

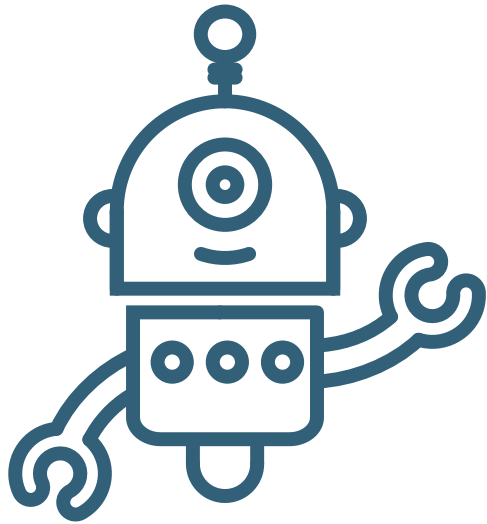
19.8.2025

# RAG Playground

You can delete all the  
info/tips slides once you're  
done with making your  
presentation



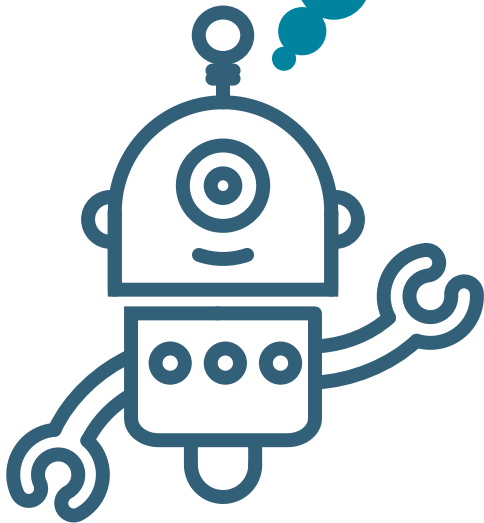
**GOFORÉ**



Hey ChatGPT, who is hosting the upcoming RAG meetup in Salzburg?



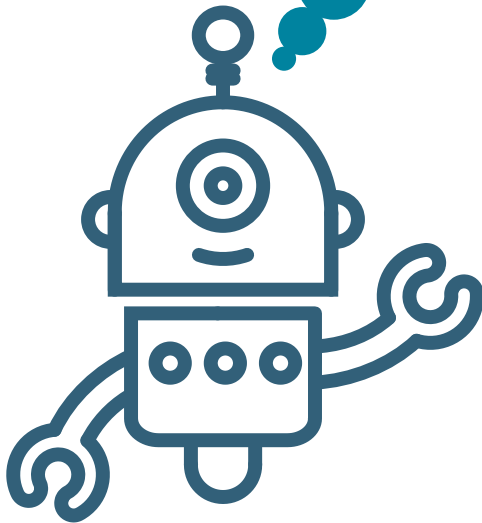
I have an information cutoff and I don't even know what date it is – I cannot certainly know this :/



Hey ChatGPT, who is hosting the upcoming RAG meetup in Salzburg?

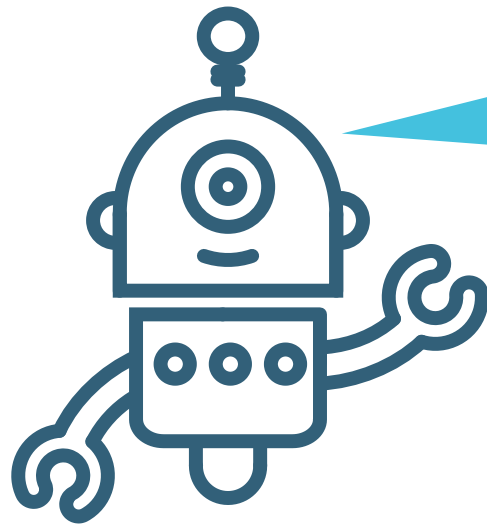


Buuut I have been trained  
to answer the user  
request to their  
satisfaction – I got this,  
I'll make something up!



Hey ChatGPT, who is  
hosting the upcoming  
RAG meetup in Salzburg?

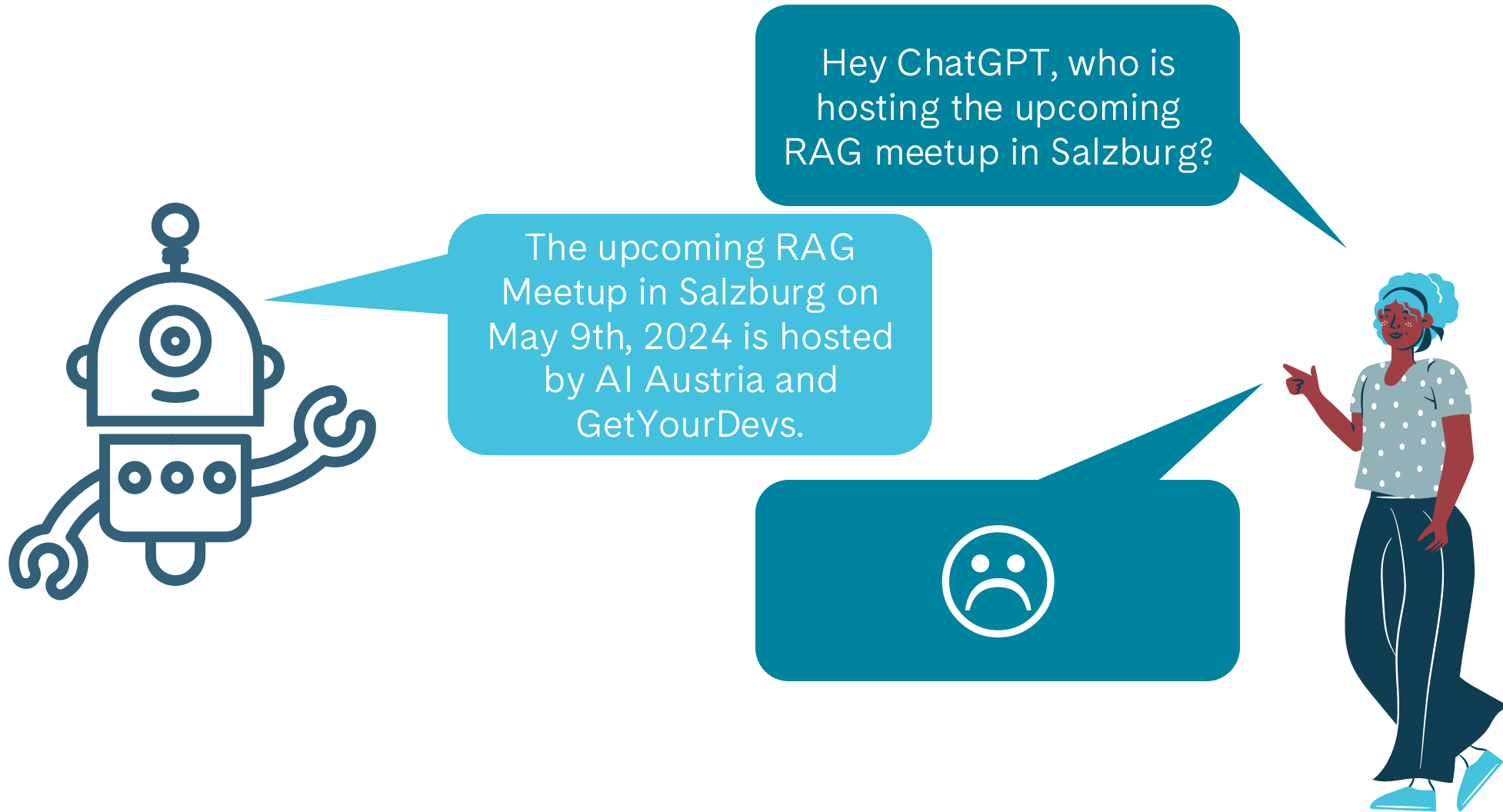




The upcoming RAG Meetup in Salzburg on May 9th, 2024 is hosted by AI Austria and GetYourDevs.

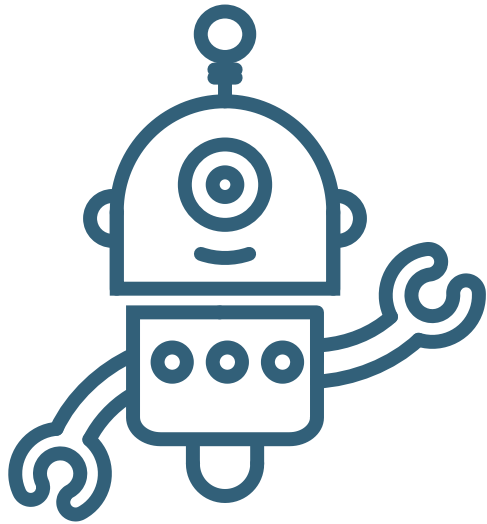
Hey ChatGPT, who is hosting the upcoming RAG meetup in Salzburg?







Let's try again

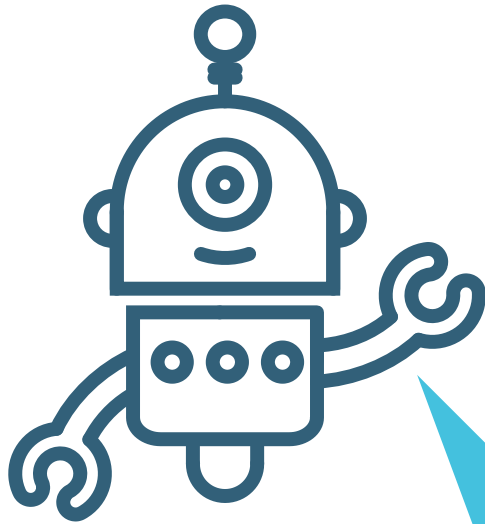


Hey ChatGPT, who is hosting the upcoming RAG meetup in Salzburg?










The hosts are Julian F.  
and two others—as listed  
on the meetup page  
[meetup.com](https://www.meetup.com).

User Question: Hey  
ChatGPT, who is  
hosting the upcoming  
RAG meetup in  
Salzburg?

Relevant Context:

**RAG from the Playground Up @ Gofore**

Hosted By  
Julian F. and 2 others



**Software  
Craft and  
Testing  
Community  
Salzburg**

4.9 ★★★★★ 100 reviews

Tuesday, August 18,  
2020  
6:00 PM to 9:00 PM  
CEST

Gofore Austria  
Imstbrucker  
Bundesstraße 71, 5020  
Salzburg, Salzburg

**Details**  
In this session, we'll build a basic **Retrieval-Augmented Generation (RAG)** setup from scratch. The goal is to connect an AI tool and data source of your choice, such that the tool's responses take your provided external information into account.

**What Will We Do?**  
We'll work in small groups or pairs to create a simple RAG pipeline. This includes



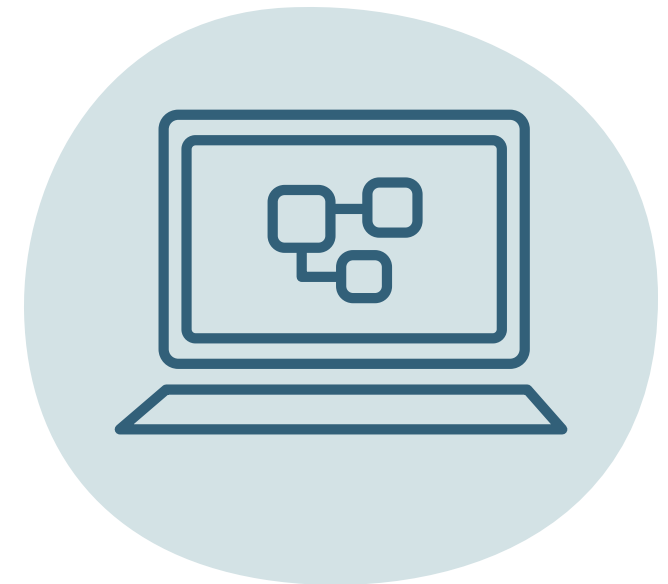
# What is this “RAG magic”?

- Intercept user query
  - The user thinks they are sending the prompt directly to the LLM, but we intercept it and add information that allow the LLM to answer correctly
    - A first step here can be to insert static additional context and verify that the model processes it
- Information Lookup
  - Any sort of information lookup will work – be creative
    - First step can be passing certain parts of the dataset based on keywords
  - Classic approach:
    - Create chunks of information (sentences, paragraphs, ...)
    - Use a Vector DB to search for chunks that are semantically most similar to the user query by comparing embeddings using similarity metrics
    - You found your information relevant to the user query
- Insert additional information into context
  - Pass that extra information you retrieved to the LLM context



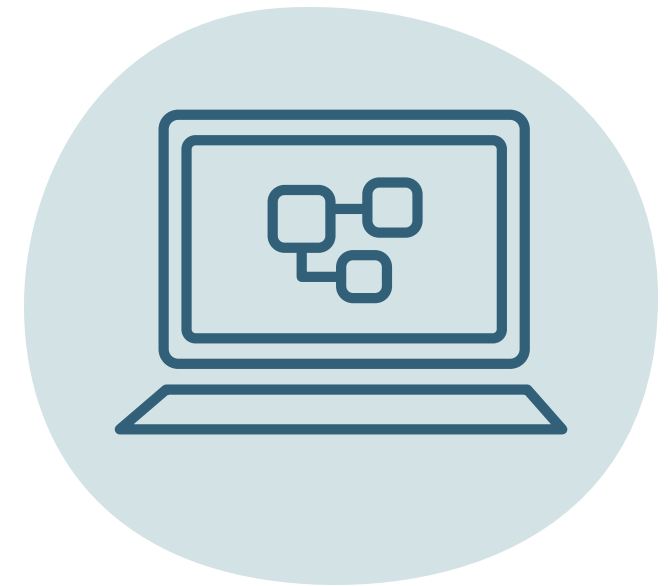
# Part 1

- Setup repo
- RAG setup
  - Bring you own data
  - Structured dataset
  - Unstructured dataset
  - Unstructured extended dataset
  - Answer the test questions
    - Think of questions the test this system
    - Make the system answer correctly
    - Everything works? Find a new question that does not work
    - Repeat

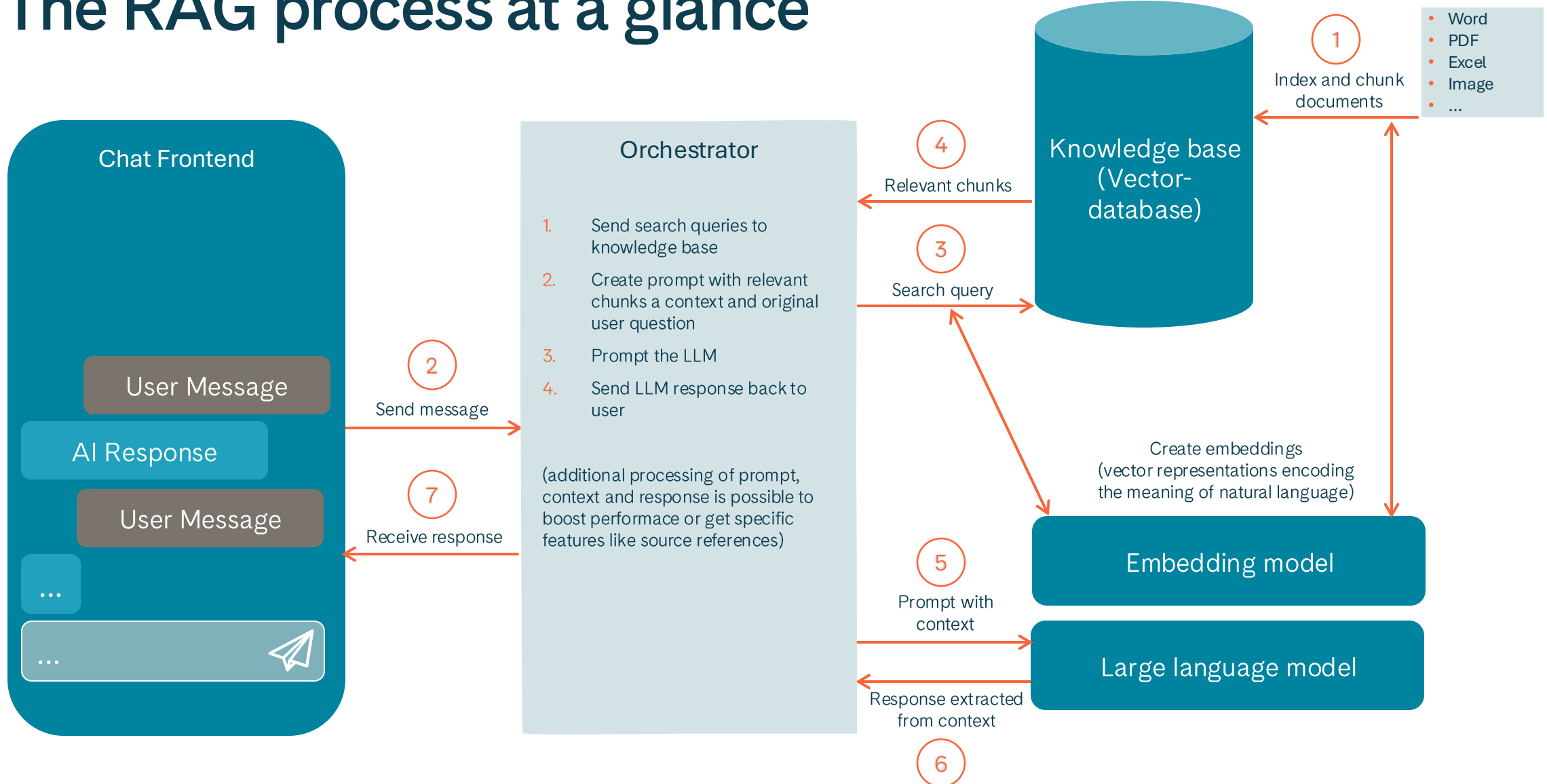


# Part 2

- Once the questions work, the sky is the limit:
  - Make the retrieval optional for queries that don't require it?
  - Minimize the number of tokens you send to the LLM API?
  - Improve the retrieval accuracy?
  - ...

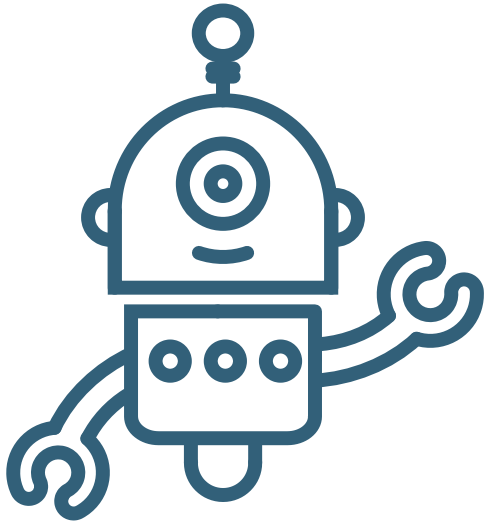


# The RAG process at a glance

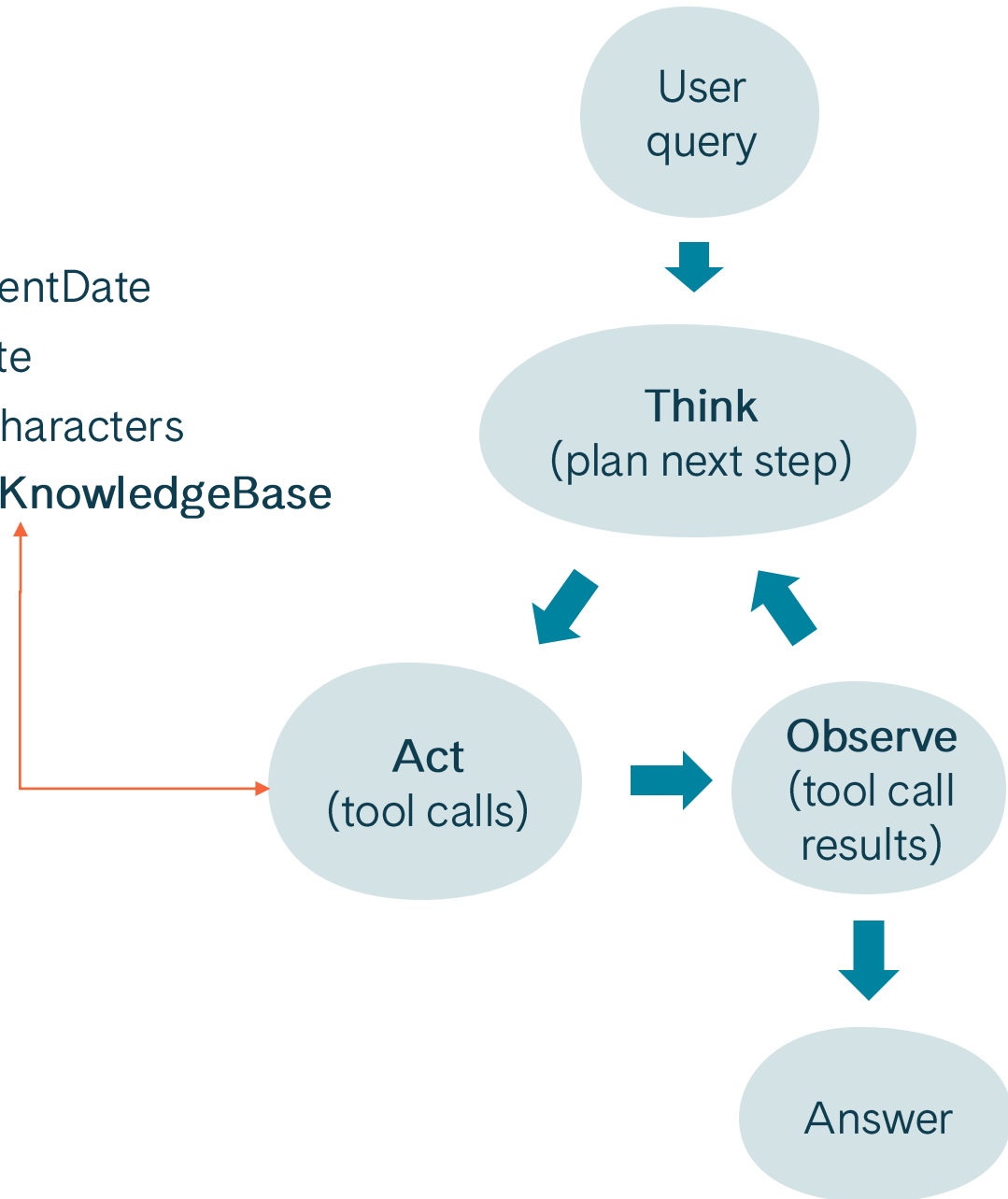


# Outlook

Agentic RAG



- Tool: getCurrentDate
- Tool: Calculate
- Tool: CountCharacters
- Tool: SearchKnowledgeBase



Pioneering  
an ethical  
digital world.

A large, solid teal circle is positioned to the right of the text, partially overlapping the word "world".

**GOFORE**