



FarmGPT projekt: A személyre szabott "ChatGPT" útvesztői

RAG, memória és eszközhasználat: hiányzó láncszemek az API-hívás és a kész megoldás között

**Zsilinszki
Anna**

dataSTREAM

2026

2026.05.13.



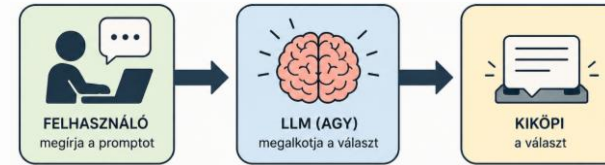
TÉVHITEK & ILLÚZIÓK

AZ AI CHATBOTOK ÉPÍTÉSÉVEL KAPCSOLATBAN

- ✗ *"Tanul a beszélgetésből." vagy "Saját adatainkkal betaníthatjuk, mint egy klasszikus ML modellt"*
- ✓ GPT: generative *pre-trained* transformer, statikus tudással, használjuk az előre tanított modellt. ML paradigmaváltás. RAG: a tanítás helyett.
- ✗ *"Emlékezik a beszélgetésekre."*
- ✓ Memóriában eltárolt beszélgetéseket a modell újra kell olvassa minden kérdésnél.
- ✗ *"Naprakész információkkal rendelkezik."*
- ✓ Ha bekötünk egy internetes kereső eszközt...

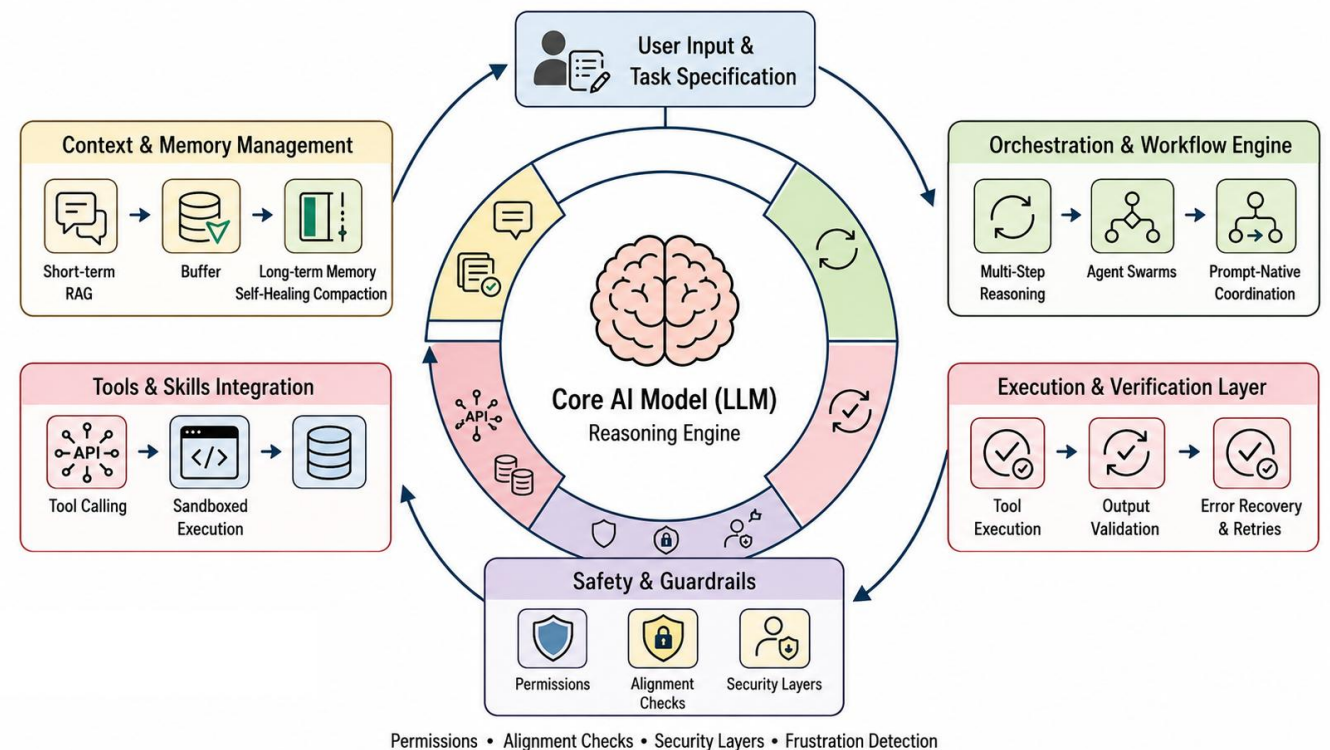
Ábra forrás: Medium: <https://medium.com/@savelis.pedro/ai-harness-engineering-what-512-000-lines-of-claude-code-leak-taught-us-e7809a9cef04>

AMIT GONDOLUNK:



LLM ✗ ChatGPT/Gemini/Claude

ÉS AMI VALÓJÁBAN TÖRTÉNIK:



ChatGPT = LLM + kiegészítő komponensek

AZ ILLÚZIÓK OKAI

MODELL BELSŐ TULAJDONSÁG

- Pre-trained (statikus) tudás
- Súlyok
- Reasoning képesség
- Context window
- Multimodalitás

Amivel ezek módosíthatók:

- Fine-tuning
- Újratanítás

MODELL KÖRÉ ÉPÍTETT RENDSZER

- Query rewriting
- System prompt
- Guardrails
- Tool use
- Web search
- Orchestration
- Conversation history
- RAG
- ...

← valamennyi átjárás van



FARMGPT: ELKÉPZELÉS

Szeretnénk egy saját személyre szabott chatbotot, ami olyan okos mint a chatGPT, csak a saját adatainkból dolgozik.

- Diverz, több nyelvű, nagy mennyiségű, változatos minőségű, nyelvezetű anyag.
- Folyamatosan bővülő tudásanyag.
- Ha szükséges meghív időjárás, talajnedvesség adatokat a válaszhoz.

Nem egy sima Q&A eszköz:

- Át kell lássa a dokumentumok összefüggéseit.
- Komplex következtetések levonása.

farmGPT

Üdvözöllek! FarmGPT vagyok, az agroökológiai átmenet szakértője. Jelenleg a komposztálás témakörében tudok neked segíteni. Mi érdekelne?

Mivel keverjek össze egy kilogramm répahéjat ahhoz, hogy ideális komposztot kapjak? Röviden és lényegretörően válaszolj!

A répahéj nitrogénben gazdag, ezért szénben gazdag, száraz barna anyagokkal kell keverni az optimális komposztáláshoz. Az ideális szén-nitrogén (C:N) arány 25-40:1, ami gyakorlatban kb. 2-3 rész barna anyagot jelent 1 rész répahéjhoz. Tehát 1 kg répahéjhoz 2-3 kg szalmát, száraz levelet vagy faforgácsot érdemes keverni. Fontos az anyagok aprítása és a nedvességtartalom 40-60% közötti biztosítása a hatékony lebomláshoz, elkerülve a rothadást és a túl magas hőmérsékletet. Ez a keverék megfelelő levegőztetést és mikrobiális aktivitást biztosít a gyors és szagmentes komposztáláshoz.

1 2 3 4 5 6 7 8 9 10

A ChatGPT nem egy modell, hanem egy hatalmas, finomhangolt gépezet, amit nekünk az API hívás után darabonként kell **összelegőznünk** ahhoz, hogy valami hasonlót kapjunk eredményül.

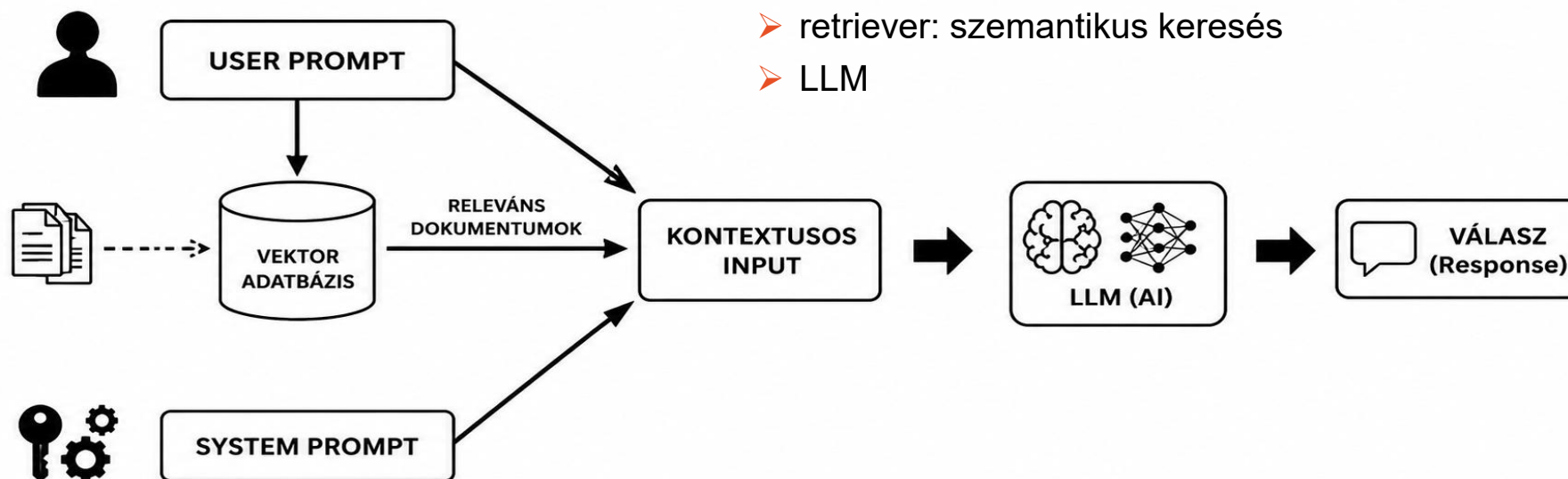
FARMGPT MEGVALÓSÍTÁS: KEZDETEK

Kiindulás: basic (naive) RAG

Komponensek:

- előfeldolgozó (adatbetöltés, adattisztítás, chunking)
- embedding model
- vektoradatbázis
- retriever: szemantikus keresés
- LLM

RAG alapkoncepció:



FARMGPT MEGVALÓSÍTÁS: KEZDETEK

➤ Előfeldolgozási komplikációk:

- többhasábos, szigetes dokumentumok
- táblázatok, képek, vegyes struktúrák
- A kinyerés mindig lineárisan történik: jelentős kézi előfeldolgozás igénye.
- A FarmGPT-nél nem egy egyszeri feladat, bővülő tudástár.

➤ Chunkolás problematikája:

- eltérő dokumentumoktípusok
- túl kicsi – kontextus veszteség
- túl nagy – nem elég specifikus válasz
- szemantikai chunkolás, LLM-alapú chunkolás (drágább)

➤ Válaszok minősége:

- Bizonyos dokumentumokat preferál.
- Drágább modell: gpt 4o mini -> gpt 4.1 mini
- Query rewrite
- LLM reranking



FARMGPT MEGVALÓSÍTÁS: KEZDETEK

- Az embedding többnyelvűség problémája:
 - embedding model a vektortérbe picit közelebb teszi az azonos nyelvű szövegeket egymáshoz: kevésbé releváns találat előrébb kerül ha a nyelv/nyelvezet azonos
 - multilingual embedding model (*cohere: embed-multilingual-v3*)
 - query translate
- Memória:
 - conversation history, rövid és hosszútávú memória
 - session kezelés
 - jelenleg: görgetett summary előző 3 beszélgetés kulcsszavai (LLM)
- Infrastruktúra:
 - üzemeltetés
 - tesztelés (automatikus, kézi)
 - javítás

Python, langchain, pinecone vdb, openai gpt 4.1 mini, cohere embedding



FARMGPT MEGOLDÁS JELENLEG

NAIVE RAG HELYETT ADVANCED RAG:

1. QUERY (USER PROMPT)
2. QUERY REWRITE
3. TRANSLATE PROMPT
4. RETRIEVAL FROM VDB
5. RERANKING RETRIEVED CHUNKS
6. SYSTEM PROMPT
7. LLM
8. ANSWER
9. SUMMARY OF CONVERSATION HISTORY
10. NEXT QUESTION

Token használat exponenciálisan el tud szállni!
Pedig itt még eszköz használat vagy ágensek nem is kerültek képbe...

Következő lépés: Eszközhasználat:

- helyzetfelismerést követő eszközmeghívás
- időjárási adatok
- Kalkulátor (C-N arány)
- talajnedvesség adatok stb.



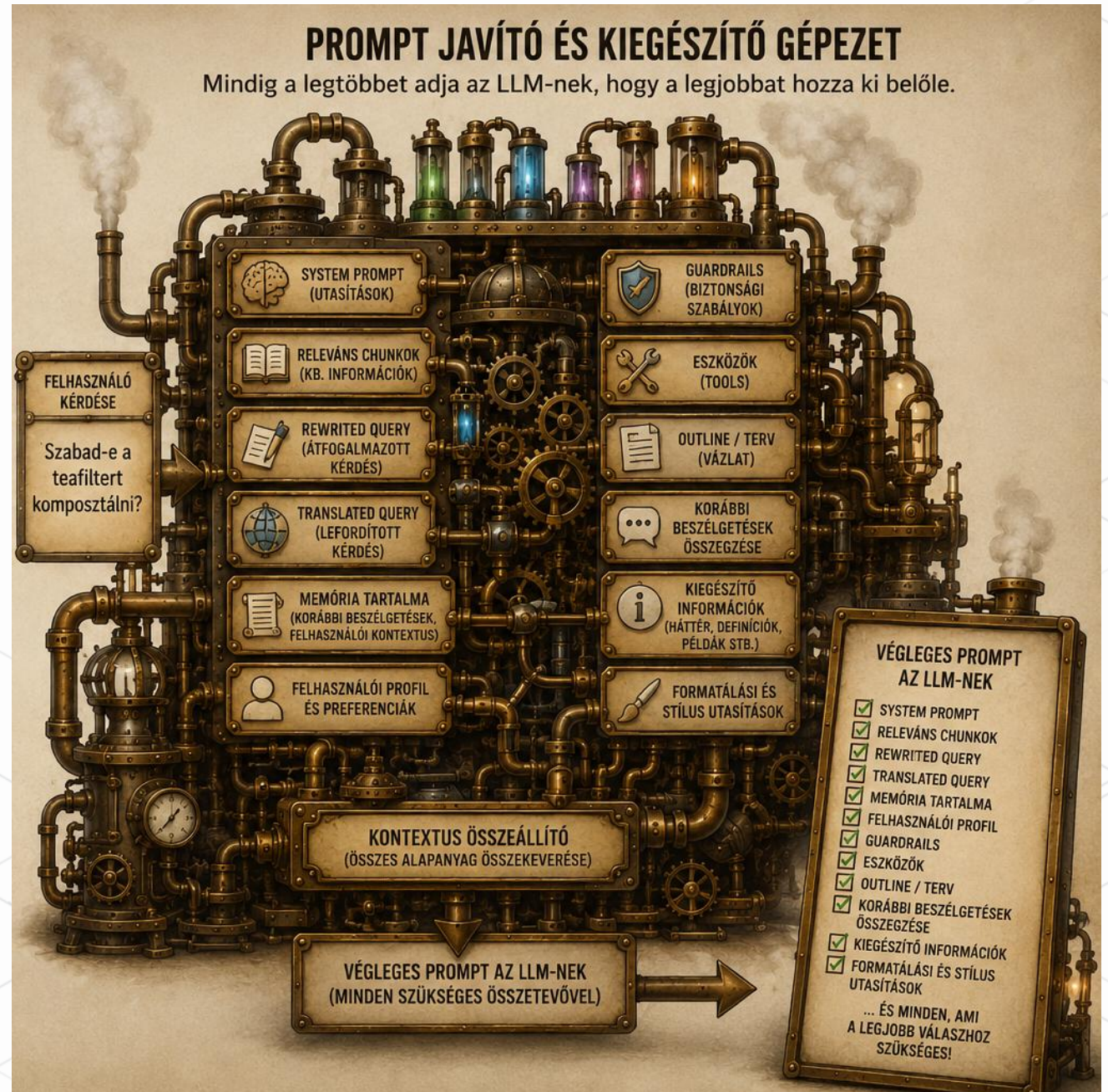
ZÁRSZÓ

Az AI fejlődése nem csak modellfejlődés
→ a „varázslat” nagy része a köré épített rendszerből jön.

A felhasználó egyetlen intelligenciát lát
→ valójában együttműködő komponensek hálózata működik.

Saját AI rendszer fejlesztésénél:
→ az architektúra tudatos megtervezése nem megspórolható.

A FarmGPT moduláris architektúrája lehetővé teszi, hogy a rendszer a résztvevők igényeihez rugalmasan alkalmazkodva, iteratív fejlesztési és tesztelési ciklusokon keresztül nyerje el végleges formáját.



Köszönöm a figyelmet!

Kérdések?

E-mail: azsilinszki@clementine.hu

