

V ére, keď veľké jazykové modely (LLM) a pokročilé metódy ako Retrieval-Augmented Generation (RAG) a fine-tuning na doménových dátach zásadne menia spôsob, akým podniky a inštitúcie pracujú s informáciami, je nevyhnutné pochopiť ich limity a riziká. Práve na to sa zameriame v tomto článku. Zodpovieme si otázky, čo sú AI halucinácie, prečo k nim dochádza, kedy je riziko ich dopadu najvyššie a prečo by niektoré kľúčové use-cases mali mať povinné schválenie človekom.

Čo sú AI halucinácie a prečo LLM halucinujú?

„AI halucinácie“ sú fenomén, pri ktorom LLM generuje obsah, ktorý nie je založený na faktoch, ale vzniká z modelovej pravdepodobnosti slovných sekvencií. Inými slovami, model namiesto presnej informácie vytvorí niečo, čo znie presvedčivo, ale je nepravdivé alebo neexistujúce.

Mechanizmus halucinácií v LLM

- **Pravdepodobnosti namiesto faktov:** LLM sú trénované na obrovskom množstve textov, kde učenie prebieha cez odhad pravdepodobnosti nasledujúceho slova na základe kontextu. Tento prístup nezaručuje, že výstup vždy presne zodpovedá realite.
- **Nedostatočná faktická konzistencia:** Model neobsahuje „databázu faktov“, ale vzory jazyka. Ak sa model stretne s otázkou nad rámec svojho trénovacieho korpusu alebo je otázka veľmi špecifická, často vygeneruje „najpravdepodobnejší“ text, ktorý nemusí byť pravdivý.
- **Neurčité a nejednoznačné otázky:** Pri nejasných vstupoch môže model extrapolovať a vytvoriť svoj vlastný záver, čo často vedie k halucináciám.

Kedy je riziko halucinácií najvyššie?

Nie všetky use-cases trpia rovnakým rizikom. Podstatné je posúdiť dôsledky nesprávnej alebo mylnej informácie.

Use-cases s vysokým rizikom

1. Právne odpovede

Právne poradenstvo vyžaduje precíznosť a aktuálnosť. Nesprávne informácie môžu viesť k finančným stratám, porušeniu zákonov alebo aj súdnym sporom.

2. Finančné rozhodnutia

Investičné odporúčania alebo poradenstvo týkajúce sa daní či úverov ovplyvňujú finančnú stabilitu osôb a firiem. Nesprávne alebo zavádzajúce informácie môžu mať katastrofálne následky.

3. Zdravotnícke odporúčania

Poskytovanie lekárskeho odporúčania, diagnóz alebo liečebných plánov vyžaduje odbornosť a kontrolu kvality. Halucinácie tu môžu ohroziť zdravie a život pacientov.

4. Multi-step výskum a odborné analýzy

Komplexné výskumné projekty, ktoré vyžadujú postupné overovanie informácií a zdrojov, môžu byť zničené jedinou chybnou či halucinovanou informáciou.

Rola RAG a fine-tuningu na doménových dátach pri redukcii rizika

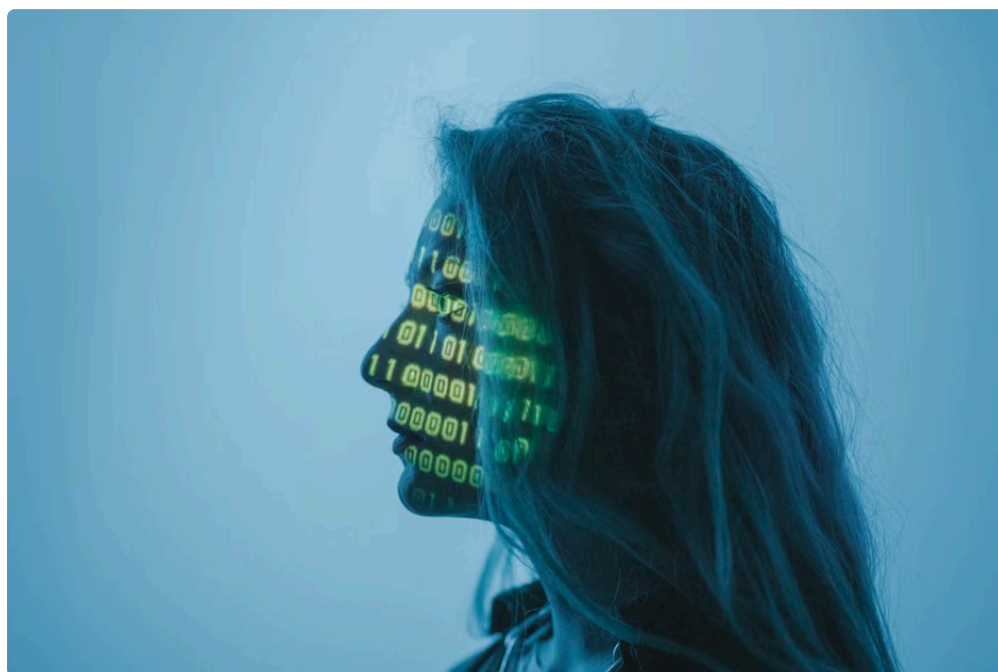
Moderné prístupy ako Retrieval-Augmented Generation (RAG) a fine-tuning na doménových špecifických dátach dokážu zlepšiť kvalitu a spoľahlivosť odpovedí LLM, avšak neeliminuje potrebu ľudského overenia v kritických oblastiach.

Retrieval-Augmented Generation (RAG)

- **Fungovanie:** RAG kombinuje tradičné vyhľadávanie v knowledge base s generovaním odpovedí, čím umožňuje prístup ku konkrétnym a aktuálnym faktom počas generovania textu.
- **Kvalita knowledge base:** Kľúčom k zníženiu halucinácií je kvalitná, aktualizovaná a konzistentná databáza faktov, ktorá je základom pre retrieval. Ak knowledge base obsahuje nepresné alebo neúplné informácie, model môže halucinovať aj na základe týchto údajov.
- **Limitácie:** RAG je takisto závislý na správnosti a pokrytí knowledge base. V nových alebo nepreskúmaných oblastiach môže dochádzať k výpadkom.

Fine-tuning na doménových dátach

Fine-tuning znamená doladenie LLM pomocou dát špecifických pre daný odbor alebo spoločnosť. Toto nasadenie má tieto výhody:



- Zvýšenie presnosti a relevantnosti odpovedí.
- Minimalizácia generovania nesprávnych alebo nezmyselných odpovedí.
- Väčšia schopnosť porozumieť kontextu špecifického odboru.

Avšak, ani fine-tuning samotný neodstraňuje riziko halucinácií. Model sa stále riadi pravdepodobnosťami a preto sa odporúča kombinovať fine-tuning s RAG a povinným ľudským overením v kritických scenároch.



Ktoré use-cases by mali mať povinné schválenie človekom? Mini-checklist

Pri posudzovaní, kedy je nevyhnutné zapojiť človeka do procesu validácie odpovedí AI, odporúčam použiť nasledujúci checklist:

Kritérium Popis Odôvodnenie Vysoké dopady rozhodnutia Finančný, právny, zdravotný alebo reputačný zásah. Chyba môže viesť k závažným stratám alebo škodám. Potreba presnosti a konzistencie Odpoveď musí byť fakticky overiteľná a jednoznačná. Nepripustiteľnosť neurčitosti alebo nejasností. Dlhodobé právne alebo regulačné záväzky Možnosť postihu za chybné informácie. Vyžaduje auditovateľnosť a transparentnosť. Viacstupňový rozhodovací proces Zložité výskumné alebo analytické úlohy vyžadujúce validáciu viacerých zdrojov. Znižuje riziko kumulatívnych chýb či halucinácií.

Záver

LLM, RAG a fine-tuning sú silné nástroje, ktoré menia pravidlá hry v spracovaní a generovaní informácií. Napriek tomu je potrebné realisticky pristupovať k ich limitáciám, najmä v témach s vysokým rizikom, ako sú **právne odpovede, finančné rozhodnutia a zdravotnícke odporúčania**. V týchto oblastiach by mali byť všetky automatizované výstupy podmienené *povinným schválením človekom* – expertom, ktorý má kompetenciu overiť kvalitu a faktickú správnosť odpovedí.

Praktický rekomendovaný postup:

1. Implementujte RAG na vyhľadávanie faktov v kvalitnej a pravidelne aktualizovanej knowledge base.
2. Fine-tunujte model na doménových dátach, aby lepšie rozumel špecifikám a terminológií.
3. Nasadzujte systém s monitoringom nepresností a možných halucinácií.
4. Zavedte workflow s jasným manuálnym schválením pre kritické odpovede.

Týmto spôsobom maximalizujete prínos AI pri minimalizácii rizík pre koncového používateľa aj jeho organizáciu.

Zdroj a testovanie

- Na pochopenie halucinácií odporúčam štúdie od OpenAI (<https://openai.com/research/>).
- Testovanie spoľahlivosti RAG a fine-tuning modelov vykonávajte na reálnych doménových dátach so známou pravdou (gold standard dataset), prípadne s odbornou verifikáciou.
- Monitorujte produkčné chatboty pravidelnými auditmi odpovedí a zaznamenaním prípadných halucinácií pre spätné učenie modelu.