

Etické a bezpečnostné aspekty prompt engineeringu

Autor: Marek Čederle

Vedúci práce: Ing. Peter Bakonyi, PhD.

Fakulta informatiky a informačných technológií

Slovenská Technická Univerzita v Bratislave

10.06.2025



Úvod do problematiky

- ▶ rýchly pokrok v oblasti umelej inteligencie (UI)
- ▶ prompt engineering
- ▶ jailbreaking

Cieľ práce

- ▶ analyzovať etické a bezpečnostné riziká prompt engineeringu a navrhnúť usmernenia pre zodpovedné používanie UI

Identifikované riziká

- ▶ etické riziká — dezinformácie, diskriminácia
- ▶ morálne riziká — nevhodný obsah, návody na výrobu zbraní
- ▶ bezpečnostné riziká — generovanie škodlivého kódu, sociálne inžinierstvo (phishing, deepfake)
- ▶ jailbreaking — obchádzanie bezpečnostných opatrení

Legislatíva

- ▶ EÚ — Akt o UI, kategorizácia podľa rizika
- ▶ USA — štátne zákony, chýba federálny zákon
- ▶ Čína — komplexná legislatíva, regulácia deepfakes a generatívnej UI

Návrh riešenia

Navrhnuté usmernenia budú pozostávať z:

1. úvodu do prompt engineeringu
2. odporúčaní pre etické a férové používanie systémov UI
3. odporúčaní na zlepšenie transparentnosti systémov UI

Experimentovanie

Vybrané boli voľne dostupné neplatené verzie pre modely:

- ▶ ChatGPT
- ▶ Microsoft Copilot
- ▶ DeepSeek
- ▶ Perplexity

Experimentovanie (2)

Experimenty:

- ▶ generovanie malware (ransomware)
- ▶ testovanie cenzurovania odpovedí
- ▶ generovanie dezinformácií
- ▶ sociálne inžinerstvo (phishing email)

Vyhodnotenie

Dotazník:

- ▶ Počet respondentov: 75
- ▶ Najznámejší model: ChatGPT (89%) a zároveň najpoužívanější (81%)

TABUĽKA Č.1: Vnímanie rizík UI

Posudzované riziko	Priemer
<i>Šírenie dezinformácií</i>	7.39
<i>Krádež identity</i>	7.05
<i>Generovanie škodlivého obsahu</i>	7.73
<i>Generovanie malware</i>	6.32

0 = žiadne riziko

10 = veľmi vysoké riziko

Vyhodnotenie (2)

- ▶ Copilot — najviac odolný voči jailbreakom (nie však na 100 %)
- ▶ ostatné modely po jailbreaknutí takmer vždy vygenerovali žiadaný obsah
- ▶ potreba lepších ochranných mechanizmov voči jailbreakom

Vyhodnotenie (3)

Navrhnuté stratégie na zlepšenie bezpečnostných mechanizmov:

- ▶ na úrovni promptu – prísnejšia kontrola vstupov, transformácie promptov
- ▶ na úrovni modelu – safety fine-tuning, pruning, moving target defense
- ▶ vhodné je kombinovať viacero stratégií

Požiadavky na etickú a dôveryhodnú UI

- ▶ Ľudský faktor a dohľad
- ▶ Technická odolnosť a bezpečnosť
- ▶ Správa súkromia a údajov
- ▶ Transparentnosť
- ▶ Nediskriminácia a spravodlivosť
- ▶ Spoločenský a environmentálny blahobyt
- ▶ Zodpovednosť

Zhrnutie

- ▶ UI prináša nové etické a bezpečnostné výzvy
- ▶ nedostatočné obranné mechanizmy modelov UI voči jailbreakom
- ▶ vytvorené usmernenia pre etickú a dôveryhodnú UI

Pripomienky oponenta práce (I.)

Dôvod výberu modelov na experimentovanie:

- ▶ ChatGPT – najznámejší a najpoužívanejší LLM (potvrdené výsledkami dotazníka)
- ▶ DeepSeek – čínsky open-source model, vhodný na analýzu cenzúry
- ▶ Copilot – založený na technológii ako ChatGPT, umožnil porovnanie bezpečnostných mechanizmov medzi Microsoft a OpenAI implementáciou
- ▶ Perplexity – verejne dostupný model s možnosťou testovania bez registrácie

Pripomienky oponenta práce (I.)

Ďalšie zvažované modely:

- ▶ Gemini – vyžadovala telefónne overenie, pričom som nechcel viazať informácie na účty kde budem testovať problematický obsah
- ▶ Claude (Sonnet) – v čase testovania nebolo možné vytvoriť (bezplatný) účet

Pripomienky oponenta práce (II.)

Overovanie experimentov:

- ▶ výstupy som analyzoval manuálne
- ▶ pri detekcii dezinformácií a censorship bias, som odpovede modelov porovnával s informáciami z verejne dostupných internetových zdrojov, pričom som vyhľadával v uznávaných médiách
- ▶ pri generovaní malware som analyzoval vygenerovaný kód na či by sa program správal ako ransomware (šifrovanie súborov, odoslanie kľúča útočníkovi)
- ▶ pri generovaní phishing emailu som analyzoval obsah emailu, a tón správy na podobnosť so známymi phishing emailami

Pripomienky oponenta práce (III.)

Neuvedenie promptov v prílohách:

- ▶ jailbreak prompty, ktoré som použil, sú citované v použitých zdrojoch
- ▶ ostatné vlastné prompty boli iba ako screenshot a to nie pre všetky experimenty; tieto som do prílohy nezahrnul
- ▶ zlepšila by sa transparentnosť ale nie celkom opakovateľnosť (výstupy LLM sú nedeterministické)

Pripomienky oponenta práce (IV.)

Obmedzenia, ktoré som mal, no výslovne som ich v práci neuviedol:

- ▶ použil som výhradne voľne dostupné verzie modelov
- ▶ použil som modely, kde bola možnosť registrácie bez asociácie osobných/kontaktných údajov

Pripomienky oponenta práce (V.)

Censorship bias:

- ▶ censorship bias pri ostatných modeloch (okrem DeepSeek) som testoval tak, že som sa zameral na to, či modely (pôvodu z USA) odmietnu generovať obsah týkajúci sa tém ako Wikileaks alebo vojenské zásahy USA
- ▶ nemali problém s generáciou tohto obsahu kde uvádzali informácie zverejnené na Wikileaks, nemali ani problém generovať obsah ohľadom témy s ktorou som testoval Deepseek