Text Generation

applications and tools

Bo Kang, Thomas Demeester, Tijl De Bie

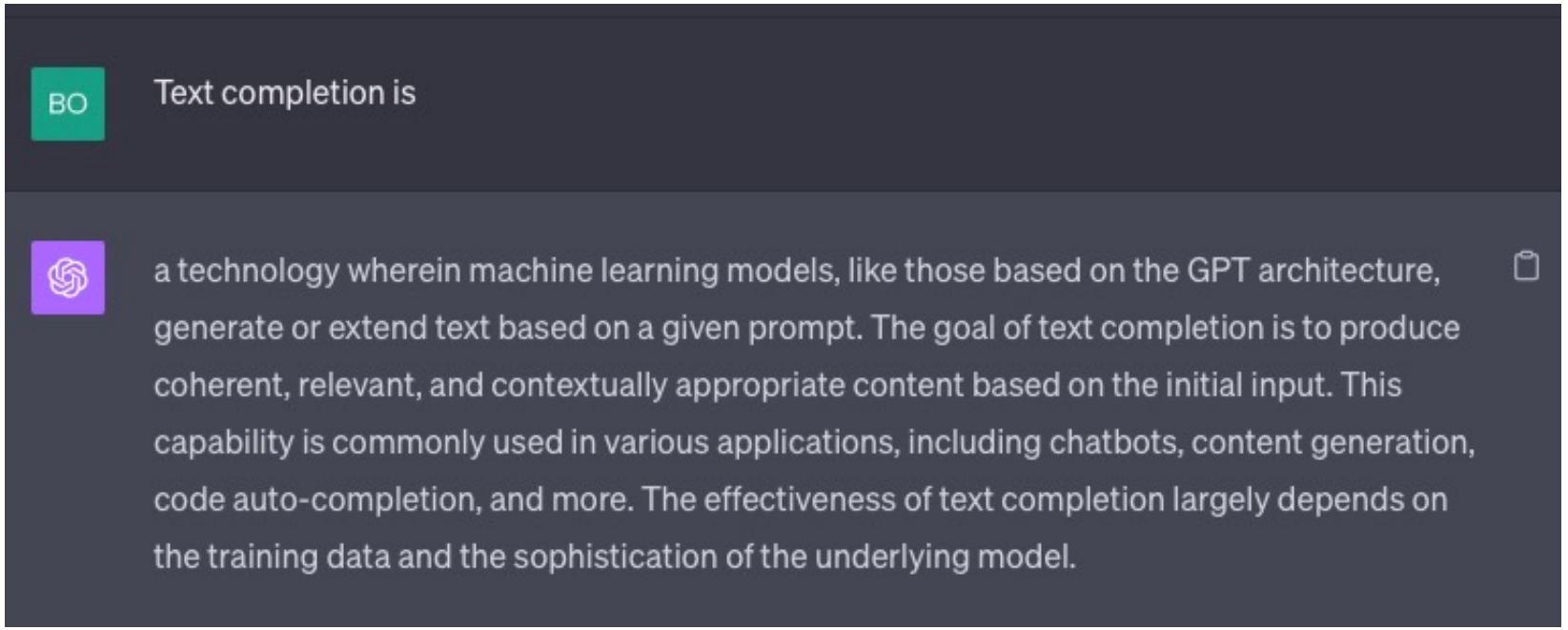
Outline

- Applications overview
- Tools overview
- Application: Retrieval Augmented Generation (RAG)
- Demo: Chat with Helpdesk
- References

An overview of LLM applications

- Text completion
- Content generation and summary
- Coding assistant
- QA (Vague Question, Visual Query)
- Agents
- Many others

Text completion



Content generation and summary

BO

I would like you to act as a professor in AI from Europe. You have abundant experience in developing and delivering AI courses to undergraduate as well as graduate level students. Next semester, the university is planning to give the graduate student a course on AIGC. Your first task is to develop a course description and discuss with the educational director of the university. There are two options for the main theme of the course. Option 1: Large Language Model (LLM), Option 2: AI generated content (AIGC). Certainly Option 2 is more broad than Option 1. Lets explore each option step by step.

Of course, here's the updated outline with section 2.2 split into two separate sections:

2. Course Outline

2.1 Introduction to AI and AIGC

2.2 Basic Understanding of Text-Based AIGC (LLM)

2.2.1 Theory and Design of Text-Based AIGC Models (LLMs)

2.2.2 Basic Use-Cases and Applications of Text-Based AIGC (LLMs)

Content generation and summary

Sure thing! Here's a more concise version, followed by a TL;DR.

Concise Version:

Artificial Intelligence (AI) is transforming various domains, with AI Generated Content (AIGC) and its subset, Large Language Models (LLMs), becoming key areas of this revolution. We propose two standalone courses: one provides an exploration of the broad AIGC field, while the other offers an in-depth focus on LLMs. Both courses emphasize AI trustworthiness, aligning with principles outlined in the EU AI Act. These proposals aim to equip students with knowledge and skills necessary for tomorrow's technological landscape, ensuring ethical AI usage. We invite the educational steering board and the director to review these course options, as they prepare students to be thought leaders in the evolving AI field.

TL;DR:

We propose two courses exploring AI Generated Content (AIGC) and its subset, Large Language Models (LLMs), both emphasizing AI trustworthiness principles from the EU AI Act. These courses aim to prepare students for the evolving AI landscape, ensuring a balance of technological and ethical understanding. We invite the steering board and the director to review these options.

Code assistant

```python
import logging
import chainlit as cl
import pandas as pd
from chainlit.input_widget import Select
from transformers import AutoTokenizer, AutoModel


from constants import DATA_FOLDER, CHAT_MODES, EMBEDDING_MODEL_DEVICE, EMBEDDING_MODEL_REPO_NAME
from utils import get_corpus, get_corpus_embeddings, query, retrieval_augmented_query


logging.basicConfig(level=logging.INFO)

logging.info("Loading embedding model")
EMBEDDING_TOKENIZER = AutoTokenizer.from_pretrained(EMBEDDING_MODEL_REPO_NAME)
```

QA: Vague Question

BO

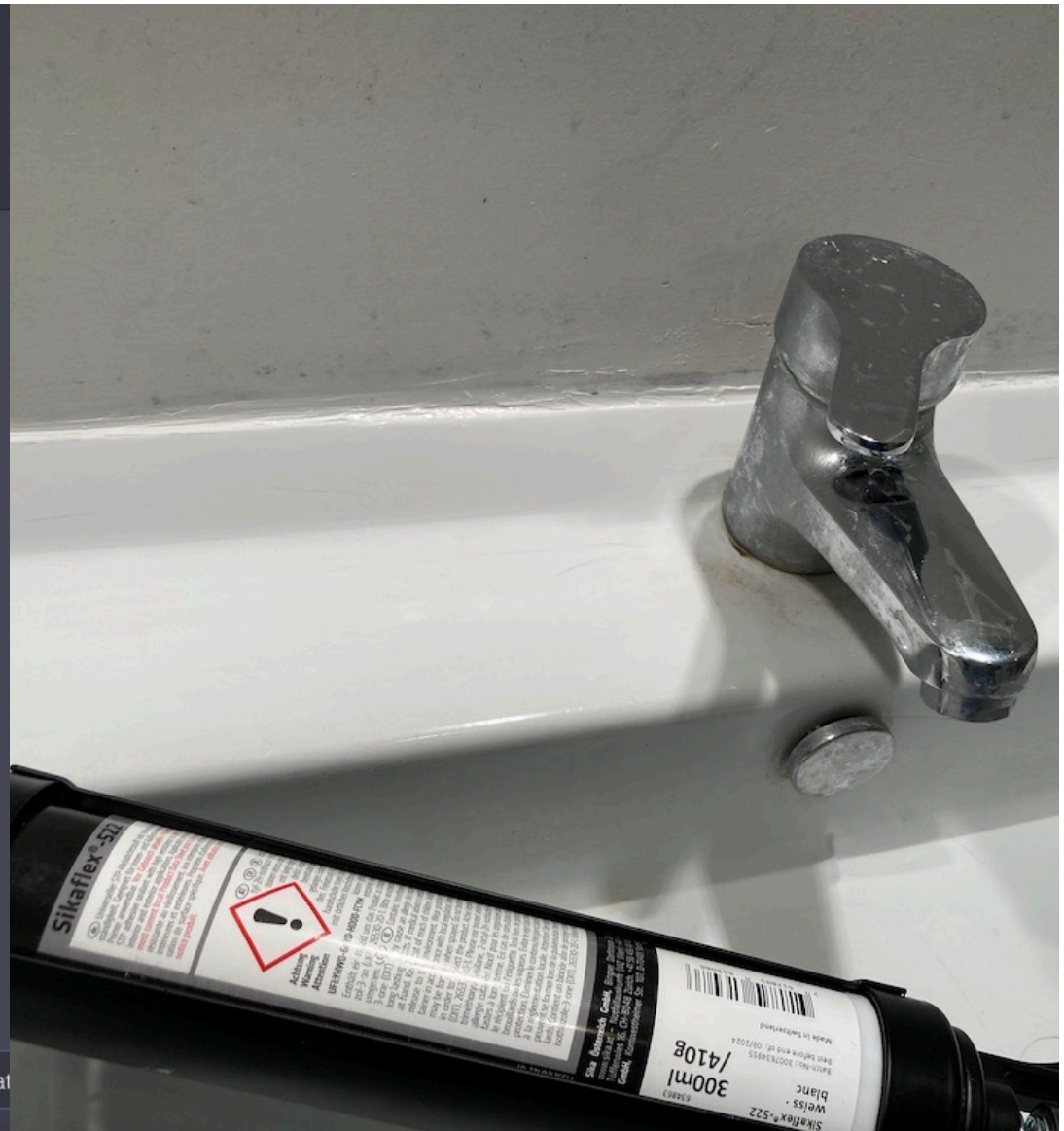
There is water leakage in my bathroom, which is caused the the water spilled over from my wash sinks. Between the sink and the wall, there is a small gap, which is used to be filled by a white substance. The substance was gone after years. Now I want to refill the gap, do you know what is the name of this substance?

Yes, the white substance you're referring to that fills the gap between your sink and the wall is likely caulk, specifically bathroom or silicone caulk. Over time, caulk can degrade, peel, or mold, especially in wet areas like bathrooms. When this happens, it can create a pathway for water to seep through, leading to potential water damage.

If you're planning to re-caulk the gap, here's a basic step-by-step guide:

1. **Preparation**:
   - Clean the area around the gap to ensure it's free of dirt, grime, and mold. This helps the new caulk adhere properly.
   - Remove the old caulk using a scraper, utility knife, or caulk removal tool. Be careful not to scratch or damage the sink or wall.
2. **Choose the Right Caulk**:
   - For bathrooms, it's advisable to use a silicone-based caulk because it's water-resistant and more flexible than other types, accommodating any slight movement between the sink and wall.
3. **Application**:
   - Cut the tip of the caulk tube at an angle. The size of the hole will determine the bead size. For a small gap, a smaller hole is appropriate.
   - Load the caulk tube into a caulking gun.
   - Start at one end of the gap, apply steady pressure to the trigger, and move the gun

Regenerat



QA: Visual Query

GPT-4

Please carefully analyze this driving scene.

Of course! Here's an analysis of the driving scene:

1. **Road Type and Condition**: The road appears to be a multi-lane highway, specifically a divided highway, judging from the central barriers and the multiple lanes designated for one direction of traffic. The road looks well-maintained and has clear lane markings.
2. **Vehicles**:
   - There are three vehicles visible.
   - The vehicle in the foreground is a green truck with a license plate "PFS 06632".
   - Further back, there's another, larger orange-yellow truck.
   - There's also a third smaller vehicle following the larger truck.
3. **Environment**:
   - The visibility seems to be slightly diminished, potentially due to fog, mist, or a similar atmospheric condition.
   - There are wind turbines in the distance, indicating the presence of a wind farm nearby. This suggests that this area might be known for its windy conditions or is trying to harness renewable energy.
4. **Objects on the Road**: There is an object on the road in the foreground that looks like debris or potentially an animal. This can pose a hazard for oncoming vehicles.
5. **Safety Precautions**: There is a barrier between the two directions of traffic, which is a safety measure to prevent head-on collisions. The barrier also appears to have some spacing for water flow, possibly to prevent water accumulation during rainy conditions.
6. **Curvature**: The road ahead curves to the right.

It's always crucial for drivers to remain alert and be aware of their surroundings, especially in conditions with reduced visibility or when there are obstacles on the road.

Agents

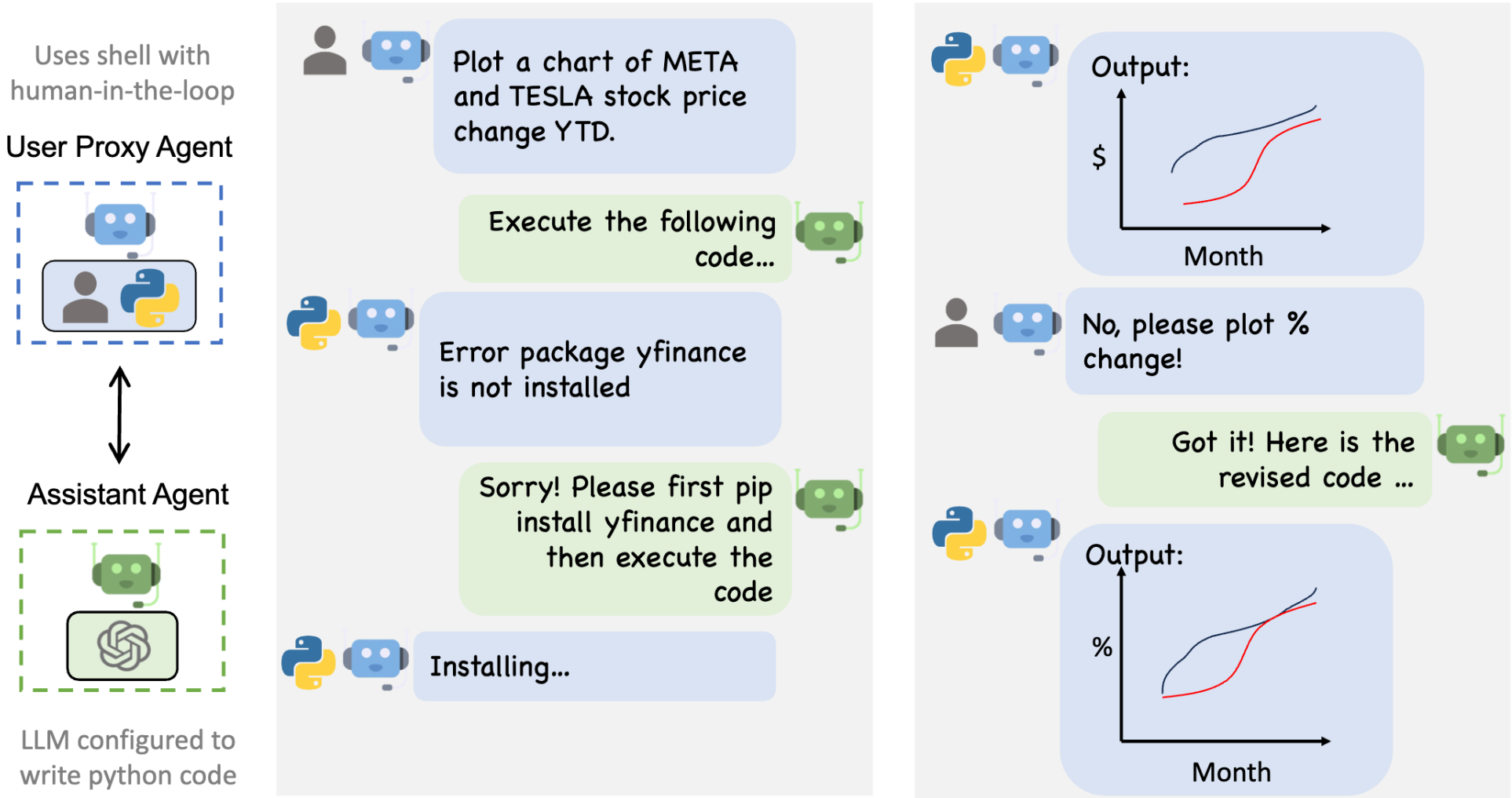


Figure: an example conversation with AutoGen[6]

An overview of LLM tools

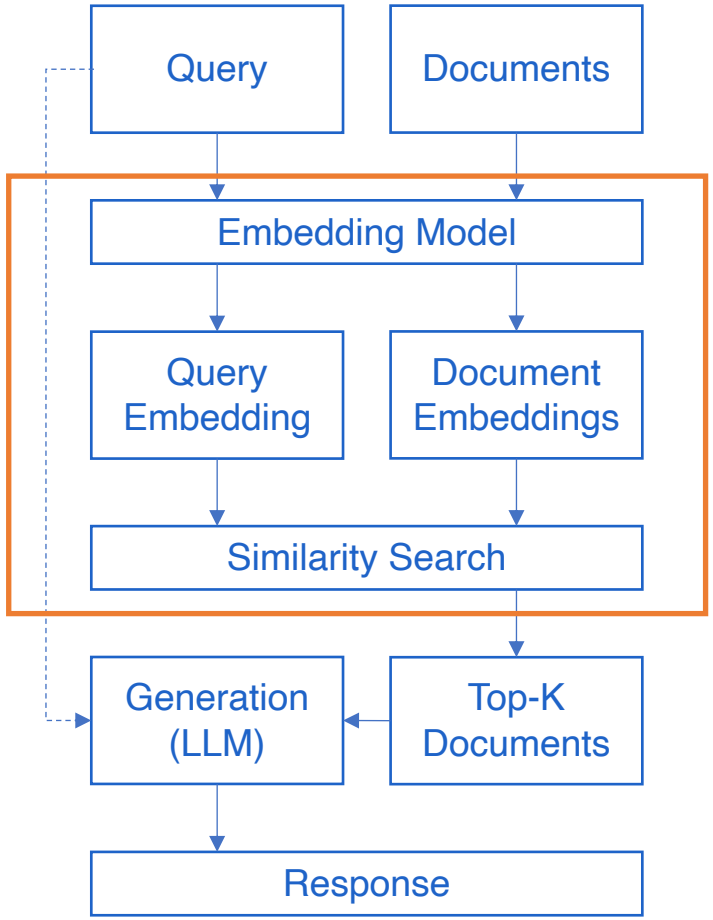
- Commercial
  - OpenAI: ChatGPT, GPT4(V), Finetuning, Embedding…
  - Cohere: Semantic retrieval, Enterprise oriented finetuning, Chat…
  - Others: Claude (Anthropic), Bard (Google)

- Open source
  - Huggingface[1]: model repositories, completion, basic tasks
  - Langchain[2]: tools, prompt templates, chain of actions, memory, retrieval, agents, chain hub
  - Llama index[3], haystack[4]: retrieval
  - AutoGPT[5], AutoGen[6]: agents

Application: Retrieval Augmented Generation (RAG)

- Motivation
  - Hallucination
  - Temporal facts
  - Efficiency and Cost

- Idea: generate response with retrieved relevant documents as context

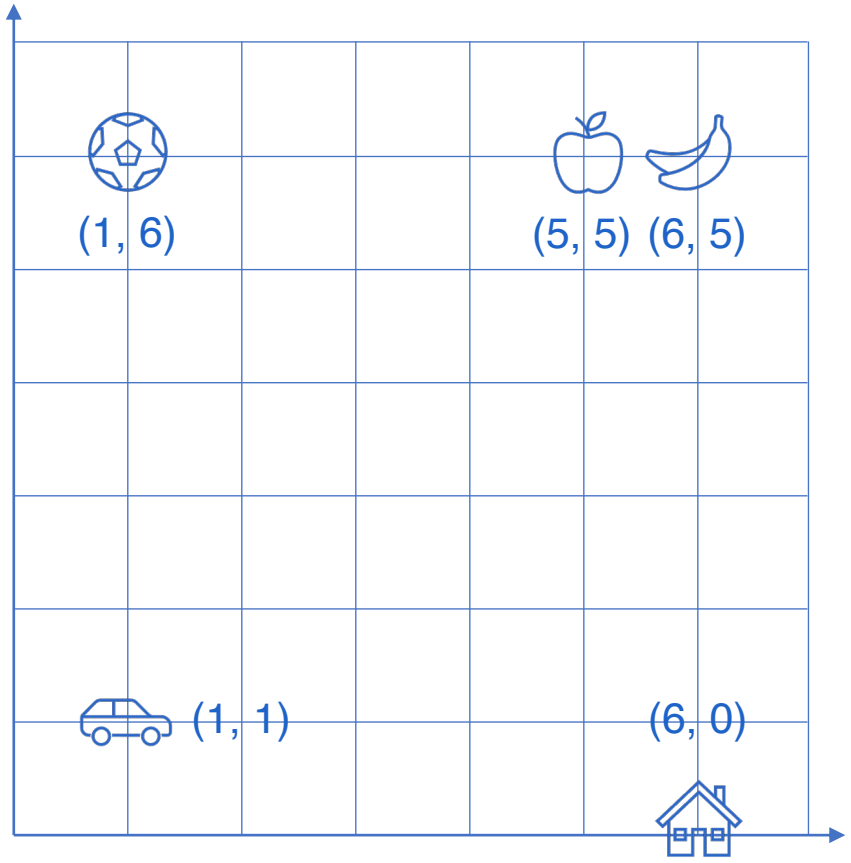


So, what is embedding?

Figure: architecture a typical RAG system

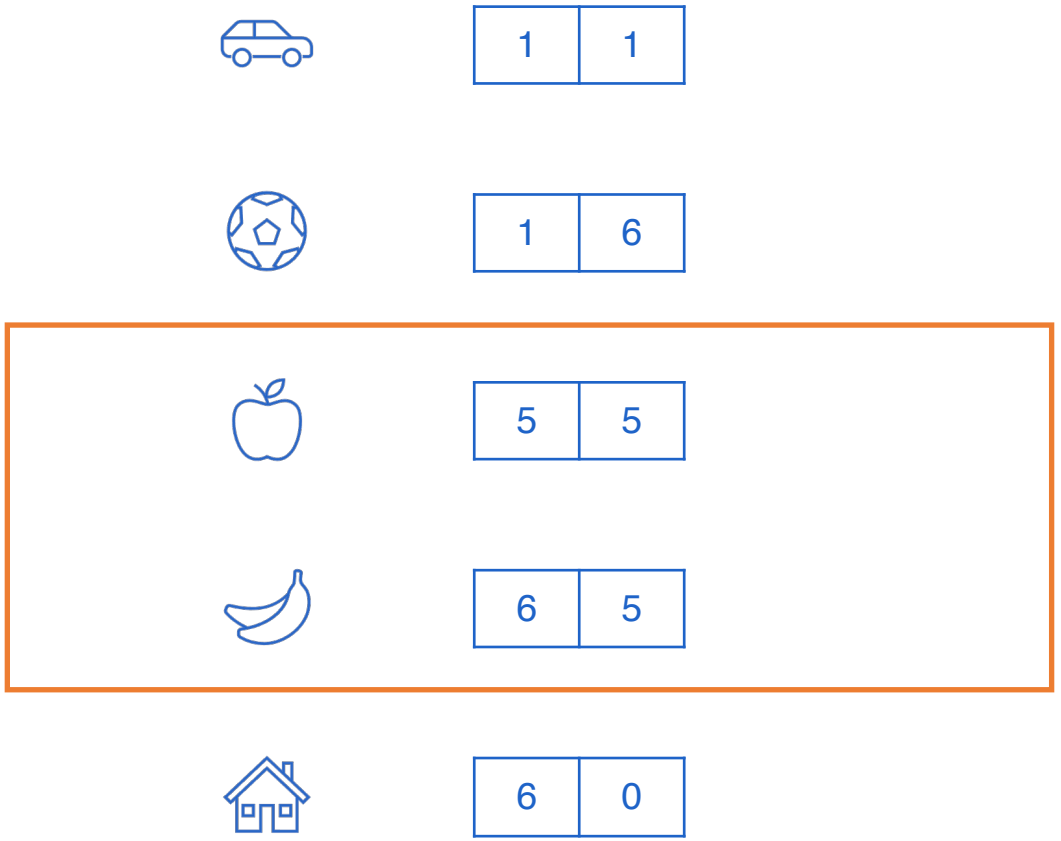
What is embedding?[7]

- Where would you put 🍎 ?
- Use coordinates to quantify the similarities



What is embedding?[7]

- Where would you put 🍎 ?
- Use coordinates to quantify the similarities



What is embedding?[7]

- Where would you put  ?
- Use coordinates to quantify the similarities

0.1	0.6	…	-0.5	1.9

1.3	2.4	…	1.4	-0.5

0.7	-0.1	…	1.1	-1.9

0.6	-0.1	…	1.2	-2.0

1.7	0.9	…	-1.6	0.3

What is embedding?[7]

- Where would you put  ?
- Use coordinates to quantify the similarities
- Embeddings: a unified way to represent objects in neural networks

This is great!	0.1	0.6	…	-0.5	1.9
Today is Friday.	1.3	2.4	…	1.4	-0.5
Hi, what's up?	0.7	-0.1	…	1.1	-1.9
Hello, how are you?	0.6	-0.1	…	1.2	-2.0
I love embeddings.	1.7	0.9	…	-1.6	0.3

Application: Retrieval Augmented Generation (RAG)

- Motivation
  - Hallucination
  - Temporal facts
  - Efficiency and Cost
- Idea: generate response with retrieved relevant documents as context
- Why it helps? More reliable, up-to-date, fine-tune unnecessary

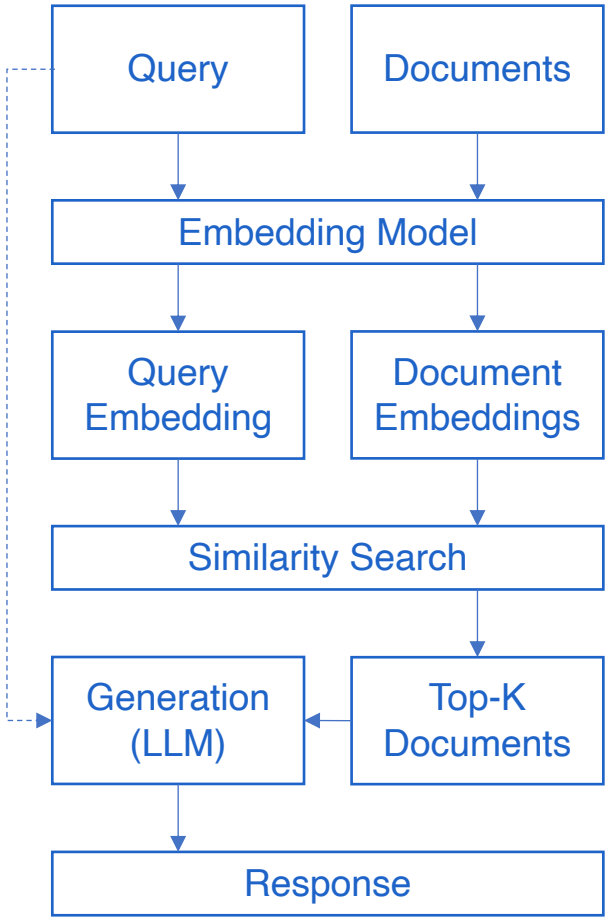


Figure: architecture a typical RAG system

Demo: Ask UGent Helpdesk

- Scrape webpages
- Transform into documents
- Index embeddings and queries
- Search for relevant documents
- Generate response

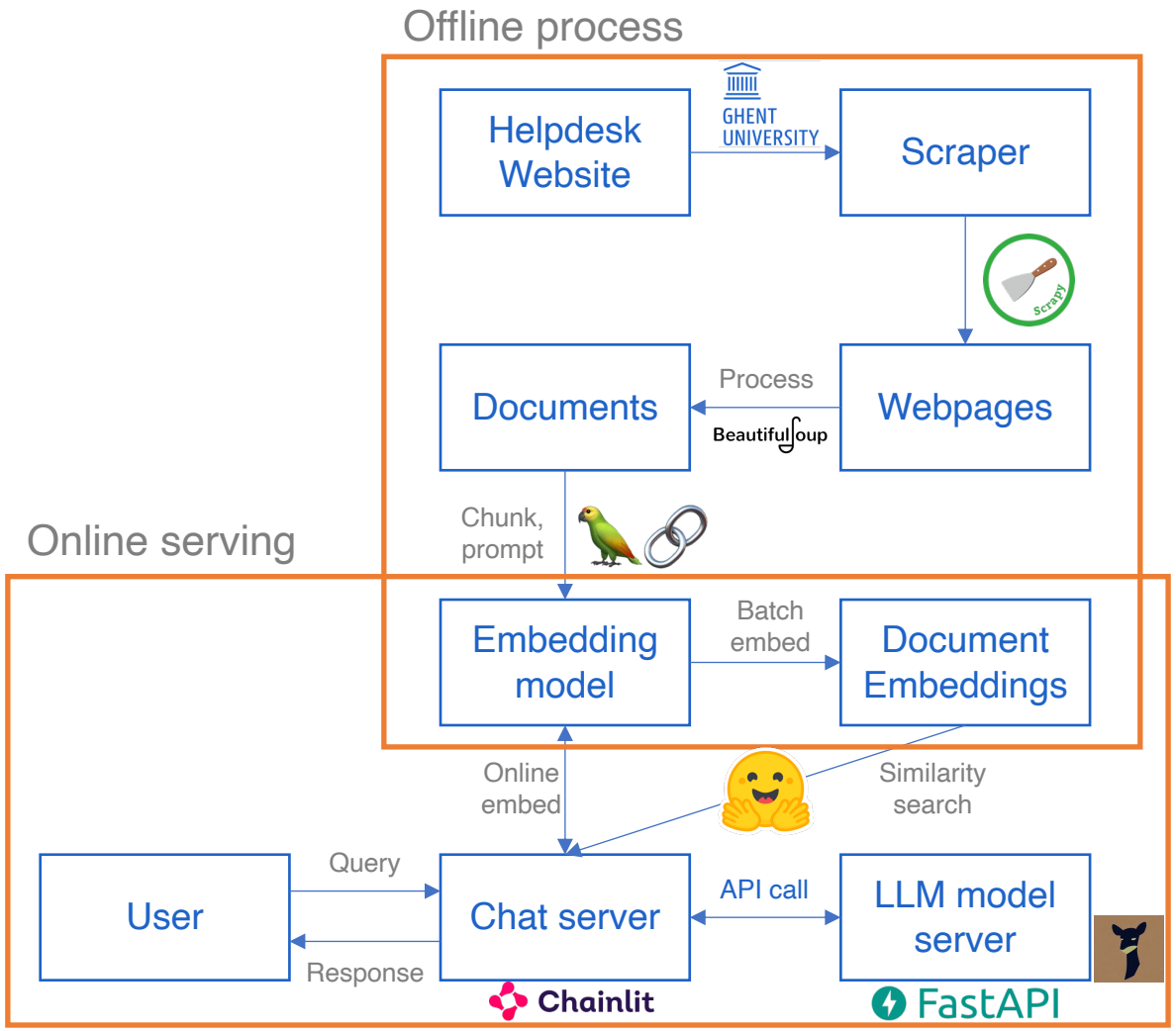


Figure: architecture of Ask UGent Helpdesk app

References

1. https://huggingface.co/
2. https://python.langchain.com/docs/get_started/introduction
3. https://docs.llamaindex.ai/en/stable/
4. https://github.com/deepset-ai/haystack
5. https://github.com/Significant-Gravitas/AutoGPT
6. https://microsoft.github.io/autogen/
7. Luis Serrano, What Are Word and Sentence Embeddings?

Thanks for your attention