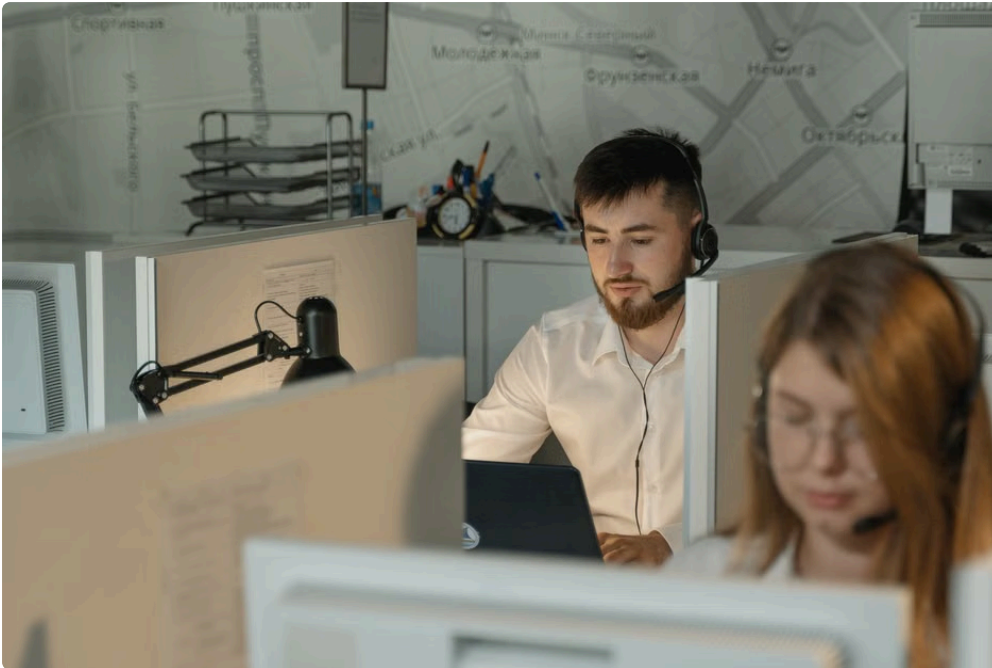


V dobe, keď sa umelá inteligencia a najmä veľké jazykové modely (LLM) stávajú bežnými nástrojmi pre získavanie informácií a podporu rozhodovania, otázka dôveryhodnosti odpovedí naberať nový význam. Ako zabezpečiť, aby každý dôležitý fakt v odpovedi mal svoj zdroj? Prečo LLM často „halucinujú“ (vymýšľajú si fakty) a ako to riešiť pomocou moderných techník ako **Retrieval-Augmented Generation (RAG)** a **fine-tuning** na doménových dátach? Tento blog sa detailne venuje problematike vyžadovania *citácií* a garantovaniu *grounded odpovedí*.

Čo sú AI halucinácie a prečo LLM halucinujú?

Termín **halucinácia** v oblasti LLM označuje situáciu, kedy model generuje informácie, ktoré nie sú pravdivé alebo existujú len v modelovej špekulácii, bez opory v skutočných dátach.



Pravdepodobnosti vs fakty

- **Generatívny princíp LLM:** Modely sú trénované na predikcii pravdepodobnosti nasledujúceho slova na základe predchádzajúceho kontextu.
- **Nepretržitá generalizácia:** LLM nepoužívajú explicitnú databázu faktov. Namiesto toho extrapolujú na základe štatistík z tréningových dát.
- **Neistota v dátach:** Ak model nemá natívny prístup k overeným faktom počas generovania, môže zvoliť pravdepodobnostne vhodný, no nesprávny fakt.

Preto halucinácie nie sú chybou softvéru v tradičnom zmysle, ale logickým dôsledkom spôsobu, akým LLM fungujú.

Kedy je riziko halucinácií najvyššie?

Nie všetky oblasti umožňujú „nejakú“ chybu. Existujú domény, kde nesprávne informácie môžu mať kritické dôsledky.

- **Právo:** Nesprávna interpretácia paragrafov alebo právnych precedентов môže znamenať právne riziko alebo nesprávne rozhodnutie.

- **Medicína:** Falošne uvedené liečebné postupy alebo diagnózy môžu ohroziť život pacienta.
- **Financie:** Nesprávne finančné dáta alebo predpovede môžu viesť k značným stratám a nesprávnym investíciám.
- **Multi-step výskum:** Pri zložitej syntéze informácií z viacerých zdrojov môže halucinácia spôsobiť lavínu nesprávnych záverov.

Prečo vyžadovať citácie? Výhody a zásadný význam

Citácie, resp. zdrojové odkazy, sú kľúčové pre overiteľnosť každej informácie. Výhody techniky **requiring citations** v odpovediach sú:

1. **Zvýšená dôveryhodnosť:** Používateľ presne vie, odkiaľ informácia pochádza.
2. **Možnosť overenia:** Zdrojové odkazy umožnia jednoducho validovať správnosť tvrdenia.
3. **Prevenca halucinácií:** Model je nútený vyhľadávať fakty, než generovať voľné špekulácie.
4. **Transparentnosť:** Odpoveď sa stáva auditovateľnou pre interný aj externý monitoring.

Technické riešenia: RAG a fine-tuning na doménových dátach

Retrieval-Augmented Generation (RAG)

RAG predstavuje moderný prístup, ktorý kombinuje **retrieval** (vyhľadávanie relevantných dokumentov) s **generatívnym modelom**:

- **Proces:**
 1. Užívateľ položí otázku.
 2. Retrieval modul vyhľadá relevantné dokumenty z knowledge base alebo interných zdrojov.
 3. Generatívny LLM použije tieto dokumenty ako kontext pre zodpovedanie otázky.
 4. Výstup obsahuje odpoveď i odkazy na zdroje.
- **Výhoda:** Generované texty sú „grounded“ v reálnych dátach, čím výrazne klesá riziko halucinácií.

Fine-tuning na doménových dátach

Fine-tuning znamená dodatočné trénovanie modelu na špecifických dátach z danej oblasti (právo, medicína, financie), pričom model získava špecifickú terminológiu a faktické znalosti.

- **Zvýšená presnosť:** Model lepšie rozumie kontextu a terminológii domény.
- **Často kombinované s RAG:** Fine-tuned model dokáže lepšie spracovať vyhľadané dokumenty a vytvoriť kvalitnejšiu odpoveď.
- **Možnosť vynucovania citácií:** Model môže byť trénovaný na šablóny výstupov, ktoré vyžadujú explicitné zdrojovanie.

Mini-checklist: ako v praxi vynútiť citácie v odpovediach LLM

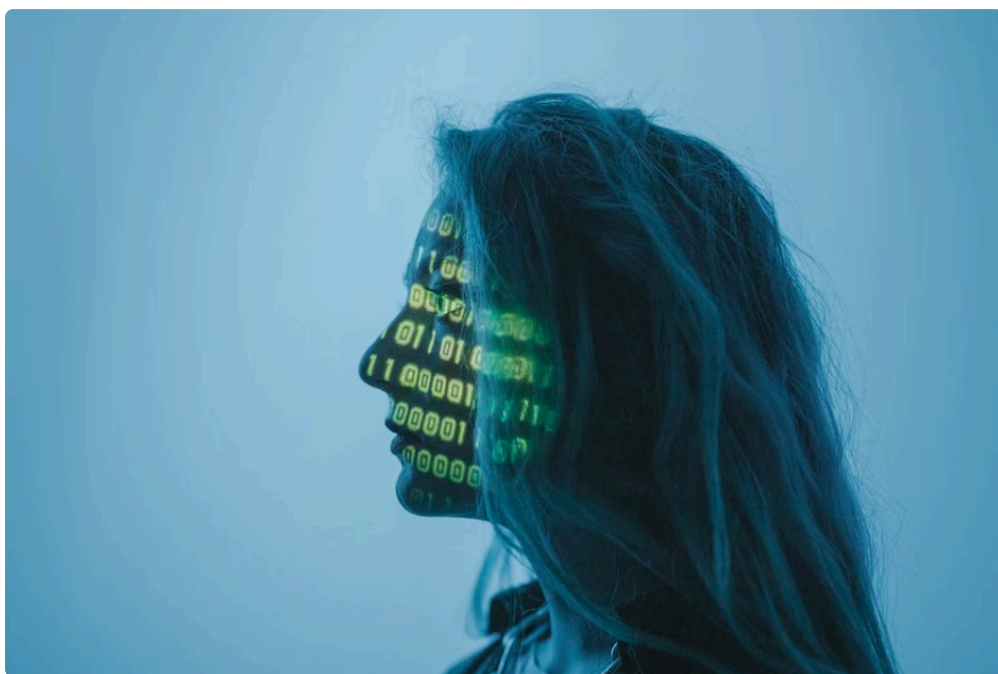
1. **Zabezpečiť kvalitnú knowledge base:** obsahuj kvalitné, validované a aktuálne zdroje súvisiace s doménou používania.

2. **Implementovať retrieval modul:** vybrať robustný vyhľadávací engine (Elasticsearch, Pinecone, Weaviate a pod.), ktorý rýchlo nájde relevantné dokumenty.
3. **Prepojiť retrieval s generáciou (RAG):** generátor odpovedí musí mať prístup ku kontextu vyhľadaných dokumentov a používať ich ako referenciu.
4. **Upraviť prompt:** explicitne požiadať model, aby ku každej dôležitej informácii vždy pridal zdrojový odkaz.
5. **Fine-tuning:** využiť doménové dáta na učenie modelu správneho formátovania odpovedí so zdrojmi.
6. **Monitoring a validácia:** sledovať a testovať frekvenciu a kvalitu citácií v produkčných odpovediach.
7. **Užívateľské rozhranie:** zobrazovať citácie pre používateľov v prehľadnej forme, napr. klikateľné odkazy, poznámky pod čiarou.

<https://www.invelity.com/blog/ako-znizit-halucinacie-ai/>

Záver: Bez citácií to nejde

Halucinácie veľkých jazykových modelov nie sú iba vecou improvizácie, ale pri niektorých použitých prípadoch doslova hrozbou. Preto je nevyhnutné pristupovať k odpovediam LLM kriticky a vždy



- žiadame **requiring citations**,
- využívame techniky **retrieval-augmented generation**,
- a **finetunujeme modely** na overené doménové dáta.

Tak môžeme výrazne znížiť nebezpečenstvo nesprávnych tvrdení a zvýšiť spoľahlivosť automatizovaných odpovedí, čo je základným predpokladom pre nasadenie AI riešení do kritických oblastí.

Autor: Expert na LLM riešenia s 9-ročnou praxou v digitálnych produktoch a 4-ročným skúsenosťami s AI tímami, produkčnými chatbotmi a knowledge base systémami.