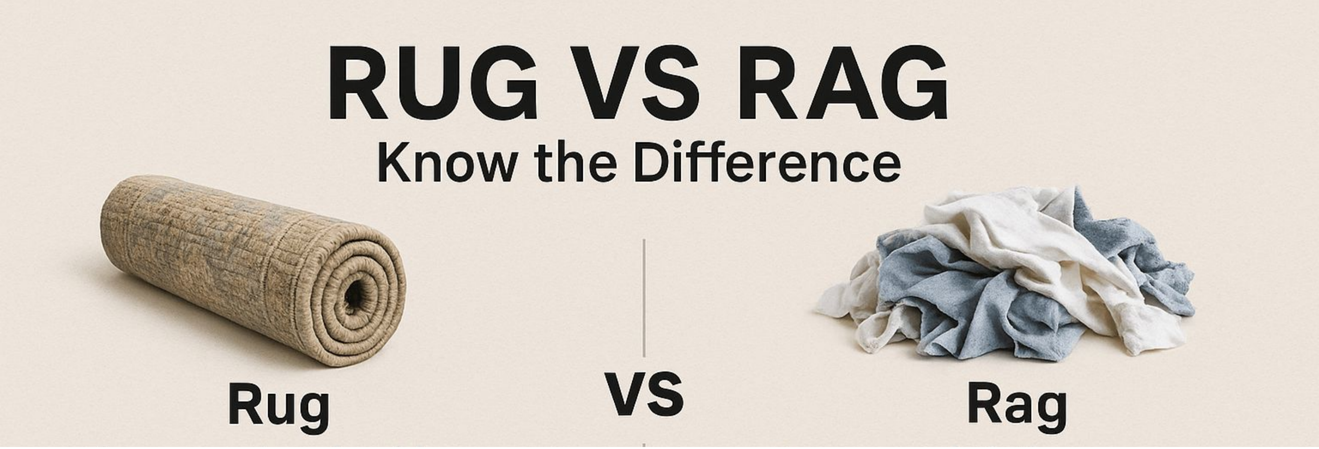
Aldrov/Vítkovice, 2025-09-16

CZDEMOS4AI and RAG



Martin Popel

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

CC BY SA
unless otherwise stated

Aldrov/Vítkovice, 2025-09-16

CZDEMOS4AI and RAG

CZDEMOS4AI = TAČR Sigma (2024–2029) project
RAG = Retrieval-Augmented Generation

Martin Popel

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

unless otherwise stated

CZDEMOS4AI TAČR Sigma (2024–2029) project

https://prg.ai/projekty/czdemos4ai

A beneficial multi-agent AI avatar in a small democratic society

avatar = chatbot (so far)

- **WP1: Historicko-edukační asistent**
- **WP2: Bezpečnostně-mediální asistent**
- **WP3: Právní asistent**
- **WP4: Duševně-zdravotní asistent**

FAKULTA SOCIÁLNÍCH VĚD
Univerzita Karlova

FAKULTA HUMANITNÍCH STUDIÍ
Univerzita Karlova

FILOZOFICKÁ FAKULTA
Univerzita Karlova

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

1. LÉKAŘSKÁ FAKULTA
Univerzita Karlova

Ipsos

ÚSTAV STÁTU A PRÁVA
Akademie věd ČR

prg.ai

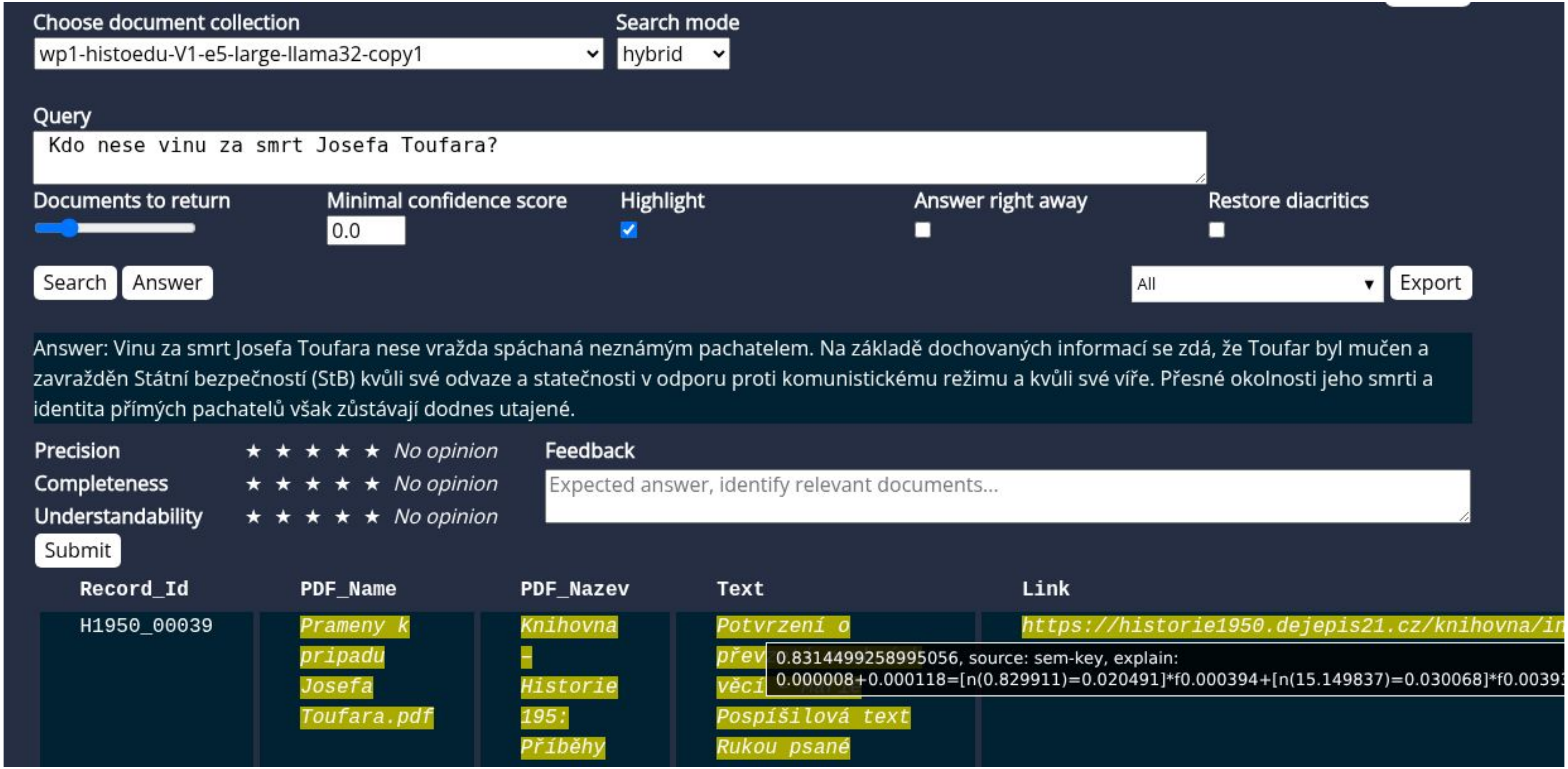
MAMA AI

ČTK ČESKÁ TISKOVÁ KANCELÁŘ

MUNI

PHYSTER

First prototype (MSearch UI by MAMA AI)



Choose document collection: wp1-histoedu-V1-e5-large-llama32-copy1

Search mode: hybrid

Query: `Kdo nese vinu za smrt Josefa Toufara?`

Documents to return

Minimal confidence score: 0.0

Highlight ☑

Answer right away ☐

Restore diacritics ☐

Answer: Vinu za smrt Josefa Toufara nese vražda spáchaná neznámým pachatelem. Na základě dochovaných informací se zdá, že Toufar byl mučen a zavražděn Státní bezpečností (StB) kvůli své odvaze a statečnosti v odporu proti komunistickému režimu a kvůli své víře. Přesné okolnosti jeho smrti a identita přímých pachatelů však zůstávají dodnes utajené.

Precision	★ ★ ★ ★ ★	*No opinion*
Completeness	★ ★ ★ ★ ★	*No opinion*
Understandability	★ ★ ★ ★ ★	*No opinion*

Feedback: Expected answer, identify relevant documents...

Record_Id	PDF_Name	PDF_Nazev	Text	Link
`H1950_00039`	*Prameny k pripadu Josefa Toufara.pdf*	*Knihovna - Historie 195: Příběhy*	*Potvrzení o přev věci Pospíšilová text Rukou psané*	*https://historie1950.dejepis21.cz/knihovna/in*

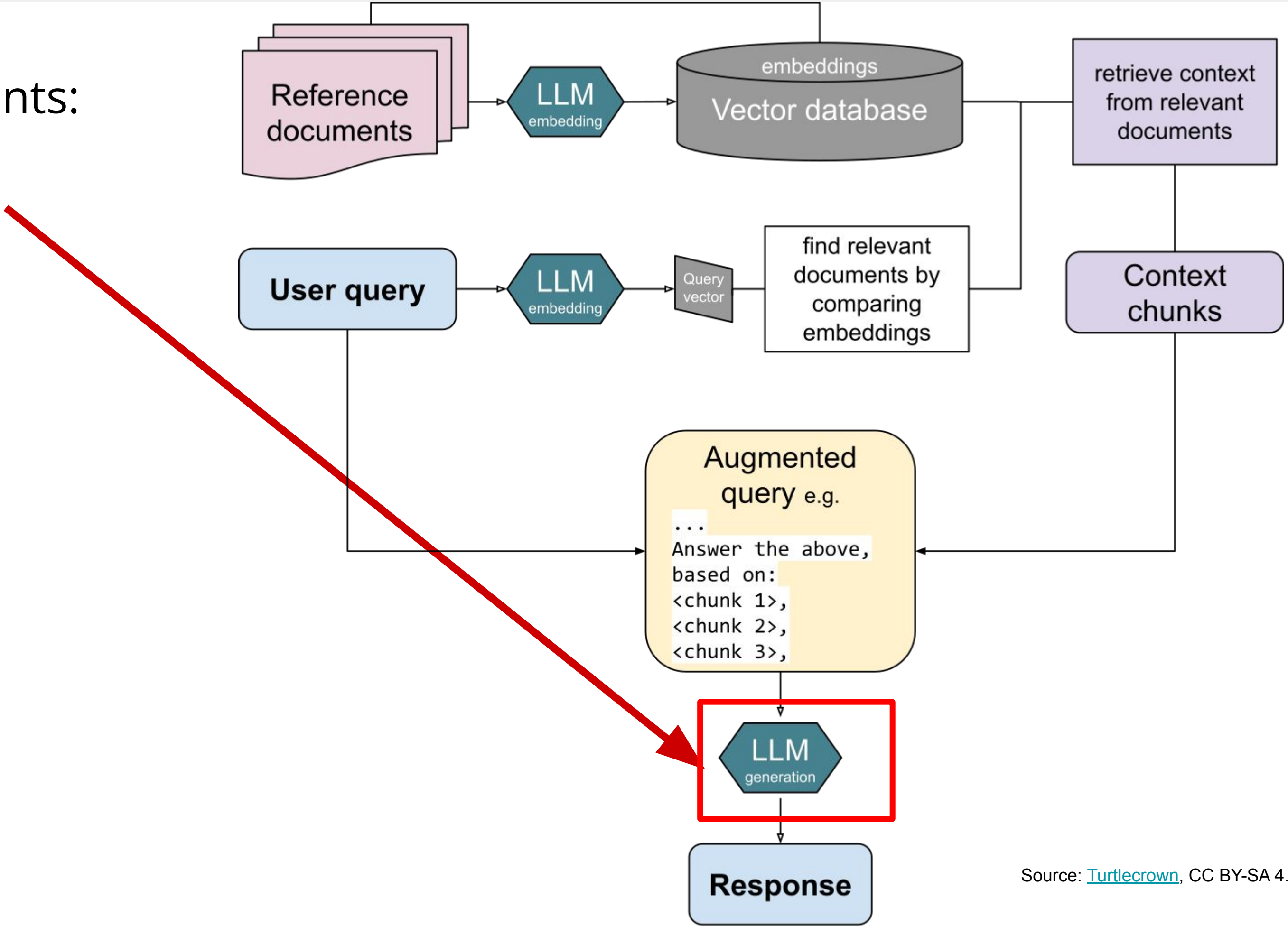
0.8314499258995056, source: sem-key, explain:
0.000008+0.000118=[n(0.829911)=0.020491]*f0.000394+[n(15.149837)=0.030068]*f0.00393

RAG = Retrieval-Augmented Generation

RAG architecture

Three main components:

- Generator (LLM)
-
-

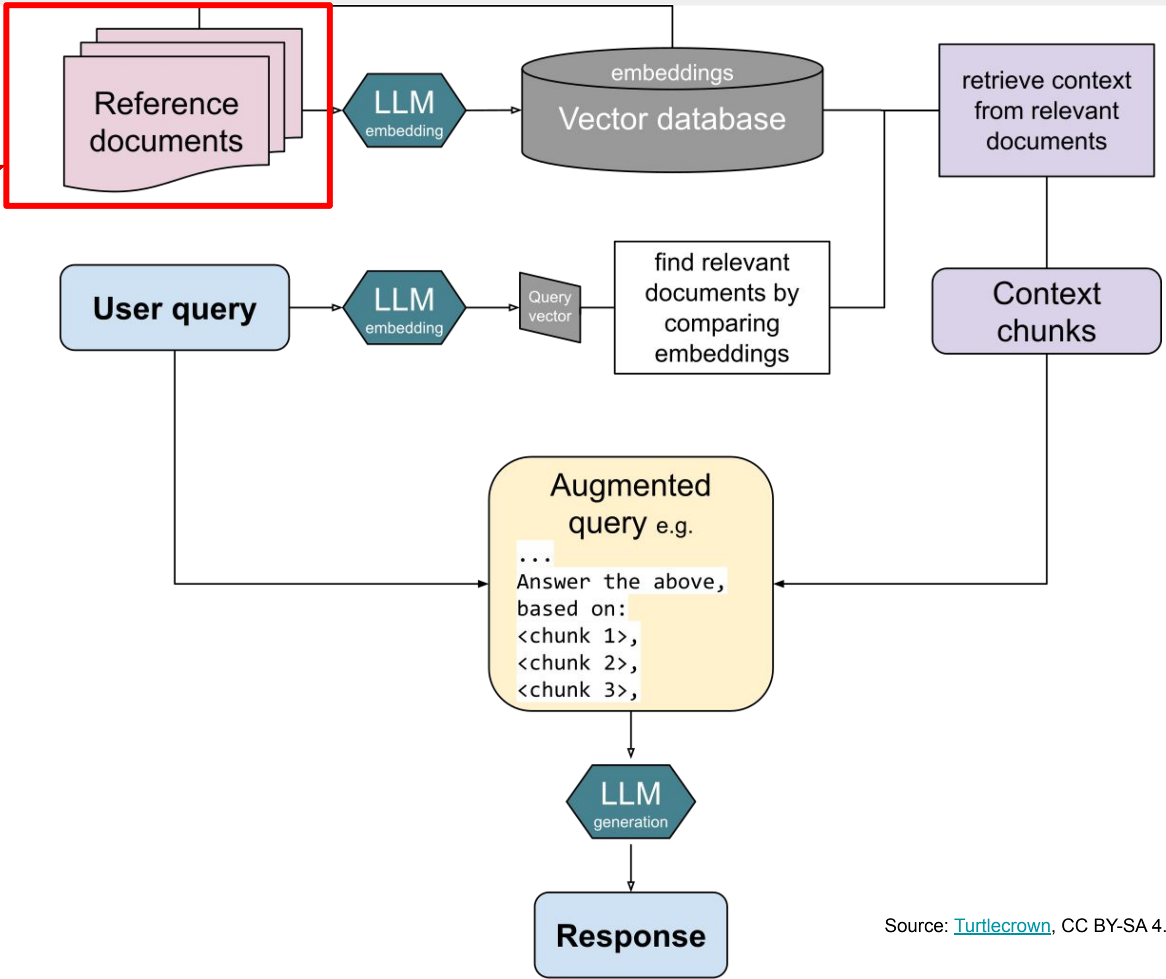


Source: Turtlecrown, CC BY-SA 4.0

RAG architecture

Three main components:

- Generator (LLM)
- Documents
-

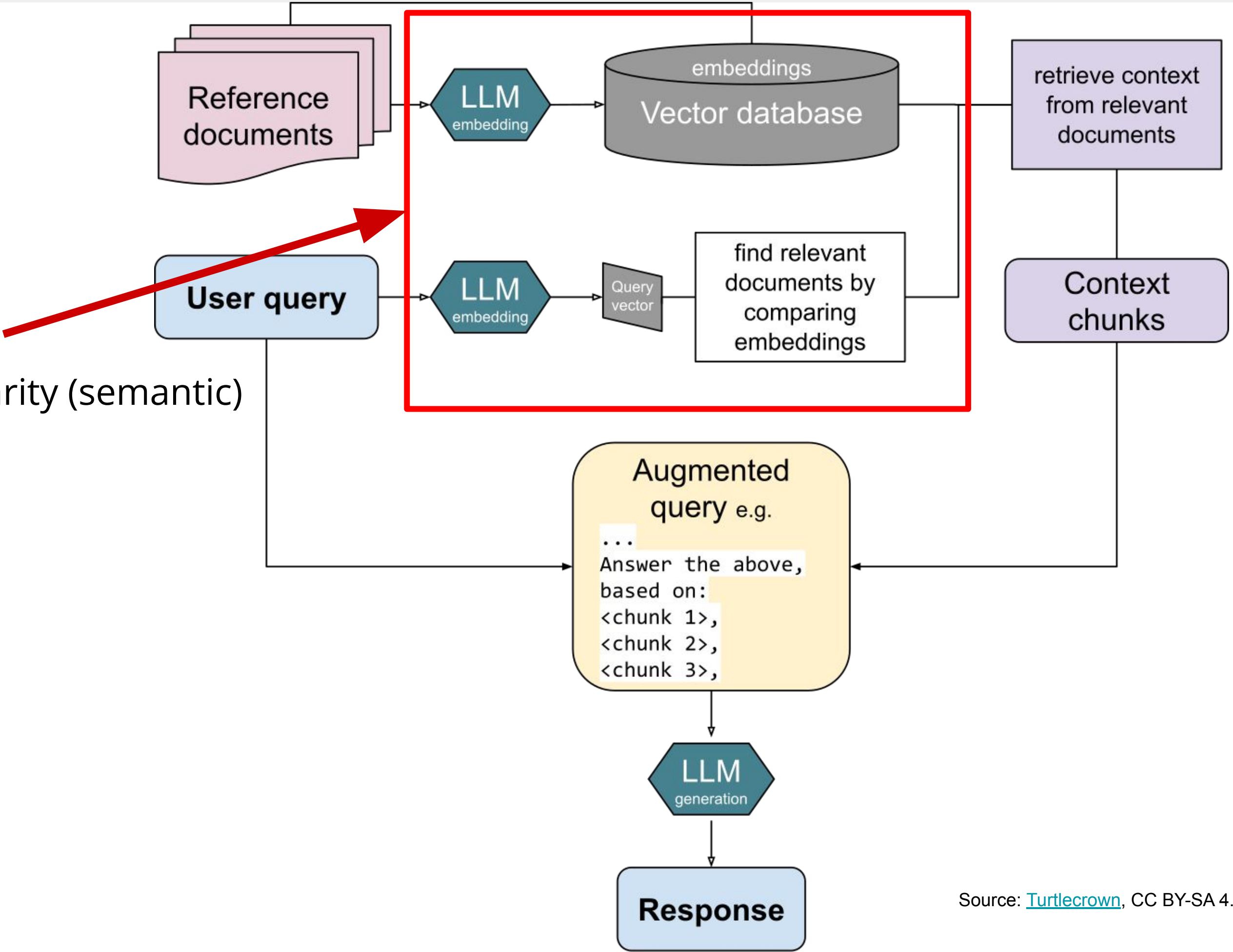


Source: Turtlecrown, CC BY-SA 4.0

RAG architecture

Three main components:

- Generator (LLM)
- Documents
- Retrieval component
    - Typically cosine similarity (semantic)

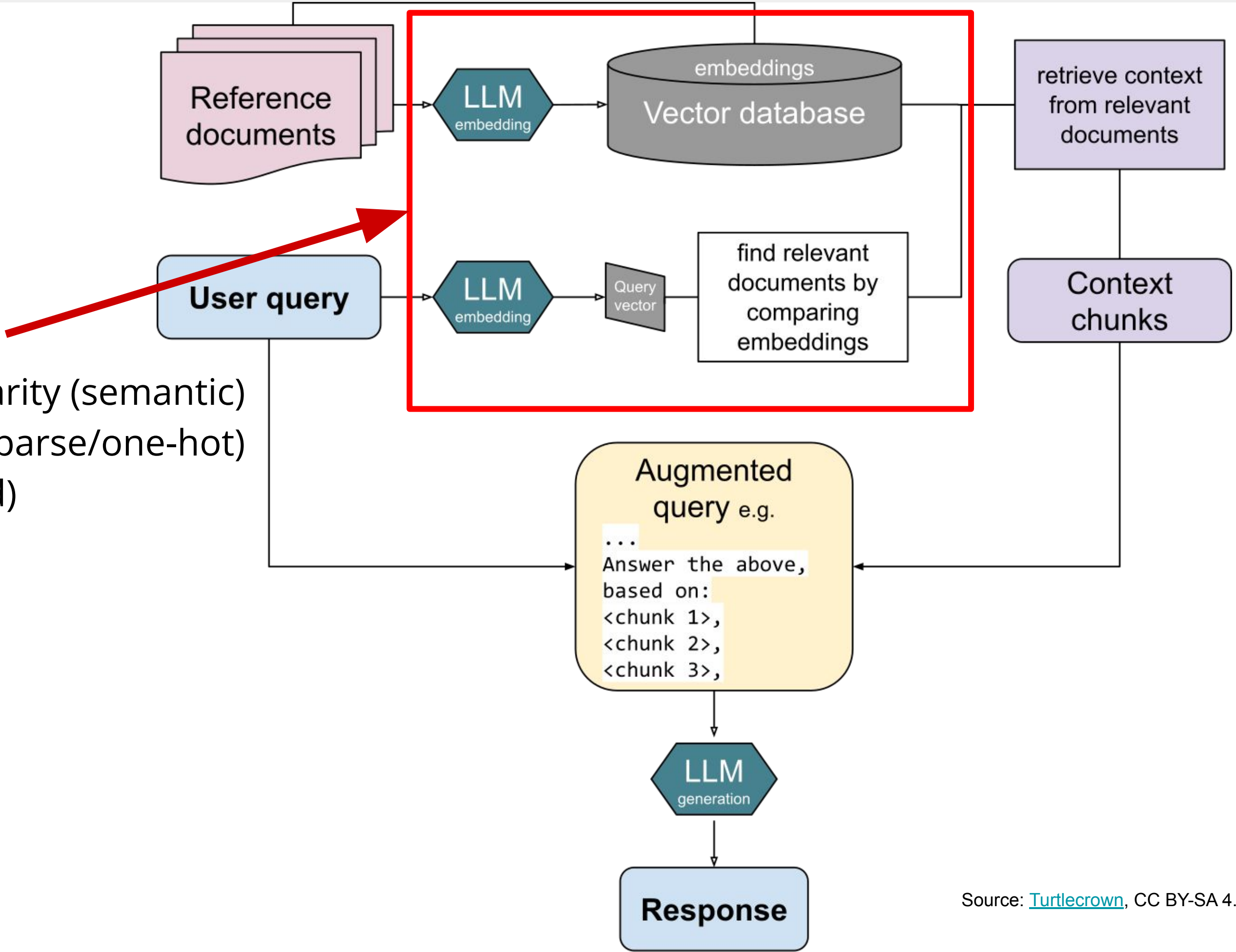


Source: Turtlecrown, CC BY-SA 4.0

RAG architecture

Three main components:

- Generator (LLM)
- Documents
- Retrieval component
  - Typically cosine similarity (semantic)
  - Or keyword search (sparse/one-hot)
  - Combine both (hybrid)

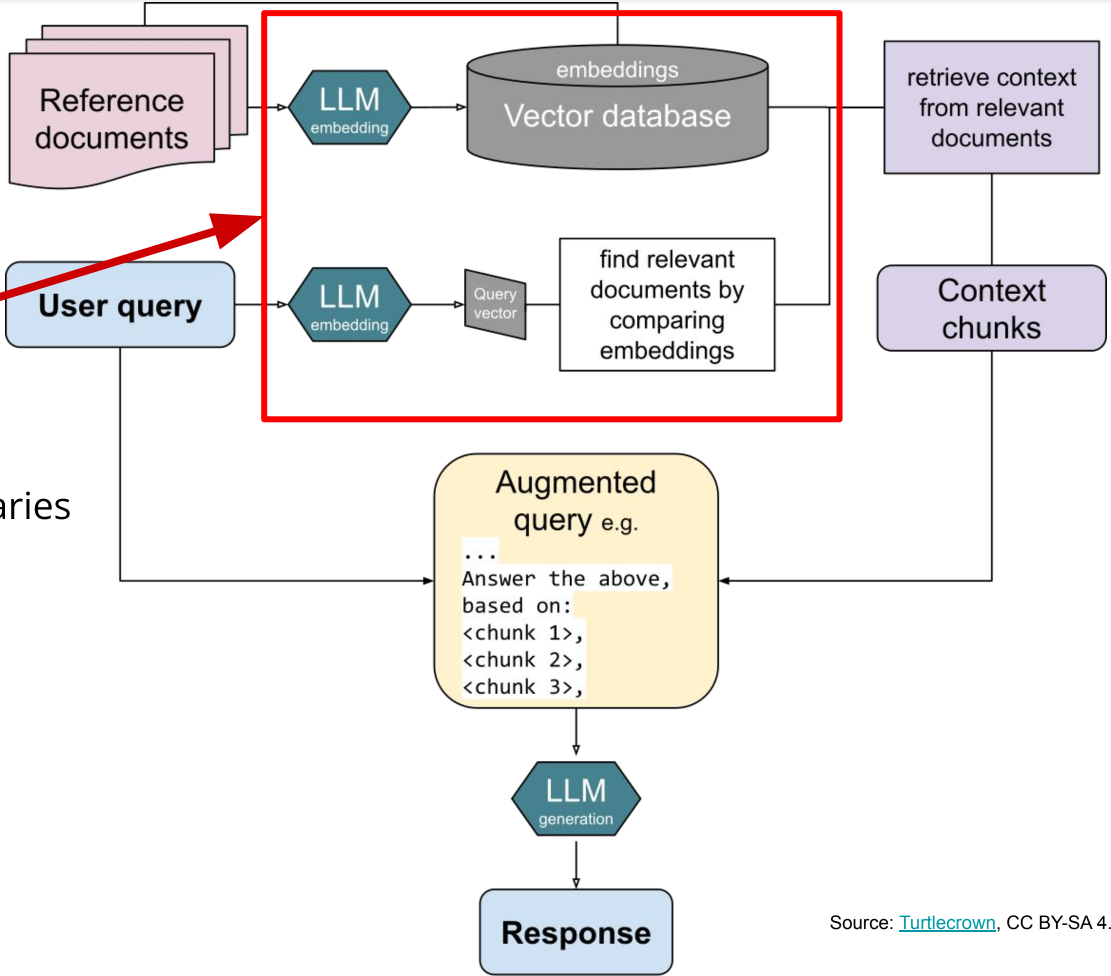


Source: Turtlecrown, CC BY-SA 4.0

RAG architecture

Three main components:

- Generator (LLM)
- Documents
- Retrieval component
  - Typically overlapping chunks
  - fixed length or based on sent/par/page/chapter boundaries



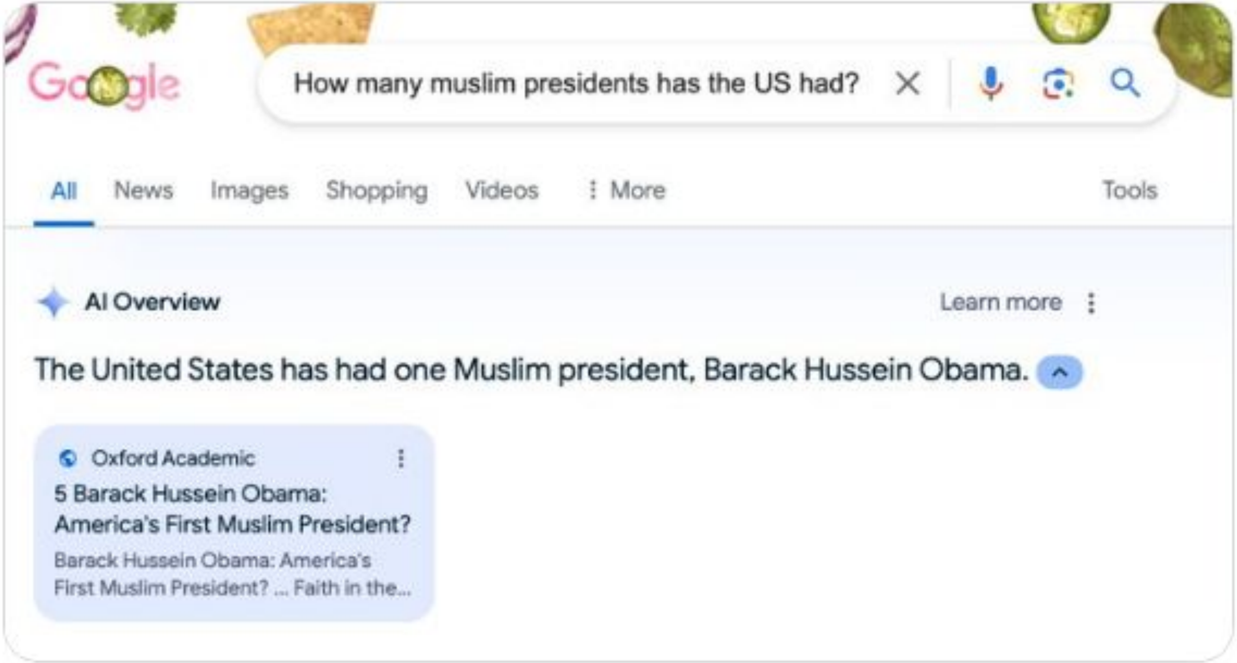
Source: Turtlecrown, CC BY-SA 4.0

RAG advantages

- updated domain-specific knowledge
    - Vector DB may include documents absent from the LLM training data.
    - Adding documents to the DB is much faster than LLM retraining/finetuning.
    - RAG is suitable for non-public (secret) documents.
    - RAG may give better results even if all documents were in LLM training data.
- less hallucinations, more cited sources
    - RAG reduces AI hallucinations.
    - The prompt template may contain instruction
      "Do not hallucinate and do not provide any information not included below."
    - The generator LLM can include links to the cited sources and/or exact citations (chunks).
- can use web search
    - but loosing the curated/reliable DB advantage

RAG limitations

- Not almighty
    - even when using reliable (factually correct) documents



How many muslim presidents has the US had?

The United States has had one Muslim president, Barack Hussein Obama.

Oxford Academic

5 Barack Hussein Obama: America's First Muslim President?

Barack Hussein Obama: America's First Muslim President? ... Faith in the...

Source: Melanie Mitchell

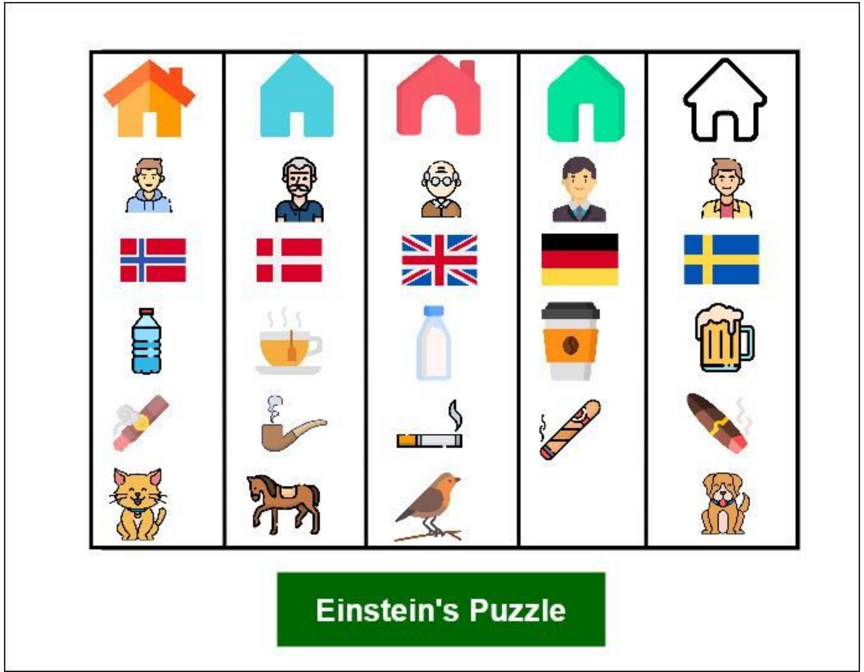
RAG limitations

- Not almighty
  - even when using reliable (factually correct) documents
  - Users may have too high expectations. RAG is not suitable for all tasks.
  - Does not work when the relevant documents are not similar to the query.

1. There are five houses.
2. The Englishman lives in the red house.

…

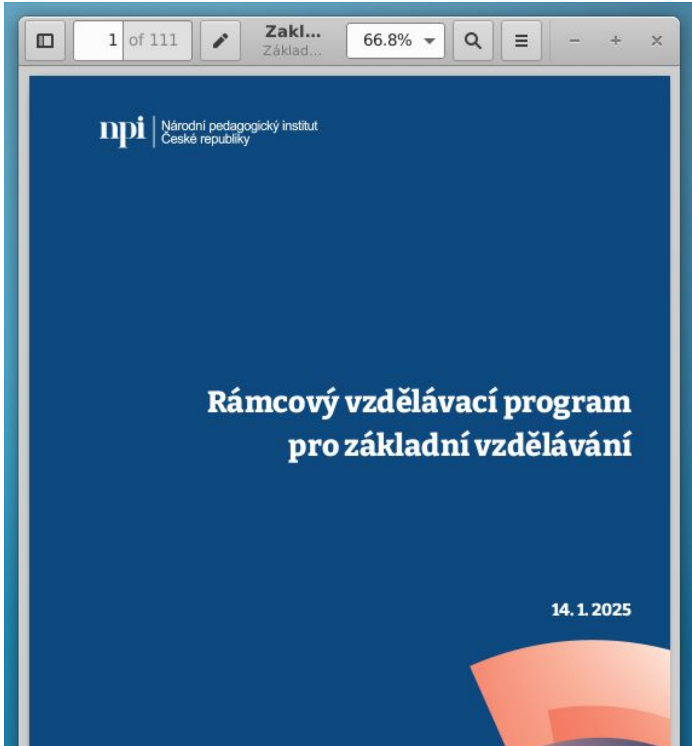
Now, who drinks water? Who owns the zebra?



RAG limitations

- Not almighty
  - even when using reliable (factually correct) documents
  - Users may have too high expectations. RAG is not suitable for all tasks.
  - Does not work when the relevant documents
    are not similar to the query.

    Teaching guidelines (Inquiry-based learning of history)
    are relevant for all WP1 questions (answers),
    so it (summary of the 111 pages) is more suitable
    for the system prompt.



npi | Národní pedagogický institut České republiky

Rámcový vzdělávací program
pro základní vzdělávání

14. 1. 2025

RAG limitations

- Not almighty
  - even when using reliable (factually correct) documents
  - Users may have too high expectations. RAG is not suitable for all tasks.
  - Does not work when the relevant documents are not similar to the query.
- LLM knows, but RAG does not
  - The instruction "Do not hallucinate" may worsen the results for some questions.
- Higher HW requirements (speed and memory)
  - Longer context (n_chunks * chunk_length)
  - Dilema: forget older chat history or decrease n_chunks or chunk_length?

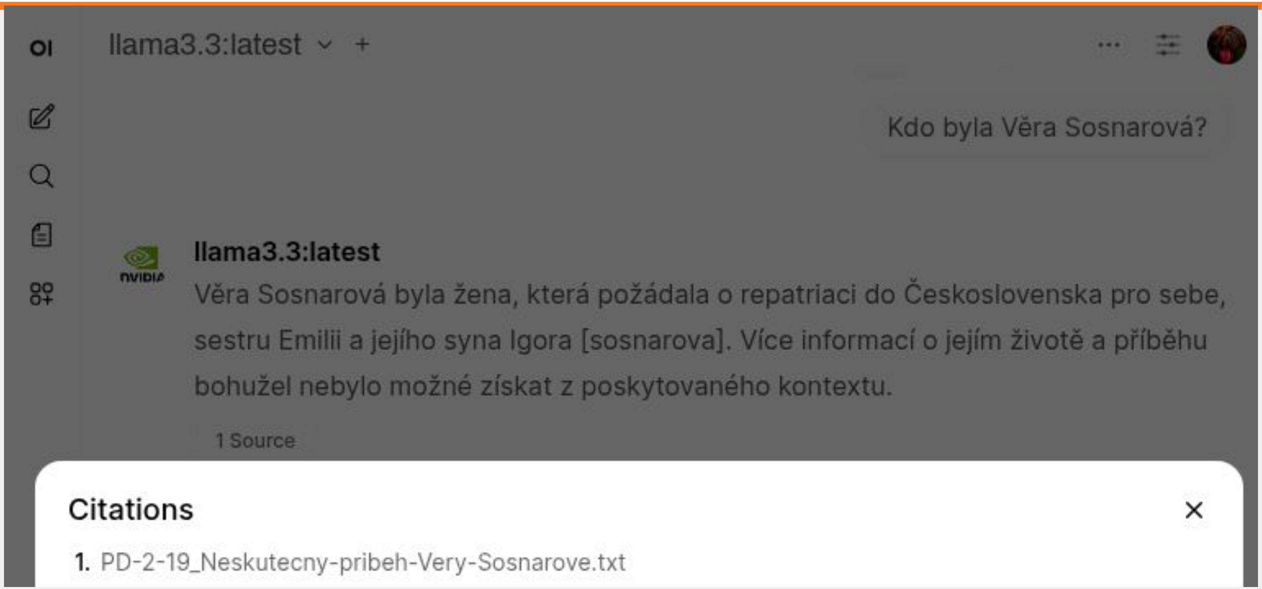
RAG improvements

- Retrieval reranking or filtering
  - cosine similarity is fast but too simple, we can use full LLM
- Query boosting
  - augmenting the query for the retrieval
- Adaptations for multiturn dialogs
  - keep previously used chunks in context
- Agentic systems
- Hierarchical RAG
- …

Not really a

Summary

- RAG is tremendous and may be beneficial for small democratic society.
- Our prototype is not public yet, but you can try ai.ufal.mff.cuni.cz and its integrated mini RAG (Workspace → New knowledge base).



Kdo byla Věra Sosnarová?

llama3.3:latest

Věra Sosnarová byla žena, která požádala o repatriaci do Československa pro sebe, sestru Emilii a jejího syna Igora [sosnarova]. Více informací o jejím životě a příběhu bohužel nebylo možné získat z poskytovaného kontextu.

1 Source

Citations

1. PD-2-19_Neskutecny-pribeh-Very-Sosnarove.txt