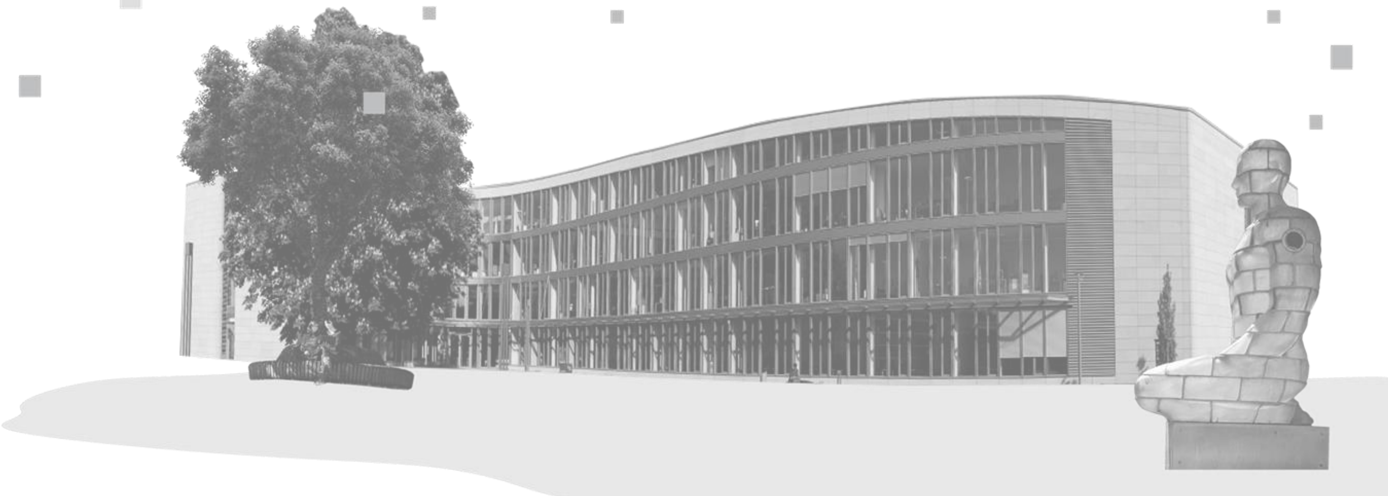
Efficient Information Retrieval from Documents with AI

Local Retrieval-Augmented Generation

Hanno Müller
hanno.mueller@hpi.de

KI Service Zentrum by Hasso-Plattner-Institut
HPI

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt



Design IT.
Create Knowledge.

www.hpi.de

KI Service Zentrum by Hasso-Plattner-Institut HPI

Who are you?

What do you know about RAG?

From **which documents** do you want to retrieve information?

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

KI Service Zentrum by Hasso-Plattner-Institut HPI

Motivation

- AI Service Centre documents over **1,000 hours** of interaction with clients from a wide range of domains, including:
  - SMEs / NGOs / associations / large enterprises / freelancers
  - Start-Ups
  - Academia
  - Municipal, state, and federal authorities
- Large share of requests revolved around questions about **RAG**, such as its benefits and feasibility.
- AI Service Centre has partnered up for RAG **pilot projects** with Deutscher Bundestag, Landtag Brandenburg, Brandenburger Umweltministerium, fghgsd (IT Baseline Protection)

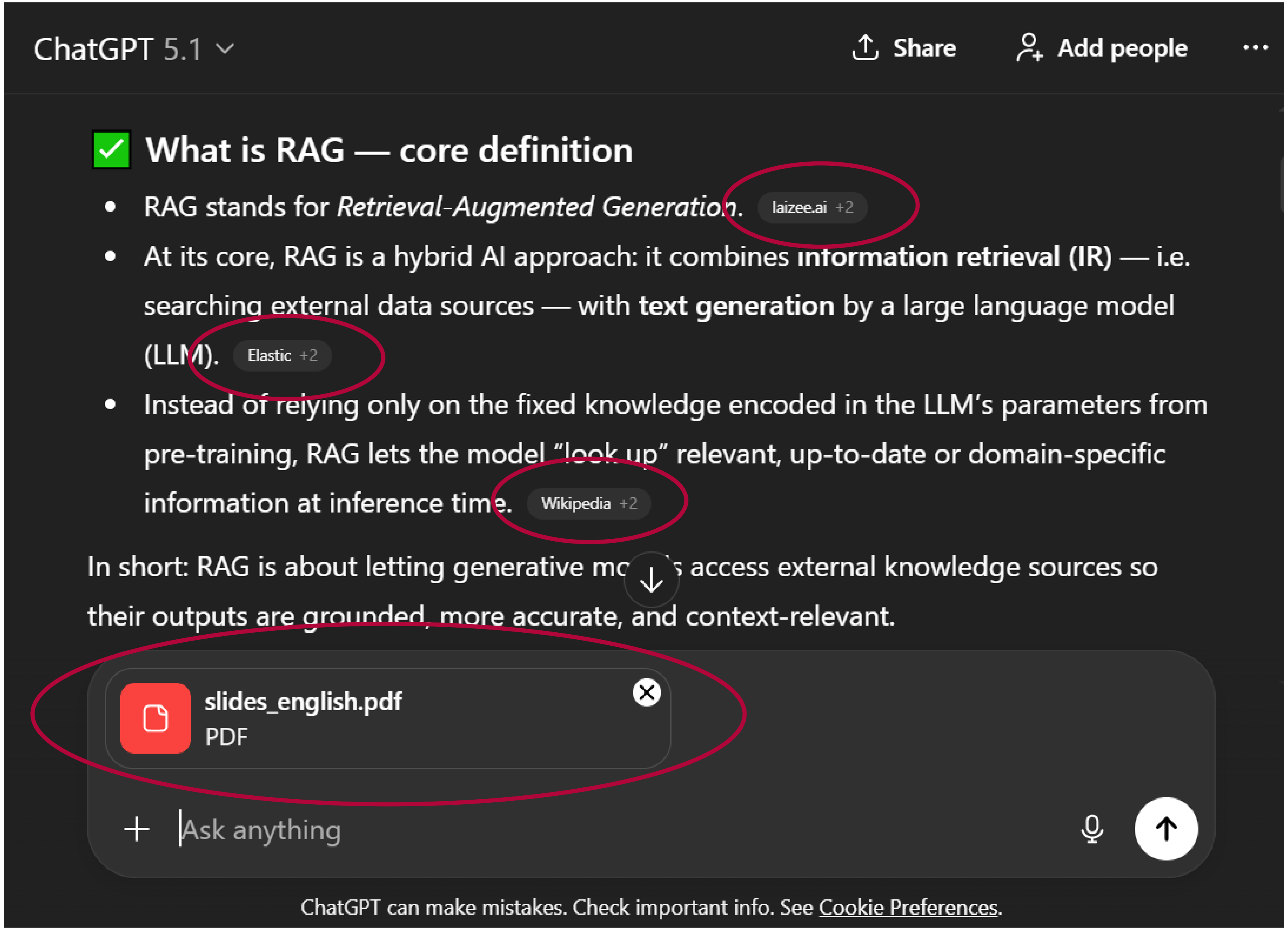
Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

What is RAG?

- Retrieval-Augmented Generation (RAG) enhances LLM performance by enabling LLMs to access **external information** from multiple sources.

Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt



Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt

What is RAG?

- Retrieval-Augmented Generation (RAG) enhances LLM performance by enabling LLMs to access **external information** from multiple sources.

  → *Generation that is augmented by retrieval*

  1. **Information Retrieval** from, e.g., internet, uploaded documents, local documents, live databases such as SQL
  2. **Text Generation**, taking into account retrieved information

- Seperation of data and model
- Answers go beyong training data without additional model training
- Less hallucinations

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

External Systems

- RAG run externally / in cloud (e.g., SaaS applications)

PRO:

- Easy to scale
- No hardware or maintenance
- Advanced toolkit available

CONTRA:

- May be expensive
- Dependence on provider
- Data leaves your environment → privacy concerns

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

On-premise Systems

- RAG run on-premise (servers you control)

PRO:

- High data protection & compliance
- Predictable (fixed) costs
- Integration with internal systems

CONTRA:

- High infrastructure cost (investment cost & maintenance)
- Slower innovation cycles
- Difficult to scale (but hybrid on-premise/clouse possible)

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Local Systems

- RAG run locally (e.g., laptop or tower with dedicated GPU)

PRO:

- Full data protection & compliance
- Fast prototyping
- (Almost) no ongoing costs

CONTRA:

- Limited performance w.r.t. model size & latency
- Not scalable
- May require strong technical skills

Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt

LEARNING GOALS

Gefördert durch:
Bundesministerium
für Forschung, Technologie
und Raumfahrt

Learnings

- Understand the fundamentals of RAG
  - Embeddings
  - Vector databases
  - Information retrieval
  - Prompt engineering
- Increase awareness for the requirements for (your own) RAG
  - e.g., document preprocessing
  - e.g., creating vector databases
  - e.g., prompt engineering

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Agenda

1. General Introduction (20 minutes)
2. Demonstration of RAG (20 minutes)
3. Interactive Notebook (20 minutes)
4. Customise RAG (20 minutes)
5. RAG Evaluation (20 minutes)
6. Discussion (20 minutes)

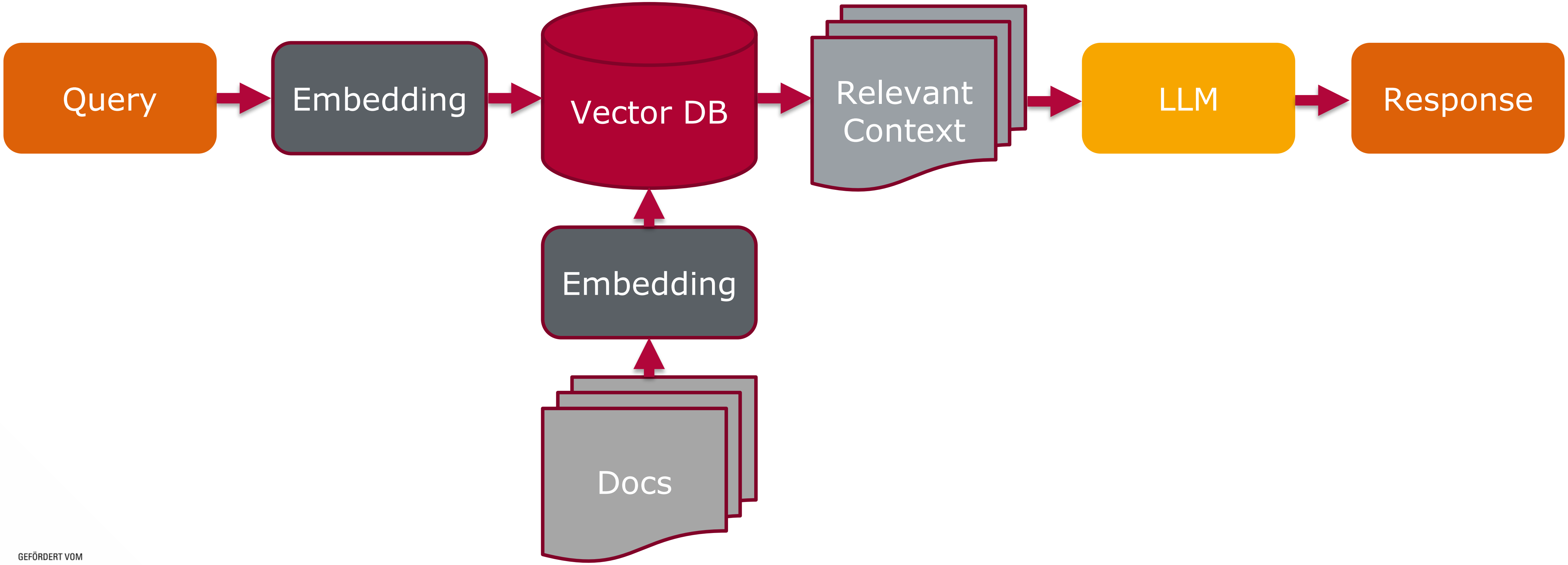
Gefördert durch:
Bundesministerium
für Forschung, Technologie
und Raumfahrt

KI Service Zentrum by Hasso-Plattner-Institut HPI

FUNDAMENTALS

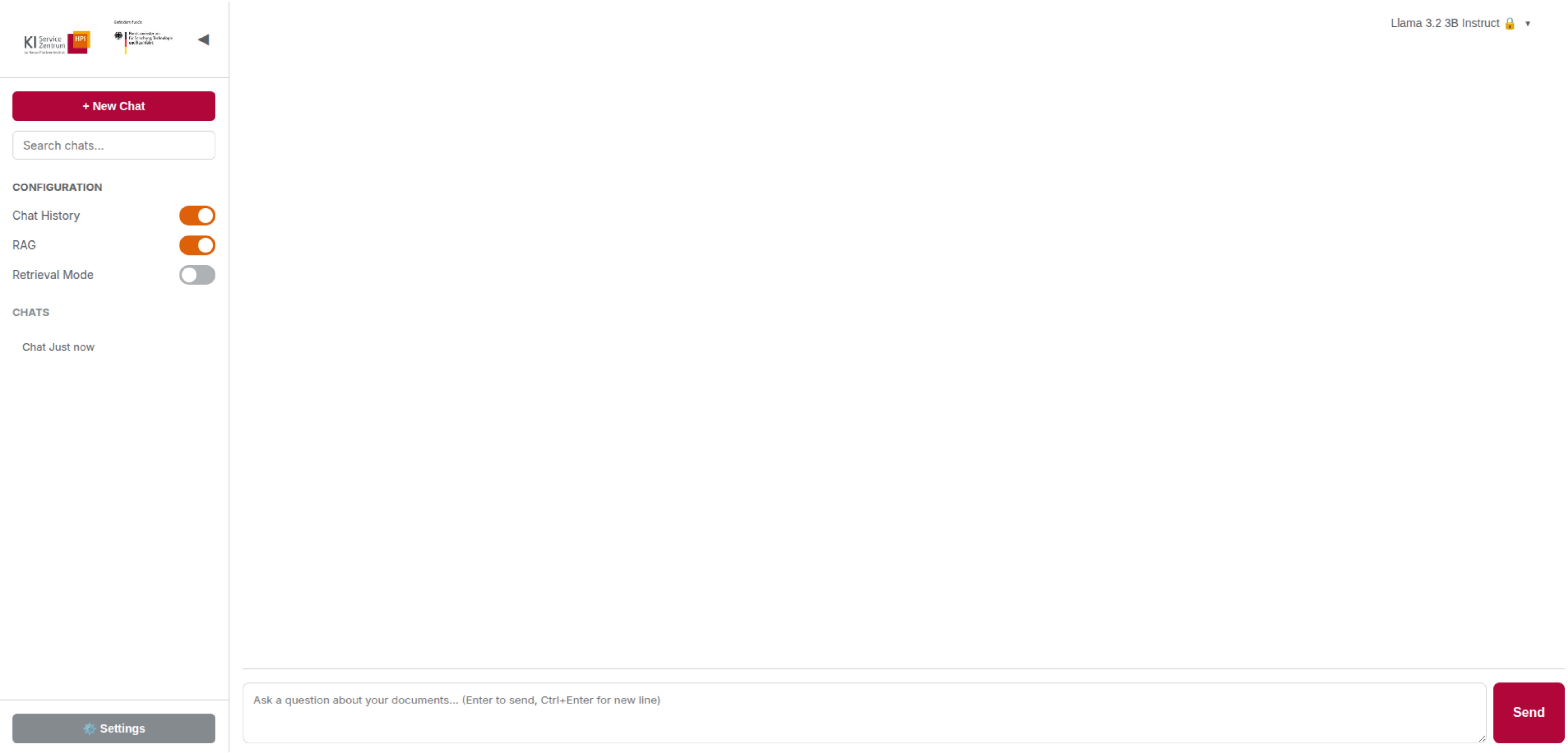
Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt



KI Service Zentrum
by Hasso-Plattner-Institut
HPI

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung


KI Service Zentrum
HPI
+ New Chat
Search chats...
CONFIGURATION
Chat History
RAG
Retrieval Mode
CHATS
Chat Just now
Settings
Llama 3.2 3B Instruct
Ask a question about your documents... (Enter to send, Ctrl+Enter for new line)
Send

Hands-On

KI Service Zentrum by Hasso-Plattner-Institut HPI

https://github.com/aihpi/workshop-rag/blob/main/learning-materials/01_workflow.ipynb

https://github.com/aihpi/
→ repositories
→ workshop-rag
→ learning-materials
→ 01_workflow.ipynb

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Data

- ‚Good' documents are **machine readable** and enable easy extraction of structure (sections, subsections, headers, metadata)
- ‚Bad' documents require **manually crafted processing algorithms**; algorithms may exist but usually function unsatisfactory
- Documents can be categorised in:
  - Text documents (including Tables)
  - Visual documents (scanned documents, figures, diagrams, illustration)

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Evaluation

- RAG can perform insufficiently due to
  1. **Retrieval** problems
     - Retrieving irrelevant information
     - Missing relevant information
     - ‘Lost in the middle’[1]
  2. **Augmentation** problems
     - e.g., gaps between retrieved information are not coherently filled
  3. **Generation** problems
     - e.g., hallucinations

[1] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157-173.

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

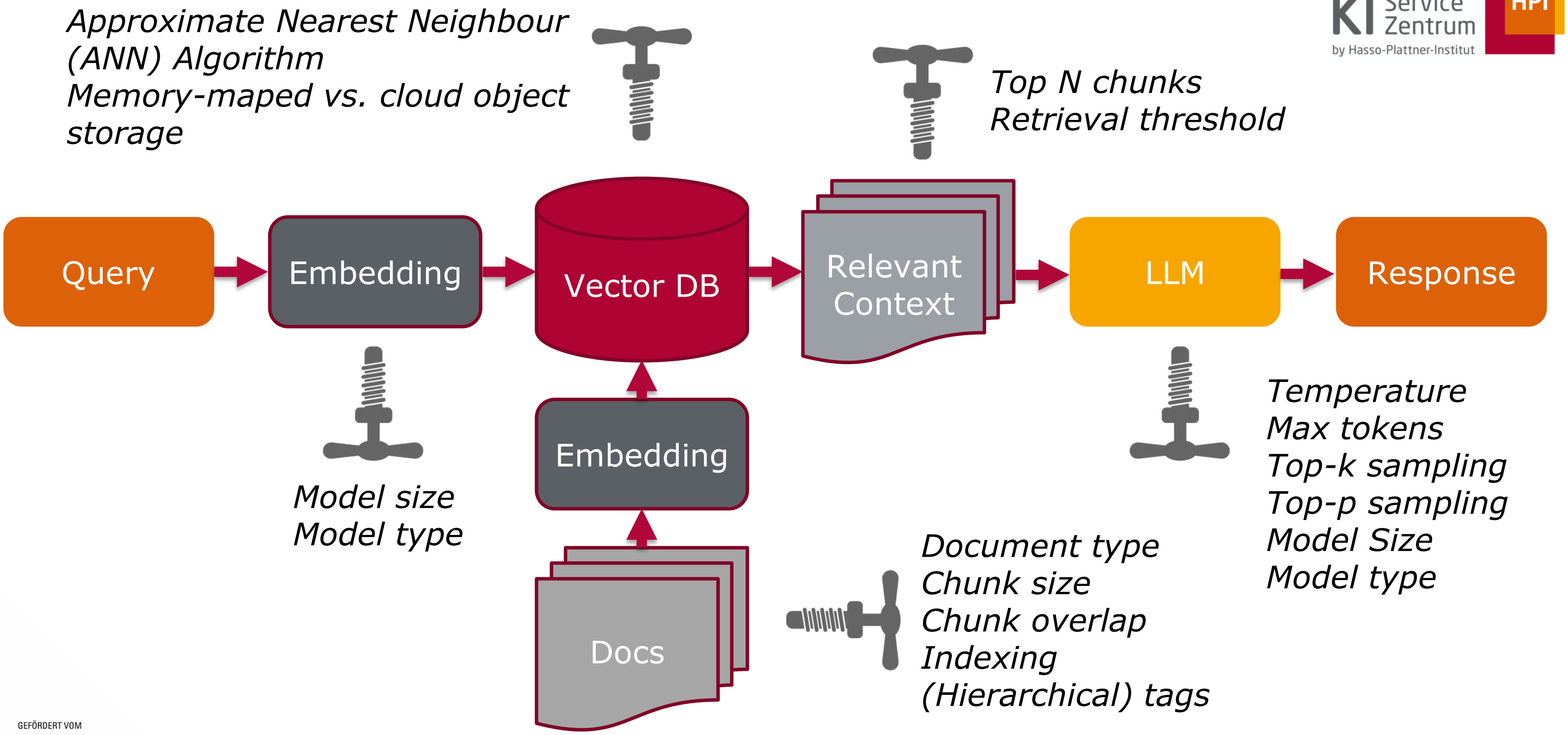
KI Service Zentrum
by Hasso-Plattner-Institut

Evaluation Frameworks

- Evaluation frameworks enable to test RAG system on predefined context data and questions answer pairs, e.g.,
  - https://docs.**ragas**.io/en/stable/
  - https://www.**quotientai**.co/

- Performance may drastically differ between **test data vs. real data**
  → create individual test data
  → manually investigate performance
    - Deconstruction of retrieval, augmentation, generation
    - User feedback

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

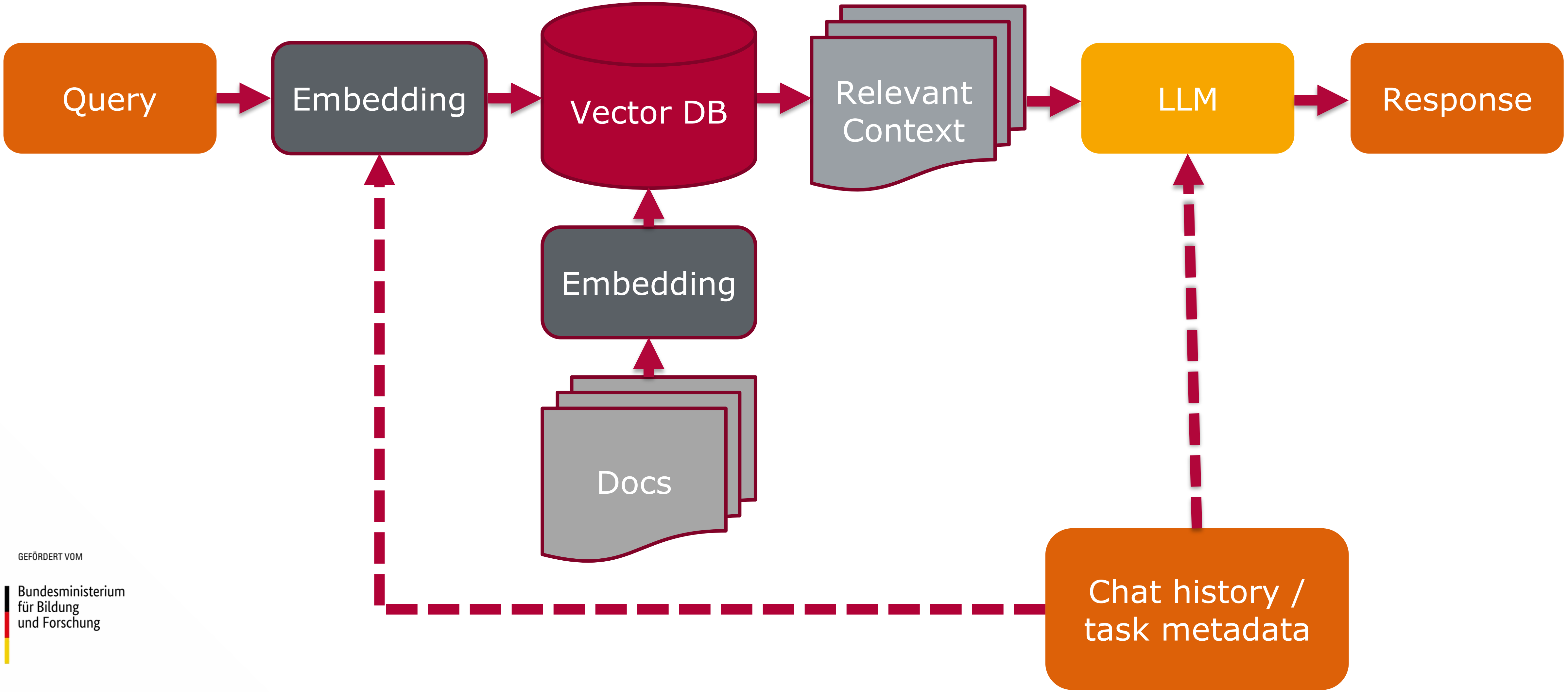


KI Service Zentrum by Hasso-Plattner-Institut HPI

GEFÖRDERT VOM

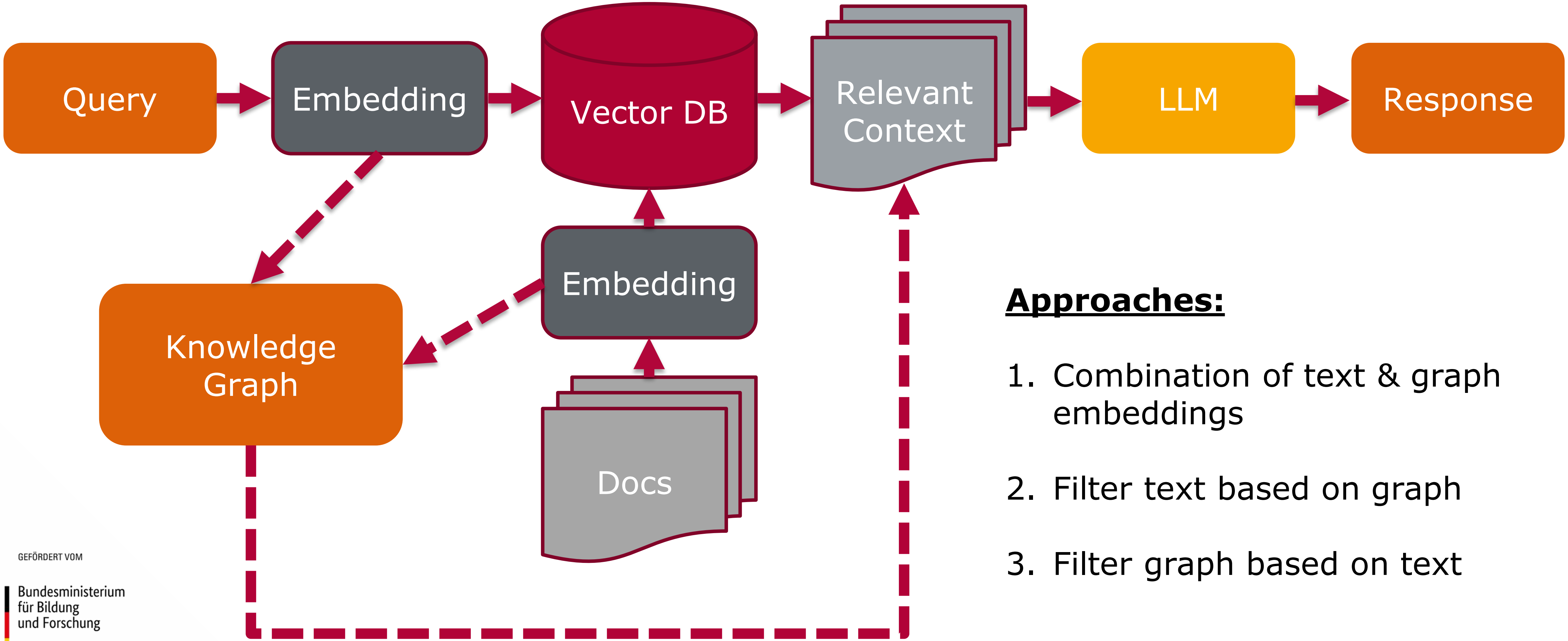
Bundesministerium
für Bildung
und Forschung

Context RAG



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Graph RAG



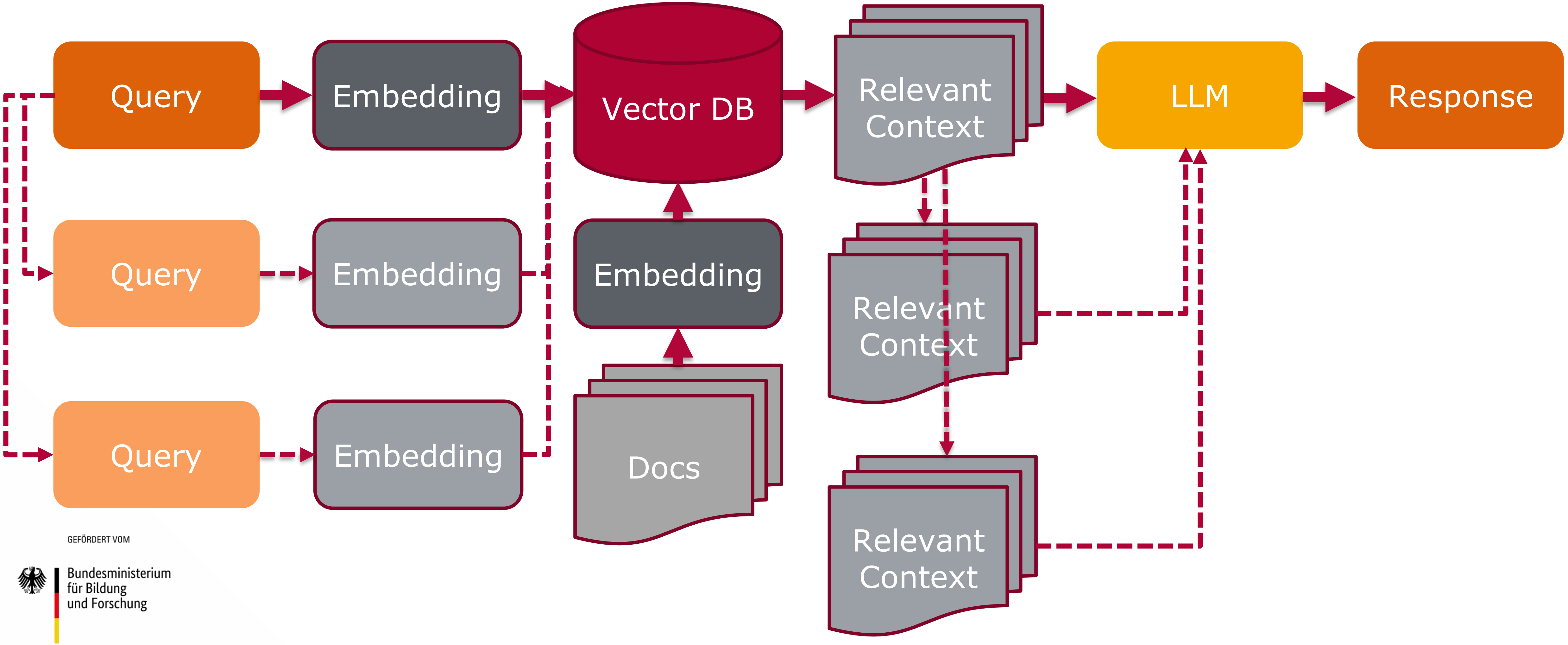
Approaches:

1. Combination of text & graph embeddings
2. Filter text based on graph
3. Filter graph based on text

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Query Transformations

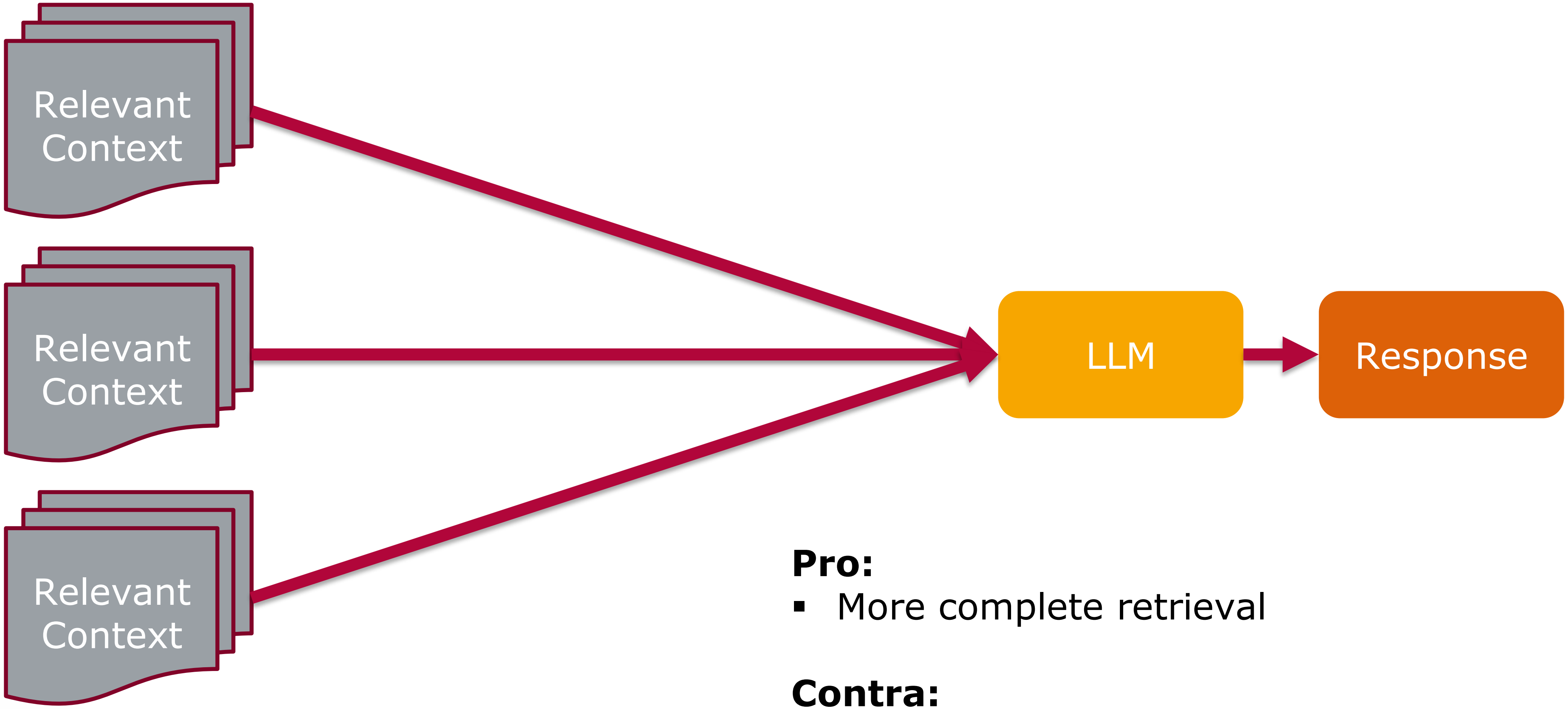
KI Service Zentrum by Hasso-Plattner-Institut HPI



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Multi Query

KI Service Zentrum by Hasso-Plattner-Institut HPI



Pro:
- More complete retrieval

Contra:
- 'Lost in the middle'
- More irrelevant retrieval
- Challenging for LLM context window

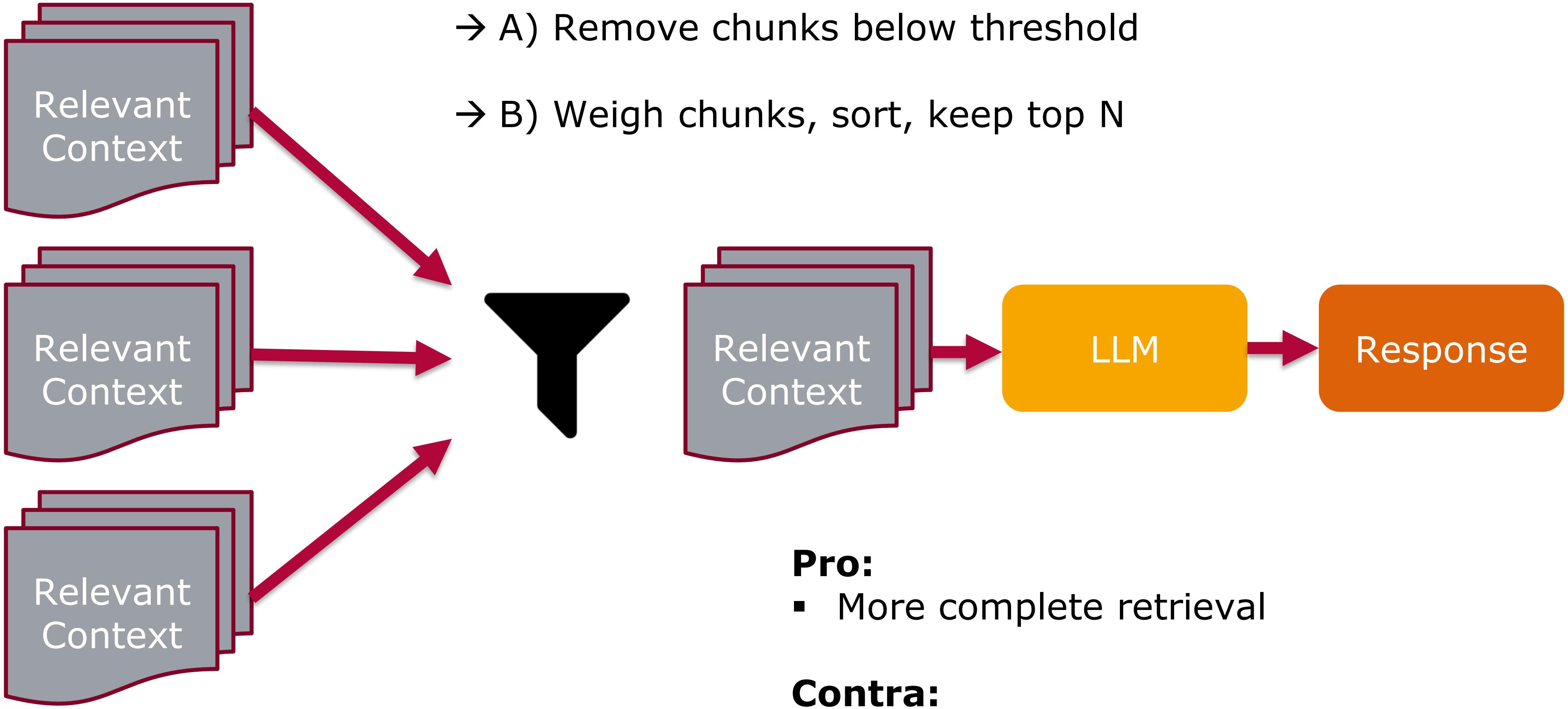
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Filtering

Filtering can be based on how often an information chunk is retrieved.

→ A) Remove chunks below threshold

→ B) Weigh chunks, sort, keep top N



Pro:

- More complete retrieval

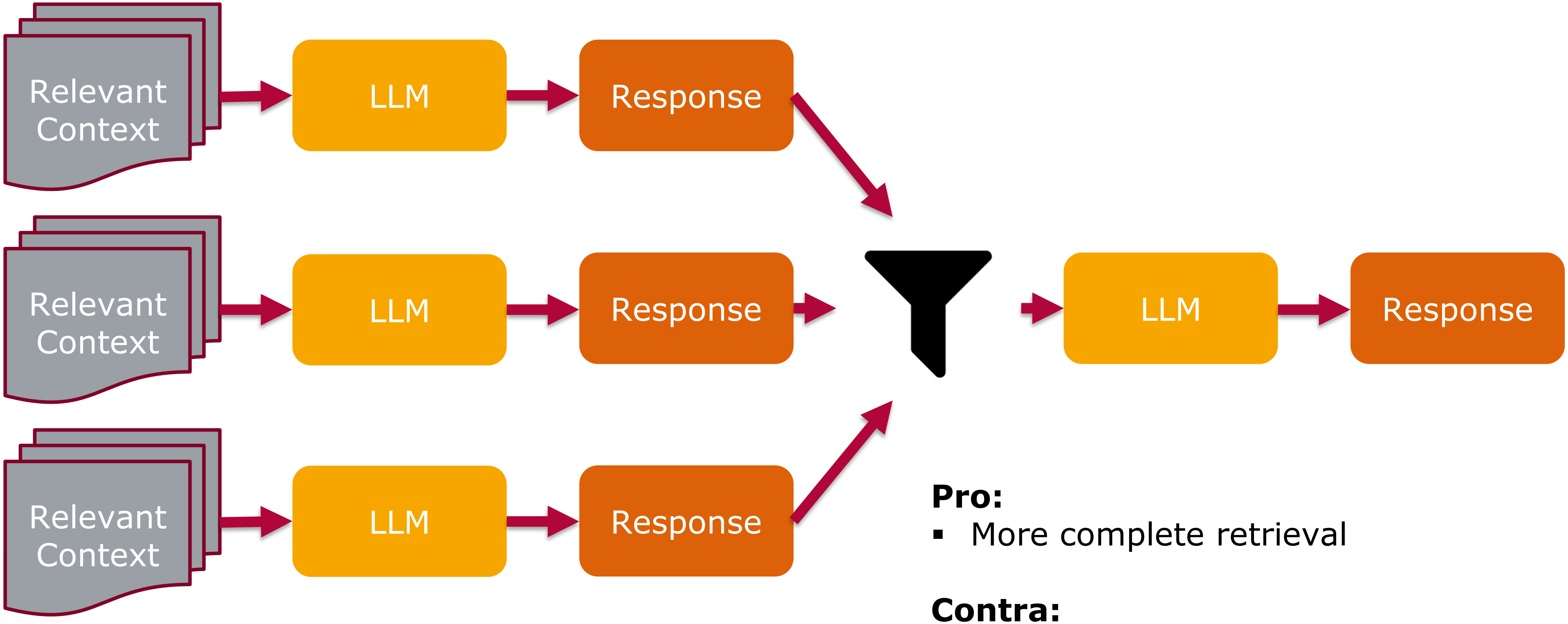
Contra:

- Only works when all relevant information can be retrieved from (paraphrased) query

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Multi Response

KI Service Zentrum
by Hasso-Plattner-Institut
HPI



Pro:
- More complete retrieval

Contra:
- High LLM usages → large costs
- Seldomly better than simpler architectures

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

KI Service Zentrum by Hasso-Plattner-Institut
HPI

DISCUSSION

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

KI Service Zentrum
by Hasso-Plattner-Institut

kisz@hpi.de

hpi.de/kisz

Your opinion is relevant!

QR code to feedback form

Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt