

Here's what kills me: v posledných rokoch umelá inteligencia a veľké jazykové modely (llm) významne menia spôsob, akým firmy pracujú s informáciami. Amazon Web Services (AWS) napríklad sľubuje verifikačnú presnosť až na úrovni 99 % v rámci svojich AI nástrojov, najmä keď využívame techniky ako Retrieval-Augmented Generation (RAG) alebo fine-tuning na doménových dátach. Avšak, kde je skutočná pravda? Kedy môžeme takýmto číslom veriť a kedy je naopak nutné byť veľmi opatrní?

Čo sú AI halucinácie a prečo vznikajú?

Halucinácie v kontexte veľkých jazykových modelov znamenajú generovanie nesprávnych alebo nepravdivých informácií, ktoré však znejú veľmi presvedčivo. Tieto chyby môžu spôsobiť, že model "vymýšľa" fakty, ktoré v tréningových dátach neexistujú, alebo nesprávne interpretuje dostupné znalosti.

Pravdepodobnosti vs fakty

- **LLM generujú obsah na základe pravdepodobností slov** – model predikuje, aké slovo alebo fráza s najväčšou pravdepodobnosťou nasleduje podľa vzoru v jeho tréningových dátach.
- **Model však nemá vlastné "pochopenie" faktov** – nerozoznáva pravdu od nepravdy, iba predikuje štatisticky vhodné výrazy.
- **Výsledkom môžu byť halucinácie – omyly alebo zatajené fakty**, ktoré môžu byť problémom hlavne v oblastiach, kde je presnosť kľúčová.

Preto sa viaceré spoločnosti snažia tieto pravdepodobnostné modely dopĺňať databázami faktov a systémami overovania, čo však stále nie je triviálny proces.

Kedy je riziko halucinácií najvyššie?

Nie všetky oblasti práce s AI majú rovnaký stupeň rizika spojený s halucináciami. Sú situácie, kde aj malé chybné informácie môžu mať závažné dôsledky:

1. **Právo:** Nesprávne interpretované zákony či právne fakty môžu viesť k chybám v rozhodnutiach alebo poradiach.
2. **Medicína:** Diagnózy, liečebné rady alebo výsledky výskumu nesmú byť založené na neoverených alebo nepresných údajoch.
3. **Finance:** Nesprávne finančné analýzy alebo predpovede môžu viesť k značným finančným stratám.
4. **Multi-step výskum:** Kumulatívne chyby pri zložitejších výskumných úlohách, kde výsledky jednej fázy ovplyvňujú ďalšie kroky, môžu eskalovať.

V týchto oblastiach je verifikačná presnosť zásadná a preto sú metódy na overovanie tvrdení nevyhnutné.



RAG a kvalita knowledge base – ako zvýšiť verifikačnú presnosť?

Retrieval-Augmented Generation (RAG) je prístup, ktorý kombinuje generatívne schopnosti LLM so schopnosťou vyhľadávať relevantné informácie z externej databázy (knowledge base). Ako to pomáha?

- **Zníženie halucinácií:** Model sa neodvoláva len na interné váhy a pravdepodobnosti, ale čerpá z overených zdrojov.
- **Kvalita knowledge base:** Kľúčová je relevantnosť a aktuálnosť dát v knowledge base – čím kvalitnejšie sú dáta, tým lepšia je verifikačná presnosť.
- **RAG umožňuje rýchle doplnenie nových znalostí** bez nutnosti rozsiahleho trénovania modelu od nuly.

V praxi je však nutné neustále monitorovať kvalitu knowledge base, pretože ak sú v nej chyby alebo zastarané informácie, model ich môže potenciálne "preklopiť" do generovaných odpovedí.

Fine-tuning na doménových dátach

Fine-tuning znamená ďalšie tréningové kolo modelu na špecifických dátach, ktoré sa blížia k cieľovej doméne použitia. Toto vedie k:



- **Vyššej špecializovanosti modelu** – lepšie chápe terminológiu, vzory aj špecifické pravidlá.
- **Zníženiu halucinácií** v rámci domény, pretože model "vie", aké informácie sú relevantné a oveľa častejšie vyberá fakty z tréningových dát.
- **Obmedzeniu generalizácie**, ktorá môže byť na škodu práve pri špecifických úlohách.

Odporúčam však nechodiť "na doraz" a vždy model testovať – fine-tuning nerieši problém halucinácií úplne, ale iba ho výrazne znižuje v rámci domény.

Produkčné testovanie a overovanie tvrdení – kde je zdroj a ako testovať?

Bez produkčného testovania a monitoringu je nereálne veriť deklarovanej verifikačnej presnosti, aj keď AWS či iné spoločnosti uvádzajú vysoké čísla. Čo zahŕňa kvalitný pipeline overovania:

1. **Jasné metriky a definície:** Čo znamená "99 % verifikačná presnosť"? Aký je konkrétny dataset, podľa čoho sa hodnotila?
2. **Benchmark dátové sady:** Použiť odvodené a verejne známe dataset-y, kde vieme presne, čo je pravdou.
3. **A/B testovanie:** Porovnanie modelu v produkcii oproti baseline alebo alternatívnym riešeniam.
4. **Monitorovanie v reálnom čase:** Logging omylov, žiadostí o spätnú väzbu od užívateľov a automatizované detekcie nereálnych odpovedí.
5. **Iteratívna aktualizácia knowledge base:** Zpracovanie nových, overených dát na základe feedbacku z produkcie.

Mini-checklist pre praktické overovanie verifikačnej presnosti:

- Máme jasný dataset s pravdivými odpoveďami.
- Testujeme model na nezávislej sade mimoriadne podrobne.
- Vyhodnocujeme nielen presnosť, ale aj typy chýb (fakty vs. štýl).
- Implementujeme monitoring v produkčnom prostredí.
- Pravidelne aktualizujeme dáta a model podľa zistených chýb.

Záver: Kedy môžeme veriť deklarovanej presnosti AI a kedy nie?

99 % verifikačná presnosť v AI nástrojoch ako tie od AWS je technicky možná, ale iba za špecifických podmienok. Musíme sa pýtať:

- Kde je **zdroj údajov**, na ktorých bola táto presnosť nameraná?
- Ako presne prebiehalo **testovanie v produkčnom prostredí**?
- Akým spôsobom je **riešený problém halucinácií** v našej konkrétnej doméne?
- Je k dispozícii kvalitný systém **RAG alebo fine-tuning**, ktorý zabezpečuje adaptáciu na reálne znalosti?

V <https://www.invelity.com/blog/ako-znizit-halucinacie-ai/> oblastiach s vysokým rizikom ako právo, medicína či financie by sme nemali prijímať čísla bez overenia a dôkladného testovania. Komplexný prístup zahŕňajúci RAG, kontrolu knowledge base a fine-tuning na doménových dátach je základom pre dôveru v AI systémy. Bez systému produkčného testovania a monitoringu je takáto dôvera naivná a môže viesť k závažným následkom.

Preto je moje odporúčanie: **neprijímať verifikačnú presnosť na slovo**, žiadať zdroje, jasné procesy testovania a režim, ako sa tieto čísla overujú v reálnych podmienkach. Len tak môžeme bezpečne využiť potenciál AI vo firmách a kritických oblastiach.