

PHUSE EU Strasbourg 2024

**Unlocking the potential of your data with RAG while upholding
privacy:** use cases with business impact at Roche.

Adam Forys, Principal Data Scientist
Mathieu Cayssol, Data Scientist

Table of contents

What are LLMs and their benefits?

Why Retrieval Augmented Generation?

What is a RAG?

1. Architecture
2. Generate embeddings
3. Cosine similarity
4. Prompt crafting
5. Answer generation

Business case:

1. OCEAN assistant
2. Code suggestions

What are LLMs and what is a RAG?

What are LLMs

Definition and benefits

Definition

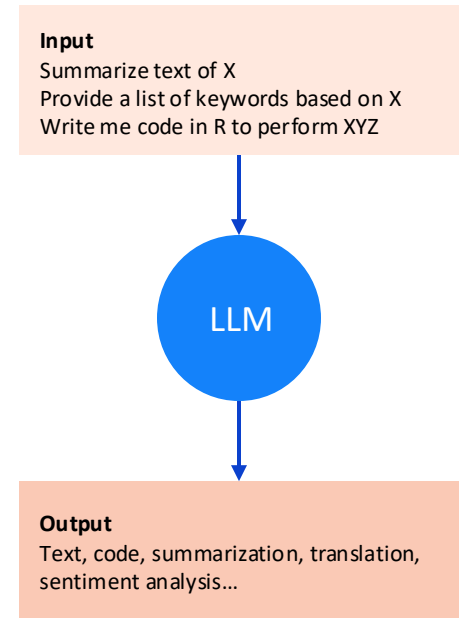
Large language models (LLMs) are **deep neural network** trained on immense amounts of data making them capable of **understanding and generating natural language** and other types of content.



Capabilities

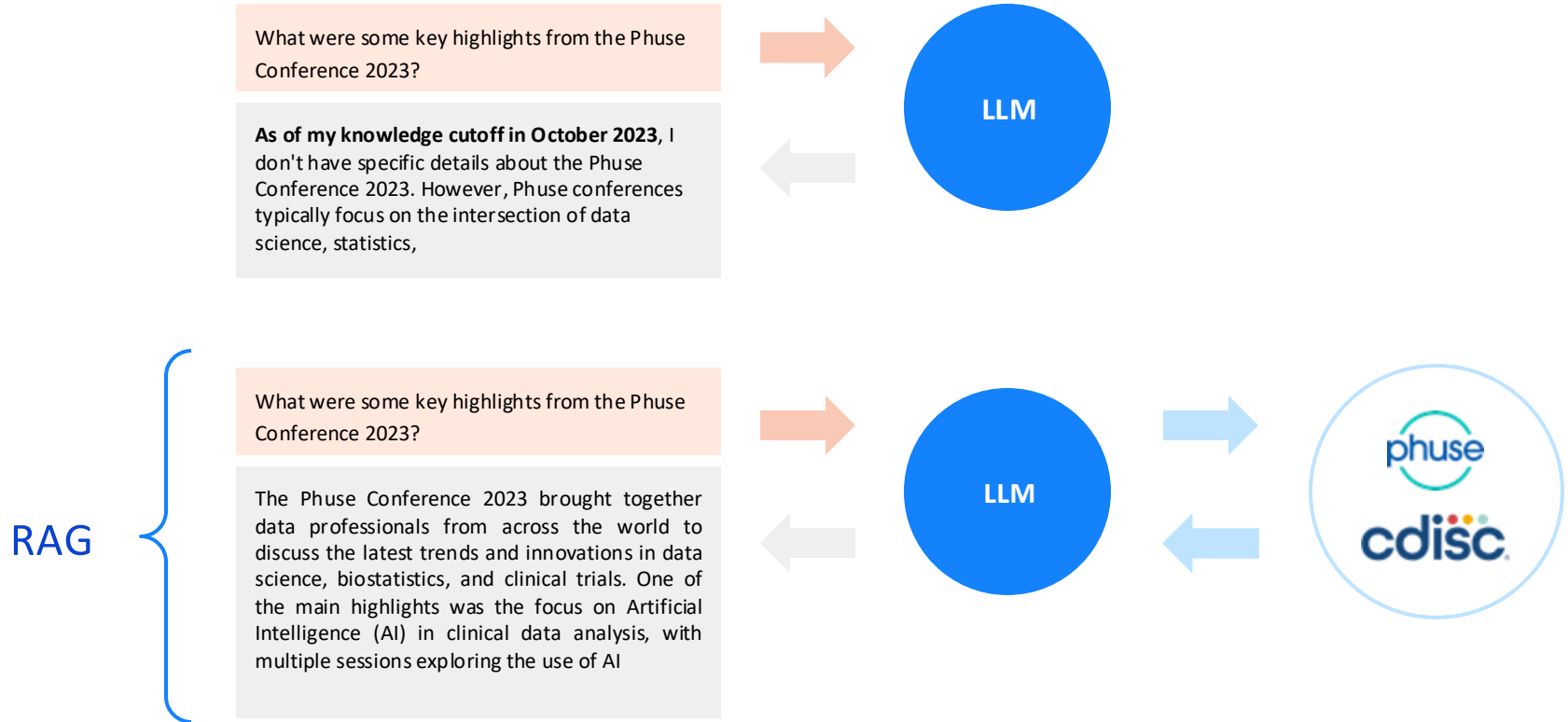
- Text generation
- Content summarization
- AI assistants
- Code generation
- Sentiment analysis
- Language translation

High-level overview



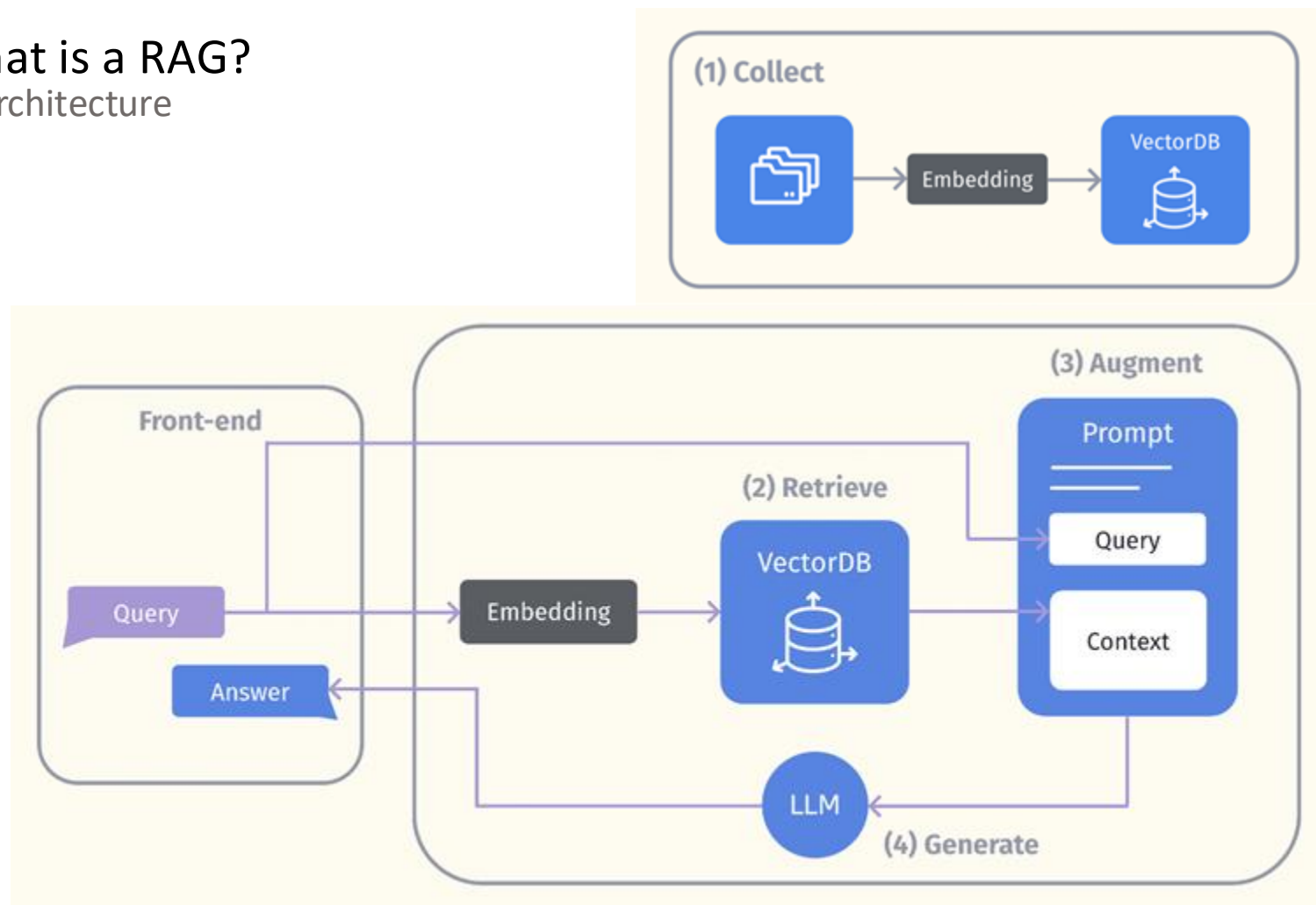
Why Retrieval Augmented Generation (RAG)?

LLM + Knowledge base



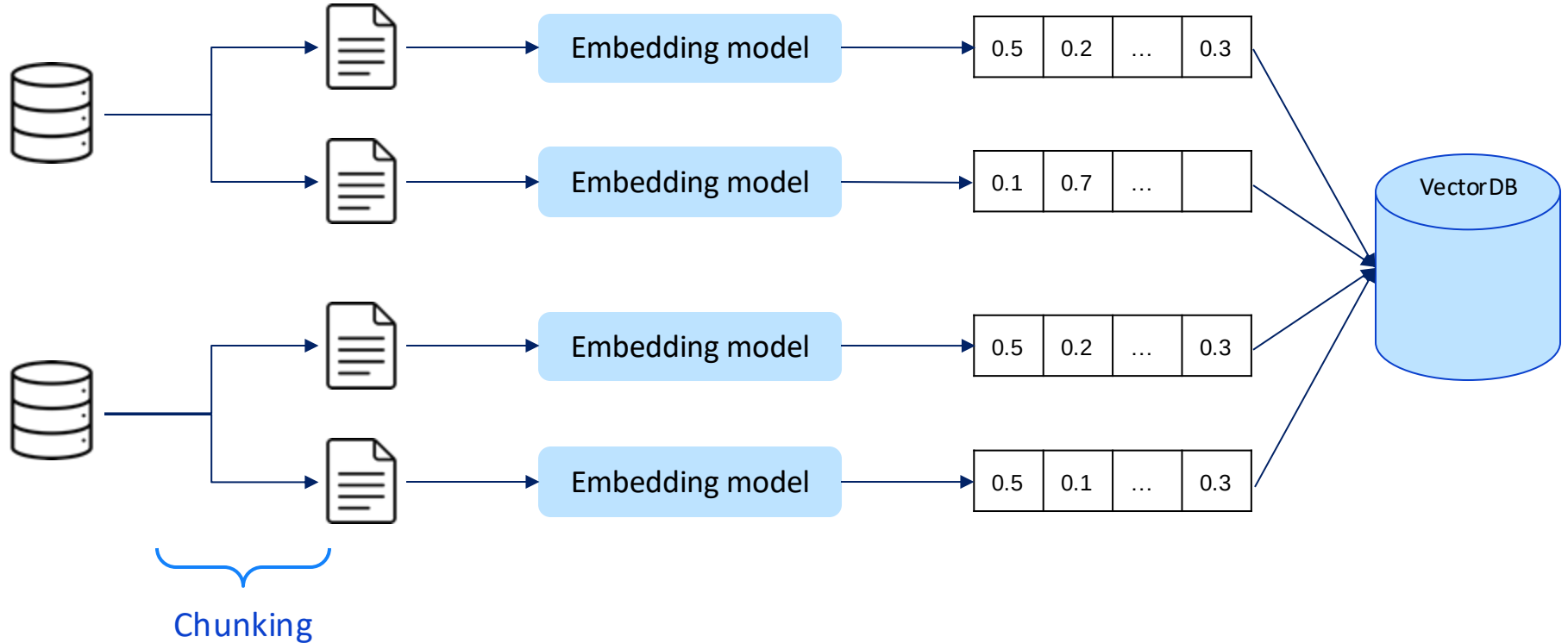
What is a RAG?

1. Architecture



What is a RAG?

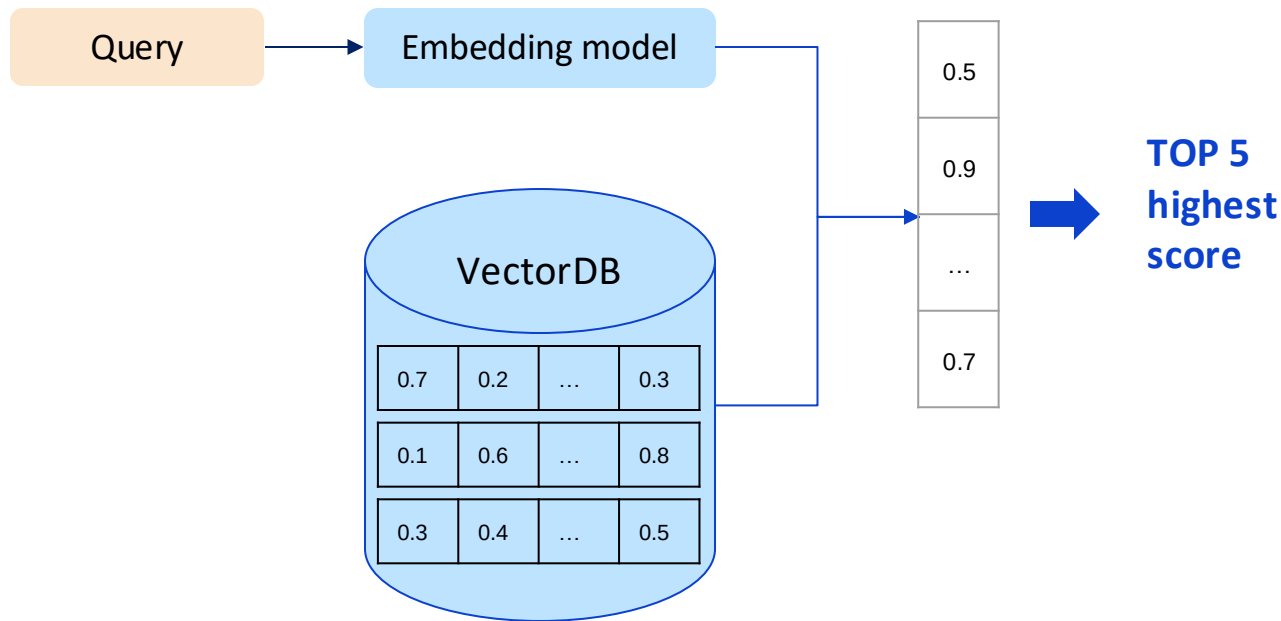
2. Generate embeddings



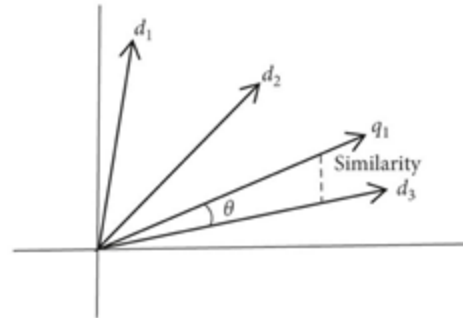
What is a RAG?

3. Cosine similarity

Vector of normalized cosine similarity score
between the {query} and {documents}



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$



What is a RAG?

4. Prompt crafting

[SYSTEM]

You are a helpful assistant. Answer the question based on the provided context from conference materials and industry news.

[USER]

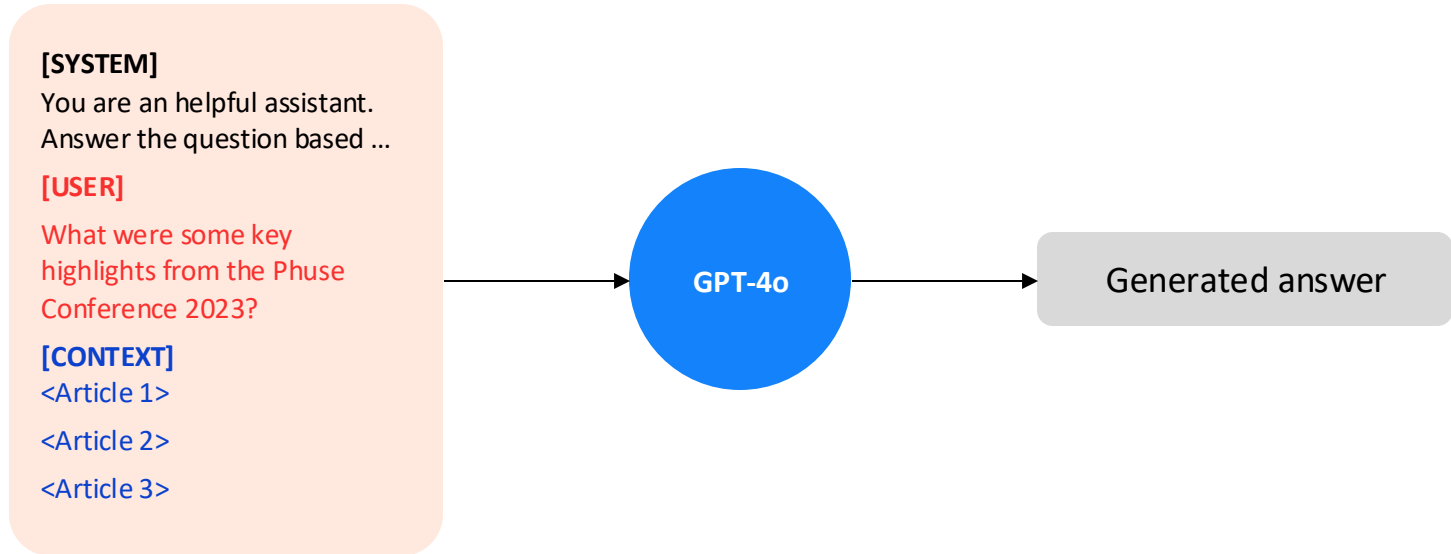
What were some key highlights from the Phuse Conference 2023?

[CONTEXT]

The Phuse Conference 2023 brought together data professionals from across the world to discuss the latest trends and innovations in data science, biostatistics, and clinical trials. One of the main highlights was the focus on Artificial Intelligence (AI) in clinical data analysis, with multiple sessions exploring the use of AI to improve patient outcomes, streamline trial processes, and enhance data quality. Additionally, the conference emphasized regulatory challenges, with key speakers from the FDA and EMA providing insights into evolving guidelines for data submission and compliance in the pharmaceutical industry.

What is a RAG?

5. Answer generation



Business case: OCEAN Assistant

OCEAN assistant

Business context

New platform

Migration to OCEAN platform (~90% migration goal)



New languages

From SAS to SAS+R



New processes/tools

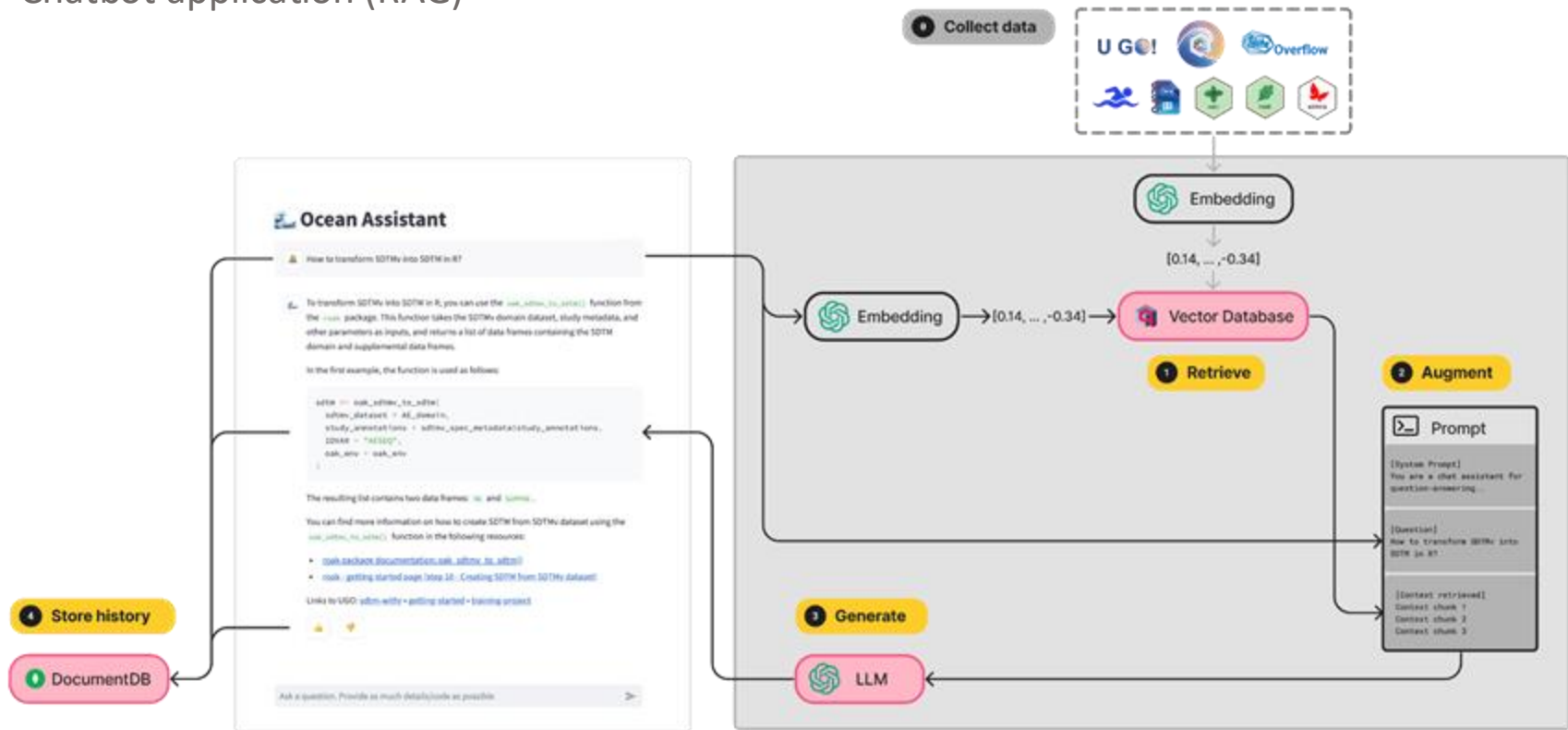
Git workflow, GitLab, orchestration...



➡ **Steep learning curve on new tools, despite the documentation & training materials**

OCEAN assistant

Chatbot application (RAG)



OCEAN assistant

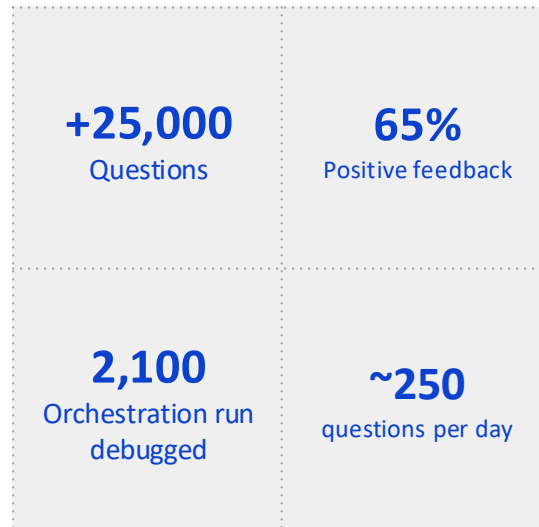
Capabilities and metrics

Capabilities

- Summarize content from the documentation
- Retrieve documentation link
- Generate simple snippet of code
- Solve potential errors you encounter
- Explain processes
- Retrieve forum discussions

Example of questions:

- What is the orc command?
- How to create AE domain with roak?
- WorkflowError: MissingInputException: Missing input files for rule adsl
- How to deal with merge conflict? Explain step-by-step.
- Do I need to manually set dictionary versions when creating TS domain
- What is the dataflow for non-CRF data in OCEAN?



Demo

Agents

Ocean Assistant

Orchestration Debugger

NExpert ①

User Guide

Data sources

Contact

 OCEAN Assistant **Easy Start**

Ocean Assistant is an AI tool for answering questions using Ocean's documentation, accessible 24/7 within the Roche network.

Example of questions:

- What is the orc command?
- How to create AE domain with roak?
- WorkflowError: MissingInputException: Missing input files for rule adsl
- How to deal with merge conflict? Explain step-by-step.
- Do I need to manually set dictionary versions when creating TS domain

 Capabilities

Ocean Assistant:

- Retrieve documentation link
- Generate simple snippet of code
- Solve potential errors you encounter
- Understand processes
- Retrieve Roverflow posts
- Question upstream to Ocean

Orchestration Debugger:

- Error extraction from the logs
- Potential solution for troubleshooting
- Retrieve Roverflow posts

 Limitations

- Ocean Assistant is a genAI app that uses Retrieval-Augmented Generation (RAG).
- Large Language Models (LLMs) are prone to hallucination.
- The retrieved information might be misleading.
- The quality of the answer is bound to what is available in the documentation.

Ask a question. Provide as much details/code as possible



Business case: code suggestions

Code suggestions

Business context

Existing tools

- GitHub Copilot
- GitLab Duo
- Codium

Common IDE

New functionalities in RStudio IDE

- Support for GitHub Copilot
- API for custom extensions

Knowledge Repository

All R package documentations, code snippets, templates, studies codebase with extra metadata

 **Improve the user productivity**



Code suggestions

Build knowledge base

The **roxygen2** package allows for the creation of custom roclets.

It is necessary to extract data from **multiple R packages** and develop a generalized solution that can be **reused** in many use cases.

The vector storage is required to be **maintained** and **kept up-to-date**, which is crucial for accurate suggestions.

Enhancing user **productivity** is a key objective.

We need a **simple format** to **manage** the **versions** of the **packages effectively**, which will facilitate better version control and dependency management.

The internal CRAN/PPM (Posit Package Manager) **repository** serves as a reliable source of truth regarding the **currently used packages**, ensuring consistency and accuracy.



Code suggestions

Build knowledge base

- **Enhanced Roxygen Functionality:** roxytractor extends roxygen package with a custom roclat.
- **Structured Documentation:** Extracts roxygen package documentation into a tabular format for easy reference.
- **Code Snippet Extraction:** Provides tools for extracting code templates and snippets
- **Internal Dataset Sharing:** Reuse and share structured datasets within your organization.



```
packages <- readLines(system.file("roxytractor_packages", package = "roxytractor"))

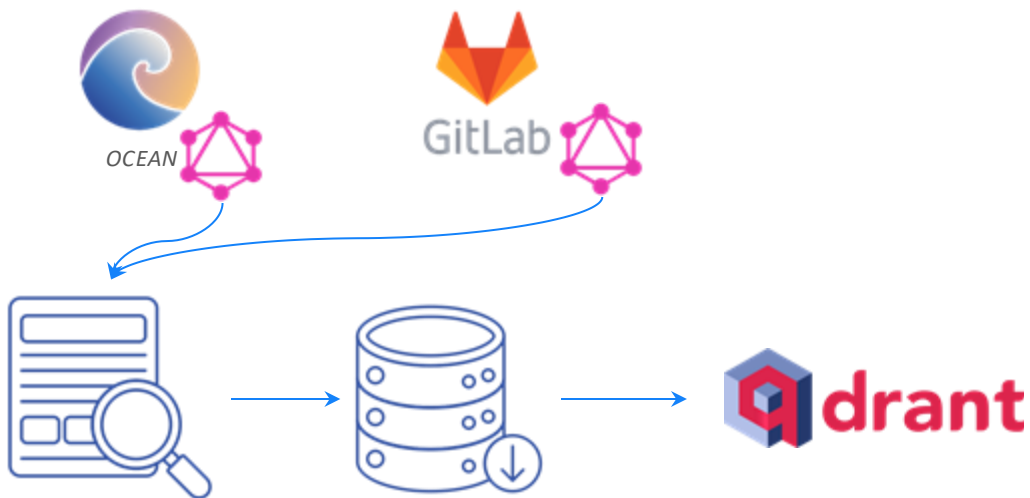
library(roxytractor)

extract_from_packages(
  packages = packages,
  output_format = "parquet",
  output_path = "package_examples.parquet",
  append = FALSE,
  load_code = "installed"
)
```

Code suggestions

Build knowledge base

- Gather Code from the Research Team
- Enhance Code with Metadata (package version, study status, pipeline status, creation date)
- OCEAN platform adopts FAIR data principles



```

1
2 query GetGroupRepositories($groupPath: ID!, $path: [String!]!) {
3   group(fullPath: $groupPath) {
4     id
5     name
6     fullPath
7     projects {
8       count
9       pageInfo {
10        startCursor
11        endCursor
12        hasNextPage
13      }
14      nodes {
15        id
16        name
17        httpUrlToRepo
18        sshUrlToRepo
19        lastActivityAt
20        visibility
21        createdAt
22        repository {
23          rootRef
24          blobs_devel: blobs(paths: $path, ref: "devel") {
25            nodes {
26              name
27              path
28              rawBlob
29            }
30          }
31        }

```

Code suggestions

Custom extension

- RStudio **Add-in** for Enhanced Code Completion
- **Non-blocking** Asynchronous HTTP Requests
- Supports for many API **providers**
- **Flexibility** with various data sources



```
test2.R+ x
Source on Save

1 # Create app template using teal package
2 library(teal)

app <- init(
  data = teal_data(IRIS = iris, MTCARS = mtcars),
  modules = example_module()
)

if (interactive()) {
  shinyApp(app$sui, app$server)
}
```

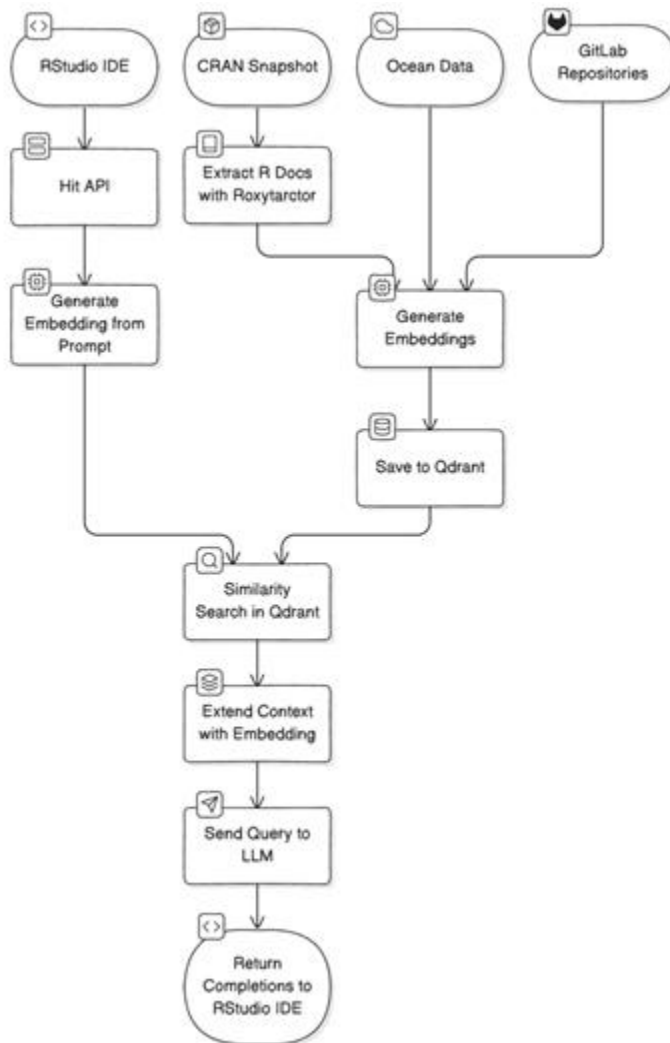
2:1 (Top Level) :

Code suggestions

RAG Flow

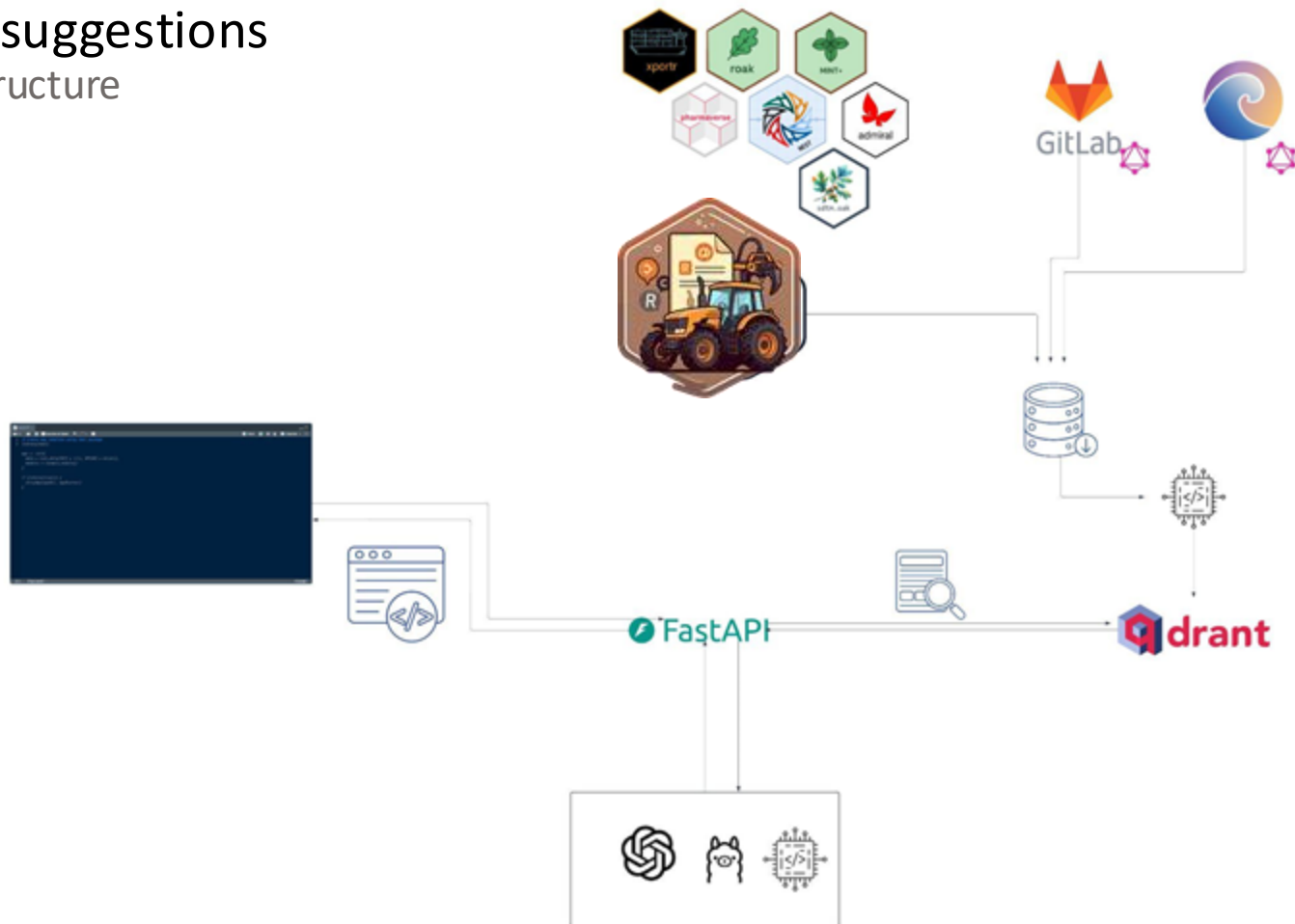
What I've learned?

- Prioritize data preparation.
- Keep R package documentation **up-to-date** and **well-maintained**.
- Applying the **FAIR principles** in your project can **save** thousands of hours of **manual work**.



Code suggestions

Infrastructure



Demo - code suggestions

```
example.R* x example_2.R x example_3.R x
Source on Save Run ↑ ↓ Source
1 # Name: ADAE
2 #
3 # Label: Adverse Event Analysis Dataset
4 #
5 # Input: ae, adsl, ex_single
6 library(admiral)
7 library(pharmaversesdtm) # Contains example datasets from the CDISC pilot project
8 library(dplyr)
9 library(lubridate)
10
11 # Load source datasets ----
12
13 ae <- pharmaversesdtm::ae
14 suppaes <- pharmaversesdtm::suppaes
15 adsl <- admiral::admiral_adsl
16 ex_single <- admiral::ex_single
17
18 ae <- convert_blanks_to_na(ae)
19 ex <- convert_blanks_to_na(ex_single)
20
21 # Get list of ADSL vars required for derivations
22 adsl_vars <- exprs(TRTSDT, TRTEDT, DTHDT, EOSDT)
23
24 adae <- ae %>%
25   # join adsl to ae
26   |
27
```

Q&A

Doing now what patients need next