompt	24
4.9	Výstup modelu s nadbytečným textem.	26

1101001
10101100001110010 1100001
101011010101 10



11010011101101001
0110000110101
111000101011101