KONTROLOVATELNOST ROZHODOVÁNÍ AI

chyby v tradičním SW

vs.

LLM

(ChatGPT, Claude, Gemini...)

Tradiční SW

- Příčinou chyby je obvykle špatný kód, HW chyba v infrastruktuře, špatná konfigurace nebo chyba uživatele.
- Chyba je konzistentní, opakovatelná.
- Reprodukce je snadná, a to opakovaně se stejným výsledkem.
- Zdrojů k prozkoumání bývá dostatek - aplikační, síťové a systémové logy.

LLM

**Velký jazykový model,
např. ChatGPT, Claude, Gemini…**

- **Obtížně reprodukovatelné** – na stejnou otázku může pokaždé podat jinou odpověď, míra úspěšnosti zopakování závisí na kontrole nad verzí modelu, nastavení temperature, kompletní prompt a vyhledávací kontext

- **Degradace nebo postupné odchýlení modelu** je obtížně detekovatelné, dokud nedojde k závažné chybě

LLM

**Velký jazykový model,
např. ChatGPT, Claude, Gemini…**

- **Zdroje k prozkoumání** – je potřeba ukládat velké množství informací jako logy vyhledávání a dovozování, prompt, kontext, dedukce agentů (většina těchto informací se defaultně neukládá).

Co zachovat pro možnost případné rekonstrukce chybového chování AI:

- **Spárované dotazy a odpovědi**, včetně časové posloupnosti
- **Kompletní systémový prompt**, uživatelský vstup, další automatizované kroky. Komplexní systémy jsou schopné dynamicky vytvářet prompt z kombinace šablon, uživatelských dotazů a obsahu dokumentů (RAG – Retrieval-augmented Generation).

- **Logy RAG systémů**: odkazy na konkrétní dokumenty nebo jejich části, dotaz a skóre, které vedlo k jejich výběru. Všechny tyto záznamy by měly být uloženy společně s generovanou odpovědí.

- **Záznamy o verzích modelů**, použitých pro získání dat. Bez této informace nemusí být možné rozhodnout, zda chyba nastala kvůli špatnému zadání dotazu nebo změnou modelu pro získávání dat.

- **Záznamy akcí agentů** – jaké použili nástroje a proč, jak delegovali úkoly a komu, jak úkoly řadili za sebe.

- **Detailní informace o modelu** – verze, ladění, konfigurace, nastavení mantinelů.

Pro **konkrétní konzultaci**

jsme tu pro Vás

cyberex

www.cyberex.cz

info@cyberex.cz

tel. +420 211 150 066