ZÁRÓFELADAT - RAG ALAPÚ AI ASSZISZTENS FEJLESZTÉSE

Fejlessz egy saját RAG (Retrieval-Augmented Generation) alapú AI asszisztenst, amely képes dokumentumok feldolgozására, releváns információk visszakeresésére és intelligens válaszok generálására.

KÖTELEZŐ KÖVETELMÉNYEK

1. RAG RENDSZER ARCHITEKTÚRA

- Dokumentum feldolgozás és chunking stratégia implementálása
- Embedding modell választása és integrálása (bármilyen: OpenAI, Cohere, lokális, stb.)
- Vektor adatbázis implementáció (pgvector, Qdrant, Pinecone, vagy egyéb)
- Retrieval és reranking mechanizmus
- LLM integráció válaszgeneráláshoz (API-based vagy self-hosted)

2. ALKALMAZÁS FUNKCIÓK

- Webes felület (Next.js, Streamlit, Gradio, vagy egyéb)
- Dokumentum feltöltés és kezelés
- Streaming válaszok támogatása
- Session/conversation management

3. HÁROMSZINTŰ EVALUATION FRAMEWORK

A teszt eset számok csak minimum követelmények, nyugodtan lehettek alaposabbak is.

3.1 RAG SZINTŰ ÉRTÉKELÉS

- Retrieval minőség mérése (precision, recall, MRR)

- Embedding modell teljesítmény tesztelése
- Chunking stratégia hatékonyságának mérése
- Minimum 20 teszteset

3.2 PROMPT SZINTŰ ÉRTÉKELÉS

- Single-turn eval
- Context relevance mérése
- Hallucináció detektálás
- LLM-as-Judge implementáció
- Minimum 15 teszteset

3.3 ALKALMAZÁS SZINTŰ ÉRTÉKELÉS

- Teljes user journey tesztelése
- Response quality értékelés
- Latency és performance metrikák
- User satisfaction szimulálása
- Minimum 10 komplex teszteset

4. MONITORING ÉS ANALITIKA

- Token használat és költség tracking
- Latency metrikák (first token, total response time)

VÁLASZTHATÓ BÓNUSZ FUNKCIÓK (plusz pontért)

- Fine-tuned embedding modell használata
- Felhasználói feedback gyűjtés és elemzés

- Advanced reranking (cross-encoder vagy ColBERT)

VIDEÓ PREZENTÁCIÓ KÖVETELMÉNYEK

Készíts 2 db Loom videót a projektről:

1. VIDEÓ - TECHNIKAI BEMUTATÓ (5 perc)
   - Architektúra áttekintés (1 perc)
   - RAG pipeline működésének bemutatása (1.5 perc)
   - Evaluation framework demonstráció (1.5 perc)
   - Monitoring dashboard bemutatása (1 perc)

2. VIDEÓ - FELHASZNÁLÓI DEMO (5 perc)
   - Dokumentum feltöltés és feldolgozás (1 perc)
   - Különböző típusú kérdések megválaszolása (2 perc)
   - Hibakezelés és edge case-ek bemutatása (1 perc)
   - Teljesítmény és válaszminőség demonstrálása (1 perc)

LEADÁSI KÖVETELMÉNYEK

1. GitHub repository a teljes forráskóddal
2. README.md telepítési és használati útmutatóval
3. Evaluation eredmények dokumentációja
4. Loom videó linkek

Sikeres projektkészítést kívánunk!

Kérdések esetén keress fel a kurzus Discord csatornáját.