

V nasledujúcom článku sa detailne pozrieme na to, ako vybudovať efektívny **feedback loop**, v ktorom support tím aktívne *labeluje odpovede* generované AI systémom ako správne alebo zlé. Tento proces je kľúčový pre **zlepšovanie modelu**, obzvlášť ak pracujeme s veľkými jazykovými modelmi (LLM) a technológiami ako RAG (Retrieval-Augmented Generation) a fine-tuning na doménových dátach.

Čo sú AI halucinácie a prečo LLM halucinujú?

Termín *AI halucinácie* označuje situácie, kedy AI model – prevažne veľký jazykový model – generuje informácie, ktoré nie sú pravdivé, často abstraktné alebo vymyslené. Ide o kritický problém najmä v produkčných nasadeniach, kde je presnosť informácií nevyhnutná.

Prečo k tomu dochádza?

- **Pravdepodobnostný charakter LLM:** LLM generujú odpovede na základe pravdepodobnosti, nie overených faktov. Model vypočíta, ktorá sekvencia slov je najpravdepodobnejšia podľa naučených vzorov, ale nemá priamy prístup k faktickým databázam alebo overeným zdrojom.
- **Chýbajúce doménové špecifiká:** Ak model nie je dostatočne naučený na špecifických dátach, môže extrapolovať alebo robiť chyby, pretože nepozná natívnu štruktúru znalostí danej oblasti.
- **Neurčitý vstup alebo dlhé multi-step otázky:** Komplexné otázky s viacerými krokmi často vedú k nahromadeniu chýb, ktoré eskalujú halucinácie.

Kedy je riziko halucinácií najvyššie?

Niektoré oblasti sú omnoho citlivejšie na nesprávne informácie vytvorené AI. Nasledujúce sektory žiadajú maximálnu pozornosť a precízne nastavenie workflow:

- **Právo:** Nesprávne právne rady môžu spôsobiť vážne právne následky a poškodenie reputácie.
- **Medicína:** Nesprávne medicínske informácie môžu ohroziť životy pacientov.
- **Financie:** Nepresné finančné odporúčania môžu viesť k výrazným stratám a regulačným problémom.
- **Multi-step výskum:** Vedecké a technické domény vyžadujú precíznosť v krokoch a uvedomovanie si závislostí, aby sa eliminovala kumulatívna chyba v odpovediach.

Úloha RAG a kvalita knowledge base v prevencii halucinácií

RAG (Retrieval-Augmented Generation) kombinuje schopnosť generovať text s vyhľadávaním relevantných dát z externej knowledge base. Vďaka tomu:



- Model môže nadviazať generovanie na aktuálne, overené fakty dostupné v knowledge base.
- Znižuje sa riziko úplných halucinácií – odpovede sú podložené konkrétnymi dokumentmi alebo údajmi.
- Kvalita knowledge base je preto kriticky dôležitá – musí byť aktualizovaná, overená a relevantná k doméne.

Bez dobrej quality control nad knowledge base sa môže RAG systém stať zdrojom nepresností rovnako ako čistý LLM.

Prečo je feedback loop a labelovanie odpovedí dôležité?

Bez spätného väzobného mechanizmu sú chyby v odpovediach neviditeľné a model sa nemá ako poučiť. Implementácia robustného procesu, kde support tím označuje odpovede ako správne alebo zlé, prináša:

1. **Identifikáciu problematických odpovedí** v reálnom čase.
2. **Podklad pre fine-tuning modelu** na doménových dátach, aby sa minimalizovali opakované chyby.
3. **Zlepšenie knowledge base** – odkrytie chýbajúcich alebo neaktuálnych informácií.
4. **Transparentnosť pri produkčnom nasadení** a lepší prehľad o kvalite AI odpovedí.

Konkrétny návod na nastavenie workflow s označovaním odpovedí support tímom

Nasledujúci checklist pomôže štruktúrovať správny workflow, ktorý možno nasadiť v podniku využívajúcom RAG a fine-tuning na doménových dátach:



1. Vytvorenie jednotného rozhrania na labelovanie odpovedí

- Zabezpečte jednoduchý a intuitívny UI pre support tím, kde môžu označiť odpovede ako *správne* alebo *zlé*.
- Pridajte možnosť krátkeho komentára, kde môžu uviesť dôvody nesúhlasu alebo doplniť správnu odpoveď.
- Zahŕňajte timestamp a ID konverzácie pre sledovateľnosť.

2. Reálne monitorovanie a reporting

- Pravidelne vyhodnocujte štatistiky správnosti odpovedí – percento labelov „správne“ vs „zlé“.
- Vytvorte dashboard pre vedenie aj vývojárov, kde môžu sledovať trendy kvality odpovedí.

3. Feedback loop implementácia do vývoja modelu

- Labelované odpovede vyberajte ako tréningové dáta pre **fine-tuning** modelu.
- Oboznámte vývojový tím s pravidelnými update-ami a testujte zmeny na validácii pred produkčným nasadením.
- Automatizujte cyklus aktualizácie modelu a knowledge base tak, aby sa chyby postupne eliminovali.

4. Kvalitná knowledge base ako základ pre RAG

- Udržujte knowledge base pravidelne aktualizovanú, presnú a špecifickú pre danú doménu.
- Využívajte efektívne nástroje na indexovanie a vyhľadávanie, aby retrieval čas bol rýchly a relevantný.

5. Školenie support tímu na správne labelovanie

- Zabezpečte, aby tím rozumel, čo znamená správna vs nesprávna odpoveď.
- Definujte jasné pravidlá a príklady pre konzistentné labelovanie.
- Zohľadnite multi-step otázky a ich hodnotenie správania modelu.

Ako otestovať nastavený workflow?

Kľúčové je validovať, či workflow funguje od labelovania až po zlepšenie modelu:

1. **Zbierka testovacích prípadov:** Simulujte rôzne druhy otázok (jednoduché, komplexné, multi-step, citlivé domény).
2. **Interný audit:** Nechajte viacero členov support tímu nezávisle označiť rovnaké odpovede a otestujte konzistenciu labelovania.
3. **Štatistická analýza:** Sledujte mieru úspešnosti modelu pred a po aktualizácii podľa feedbacku.
4. **Monitoring produkcie:** Zapojte monitoring s alertmi, ak podiel zlých odpovedí stúpne nad definovanú hranicu.

Zhrnutie

Fáza Čo robiť Prečo Labelovanie odpovedí Support tím označuje odpovede ako správne alebo zlé Získanie dát pre spätnú väzbu a identifikáciu chýb Monitoring a reportovanie Priebežné vyhodnocovanie kvality a trendov Prehľad o výkone modelu v produkcii Fine-tuning modelu Pravidelne tréning na označených dátach zo support tímu Minimalizácia halucinácií a zvýšenie presnosti Udržiavanie knowledge base Aktualizácia a kontrola kvality zdrojových dát pre RAG Základ pre relevantné a pravdivé odpovede Školenie support tímu Presne definované pravidlá labelovania a školenia Konzistentné a hodnotné spätné väzby <https://www.invelity.com/blog/ako-znizit-halucinacie-ai/>

Ak sa vyhnete nejasnostiam ("nejako" a "asi" nemajú miesto v špecifikácii feedback loopu) a zabezpečíte jasné zdroje dát aj pravidelný monitoring, eliminujete najväčšie riziká AI halucinácií a váš systém sa stane robustným nástrojom pre podporu vášho tímu.

Upozornenie: Kde je zdroj a ako to otestujeme?

Každý nástroj a postup nastavenia workflow by mal byť zdokumentovaný so zdrojmi tréningových dát, presným popisom pravidiel labelovania a s definovanými metrikami úspešnosti. Testovanie prevedieme podľa checklistu vyššie s jasne stanovenými KPI a pravidelnými revíziami na produkčných dátach.

Here's what kills me: takto získate skutočný pohľad na kvalitu a efektivitu ai asistentov vo vašom podniku, nie len abstraktné dohady a prania.