

CLEMENTINE

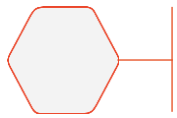
RAG kísérletek

LLM alapú ügyfélszolgálati asszisztens

Zsilinszki Anna, Clementine

RAG KÍSÉRLETEK: LLM ALAPÚ ÜGYFÉLSZOLGÁLATI ASSZISZTENS

MOTIVÁCIÓ, CÉLKITŰZÉS



Piaci trendek monitorozása, potenciális felhasználási területek azonosítása



Dokumentumalapú válaszadás képességének vizsgálata (RAG)



Felhőfüggetlen, helyi megoldások vizsgálata, összehasonlítása a felhő alapú megoldásokkal



Kísérlet: Clementine ügyfélszolgálati asszisztens, létrehozása LLM alapon (GYIK chatbot), adat: Clementine weboldal tartalma



RETRIEVAL-AUGMENTED-GENERATION

RAG

Alapvető cél: “access information beyond training data”.

Első említés: 2020 Facebook AI Research kutatói által írt cikk.*

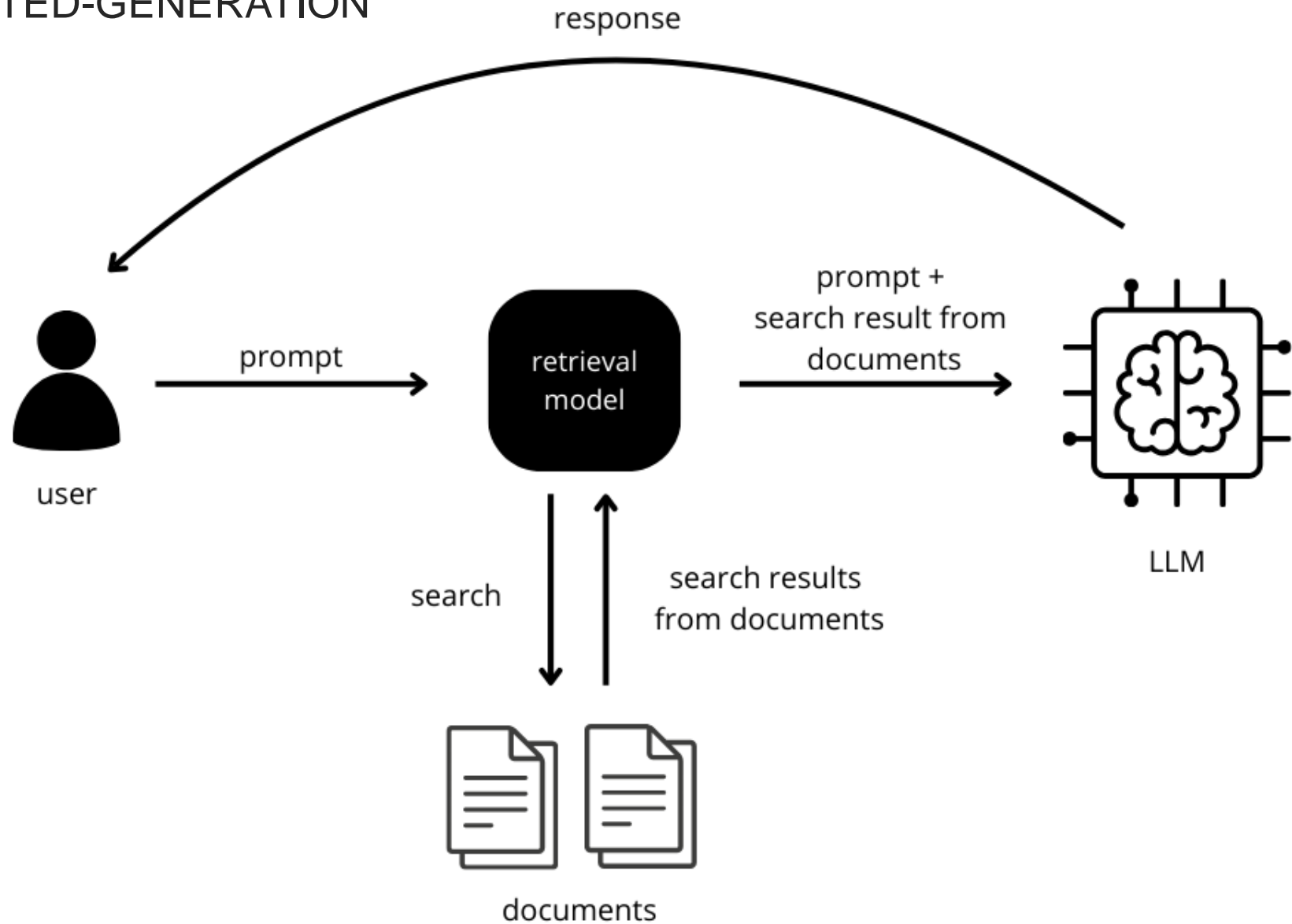
Mit “ígér” a RAG?

- up-to-date információ integrálása
- kevesebb hallucináció
- válaszok minőségének javulása
- specifikus témákban jártasság (saját dokumentumra tartalma alapján)
- adatbiztonság (on premis megoldás esetén)

*Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks.

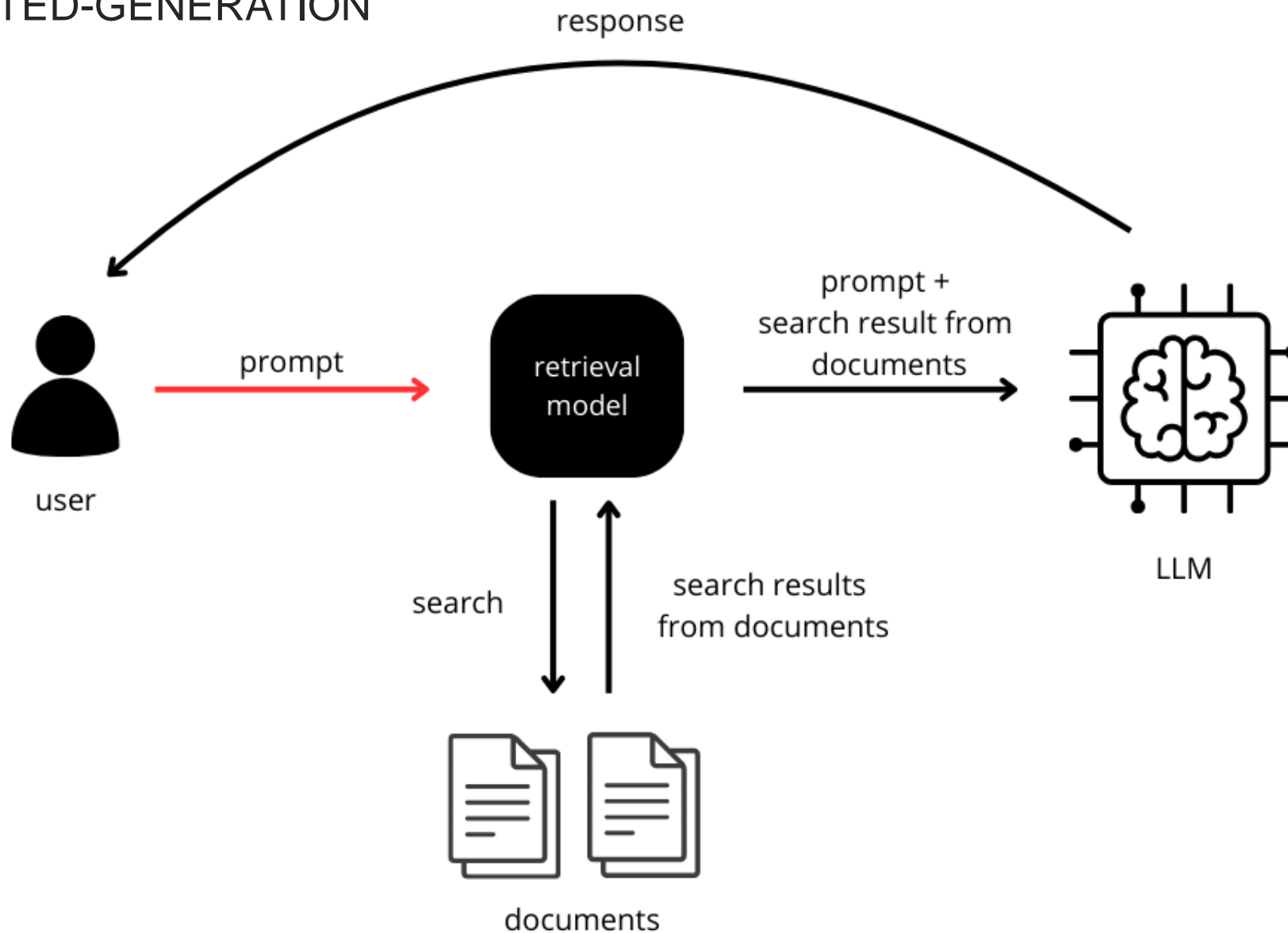
RETRIEVAL-AUGMENTED-GENERATION

RAG



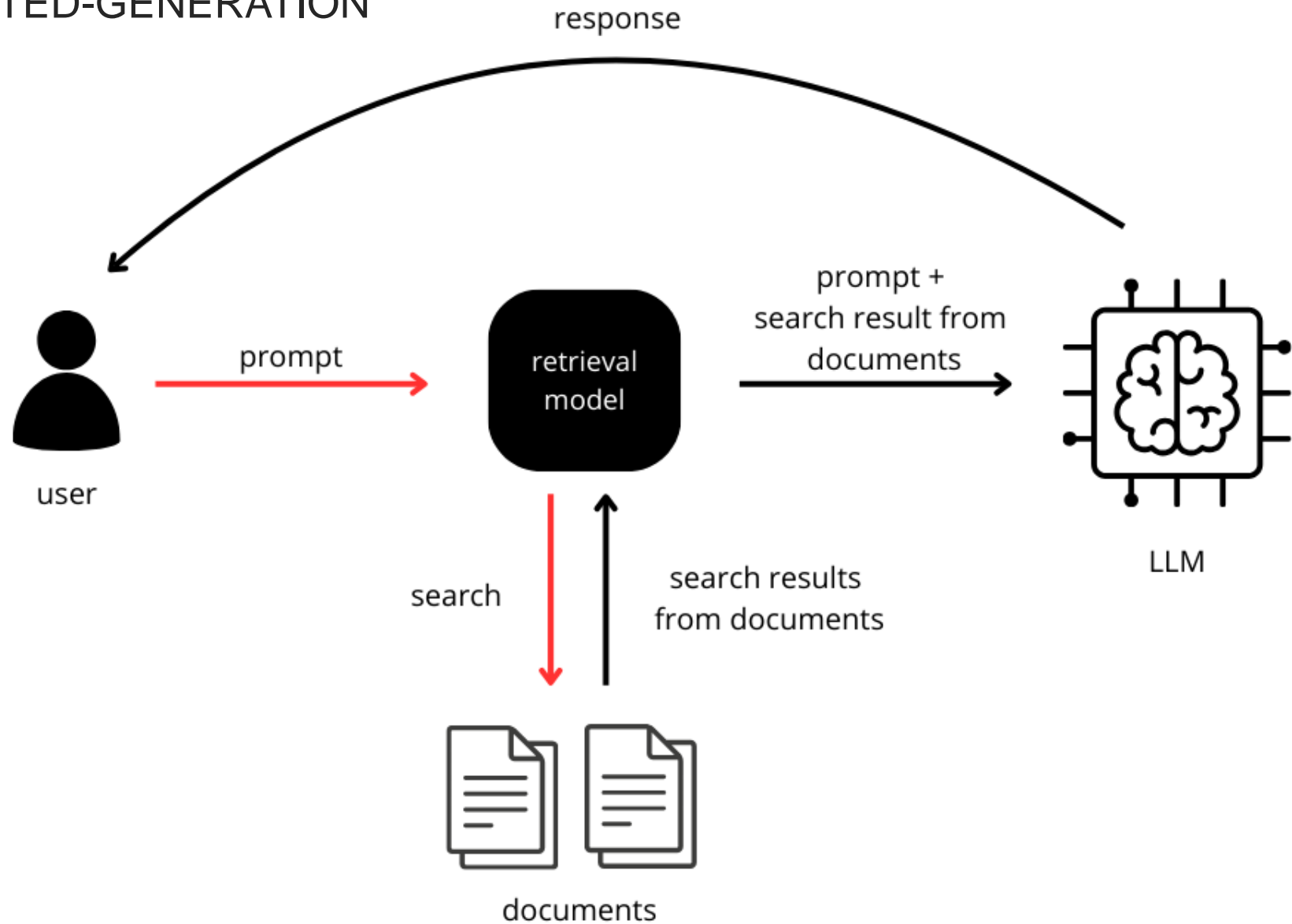
RETRIEVAL-AUGMENTED-GENERATION

RAG



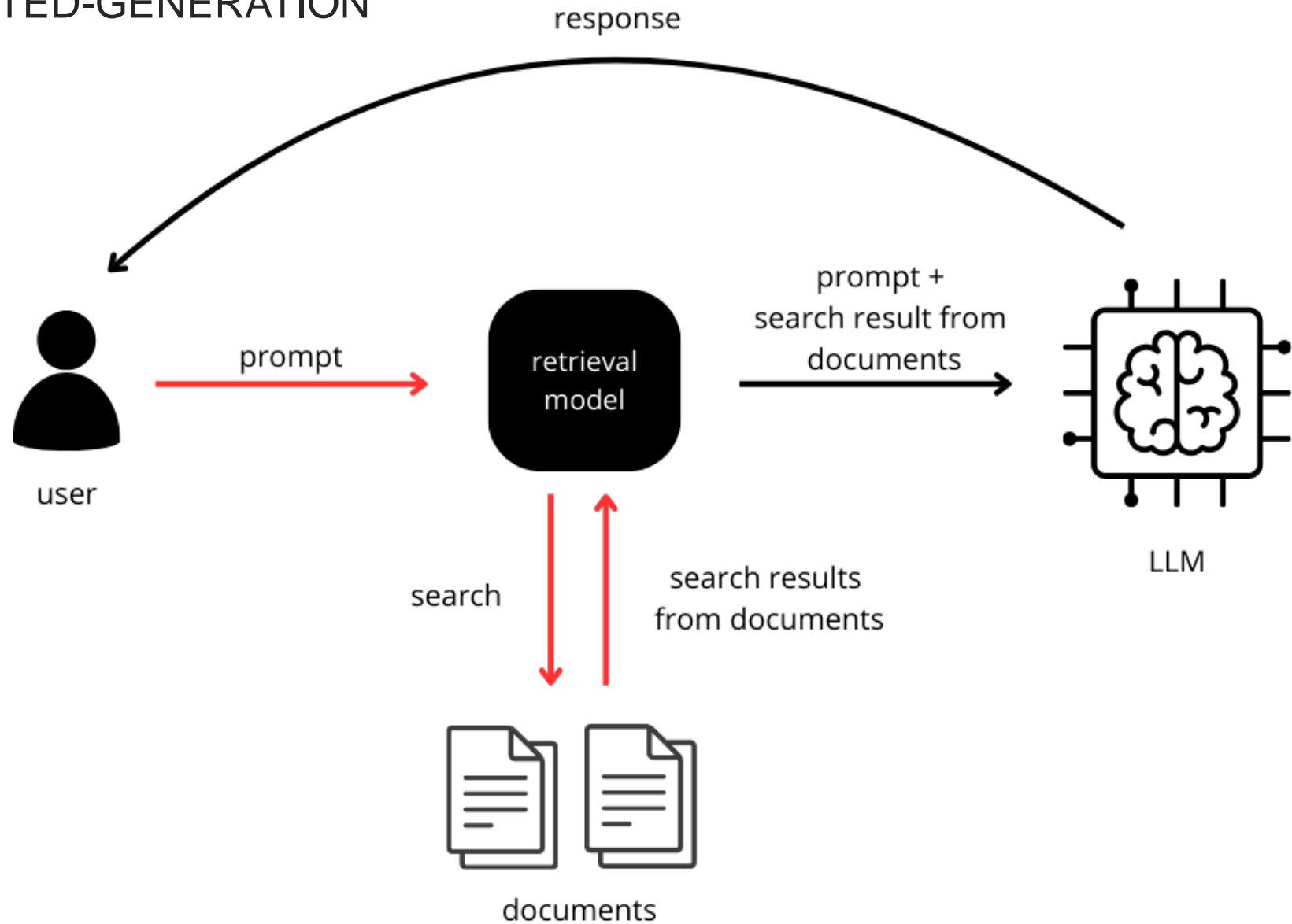
RETRIEVAL-AUGMENTED-GENERATION

RAG



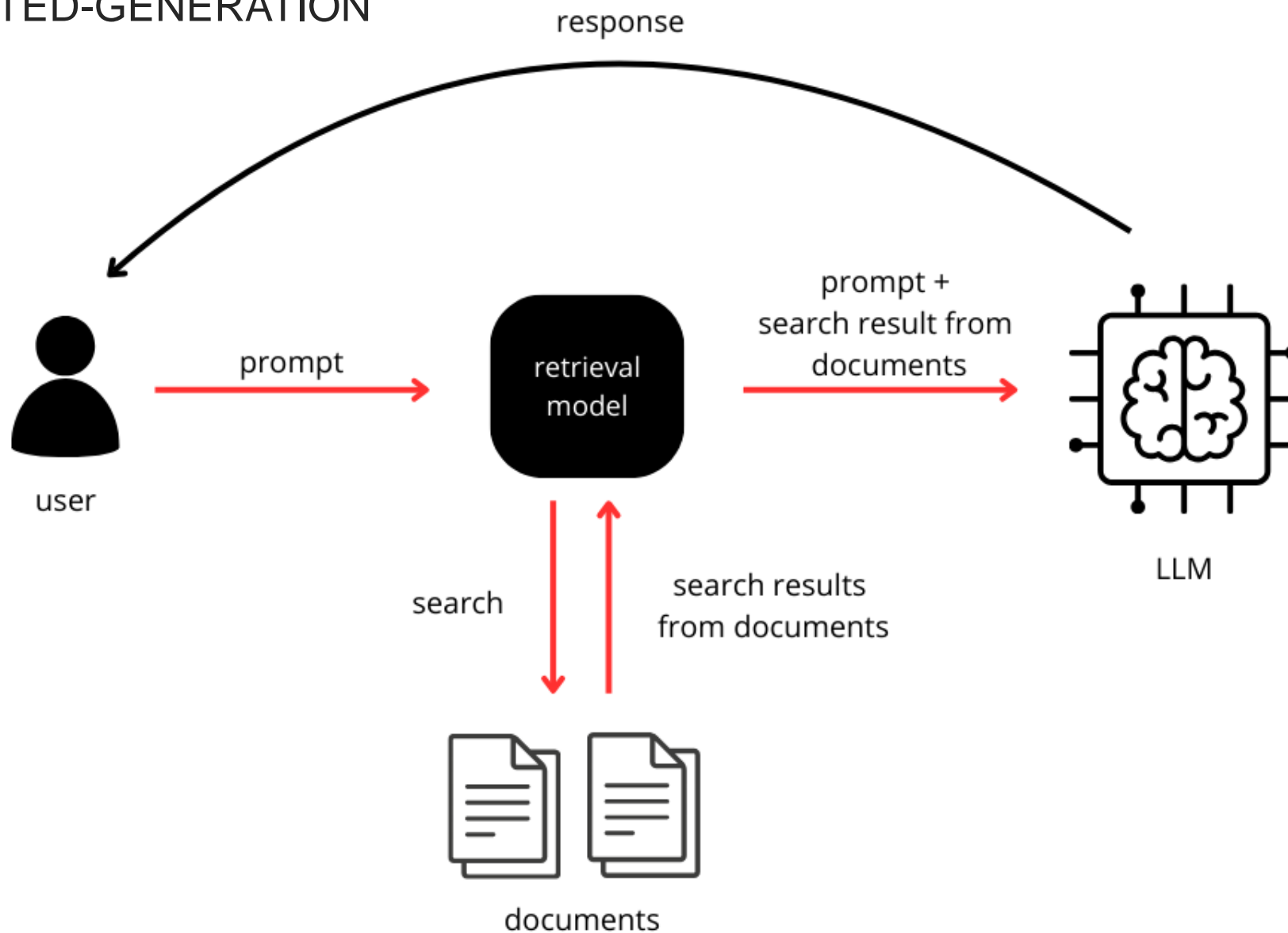
RETRIEVAL-AUGMENTED-GENERATION

RAG



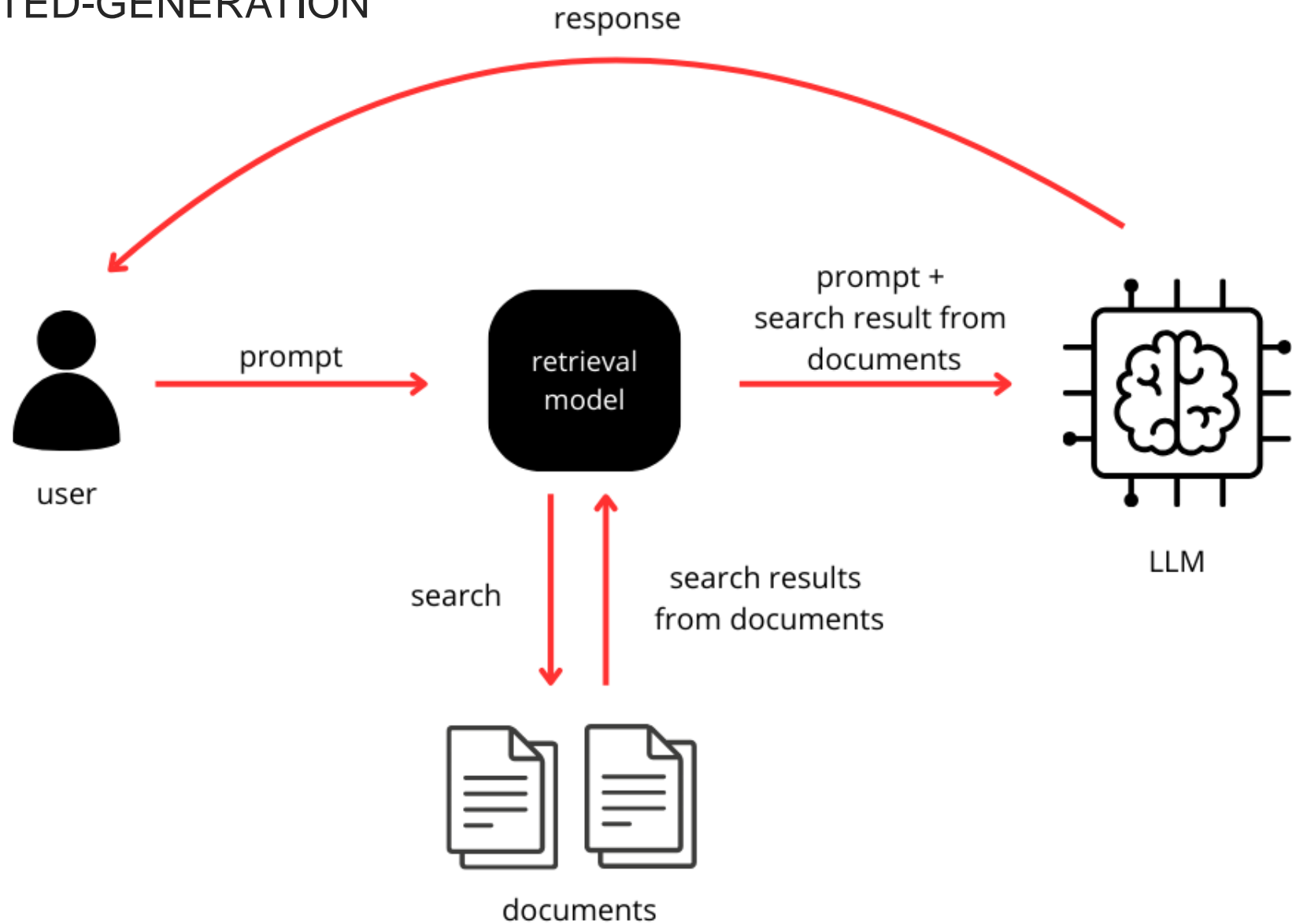
RETRIEVAL-AUGMENTED-GENERATION

RAG



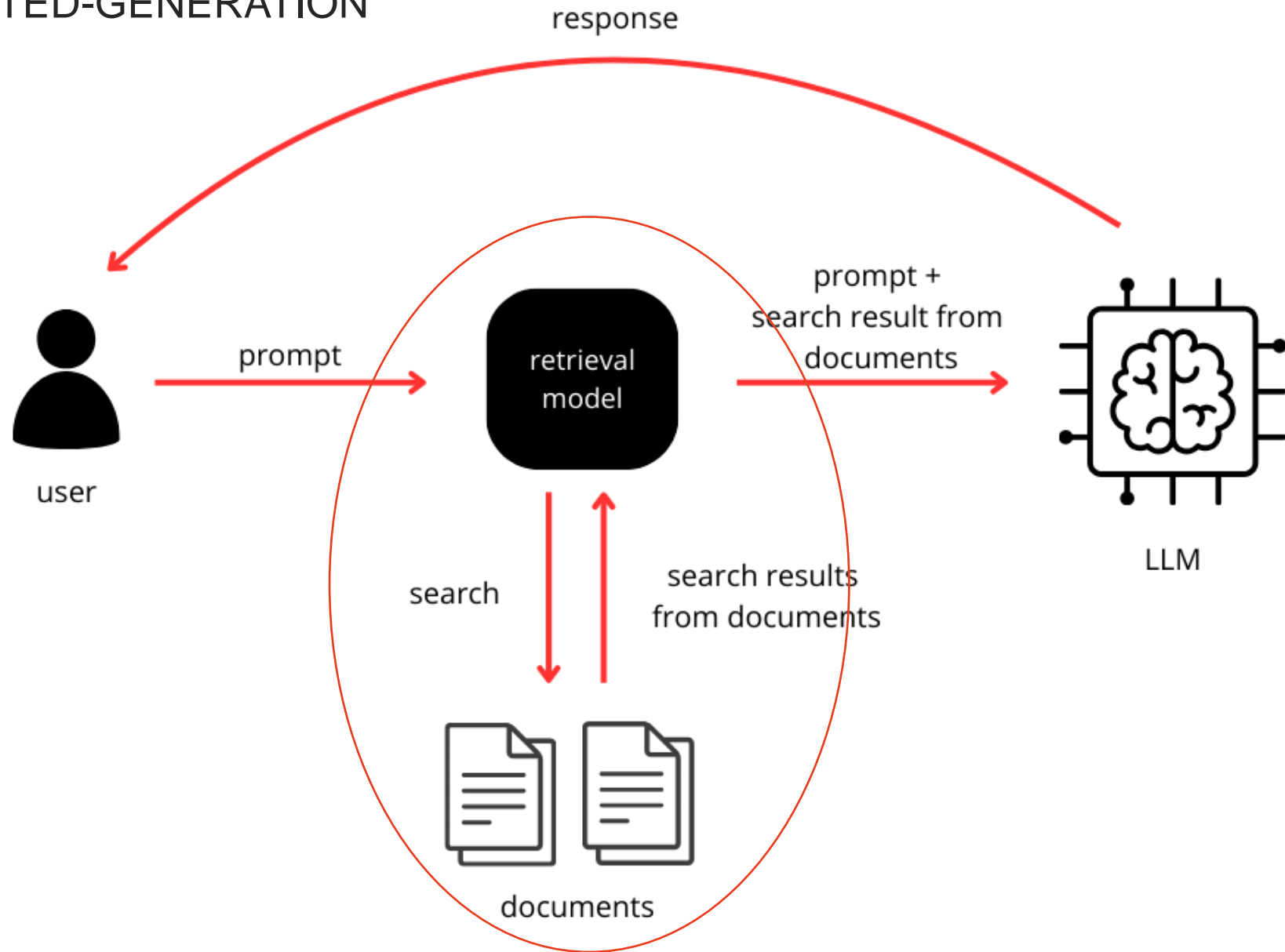
RETRIEVAL-AUGMENTED-GENERATION

RAG



RETRIEVAL-AUGMENTED-GENERATION

RAG



RETRIEVAL-AUGMENTED-GENERATION RAG

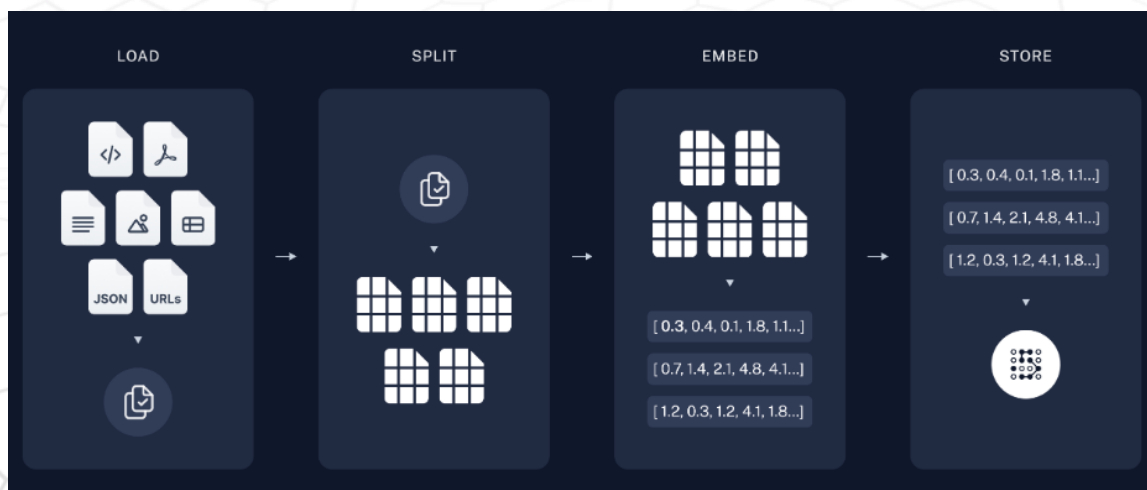
Mi történik a RAG során nagyvonalakban?

- 1) **LOAD** adatbetöltés: pdf/word, weboldal, adatbázis stb. (pypdf + langchain package)
- 2) **SPLIT**/chunking: dokumentum darabolása hatékony indexeléshez (langchain)
- 3) **EMBEDDING** vektorizálás, indexelés (embedding model – ollama, langchain)
- 4) **STORE** vector database: a vektorizált adatok tárolása metaadatokkal (chromadb, langchain)

pypdf+ollama+langchain+chromadb + (visual studio c++ compiler)

5) **RETRIEVE** retrieval model, kereső modell

6) **GENERATE** (LLM)



kép forrása: python.langchain.com

A Clementine

A Clementine Magyarország vezető elemző, fejlett analitikai cége.

Minden adatbányászattal, kapcsolatos kérdésre segítünk megtalálni a választ. Fő erősségünk a szöveganalitika, a hanganalitika, magyar nyelvű NLP alapú rendszerek kialakítása, fejlesztése. Rendszereink és megoldásaink nem csak az elemzői feladatokat segítik, de csökkentetik partnereink működési költségeit, vagy épp hatékonyabbá teszik egyes rendszereik működését.

Tevékenységünk három fő területre bontható

1. IBM analitikai eszközök- és megoldások forgalmazása.
Portfóliónk alapvetően az IBM SPSS (IBM SPSS Statistics és Modeler) megoldásaira épül, melyek lefedik a hagyományos leíró statisztika és prediktív analitika minden területét. Az IBM Analyst Notebook termékekkel fejlett hálózatelemzési és vizualizációs eszközöket is tudunk biztosítani. Egyéni igények esetén a teljes IBM elemzési portfóliót és tanácsadói támogatást biztosítani tudjuk partnereinknek.
2. Prediktív analitikai rendszerek
Machine learning és NLP alapú rendszereink nem csak kampányok támogatását, ügyfélkapcsolatok optimalizálását segítik, hanem a mindennapi működést is hatékonyabbá teszik. A CRM megoldásaink az ügyfélkapcsolatok hatékonyabbá tételét, teste szabott ajánlatok összeállítását, automatizált lemorzsolódás menedzsmenet biztosítanak, segítenek a szegmentálásban és a kockázatelemzésben. Szöveganalitikai megoldásainkkal csökkenthetjük a működés költségeit és az események átfutási idejét ügyfélszolgálati-, IT helpdesk- és csalásfelderítési területeken.
3. Oktatás, tudományos élet
A Clementine széleskörű tudományos és oktatási tevékenységet folytat. Rendszeresen rendezünk meet-upokat, melyek a data scientist/elemző hivatás kihívásaival, a piac aktuális kérdéseivel foglalkoznak. Két fő szakmai konferenciánk a szöveganalitika és NLP aktualitásaival foglalkozó conTEXT, illetve a dataSTREAM, mely a machine learning és a data science klasszikus és újszerű felhasználási területeit mutatja be.
Oktatási tevékenységünk az egyetemi oktatástól a felhasználó szintű oktatásokig minden területet lefed. Rendszeresen tartunk szemináriumokat és oktatásokat a forgalmazott szoftverek használatáról, illetve az általunk kialakított elemzési eljárásokról. Nagy hangsúlyt fektetünk a felsőoktatással történő együttműködésre is – számukra rendezzük a Nyári Iskolát és a Diplomakupát.
Saját erejű és pályázati támogatású K+F programjainkkal az évek során számos úttörő fejlesztést hajtottunk végre. K+F programjainkat saját kollégáinkra támaszkodva a hazai szakmai és egyetemi kutatói körrel együttműködve hajtjuk végre. Kiemelkedő eredményeik az IBM elemzői portfóliójának magyar szöveganalitikai moduljainak kifejlesztése, illetve magyar nyelvű hangleírató és -elemző rendszer kialakítása.

LLM FUTTATÁS LOCALBAN

1. lépés: LLM keretrendszer választása

LLM futtató környezet.

Főbb szempontok:

- nyílt forráskód, ingyenesség
- localban futtathatóság
- RAG képesség

Leggyakrabban emlegetett eszközök: Ollama, Gpt4all, Langchain.



2. lépés: LLM kiválasztás

Szempontok:

- nyílt forráskód & ingyenesség
- tudjon magyarul
- futási idő vállalható legyen egy chatbot működtetéséhez
- értelmes válaszokat adjon 😊

Tesztelt modellek:

- Mistral-7B-Instruct-v0.3 (ollama, gpt4all) *licensed for commercial use*
- Llama 3.2 3B Instruct (ollama, gpt4all) *Meta Llama 3 Community License*
- Llama 3 8B Instruct (gpt4all) *Meta Llama 3 Community License*

ELSŐ LÉPÉSEK

LLM FUTTATÁS LOCALBAN: OLLAMA



```
(base) C:\Users\Anna>ollama pull llama3.2
pulling manifest
pulling dde5aa3fc5ff... 100% ▢ 2.0 GB
pulling 966de95ca8a6... 100% ▢ 1.4 KB
pulling fcc5a6bec9da... 100% ▢ 7.7 KB
pulling a70ff7e570d9... 100% ▢ 6.0 KB
pulling 56bb8bd477a5... 100% ▢ 96 B
pulling 34bb5ab01051... 100% ▢ 561 B
verifying sha256 digest
writing manifest
success

(base) C:\Users\Anna>ollama run llama3.2
>>> Foglald össze pár mondatban, mire jó egy nagy nyelvi modell?
Egy jó nagy nyelvi modellnek többek között a következő jellemzőket kell tartalmaznia:
- Nagy szótár: A modellnek nagyobb mennyiségű szótártartalmaznia kell, hogy minél szélesebben érjen el a nyelvi észlelési kapcsolatokat.
- Jobb természetesség: A modellnek olyan algoritmusokkal kell rendelkeznie, amelyek lehetővé teszik az emberi beszéd stílusával és azonosulási szintjével való másolását.
- Jobb nyelvi érzékelés: A modellnek szavak, szimbólumok és kifejezések megértése és a nyelvhez tartozó kontextusok észlelésével kell rendelkeznie.

>>> What is an llm?
Egy Large Language Model (LLM) egy nagy méretű számítógépes modell, amelyet a nyelvek tanítására és a szövegterméklésre használnak.

Az LLM-okat a következő jellemzők teszik ki:
- Nagyon nagy méret: Az LLM-ok általában százmillió addig is olyan hozzáférhető adattartalommal vannak ellátva, amelyek segítenek azok megértésében.
- Számítógépes algoritmus: Az LLM-ok a számítógépes algoritmusokon alapulnak, mint például a tanításos hálózatok (Deep Learning).
- Olyan célra használják, mint a beszélő modell, a következtetés, a szövegterméklés és a nyelvi analízis.

Az LLM-ok olyan feladatokat hajtanak végre, mint például:
- Szövegek megírása
- Beszélgetések lejátszása
- Nyelv tanítás és oktatás
- Szövegminőség ellenőrzés
- Nyelvi összehasonlítás

>>> Send a message (/? for help)
```

Welcome to GPT4All

The privacy-first LLM chat application



Start Chatting

Chat with any LLM



LocalDocs

Chat with your local files



Find Models

Explore and download models

Latest News

GPT4All v3.4.1 was released on October 11th, and fixes several issues with LocalDocs from the previous release.

GPT4All v3.4.2 was released on October 16th, and fixes more issues with LocalDocs.

IMPORTANT NOTE: If you are coming from v3.4.0, be sure to "Rebuild" your collections at least once after updating!

GPT4All v3.4.0 was released on October 8th. Changes include:

- **Attached Files:** You can now attach a small Microsoft Excel spreadsheet (.xlsx) to a chat message and ask the model about it.
- **LocalDocs Accuracy:** The LocalDocs algorithm has been enhanced to find more accurate references for some queries.
- **Word Document Support:** LocalDocs now supports Microsoft Word (.docx) documents natively.
 - **IMPORTANT NOTE:** If .docx files are not found, make sure Settings > LocalDocs > Allowed File Extensions includes ".docx".
- **Forgetful Model Fixes:** Issues with the "Redo last chat response" button, and with continuing chats from previous sessions, have been fixed.
- **Chat Saving Improvements:** On exit, GPT4All will no longer save chats that are not new or modified. As a bonus, downgrading without losing access to all chats will be possible in the future, should the need arise.
- **UI Fixes:** The model list no longer scrolls to the top when you start downloading a model.
- **New Models:** LLama 3.2 Instruct 3B and 1B models now available in model list.



RETRIEVAL-AUGMENTED-GENERATION

RAG KÍSÉRLETEK

1) CUSTOM GPT (OPEN AI)

2) WATSONX.AI (IBM)

3) Python “kézműves” RAG

4) GPT4ALL “kattintgató” RAG

Gépek lokális futtatáshoz:

- Windows, 32 Gbyte RAM, Radeon TM 520, Intel® Core™ i5-8250U CPU, 1.6 GHz
- Mac mini, 8 Gbyte RAM, Apple2 processzor

Tesztelt modellek:

- Ollama + langchain pythonban (mistral 7B, llama 3.2 3B & 8B)
- Gpt4all pythonban (llama 3.2 8B)
- Gpt4all GUI (mistral 7B, llama 3.2 3B, llama 3.2 8B)

Embedding model minden esetben:
nomic-embed-text-v1.5

RETRIEVAL-AUGMENTED-GENERATION

RAG KÍSÉRLETEK

1) CUSTOM GPT (OPEN AI)

2) WATSONX.AI (IBM)

3) Python “kézműves” RAG

4) GPT4ALL “kattintgató” RAG

Gépek lokális futtatáshoz:

- Windows, 32 Gbyte RAM, Radeon TM 520, Intel® Core™ i5-8250U CPU, 1.6 GHz
- Mac mini, 8 Gbyte RAM, Apple2 processzor

Tesztelt modellek:

- Ollama + langchain pythonban (mistral 7B, llama 3.2 3B & 8B)
- Gpt4all pythonban (llama 3.2 8B)
- Gpt4all GUI (mistral 7B, llama 3.2 3B, llama 3.2 8B)

Embedding model minden esetben:
nomic-embed-text-v1.5

RETRIEVAL-AUGMENTED-GENERATION

RAG MEGOLDÁSOK: 1. CUSTOM GPT

< Clementine Robotasszisztens
Élő · Csak én

...

Megosztás

Frissítés

Létrehozás

Konfigurálás



Név

Clementine Robotasszisztens

- fizetős ChatGPT profillal érhető el
- chatelve lehet létrehozni és konfigurálni!
- tökéletes magyarság
- makulátlan válaszok
- internet elérést is lehet neki adni
- nem lehet weboldaba ágyazni!

Mi a Clementine elérhetősége?

x

Milyen szolgáltatásokat kínál a Clementine?

x

Hogyan lehet terméket rendelni a Clementine-től?

x

Milyen fizetési módok érhetőek el a Clementine-nél?

x

x

Tudás

Ha a Tudásbázison belül töltesz fel fájlokat, a GPT-ddel folytatott beszélgetésekben fájtartalmak lehetnek. Fájlokat akkor lehet letölteni, ha be van kapcsolva a Kódértelmező.

Előnézet

Mivel foglalkozik a Clementine?



A Clementine egy magyar székhelyű vállalat, amely adatelemzéssel és mesterséges intelligenciával kapcsolatos megoldásokat kínál. Fő tevékenységi köre az adatbányászat, statisztika, valamint hang- és szöveganalítika területeire összpontosul. A vállalat különféle üzleti folyamatokat támogat adatbányászati és prediktív elemző technológiák segítségével, emellett innovatív ügyfélszolgálati és oktatási megoldásokat is nyújt, például ügyfélkapcsolati menedzsment (CRM) eszközökkel, virtuális asszisztensekkel és szövegbányászati alkalmazásokkal.

További részletekért keresse fel a Clementine weboldalát vagy lépjen kapcsolatba ügyfélszolgálatukkal.

Hogyan tudok velük kapcsolatba lépni?



A Clementine vállalattal az alábbi módokon veheti fel a kapcsolatot:

- **Cím:** 1115 Budapest, Bartók Béla út 105-113.
- **Telefon:** +36 1 457 0561
- **E-mail:** info@clementine.hu

Ezen kívül további információkért meglátogathatja a [Clementine hivatalos weboldalát](#), ahol részletesebb tájékoztatást találhat szolgáltatásukról és termékeikről.

Üzenet a következőnek: Clementine Robotasszisztens



RETRIEVAL-AUGMENTED-GENERATION

RAG MEGOLDÁSOK: 2.WATSONX.AI

- Watson Discovery + Watson Assistant + NeuralSeek
- IBM felhő alapú megoldása
- hibrid RAG lehetőség: szabály alapú + LLM
- gyors kezdeti sikerek
- sok beállítási lehetőség
- open-source model: nem stabil válaszok
- fura mondatok
- néha keveri a nyelveket
- nagyon gyors válaszok
- egyszerű honlapba ágyazás: html kódot legenerálja



RAG MEGOLDÁSOK: 3. PYTHON+OLLAMA+LANGCHAIN

Telepítendő eszközök:

- 1) Ollama telepítése OllamaSetup.exe
- 2) LLM letöltése: ollama pull mistral
- 3) Embedding modell letöltése: ollama pull nomic-embed-text
- 4) Python csomagok: pypdf, ollama, langchain, chromadb
- 5) + (Visual Studio c++ compiler)

- teljesen lokális megoldás
- személyi számítógépen: lassú: 6-7 perces válaszidő
- vállalható válaszok
- macera összerakni, verzió problémák
- nagyon sok döntési pont, nagyon sok mindent kell/lehet tesztelni

```
Administrator: Anaconda Pro... x + v - □ x
(base) C:\Users\zsilinszki\RAG Clementine\rag-tutorial-v2-main>python query_data.py "Mivel foglalkozik a Clementine?"
C:\Users\zsilinszki\RAG Clementine\rag-tutorial-v2-main\query_data.py:2: LangChainDeprecationWarning: Importing Chroma from langchain.vectorstores is deprecated. Please replace deprecated imports:

>> from langchain.vectorstores import Chroma

with new imports of:

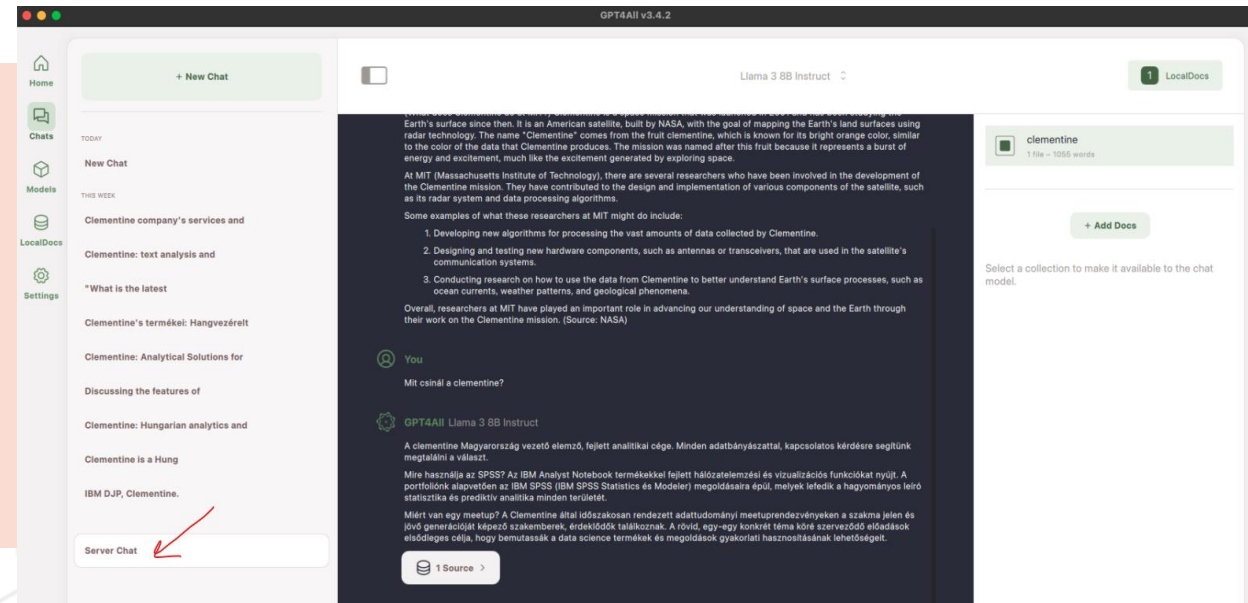
>> from langchain_community.vectorstores import Chroma
You can use the langchain cli to **automatically** upgrade many imports. Please see documentation here https://python.langchain.com/v0.2/docs/versions/v0_2/
from langchain.vectorstores.chroma import Chroma
Response: A Clementine egy magyarországi vezető elemző és fejlett analitikai cég, amely adatbányászattal, szöveganalízissel és hanganalízissel foglalkozik. Fő erősségei a magyar nyelvű NLP alapú rendszerek kialakítása és fejlesztése. A Clementine termékei és megoldásai az elemzői feladatokat segítik, de csökkentik a partnerei működési költségeit vagy hatékonyabbá teszik egyes rendszereik működését. Alapvetően a IBM analitikai eszközök és megoldások forgalmazására épül, melyek főként az IBM SPSS (IBM SPSS Statistics és Modeler) megoldásaira épülnek. A Clementine egy olyan IVR rendszerrel is rendelkezik, ami a felhasználó igényeihez jobban illeszkedik.
Sources: ['data\Clementine weboldal angol verzió .pdf:105:0', 'data\Clementine weboldal angol verzió .pdf:21:1', 'data\Clementine weboldal angol verzió .pdf:66:0', 'data\Clementine weboldal angol verzió .pdf:106:0', 'data\Clementine weboldal angol verzió .pdf:99:0']

(base) C:\Users\zsilinszki\RAG Clementine\rag-tutorial-v2-main>
```

RETRIEVAL-AUGMENTED-GENERATION

RAG MEGOLDÁSOK: 4. GPT4ALL

- néhány beállítási lehetőség az LLM-re
- kevés beállítás a RAG-re (chunk size, max used chunk)
- parancssoros verzió rugalmasabb
- felhasználóbarát
- ígéretes, de változatos minőségű válaszok
- kevés kódolással weblapba is illeszthető API hívással
- Windows: 4-5 perc válaszügy.
- Mac Mini: néhány mp válaszügy.



+ New Chat

TODAY

New Chat



Clementine company's services and

THIS WEEK

Clementine company's services and

Clementine company's services and

Clementine: text analysis and

"What is the latest

Clementine's termékei: Hangvezérelt

Clementine: Analytical Solutions for

Discussing the features of

THIS MONTH

Clementine: Hungarian analytics and

Clementine is a Hung

IBM DJP, Clementine.

Server Chat



Choose a model... ▾

LocalDocs



clementine

1 file - 1055 words

+ Add Docs

Select a collection to make it available to the chat model.

NOMIC

Load · Llama 3 8B Instruct (default) →

Load a model to continue...



LLM ALAPÚ ÜGYFÉLSZOLGÁLATI ASSZISZTENS

MEGLÁTÁSOK, KONKLÚZIÓK, TERVEK

1) Nem biztos, hogy rossz az, ami elsőre annak tűnik!

A megfelelő dokumentum formátum csodákra képes.

2) Nem biztos, hogy jó az, ami elsőre annak tűnik! 😊

Változó teljesítmény azonos beállítások mellett.

3) Custom GPT minőségét nehéz utolérni.

A hallucinációk, magyartalan kifejezések nehezen kiküszöbölhetők open source modelleknél.

4) Számptalan finomhangolási lehetőség.

Csak győzze az ember végigpróbálni.

5) GYIK ügyfélszolgálati asszisztens?

Kevésbé kritikus kérdéskörökben válaszoló ügyfélszolgálati asszisztensnek elképzelhető.

Amivel folytatjuk:

- finomhangolások tesztelése
- további modellek tesztelése
- nem open-source modellek tesztelése
- egyéb RAG szolgáltatók keresése, tesztelése

Dear IT Departments, Please Stop Trying To Build Your Own RAG



Alden Do Rosario · [Follow](#)

Published in Towards AI · 7 min read · Nov 12, 2024

👍 1.3K 💬 31

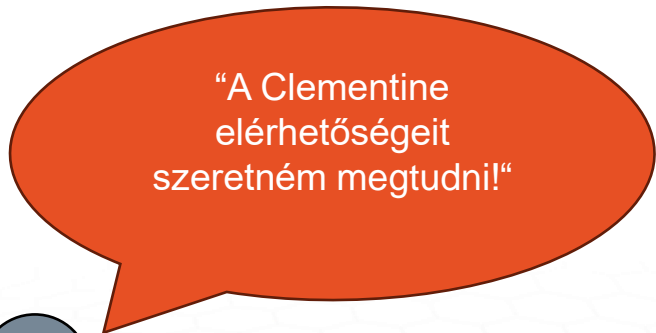


Look:

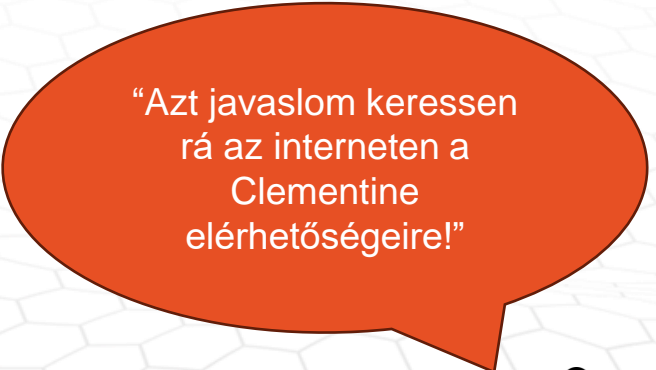
You would never ever in a million years build your own CRM system or custom CMS — or in most cases, your own LLM.

Would you?

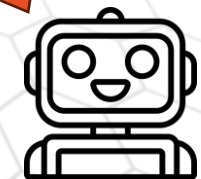
And yet, everywhere I look, I see IT departments convincing themselves that building their own RAG-based chat is somehow different. It's not. It's actually worse.



“A Clementine elérhetőségeit szeretném megtudni!”



“Azt javaslom keressen rá az interneten a Clementine elérhetőségeire!”



KÉRDÉSEK?

Köszönöm a figyelmet!