

V súčasnosti je Retrieval-Augmented Generation (RAG) jednou z najpopulárnejších metód, ako využiť veľké jazykové modely (Large Language Models – LLM) v kombinácii s externými databázami znalostí (knowledge base) na produkčné aplikácie. Mnohí však stále narážajú na problém halucinácií, teda prípadov, keď model generuje nesprávne alebo úplne vymyslené informácie, aj keď má dostupnú knowledge base. Čo stojí za týmito halucináciami? Prečo RAG nie je zázračné riešenie? A ako kvalita dát, relevance retrievalu a segmentácia dokumentov ovplyvňujú výsledky? Na to sa pozrieme v tomto článku.

Čo sú AI halucinácie?

AI halucinácia je jav, kedy generatívny model poskytne nepravdivé, nepresné alebo _vymyslené_ dáta, ktoré však na prvý pohľad pôsobia vierohodne a dôveryhodne. Halucinácie nie sú len drobné chyby – môžu spôsobovať závažné problémy najmä v kritických oblastiach, ako sú právo, medicína alebo financie.

Mini-checklist halucinácií:

- Generovanie faktov, ktoré neexistujú v dátach.
- Zmiešavanie alebo nesprávne interpretovanie znalostí.
- Vytváranie implikácií a súvislostí, ktoré nie sú podložené.
- Vysoká vierohodnosť textu napriek nepravdivosti obsahu.

Prečo LLM halucinuje? Pravdepodobnosti vs. fakty

Veľké jazykové modely fungujú na princípe predikcie pravdepodobnosti nasledujúceho tokenu (slova, znaku) vzhľadom na predchádzajúci kontext. Nie sú to databázy faktov alebo encyklopédie – nespracovávajú informácie v nejakej štruktúrovanej forme, ale modelujú jazykové vzory a štatistické korelácie.

To znamená, že model:



1. Vyberá slová, ktoré pravdepodobne najlepšie zapadnú do kontextu a sújazdnosti.
2. Neposkytuje garanciu pravdivosti, len pravdepodobnosť konzistentného textu.

3. Môže vytvoriť "falošné fakty", ak im štatisticky zodpovedajú daný kontext.

Tento princíp je často zdrojom halucinácií, najmä ak:

- Model nemá priamy prístup k relevantným dátam.
- Výcvikové dáta sú limitované alebo neúplné pre danú doménu.
- Model je nasadený v kontexte, kde vyžaduje multi-step logické uvažovanie podľa faktov.

Kontrast k RAG:

RAG sa snaží túto limitáciu LLM obísť tým, že do procesu generovania zapája vyhľadávanie (retrieval) relevantných dokumentov z knowledge base, čím model dostáva kontextové fakty. To však neznamená, že halucinácie úplne zmiznú – práve tu sa začína dôležitosť kvality knowledge base a toho, ako systém tieto dáta spracováva.

Kedy je riziko halucinácií najvyššie?

Mieru a kritickosť halucinácií ovplyvňuje domena použitia a charakter otázok. Najrizikovejšie oblasti sú:

Oblasť Dôvod vysokého rizika Typický príklad halucinácie Právo Nepresné právne rady môžu viesť k právnym sankciám a stratám. Model vymyslí neexistujúci paragraf alebo neplatnú výkladovú prax. Medicína Chyby v diagnostike alebo doporučeníach môžu ohroziť zdravie pacientov. Model uvedie neaktuálne liečebné postupy alebo kontraindikácie. Financie Nesprávne finančné poradenstvo môže viesť k značným finančným stratám. Model nesprávne interpretuje trhové dáta alebo kalkuluje riziká. Multi-step výskum a analýza Zložité logické závery sú náchylné na chyby, ak nie sú podložené faktami. Model nesprávne syntetizuje informácie z viacerých zdrojov, vytvára nepravdivé závery.

RAG a kvalita knowledge base – kde je problém?

Práve v kvalite a štruktúre knowledge base sa skrýva kľúč k minimalizácii halucinácií. Kvalita dát RAG ovplyvňuje, čo model dostane ako informáciu na základe retrievalu. Ak retrieval systém prináša irelevantné, neaktuálne alebo nekonzistentné dokumenty, LLM môže stále halucinovať, pretože "fakty" sú chybné alebo zavádzajúce.

Hlavné aspekty kvality dát v RAG:

1. Relevantnosť retrievalu:

- Dokumenty musia byť relevantné a obsahovo zodpovedať položeným otázkam.
- Pre zvýšenie presnosti je potrebné testovať retrieval systém pomocou hodnotení na ukázkových dotazoch.
- Priestor na zlepšenie – napríklad multi-hop retrieval, rozdelenie dotazu do viacerých častí.

2. Segmentácia dokumentov:

- Zmysluplné a koherentné potomky (chunks) dokumentov pomáhajú LLM lepšie spracovať a interpretovať kontext.
- Príliš veľké alebo príliš malé segmenty môžu viesť k strate dôležitých detailov alebo preťaženiu modelu.
- Kvalitná segmentácia by mala rešpektovať prirodzené štruktúry textu (odstavce, sekcie).

3. Aktualizácia a konzistencia dát:

- Zastaralé alebo protichodné informácie zvyšujú riziko halucinácie.
- Proces pravidelnej aktualizácie knowledge base a verifikácie zdrojov je kritický.

Fine-tuning na doménových dátach – riešenie alebo doplnok?

Fine-tuning LLM na špecifických doménových dátach predstavuje ďalší stupeň, kde model získava hlbšie pochopenie relevantných konceptov a terminológií. To výrazne zlepšuje šancu, že model bude generovať presnejšie a konzistentnejšie odpovede. Avšak:

- Fine-tuning nie je náhrada za kvalitný retrieval (kvalita dát RAG musí zostať prioritou).
- Model môže stále halucinovať, ak narába s nejednoznačnými alebo neúplnými dátami.
- Fine-tuning vyžaduje rozsiahle množstvo dobre označených a validovaných dát z danej domény.

Preto je ideálne kombinovať **fine-tuning** a **RAG** – fine-tuning zdokonalí jazykové schopnosti a doménové znalosti modelu, zatiaľ čo RAG zabezpečí aktuálny a faktický kontext z knowledge base.

Mini-checklist: Ako znížiť halucinácie v RAG systémoch

1. Skontrolovať a zvýšiť kvalitu knowledge base:

- Odstrániť neaktuálne a nekonzistentné dokumenty.
- Pravidelne aktualizovať a dopĺňať dáta.

2. Optimalizovať retrieval systém:

- Testovať a doladovať relevantnosť retrievalu.
- Implementovať viacúrovňové vyhľadávanie a filter výsledkov.

3. Premyslená segmentácia dokumentov:

- Deliť dáta podľa logických častí a kontextu hyperparametrov modelu.

4. Fine-tuning na doménových údajoch:

- Zlepšiť schopnosť modelu rozumieť špecifickým pojmom a konceptom.
- Znížiť množstvo neúplných alebo chýbajúcich znalostí.

5. Implementácia dôveryhodnosti a monitoringu:

- Sledovať produkčné halucinácie pomocou užívateľskej spätnej väzby a metrík.
- Využiť scoring relevance dokumentov a confidence skóre modelu.

Záver

Halucinácie v RAG systémoch nie sú výsledkom poruchy alebo chyby technológie, ale prirodzenou dôsledkom spôsobu, akým veľké jazykové modely pracujú s pravdepodobnosťami a faktami. RAG ponúka efektívny nástroj na zvýšenie faktickosti generovaného textu, ale jeho úspech je priamo závislý od kvality knowledge base, relevance retrievalu a vhodnej segmentácie dokumentov.

Kombinácia fine-tuning-u na doménových dátach a pravidelného monitoringu produkčných systémov predstavuje najlepšiu cestu, ako minimalizovať halucinácie a využiť plný potenciál moderných AI riešení v kritických oblastiach.

Ak túžite vybudovať systém, ktorý skutočne dôveruje faktom, nezabúdajte – **kvalita dát a dôsledné testovanie sú absolútnou nevyhnutnosťou.**



Kde je zdroj a ako to otestujeme? Testujte retrievall na vlastných dátach a simulujte kritické scenáre z reálneho použitia, len tak uvidíte, ako často a prečo halucinácie vznikajú.