

V dnešnej dobe sme svedkami explózie nástrojov založených na veľkých jazykových modeloch (LLM), ktoré dokážu odpovedať na otázky, tvoriť texty či pomáhať v rôznych oblastiach biznisu. Napriek ich úžasným schopnostiam však narážame na fenomén tzv. **AI halucinácií** — generovanie nesprávnych, a často presvedčivých odpovedí, ktoré nie sú založené na faktoch. Preto je kľúčové, aby systémy pri nejasnostiach nepodávali falošné informácie, ale používali *fallback odpovede*, ktoré zachovávajú dôveru, komunikujú bezpečne a priznávajú limity AI.

Čo sú AI halucinácie a prečo LLM halucinujú?

Halucinácia je v kontexte LLM stav, kedy model odpovie plynule a sebavedome, no poskytuje informácie, ktoré sú nepravdivé, nezmyselné alebo nemožné overiť. Na rozdiel od chýb tradičných systémov tu nie je explicitná programátorská chyba, ale problém princípu fungovania modelu.

Model vs fakt

Veľké jazykové modely sú v podstate *probabilistické generátory textu*. Na základe štatistík a pravdepodobností vyberajú slová, ktoré sú najpravdepodobnejšie v danom kontexte, pričom nemajú priamy prístup k databáze faktov ani nie sú schopné logicky overiť pravdivosť vyjadrení. Preto vzniká riziko, že model vytvorí *kombináciu slov, ktorá znie dobre, no je nesprávna*.

Typické príčiny halucinácií

- **Neúplné trénovacie dáta:** model nemusí poznať špecifické a aktuálne fakty.
- **Všeobecný tréning:** bez doménového zamerania môže generovať informačne nepresné odpovede v špecializovaných oblastiach.
- **Kontextová nejednoznačnosť:** viacznačné vstupy vedú k nesprávnej interpretácii.

Kedy je riziko halucinácií najvyššie?

Nie všetky aplikácie LLM si môžu dovoliť chyby rovnakej vážnosti. Riziko pôsobenia nesprávnych informácií je kritické v týchto oblastiach:

1. **Právo:** nesprávne právne rady môžu viesť k súdnym problémom a porušeniam predpisov.
2. **Medicína:** chybná diagnóza alebo odporúčanie liečby ohrozujú zdravie a životy pacientov.
3. **Financie:** nesprávne údaje môžu spôsobiť finančné straty či porušenia regulácií.
4. **Multi-step výskum:** nesprávne podklady môžu narušiť vedeckú integritu a výsledky.

Preto pri týchto prípadoch je nutné aplikovať metódy, ktoré znižujú pravdepodobnosť halucinácií a zároveň zaisťujú, že v prípade neistoty AI neklame.

RAG a fine-tuning ako riešenia pre lepšiu verifikáciu odpovedí

Retrieval-Augmented Generation (RAG)

RAG kombinuje výhody tradičného vyhľadávania s generatívnym modelom. Model najskôr vyhledá relevantné dokumenty v knowledge base podľa otázky, následne na základe týchto zdrojov vygeneruje odpoveď. Last month, I was working with a client who was shocked by the final bill.. Výsledkom je, že odpoveď je priamo naviazaná na overiteľné údaje, čím výrazne klesá riziko halucinácií.

Prečo je RAG dôležité?

- Znižuje neistotu LLM tým, že mu dáva konkrétne zdroje.
- Umožňuje transparentne ukázať, odkiaľ pochádza odpoveď.
- Pomáha pri aktualizácii informácií bez nutnosti kompletne trénovať model nanovo.

Fine-tuning na doménových dátach

Aj napriek použitiu RAG môže byť užitočné doladiť model na špecifických dátach firmy či odvetvia. Fine-tuning znamená, že model sa trénuje ďalej na kvalifikovaných a overených doménových dátach, čo mu zlepšuje porozumenie terminológie, kontextu a pravidiel daného odvetvia.



- Znižuje počet nesprávnych interpretácií špecifických pojmov.
- Zlepšuje konzistenciu odpovedí vo vnútri domény.
- Podporuje bezpečnú komunikáciu s používateľom.

Fallback odpoveď: základ bezpečnej komunikácie a priznania limitov AI

Fallback odpoveď je predpripravený text alebo stratégia, ktorú použijeme v prípadoch, keď AI nie je schopná zabezpečiť správnu, overiteľnú alebo spoľahlivú informáciu. Mnohé spoločnosti si neuvedomujú, že **fallback nie je prehra – je to esenciálna časť zodpovednosti** pri nasadzovaní AI riešení.



Mini-checklist pre dobrý fallback text

1. **Priznajte limitácie:** text explicitne hovorí, že AI nemusí poznať odpoveď alebo že táto téma je komplikovaná.
2. **Nevytvárajte dojem falošnej istoty:** vyhnite sa slovám ako „samozrejme“, „určite“ alebo „jednoznačne“, ak nie ste stopercentne istí.
3. **Navrhňte ďalší krok:** komunikujte, že otázka má byť overená človekom, prípadne odporučte konzultáciu so špecialistom.
4. **Zákazky vykazujte transparentnosť:** napríklad „Táto informácia momentálne nie je v mojej databáze dostupná.“
5. **Udržte tón dôveryhodnosti:** buďte úprimní, nie negatívni alebo tendenciozni.

Príklady fallback textov

Scenár Príklad fallback odpovede Medicínske dotazy „Ľutujem, ale na základe dostupných údajov a môjho tréningu nedokážem poskytnúť presnú odpoveď. Prosím, kontaktujte kvalifikovaného zdravotníckeho odborníka.“ Právne otázky „Táto otázka patrí do oblasti práva, kde je dôležité presné poradenstvo. Odporúčam konzultáciu s právnikom, ktorý pozná Váš konkrétny prípad.“ Technické špecifikácie mimo databázy „Momentálne nemám dostatok informácií na odpoveď. Skontrolujem aktualizácie a čoskoro Vám poskytnem presnejšie údaje.“

Testovanie a validácia fallback mechanizmu

Think about it: kde je zdroj fallback odpovede? [Automated Reasoning](#) fallback texty by mali vychádzať z krízového scenáru definovaného spolu s expertmi v danom odbore a výrazne testované na reálnych otázkach, ktoré model pravidelne nezodpovie správne.

Ako to otestujeme?

- Simulujte otázky, na ktoré model špecificky nevie odpovedať – sledujte, či padne fallback a nie halucinácia.
- Monitorujte produkčné nasadenie – zbierajte metriky „confidence score“ modelu, počet fallback odpovedí a spätnú väzbu užívateľov.
- Vyhodnocujte kvalitu knowledge base a jej schopnosť podporovať RAG.

- Aktualizujte both fallback texty aj doménové dáta podľa zistených medzier.

Záver: fallback odpoveď je nevyhnutná súčasť bezpečnej AI komunikácie

V nasadzovaní AI riešení založených na LLM je dôležité si uvedomiť, že **model nie je databázou faktov, ale pravdepodobnostný generátor textu**. Halucinácie sú jej inherentnou vlastnosťou. Z tohto dôvodu je bezpečná komunikácia, ktorá priznáva limity AI, kľúčová pre udržanie dôvery používateľov.

Fallback odpoveď nie je slabosťou systému, ale jeho zodpovednosťou. Správne navrhnuté fallback texty, podporené mechanizmami ako RAG a fine-tuning, pomáhajú minimalizovať riziko zavádzajúcich informácií, zvlášť v oblastiach, kde sú chyby neakceptovateľné. Vždy všetko testujeme, evidujeme a aktívne vylepšujeme.

Pre úspešné AI nasadenie sa teda nevyhneme nasledujúcim krokovám:

- Implementovať doménovo relevantnú knowledge base s pravidelnou aktualizáciou.
- Využiť RAG ako štandardnú vrstvu pred generovaním odpovedí.
- Fine-tunovať model na doménových dátach pre presnosť a kontext.
- Navrhnuť fallback mechanizmus, ktorý explicitne komunikovať neistoty a odporúča ďalšie overovanie.
- Systematicky testovať a monitorovať produkciu AI.

Takáto metodika je nevyhnutná pre zodpovedné a výročné využívanie AI technológií v kritických oblastiach dnešného sveta.