

V ére umelej inteligencie založenej na veľkých jazykových modeloch (LLM) sa stále intenzívnejšie diskutuje o ich aplikácii v špecializovaných oblastiach. Firmy často stojí pred otázkou, či investovať do **fine-tuningu na doménových dátach** alebo zvoliť modernejšie prístupy ako **RAG (Retrieval-Augmented Generation)**. V tomto článku rozoberieme, čo sú *AI halucinácie*, prečo LLM halucinujú, kedy je riziko týchto halucinácií najvyššie a ako rozhodnúť, či je doménový fine-tuning opodstatnený alebo nie.

Čo sú AI halucinácie a prečo LLM halucinujú?

AI halucinácie sú situácie, keď LLM generuje niečo, čo vyzerá ako pravda, ale v skutočnosti je nepravdivé alebo mylné. Ide o jednu z najväčších výziev pri praktickom využití týchto modelov.

Mechanizmus halucinácií: pravdepodobnosti vs. fakty

LLM sú jazykové modely trénované na obrovských korpusoch textu, ktoré sa učia produkovať *pravdepodobnostne najpravdepodobnejšie následné tokeny* v sekvencii. Model teda *nevie priateľ faktov* tak, ako databáza, ale pracuje s pravdepodobnosťami.

- **Žiadnu explicitnú databázu faktov:** LLM neukladajú fakty ako v relačnej databáze, ale rozumejú vzorcom a štatistikám v dátach.
- **Odchýlky a nepresnosti:** Pri nekonzistentných alebo menej zastúpených informáciách môže dôjsť k omylom.
- **Výrazné riziko pri komplexnejšom uvažovaní:** Multi-step logické úlohy alebo zložité odborné témy môžu viesť k vyššiemu počtu halucinácií.

Kedy je riziko halucinácií najvyššie?

Halucinácie môžu byť nepríjemné v bežnom užívaní, no v niektorých oblastiach môžu mať fatálne dôsledky. Last month, I was working with a client who learned this lesson the hard way.. (my cat just knocked over my water). Preto je dôležité vedieť, kedy treba byť zvlášť opatrný.

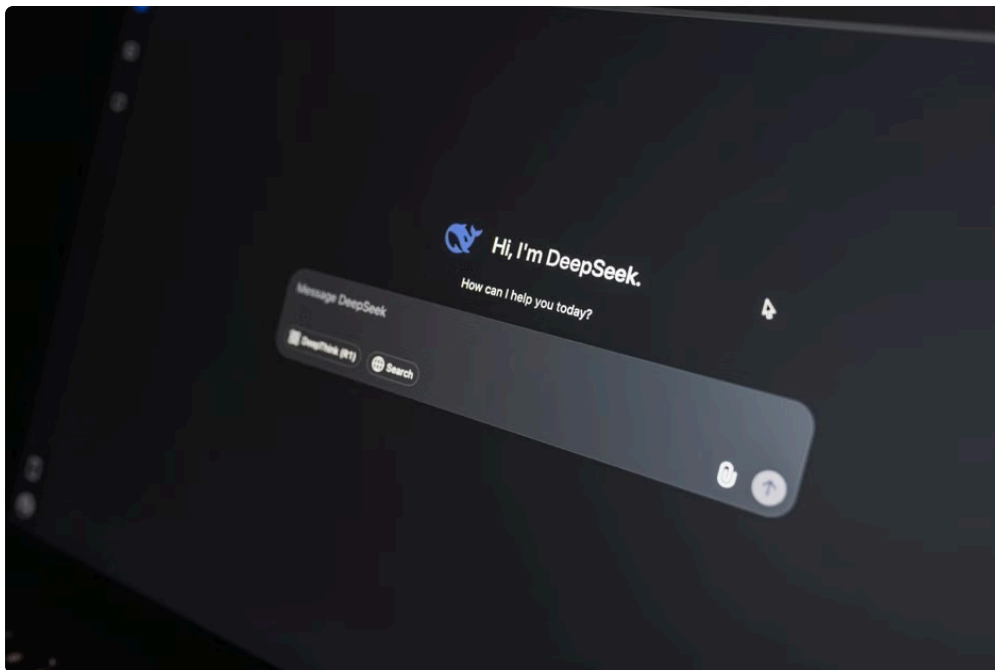
1. **Právo:** Nepravdivé právne rady môžu viesť k závažným problémom so zákonom.
2. **Medicína:** Nesprávna diagnóza alebo odporúčanie môže ohroziť zdravie a život.
3. **Financie:** Chyby pri finančných analýzach môžu viesť k stratám alebo právnym postihom.
4. **Multi-step výskum a analýzy:** Pri dlhších reťazcoch záverov sa halucinácie kumulujú, čím rastie riziko chýb.

Mini-checklist rizikových situácií:

- Oblasť so silným právno-štandardným rámcom
- Vyžaduje sa vysoká presnosť a verifikovateľnosť informácií
- Odborné alebo vedecké kontexty s nutnosťou transparentnosti zdrojov

Čo je RAG a ako zvyšuje kvalitu knowledge base?

RAG (Retrieval-Augmented Generation) je architektúra kombinujúca klasické LLM so systémom na vyhľadávanie faktov z externe spravovaných knowledge base (napr. interné dokumenty, zmluvy, vedecké štúdie).



RAG funguje v dvoch krokoch:

invelity.com

1. Dokument retrieve: Systém vyhledá relevantné kúsky z databázy/knowledge base na základe otázky.
2. Generation: LLM použije tieto overené fakty na generovanie spoľahlivej odpovede.

Výhody RAG oproti čistému LLM alebo doménovému fine-tuningu:

- **Flexibilita:** Knowledge base nie je potrebné meniť pri nových dátach, stačí aktualizovať indexing.
- **Nižšia miera halucinácií:** Model vychádza z overených vstupov.
- **Nižšie náklady:** Pri veľkých dátach nie je potrebné každý krát retrenovať model, čo je finančne a časovo náročné.

Doménový fine-tuning: kedy áno a kedy nie?

Doménový fine-tuning znamená dodatočné tréningovanie veľkého modelu na špecifické doménové dáta, aby sa model správal špecializovane a lepšie rozumel terminológii, zvyklostiam alebo štruktúre textov v danej oblasti.

Kedy dáva zmysel doménový fine-tuning?

- **Úzka, špecifická doména:** Ak štandardné LLM nedokážu správne pochopiť alebo generovať obsah (napr. veľmi technické IT, farmaceutický výskum).
- **Kontrola nad jazykovým štýlom:** Ak je kritický presný terminologický alebo štýlový súlad (napr. interné značky alebo citlivé texty).
- **Obmedzený knowledge base:** Ak vaša databáza záznamov je malá a obsahuje presne definované štruktúrované dáta, ktoré zvyšujú presnosť modelu.
- **Vysoké regulačné nároky:** Kde musíte garantovať presnosť a auditovateľnosť výsledku na úrovni samotného modelu.

Kedy je fine-tuning overkill?

- **Veľké a často sa meniace knowledge base:** Vtedy RAG umožní oveľa flexibilnejšiu a lacnejšiu aktualizáciu obsahu bez potreby retrainovania modelu.
- **Veľké náklady na tréning:** Fine-tuning veľkých modelov (zo série GPT-3, GPT-4, PaLM) je finančne a infraštruktúrne náročný.
- **Riziko degradácie generalizácie:** Príliš silný fine-tuning môže spôsobiť, že model stratí schopnosť odpovedať na otázky mimo domény.
- **Monitorovanie a rollback:** Bez kvalitného monitoringu a testovania môže byť fine-tuning zdrojom produkčných incidentov.

Náklady fine-tuningu vs. prínosy

Kriterium Fine-tuning RAG Počiatočné náklady Vysoké (výpočtová kapacita, dátová príprava) Nízke (nastavenie indexov, vyhľadávanie) Aktualizácia znalostnej bázy Vyžaduje retrainovanie alebo dodatkové techniky Jednoduchá aktualizácia indexov Flexibilita použitia Špecializované domény, stratí generalizáciu Vhodné pre široké spektrum domén Riziko halucinácií Mierne znížené, ale model stále generuje pravdepodobnosti Výrazne znížené, vychádza z overených zdrojov

Mini-checklist: Ako rozhodnúť?

1. **Máte dobre pokrytú knowledge base**, ktorú chcete výrazne využiť? Zvážte RAG.
2. **Potrebujete vysoko špecializované správanie modelu** s presným jazykom? Fine-tuning môže pomôcť.
3. **Máte finančné a technologické možnosti riadne spravovať fine-tuning** vrátane monitoringu? Ak nie, RAG je bezpečnejšia voľba.
4. **Vaša aplikácia nesmie tolerovať chyby** (medicína, právo, financie)? Kombinujte doménový fine-tuning s RAG a robustným monitoringom.

Záver

Here's what kills me: doménový fine-tuning nie je všeliek, ale cenný nástroj, ktorý má svoje miesto v konkrétnych prípadoch. Naopak, moderné techniky ako RAG umožňujú využívať LLM efektívnejšie a bez nutnosti neustáleho nákladného pretrainovania. Kľúčové je vždy rozumieť zdrojom dát, potenciálnym rizikám halucinácií a mať pripravený plán testovania a monitoringu nasadených riešení.



Bez zdrojov a testovacieho prostredia sa nikdy nepúšťajte do produkcie so špecializovaným LLM – to je mantra, ktorú by si mal každý AI tím osvojiť.