

V dobe, kedy sú veľké jazykové modely (LLM) všadeprítomné, ich spoľahlivosť sa stáva kľúčovým faktorom pre využitie v kritických oblastiach ako právo, medicína či financie. Avšak, hodnotiť ich výstupy bez adekvátneho pochopenia ich limitov môže viesť k **klamanu si metrikami** a následne k nebezpečným rozhodnutiam.

V tomto článku rozoberieme, čo sú AI halucinácie, prečo vznikajú, ako využiť nástroje ako **Retrieval-Augmented Generation (RAG)** a **fine-tuning na doménových dátach** pre zvýšenie spoľahlivosti a ako nastaviť **produkčné KPI**, ktoré reálne odrážajú kvalitu výstupov AI systémov.

Čo sú AI halucinácie a prečo LLM halucinujú?

AI halucinácie označujú situácie, kedy model generuje nesprávne, vymyslené alebo neoveriteľné informácie, ktoré môžu pôsobiť veľmi viero hodne. To je obzvlášť problematické, ak sa AI používa na zodpovedanie otázok vyžadujúcich presné fakty.

Pravdepodobnosti vs fakty

LLM ako GPT fungujú na princípe odhadovania pravdepodobnosti ďalšieho slova vo vete na základe vstupného textu a naučených vzorov z rozsiahlych dátových korpusov. **Nepracujú s databázou faktov, ale s pravdepodobnostnou štatistikou slovných vzorov.**

- **Pravdepodobnosť:** Model predikuje, ktoré slová, frázy alebo vety sú najpravdepodobnejšie nasledovať na základe trénovacích dát.
- **Fakty:** Skutočné overiteľné informácie často vyžadujú presné dátové zdroje, ktoré model nemusí mať pri generovaní po ruke.

Preto môžu modely „halucinovať“ — vytvárať uveriteľné texty, ktoré však nemajú oporu v realite.

Kedy je riziko halucinácií najvyššie?

Nie každý prípad použitia AI si môže dovoliť akceptovať nesprávnu alebo invelity.com spornú odpoveď. Riziko nesprávnych informácií rastie so zložitou, dôsledkom chyby a špecializáciou oblasti. Medzi kľúčové rizikové oblasti patria:

1. **Právo:** Nepravdivé právne rady môžu viesť k vážnym dôsledkom, vrátane porušenia zákona alebo finančných strát.
2. **Medicína:** Nesprávna diagnóza alebo odporúčanie liečby môže byť život ohrozujúce.
3. **Financie:** Nesprávne finančné analýzy alebo odporúčania môžu viesť k obrovským stratám.
4. **Multi-step výskum a vývoj:** Komplexné úlohy s posloupnosťou krokov ľahko vedú k eskalácii chýb, keď sa jedna nepresnosť prenáša ďalej.

V týchto oblastiach je nevyhnutné umiestniť AI systémy pod prísne monitorovanie spoľahlivosti a zabezpečiť, že ich metriky reálne odrážajú schopnosť generovať presné a overiteľné výsledky.



RAG (Retrieval-Augmented Generation) a kvalita knowledge base

Aby sme znížili výskyt halucinácií a zvýšili spoľahlivosť AI, je efektívne kombinovať modely LLM s externými dátovými zdrojmi. Jedným z najvýkonnejších prístupov je **Retrieval-Augmented Generation (RAG)**.

Čo je RAG?

RAG je technika, ktorá napája LLM o relevantné informácie z rozsiahlej knowledge base (KB) prostredníctvom retrieval modulu. Tento modul vyhľadáva dokumenty alebo útržky textu, ktoré sú k téme relevantné, a tieto "fakty" sú následne použité ako kontext pre generovanie odpovedí v LLM.

- **Výhody:** Model už nemusí "vymýšľať" fakty z hlavy, ale čerpá z aktuálnych a overiteľných zdrojov.
- **Kvalita knowledge base:** Kritickým faktorom je ako kvalitná, aktuálna a doménovo relevantná je knowledge base, na ktorú sa retrieval moduly spoliehajú.

Vplyv na spoľahlivosť a metriky

Implementácia RAG významne zlepšuje:

- **Reálnu presnosť výstupov**, keďže odpovede sú založené na overiteľných zdrojoch.
- **Sledovanie chýb**, ktoré je možné priamo prepojiť s kvalitou použitých dokumentov.
- **Moznosť auditu**, pretože poznáme konkrétne zdroje, z ktorých AI čerpala.

Fine-tuning na doménových dátach ako cesta k zlepšeniu evaluácie LLM

Ďalšou metódou zvýšenia spoľahlivosti je **fine-tuning modelu na doménových dátach**. Znamená to, že existujúci veľký model sa ďalej trénuje špecificky na dátach z cieľovej oblasti (napr. lekárske správy, právne predpisy, finančné výkazy).

Výhody fine-tuningu

- **Lepšia doménová špecifickosť:** Model rozumie terminológii, štruktúre a relevantným faktom lepšie.

- **Zníženie výskytu halucinácií v špecifickej oblasti**, pretože model má k dispozícii viac dát relevantných pre daný kontext.
- **Možnosť cieleného hodnotenia a metriky**: evaluácia prebieha na relevantných úlohách a otázkach, ktoré odrážajú reálne scenáre využitia.

Ako testovať a overovať spoľahlivosť po fine-tuningu?

- **Benchmark testy so známymi odpoveďami**: vyberáme otázky s overenými správnymi odpoveďami, ktoré pokrývajú rôzne scénare.
- **Testy s adversariálnymi vzormi**: zámerné zavádzajúce otázky na odhalenie slabých miest modelu.
- **Užívateľský feedback a incident management**: monitorovanie produkčných incidentov, analyzovanie príčin a ich zapracovanie do ďalších vylepšení modelu.

Aké metriky a KPI použiť na meranie spoľahlivosti AI systémov?

Pre profesionálnu evidenciu kvality a spoľahlivosti AI systémov nestačia len základné metriky ako *accuracy* alebo *BLEU* (pre texty). Potrebujeme KPI, ktoré zohľadňujú špecifiká LLM a ich aplikácií.

Metrika/KPI Popis Príklad použitia Fact-Checking Precision Podiel správnych faktických odpovedí overených z knowledge base alebo odborníkov. Kolik % odpovedí správne korešponduje s dokumentáciou alebo referenciami. Recall na doménových entitách Schopnosť modelu rozpoznať a zapracovať do odpovede správne entitné informácie. Vyhľadávacie otázky v medicíne a správne zahrnutie diagnóz, liekov. Incident Rate Počet prípadov, kedy AI generovala nesprávnu alebo riskantnú odpoveď v produkcii. Sledovanie chýb v právnom chatbotovi počas mesiacov. Time-to-Detect (TTD) & Time-to-Resolve (TTR) Rýchlosť identifikácie a opravy produkčných chýb. Monitorovanie incidentov a reakčný čas AI tímu. User Satisfaction Score Zhodnotenie užívateľov na základe ich skúsenosti s kvalitou odpovedí. Prieskumy spokojnosti po interakcii s AI systémom.

Mini-checklist pre spoľahlivú evaluáciu AI systému:

1. Overte, či metriky zahŕňajú fakty a pravdivosť, nie len štylistickú kvalitu výstupu.
2. Vykonajte hodnotenie na doménovo relevantných dátach a scénároch.
3. Nastavte monitoring na zachytenie *real-world* incidentov v produkcii.
4. Zapojiť spätnú väzbu od koncových užívateľov pravidelne.
5. Implementujte RAG, alebo iné retrieval systémy, aby sa minimalizovali halucinácie.
6. Pokračujte v fine-tuningu a model aktualizujte podľa produkčných poznatkov.

Záver

Meranie spoľahlivosti AI systému bez klamania si metrikami je výzvou, ktorá si vyžaduje komplexný prístup kombinujúci:

- pochopenie mechanizmov vzniku AI halucinácií,
- nasadenie techník ako RAG na dodanie overiteľných podkladov,
- fine-tuning modelov na špecifické doménové dáta,
- a hlavne vytvorenie produktového monitoringu s KPI orientovanými na reálne riziká.

Iba tak vieme využiť AI systémy v oblastiach s vysokou zodpovednosťou bezpečne a efektívne.

Kde je zdroj a ako to testujeme? – Tento text stavia na poznatkoch z publikovaných štúdií o LLM (napr. OpenAI research), praktických aplikáciách retrieval-augmented generation (Lewis et al., 2020) a overených postupoch z oblasti AI incident managementu. Testovanie prebieha na benchmarkoch špecifických pre danú doménu a v reálnych produkčných prostrediach so zavedeným monitoringom incidentov a spätnou väzbou užívateľov.

