

The Levels of RAG



Jeroen Overschie

PyData Eindhoven — 11th of July 2024



Jeroen Overschie
Machine Learning Engineer

Why RAG ?

- **Knowledge freshness**
 - Traditional LLMs are trained on static data, which may be outdated or incomplete.
 - RAG allows the model to access and retrieve information from external sources, ensuring it provides up-to-date and relevant answers.
- **Domain-specific expertise**
 - RAG enables the model to leverage specialized knowledge bases or databases, enhancing its performance in specific domains.
- **Reduced hallucinations**
 - By grounding the model's responses in retrieved information, RAG helps reduce the risk of generating false or misleading content (hallucinations).
- **Improved accuracy**
 - RAG can significantly improve the model's accuracy by providing it with relevant context and supporting evidence for its answers.
- **Flexibility and adaptability**
 - RAG systems can be easily updated and adapted to new information sources, allowing them to stay current and responsive to changing requirements.
- **Enhanced transparency**
 - RAG provides a clear path from the user's question to the retrieved information, making the model's reasoning process more transparent and explainable.
- **Cost efficiency**
 - RAG can be more cost-effective than fine-tuning large LLMs for specific tasks, as it leverages pre-trained models and external data sources.

Why RAG ?



RAG
paper^[1]



ChatGPT
launch

2019

2020

2021

2022

2023

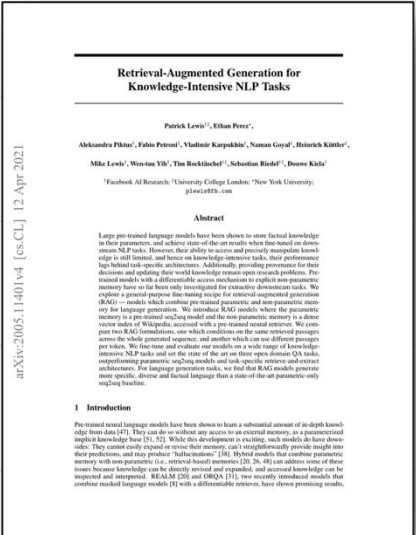
2024

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks – Lewis et al.

Why RAG ?



“Retrieval Augmented Generation”



ChatGPT
launch

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks – Lewis et al.

Why RAG ?

Google Trends



“Retrieval Augmented Generation”

✓ Interest



ChatGPT launch

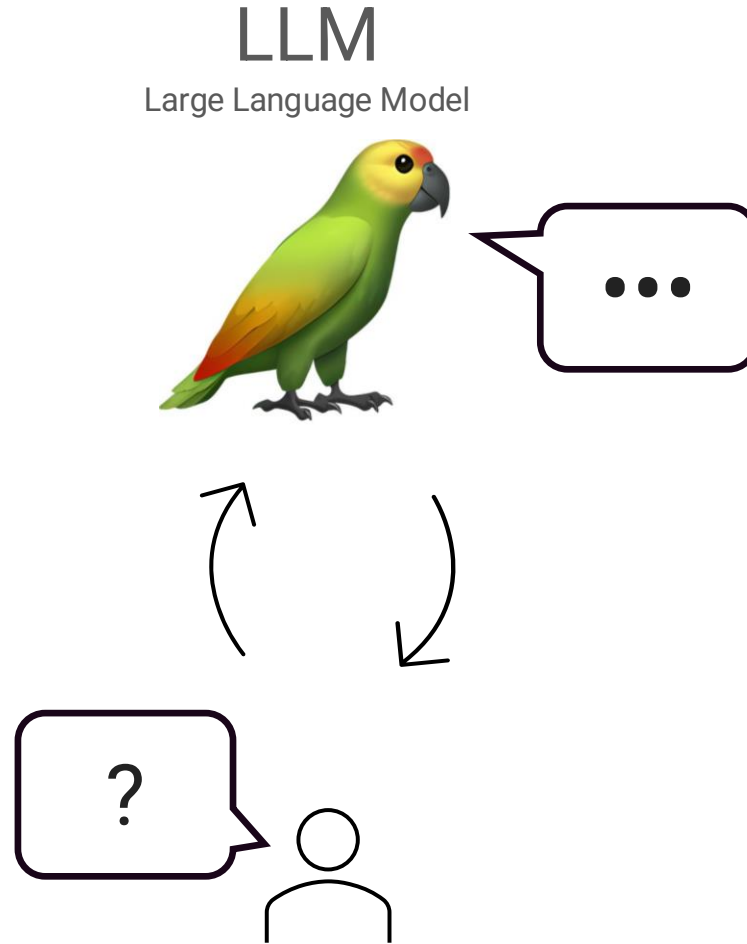
[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks – Lewis et al.
[2] Google Trends index for search query “Retrieval Augmented Generation” worldwide.

Why RAG ?

- **Knowledge freshness**
 - RAG can be updated with new information without retraining the model.
 - This is particularly useful for applications where the knowledge base is constantly changing.
- **Domain-specific expertise**
 - RAG can be tailored to specific domains or tasks by using relevant documents as input.
 - This allows for more accurate and context-aware responses.
- **Transparency and explainability**
 - RAG systems can provide references to the source documents used to generate the response.
 - This helps users understand the basis of the information provided.
- **Reduced hallucinations**
 - By grounding the model's output in real-world data, RAG can significantly reduce the risk of generating false or misleading information.
- **Scalability**
 - RAG can be scaled to handle large volumes of data and complex queries.
 - This makes it suitable for enterprise-level applications.
- **Integration with existing systems**
 - RAG can be easily integrated with existing databases, APIs, and other data sources.
 - This allows for seamless data flow and improved system efficiency.

Why RAG ?

Model interactions
without RAG



Why RAG ?

Model interactions without RAG

Prompt

What was the result for the European cup 2024 game Netherlands against England?

Prompt

What is your knowledge cut-off?

LLM

Large Language Model



Answer

The European Cup 2024 has not yet taken place, so there is no result for the game between Netherlands and England..^[1]

[1] Produced with [gemini-1.0-pro-001](#), temperature = 0.0, top_p = 1.0

Answer

My knowledge cutoff is April 2023.

This means that I do not have access to real-time information or events that have occurred after this date.

However, I am able to provide information and answer questions based on my knowledge up to April 2023. If you have questions about events or information beyond this date, I may not be able to provide an accurate or up-to-date response.

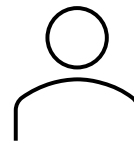
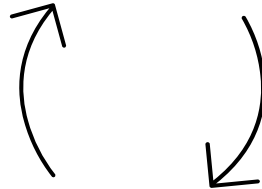
Is there anything else I can help you with today?^[1]

[1] Produced with [gemini-1.0-pro-001](#), temperature = 0.0, top_p = 1.0

Why RAG ?

Model interactions
without RAG

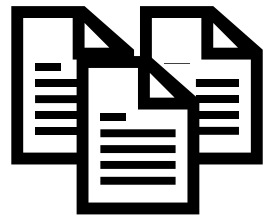
LLM
Large Language Model



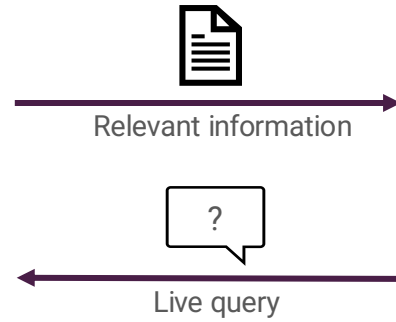
- × Outdated knowledge
- × No internal company data
- × Hallucinations

Why RAG ?

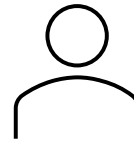
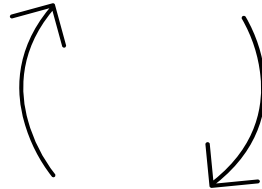
Model interactions
with **RAG**



Documents
/ database



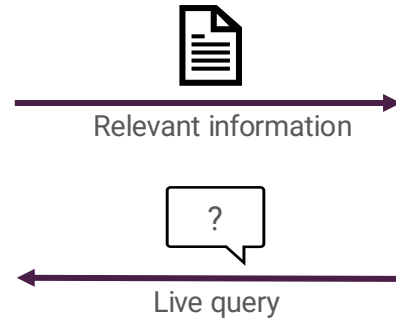
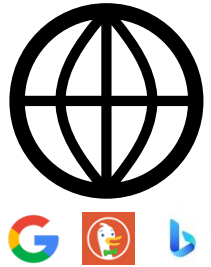
LLM
Large Language Model



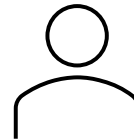
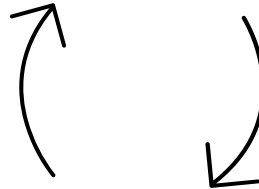
Why RAG ?

Model interactions
with **RAG**

Search engine



LLM
Large Language Model



Why RAG ?

Model interactions
with RAG

Prompt

What was the result for the
European cup 2024 game
Netherlands against England?

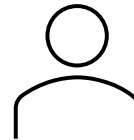
LLM
Large Language Model



Answer

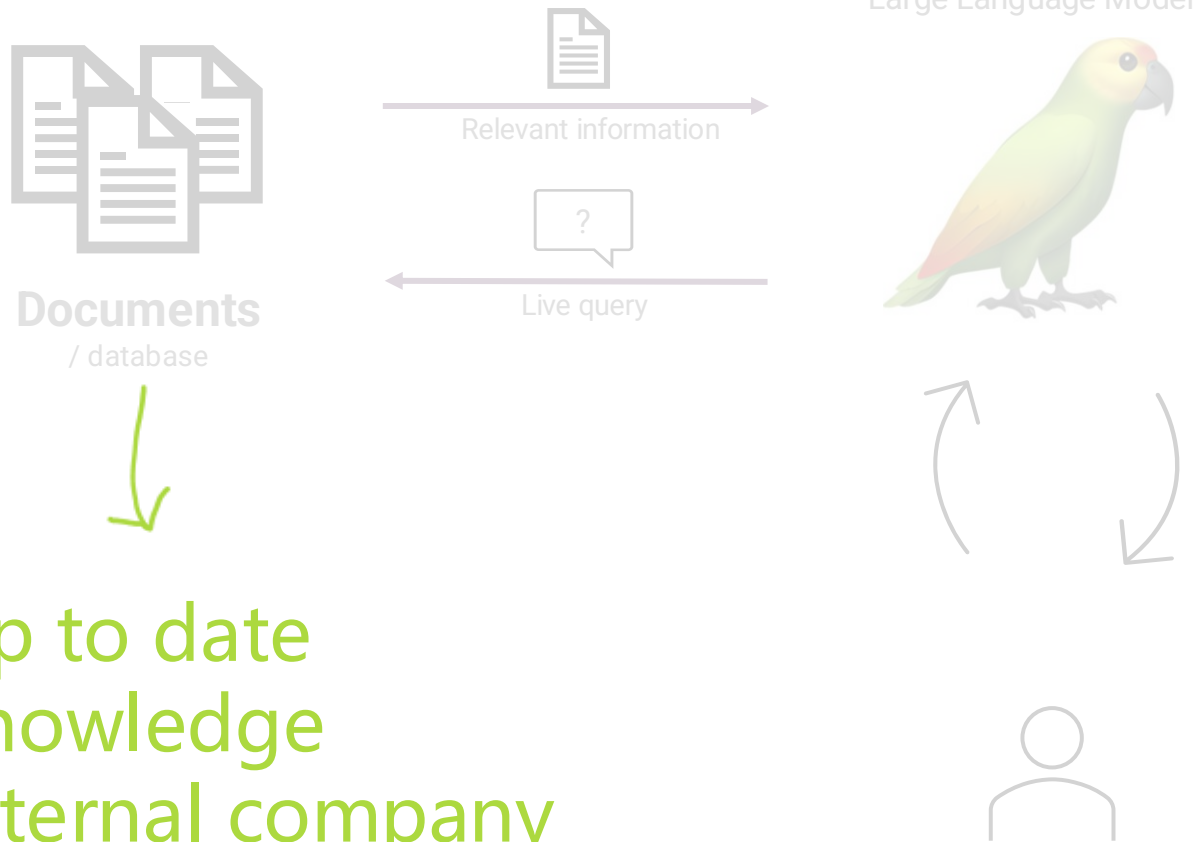
England beat Netherlands 2-1 in the
second semi-final of the UEFA European
Football Championships, Euro 2024.^[1]

[1] Produced with [gemini-1.0-pro-001](#), temperature = 0.0, top_p = 1.0,
grounded in [Google Search](#)



Why RAG ?

Model interactions
with **RAG**



- ✓ Up to date knowledge
- ✓ Internal company data
- ✓ Factual answers

RAG is useful,
so how does it
work?

Let's together understand RAG
and its **levels of difficulty**.

The Levels of **RAG**



Level 1
Basic RAG



Level 2
Hybrid search



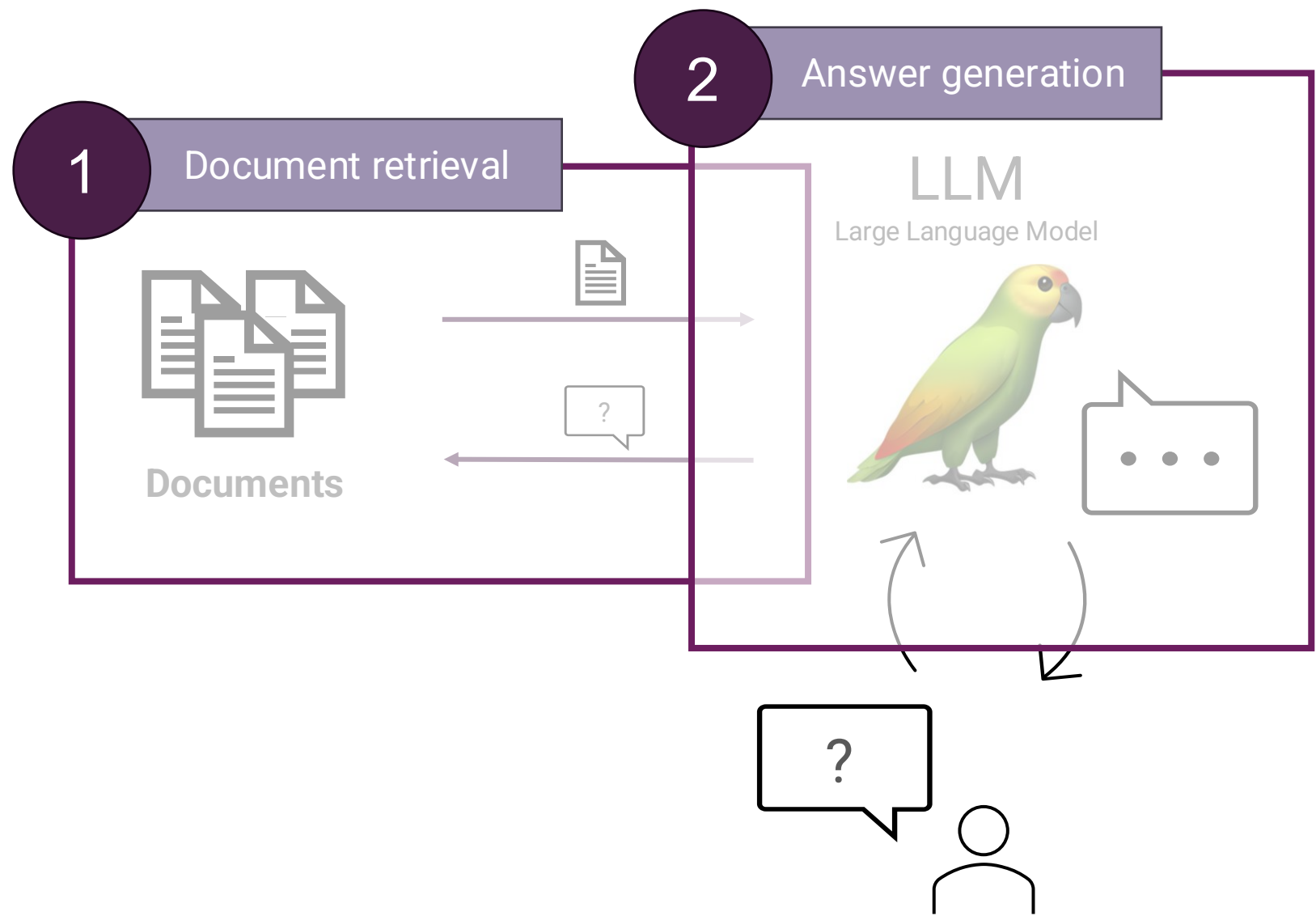
Level 3
Advanced data formats



Level 4
Multimodal

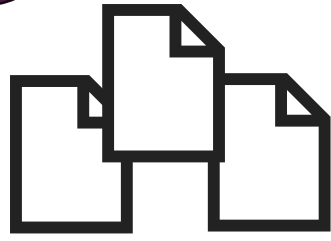
Level 1

Basic RAG



1

Document retrieval



Documents

Matching document texts and metadata



API

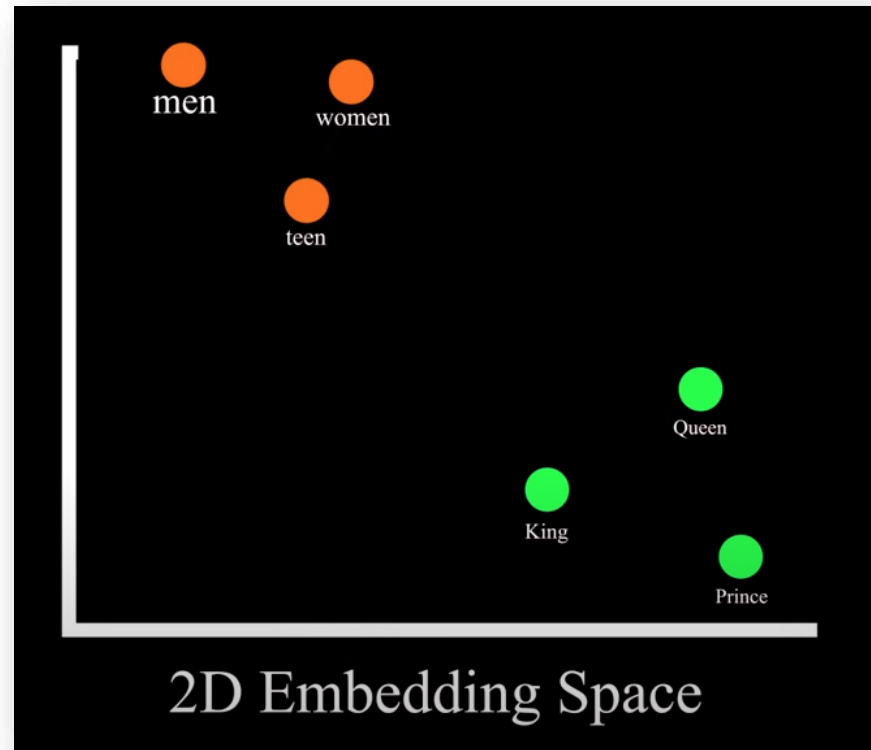
Question

?



Convert search
query to embedding

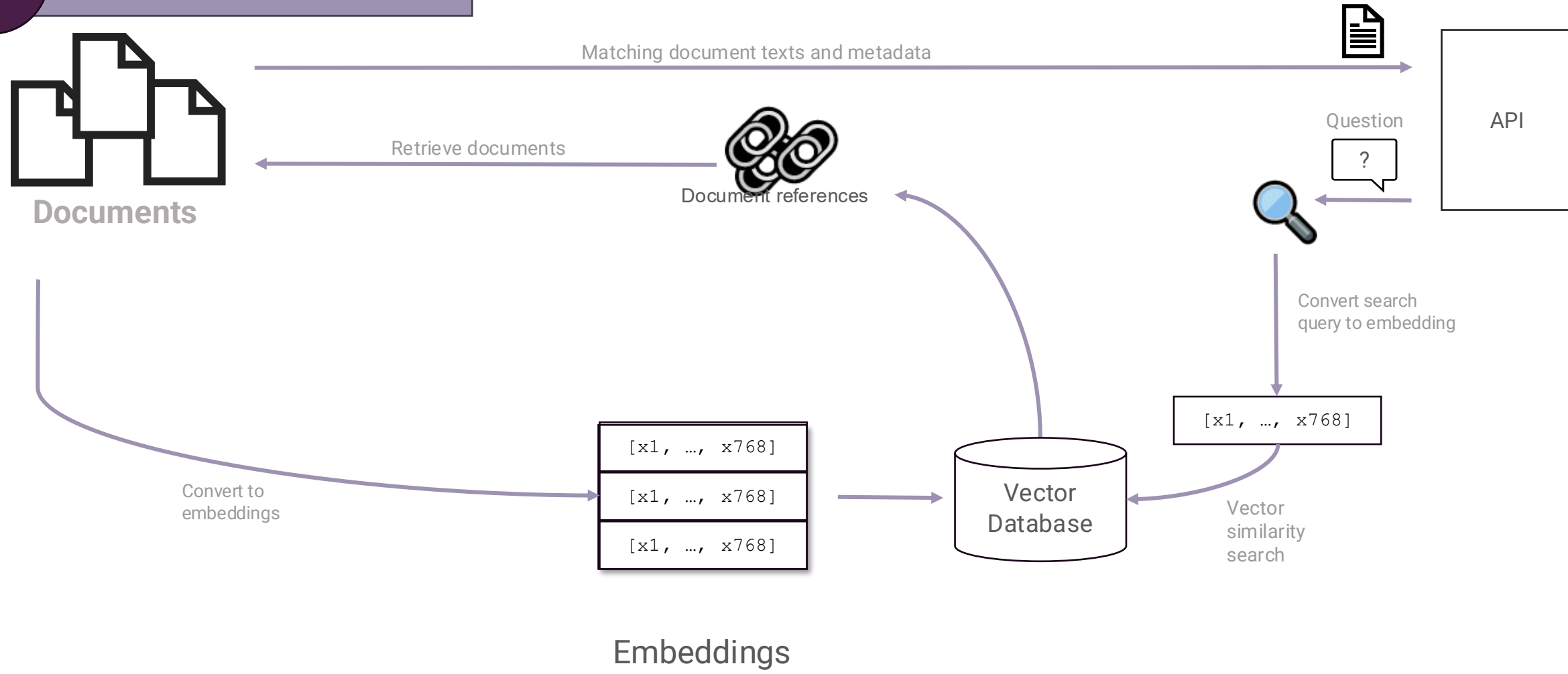
$[x_1, \dots, x_{768}]$



<https://images.app.goo.gl/canHYcAF5D2kjmmdA>

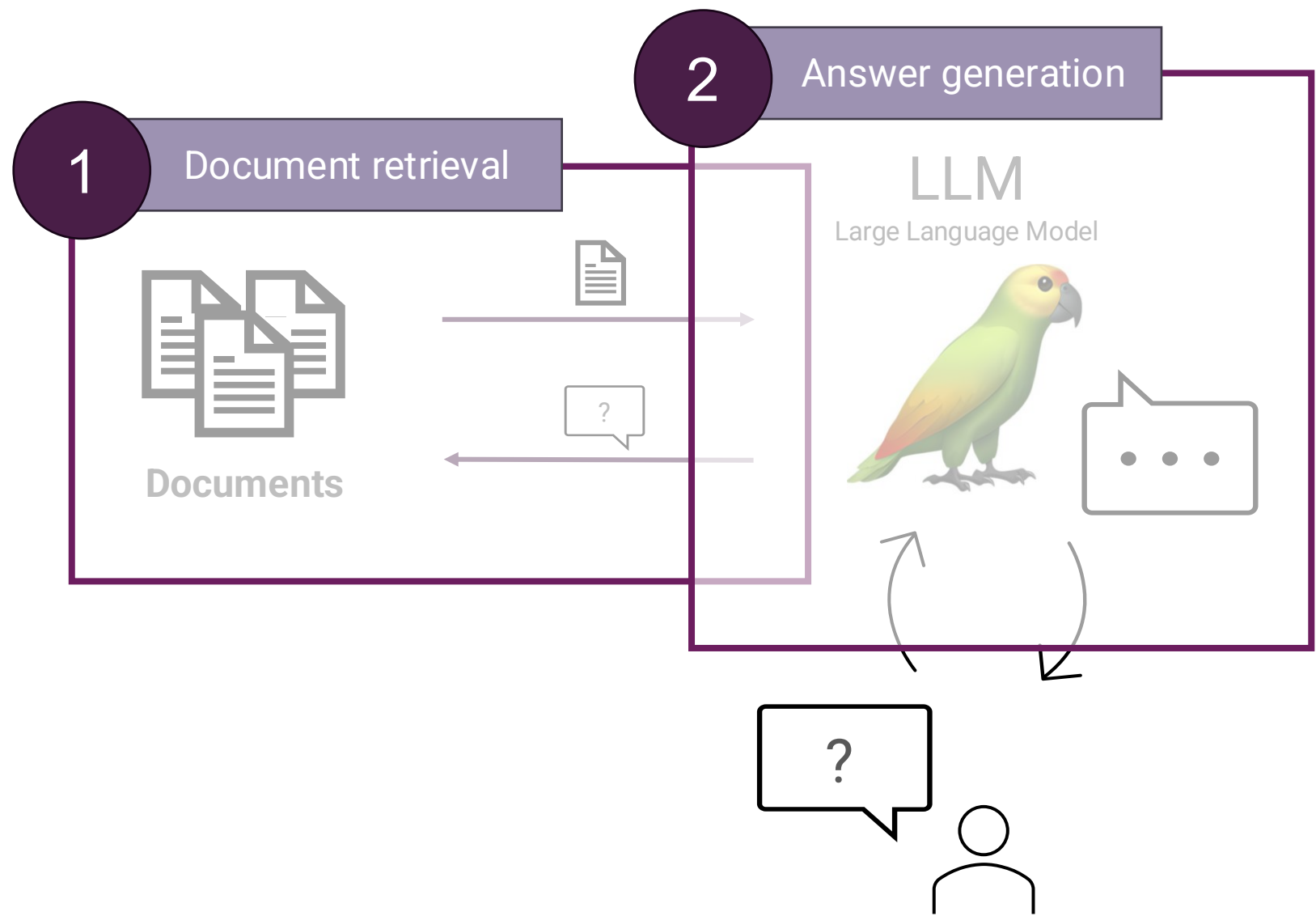
1

Document retrieval



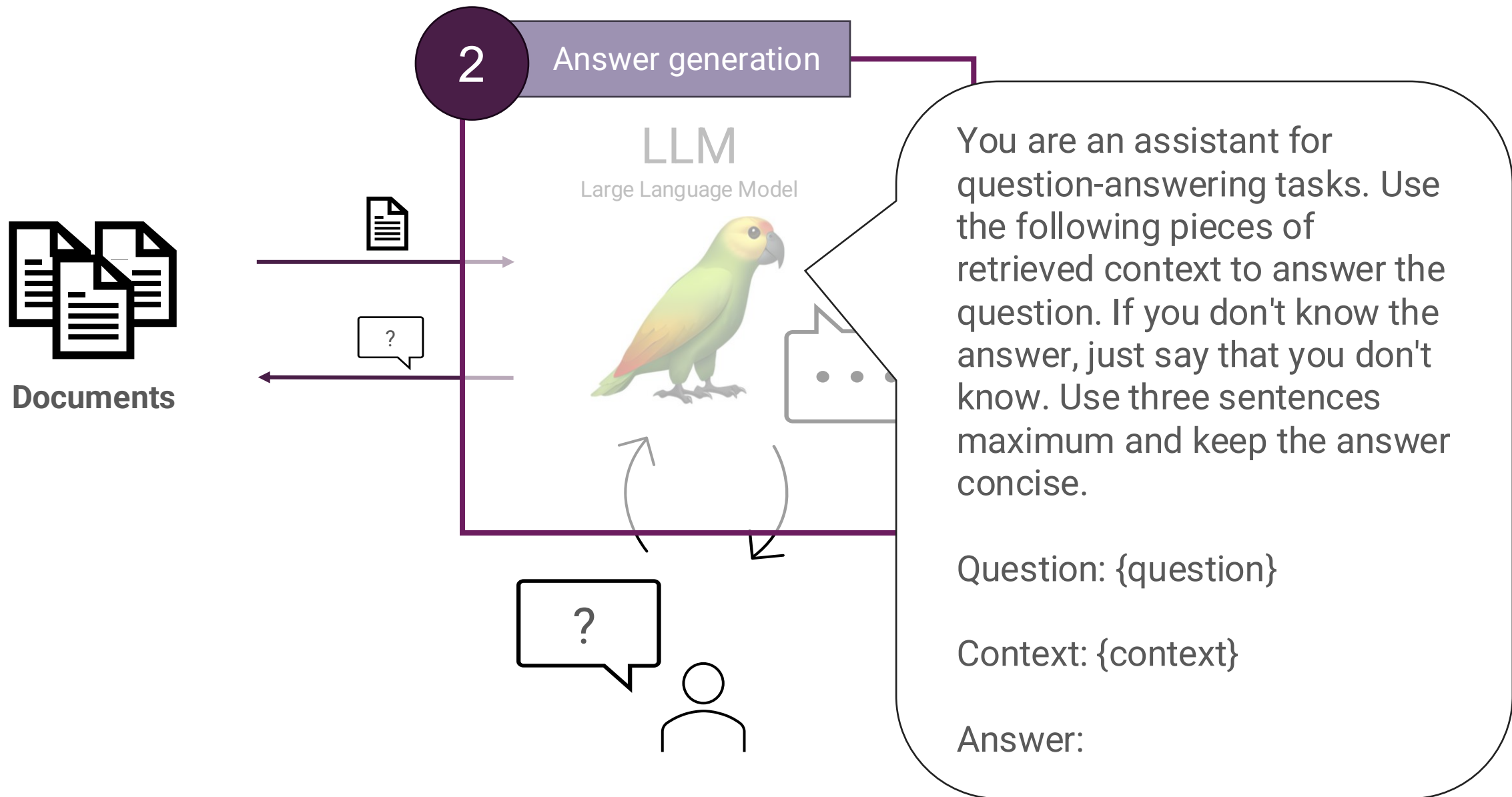
Level 1

Basic RAG



Level 1

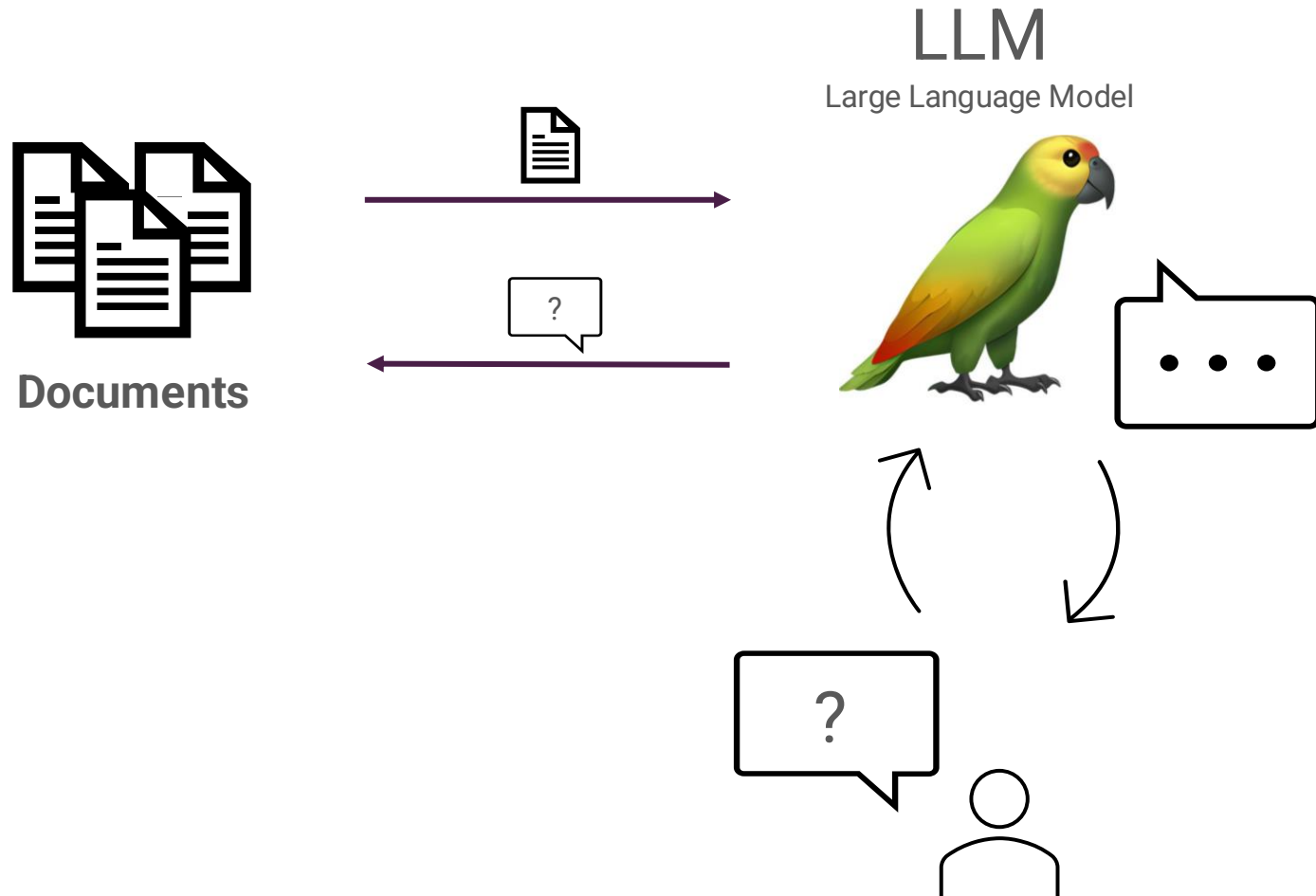
Basic RAG



Level 1

Basic RAG

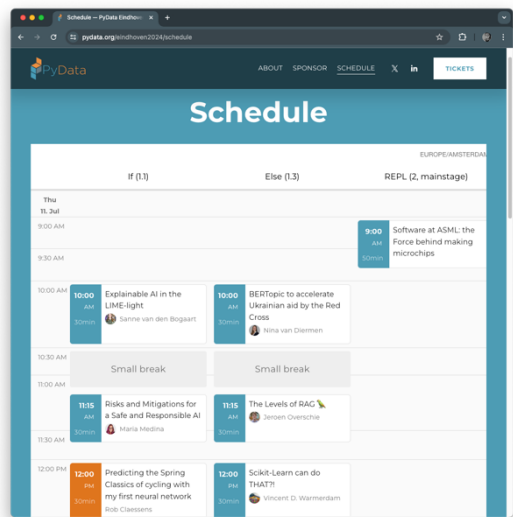
Let's give it a spin!



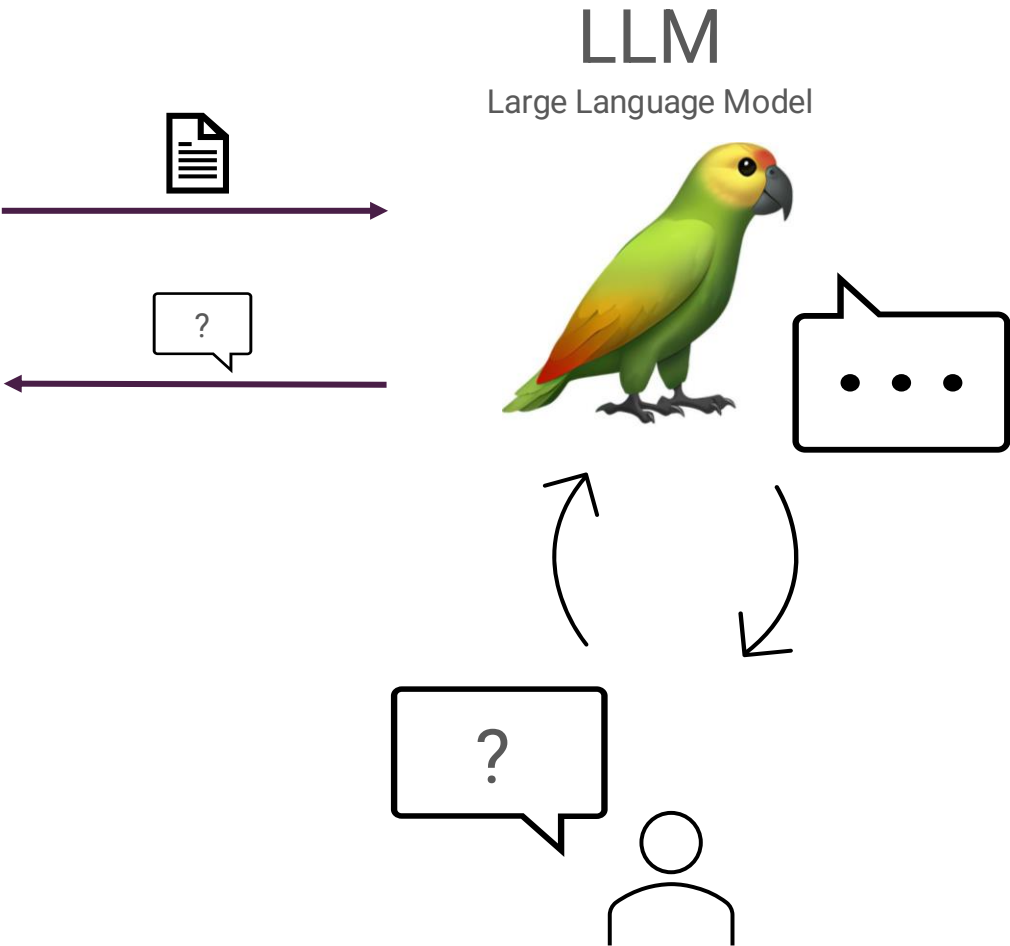
Level 1

Basic RAG

Let's give it a spin!

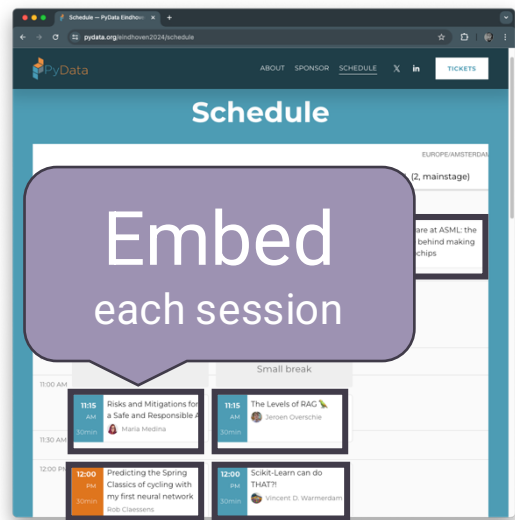


PyData Eindhoven 2024
Schedule



Level 1
Basic RAG

Let's give it a spin!



PyData Eindhoven 2024
Schedule

Enhancing ...
Computer vision ...
Predicting the ...
How I lost 1000€ ...



LLM
Large Language Model



You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.

Question: Which talks are related to sports?

Context:

Enhancing Event Analysis at Scale: Leveraging Tracking Data in Sports.
Learn how to automate the generation of contextual metrics from tracking data to enrich event analysis, handling the influx of games arriving daily in an efficient way by scaling-out the entire architecture.

Description

In the dynamic landscape of sports analytics, the integration of tracking data has opened new frontiers for in-depth event analysis. Yet, the use of this data remains a bottleneck, particularly when dealing with a large volume of games. Indeed, such computation is either too expensive or too long. The focus of the presentation will be on automating the generation of these contextual metrics at scale, and their usage by professionals and decision-makers.

The presentation will showcase an architecture and an automated pipeline designed to handle the influx of games. Leveraging Python and cloud computing services such as message queues, we efficiently manage incoming game data by scaling the infrastructure based on the workload, ensuring optimal performance during peak period while minimizing costs during quieter times. The presentation will strike a balance between technical depth and practical application. Attendees will gain insights into the architecture required to efficiently process hundreds of games weekly, while accommodating the thousands already present in the database. The advantage granted by this method will be quantified in terms of time and resources to inform data scientists and data engineers the efficiency they could reach.

Timeslot

2024-07-11T14:45:00+02:00 with a duration of 00:30

Room

If (1.1)

Speaker

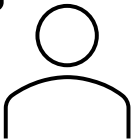
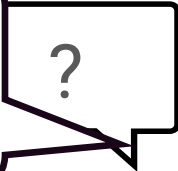
Yannis MOUDERE

I am a French Data Scientist, holding an engineering diploma from Telecom Paris and a Master's degree from Institut Polytechnique de Paris in Applied Mathematics and Data Science.
At the end of my studies, I completed a Data Science internship at Parma Calcio 1913. I now serve as a full-time Data Scientist at the club, working on leveraging tracking data.

Answer:

Prompt

Which talks are related to sports?
return a bullet list



Answer

Here's a list of talks related to sports:

- **Enhancing Event Analysis at Scale: Leveraging Tracking Data in Sports.**
- **Computer vision at the Dutch Tennis Federation: Utilizing YOLO to create insights for coaches**
- **Predicting the Spring Classics of cycling with my first neural network**
- **How I lost 1000€ betting on CS:GO with machine learning and Python**

Level 1

Basic RAG

Let's give it a spin!

Schedule

Embed each session

PyData Eindhoven 2024 Schedule

Which talks are return a bullet list

12:00 PM	12:00 PM 30min	Predicting the Spring Classics of cycling with my first neural network Rob Claessens PySport - Sports Analytics
12:30 PM	Lunch	
2:00 PM	2:00 PM 30min	Computer vision at the Dutch Tennis Federation: Utilizing YOLO to create insights for coaches Max Brouwer, Matthias Kempe PySport - Sports Analytics
2:30 PM	2:45 PM 30min	Enhancing Event Analysis at Scale: Leveraging Tracking Data in Sports. Yannis MOUDERE PySport - Sports Analytics
3:00 PM		

PyData Eindhoven 2024

Home About Sponsor Schedule Tickets Jeroen Overschie

How I lost 1000€ betting on CS:GO with machine learning and Python

07-11, 16:05–16:35 (Europe/Amsterdam), If (1.1)

People have been using machine learning for **sports** betting for decades. Logistic regression applied to horse racing made someone a multi-millionaire in the 80s. While fun, betting is a losing proposition for most. The house always wins, right?

With a friend, I thought we could beat the house in e-**sports** by leveraging modern ML tools like LightGBM. E-**sports** betting is less sophisticated than football or horse racing i.e. the market is less efficient. There is a lot of online data and unknown teams. It was a space ripe for money-making, or so we thought.

First, I will explain the theory behind e-**sports** betting with ML: what is an edge, financial decision-making, the expected value and decision rule for one bet, multiple bets with the Kelly criterion, probability calibration and the winner's curse.

Then, I will explain how we built a web scraper to extract features, developed a probabilistic classifier using LightGBM, defined betting rules using the Kelly criterion, backtested it with a positive ROI, and then lost actual money, with many priceless lessons coming out of it.

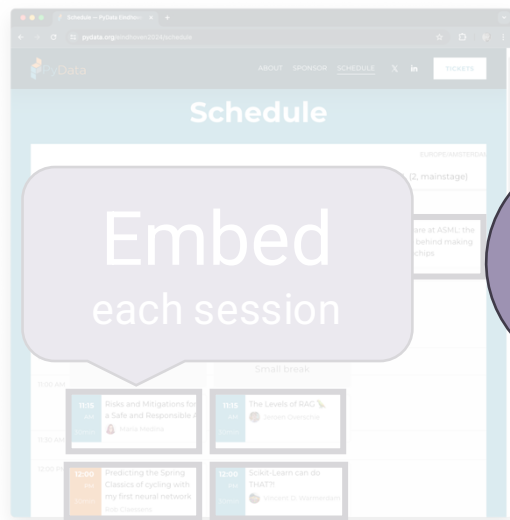
This presentation goes in-depth on how to use ML for e-**sports** betting and the pitfalls one might fall in. More broadly, I try to connect ML with financial decision-making, which can be applied in other domains too (credit, fraud, marketing), targeting data scientists and ML practitioners who are interested in financial applications.

Answer

Here's a list of talks related to sports:

- Enhancing Event Analysis at Scale: Leveraging Tracking Data in Sports.
- Computer vision at the Dutch Tennis Federation: Utilizing YOLO to create insights for coaches
- Predicting the Spring Classics of cycling with my first neural network
- How I lost 1000€ betting on CS:GO with machine learning and Python

Let's give it a spin!



PyData Eindhoven 2024
Schedule

Prompt

Which talks are related to sports?
return a bullet list

How to nicely
visualise? 
Logging / monitoring?

Answer

Here's a list of talks related to sports:

- Enhancing Event Analysis at Scale: Leveraging Tracking Data in Sports.
- Computer vision at the Dutch Tennis Federation: Utilizing YOLO to create insights for coaches
- Predicting the Spring Classics of cycling with my first neural network
- How I lost 1000€ betting on CS:GO with machine learning and Python

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.

Question: Which talks are related to sports?

Context:

Enhancing Event Analysis at Scale: Leveraging Tracking Data in Sports.
How to automate the generation of contextual metrics from tracking data to enrich event analysis, handling the influx of games being daily in an efficient way by scaling-out the entire architecture.

In the landscape of sports analytics, the integration of tracking data has opened new frontiers for in-depth event analysis. However, data remains a bottleneck, particularly when dealing with a large volume of games. Indeed, such computation is often too long. The focus of the presentation will be on automating the generation of these contextual metrics at scale for professionals and decision-makers.

We showcase an architecture and an automated pipeline designed to handle the influx of games. Leveraging streaming services such as message queues, we efficiently manage incoming game data by scaling the processing workload, ensuring optimal performance during peak period while minimizing costs during quieter times. We strike a balance between technical depth and practical application. Attendees will gain insights into the efficient processing of hundreds of games weekly, while accommodating the thousands already present in the database. The results granted by this method will be quantified in terms of time and resources to inform data scientists and data engineers they could reach.

Duration: 00:30

He holds an engineering diploma from Telecom Paris and a Master's degree from Institut Polytechnique de Paris in Statistics and Data Science.

He completed a Data Science internship at Parma Calcio 1913. I now serve as a full-time Data Scientist at the company specializing in tracking data.

Level 1

Basic RAG

LLM monitoring platforms:



Phoenix



Langfuse



LangSmith

levels-of-rag - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63?timeModel=%7B"duration"%3A"7d"%7D

Stats
Last 7 days
RUN COUNT
151
TOTAL TOKENS
98,350 / \$0.00
MEDIAN TOKENS
13
ERROR RATE
2%
% STREAMING
0%
LATENCY
P50: 1.28s P99: 7.04s

Personal > Projects > levels-of-rag

DEVELOPER

levels-of-rag

ID Data Retention 14d Add Rule Edit

Runs Threads Monitor Setup

1 filter Last 7 days Root Runs LLM Calls All Runs Columns

		✓	Name	Input	Output	Start Time	Latency	Dataset	Annotation Queue	Tokens
	>	✓	RunnableSequence	which talks are related...	Here is a list of tal...	02/07/2024, 21:33:13	4.38s			2,914
	>	✓	RunnableSequence	which talks are related...	Here are the talks...	02/07/2024, 21:33:02	6.52s			2,993
	>	✓	RunnableSequence	which talks are related...	Here's a list of tal...	02/07/2024, 21:31:25	5.64s			2,922
	>	✓	RunnableSequence	which talks are related...	The following talk...	02/07/2024, 21:29:33	3.53s			2,917
	>	✓	RunnableSequence	which talks are related...	Here are the talks...	02/07/2024, 21:29:19	6.07s			2,978
		✓	Retriever	which talks are related...	[{"page_content": "...	02/07/2024, 21:27:34	0.29s			0
		✓	Retriever	which talks are related...	[{"page_content": "...	02/07/2024, 21:27:06	0.25s			0
		✓	ChatGoogleGenerativeA	human: hello!	ai: Hello! How can ...	02/07/2024, 21:26:00	1.69s			13
	>	✓	RunnableSequence	which talks are related...	The following talk...	02/07/2024, 21:25:23	3.94s			2,901
	>	✓	RunnableSequence	which talks are related...	The following talk...	02/07/2024, 21:24:09	7.57s			2,930
		✓	ChatGoogleGenerativeA	human: what is the RA...	ai: Please provide ...	02/07/2024, 21:24:02	5.51s			158
		✓	Retriever	which talks are related...	[{"page_content": "...	02/07/2024, 21:23:59	0.28s			0
		✓	Retriever	which talks are related...	[{"page_content": "...	02/07/2024, 21:23:52	0.28s			0
		✓	ChatGoogleGenerativeA	human: hello!	ai: Hello! How can ...	02/07/2024, 21:23:38	1.87s			13
		✓	ChatGoogleGenerativeA	human: what talks are ...	ai: Unfortunately, I...	02/07/2024, 21:22:25	4.17s			130
		✓	ChatGoogleGenerativeA	human: hello!	ai: Hello! What can...	02/07/2024, 21:22:05	1.20s			14

Filter Shortcuts

Metadata

☐ ls_model_name

☐ ls_model_name == "gpt-3.5-..."

☐ ls_model_name == "gpt-35-t..."

☐ ls_model_name == "models/..."

☐ ls_model_name == "models/..."

☐ ls_model_type

☐ ls_model_type == "chat"

☐ ls_provider

☐ ls_provider == "google_genai"

☐ ls_provider == "openai"

☐ ls_temperature

☐ ls_temperature == 0

☐ ls_temperature == 0.7

levels-of-rag - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63?timeModel=%7B"duration"%3A"7d"%7D&peek=584da75b-d945-4062...

Personal > Projects > levels-of-rag

levels-of-rag

Runs Threads Monitor Setup

1 filter Last 7 days Root Runs LLM Calls All Runs

	>	✓	Name	Input
	>	✓	RunnableSequence	which talks are
	>	✓	RunnableSequence	which talks are
	>	✓	RunnableSequence	which talks are
	>	✓	RunnableSequence	which talks are
	>	✓	RunnableSequence	which talks are
	>	✓	RunnableSequence	which talks are
	>	✓	Retriever	which talks are
	>	✓	Retriever	which talks are
	>	✓	ChatGoogleGenerativeA	human: hello!
	>	✓	RunnableSequence	which talks are
	>	✓	RunnableSequence	which talks are
	>	✓	ChatGoogleGenerativeA	human: what is
	>	✓	Retriever	which talks are
	>	✓	Retriever	which talks are
	>	✓	ChatGoogleGenerativeA	human: hello!
	>	✓	ChatGoogleGenerativeA	human: what ta
	>	✓	ChatGoogleGenerativeA	human: hello!

TRACE

Collapse Stats Most relevant

RunnableSequence 4.38s 2,914

map:key:context 0.50s

Retriever 0.49s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGe... models/gemi... 3.88s

Some runs have been hidden. [Show 3 hidden runs](#)

RunnableSequence

Run Feedback Metadata

Input

1 input: which talks are related to sports? return a bullet list

YAML

Output

Here is a list of talks related to sports:

- Enhancing Event Analysis at Scale: Leveraging Tracking Data in Sports.
- Computer vision at the Dutch Tennis Federation: Utilizing YOLO to create insights for coaches.
- Predicting the Spring Classics of cycling with my first neural network.
- How I lost 1000€ betting on CS:GO with machine learning and Python.

START TIME07/02/2024, 09:33:13 PM

END TIME07/02/2024, 09:33:17 PM

TIME TO FIRST TOKENN/A

STATUSSuccess

TOTAL TOKENS2,914 tokens

LATENCY4.38s

TYPESequence

levels-of-rag - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63?timeModel=%7B"duration"%3A"7d"%7D&peek=288a980d-a442-4080...

levels-of-rag

Runs Threads Monitor Setup

1 filter Last 7 days Root Runs LLM Calls All Runs

	Name	Input
	RunnableSequence	which talks are
	RunnableSequence	which talks are
	RunnableSequence	which talks are
	RunnableSequence	which talks are
	RunnableSequence	which talks are
	RunnableSequence	which talks are
	Retriever	which talks are
	Retriever	which talks are
	ChatGoogleGenerativeA	human: hello!
	RunnableSequence	which talks are
	RunnableSequence	which talks are
	ChatGoogleGenerativeA	human: what is
	Retriever	which talks are
	Retriever	which talks are
	ChatGoogleGenerativeA	human: hello!
	ChatGoogleGenerativeA	human: what ta
	ChatGoogleGenerativeA	human: hello!

TRACE

Collapse Stats Most relevant

RunnableSequence 4.38s 2,914

map:key:context 0.50s

Retriever 0.49s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGe... models/gemi... 3.88s

Some runs have been hidden. [Show 3 hidden runs](#)

Retriever

Run Feedback Metadata

Input

1 query: which talks are related to sports? return a bullet list

YAML

Output

DOCUMENTS 4

Enhancing Event Analysis at Scale: Leveraging Trac... +20

Computer vision at the Dutch Tennis Federation: Ut... +20

Predicting the Spring Classics of cycling with my fir... +20

How I lost 1000€ betting on CS:GO with machine lea... +20

START TIME 07/02/2024, 09:33:13 PM

END TIME 07/02/2024, 09:33:13 PM

TIME TO FIRST TOKEN N/A

STATUS Success

TOTAL TOKENS 0 tokens

LATENCY 0.49s

TYPE Retriever

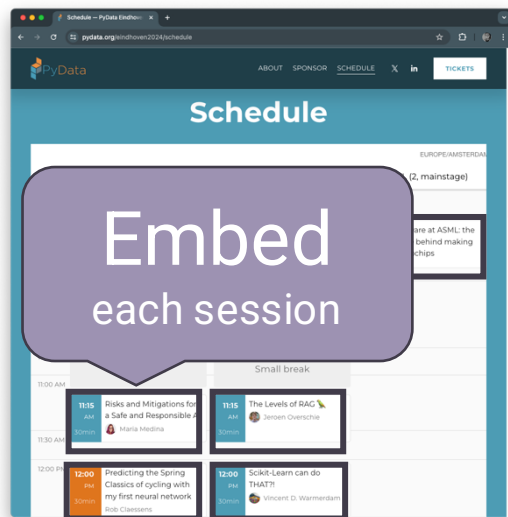
TAGS GoogleGenerativeAIEmbeddings

seq:step:1 FAISS

Level 1

Basic RAG

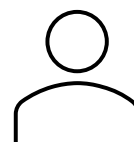
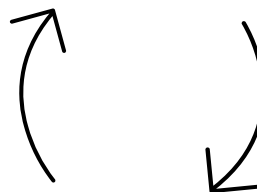
That was nice. But can we get more **detailed**?



PyData Eindhoven 2024
Schedule



LLM
Large Language Model



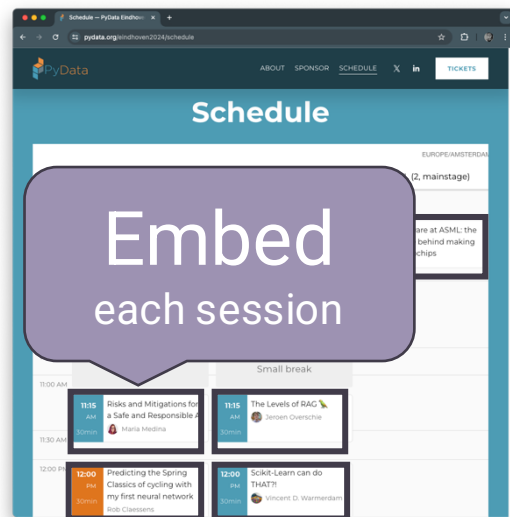
- ✗ Embedding a **large** piece of text cause embeddings to get *saturated* and lose meaning

- ✗ Imprecise citations
- ✗ Large context → high cost

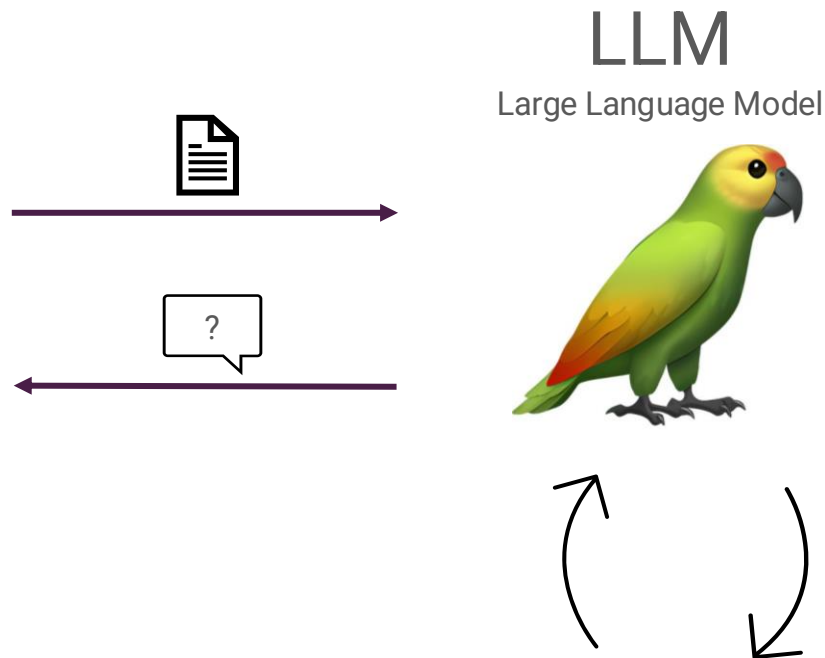
Level 1

Basic RAG

That was nice. But can we get more **detailed**?



PyData Eindhoven 2024
Schedule



✗ Embedding a **large** piece of text cause embeddings to get *saturated* and lose meaning

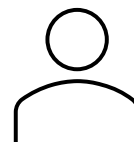
✗ Imprecise citations

✗ Large context → high cost



Divide up text in smaller pieces, then embed.

= *Chunking*



Maximizing marketplace experimentation:
switchback design for small samples and
subtle effects

Conventional A/B testing often falls short
in industries such as airlines, ride-
sharing, and delivery services, where
challenges like small samples and subtle
effects complicate testing new features.

Maximizing marketplace experimentation:
switchback design for small samples and
subtle effects ¶

¶

Conventional A/B testing often falls short
in industries such as airlines, ride-
sharing, and delivery services, where
challenges like small samples and subtle
effects complicate testing new features. ¶

Character Text Splitter  
with
Chunk size: 25
Overlap: 0

Maximizing marketplace experimentation:
switchback design for small samples and
subtle effects ¶

¶

Conventional A/B testing often falls short
in industries such as airlines, ride-
sharing, and delivery services, where
challenges like small samples and subtle
effects complicate testing new features. ¶

Recursive Character Text Splitter 🦜🔗
with
Chunk size: 25
Overlap: 0

Maximizing marketplace experimentation:



LangChain

[Integrations](#)

[API Reference](#)

[More](#)

v0.2



[Introduction](#)

[Tutorials](#)

[How-to guides](#)

[Conceptual guide](#)

[Ecosystem](#)

[LangSmith](#)

[LangGraph](#)

[Versions](#)

[Overview](#)

[Release Policy](#)

Text splitters

Text Splitters take a document and split into chunks that can be used for retrieval.

- How to: recursively split text
- How to: split by HTML headers
- How to: split by HTML sections
- How to: split by character
- How to: split code
- How to: split Markdown by headers
- How to: recursively split JSON
- How to: split text into semantic chunks
- How to: split by tokens

Level 1

Basic RAG

level_0_basic_rag.ipynb X

```
vectorstore_chunked.similarity_search(  
    query="
```

a methodology for comparing two versions of a webpage or app against each other to determine which one performs better

"

```
)
```

OUTPUT

```
[
```

Document(page_content='and recognize its advantages over conventional A/B testing for more informed decisions and improved', ...),

Document(page_content='design as a practical alternative to conventional A/B testing. By addressing small sample size', ...),

Document(page_content='It addresses traditional A/B testing's limitations on small samples and subtle effects, highlighted', ...),

Document(page_content='Conventional A/B testing often falls short in industries such as airlines, ride-sharing, and', ...)

```
]
```

Citations

Maximizing marketplace experimentation: switchback design for small samples and subtle effects

.ical

07-11, 14:45–15:15 (Europe/Amsterdam), Else (1.3)

Conventional A/B testing often falls short in industries such as airlines, ride-sharing, and delivery services, where challenges like small samples and subtle effects complicate testing new features. Inspired by its significant impact in leading companies like Uber, Lyft, and Doordash, we introduce the switchback design as a practical alternative to conventional A/B testing. By addressing small sample size limitations and the need to detect subtle effects quickly, this approach boosts statistical power while reducing variability and interference. We guide the audience through the challenges of marketplace experimentation and implementing this approach, period length optimization and switch frequency using a case study from the airline industry.

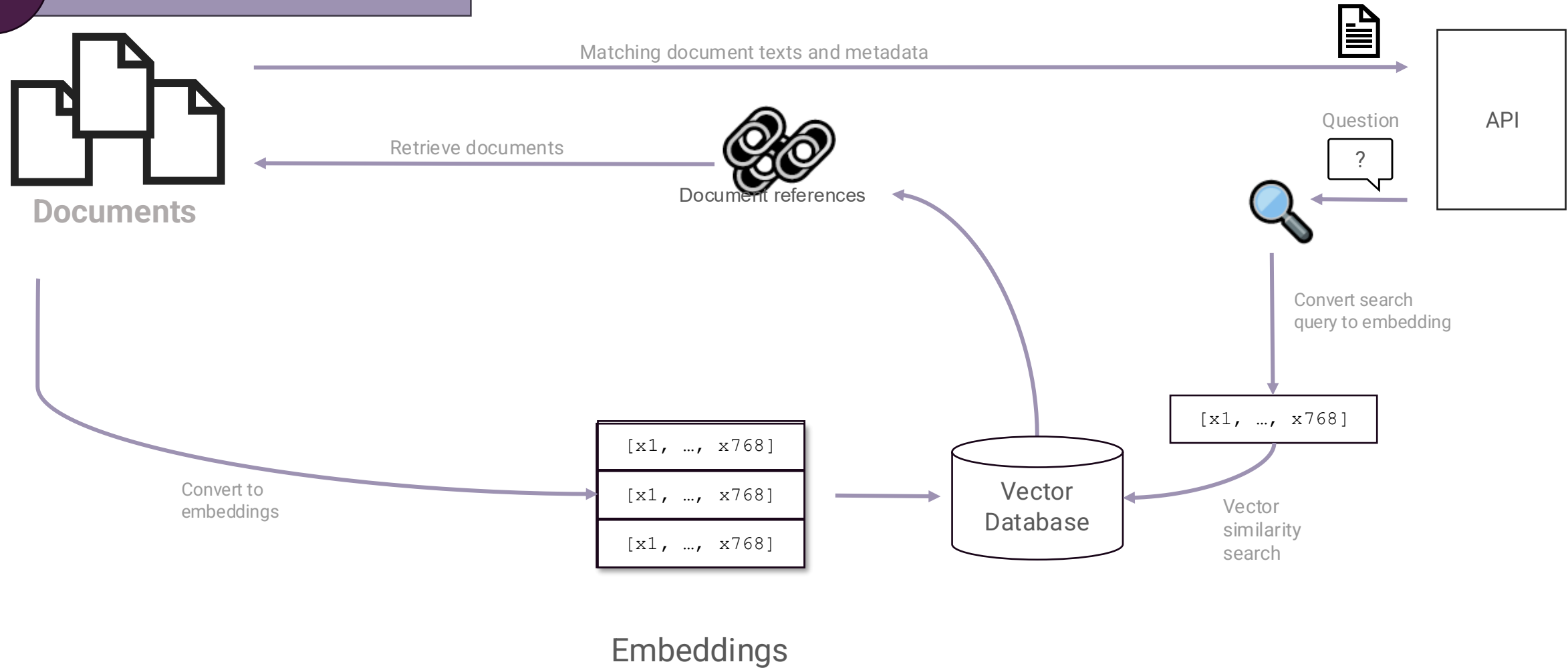
In this talk, we introduce switchback design, a method that addresses key challenges in marketplace experimentation faced by sectors like airlines and ride-sharing. It addresses traditional A/B testing's limitations on small samples and subtle effects, highlighted by successes in companies like Uber, Lyft, and Doordash. This approach, crucial for accurate effect measurement, aims to boost decision-making and operational optimization, supported by a practical case study.

Talk outline:

- Introduction (5 mins): Quick intro to our expertise and the topic's relevance.
- Challenges (5 mins): Overview of marketplace experimentation challenges.
- Switchback Design (5 mins): Basics of switchback design vs. traditional A/B testing.
- Case Study (10 mins): Real-world application in the airline industry, showcasing benefits and drawbacks.
- Key Takeaways (5 mins): Summary of our talk.
- Q&A (10 mins): Open discussion for clarifications and further insights.

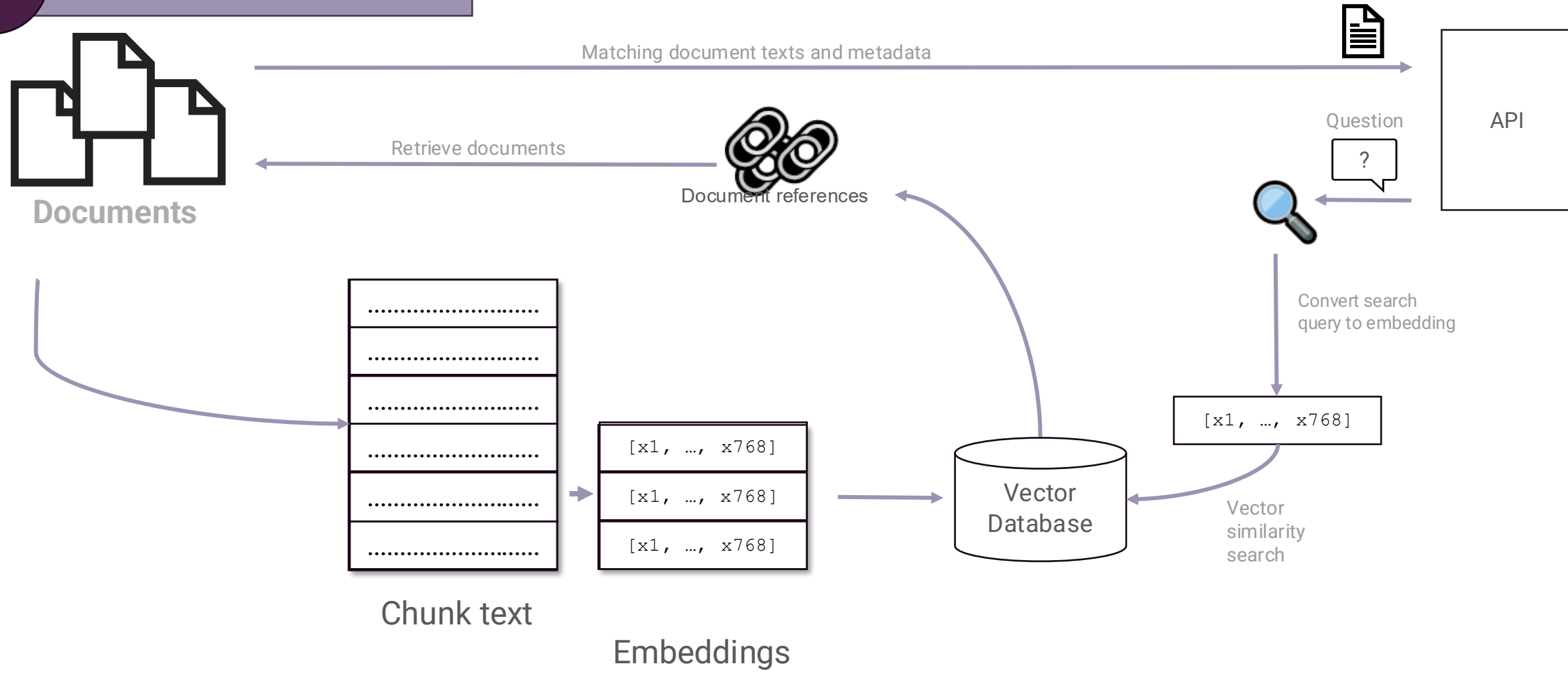
1

Document retrieval



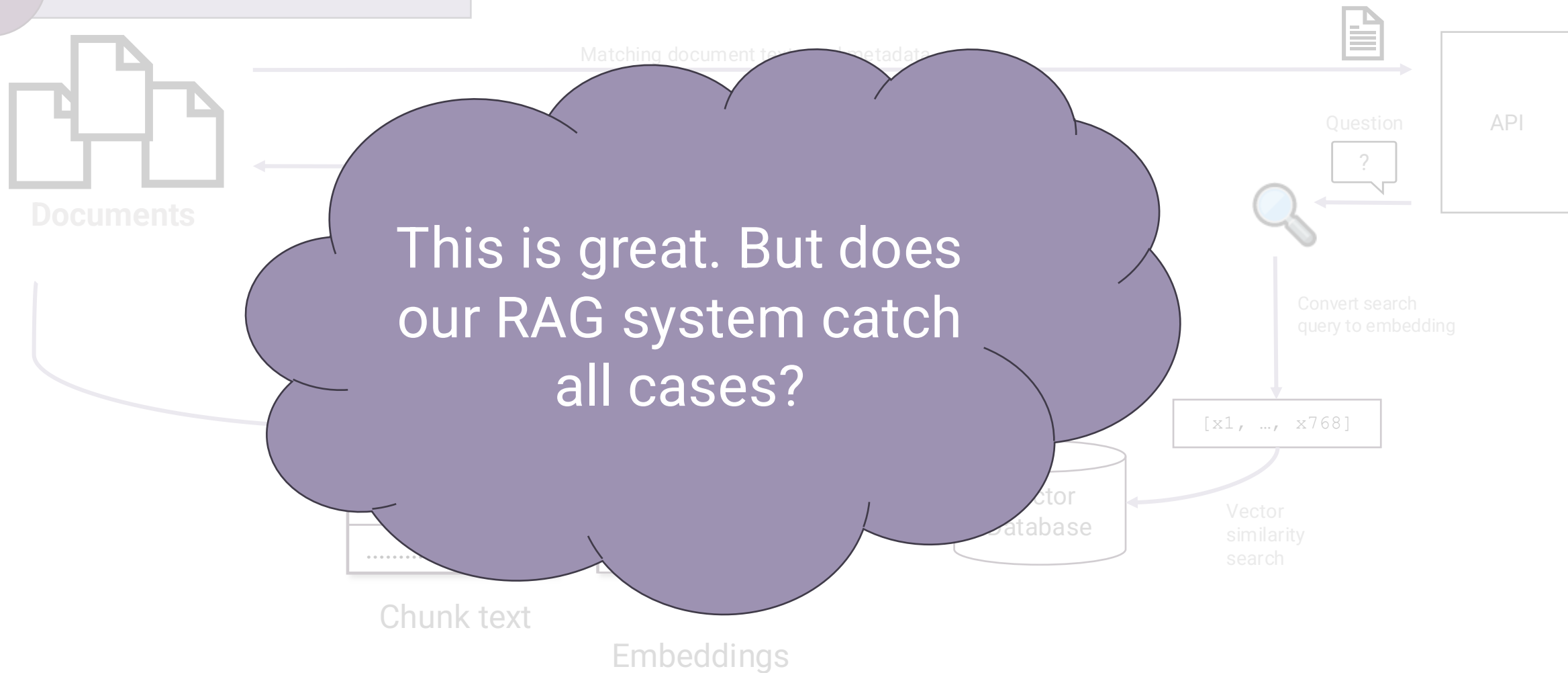
1

Document retrieval



1

Document retrieval



Level 1

Basic RAG

RunnableSequence - LangSmi

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/4e6a6c2b-4512-4d46-a48a-a15cbb9e...

Paused

Personal > Projects > levels-of-rag > RunnableSequence

PlaygroundAdd to DatasetAdd to Annotation QueueShareAnnotateDEVELOPER

TRACE

CollapseStatsMost relevant

RunnableSequence 4.59s 2,123

map:key:context 0.77s

Retriever 0.77s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGenerativeAI models/gemini-1.5-pro 3.82s

Some runs have been hidden. [Show 3 hidden runs](#)

RunnableSequence

RunFeedbackMetadata

Run IDTrace ID

Input

1input: What's the talk starting with Maximizing about?

YAML

Output

The talk starting with "Maximizing" is about less common but noteworthy features of scikit-learn and its ecosystem. It covers topics such as sparse datasets, larger-than-memory datasets, sample weight techniques, and more. The speaker aims to highlight these features and potentially live code some examples.

START TIME

07/02/2024, 10:35:28 PM

END TIME

07/02/2024, 10:35:33 PM

TIME TO FIRST TOKEN

N/A

STATUS

Success

TOTAL TOKENS

2,123 tokens

LATENCY

4.59s

TYPE

Sequence

Level 1

Basic RAG

RunnableSequence - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/4e6a6c2b-4512-4d46-a48a-a15cbb9e...

Paused

Input

1

input: What's the talk starting with Maximizing about?

YAML

Some runs have been hidden. [Show 3 hidden runs](#)

The talk starting with "maximizing" is about less common but noteworthy features of scikit-learn and its ecosystem. It covers topics such as sparse datasets, larger-than-memory datasets, sample weight techniques, and more. The speaker aims to highlight these features and potentially live code some examples.

Success

TOTAL TOKENS
2,123 tokens

LATENCY
4.59s

TYPE
Sequence

Level 1

Basic RAG

RunnableSequence - LangSmi x +

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/4e6a6c2b-4512-4d46-a48a-a15cbb9e... ☆ ⌵ Paused ⋮

Input ▾

1

input: What's the talk starting with Maximizing about?

YAML ⌵

Some runs have been hidden. [Show 3 hidden runs](#)

The talk starting with "Maximizing" is about less common but noteworthy features of scikit-learn and its ecosystem. It covers topics such as sparse datasets, larger-than-memory datasets, sample weight techniques, and more. The speaker aims to highlight these features and potentially live code some examples.

Success

TOTAL TOKENS
2,123 tokens

LATENCY
⌚ 4.59s

Output ▾

The talk starting with "Maximizing" is about less common but noteworthy features of scikit-learn and its ecosystem. It covers topics such as sparse datasets, larger-than-memory da more. The speaker aims to highlight some examples.

2:45 PM 30min

Maximizing marketplace experimentation: switchback design for small samples and subtle effects

 Nazli M. Alagoz, Joël Gastelaars

12:00 PM 30min

Scikit-Learn can do THAT?!

 Vincent D. Warmerdam

Level 1 Basic RAG

Retriever - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/fadb1c03-ef6c-46db-9547-3da7dc0a7...

Personal > Projects > levels-of-rag > RunnableSequence > ... > Retriever

Playground Add to Dataset Share Annotate DEVELOPER

TRACE

Collapse Stats Most relevant

RunnableSequence 4.59s 2,123

map:key:context 0.77s

Retriever 0.77s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGenerativeAI models/gemini-1.5-pro 3.82s

Some runs have been hidden. [Show 3 hidden runs](#)

Retriever

Run Feedback Metadata

Input

1 query: What's the talk starting with Maximizing about?

YAML

Output

DOCUMENTS 4

- # The Levels of RAG LLM's can be supercharged using a ... +20
- # Scikit-Learn can do THAT?! Many of us know scikit-learn f... +20
- # Evaluating LLM Frameworks Large Language Models are ... +20
- # Software at ASML: the Force behind making microchips l... +20

START TIME 07/02/2024, 10:35:28 PM

END TIME 07/02/2024, 10:35:29 PM

TIME TO FIRST TOKEN N/A

STATUS Success

TOTAL TOKENS 0 tokens

LATENCY 0.77s

TYPE Retriever

TAGS GoogleGenerativeAIEmbeddings seq:step1 FAISS

How can we fix this?

The Levels of RAG



Level 1
Basic RAG



Level 2
Hybrid search



Level 3
Advanced data formats



Level 4
Multimodal

Level 2

Hybrid search

Retriever - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/fadb1c03-ef6c-46db-9547-3da7dc0a7...

Paused

Personal > Projects > levels-of-rag > RunnableSequence > ... > Retriever

Playground Add to Dataset Share Annotate DEVELOPER

TRACE

Collapse Stats Most relevant

RunnableSequence 4.59s 2,123

map:key:context 0.77s

Retriever 0.77s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGenerativeAI models/gemini-1.5-pro 3.82s

Some runs have been hidden. Show 3 hidden runs

Retriever

Run Feedback Metadata

Input

1 query: What's the talk starting with Maximizing about?

YAML

Output

DOCUMENTS 4

The Levels of RAG LLM's can be supercharged using a ... +20

Scikit-Learn can do THAT?! Many of us know scikit-learn f... +20

Evaluating LLM Frameworks Large Language Models are ... +20

Software ASML: the Force behind making microchips l... +20

START TIME 07/02/2024, 10:35:28 PM

END TIME 07/02/2024, 10:35:29 PM

TIME TO FIRST TOKEN N/A

STATUS Success

TOTAL TOKENS 0 tokens

LATENCY 0.77s

TYPE Retriever

TAGS GoogleGenerativeAIEmbeddings seq:step:1 FAISS

2:45 PM 30min

Maximizing marketplace experimentation: switchback design for small samples and subtle effects

Nazli M. Alagoz, Joël Gastelaars

Level 2 Hybrid search

Keyword search
algorithms

tf-idf
term
frequency-
inverse
document
frequency

BM25
Okapi BM25

Retriever - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/fadb1c03-ef6c-46db-9547-3da7dc0a7...

Paused

Personal > Projects > levels-of-rag > RunnableSequence > ... > Retriever

Playground Add to Dataset Share Annotate DEVELOPER

TRACE

Collapse Stats Most relevant

RunnableSequence 4.59s 2,123

map:key:context 0.77s

Retriever 0.77s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGenerativeAI models/gemini-1.5-pro 3.82s

Some runs have been hidden. Show 3 hidden runs

Retriever

Run Feedback Metadata

Input

1 query: What's the talk starting with Maximizing about?

YAML

Output

DOCUMENTS 4

The Levels of RAG LLM's can be supercharged using a ... +20

Scikit-Learn can do THAT?! Many of us know scikit-learn f... +20

Evaluating LLM Frameworks Large Language Models are ... +20

Software as a Service (SaaS): the Force behind making microchips I... +20

START TIME 07/02/2024, 10:35:28 PM

END TIME 07/02/2024, 10:35:29 PM

TIME TO FIRST TOKEN N/A

STATUS Success

TOTAL TOKENS 0 tokens

LATENCY 0.77s

TYPE Retriever

TAGS GoogleGenerativeAIEmbeddings seq:step:1 FAISS

2:45 PM

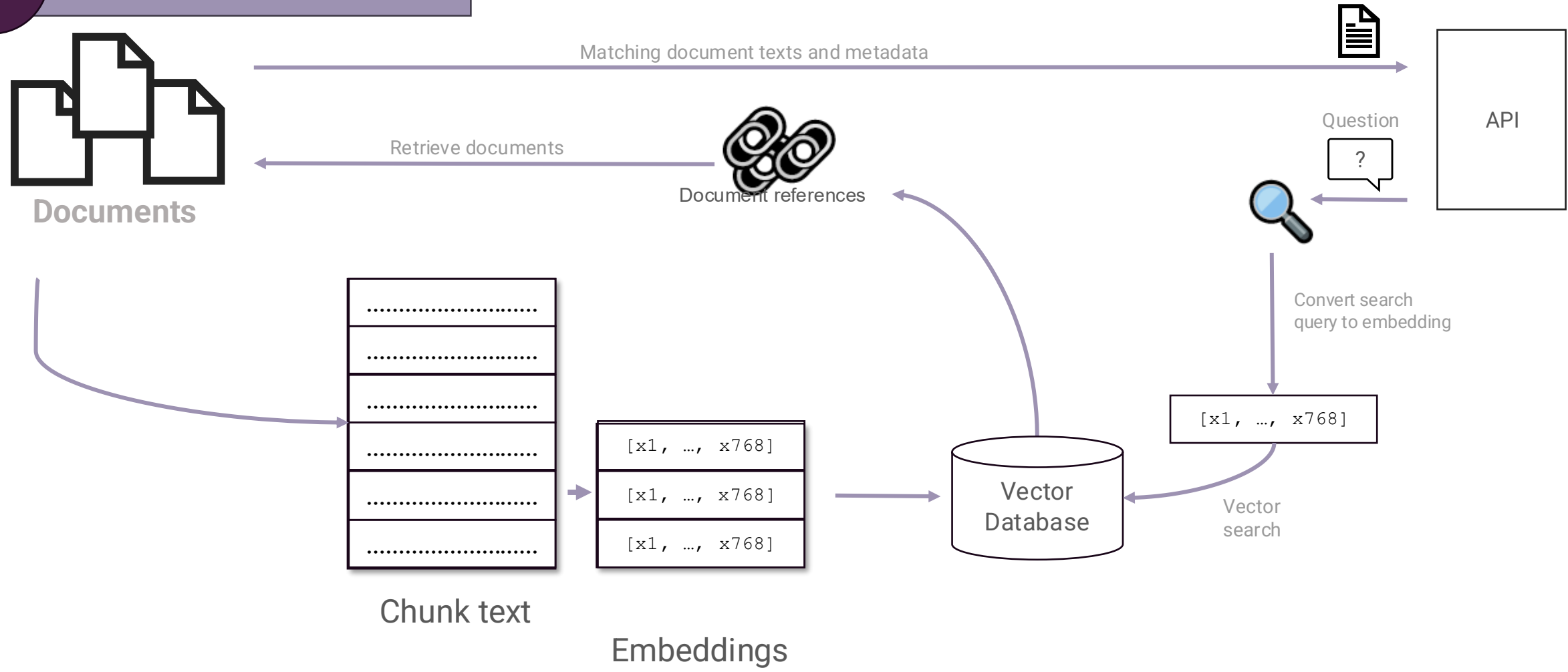
30min

Maximizing marketplace experimentation: switchback design for small samples and subtle effects

 Nazli M. Alagoz, Joël Gastelaars

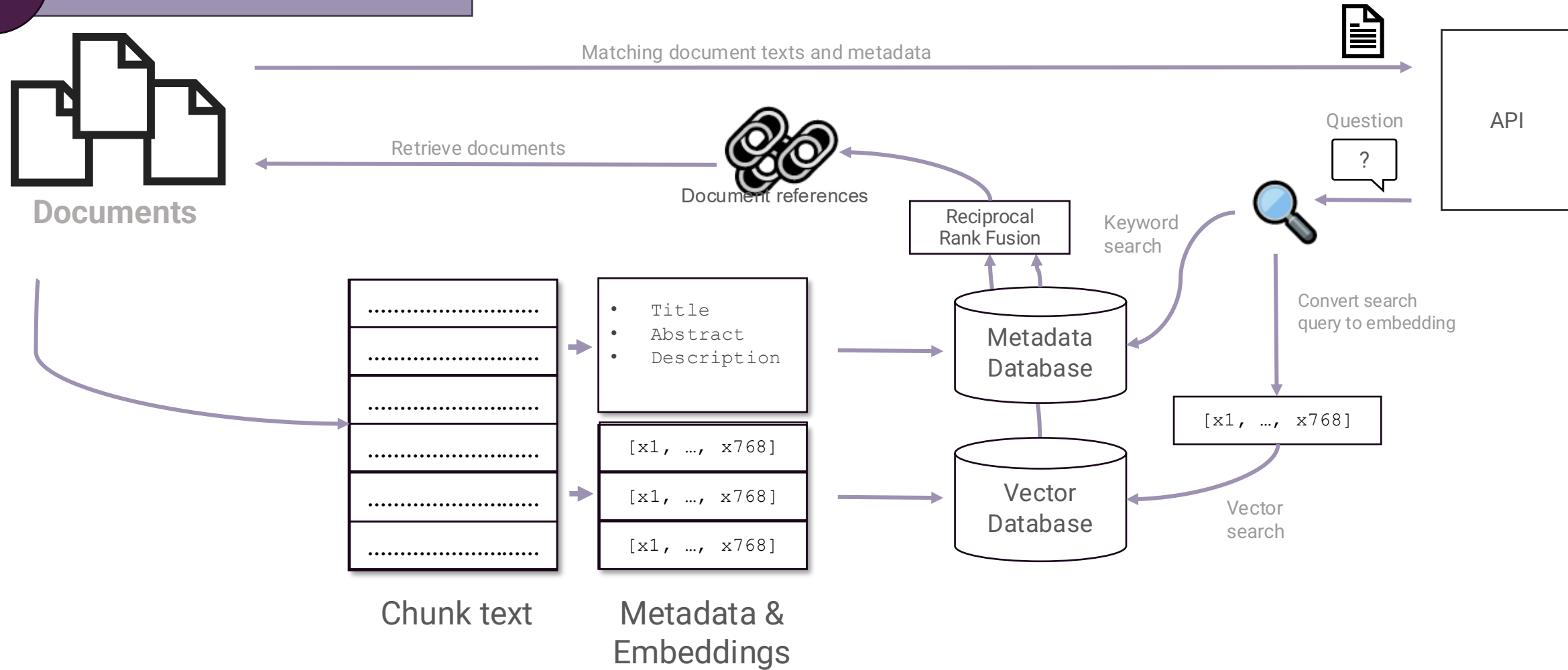
1

Document retrieval

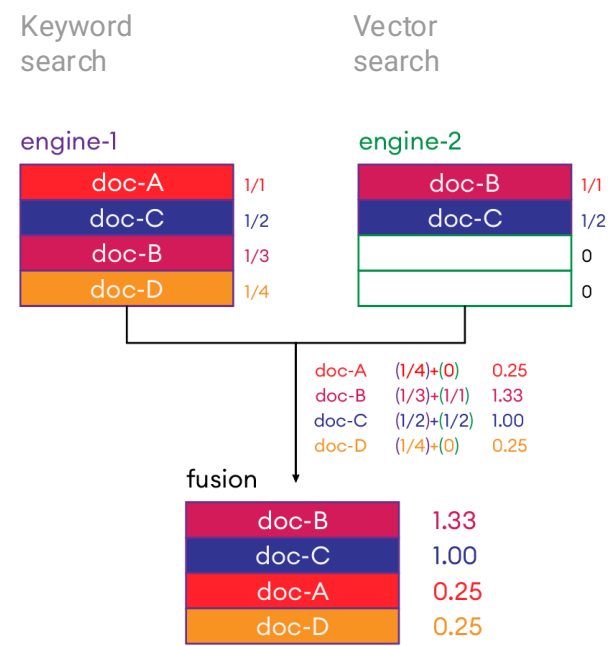


1

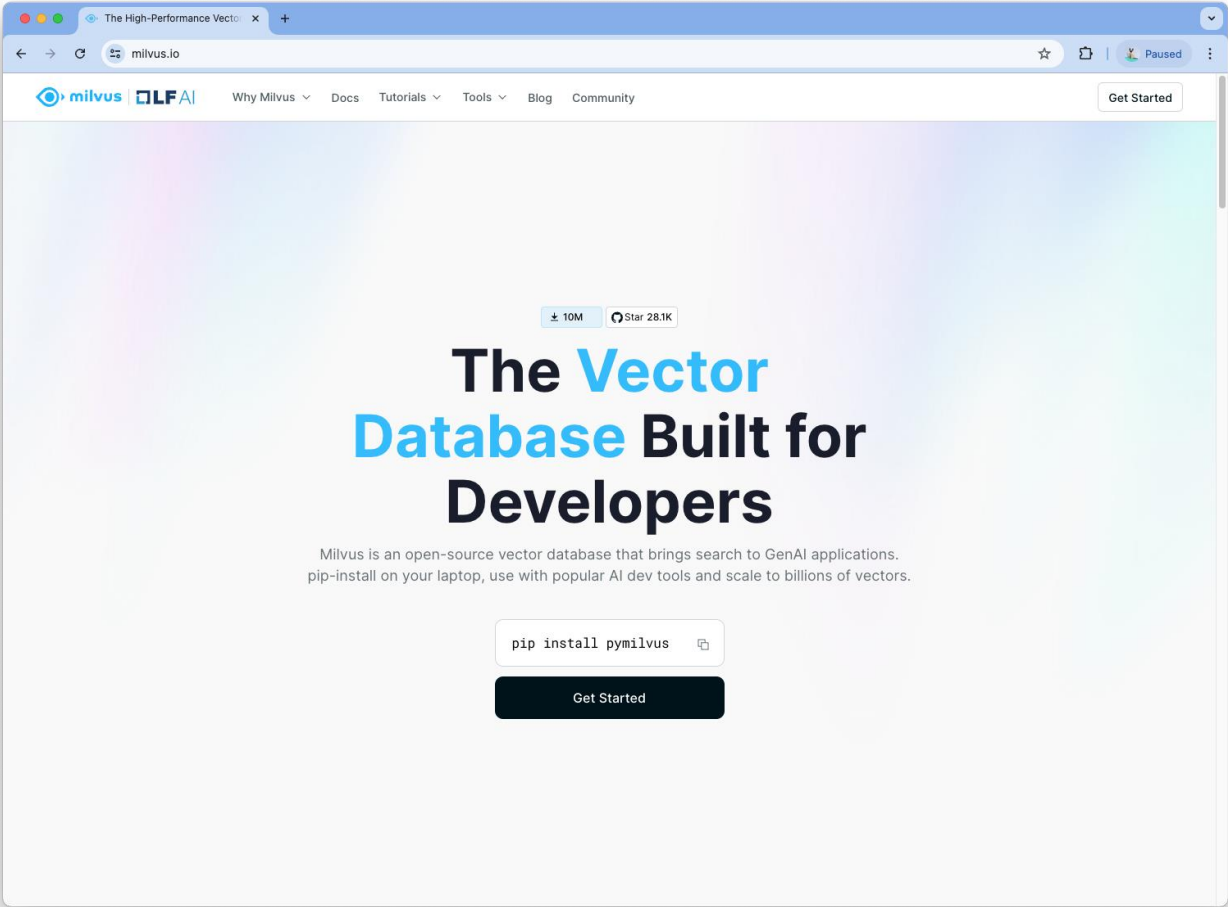
Document retrieval



Reciprocal Rank Fusion



<https://images.app.goo.gl/iQQfwHGhw3swKr948>



<https://milvus.io/>

Level 2
Hybrid search

Retriever - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/fadb1c03-ef6c-46db-9547-3da7dc0a7...

Paused

Personal > Projects > levels-of-rag > RunnableSequence > ... > Retriever

PlaygroundAdd to DatasetShareAnnotateDEVELOPER

TRACE

CollapseStatsMost relevant

RunnableSequence 4.59s 2,123

map:key:context 0.77s

Retriever 0.77s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGenerativeAI models/gemini-1.5-pro 3.82s

Some runs have been hidden. [Show 3 hidden runs](#)

Retriever

RunFeedbackMetadata

Input

1query: What's the talk starting with Maximizing about?

YAML

Output

DOCUMENTS 4

The Levels of RAG LLM's can be supercharged using a ... +20

Scikit-Learn can do THAT?! Many of us know scikit-learn f... +20

Evaluating LLM Frameworks Large Language Models are ... +20

Software as a Service (SaaS): the Force behind making microchips l... +20

START TIME07/02/2024, 10:35:28 PM

END TIME07/02/2024, 10:35:29 PM

TIME TO FIRST TOKENN/A

STATUSSuccess

TOTAL TOKENS0 tokens

LATENCY0.77s

TYPERetriever

TAGSGoogle Gemini v1.5 Embeddings, Reptile, F1SS

Without hybrid search

Level 2

Hybrid search

Retriever - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/92ff66ab-1883-4faa-8888-50214792b...

Paused

Personal > Projects > levels-of-rag > RunnableSequence > ... > Retriever

PlaygroundAdd to DatasetShareAnnotateDEVELOPER

TRACE

CollapseStatsMost relevant

RunnableSequence 3.67s 3,448

map:key:context 0.41s

Retriever 0.40s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatGoogleGenerativeAI models/gemini-1.5-pro 3.25s

Some runs have been hidden. [Show 3 hidden runs](#)

Retriever

RunFeedbackMetadata

Run IDTrace ID

Input

1query: What's the talk starting with Maximizing about?

YAML

Output

DOCUMENTS 4

Cloud? No Thanks! I'm Gonna Run GenAI on M... 450845329692046494

Computer vision at the Dutch Tennis Federatio... 450845329692046484

BERTopic to accelerate Ukrainian aid by the R... 450845329692046488

Maximizing marketplace experimentation: swit... 450845329692046492

START TIME07/01/2024, 05:29:02 PM

END TIME07/01/2024, 05:29:03 PM

TIME TO FIRST TOKENN/A

STATUSSuccess

TOTAL TOKENS0 tokens

LATENCY0.40s

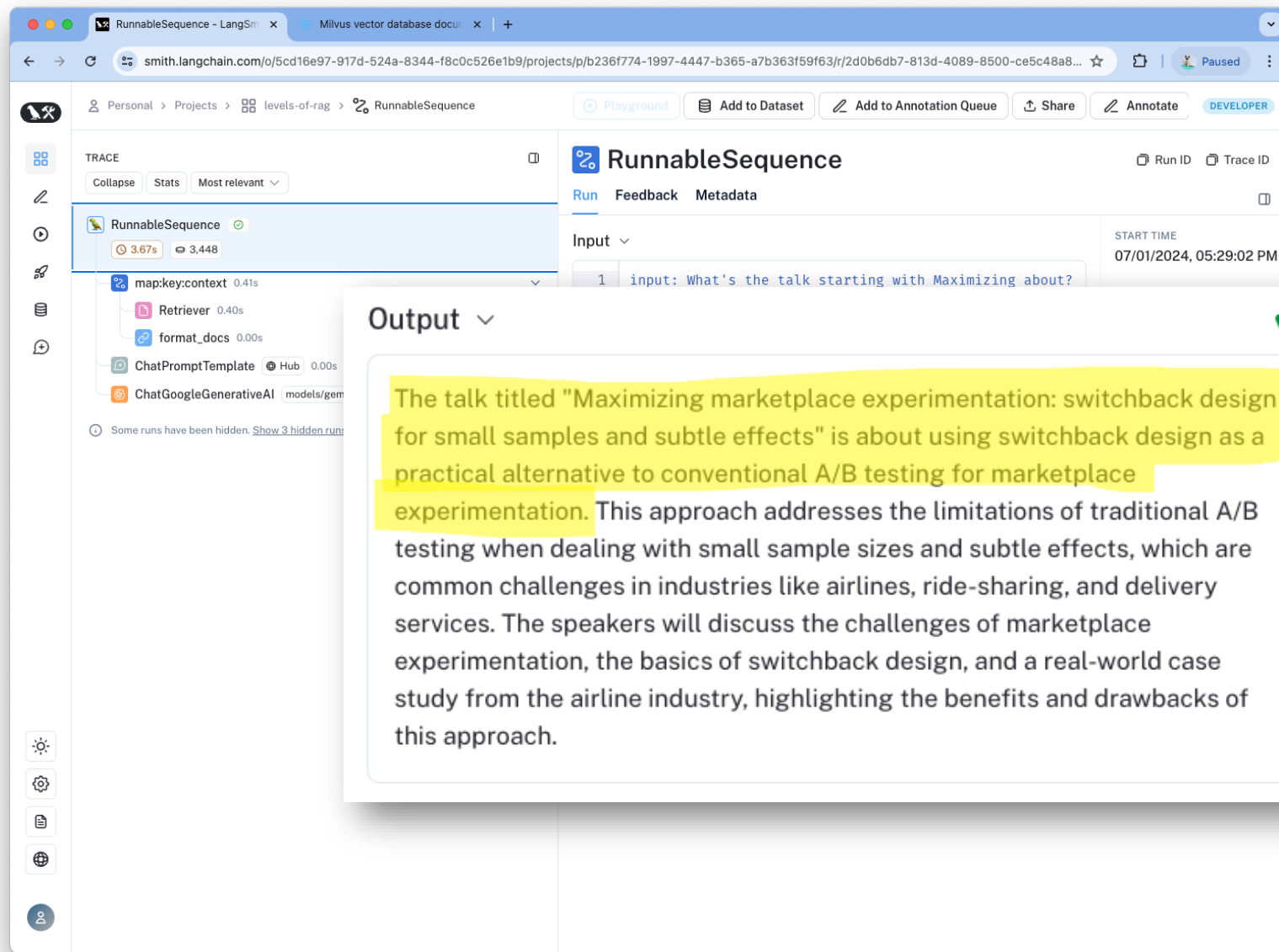
TYPERetriever

TAGSseq:step:1

✓ With hybrid search

Level 2

Hybrid search

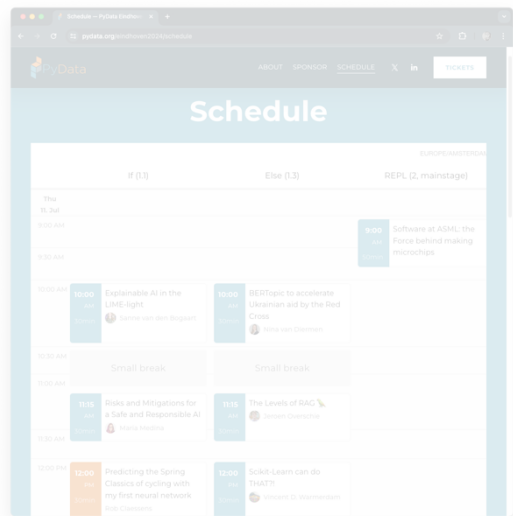


The screenshot displays the RunnableSequence web interface. The top navigation bar includes a 'PLAYGROUND' button and several action buttons: 'Add to Dataset', 'Add to Annotation Queue', 'Share', 'Annotate', and a 'DEVELOPER' badge. The main area is divided into a left sidebar and a right pane. The sidebar shows a 'TRACE' section with a 'RunnableSequence' entry highlighted, indicating a duration of 3.67s and a token count of 3,448. Below this, a flow diagram shows the sequence of operations: 'map:key:context' (0.41s), 'Retriever' (0.40s), 'format_docs' (0.00s), 'ChatPromptTemplate' (0.00s), and 'ChatGoogleGenerativeAI' (models/gemini-1.5-pro-latest). The right pane, titled 'RunnableSequence', shows the 'Run' tab with an input field containing 'input: What's the talk starting with Maximizing about?'. The 'START TIME' is listed as '07/01/2024, 05:29:02 PM'. A green checkmark is placed to the right of the output window. The output window, titled 'Output', displays the following text:

The talk titled "Maximizing marketplace experimentation: switchback design for small samples and subtle effects" is about using switchback design as a practical alternative to conventional A/B testing for marketplace experimentation. This approach addresses the limitations of traditional A/B testing when dealing with small sample sizes and subtle effects, which are common challenges in industries like airlines, ride-sharing, and delivery services. The speakers will discuss the challenges of marketplace experimentation, the basics of switchback design, and a real-world case study from the airline industry, highlighting the benefits and drawbacks of this approach.

Level 2

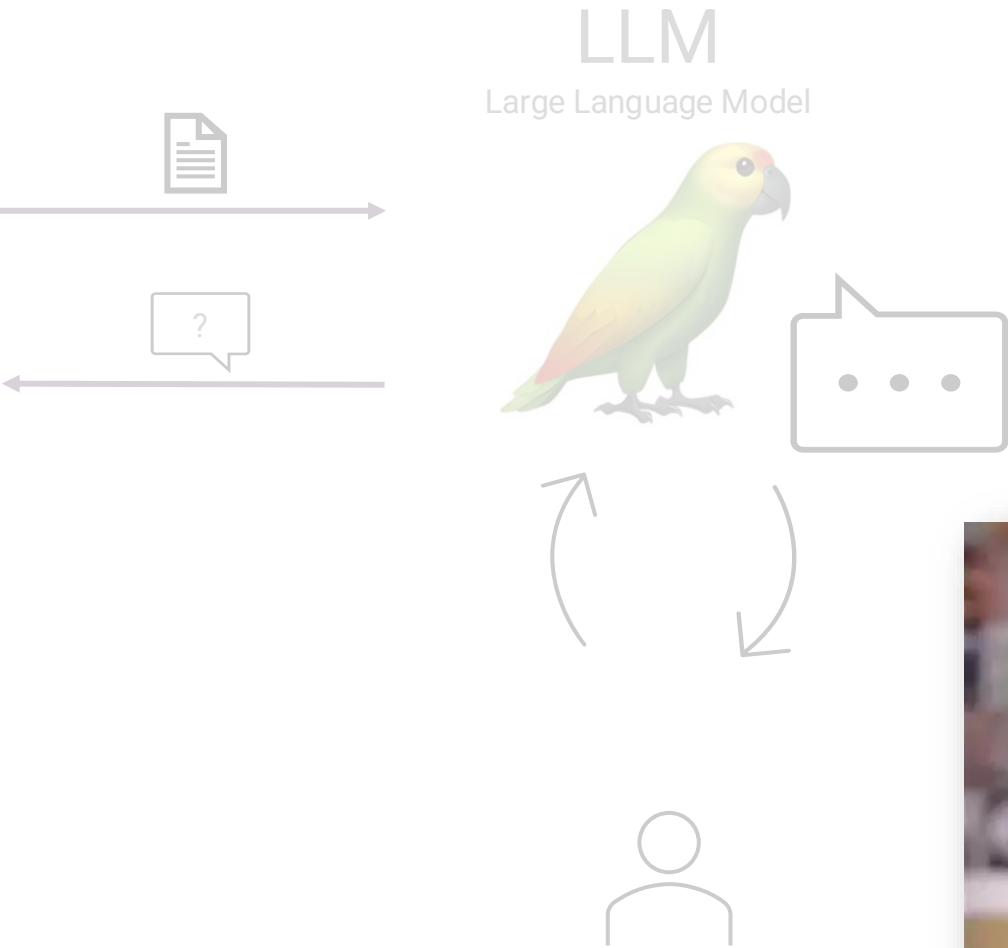
Hybrid search



The screenshot shows the 'Schedule' page of the PyData Eindhoven 2024 website. The page has a header with 'PyData' and navigation links for 'ABOUT', 'SPONSOR', 'SCHEDULE', and 'TICKETS'. The main content is a table with columns for 'if (1.1)', 'Else (1.3)', and 'REPL (2, maintage)'. The table lists various sessions with their titles, speakers, and times.

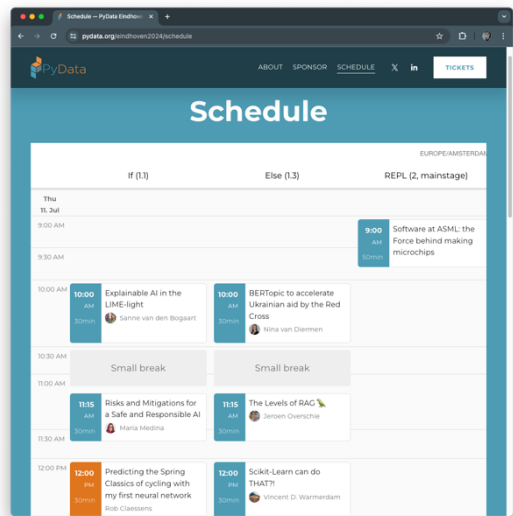
Schedule		
if (1.1)	Else (1.3)	REPL (2, maintage)
Thu 11 Jul 10:00 AM		10:00 AM Software at ASML: the Force behind making microscopes
10:30 AM	10:00 AM Explainable AI in the LIME-light Barrie van den Bogaert	10:00 AM BEDTopic to accelerate Ukesman aid by the Red Cross Tina van Oort
10:30 AM	Small break	Small break
11:15 AM	11:15 AM Risks and Mitigations for a Safe and Responsible AI Hans Hees	11:15 AM The Levels of RAG Steven Christen
12:00 PM	12:00 PM Predicting the Spring Classics of cycling with my first neural network Bart Claessens	12:00 PM Gokul-Learn can do THAT? Vishwanath D. Venkatasubramanian

PyData Eindhoven 2024
Schedule



Level 2

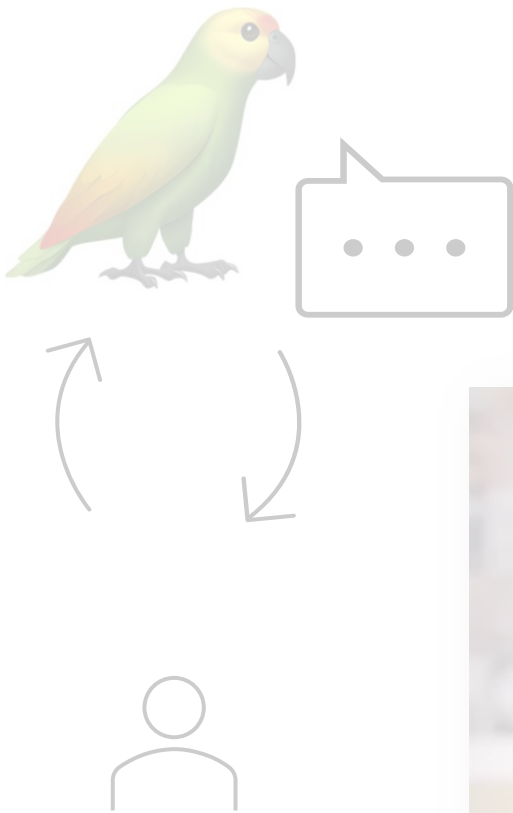
Hybrid search



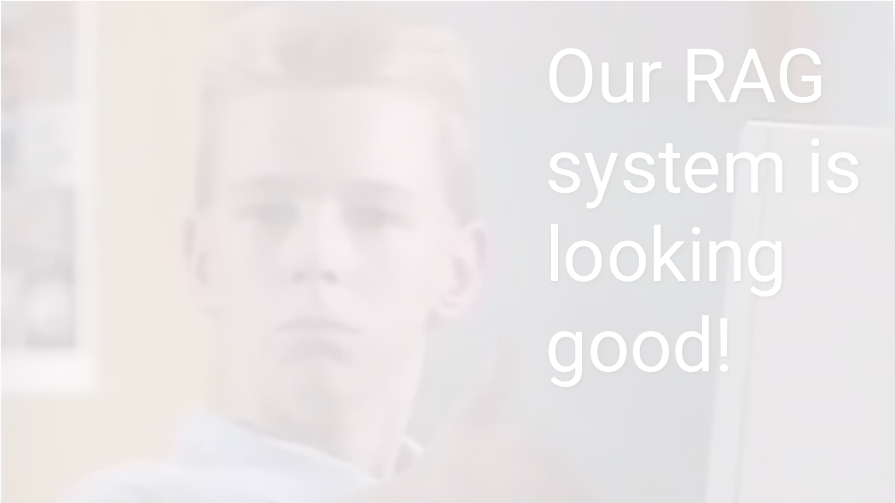
PyData Eindhoven 2024
Schedule



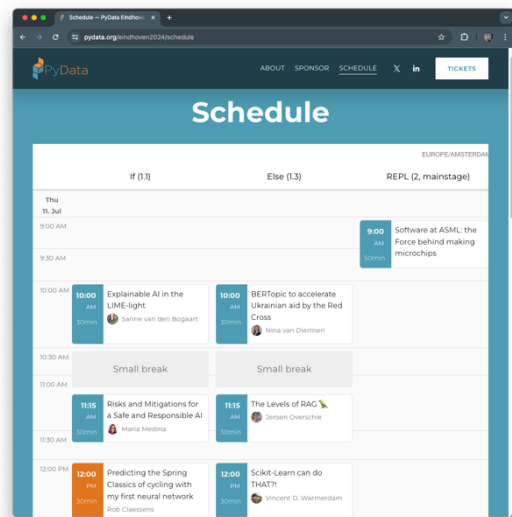
LLM
Large Language Model



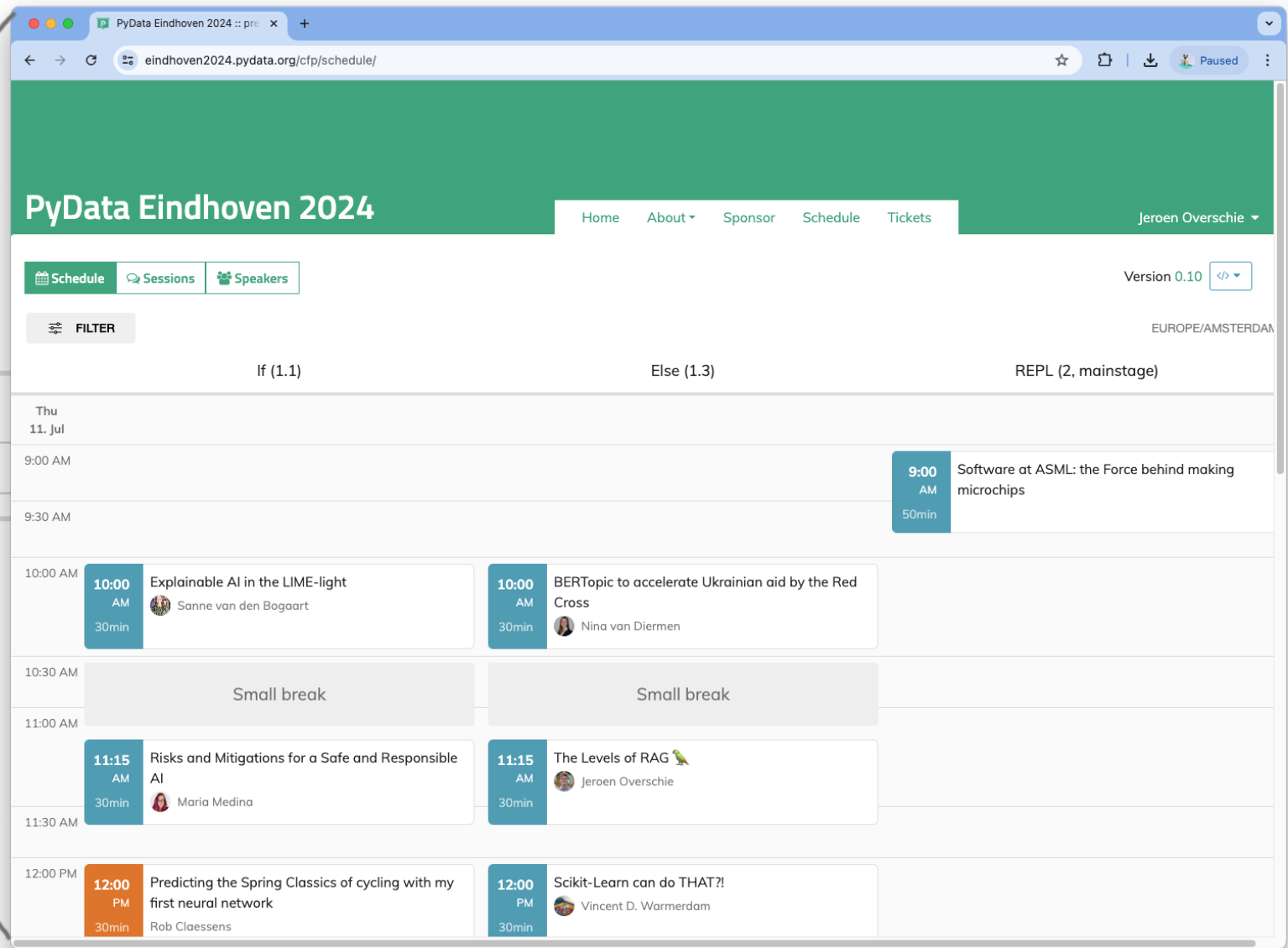
Actually, how
did I load in
the PyData
schedule?



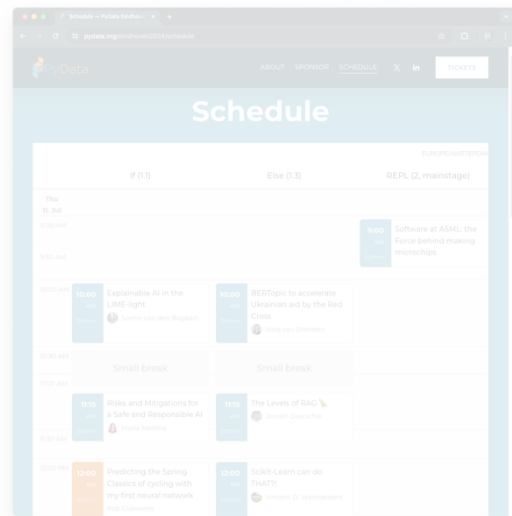
Level 2
Hybrid search



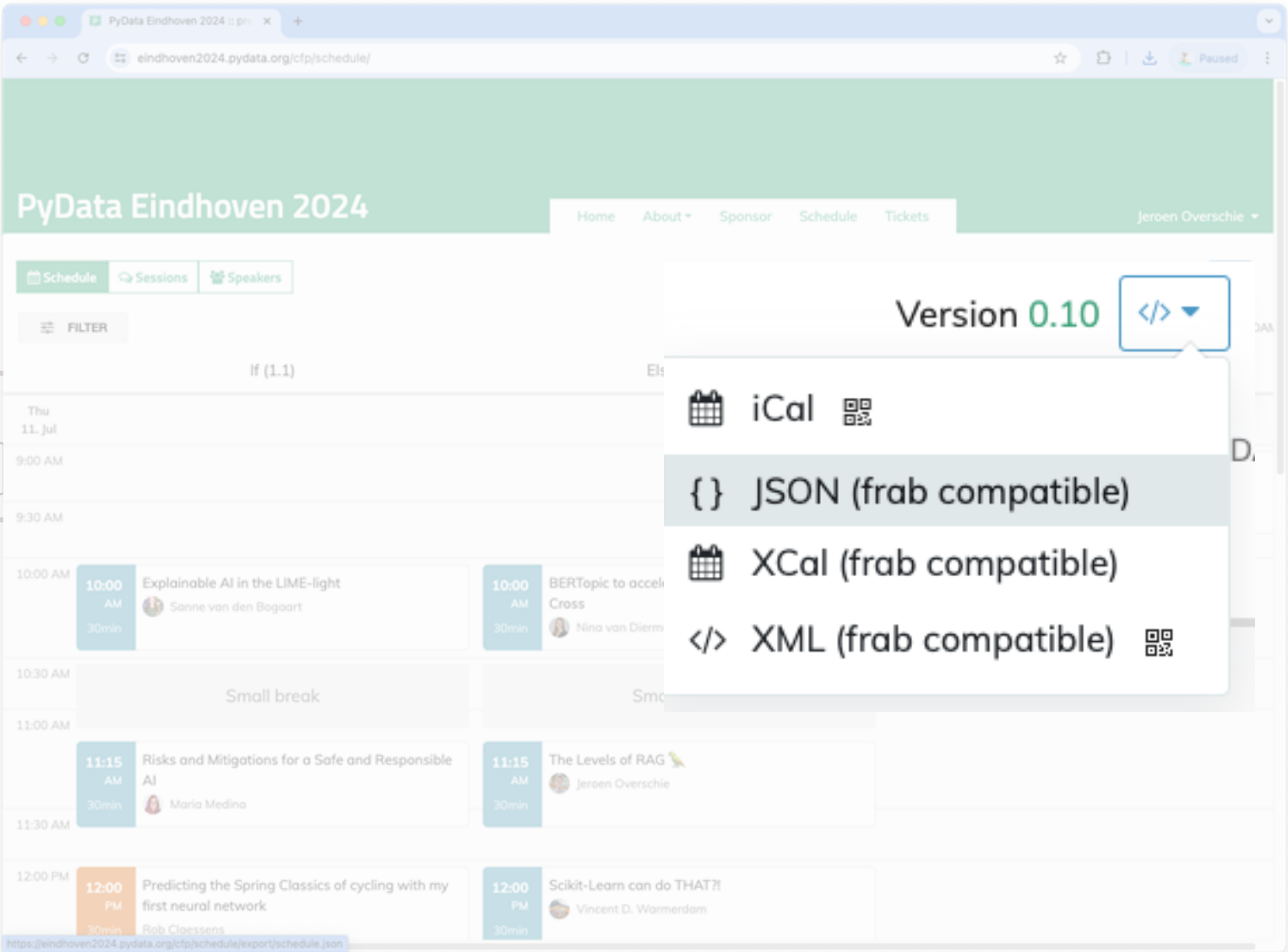
PyData Eindhoven 2024
Schedule



Level 2
Hybrid search
Let's give it a spin!



PyData Eindhoven 2024
Schedule



If all data was structured

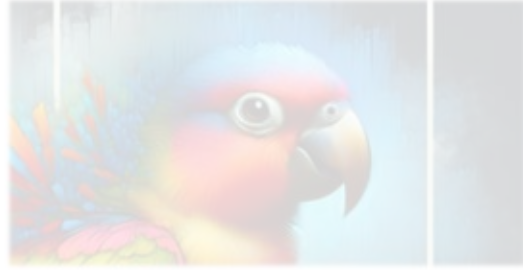
THAT WOULD BE GREAT



The Levels of **RAG**



Level 1
Basic RAG



Level 2
Hybrid search



Level 3
Advanced data formats



Level 4
Multimodal

Level 3
Advanced
data formats

Let's take PDF's
as an example

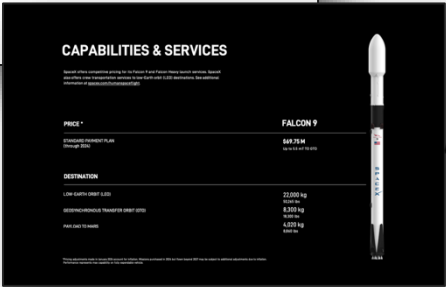
LangChain Integrations API Reference More pdf0/34 Search				
Providers	Anthropic	PyMuPDFLoader	Load PDF files using PyMuPDF.	✓✗
	AWS	PyPDFDirectoryLoader	Load a directory with PDF files using pypdf and chunks at character level.	✗✗
PyPDFLoader		Load PDF using pypdf into list of documents.		
OpenAI		PyPDFium2Loader	Load PDF using pypdfium2 and chunks at character level.	✓✗



Level 3
Advanced
data formats

Let's take PDF's
as an example

LangChain				
Integrations API Reference More				
pdf 0/34 Search				
Providers				
Anthropic	PyMuPDFLoader	Load PDF files using PyMuPDF.	✓	✗
AWS	PyPDFDirectoryLoader	Load a directory with PDF files using pypdf and chunks at character level.	✗	✗
PyPDFLoader Load PDF using pypdf into list of documents.				
OpenAI	PyPDFium2Loader	Load PDF using pypdfium2 and chunks at character level.	✓	✗



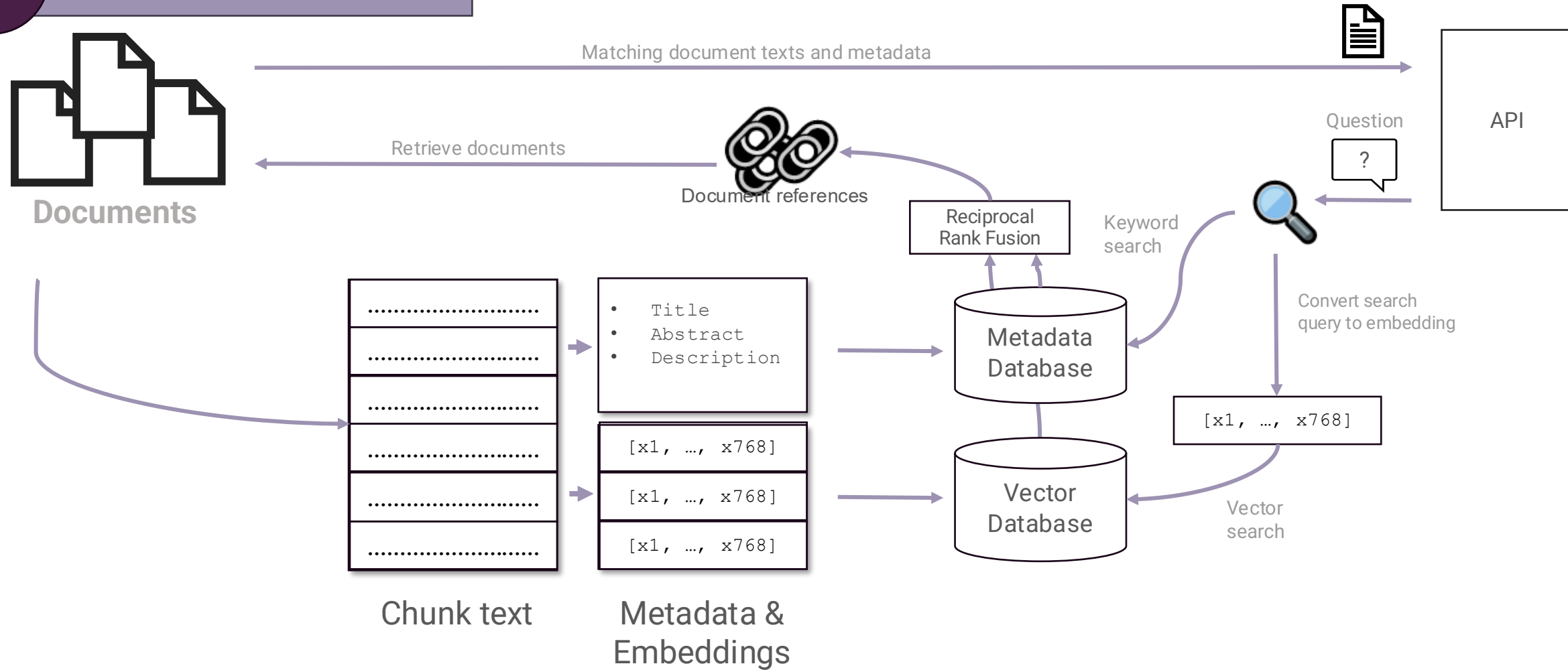
© Space Exploration Technologies Corp. All rights reserved.
COPYRIGHT
Subject to the existing rights of third parties, Space Exploration Technologies Corp. (SpaceX) is the owner of the copyright in this work, and no portion hereof is to be copied, reproduced, or disseminated without the prior written consent of SpaceX.

Falcon 9 Launch Vehicle NAFCOM Cost Estimates August 2011
NASA Associate Deputy Administrator for Policy 1

CAPABILITIES & SERVICES
FALCON 9 DESTINATIONS
SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low-Earth orbit (LEO) destinations. See additional information at spacex.com/humanspaceflight.
LOW-EARTH ORBIT (LEO)
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)
PAYLOAD TO MARS 22,000 kg
50,265 lbs
8,300 kg
18,300 lbs
4,020 kg
8,860 lbs
STANDARD PAYMENT PLAN (through 2024) \$69.75 MUP to 5.5 mT
TO GTO PRICE * *Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation. Performance represents max capability on fully expendable vehicle.

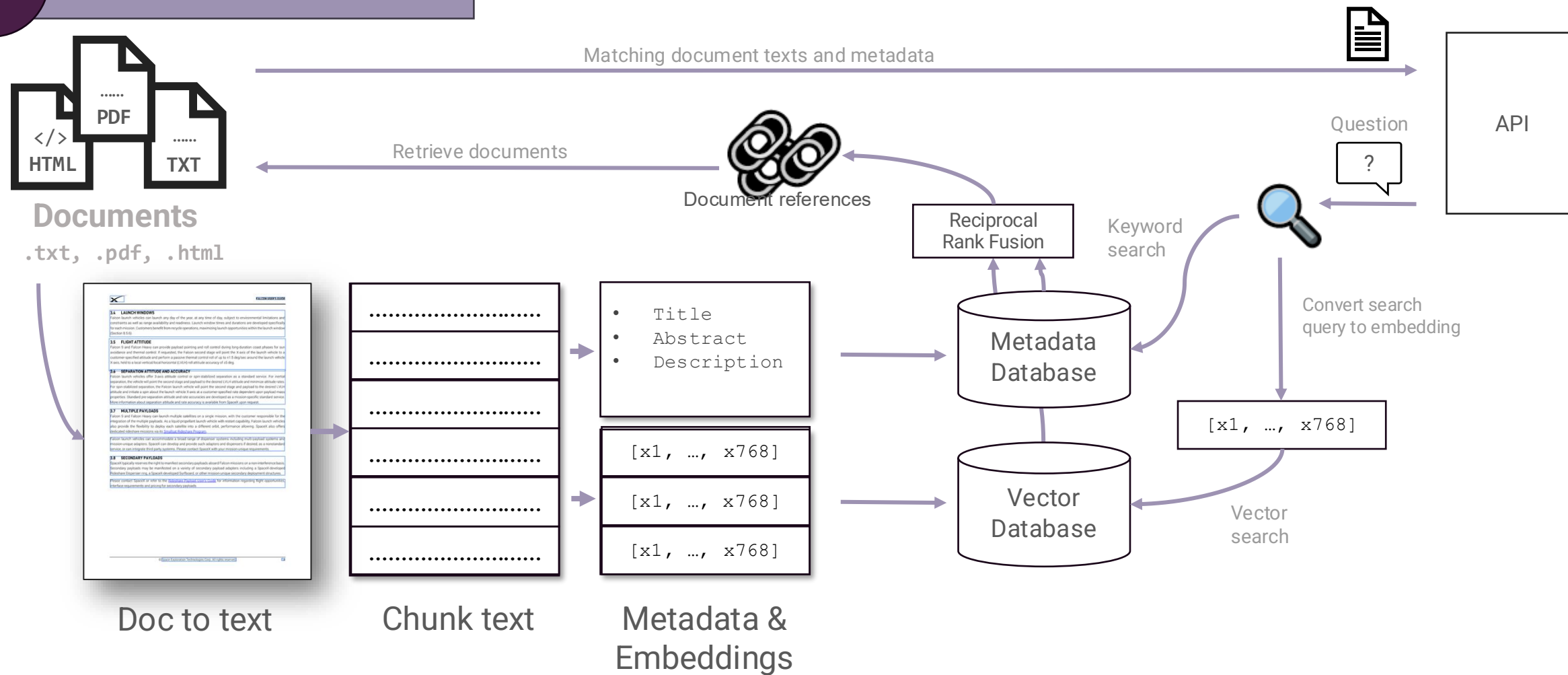
1

Document retrieval



1

Document retrieval



Level 3

Advanced
data formats

RunnableSequence - LangSm x +

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/74ea9bf0-2c20-4b76-b6c2-bb2519610... ☆ | Paused ⋮

Personal >

TRACE

Collapse ⌵

RunnableS

2.55s

map:ke

Re

fo

ChatP

ChatOpenAI gpt-3.5-turbo 1.94s

Some runs have been hidden. [Show 3 hidden runs](#)

Input ⌵

1 input: whats the price for a falcon 9 rocket launch with the standard payment plan?

YAML ⌵

Output ⌵

The price for a Falcon 9 rocket launch with the standard payment plan is

Annotate DEVELOPER

Run ID Trace ID

START TIME 7/03/2024, 02:48:13

END TIME 7/03/2024, 02:48:15

TIME TO FIRST TOKEN /A

STATUS Success

TOTAL TOKENS 281 tokens / \$0.0006975

LATENCY 2.55s

TYPE Sequence

PRICE *	FALCON 9
STANDARD PAYMENT PLAN (through 2024)	\$69.75 M Up to 5.5 mT to GTO
LOW-EARTH ORBIT (LEO)	22,000 kg 52,245 lbs
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)	8,300 kg 18,300 lbs
PAYLOAD TO MARS	4,020 kg 8,860 lbs

Level 3

Advanced
data formats

RunnableSequence - LangSm x +

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/721766e1-d736-4f28-8573-aada46130... ☆ | Paused ⋮

Personal >

TRACE

Collapse ⌵

St...

Runnables

map:k

R

fc

ChatPromptTemplate

ChatOpenAI gpt-3.5-turbo 1.61s

Some runs

Input ⌵

1 input: what is the payload I can carry with the Falcon 9 rocket to Mars?

YAML ⌵

Output ⌵

The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

Annotate DEVELOPER

Run ID Trace ID

START TIME 07/03/2024, 02:48:25 PM

END TIME 07/03/2024, 02:48:36 PM

TIME TO FIRST TOKEN N/A

STATUS Success

TOTAL TOKENS 800 tokens / \$0

LATENCY 11.04s

TYPE Sequence

CAPABILITIES & SERVICES

SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low-Earth orbit (LEO) destinations. See additional information at [spacex.com/human-spaceflight](https://www.spacex.com/human-spaceflight).

PRICE * FALCON 9

STANDARD PAYMENT PLAN (through 2024) \$69.75 M Up to 5.5 mT TO LEO

DESTINATION

PAYLOAD TO MARS

4,020 kg
8,860 lbs

<https://www.spacex.com/media/Capabilities&Services.pdf>

Level 3

Advanced data formats

RunnableSequence - LangSm

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/721766e1-d736-4f28-8573-aada46130...

Paused

Personal > Projects > levels-of-rag > RunnableSequence

PlaygroundAdd to DatasetAdd to Annotation QueueShareAnnotateDEVELOPER

TRACE

CollapseStatsMost relevant

RunnableSequence

11.04s800

map:key:context9.42s

Retriever9.41s

format_docs0.00s

ChatPromptTemplateHub0.00s

ChatOpenAIGPT-3.5-turbo1.61s

Some runs have been hidden. [Show 3 hidden runs](#)

RunnableSequence

RunIDTrace ID

RunFeedbackMetadata

Input

1

input: what is the payload I can carry with the Falcon 9 rocket to Mars?

YAML

Output

The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

START TIME

07/03/2024, 02:48:25 PM

END TIME

07/03/2024, 02:48:36 PM

TIME TO FIRST TOKEN

N/A

STATUS

Success

TOTAL TOKENS

800 tokens / \$0.000462

LATENCY

11.04s

TYPE

Sequence

Level 3

Advanced
data formats

ChatOpenAI - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/ec4e011c-f620-442e-ad8e-f43688eaa...

Paused

Personal > Projects > levels-of-rag >

TRACE

Collapse

Stats

Most relevant

RunnableSequence

11.04s 800

map:key:context 9.42s

Retriever 9.41s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatOpenAI gpt-3.5-turbo 1.61s

Some runs have been hidden. [Show 3 hidden runs](#)

Settings

Tools

Docs

Help

Profile

Context: CAPABILITIES & SERVICES

FALCON 9 DESTINATIONS

SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low-Earth orbit (LEO) destinations. See additional information at spacex.com/humanspaceflight.

LOW-EARTH ORBIT (LEO) GEOSYNCHRONOUS TRANSFER ORBIT (GTO) PAYLOAD TO MARS

22,000 kg50,265 lbs8,300 kg18,300 lbs4,020 kg8,860 lbs

STANDARD PAYMENT PLAN (through 2024) \$69.75 M

Up to 5.5 mT TO GTO PRICE *

*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation. Performance represents max capability on fully expendable vehicle.

© Space Exploration Technologies Corp. All rights reserved.

13

The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

Level 3

Advanced
data formats

ChatOpenAI - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/ec4e011c-f620-442e-ad8e-f43688eaa...

Paused

Personal > Projects > levels-of-rag >

TRACE

Collapse

Stats

Most relevant

RunnableSequence

11.04s 800

map:key:context 9.42s

Retriever 9.41s

format_docs 0.00s

ChatPromptTemplate Hub 0.00s

ChatOpenAI gpt-3.5-turbo 1.61s

Some runs have been hidden. [Show 3 hidden runs](#)

🔍

⚙️

📄

🌐

👤

Context: CAPABILITIES & SERVICES

FALCON 9 DESTINATIONS

SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low-Earth orbit (LEO) destinations. See additional information at spacex.com/humanspaceflight.

LOW-EARTH ORBIT (LEO) GEOSYNCHRONOUS TRANSFER ORBIT (GTO) PAYLOAD TO MARS 22,000 kg 50,265 lbs 8,300 kg 18,300 lbs 4,020 kg 8,860 lbs

STANDARD PAYMENT PLAN (through 2024) \$69.75 MU up to 5.5 mT TO GTO PRICE *

*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation. Performance represents max capability on fully expendable vehicle.

© Space Exploration Technologies Corp. All rights reserved.

13

The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

Level 3

Advanced data formats

ChatOpenAI - LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/projects/p/b236f774-1997-4447-b365-a7b363f59f63/r/ec4e011c-f620-442e-ad8e-f43688eaa...

Paused

Personal > Projects > levels-of-rag > RunnableSequence > ChatOpenAI

PlaygroundAdd to DatasetShareAnnotateDEVELOPER

TRACE

CollapseStatsMost relevant

RunnableSequence11.04s800

map:key:context9.42s

Retriever9.41s

format_docs0.00s

ChatPromptTemplateHub0.00s

ChatOpenAIgpt-3.5-turbo1.61s

Some runs have been hidden. [Show 3 hidden runs](#)

ChatOpenAI

Run IDTrace ID

RunFeedbackMetadata

Input

HUMAN

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.
Question: what is the payload I can carry with the Falcon 9 rocket to Mars?
Context: CAPABILITIES & SERVICESFALCON 9DESTINATIONS
SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceXalso offers crew transportation services tolow-Earth orbit (LEO) destinations.See additional information at spacex.com/humanspaceflight.
LOW-EARTH ORBIT (LEO)GEOSYNCHRONOUS TRANSFER ORBIT (GTO)PAYLOAD TO MARS22,000 kg50,265 lbs8,300 kg18,300 lbs4,020 kg8,860 lbsSTANDARD PAYMENT PLAN(through 2024)\$69.75 MUp to 5.5 mT TO GTOPRICE *
*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation.Performance represents max capability on fully expendable vehicle.

© Space Exploration Technologies Corp. All rights reserved. 13

Output

AI

The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

START TIME

07/03/2024, 02:48:34 PM

END TIME

07/03/2024, 02:48:36 PM

TIME TO FIRST TOKEN

N/A

STATUS

Success

TOTAL TOKENS

800 tokens / \$0.000462

LATENCY

1.61s

TYPE

LLM

TAGS

seq:step:3

Level 3

Advanced
data formats

LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/playground/r/ec4e011c-f620-442e-ad8e-f43688eaa4c8

Paused

Personal < Back to ChatOpenAI Run

DEVELOPER

Secrets & API KeysHistory

Editing ChatOpenAI

Input

HUMAN

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.
Question: what is the payload I can carry with the Falcon 9 rocket to Mars?
Context: CAPABILITIES & SERVICESFALCON 9DESTINATIONSpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceXalso offers crew transportation services tolow-Earth orbit (LEO) destinations.See additional information at spacex.com/humanspaceflight. LOW-EARTH ORBIT (LEO)GEOSYNCHRONOUS TRANSFER ORBIT (GTO)PAYLOAD TO MARS22,000 kg50,265 lbs8,300 kg18,300 lbs4,020 kg8,860 lbsSTANDARD PAYMENT PLAN(through 2024)\$69.75 MUp to 5.5 mT TO GTOPRICE *
*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation.Performance represents max capability on fully expendable vehicle.

© Space Exploration Technologies Corp. All rights reserved. 13

SpaceX requires that customers verify the mass properties of their system through measurement before shipping it to the launch site. SpaceX may request insight into

Output

ChatOpenAI gpt-3.5-turbo

Fancy rendering

AI
The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

1.61s800Annotate16 minutes ago

Start

Settings

Provider
ChatOpenAI

Model
gpt-3.5-turbo

Temperature
0

Maximum Length
-1

Top P
1

Presence Penalty
0

Frequency Penalty
0

Tools
Tools auto-selected b+ Add

Seed

JSON mode

Stop Sequences
+ Add stop sequence

Provider Options

Level 3
Advanced
data formats

LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/playground/r/ec4e011c-f620-442e-ad8e-f43688aa4c8

Personal < Back to ChatOpenAI Run

Editing ChatOpenAI

Input

HUMAN

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.
Question: what is the payload I can carry with the Falcon 9 rocket to Mars?
Context: CAPABILITIES & SERVICES
FALCON 9 DESTINATIONS
SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low-Earth orbit (LEO) destinations. See additional information at [spacex.com/humanspaceflight](#).
LOW-EARTH ORBIT (LEO) GEOSYNCHRONOUS TRANSFER ORBIT (GTO) PAYLOAD TO MARS
22,000 kg 50,265 lbs 8,300 kg 18,300 lbs 4,020 kg 8,860 lbs
STANDARD PAYMENT PLAN (through 2024) \$69.75 M Up to 5.5 mT TO GTO
PRICE *
*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation. Performance represents max capability on fully expendable vehicle.
© Space Exploration Technologies Corp. All rights reserved. 13
SpaceX requires that customers verify the mass properties of their system through measurement before shipping it to the launch site. SpaceX may request insight into

Output

ChatOpenAI gpt-3.5-turbo

Fancy rendering

AI

The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

Computer
Vision model

CAPABILITIES & SERVICES

SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low-Earth orbit (LEO) destinations. See additional information at [spacex.com/humanspaceflight](#).

PRICE *

FALCON 9

STANDARD PAYMENT PLAN (through 2024)

\$69.75 M Up to 5.5 mT TO GTO

DESTINATION	
LOW-EARTH ORBIT (LEO)	22,000 kg 50,265 lbs
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)	8,300 kg 18,300 lbs
PAYLOAD TO MARS	4,020 kg 8,860 lbs

*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation. Performance represents max capability on fully expendable vehicle.



Destination	Load Carry
LOW-EARTH ORBIT (LEO)	22,000 kg (50,265 lbs)
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)	8,300 kg (18,300 lbs)
PAYLOAD TO MARS	4,020 kg (8,860 lbs)

Level 3

Advanced data formats

LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/playground/r/ec4e011c-f620-442e-ad8e-f43688eaa4c8

Personal < Back to ChatOpenAI Run

Editing ChatOpenAI

Input

HUMAN

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that.

Output

ChatOpenAI gpt-3.5-turbo

Fancy rendering

AI

The payload you can carry with the Falcon 9 rocket to Mars is 4,020 kg

Context:

Destination Falcon 9
--- ---
LOW-EARTH ORBIT (LEO) 22,000 kg (50,265 lbs)
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)
8,300 kg (18,300 lbs)
PAYLOAD TO MARS 4,020 kg (8,860 lbs)

Markdown formatted table in context

CAPABILITIES & SERVICES

SpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceX also offers crew transportation services to low-Earth orbit (LEO) destinations. See additional information at [spacex.com/humanspaceflight](#).

PRICE

FALCON 9

STANDARD PAYMENT PLAN (through 2024)

\$69.75 M Up to 5.5 mT TO GTO

DESTINATION	
LOW-EARTH ORBIT (LEO)	22,000 kg 50,265 lbs
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)	8,300 kg 18,300 lbs
PAYLOAD TO MARS	4,020 kg 8,860 lbs

*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation. Performance represents max capability on fully expendable vehicle.



Destination	Load Carry
LOW-EARTH ORBIT (LEO)	22,000 kg (50,265 lbs)
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)	8,300 kg (18,300 lbs)
PAYLOAD TO MARS	4,020 kg (8,860 lbs)

Level 3

Advanced
data formats

LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/playground/r/ec4e011c-f620-442e-ad8e-f43688eaa4c8

Paused

Personal < Back to ChatOpenAI Run

DEVELOPER

Secrets & API KeysHistory

Editing ChatOpenAI

Input

HUMAN ↕

You are an assistant for qu
Use the following pieces o
answer the question. If you
just say that you don't kno
maximum and keep the an
Question: what is the paylo
Falcon 9 rocket to Mars?
Context:

Destination	Falcon 9
LOW-EARTH ORBIT (LEO)	22,000 kg (50,265 lbs)
GEOSYNCHRONOUS TRANSFER ORBIT (GTO)	
8,300 kg (18,300 lbs)	
PAYLOAD TO MARS	4,020 kg (8,860 lbs)

+ Message

AI

The payload you can carry with the Falcon 9 rocket to Mars is 4,020 kg (8,860 lbs).

Top P1

Presence Penalty0

Frequency Penalty0

ToolsTools auto-selected b+ Add

Seed

JSON mode

Stop Sequences+ Add stop sequence

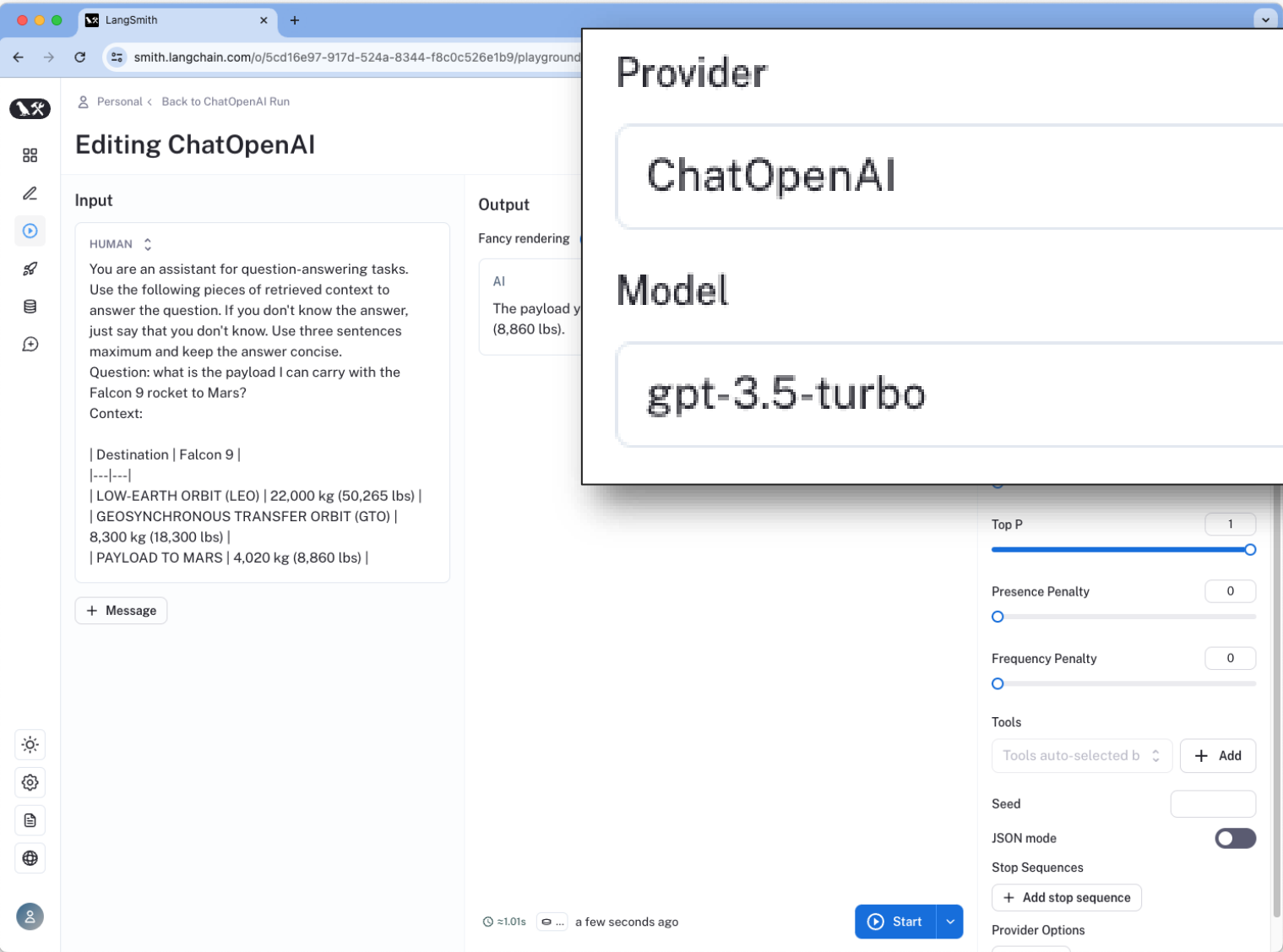
Provider Options

≈1.01s a few seconds ago

Start

Level 3

Advanced
data formats



I have to
make a
confession.

We were using
GPT-3.5 turbo
for this task.
**What about
newer models?**

The Levels of **RAG**



Level 1
Basic RAG



Level 2
Hybrid search



Level 3
Advanced data formats



Level 4
Multimodal

Level 4

Multimodal

LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/playground/r/ec4e011c-f620-442e-ad8e-f43688eaa4c8

Paused

Personal < Back to ChatOpenAI Run

DEVELOPER

Editing ChatOpenAI

Secrets & API KeysHistory

Input

HUMAN ↕

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.
Question: what is the payload I can carry with the Falcon 9 rocket to Mars?
Context: CAPABILITIES & SERVICESFALCON 9DESTINATIONSpaceX offers competitive pricing for its Falcon 9 and Falcon Heavy launch services. SpaceXalso offers crew transportation services tolow-Earth orbit (LEO) destinations.See additional information at spacex.com/humanspaceflight. LOW-EARTH ORBIT (LEO)GEOSYNCHRONOUS TRANSFER ORBIT (GTO)PAYLOAD TO MARS22,000 kg50,265 lbs8,300 kg18,300 lbs4,020 kg8,860 lbsSTANDARD PAYMENT PLAN(through 2024)\$69.75 MUp to 5.5 mT TO GTOPRICE *
*Pricing adjustments made in January 2024 account for inflation. Missions purchased in 2024 but flown beyond 2027 may be subject to additional adjustments due to inflation.Performance represents max capability on fully expendable vehicle.

© Space Exploration Technologies Corp. All rights reserved. 13

SpaceX requires that customers verify the mass properties of their system through measurement before shipping it to the launch site. SpaceX may request insight into

Output

ChatOpenAI gpt-3.5-turbo

Fancy rendering ☒

AI

The payload capacity of the Falcon 9 rocket to Mars is 8,300 kg or 18,300 lbs. SpaceX offers competitive pricing for its launch services to various destinations, including low-Earth orbit and Mars. Customers need to verify the mass properties of their payload before shipping it to the launch site.

Settings

ProviderChatOpenAI

Modelgpt-3.5-turbo

Temperature0

Maximum Length-1

Top P1

Presence Penalty0

Frequency Penalty0

ToolsTools auto-selected b + Add

Seed

JSON mode

Stop Sequences+ Add stop sequence

Provider Options

1.61s 800 Annotate 16 minutes ago

Start

Level 4

Multimodal

Input

HUMAN ↕

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.

Question: what is the payload I can carry with the Falcon 9 rocket to Mars?

Context:

CAPABILITIES & SERVICES

SpaceX offers competitive pricing for the Falcon 9 and Falcon Heavy launch services. Based on the cost of the rocket and the cost of the payload, the price of the launch is determined. The price of the launch is determined by the weight of the payload and the destination. The price of the launch is determined by the weight of the payload and the destination.

PRICE*	FALCON 9
12,000,000 USD (through 2023)	\$49.75 M (2024-2025)

DESTINATION	
LOW EARTH ORBIT (LEO)	22,000 kg (48,500 lb)
MEDIUM EARTH ORBIT (MEO)	8,300 kg (18,300 lb)
GEOSTATIONARY TRANSFER ORBIT (GTO)	4,300 kg (9,500 lb)



Answer:

Provider

ChatOpenAI

Model

gpt-4o

Temperature 0

Maximum Length -1

Top P 1

Presence Penalty 0

Frequency Penalty 0

Tools Tools auto-selected b + Add

Seed

JSON mode

Stop Sequences + Add stop sequence

Provider Options

May 13, 2024

Hello GPT-4o

We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.

[Contributions >](#) [Try on ChatGPT ↗](#) [Try in Playground ↗](#) [Rewatch live demos >](#)

SpaceX requires that customers verify the mass properties of their system through measurement before shipping it to the launch site. SpaceX may request insight into

1.61s 800 Annotate 16 minutes ago

Start

Level 4
Multimodal

LangSmith

smith.langchain.com/o/5cd16e97-917d-524a-8344-f8c0c526e1b9/playground/r/ec4e011c-f620-442e-ad8e-f43688eaa4c8

Personal < Back to ChatOpenAI Run

DEVELOPER

Secrets & API Keys

History

Editing ChatOpenAI

Input

Output

ChatOpenAI got 40.78

HUMAN

You are an assistant for ques
Use the following pieces of re
answer the question. If you d
just say that you don't know.
maximum and keep the answ
Question: what is the payload
Falcon 9 rocket to Mars?
Context:

CAPABILITIES & SERVICES

PRICE*

FALCON 9

DESTINATION

AI

The payload capacity of the Falcon 9 rocket to Mars is 4,020 kg (8,860 lbs).

Top P

1

Presence Penalty

0

Answer:

+ Message

142,008 prompt tokens / \$0.71004

23 completion tokens / \$0.000345

Time to first token: 2487ms

GPT-4o with picture input

2.92s

142,031

Annotate

25 minutes ago

2000x increase in
input prompt cost

GPT 3.5 with only text input

738 prompt tokens / \$0.000369

62 completion tokens / \$0.000093

Time to first token: N/A

1.61s

800

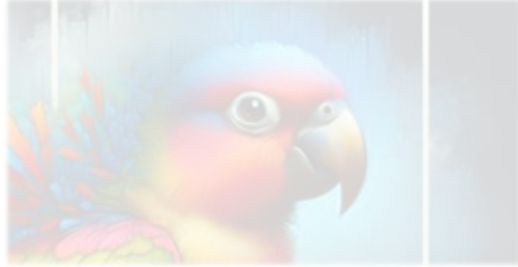
Annotate

43 minutes ago

The Levels of **RAG**



Level 1
Basic RAG



Level 2
Hybrid search



Level 3
Advanced data formats



Level 4
Multimodal

The Levels of RAG



Level 1
Basic RAG

- **Retrieval, Augmentation & Generation**
- **RAG components** : Data- collection and parsing, Chunking, Embedding, Vector database, Large Language Model (LLM)
- **Vector search**
- **Chunking**



Level 2
Hybrid search

- **Combining** vector search and keyword search.
- Sparse text search using: TF-IDF and **BM-25**
- **Reciprocal Rank Fusion**



Level 3
Advanced data formats

- HTML, Word, **PDF**
- PDF with native text but also images
- Jumbled text
- **Tables**: using Computer Vision and Markdown



Level 4
Multimodal

- Multimodal for audio, images and even video.
- Images in prompt can be very powerful. Beware of cost.

The Levels of RAG



Level 1 Basic RAG

- RAG consists out of a **Retrieval** and a **Generation** step
- The main **RAG components** are: Data- collection and parsing step, Chunking, an Embedding model, a Vector database and a Large Language Model (LLM)
- **Vector search** allows us to find source information familiar to the user query.
- **Chunking** comes in to split long pieces of text and get accurate answer citations.
- ... but what if we miss **obvious** matches? Is Vector search enough?



Level 2 Hybrid search

- Going hybrid: combining vector search and keyword search.
- Popular algorithms for text search are: TF-IDF and **BM-25**
- **Reciprocal Rank Fusion** allows us to combine keyword search and vector search.
- This gives us better performing search to prevent missing obvious matches.
- ... but now what if we have **more complicated document types**?



Level 3 Advanced data formats

- Formats like HTML, Word but especially **PDF** can lack structure and are harder to parse.
- Luckily many **PDF parsers** exist. Most importantly, sometimes we need **OCR** to read the PDF's when native text is lacking.
- However, **be cautious** when parsing PDF's: text can be presented to the LLM in **jumbled** fashion, causing the LLM to *convincingly* give *wrong* answers.
- Answering originating from **tables** can be hard to answer with in your RAG system. We can give them a special treatment by parsing them separately and inserting them as Markdown in the prompt.
- ... another way to parse tables and much more is to go **Multimodal**.



Level 4 Multimodal

- Multimodal models are around. They process not only text but also audio, images and even video.
- Images can be inserted in the prompt and presented to the LLM, opening up new possibilities. Beware of the cost.
- Good luck with building your RAG system! 🍀



Thank you!



Jeroen Overschie

Machine Learning Engineer

jeroen.overschie@xebia.com

