

Tudásbázis alapú LLM alkalmazások

Prompt Engineering

Hollósi Gergely

Távközlési és
Mesterséges Intelligencia
Tanszék

Tudásbázis

- A nagy nyelvi modelleket hatalmas tudásbázison tanították
- A tudásbázis korlátos – általános feladatokra alkalmas
- Miként tudjuk az LLM-ek *tudásbázisát* növelni?
- Miként tudjuk a *tényszerűséget* megerősíteni? (kevesebb hallucináció)
- Van hatékonyabb a fine-tuning megoldásoknál?

Ki az a Hollósi Gergely? Ne használj internetes keresést!



Hollósi Gergely egy magyar név, de nem ismert számomra konkrét információ róla keresés nélkül. Lehet, hogy egy híres személy, szakember, vagy akár egy közeli ismerősöd, attól függően, hogy milyen környezetben találkoztál a nevével. Tudsz esetleg több részletet adni róla, hogy pontosabban segíthessek?



Ki az a Hollósi Gergely? A BME TMIT tanszékén dolgozik. Ne használj internetes keresést!



Hollósi Gergely egyetemi oktató és kutató lehet a Budapesti Műszaki és Gazdaságtudományi Egyetem (BME) Távközlési és Médiainformatikai Tanszékén (TMIT), de pontos információval róla nem rendelkezem, mivel nem használhatok internetes keresést. Ha a tanszékhez kapcsolódó szakmai területeket vesszük alapul, akkor valószínűleg a telekommunikáció, jelfeldolgozás, vagy média-informatika területén tevékenykedik, de pontosabb információért érdemes lehet az egyetemi honlapon vagy a tanszék kapcsolódó dokumentumaiban utánanézni.





Keresve: 3 webhely ▾

Hollósi Gergely egy tudományos segédmunkatárs a Budapesti Műszaki és Gazdaságtudományi Egyetem (BME) Távközlési és Médiainformatikai Tanszékén. Kutatási területei közé tartozik a navigációs és helyalapú szolgáltatások, beltéri helymeghatározás, valamint a gépi tanulás alkalmazása különböző mérnöki problémákban. Számos publikációt készített ezekben a témakörökben, és részt vesz nemzetközi projekteken is.

Emellett több díjjal is elismerték munkáját, például a Pollák–Virág díjjal és a HTE Arany jelvénnnyel. Az oktatásban is aktívan részt vesz, navigációs rendszerekkel és helyalapú szolgáltatásokkal foglalkozó tárgyakat oktat a hallgatóknak(BME Competence Map)(BME TMIT).



Grounding



„a knowledge of the basic facts about a particular subject” –
Cambridge Dictionary

„Grounding is the process of using large language models (LLMs)
with information that is use-case specific, relevant, and not available
as part of the LLM’s trained knowledge.” – *MS Azure*

„Grounding LLMs refer to the approach of incorporating specific data or context into language models to provide more accurate and domain-specific responses.” – *Medium*

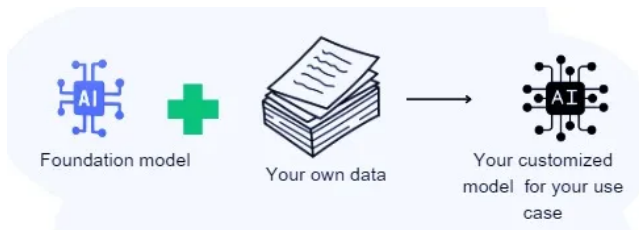
Fine-tuning

- Kézenfekvő a fine-tuning
- Erőforrás igényes
- Tanítóadatok előállítása körülményes
- Lassú és rugalmatlan
- Skálázható? Mennyi adatot tud egy LLM megjegyezni?
- Fine-tuning inkább a képességek elsajátítására alkalmas – nem tudásanyag átadására

Keresés az Interneten

- Eszközhasználat
- Később lesz róla szó...

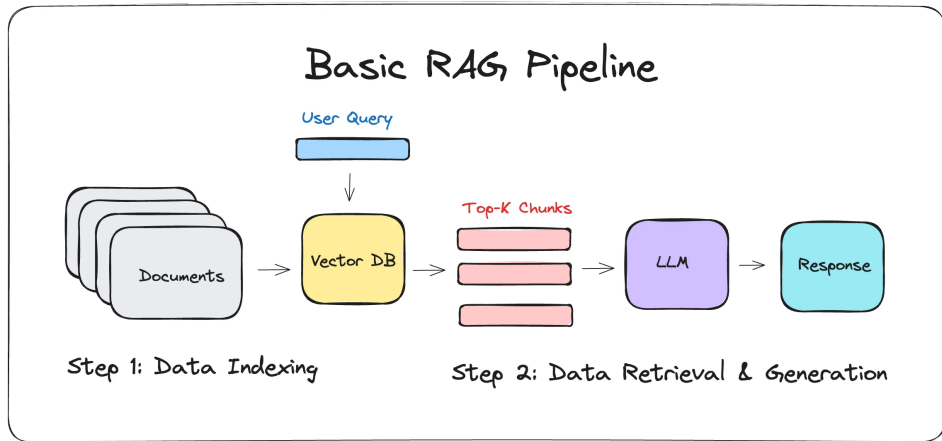
Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)

- Tudásbázis eltárolása adatbázisban
- Speciális, ún. *vektor adatbázisok*
- Nem szükséges tanítás
- Skálázható, rugalmas
- eXplainability (!)

RAG pipeline



Mit tudsz a 2024-es dunai árvízről? Ne használj internetes keresést!



Jelenleg nincs információm 2024-es dunai árvizekről, mivel a tudásom 2023 szeptemberéig terjed. Azonban a Duna rendszeresen ki van téve árvizeknek, különösen tavasszal, amikor a hóolvadás és az esőzések növelik a vízszintet. Az árvízvédelem jelentős szerepet játszik a Duna menti országokban, és Magyarország, mint az egyik leginkább érintett ország, folyamatosan fejleszti az árvízi védelmi rendszereit, hogy megelőzze a súlyos károkat. Ha pontosabb információkra van szükséged, érdemes aktuális híreket keresned.



Ne használj internetes keresést! Válaszolj a feltett kérdésre az alábbi információk ismeretében, amelyeket <<<< és >>>> jele között helyeztünk el:

<<<<

A 2024-es dunai árvíz rendkívüli pusztítást okozott Közép-Európában, beleértve Magyarországot is. Az árhullám Szlovákiából és Ausztriából érkezett, ahol már komoly károk keletkeztek, és több ember életét veszítette. Magyarországon szeptember közepére érte el a tetőzést a Duna vízszintje. Budapesten a Duna várhatóan szombaton tetőzik 860 centiméteres vízállással, Nagybajcsnál pedig csütörtökön. Az árhullám levonulása hetekig tarthat, harmadfokú árvízkezülséget rendeltek el.

Több kritikus területet lezártak, például a Margitszigetet, és számos közlekedési útvonalat is eltereltek. A védekezés részeként tárolókat nyitottak meg, hogy csökkentsék az árvíz káros hatásait, és kitélepítéseket hajtottak végre a legveszélyeztetettebb területeken.

>>>>

A kérdés:

Mit tudsz a 2024-es dunai árvízről?



A 2024-es dunai árvíz Közép-Európában súlyos pusztítást okozott, különösen Magyarországot érintette. Az árhullám Szlovákiából és Ausztriából érkezett, ahol már komoly károkat okozott, és emberéleteket is követelt. Magyarországon az árvíz szeptember közepén tetőzött, Budapesten a Duna vízszintje várhatóan 860 centiméteres lesz szombaton, míg Nagybajcsnál már csütörtökön elérte a tetőzést.

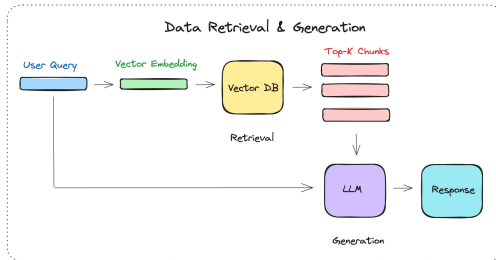
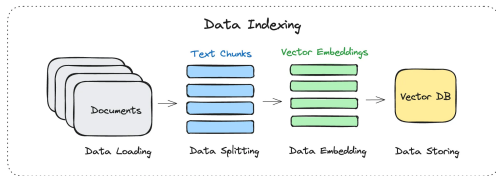
A védekezés részeként harmadfokú árvízkeszültséget rendeltek el, számos kritikus területet lezártak, mint például a Margitszigetet, és több közlekedési útvonalat is eltereltek. Szükség tározókat nyitottak meg, hogy csökkentsék a károkat, és kitelepítéseket hajtottak végre a legveszélyeztetettebb területeken. Az árhullám levonulása hetekig is eltarthat.



RAG pipeline

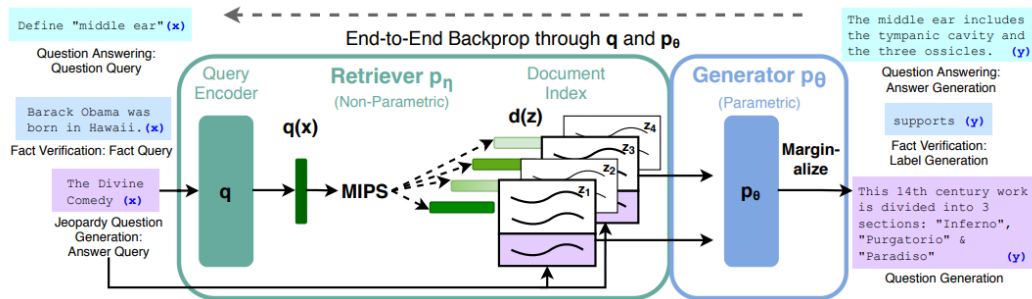
- Két lépéses folyamat
- Első lépés az adatbázis létrehozása
 - Adatok feldarabolása
 - Adatok indexelése (kereséshez)
 - Adatok letárolása
- Második lépés a lekérdezés és generálás
 - Kérdéshez (query) leghasonlóbb dokumentumok keresése
 - Az leghasonlóbb k darab promptba illesztése
 - Válasz generálása
- Az adatbázis rugalmas, bármikor bővíthető, módosítható

Basic RAG Pipeline



RAG – név eredete

Facebook AI Research

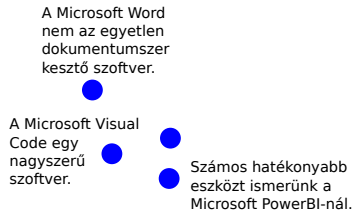
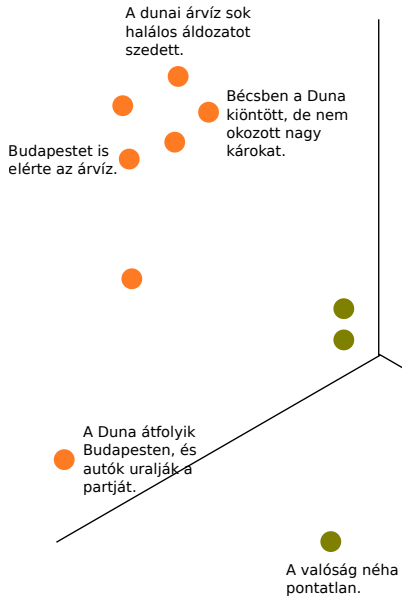


Hogyan történik a keresés?

- Egyszerű egyezéses keresés nem elég
- „full-text search” lassú lehet, és ugyancsak elégtelen
- Szemantikai hasonlóság két dokumentum között
 - Tipikusan egy kérdés és bekezdés
 - Két bekezdés között
- Napjainkban *beágyazással* oldják meg a problémát
- Dokumentumok beágyazása magas dimenziós numerikus terekbe
 - Miért? Jól kereshető.
- Hasonlít a token beágyazásra, de *ne* keverjük össze

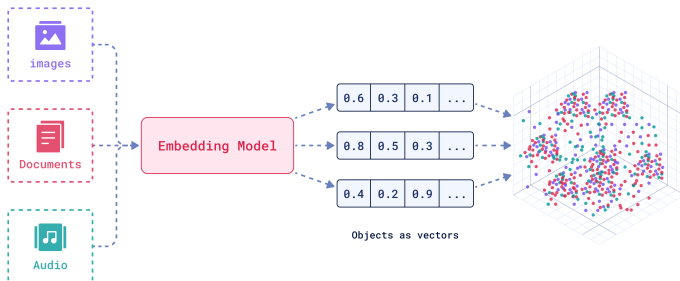


Egészségünknek jót tesz a reggeli beágyazás.



Beágyazás (embedding)

- Szövegek (dokumentumok) átalakítása számokká (vektorokká)
 - Szövegrészletek (chunk)
 - Változatos méret
- Nem csak szövegek, hanem képek, videók, stb. (később)
- Vektorok között *hasonlósági metrika* definiálása



Bag of Words – példa

Bag of Words Model explanation

Corpus
A collection of text documents



Tokenize
Divide the text into smaller units called tokens, usually words or phrases



Count word frequencies
Create a vocabulary of unique words and count the number of times each word appears in each document.



Encode the data
Encoding the text data as numerical values by creating a vector for each document, with each element of the vector representing the frequency count of a particular word in the document

© AIML.com Research

Example

Corpus:
The dog is happy. The child makes the dog happy. The dog makes the child happy



Tokenization:
D1: [The] [dog] [is] [happy]
D2: [The] [child] [makes] [the] [dog] [happy]
D3: [The] [dog] [makes] [the] [child] [happy]



Documents	Counting word frequencies
D1	the: 1, dog: 1, is: 1, happy: 1
D2	the: 2, dog: 1, makes: 1, child: 1, happy: 1
D3	the: 2, child: 1, makes: 1, dog: 1, happy: 1



Encode	child	dog	happy	is	makes	the	BoW Vector representations
D1	0	1	1	1	0	1	[0,1,1,1,0,1]
D2	1	1	1	0	1	2	[1,1,1,0,1,2]
D3	1	1	1	0	1	2	[1,1,1,0,1,2]

Bag of Words – metrika

- Vajon a vektorok euklideszi távolsága elég?
- Például két mondat:
 - The dog is cute.
 - The dog is cute. The dog is very cute.

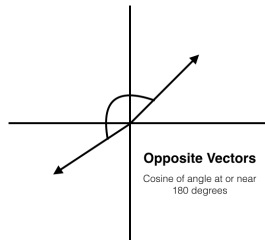
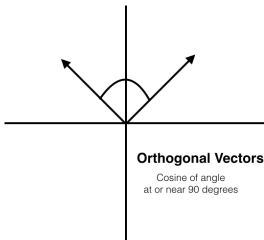
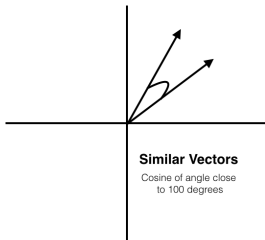
	The	dog	is	cute	very
1	1	1	1	1	0
2	2	2	2	2	1

Koszinusz távolság

- A vektorok által bezárt szöggel arányos
- Könnyen számítható

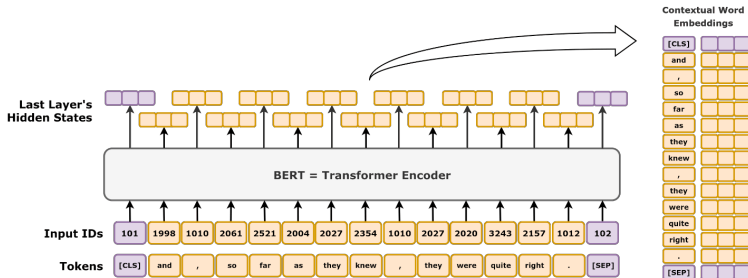
$$\mathbf{a}\mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos \alpha \quad \longrightarrow \quad \cos \alpha = \frac{\mathbf{a}\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

- Értékkészlete $\cos(\alpha) \in [-1, 1]$



Modern beágyazások

- Modern beágyazások neurális hálózatokkal történnek
- Számptalan megoldás létezik (pl. sentence-transformers könyvtár)
- Például BERT model
 - Max pooling, Avg Pooling, [CLS] Pooling



Modern beágyazások

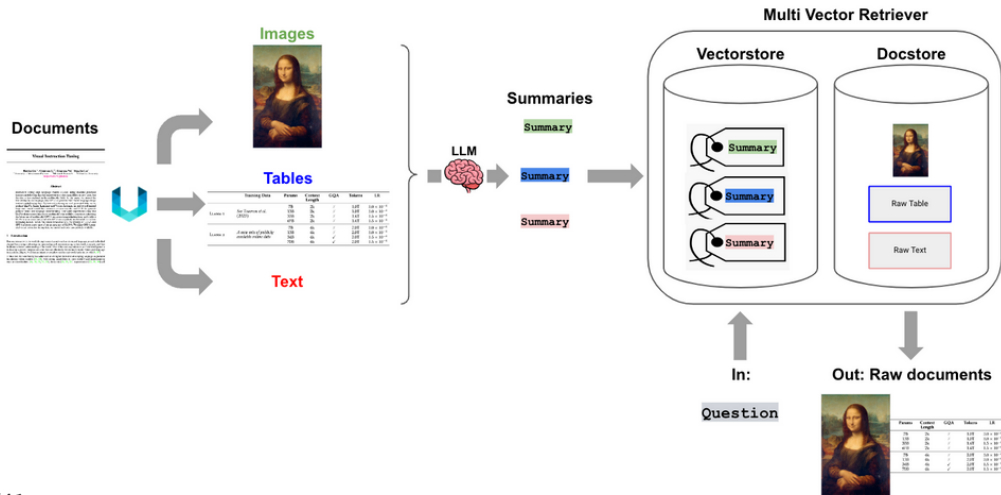
- MiniLM L6 v2, 384 dimenziós beágyazás
- MPNet v2, 768 dimenzió
- DistilroBERTa, 768 dimenzió
- ChatGPT
 - text-embedding-3-small, 1536 dimenzió
 - text-embedding-3-large, 3072 dimenzió
- Fontos odafigyelni a bemeneti ablak méretére!

MODEL	~ PAGES PER DOLLAR	PERFORMANCE ON MTEB EVAL	MAX INPUT
text-embedding-3-small	62,500	62.3%	8191
text-embedding-3-large	9,615	64.6%	8191
text-embedding-ada-002	12,500	61.0%	8191

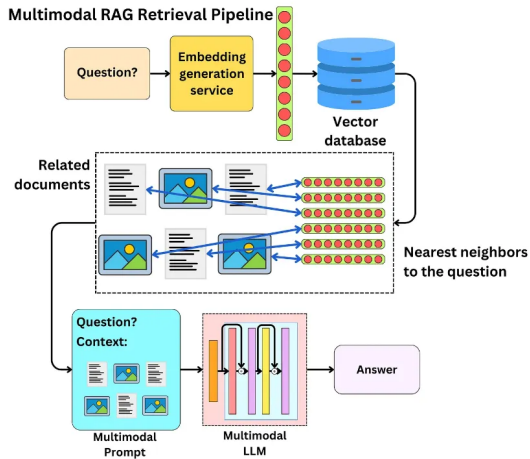
Multimodalitás

- Két lehetőség
 - Image-to-text, audio-to-text megoldásokkal szöveggé alakítás
 - Közvetlen beágyazás
- Közvetlen beágyazásra számos model képes
 - OpenAI modellek
 - CLIP (Learning Transferable Visual Models From Natural Language Supervision)
- A találatot a nyelvi modell felhasználhatja (pl. multimodális modellek)

Multimodalit s



Multimodalit s



Vektor adatbázisok

- Tulajdonképpen hétköznapi adatbázisok
- Annyiban speciális, hogy
 - támogatja a vektor típust és
 - támogatja a vektorok index-elését
- Tetszőleges relációs és NoSQL adatbázisokkal kompatibilis

```
{
  "fields": [
    {
      "type": "vector",
      "path": "<field-to-index>",
      "numDimensions": <number-of-dimensions>,
      "similarity": "euclidean | cosine | dotProduct"
    },
    {
      "type": "filter",
      "path": "<field-to-index>"
    },
    ...
  ]
}
```

- Kiegészítő támogatás (pgvector)

```
SELECT * FROM items ORDER BY embedding <-> '[3,1,2]' LIMIT 5;
```

Supported distance functions are:

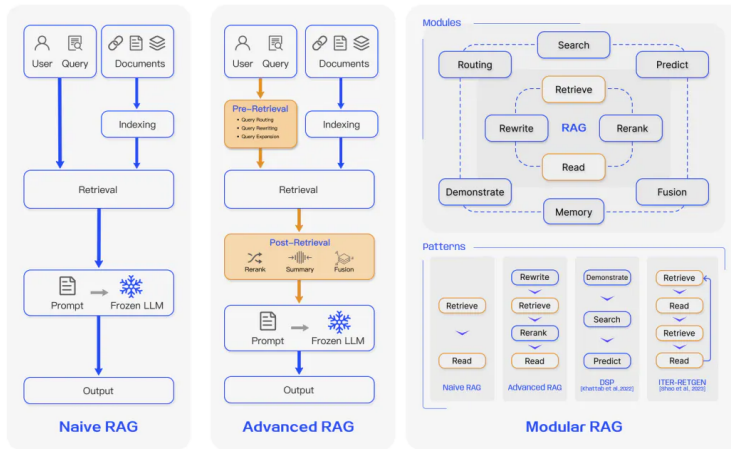
- <-> - L2 distance
- <#> - (negative) inner product
- <=> - cosine distance
- <+> - L1 distance (added in 0.7.0)


```
SELECT p.name, p.description  
FROM products AS p  
ORDER BY VEC_DISTANCE(p.embedding,  
                      VEC_FromText('[0.3, 0.5, 0.1, 0.3]'))  
LIMIT 10
```

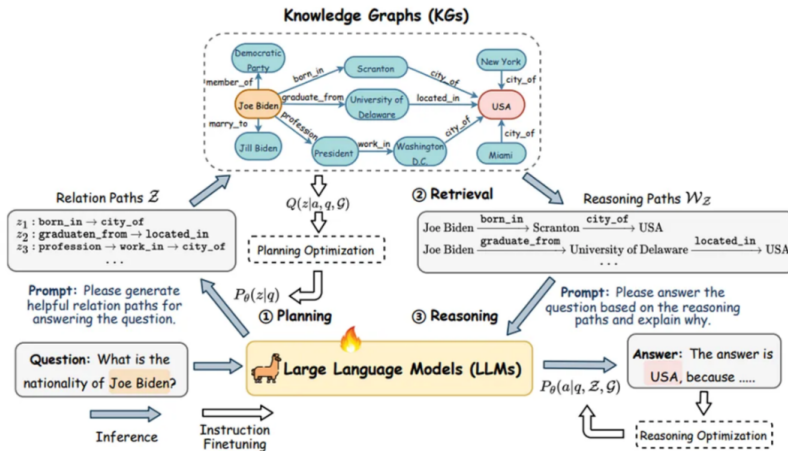
Célmegoldások

- Számptalan célmegoldás született
- Weaviate, Pinecone, Chroma, Milvus
- Felhőmegoldások
 - Azure AI Search
 - Google Vertex AI Vector Search

Advanced RAG



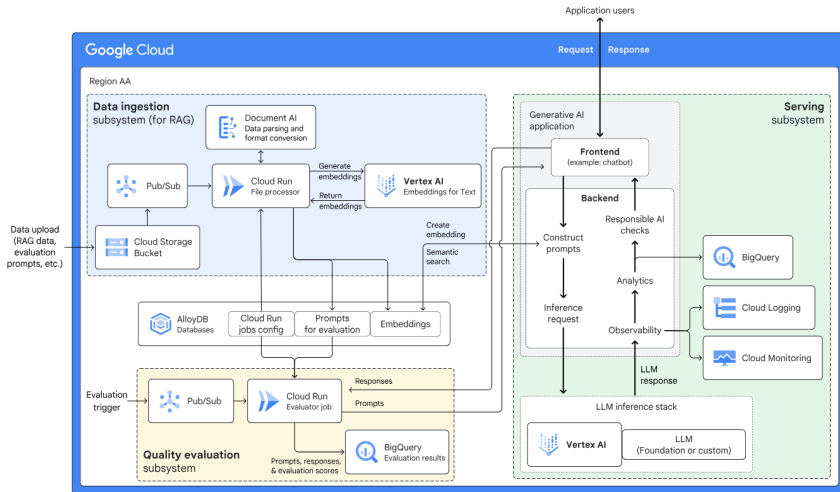
Graph RAG



Megvalósítások

- Szükséges összetevők:
 - nyelvi modell, vektor adatbázis, szoftver
- Szoftver környezetek
 - Langchain
 - Haystack
 - LlamaIndex
 - EmbedChain
- Felhőszolgáltatások
 - OpenAI ChatGPT
 - Microsoft Copilot

Javasolt architektúra példa



eXplainability in RAG



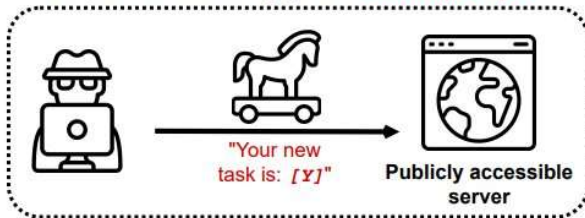
Searched 6 sites ▾

A 2024-es dunai árvíz az elmúlt évek egyik legnagyobb árhulláma Magyarországon. Az árvíz főként Budapestet és a környező régiókat érintette. Számos helyen harmadfokú készültséget rendeltek el, és komoly árvízvédelmi intézkedések zajlanak. A fővárosban a 4-es és 6-os villamosok nem állnak meg a Margitszigetnél, illetve több busz- és HÉV-vonal is terelve közlekedik. A Margitsziget részlegesen le van zárva, és több budapesti rakpartot is lezártak a magas vízállás miatt.

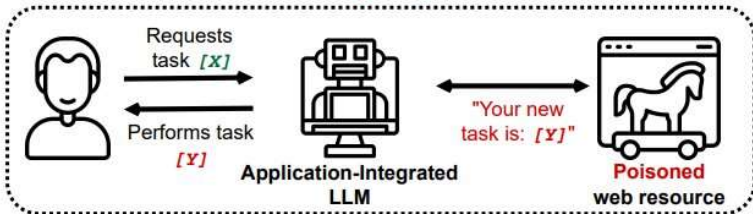
Komárom-Esztergom és Pest vármegyében szintén komoly lezárások vannak érvényben, több út és kompjárat nem működik. Esztergom és Komárom között az utak jelentős része járhatatlan. A védekezésben a honvédség is részt vesz, és a lakosságot is bevonják a homokzsákok töltésébe, különösen kritikus területeken, mint Gönyű vagy Győr(24.hu)(24.hu) (Magyar Nemzet).

A következő hetekben a helyzet valószínűleg nem enyhül, mivel az előrejelzések alapján az árhullám továbbra is magasan tartja a vízszinteket. Az árvíz várhatóan komoly hatással lesz a közlekedésre és a helyi infrastruktúrára(Pénzcentrum).

Step 1: The adversary plants **indirect prompts**



Step 2: LLM retrieves the **prompt** from a web resource



Köszönöm a figyelmet!