

V dobe, keď umelá inteligencia (AI) a veľké jazykové modely (LLM) prenikajú do náročných a vysoko regulovaných oblastí, ako je právo, medicína či financie, otázka bezpečnosti a spoľahlivosti modelov naberá úplne nový rozmer. Jedným zo zásadných konceptov, ktoré pomáhajú znížiť riziko nesprávnych rozhodnutí, je **uncertainty estimation** – odhad neistoty modelu. V tomto článku sa pozrieme na to, čo uncertainty estimation znamená, prečo LLM halucinujú, aké sú najväčšie riziká v regulovaných odvetviach, a ako možno pomocou nástrojov ako RAG a fine-tuning zvýšiť bezpečnosť AI systémov.

Čo sú AI halucinácie a prečo LLM halucinujú?

Výraz *AI halucinácie* označuje situácie, keď model vytvorí odpoveď, ktorá je buď úplne nesprávna, neexistujúca alebo nepravdivá, hoci môže znieť dôveryhodne a presvedčivo. To je v podstate problém, ktorý pramení z princípu fungovania veľkých jazykových modelov.

Pravdepodobnosti vs fakty

- **LLM nie sú databázy faktov.** Modely ako GPT, BERT či iné neuralné siete vychádzajú z pravdepodobností výskytu slov a fráz v textovom korpuse, nie z explicitne overených faktických údajov.
- **Predpovedajú ďalšie slovo na základe vzoru.** Ich "vedomosti" sú implicitné a štatistické, preto často generujú obsah, ktorý je štatisticky pravdepodobný, nie nevyhnutne pravdivý.
- **Neistota sa nekomunikuje.** Štandardné LLM často neuvádzajú, ako veľmi si model istý svojou odpoveďou je – tzv. odhad neistoty chýba.

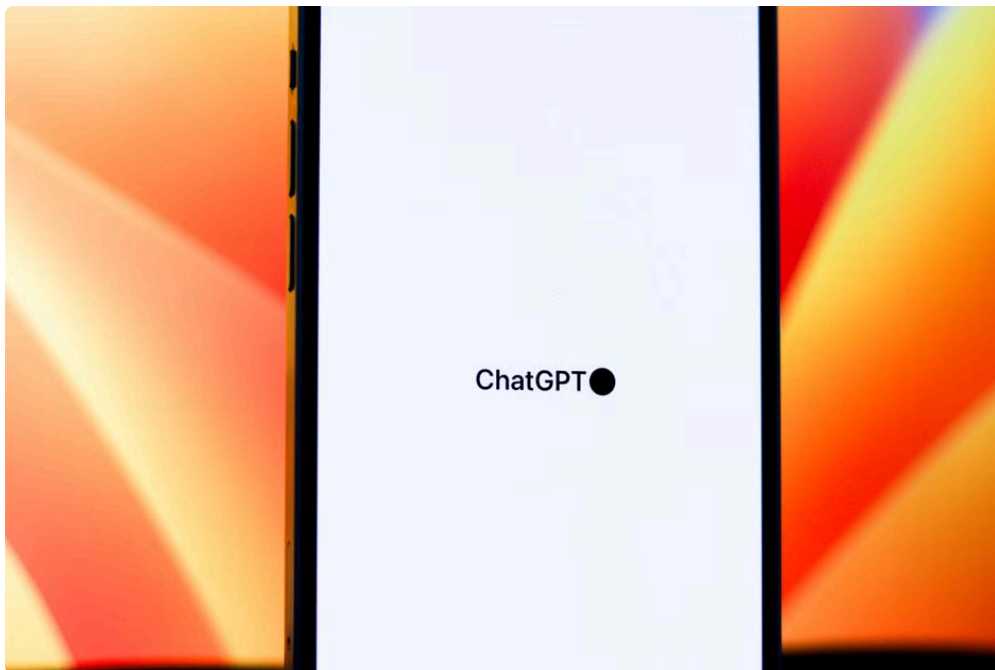
Práve absencia **uncertainty estimation** vedie k riziku, že užívateľ neodlíši správnu odpoveď od halucinácie, čo môže mať v kritických aplikáciách katastrofálne dôsledky.

Prečo je uncertainty estimation kritická v regulovaných odvetviach?

Nie všetky odvetvia sú rovnaké – tam, kde môžete tolerovať pár nesprávnych informácií (napríklad v bežných konverzačných chatbotov), v regulovaných sektoroch riziko omylu výrazne stúpa. Tu je prehľad oblastí s najvyšším rizikom:

1. **Právo** – nesprávna právna rada môže viesť k strate majetku alebo porušeniu zákonov.
2. **Medicína** – nesprávna diagnóza alebo liečba môže vážne ohroziť život pacienta.
3. **Financie** – chybné investičné rady alebo nesprávna interpretácia finančných údajov môžu viesť k značným stratám.
4. **Multi-step výskum** – veda a výskum často vyžadujú overenie viacerých krokov, kde akákoľvek chyba sa kumuluje a skresľuje výsledný záver.

V týchto oblastiach sa požaduje nielen vysoká kvalita odpovedí, ale aj transparentnosť ohľadom ich spoľahlivosti. Tu nastupuje na scénu **uncertainty estimation** – schopnosť modelu nielen odpovedať, ale aj vyjadriť, ako veľmi je si istý správnosťou svojej odpovede.



Uncertainty estimation: Čo to vlastne je?

Uncertainty estimation je technika merania a kvantifikácie neistoty, ktorú model má v generovanej výstupnej hodnote alebo rozhodnutí. V kontexte LLM znamená možnosť, že model vie indikovať, či je odpoveď pevne podložená dátami, alebo len špekuláciou.

- **Epistemická neistota:** Nepochopenie alebo nedostatok znalostí modelu o danej problematike. Táto neistota môže klesať s viac-doménovými dátami alebo fine-tuningom.
- **Aleatorická neistota:** Neistota vyplývajúca z inherentnej náhodnosti dát (napr. nepresné, neúplné, alebo protichodné informácie v knowledge base).

Prečo odhad neistoty pomôže?

- Model dokáže upozorniť používateľa na neisté odpovede, čím šetrí čas pri verifikácii.
- V regulovaných odvetviach znižuje riziko právnych, finančných či medicínskych chýb.
- Umožňuje inteligentnú eskaláciu (napríklad kontakt priamo s expertom, keď je neistota príliš vysoká).

RAG (Retrieval-Augmented Generation) a kvalita knowledge base

Jedným z efektívnych prístupov, ako zlepšiť kvalitu odpovedí a zároveň znížiť neistotu LLM, je využitie metódy **RAG (Retrieval-Augmented Generation)**. Tento prístup kombinuje generovanie odpovedí s vyhľadávaním relevantných faktických informácií v trusted knowledge base.

Ako RAG pomáha:

- Model pred samotným generovaním prehľadá dokumenty, databázy, interné knowledge base, aby získal konkrétne a fakticky podložené dáta.
- Odpoveď je generovaná na základe reálnych, overených informácií, nie len z patternu textu.
- Keď knowledge base nie je dostatočne kvalitná alebo aktuálna, model môže signalizovať vyššiu neistotu alebo požadovať doplňujúce overenie.

Kvalita knowledge base je kľúčová

Bez kvalitnej, aktuálnej a dobre udržiavanej knowledge base je RAG menej efektívny a riziko halucinácií napriek nemu pretrváva. Preto regulované inštitúcie musia investovať do:

- Verifikácie dát a ich pravidelnej aktualizácie.
- Štruktúrovania informácií a dôkladnej kategorizácie obsahu.
- Monitoringu a auditovania odpovedí AI systémov v produkcii.

Fine-tuning LLM na doménových dátach

Aby modely lepšie rozumeli špecifikám daného odvetvia, často sa využíva postup **fine-tuning** – doladovanie modelu na dátach konkrétnej domény, ktoré sú presne relevantné k úlohe (napríklad právne dokumenty, medicínske štúdie alebo finančné správy).

Výhody fine-tuningu

- Zvyšuje odbornú presnosť odpovedí a znižuje epistemickú neistotu.
- Pomáha modelu rozpoznať a vyhýbať sa generovaniu halucinácií v rámci konkrétnej domény.
- Vytvára špecifické pravidlá a preferencie, ktoré sú v súlade s internými predpismi a reguláciami.

Pri fine-tuningu by však nikdy nemal chýbať robustný systém na odhad neistoty – ak model po doladovaní stále generuje **invelity** neisté odpovede, riadiaci systém to musí vedieť zachytiť a riešiť.

Mini-checklist pre nasadenie AI riešení v regulovaných odvetviach

1. **Definujte jasné KPI:** Presné metriky kvality odpovedí, tolerancia chýb a úrovne neistoty.
2. **Investujte do kvalitnej knowledge base:** Overené, aktualizované a kategorizované dáta.
3. **Implementujte RAG:** Na podloženie každého výstupu faktami z knowledge base.
4. **Vykonajte fine-tuning modelu:** Na doménové dáta, reflektujúce špecifiká vašej oblasti.
5. **Zavedenie uncertainty estimation:** Jasná komunikácia neistoty, eskalácia pri vysokej neistote.
6. **Monitorovanie produkčných incidentov:** Pri každom podozrení na halucináciu automatická notifikácia alebo vyhodnotenie.
7. **Pravidelný audit a update:** Just-in-time update modelu a knowledge base.

Záver

V regulovaných odvetviach nie je AI "len" nástrojom, ale svojím spôsobom partnerom, ktorý môže mať zodpovednosť za závažné rozhodnutia. Preto je nevyhnutné, aby umelá inteligencia nebola len zdrojom odpovedí, ale aby vedela povedať aj, keď si nie je istá – a tým výrazne znížila riziko spôsobenia škody.

Uncertainty estimation sa stáva kľúčovým prvkom pre dôveryhodnosť a bezpečnosť AI riešení. Kombinácia RAG pre vyhľadávanie faktov, fine-tuningu na presné doménové dáta a dôslednej správy knowledge base vytvára základ pre spoľahlivú a bezpečnú umelú inteligenciu aj v tých najprísnejšie regulovaných oblastiach.



Nezabúdajme: v AI nikdy nejde o "asi" alebo "nejako" – vždy vyžadujeme jasný zdroj, odhad neistoty a spôsob, ako to otestovať.