tesena
SMART TESTING

SROVNÁNÍ AI-CHATBOTŮ PRO POUŽITÍ V TESTERSKÉ PRAXI

DUBEN 2024

David Sedláček, Marina Mulyukina, Jakub Benešovský,
Petra Ďurková, Lucie Lavičková

www.tesena.com
info@tesena.com

Abstrakt

Tato e-kniha se zaměřuje na porovnání a hodnocení různých chatbotů s cílem zjistit jejich použitelnost při testování softwaru a skriptování. Zkoumali jsme efektivitu práce a výstupy chatbotů ChatGPT 3.5, ChatGPT 4.0, Gemini a Microsoft Copilot v různých oblastech, jako je generování obsahu, kódování a diskuse.

Obsah je rozdělen do čtyř hlavních sekcí: generování obsahu, generování nápadů, zkoumání a kódování. Každá kapitola zahrnuje porovnání schopností jednotlivých chatbotů při zpracování úloh, generování odpovědí a komunikaci s uživatelem.

Tento článek je určen pro profesionály v oblasti softwarového inženýrství, vývojáře, testery a studenty. Předchozí zkušenosti s chatboty nejsou nutné, avšak základní znalosti z oblasti programování budou užitečné.

Je důležité poznamenat, že informace uvedené v tomto článku jsou časově citlivé, protože chatboti se neustále zlepšují a vyvíjejí. Na závěr tohoto článku budou mít čtenáři jasnou představu o schopnostech jednotlivých chatbotů a budou schopni tyto technologie efektivně využívat ve svých vlastních projektech.

Obsah

Úvod

V současné době je dostupnost AI-chatbotů (dále jen chatbotů) stále širší a rychlý rozvoj této technologie přináší nové možnosti v různých odvětvích. Jejich vývoj a nasazení se stává běžnou praxí, a to zejména v zájmu zlepšení efektivity práce. Avšak s rychlým růstem možností se rovněž zvyšuje složitost výběru nejvhodnějšího nástroje pro konkrétní potřeby.

Sledování technologického vývoje chatbotů je pro většinu nejspíše výzvou, kterou nezvládají udržet. Nové funkce, algoritmy a platformy se neustále objevují, a to vytváří dynamické prostředí při výběru těchto nástrojů.

Při rozhodování o výběru chatbotu je tedy vhodné provést komplexní srovnání dostupných možností. To vyžaduje nejen porovnání funkcionalit a vlastností jednotlivých nástrojů, ale i zhodnocení jejich schopnosti splnit konkrétní potřeby a cíle organizace či projektu.

V tomto článku se proto zaměříme na zkoumání široce dostupných chatbotů a jejich srovnání a identifikujeme ty, které nejlépe vyhovují potřebám testerů a přinášejí maximální efektivitu v praxi.

Porovnání chatbotů nám pomůže s výběrem nejvhodnějšího nástroje pro konkrétní případy užití. Tento aspekt je klíčový pro efektivní použití chatbotů v praxi, kde je důležité zvolit nástroj, který co nejlépe vyhovuje našim potřebám a cílům.

Příchod chatbotů mění požadavky na znalosti testerů a zvyšuje efektivitu práce tím, že umožňuje rychlejší a efektivnější testování. Také zjednodušují pracnost automatizace tím, že usnadňují analýzu, úpravu i tvorbu skriptů. Proto vnímáme jako důležité je podrobit širokému spektru úkolů pro zmapování jejich dovedností.

Cílem našeho průzkumu je identifikovat nejvhodnější chatboty pro potřeby testerů v různých situacích a provést srovnání populárních chatbotů s ohledem na jejich schopnost splnit konkrétní potřeby a cíle v oblasti testování.

Proto jsme definovali následující otázky, které se snažili co nejlépe zodpovědět v tomto článku:

1. V jakých situacích mohou chatboti podporovat potřeby a cíle testerů?
2. Jaké jsou nejpopulárnější chatboty používané v oblasti testování?
3. Jak se liší jejich schopnosti splnit specifické úkoly?
4. Jaké jsou výhody a nevýhody každého z identifikovaných chatbotů z hlediska potřeb a cílů testerů?

1. Metodika

V této sekci článku uvádíme výběr chatbotů pro srovnání a co za rozhodnutím stálo. Zaměřujeme se zde na zvolenou metodiku a postup, kterými hodnocení provádíme. Představujeme aspekty kvality a výběr kritérií, jimiž jsme posuzovali kvalitu a efektivitu chatbotů. Poskytujeme podrobný pohled na naše hodnotící postupy, které jsme při hodnocení jejich schopností aplikovali.

Výběr chatbotů pro naše srovnání byl založen na jejich dostupnosti a popularitě, a to nejen v rámci naší testerské komunity, ale také mezi kolegy, kteří s námi pracují na různých projektech. Chatboty s omezenou dostupností, jako je Meta a Anthropic, jsme z výběru vyřadili. Nezabývali jsme se ani platformami pro tvorbu AI asistentů, z velkých firem tak vypadl Amazon.

Zvolili jsme produkty od OpenAI, Google a Microsoft. ChatGPT jsme zařadili hned v obou variantách, volně dostupná verze 3.5 i placená verze 4.0. Google s nedávným přejmenováním Barda na Gemini poskytuje také dvě varianty, volně dostupnou a Gemini Advanced zahrnuli jsme pouze základní variantu. Posledním je Microsoft Copilot, dříve označovaný jako Bing AI, který po přechodu na integraci s GPT-4 zůstává stále zdarma dostupný v prohlížeči Edge.

V poslední době vzniká celá řada chatbotů specializovaných pro testování, my se zde ale chceme zaměřit na univerzální nástroje. Těm specializovaným dáme nějaký čas dozrát a srovnáme je návazně.

Při posuzování se zaměřujeme na několik aspektů kvality, které přispívají k celkovému hodnocení. Hlavní aspekty kvality mají pozitivní přínos k hodnocení, jsou jimi:

1. **Porozumění zadání**, hodnotí se schopnost interpretovat a řešit problémy obsažené v zadání a schopnost správně identifikovat a interpretovat kontext. Kritériem úspěchu je, zda výstup odpovídá obsahu zadání a jestli správně zohledňuje kontext.
2. **Kompletnost výstupu**, hodnotí se zde rozsah a úplnost výstupu v porovnání s požadavky definovanými v zadání. Kritérium úspěchu je, zda je výstup dostatečně obsáhlý a pokrývá všechny klíčové body nebo požadavky.
3. **Adaptace formy výstupu**, hodnotí se schopnost přizpůsobit formu výstupu podle definice cílového publika nebo specifik úkolu. Kritériem úspěchu je schopnost srozumitelně komunikovat informace.
4. **Reakce na dodatečné požadavky**, hodnotí se schopnost reagovat na nové požadavky nebo změny v kontextu a adekvátně je začlenit do výstupu. Kritériem úspěchu je schopnost flexibilně reagovat na nové informace při zachování kvality a relevance výstupu.
5. **Chybovost zpracování**, hodnotí se míra, do jaké výstup zahrnuje faktické chyby, logické chyby, nepřesnosti, nekonzistence a regrese. Kritériem je množství a závažnost identifikovaných chyb ve výstupu.
6. **Efektivita použití pro daný účel**, hodnotí se schopnost dosáhnout zamýšlených výsledků s minimálním časem a úsilím. Kritériem úspěch je, zda je použití daného nástroje efektivnější než konvenční zpracování.

Vedlejší aspekty hodnocení ovlivňují pouze pokud mají negativní dopad na praktické užívání nástroje, řadíme mezi ně například **čitelnost, jazykové dovednosti, rychlost odezvy, limitace** a podobné. Rozhodli jsme se nehodnotit nástroje z pohledu bezpečnosti a ochrany soukromí, implementace v komerčním prostředí, schopnosti dlouhodobého učení ani podpory týmové spolupráce.

S výsledky jednotlivých testů jsme dále sledovali vybrané metriky relevantní účelu způsobu užití chatbotů. Primární metrikou je **délka dialogu**, která třídí testy podle počtu vyměněných dotazů na dlouhé, krátké a okamžitě uzavřené. Sekundárními metrikami jsou **frekvence opakování požadavku a četnost opakování dialogu**.

Vyhodnocení těchto metrik není triviální. Délka dialogu, v závislosti na tematické oblasti, může být krátká, někdy i dlouhá a stále se může jednat o kvalitní a produktivní dialog. V některých případech pouze zkoušíme kvalitu reakcí, nebo hloubku znalostí chatbotu a dialog se tím nutně natáhne. Delší dialogy také mohou mít za příčinu problémy s pochopením zadání nebo jeho plněním.

Frekvence opakování požadavku vyšší než jedenkrát za dialog, indikuje možné problémy v pochopení úkolu nebo kompletnosti výstupu. V delších a zejména nelineárních konverzacích může explicitní vyjádření původního cíle být přirozenou potřebou v dialogu.

Nenulová četnost opakování dialogu je obvykle negativním jevem, pramenícím z nepochopení zadání, kontextu, v zacyklení konverzace nebo ve ztrátě návaznosti konverzace. Jelikož nepochopení zadání a kontextu často plyne ze špatné kvality na vstupu, tak je v některých případech obnovení konverzace v nové instanci tolerovatelné.

Testy jsme navrhovali a vybírali tak, aby pokrývaly široké spektrum úloh, které při testování typicky řešíme. Cílem bylo vybrat nejen úlohy, kde je užití chatbotů běžné, ale i takové, kde se většina našich kolegů stále spoléhá na dovednosti a znalosti své nebo svých kolegů. Chatboty takto srovnáváme s obrazem ideálního asistenta a mentora zároveň.

Došli jsme k **sadě šedesáti sedmi úloh pro chatboty**. Ty jsme rozdělili do čtyř oblastí podle povahy úlohy. Jednou z oblastí je generování obsahu, kde chceme po asistentovi generický obsah, typická řešení, šablony, osnovy, doplnění textací, jejich odvození a podobné.

Další oblastí je generování nápadů, ta se od generování obsahu liší hlavně v očekáváních variací výstupů. Jde o úlohy, kde hledáme inspiraci, varianty, možnosti akcí v dané situaci, například návrhy na pokrytí testy, pravděpodobná rizika a způsoby jejich pokrytí, možné příčiny selhání.

Specifickou oblastí je zkoumání, tyto testy prověřují chatbota v roli mentora. Vžíváme se do pozice juniornějších testerů, abychom si nechali vysvětlovat pojmy, koncepty, techniky a často směřujeme dialog na návod k praktické implementaci. Celkově se dialogy zkoumání buď ubírají k specializaci v rámci tématu nebo k jeho generalizaci.

Poslední oblast označujeme za kódování a spadají sem veškeré úlohy spojené s automatizovaným testováním, skriptováním, integračním testováním i testováním databází. Zde jsme zvolili pouze reprezentativní sadu testů, protože se použití chatbotů při vývoji a automatizaci věnuje již řada článků a také protože by si téma zasloužilo samostatné a specifické pokrytí.

1. Generování obsahu - Chceme po asistentovi generický obsah, typická řešení, šablony, osnovy, doplnění textací, jejich odvození a podobné
2. Generování nápadů - Jde o úlohy, kde hledáme inspiraci, varianty, možnosti akcí v dané situaci, například návrhy na pokrytí testy, pravděpodobná rizika, způsoby jejich pokrytí a možné příčiny selhání.
3. Zkoumání - Tyto testy prověřují chatbota v roli mentora. Vžíváme se do pozice juniornějších testerů, abychom si nechali vysvětlovat pojmy, koncepty, techniky a často směřujeme dialog na návod k praktické implementaci. Celkově se dialogy zkoumání buď ubírají k specializaci v rámci tématu nebo k jeho generalizaci.
4. Kódování - Tady spadají veškeré úlohy spojené s automatizovaným testováním, skriptováním, integračním testováním i testováním databází. Zde jsme zvolili pouze reprezentativní sadu testů, protože se použití chatbotů při vývoji a automatizaci věnuje již řada článků a také protože by si téma zasloužilo samostatné a specifické pokrytí.

V našem podání se zadání testu skládá z **nepovinného kontextu a z úlohy**. Některé testy mají cíleně kontext velmi obsáhlý, jiné stručný a jiné neexistující, abychom simulovali použití na projektech s různými přístupy k dokumentaci. Vzhledem k povaze chatbotů naše testy **nemají očekávané výsledky**, závisí na našem expertním pohledu, zda budeme s výstupy spokojeni.

Jelikož si při práci s chatboty postupně osvojujeme tipy, jak s konkrétním nástrojem pracovat, jak předcházet nedorozumění a přizpůsobujeme své požadavky jeho dovednostem, přistoupili jsme k definici pevného zadání pro všechny chatboty. Tím sice přicházíme o nejlepší možné výsledky, které daný chatbot dokáže poskytnout, předcházíme tím ale zkreslení přebytkem nebo nedostatkem expertízy s nástrojem. Naše zadání úloh je proto formulováno stejně, jako bychom ho předávali kolegovi.

Samotný test provádíme vždy **v novém dialogu, zahajujeme zadáním kontextu a úlohy do jedné zprávy a dále reagujeme na výstupy**, dokud nejsme s výstupem spokojeni nebo nemáme jasný obraz, jakým způsobem a s jakou náročností bychom se dobrali uspokojivého výsledku. Záznam konverzace si držíme pro finální hodnocení s nepovinnou poznámkou o okamžitých dojmech.

Při vyhodnocování dodržujeme, že **řešení všech chatbotů stejného zadání úlohy vyhodnocuje jeden expert**. Na základě popsaných metrik řešení hodnotíme na škále od 1 (použitelné v praxi bez zásadních změn) do 5 (nepoužitelné). Chatboty posuzujeme vždy nejprve pro praktickou použitelnost a následně vzájemně.

Předpokládáme jednak, že každý chatbot může mít své silné a slabé stránky, ale také, že není praktické přepínat mezi několika nástroji pro individuální úlohy. Proto hodnotíme

celé oblasti zvlášť průměrem dosažených výsledků testů. Nehledáme unikátní příležitosti, kde by byl jeden z konkrétních nástrojů silnější.

Tím jsme představili metodiku, kterou jsme použili k hodnocení vybraných chatbotů a nastínili jsme oblasti, na které jsme se zaměřili při vyhodnocování jejich vlastností a schopností. Dále se zaměříme na samotné výsledky testování a závěry, ke kterým jsme došli.

2. Srovnání nástrojů

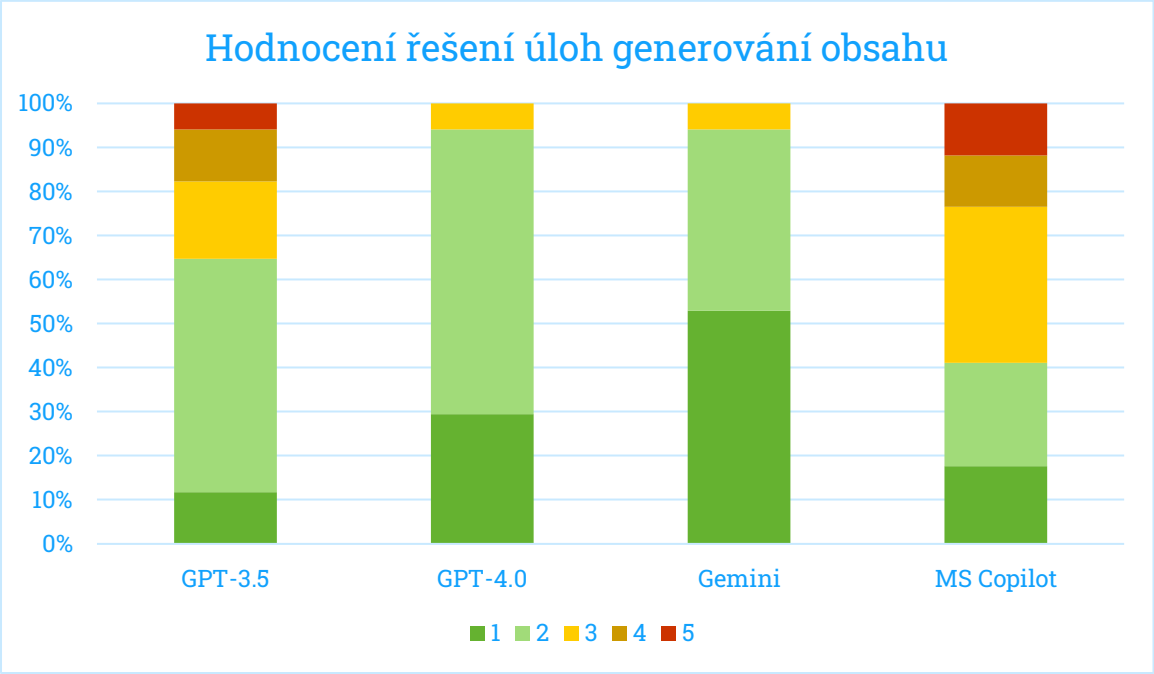
Dále popisujeme, jak si jednotlivé nástroje poradily s našimi úkoly. Probíráme jednu oblast za druhou, v úvodu každé z nich také uvádíme přehled dosažených hodnocení každým z chatbotů. Celkové hodnocení chatbota pro danou oblast uvádíme na konci každé sekce ihned po shrnutí získaných poznatků a dojmů.

2.1. Generování obsahu

V oblasti generování obsahu nás zajímalo, jak dobře nám mohou chatboti pomoct s přípravou obsahu ať už generického nebo specifického, uzpůsobit ho kontextu nebo detailnímu zadání. Hodnotili jsme textace do testovacích strategií a plánů, do uživatelských příruček i testovacích scénářů. Generovali release notes, shrnutí avysvětlení pojmů i přístupů, kontrolovali pokrytí požadavků testy, kategorizovali defekty a hledali duplicity.

Zvolili jsme přístup zkoumání na příkladě fiktivního bankovního projektu, měnili jsme jeho zaměření, velikost i jakým životním cyklem vývoje softwaru se řídí. Simulovali jsme změny v průběhu projektu a problémy s nimi spojené, které běžně řešíme na našich projektech.

Ačkoli už samotná příprava návrhu textace je výrazným usnadněním práce, protože je snazší a baví nás více revidovat výstupy cizí práce a hledat v nich nedostatky, než je vytvářet, přesto jsme v našich testech šli dále a snažili se velmi kritickým pohledem jak autora, tak konzumenta ohodnotit použitelnost a přesnost výstupů.



ChatGPT 3.5

Většina odpovědí měla své nedostatky, značnou část bylo možné napravit dalším dotazováním. Občas nepochopil otázku a odpověděl na něco jiného, upřesnění otázky obvykle pomohlo. Velké problémy měl se zpracováním větších datových souborů a vytvářením vlastních závěrů, například pro release notes.

Generování textací s ním nebylo vyloženě snadné, ale stále bylo účinné. Zejména poté co se naučíte jeho limitace, běžné problémy a jak drobné úkoly musí být, tak je práce celkem snadná a odsýpá. Stále je samozřejmě nutné si vyhradit čas a energii na revizi a korekci textů, jeho výstupy slouží dobře jako hrubá kostra nebo námět, ale samostatně použitelné příliš nejsou.

Hodnocení: 2.50

ChatGPT 4.0

Odpovědi tohoto modelu byly většinou velmi kvalitní a často se dala přímo použít první odpověď. Model dobře chápe komplexní i stručná zadání a poskytuje přesné odpovědi. Bývá potřeba specifikovat konkrétní formát odpovědi, pohled na problematiku nebo doplňující dotaz, aby jeho výstupy pokrývaly plně naše potřeby.

Práce s ním je pro generování obsahu výrazně snazší než s ChatGPT 3.5, stačí méně dotazů, úprav i manuálních korekcí pro dosažení vhodného výstupu. Také vzácněji narazíte na témata, která by prostě nezvládal pokrýt, nebo se u nich ztrácel v kontextu.

Hodnocení: 1.76

Gemini

Gemini zde dosáhl nejlepšího hodnocení. Jeho odpovědi byly kreativní a zodpovídaly naše dotazy. Naše úkoly byly čistě s textovým zadáním a tam exceloval. V několika případech, kdy jsme mu dali část zadání v obrázku, jsme dosahovali horších výsledků, ale sloučení do jednoho textového zadání nám vždy opět zvedlo kvalitu výstupu.

Oproti ChatGPT modelům se jeho odpovědi výrazně odlišovaly formou, kdy Gemini poskytuje často víc doplňujících informací strukturovaných do pojmenovaných sekcí. V takových případech se nám snadno pracovalo i s delšími texty, které by v jiném podání byly zdlouhavé a nepřehledné.

Občas jsme s Gemini narazili na zadání, které kombinovalo mnoho požadavků, cílů a pravidel, kde bylo složité se snadno dobrat optimální odpovědi. Takové zadání byly složité pro všechny z testovaných chatbotů, ověřeným postupem je zadání rozdělit do několika konverzačních toků.

Hodnocení: 1.50

Microsoft Copilot

Microsoft Copilot často nepochopil položený dotaz a jeho odpovědi pak postrádaly zásadní informace. Bylo nutné velmi pečlivě kontrolovat získané odpovědi, často se totiž zdálo, že zodpovídají dotaz, ale při bližším zkoumání obsahovaly nesmysly, reference a obecné formulace.

Limitace na 4000 znaků vstupní zprávy nám mírně bránila v analýze větších datových souborů, obvykle je možné ji alespoň rozdělit a posílat po částech, ale tento proces byl s Copilotem obzvláště krkolomný.

Další složitostí je správně zvolit mezi styly konverzace (kreativní, vyvážený, přesný) pro daný dotaz, vyžaduje to experimentaci a jisté zkušenosti s jednotlivými styly, aniž bychom měli jistotu, že daný styl bude nejpřínosnější.

Očekávali jsme velkou sílu kreativního stylu, přece jen síla Gemini byla právě v kreativitě, často jsme ale došli k závěru, že přesný styl nám pro tvorbu obsahu vyhovuje lépe. V moment, kdy revidujeme a doplňujeme textace, je snazší doplňovat detaily, než redukovat prázdné fráze a obraty z každé věty.

Hodnocení: 2.80

Při generování obsahu ChatGPT 3.5 a Microsoft Copilot ukázaly určité nedostatky v chápání zadání a přesnosti výstupů. Používat určitě jdou, musíme se jim ale trochu přizpůsobovat a spokojenost s výstupem není příliš vysoká.

Naopak, ChatGPT 4.0 a Gemini poskytly výrazně kvalitnější a relevantnější odpovědi. Gemini je kreativní a uživatelsky přívětivý, zatímco ChatGPT 4.0 je univerzálnější a celkem konzistentní.

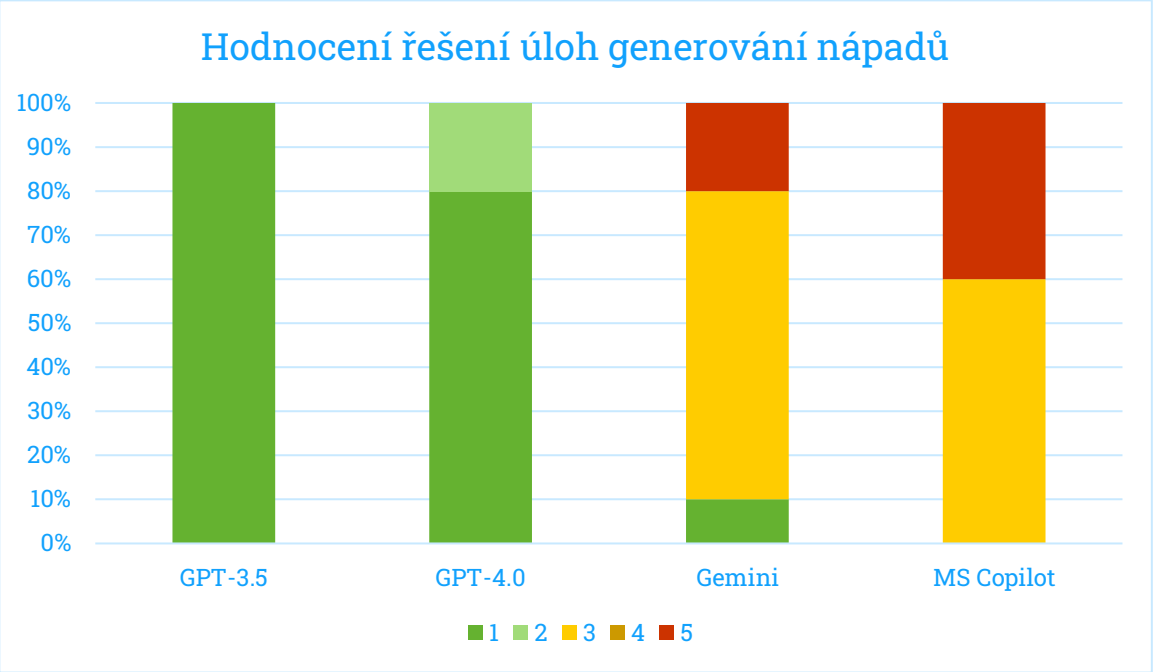
Ve většině případů bychom dali přednost Gemini, ale pro komplexnější úlohy, jakými jsou typicky tvorba tabulek a transformace informací, často zvládneme s menším úsilím skrz ChatGPT 4.0.

2.2. Generování nápadů

Všichni máme různé znalosti a zkušenosti, které nám pomáhají se přizpůsobit situaci a poskytnout velkou přidanou hodnotu s malou námahou. Dokážeme se výrazně lišit v přidané hodnotě v závislosti na tématu nebo specifickém úhlu pohledu. Pokud jsme se nějaké oblasti dlouho nevěnovali, může nám také déle trvat si vzpomenout na záludnosti, jaké nás mohou potkat.

Ať už hledáme inspiraci v tom, co nebo jak by bylo vhodné podrobit kritickému pohledu specialisty na kvalitu, jaké scénáře mohou nastat, jakým způsobem by šlo daný úkol řešit, nebo jenom hledáme ujištění, že jsme na nic důležitého nezapomněli, tak nám Chatboti mohou velmi pomoct.

Nemělo by valný smysl kontrolovat, zda nám kdejaký chatbot zvládne podat učebnicové řešení, proto se v našich testech soustředíme na konzultační použitelnost jak v obecných, tak velmi specifických oblastech testování a na generování příkladů, námětů i dat.



ChatGPT 3.5

Model ChatGPT 3.5 projevil **vysokou úspěšnost při generování přesných a relevantních dat pro různé testovací scénáře**. Ohledně přístupů a námětů poskytl vždy dost bohatý a nikoli příliš obsáhlý výstup, na který šlo snadno navazovat doplňujícími otázkami. Exceloval tak ve všech testech bez výrazných výkyvů a potřeby upřesňovat zadání.

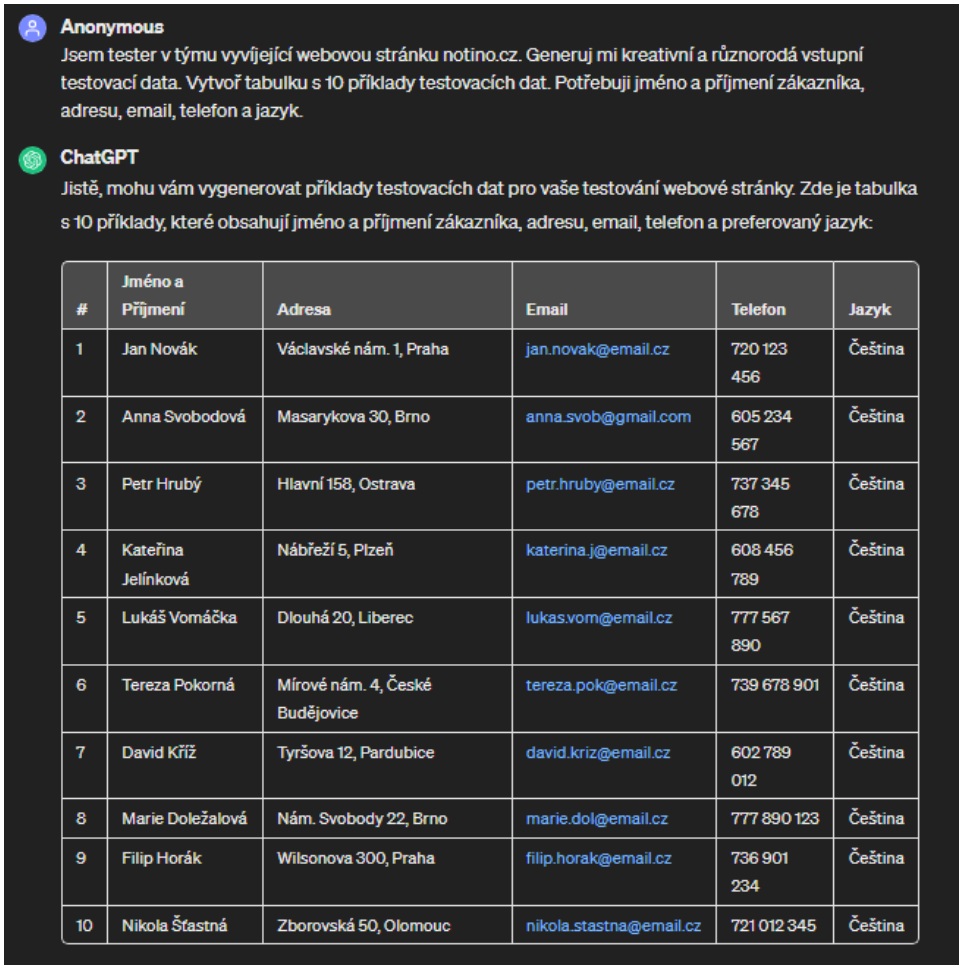
Samozřejmě najdou se úlohy z této oblasti, které by nevyřešil k naší spokojenosti, ale podle výsledků vzorových úkolů není nutné hledat silnější modely pro účely hledání nápadů.

Hodnocení: 1.00

ChatGPT 4.0

Tento model byl také velmi úspěšný, exceloval v **poskytování komplexnějších odpovědí a generování nápadů s větší jemností**. Jeho schopnost formulovat detailní a rozsáhlé odpovědi byla zřetelná. V některých případech však správně nepochopil, v čem by variace měla spočívat a bylo potřeba dalších dotazů k dosažení primárního cíle.

Obrázek **Error! Reference source not found.** ukazuje, jaká data poskytl ChatGPT 4.0 v úvodní odpovědi. Tím nedosáhl významné variace pro účely testování, ve srovnání s modelem ChatGPT 3.5 to byl velmi slabý výkon. Sice ho napravil, museli jsme ale



Anonymous

Jsem tester v týmu vyvíjející webovou stránku notino.cz. Generuj mi kreativní a různorodá vstupní testovací data. Vytvoř tabulku s 10 příklady testovacích dat. Potřebuji jméno a příjmení zákazníka, adresu, email, telefon a jazyk.

ChatGPT

Jistě, mohu vám vygenerovat příklady testovacích dat pro vaše testování webové stránky. Zde je tabulka s 10 příklady, které obsahují jméno a příjmení zákazníka, adresu, email, telefon a preferovaný jazyk:

#	Jméno a Příjmení	Adresa	Email	Telefon	Jazyk
1	Jan Novák	Václavské nám. 1, Praha	jan.novak@email.cz	720 123 456	Čeština
2	Anna Svobodová	Masarykova 30, Brno	anna.svob@gmail.com	605 234 567	Čeština
3	Petr Hrubý	Hlavní 158, Ostrava	petr.hruby@email.cz	737 345 678	Čeština
4	Kateřina Jelínková	Nábřeží 5, Plzeň	katerina.j@email.cz	608 456 789	Čeština
5	Lukáš Vomáčka	Dlouhá 20, Liberec	lukas.vom@email.cz	777 567 890	Čeština
6	Tereza Pokorná	Mírové nám. 4, České Budějovice	tereza.pok@email.cz	739 678 901	Čeština
7	David Kříž	Tyršova 12, Pardubice	david.kriz@email.cz	602 789 012	Čeština
8	Marie Doležalová	Nám. Svobody 22, Brno	marie.dol@email.cz	777 890 123	Čeština
9	Filip Horák	Wilsonova 300, Praha	filip.horak@email.cz	736 901 234	Čeština
10	Nikola Šťastná	Zborovská 50, Olomouc	nikola.stastna@email.cz	721 012 345	Čeština

Obrázek 1: Dotaz na Variace testovacích dat

k řešení jít oklikou a zabralo to zbytečně moc komunikace a času.

Je sice velmi pravděpodobné, že nedostatky ChatGPT 4.0 jsou spíše namátkové a na velkém množství dotazů by se rozdíly stíraly, ale vzhledem k tomu, že zadání úkolů pro všechny chatboty jsme drželi identická, tak nezbývá než konstatovat, že **novější model není natolik kreativní, jak bychom po zkušenostech s předchozí verzí doufali**.

Hodnocení: 1.20

Gemini

Gemini vykazoval **nižší míru úspěšnosti a často generoval nesprávné nebo neúplné odpovědi**. I když v některých oblastech dosahoval přijatelných výsledků, jeho schopnost generovat relevantní nápady pro testery byla omezenější než u modelů ChatGPT 3.5 a ChatGPT 4.0

Primárním problémem není, že by nápady nebo data nezvládl vygenerovat, nebo nedokázal navrhnout variabilitu. Největším nepřítelem tohoto modelu z naší

perspektivy byla forma výstupu. Tabulky byly občas nesmyslně sestavené, nebo část obsahu byla v návazném textu za tabulkou jako náměty na pokračování.

V kombinaci s tendencí modelu nabízet návody, jak by se úkol dal řešit obecnými kroky a občasnou anonymizací poskytnutých dat, jsme se při každodenním používání začali stále rychleji dostávat k frustraci z plynulosti práce, což vedlo k výraznému zhoršení hodnocení pro tuto oblast.

Hodnocení: 3.20

Microsoft Copilot

Microsoft Copilot vykazoval nekonzistenci v poskytovaných výsledcích a často v našich úlohách nedosahoval ani úrovně minimálního očekávání od juniorního testera. Samozřejmě většiny očekávaných odpovědí se tímto nástrojem eventuelně dobereme, ale za cenu velmi detailního specifikování úkolu, až do té míry, kdy by bylo snazší úkol vyřešit samostatně.

Věcnost některých odpovědí nás velmi zklamala, protože ani po důkladném prostudování velmi obecné odpovědi s příklady jsme se nedozvěděli nic nového a museli se ptát znovu a lépe i když se jednalo o každodenní problémy testera.

Hodnocení: 3.80

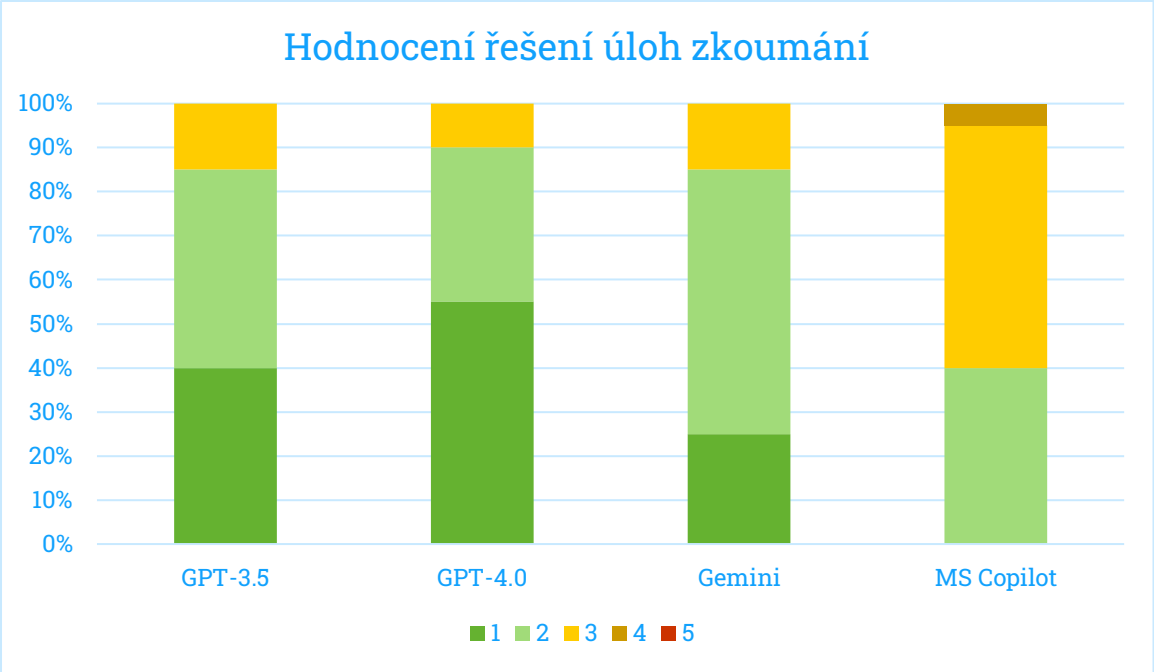
Celkově lze říct, že modely ChatGPT 3.5 a ChatGPT 4.0 jsou vhodné pro generování nápadů, námětů a dat. Gemini má v této oblasti vysoký potenciál a doporučujeme ho vyzkoušet, ale na denní bázi ho k tomuto konkrétnímu účelu asi nebudete chtít používat. Microsoft Copilot zde zaostává a potřebuje se zlepšit, zatím ho ale nemůžeme doporučit.

2.3. Zkoumání

Dalším faktorem, kterým se mezi sebou rozlišujeme a často označujeme jako senioritu jsou naše rozdíly ve znalostech, přístupech k řešení problémů a schopnostech se učit. Náš proces učení již v minulosti jednou výrazně usnadnilo online vyhledávání, nyní se nám s pomocí chatbotů odemyká další úroveň efektivity, kdy informace z mnoha zdrojů nebo obsáhlého materiálu můžeme vytáhnout bez většího úsilí ve velmi krátkém čase.

V oblasti zkoumání se naše testy zaměřují na zhodnocení použitelnosti chatbotů jednak k sumarizaci znalostních oblastí z perspektivy odborníka na kvalitu, ale také k dalšímu zkoumání takové oblasti s cílem jí lépe porozumět nebo se jí naučit.

Úlohy, které jsme chatbotům zadávali, se skládaly z jednoduchých témat, jak testovat vybrané nefunkční aspekty kvality, co může být zdrojem problému nebo jakou strategii bychom v naší situaci měli zvolit, ale i specializovaných dotazů na specifika testování konkrétních technologií či zařízení.



ChatGPT 3.5

ChatGPT 3.5 podával skvělé výsledky v naprosté většině úloh. Kamenem úrazu často ale byly úkoly návrhu negativních scénářů, kde jeho odpovědi vyjadřovaly spíše co by mohlo nefungovat než jak se pokusit narušit správné chování předmětu testování.

Mimo to ale poskytoval solidní odpovědi s občasnou potřebou upřesnění dotazu a z našeho pohledu tak patří mezi snadno použitelné a výkonné nástroje pro daný účel.

Hodnocení: 1.75

ChatGPT 4.0

ChatGPT 4.0 poskytoval lepší odpovědi v řadě oblastí dotazů na testování softwaru než jeho předchůdce. Rozdíl byl zejména znatelný při dotazování na identifikaci zranitelností, detekci výkyvů výkonu, testování integrity dat a identifikaci negativních scénářů. Přestože ukázal zlepšení v mnoha oblastech, také čelil některým výzvám v testovacích metodologiích a testování mikro služeb.

Rozdíl mezi tímto modelem a jeho předchůdcem je sice velmi malý, ale pokud používáte chatbota ke zkoumání na denní bázi, je znatelný.

Hodnocení: 1.55

Gemini

Výsledky testů byly překvapivé, navzdory očekávání, s problémy, kterým čelily GPT modely si Gemini poradil docela snadno. Naopak mu dělaly větší problémy témata jako testování API a UI/UX.

Při komunikaci s Gemini byla běžná potřeba navazovat na úvodní dotaz s požadavkem o doplnění. Také má tendenci uvádět více informací, než je potřeba a může být složitější v odpovědích najít těch pár vět, které nás skutečně zajímají, což může být unavující zejména při delších nebo pravidelnějších konverzacích.

Celkově jsme tím došli k závěru, že je určitě velmi dobře použitelný pro daný účel, je ale bohužel zastíněn praktičností obou GPT modelů, kterým dáváme při zkoumání určitě přednost. Prakticky si ale necháváme Gemini vždy bokem otevřený, pokud se zdá nějaké téma příliš těžké na GPT, tak ho Gemini běžně krásně doplní.

Hodnocení: 1.90

Microsoft Copilot

Spolupráce s Copilotem na téma zkoumání byla výrazně složitější než s ostatními testovanými chatboty. Má tendenci vracet velmi obecné odpovědi, které jsou bez dalšího upřesnění nepoužitelné. Kvalitě výsledků velmi pomáhalo, pokud byly navazující dotazy formulovány již s obecnou znalostí tématu, což ale není vždy náš případ.

Při práci s ním je určitě vhodné experimentovat s přednastavenými styly odpovědí mezi kreativním, vyváženým a exaktním, v této oblasti není jasná odpověď na to, jaký styl bychom my preferovali, na základě našich zkušeností jsme se ale začali vyhýbat vyváženému stylu.

Jednoduché dotazy zvládá zodpovídat, ale celkově jsme se s ním k uspokojivé odpovědi dobírali skrz delší konverzace, v praxi denního užívání nám šetřil jen velmi málo času a naučili jsme se spíš na něj nespoléhat než ho využívat.

Hodnocení: 2.65

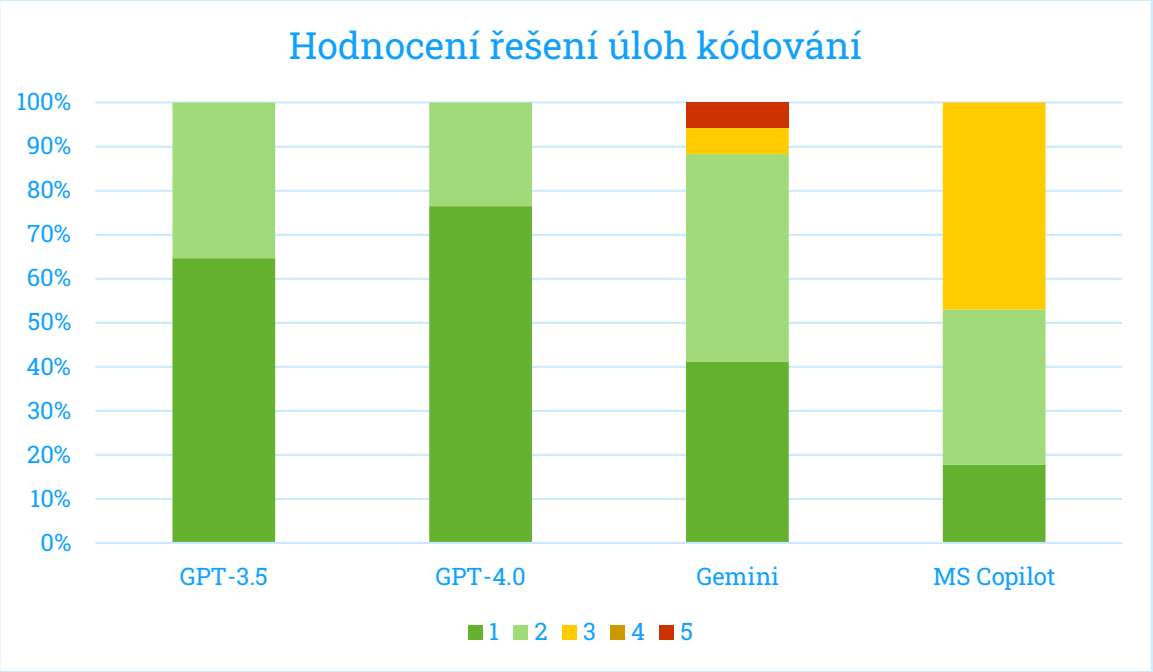
Jednoznačným vítězem je zde ChatGPT 4.0, nejlépe vyhovuje při každodenním používání a poskytuje přesné a kompletní informace. Pokud si ho nechcete platit, tak ChatGPT 3.5 zvládne ty samé úkoly na vysoké úrovni jen s potřebou trochy komunikace navíc. Gemini je určitě taky velmi dobře použitelný a jak jsme již zmiňovali, je skvělým doplňkem k používání GPT modelů. MS Copilot je sice použitelný také, ale práce s ním vyžaduje zvyk a ani pak není příliš plynulá.

2.4. Kódování

Poslední oblastí, které jsme se věnovali, byly schopnosti pomoct s kódováním pro použití při testování a automatizaci. Naše úlohy proto měly za cíl získat popis kódu, generovat komentáře nebo segmenty kódu, případně je upravit. Prověřovali jsme také schopnost kód revidovat, ohodnotit, navrhnout úpravy nebo pokrytí testy.

V poslední řadě také návrh řešení problému ať už v obecné rovině nebo přímo návrhem úpravy kódu. V našich dotazech jsme se limitovali na případy používající Python,

JavaScript, C#, SQL a regulární výrazy. Integrační testy jsme v této sekci nepokrývali, jelikož byly částečně prověřeny již v sekci zkoumání.



ChatGPT 3.5

ChatGPT 3.5 se ukázal jako vynikající nástroj pro zpracování různých skriptovacích úkolů. Má skvělou úroveň schopnosti analyzovat a komentovat kód. Také generovat vhodná testovací data. Dokáže efektivně pomoct i uživatelům, kteří si se skriptováním neví rady nebo nechtějí lámat hlavu.

Hodnocení: 1.35

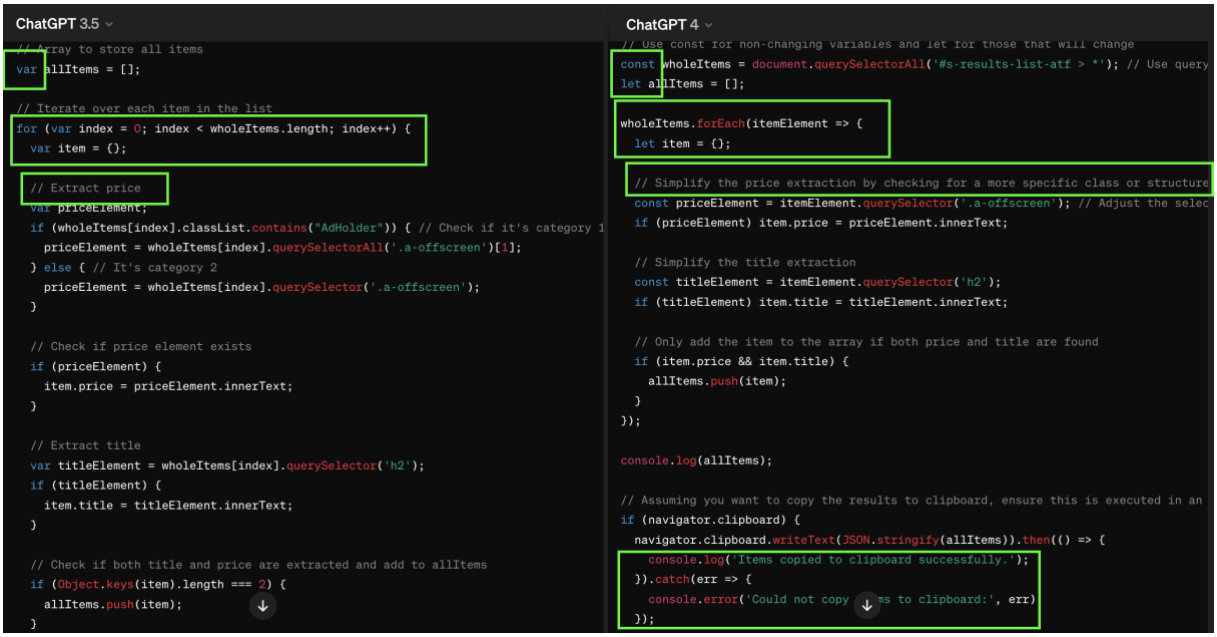
ChatGPT 4.0

Jelikož je ChatGPT 4.0 online, tak je bezesporu výhodnějším asistentem pro kódování, zejména pokud pracujete s technologiemi, které jsou novější a aktivně řešeny online. Oproti předchozí verzi může poskytnout lepší, čistší řešení, a nebo i řešení tam, kde dříve ChatGPT 3.5 nic nenavrhnul.

Abychom si rozdíl demonstrovali, přidali jsme ilustraci Obrázek 2: Srovnání úprav kódu ChatGPT 3.5 a 4.0, kde si můžete všimnout jak verze 4.0 navrhuje přehlednější a modernější způsob zápisu i logiku skriptu. Návrh skriptu verze 4.0 navíc obsahuje úpravy, které pomohou při debuggingu.

Pro základní úkoly skriptování jsou jinak jejich schopnosti srovnatelné, pokud tedy vaše nároky nejsou vysoké, nemusí se ChatGPT 4.0 zdát za svou cenu opodstatněný. V našich testech si vedl lepe než verze ChatGPT 3.5, jen výjimečně byl méně názorný.

Hodnocení: 1.24



Obrázek 2: Srovnání úprav kódu ChatGPT 3.5 a 4.0

Gemini

Po oficiálním přechodu z Bard na Gemini jsou výsledky v této oblasti určitě lepší. Původní potíže s vysvětlením, komentováním a zlepšováním kódu prakticky zmizely. Stále se může zadrhnout na některých úkolech, nám v některých příležitostech odmítl generovat testovací data.

Určitě je použitelným nástrojem, možná ale bude záležet na potřebných technologiích, jestli obstojí v konkrétních situacích. Pro použití v automatizaci testování je v našich testech těsně za GPT modely.

Hodnocení: 1.82

Microsoft Copilot

Copilot si s problémy v této oblasti poradí, ale úroveň výstupu je o poznání horší než u modelů GPT a Gemini. Mírnou složitost fungování s Copilotem představují jeho styly odpovědí. Kreativní a přesný styl odpovědí je velmi lákavý, zejména pokud se nevejdeme do dvou tisíc znaků s naším dotazem, rozdíl mezi nimi ale někdy znamená úspěch či neúspěch odpovědi.

V kódování daleko více než v ostatních oblastech na nás působily jednotlivé styly více či méně vhodné danému dotazu. Nejde jenom o košatost výrazů nebo strohost vět, ale celkové přizpůsobení výstupu cílovému publiku. Přesný styl odpovědí byl obvykle vhodnější pro osobní použití, poznámky a rychlé drobné úpravy. Kreativní styl se naopak více hodil pro výstupy sdílené s týmem nebo pokud jsme se v tématu dobře nevyznali.

Překvapil nás výstup při komentování kódu, kde sice dle našich zkušeností komentáře v přesném stylu byly stručné, někdy nejednoznačné a v kreativním stylu byly obecnější

a lépe pochopitelné, ale rozdíl byl ve vyváženém stylu. Ten navzdory našim očekáváním podal obsáhlejší a přesnější komentáře než oba předchozí styly dohromady.

Ve finále je ale výstup Copilota málokdy lepší než od ostatních srovnávaných modelů a neustálé otevírání nových oken pro změnu stylu konverzace také často narušuje plynulost spolupráce. Nejspíš proto budete preferovat jiného chatbota při práci, nicméně stále je dobře použitelný i Copilot.

Hodnocení: 2.29

Pro každodenní používání je tedy nejvýhodnější pracovat s ChatGPT 4.0, samozřejmě velmi silná je i verze 3.5, ale pokud si vyzkoušíte oba, poznáte, že je verze 4.0 příjemnější na použití. Gemini lze také velmi dobře použít, zatím je ale nutno dbát na jeho limitace ohledně testovacích dat, pokud jste ho letos ještě nevyzkoušeli, určitě mu dejte šanci. Copilot sice není nejplynulejší, ale úkoly vesměs zvládá.

Diskuse

Zkoumaní chatboti byli velmi úspěšní v řešení našich úkolů, nelze proto identifikovat specifické potřeby nebo cíle, které by testerům nemohli pomoct naplnit. Pokud lze potřebu zformulovat do dotazu a skrze textový výstup vyřešit nebo usnadnit řešení, tak lze použít některého z chatbotů.

Rozdíly v jejich schopnostech řešit naše problémy jsou znatelné a dostatečné k tomu, abychom si mohli vybírat vhodný nástroj. Některé z nich mají problém s vybranými problémy natolik znatelný, že bychom je dobrovolně na daný úkol nepoužili.

Zatímco my jsme měli při práci s hypotetickými nebo osobními úkoly absolutní volnost volby a použití chatbotů, tak v praxi to vždy není možné. Některé firmy kladou omezení na používané nástroje, poskytují vlastní varianty nebo si obstarají specifickou licenci splňující jejich požadavky, obvykle na ochranu dat. V takovém případě si příliš vybírat nemůžeme a rozhodujeme se, zda poskytnutý nástroj vůbec použít.

Ze srovnávaných nástrojů bychom žádný nezavrhli, protože naprostou většinu úkolů zvládaly a je velmi pravděpodobné, že je budou zvládat i v projektové praxi. Pokud některý z úkolů nezvládnou, nezbývá nám než přizpůsobit zadání nebo ho řešit postaru.

Naše obava, že ChatGPT 4.0 bude ve všech oblastech nejlepší se nenaplnila. Přestože výsledky jednotlivých chatbotů nejsou příliš rozmanité, stejně je možné sledovat jejich silné a slabé stránky, zda se na naše úkoly nehodí jiný z nich. Určitě nezatratíme Gemini, ani občasné použití ChatGPT 3.5, ale v případě Copilotu je nutné učení na obou stranách.

ChatGPT 3.5

Silné stránky:

- Vhodný pro většinu úkolů
- Snadno použitelný
- Silný nástroj pro kódování
- Efektivní při generování návrhů a dat

Slabé stránky:

- Komplexní úkoly
- Návrhy negativních scénářů
- Vyžaduje více komunikace
- Znalosti do ledna 2022

ChatGPT 4.0

Silné stránky:

- Skvělý pro každodenní používání
- Dobrý i pro složitější dotazy
- Online dotazy
- Práce s novějšími technologiemi

Slabé stránky:

- Někdy nesprávně interpretuje zadání
- Méně kreativní
- Placená služba

Gemini

Silné stránky:

- Generování relevantních nápadů
- Schopnost navrhnutí variability
- Zvládá složitější dotazy
- Kreativnější výstupy

Slabé stránky:

- Sestavení tabulek
- Někdy neúplné odpovědi
- Generování testovacích dat

Microsoft Copilot

Silné stránky:

- Vhodný pro jednoduché dotazy
- Kreativní styl odpovědí
- Přesný styl odpovědí

Slabé stránky:

- Vyžaduje detailní zadání
- Někdy nevěcné a obecné odpovědi
- Vyžaduje více komunikace

Pokud si můžeme vybrat, volíme nejuniverzálnější nástroj ChatGPT 4.0, pokud nejsme spokojeni s výsledky, obracíme se na Gemini nebo ChatGPT 3.5 podle typu a komplexity úlohy. Pokud není možnost používat ChatGPT, volíme v první řadě Gemini, až pak Copilot.

Vzhledem k tomu, že všechny modely kromě ChatGPT 4.0 jsou zdarma, nelze je velmi dobře srovnávat dle nákladů. Ani se nedá jednoznačně říct, že placené služby by byly natolik lepší, co se týká schopností, aby se pouze jimi opodstatnila investice do prémiové služby.

Pro interpretaci našich výsledků a závěrů mějte ale vždy na paměti, že mají jisté limitace. Každého z chatbotů jsme při práci používali několik měsíců, za tu dobu jsme si mohli vytvořit návyky práce s danými nástroji. Data jsme sbírali v lednu a únoru roku 2024, je velmi pravděpodobné, že se jejich výsledky mohly mezitím zlepšit.

Ačkoli jsme se snažili, aby naše testy byly objektivní, tak výsledky otevřených úloh nejsou dobře měřitelné a naše hodnocení odráží naši spokojenost s výstupem, zda bychom takovou odpověď akceptovali od kolegy, nebo jí byli ochotni vzít za vlastní a obhajovat. Hodnocení proto přímo vyplývá ze zkušeností, co jsme nasbírali v praxi a může být specifická.

V ideálním případě, bychom chtěli chatboty podrobit stejné úloze několikrát, ať už beze změny zadání nebo s mírnými úpravami, abychom ověřili konzistenci výstupu a nejlepší výsledky pro každý z modelů. Takový test je ale bohužel příliš časově náročný, nejspíš by z něho nasbíraná data byla již zastaralá, než bychom se dostali k jejich zpracování.

Abychom se dostali k výsledkům v rozumném čase, tak jsme se nemohli věnovat velkému množství chatbotů. Vybrali jsme si významné reprezentanty a museli jsme

vynechat jiná zajímavá jména jako Claude, Perplexity a Gemini Advanced, k nim se ovšem můžeme vrátit.

Prozatím jsme se nevěnovali ani chatbotům specificky učeným nebo upraveným pro použití při testování ani tvorbě vlastního chatbota. Také jsme se nezaměřili na nástroje specifického zaměření jako Github Copilot a obecně nástroje, které AI integrují do svých funkcí. Naší hypotézou je, že bude nejefektivnější používat více nástrojů současně a budeme si jí chtít určitě ověřit.

Jednou z oblastí, kterou chatboti typicky špatně zvládají jsou úlohy s čísly. Naše testy jsme na tuto problematiku nezaměřili, nejen proto, že nejde o specifickou oblast testování, ale zejména protože se nesnažíme nahradit kalkulačky, tabulkové procesory skripty, které jsou pro práci s čísly spolehlivější. Je to pro nás určitě kandidátem na další zkoumání.

Závěr

Na základě analýzy výsledků lze vyvodit několik obecných závěrů:

1. **Rozmanitost chatbotů:** Ověřili jsme, že nelze pouze měřit, jak dobrý chatbot je, ale že se jeho použitelnost liší v závislosti na typu úkolu.
2. **Vhodnost pro různé úkoly:** Zjistili jsme, že každý z testovaných chatbotů má své vlastní silné a slabé stránky, což umožňuje si vybírat vhodný nástroj dle preferencí a potřeb.
3. **Kombinace nástrojů:** Zjistili jsme, že v konkrétních případech je vhodné používat kombinaci chatbotů, pro dosažení lepších výsledků.

Chatboti představují užitečný nástroj při testování softwaru, jejich účinnost závisí na konkrétních požadavcích a kontextu použití. Z široce populárních nástrojů jsou nejefektivnější v první řadě ChatGPT 4.0 a za ním pak Gemini i ChatGPT 3.5. Přestože se Copilot za poslední měsíce velmi posunul, stále je méně praktický než jeho konkurence.