Module 04 - Házi Feladat: Multi-Turn Evaluáció és Iteratív Fejlesztés

Feladat Áttekintése

A házi feladat célja, hogy gyakorlati tapasztalatot szerezzetek multi-turn evaluációval és iteratív AI asszisztens fejlesztéssel. **Nem egy tökéletes AI asszisztens építése a cél**, hanem annak megértése, hogyan tudunk evaluációval javítani egy-egy konkrét aspektuson.

Kiindulási Pont

Használhatjátok:

- Az előző hét házijából származó AI asszisztenseteket (ha már bekötöttétek valamilyen adatot)
- Csak a jelenlegi egyszerű memóriával rendelkező AI asszisztenst
- Ki is vehetitek a tool hívásokat, hogy egyszerűsítsétek a független eval futtatásokat

Feladat Lépései

1. AI Asszisztens Kiválasztása és Scope Meghatározása

1.1) Válasszatok egy egyszerű, konkrét aspektust értékelésre:

- **Tone/hangnem**: túl formális vs. barátságos
- **Happy path működés**: alapvető kérdésekre ad-e értelmes választ
- **Specifikus domain knowledge**: pl. csak egy kis témakörben működik jól
- **Udvariasság**: köszön-e, használ-e udvarias fordulatokat
- **Rövidség vs. részletesség**: túl tömör vagy túl bőbeszédű

1.2) Dokumentáljátok:

- Melyik asszisztenst használjátok
- Milyen aspektust fogtok értékelni
- Mi a célotok (pl. "barátságosabb hangnemet elérni")

2. Szimulált Felhasználó Implementációja

Készítsetek 2-3 különböző persona-t illetve a fejlesztési aspektushoz tartozó megfelelő goal-t:

- Pl. türelmes tapasztalt, türelmetlen kezdő, stb.

3. Evaluációs Metrikák Definiálása

3.1) Implementáljatok egy egyszerű LLM-as-a-Judge értékelőt:

- A kiválasztott aspektusra (pl. tone) 0-3 skálán értékeljen
- Kérjetek indoklást is

3.2) Példa metrikák:

- User elégedettsége (szimulált)
- Beszélgetés sikeressége (cél teljesítése)
- Kiválasztott aspektus pontszáma (LLM judge)

4. Baseline Mérés

4.1) Futtassatok 3-5 szimulációt a jelenlegi asszisztensetekkel

- Dokumentáljátok az eredményeket
- Figyeljétek meg a mintázatokat - mi működik rosszul?

5. Iteratív Fejlesztés

5.1) Alapján a baseline eredmények alapján tegyetek egy konkrét változtatást:

- System prompt módosítás
- Egyszerű logika hozzáadása
- Temperature vagy más paraméter állítás

5.2) Mérjétek meg újra ugyanazokkal a szimulációkkal

5.3) Dokumentáljátok:

- Mit változtattatok
- Javult-e a kiválasztott aspektus

- Milyen trade-offokat tapasztaltatok

Elvárások

Amit NEM várunk el:

- Tökéletes AI asszisztenst
- Sok evaluáció futtatása (költség miatt)
- Komplex evaluációs keretrendszer
- Minden aspektus javítása

Amit elvárunk:

- **Egy konkrét aspektus** mérését és javítását
- **Dokumentált iterációs folyamat**
- **Őszinte reflexió** - mi működött, mi nem
- **Kód** a szimulációhoz és evaluációhoz
- **2-3 iteráció** maximum

Leadási Formátum

Kötelező fájlok:

1. **README.md** - dokumentáció az egész folyamatról
2. A frissített kódbázis - új persona, system prompt, stb.

README.md tartalma:

Multi-Turn Evaluáció Házi

Kiválasztott Aspektus
[Milyen aspektust értékeltetek és miért]

Baseline Eredmények
[Mit mértetek kezdetben]

Változtatások
[Mit módosítottatok és miért]

Végeredmény
[Javult-e az aspektus, milyen trade-offok]

Tanulságok
[Mit tanultatok a folyamatból]

Tippek

Debugging:

- Mentsétek el a beszélgetéseket
- Nézzétek át manuálisan is pár szimulációt
- Egyszerű logolás sokat segíthet

Scope:

- **Egy aspektus** elegendő
- 2-3 iteráció elég
- Happy path esetek is teljesen okés
- Nem kell edge case-eket kezelni

Fontos: A cél nem a tökéletes eredmény, hanem a folyamat megértése. Dokumentáljátok őszintén, ha valami nem működött - ez is értékes tanulság! A kritérium, hogy legyen 2-3 lefutattot kísérlet, benne szimulált beszélgetésekkel. Vagyis az követelmény, hogy maga az eval pipeline lefusson.