GenAI for Software Engineering 2025

Retrieval-augmented Generation

Jorge Melegati

So far…

- We only relied on the content generated by the model…
- … based on its training

- Limitations:
  - Restricted on time (end of the data)
  - No access to data of specific contexts

An analogy with professional programmers

- Carla studied Computer Science and she is considered a good programmer

- In her first job, she has to maintain an existent software system

- Even as a good programmer, she needs to learn how the code is structured

Image created by AI.

The same for an LLM

- LLMs are really good in recognizing patterns

- But they do not know anything about specific contexts not present in the training data

Programmers vs LLMs

- Example: Carla needs to use an API
  - She will probably consult the API documentation
  - Based on her knowledge she can understand the documentation and use the API

- When an LLM needs some knowledge to perform a task
  - We could provide it to the model

Retrieval Augmented Generation (RAG)

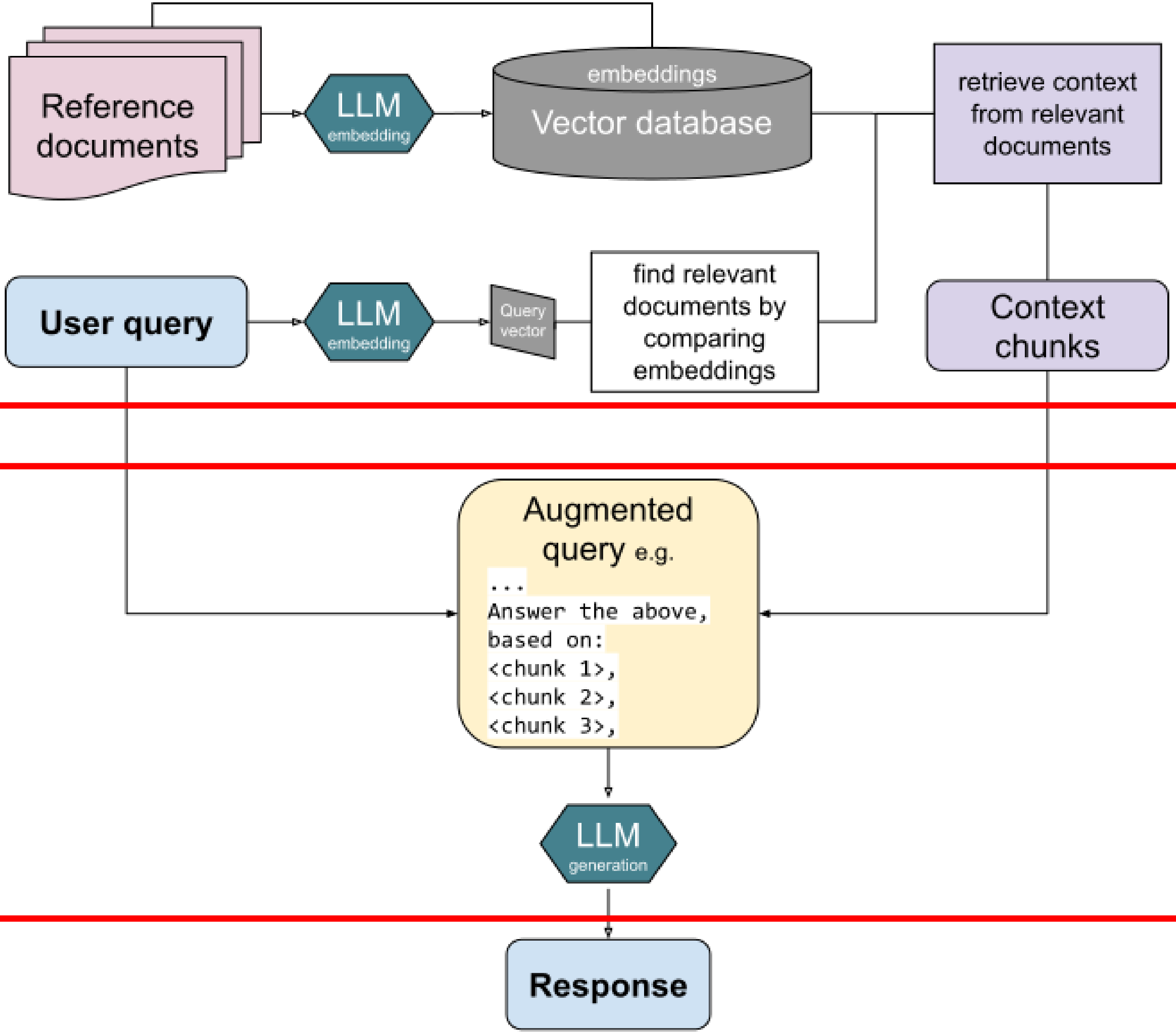
- Relies on existing knowledge artifacts to generate the answer

- Combination of generation with information retrieval

Retrieval-

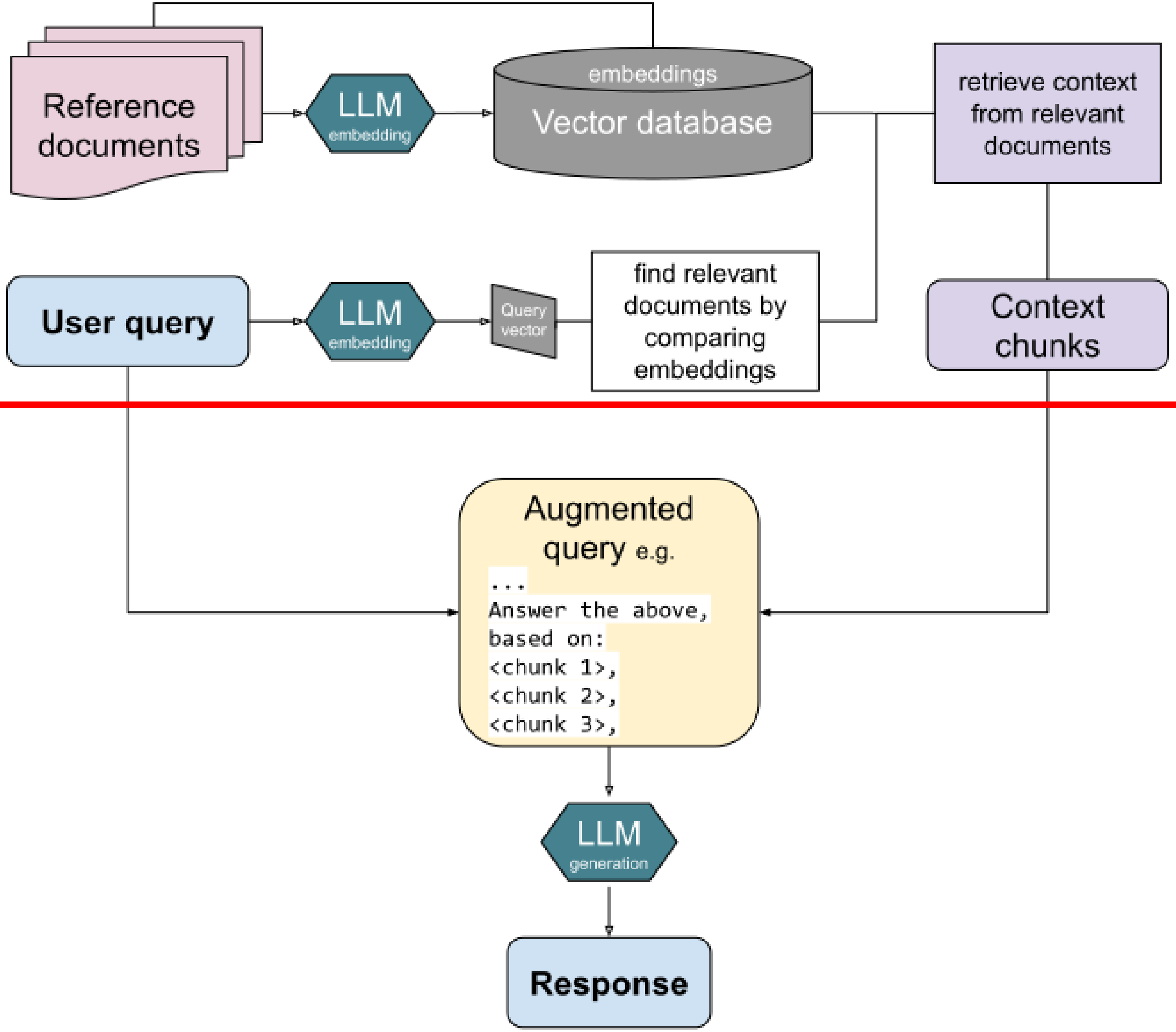
augmented

Generation



Source: Turtlecrown, https://commons.wikimedia.org/wiki/File:RAG_diagram.svg. CC-BY-SA 4.0.

Retrieval-augmented Generation



```
...
Answer the above,
based on:
<chunk 1>,
<chunk 2>,
<chunk 3>,
```

Source: Turtlecrown, https://commons.wikimedia.org/wiki/File:RAG_diagram.svg. CC-BY-SA 4.0.

Problem

How can we find **semantically** relevant documents for a given query?

Do you recall from the first lecture?

- Representing words in a vector space
- Semantically and syntactically similar words are mapped to nearby points

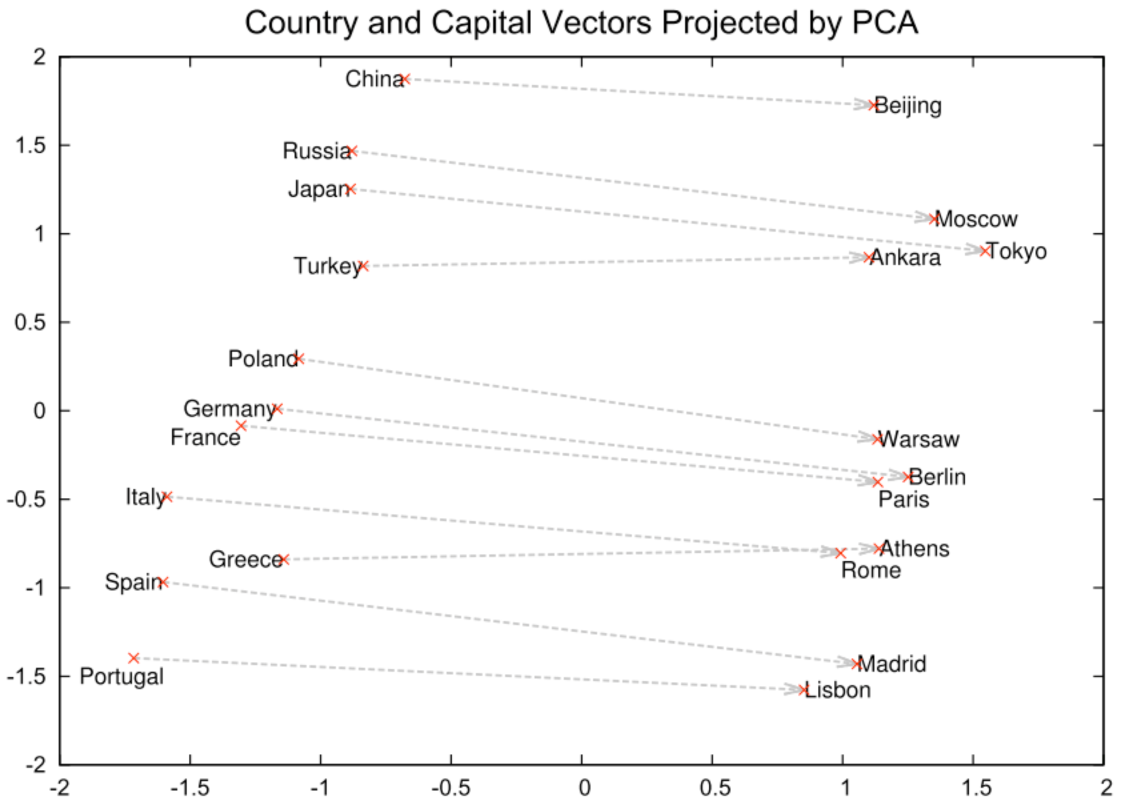


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Source: Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).

A similar idea is used in RAG

- Difference: rather than simple words, in RAG, representations of chunks of text or complete documents are stored

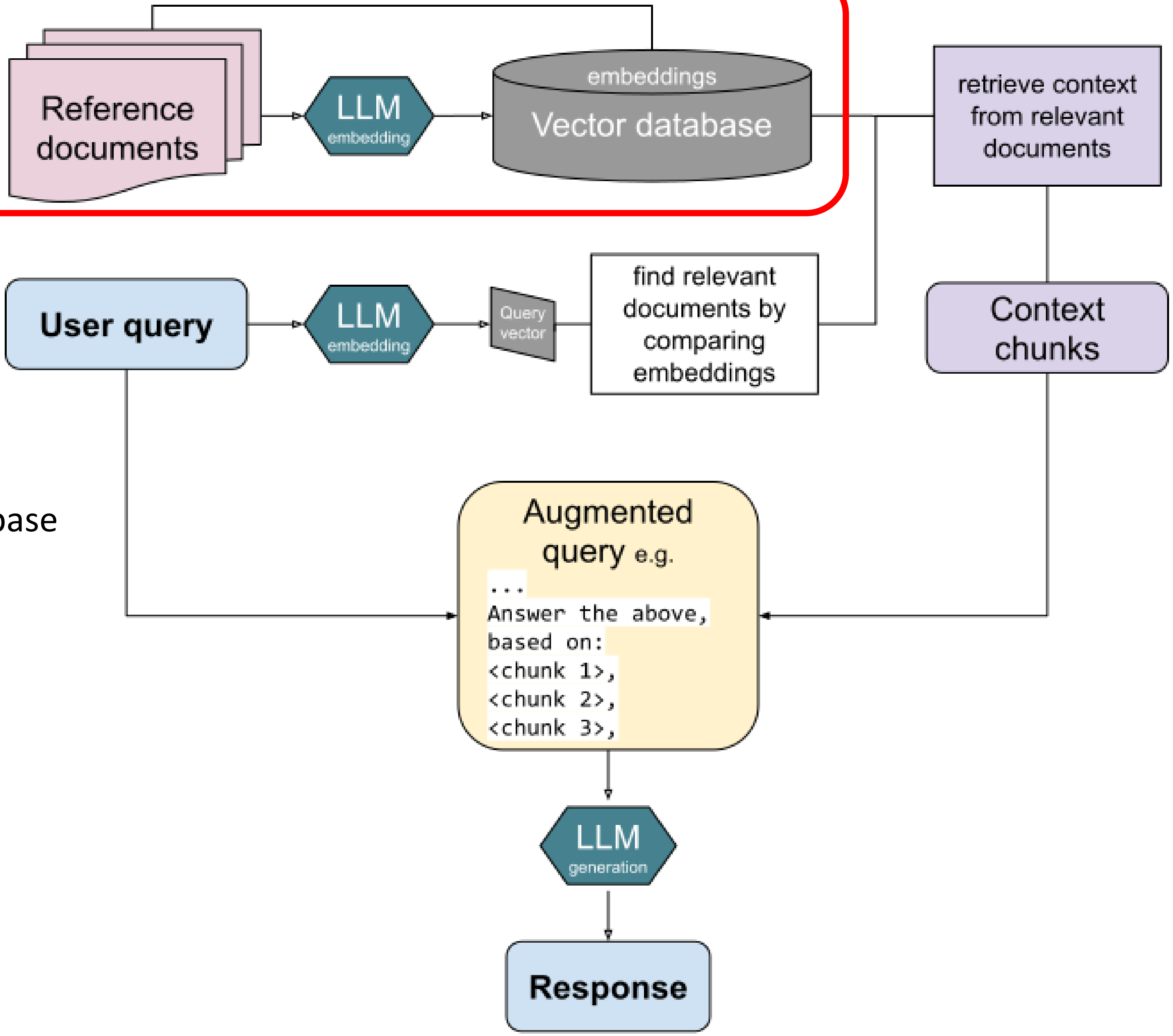
- Generally, using LLM-based embeddings
    - Not necessarily the same of the model

Vector databases

- Store documents and their associated vectors (lists of numbers)

- Similarity search
  - Example: cosine similarity

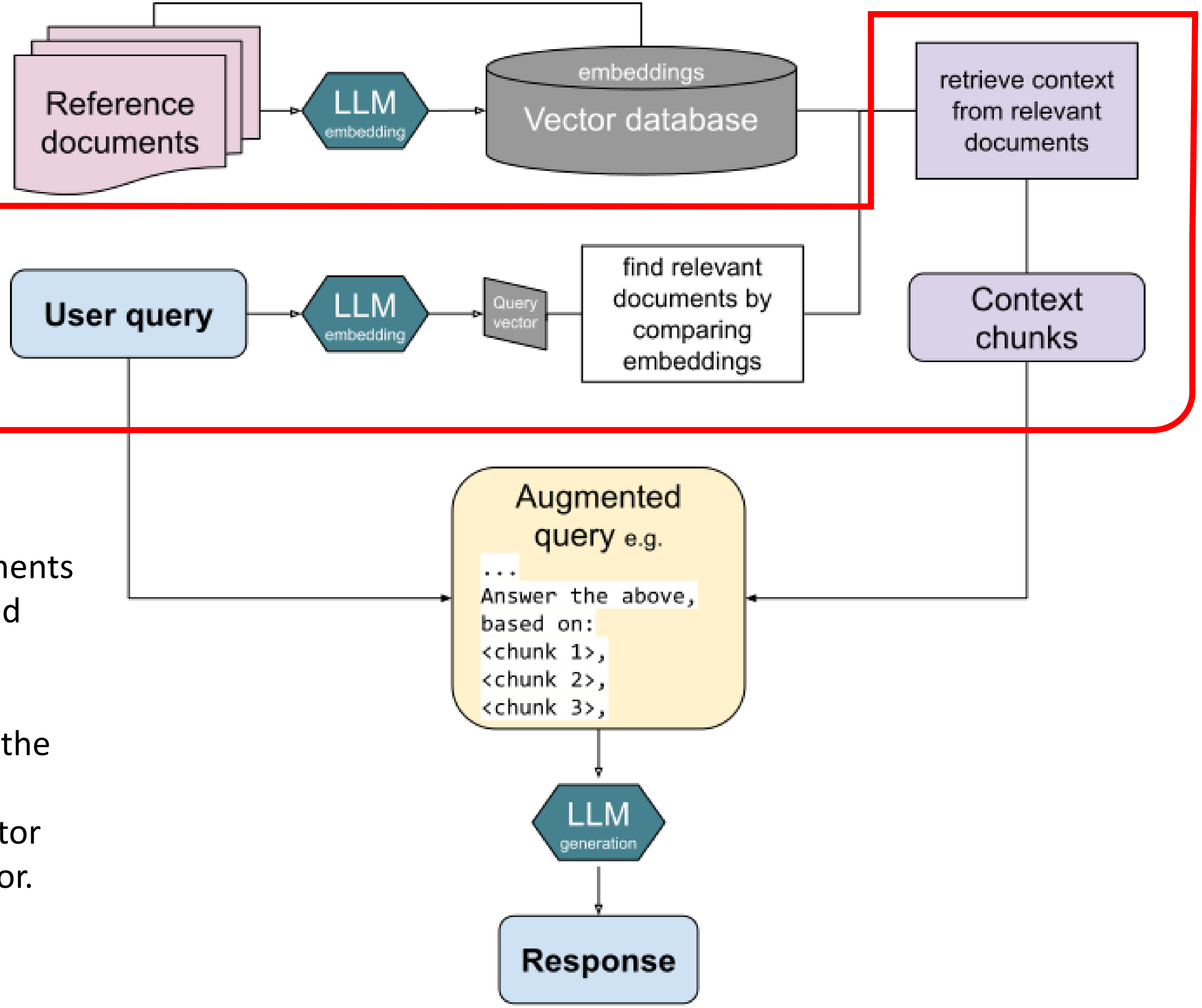
Indexing step



- All documents are added to the database
- With their respective embeddings

Source: Turtlecrown, https://commons.wikimedia.org/wiki/File:RAG_diagram.svg. CC-BY-SA 4.0.

Searching step



- For a given query, identify the documents that are the most semantically related

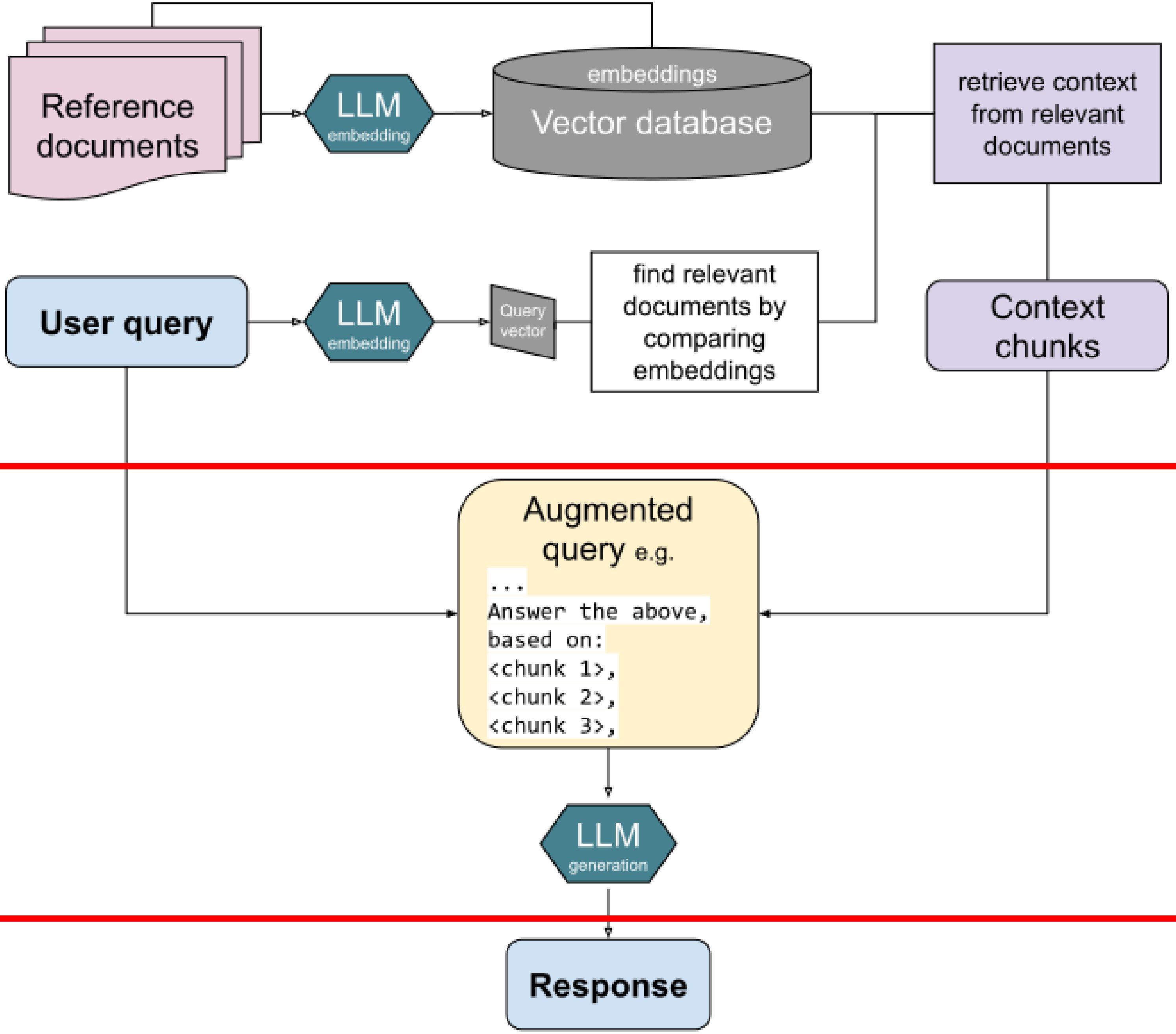
- In mathematical terms:
  - Generate a vector representing the query.
  - Locate the document in the vector space nearest to the query vector.

Source: Turtlecrown, https://commons.wikimedia.org/wiki/File:RAG_diagram.svg. CC-BY-SA 4.0.

Retrieval-

augmented

Generation



```
...
Answer the above,
based on:
<chunk 1>,
<chunk 2>,
<chunk 3>,
```

Source: Turtlecrown, https://commons.wikimedia.org/wiki/File:RAG_diagram.svg. CC-BY-SA 4.0.

Side-effect advantages

- Documents can serve two purposes:
    - Increase the explainability of the answer
    - Reduce the chance of hallucinations

Example: Generating code that uses APIs

- Problem: specific APIs are probably not known by the model

- Solution: use a RAG to provide the information need about the API calls need for generating the code

- Open in Google Colab the file RAGforAPI.ipynb from the repository melegati/gen4ai-course
  - If you prefer, you can also run locally!

Another Example

- Using LLMs and RAG for bug identification

- Paper:
  - “From Bugs to Fixes: HDL Bug Identification and Patching using LLMs and RAG”
  - https://ieeexplore.ieee.org/document/10691874/

Exercise: reading a paper

- A more complex example from ICSE 2025
- ChatGPT Inaccuracy Mitigation during Technical Report Understanding: Are We There Yet?
  - https://arxiv.org/abs/2411.07360
- Goal: understand the issues identified and how they mitigated it