goto;

DOWNLOAD & USE
THE GOTO GUIDE APP

DOWNLOAD THE APP
CREATE YOUR SCHEDULE
ASK QUESTIONS
RATE SESSIONS

Download on the App Store

GET IT ON Google Play

goto;

GOTO CPH 2024

Building Performant RAG Applications for Production

David Carlos Zachariae
Trifork A/S

Who am I ?

Name is David Carlos Zachariae

Master's Degree in Computer Science from Aarhus University in 2019 specializing in ML/NLP

Live in Aarhus with my girlfriend and my dog Ruby

Has worked at Trifork for ~5 years as a Software Pilot

Agenda.

- Motivation and use-case
- First iteration - The simple case
- Second iteration - Multiple categories of documentation
- Third iteration - Unstructured documentation
- Fourth iteration - Dynamic context and actions

Agenda.

- Motivation and use-case
- First iteration - The simple case
- Second iteration - Multiple categories of documentation
- Third iteration - Unstructured documentation
- Fourth iteration - Dynamic context and actions

Why use
Retrieval Augmented Generation ?

Phases of RAG .

Indexing

Index a knowledge base of documents using embedding models and vector databases

Phases of RAG .

Indexing

Index a knowledge base of documents using embedding models and vector databases

Retrieval

Embed query and use vector search to retrieve relevant docs

Phases of RAG .

Indexing

Index a knowledge base of documents using embedding models and vector databases

Retrieval

Embed query and use vector search to retrieve relevant docs

Generation

Generate answer based on *query* and *context* from retrieved docs

Performant RAG ?

The use-case .

Automatic generation of drafts to answer support tickets for **Fake A/S**

Agenda.

- Motivation and use-case
- First iteration - The simple case
- Second iteration - Multiple categories of documentation
- Third iteration - Unstructured documentation
- Fourth iteration - Dynamic context and actions

First iteration .

Handles a **single** product: *Fake Product*

Tickets are short and contain **one question** each

One document for the product with a list of **questions and appropriate answers**

Documentation .

Support Guide

Q: What is the Fake Product?

A: The Fake Product is a revolutionary solution that will change your life. It combines cutting-edge technology with unparalleled functionality to deliver an exceptional user experience.

Q: How does the Fake Product work?

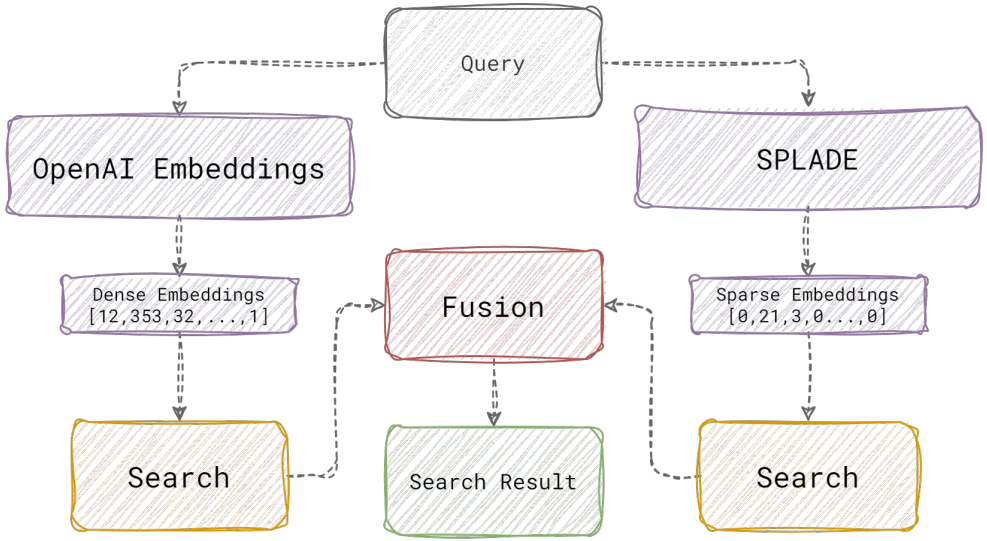
A: The Fake Product utilizes advanced algorithms and state-of-the-art hardware to perform its magic. It seamlessly integrates with your existing systems and processes, making it easy to incorporate into your workflow.

Hybrid Search .

Mixing different search approaches

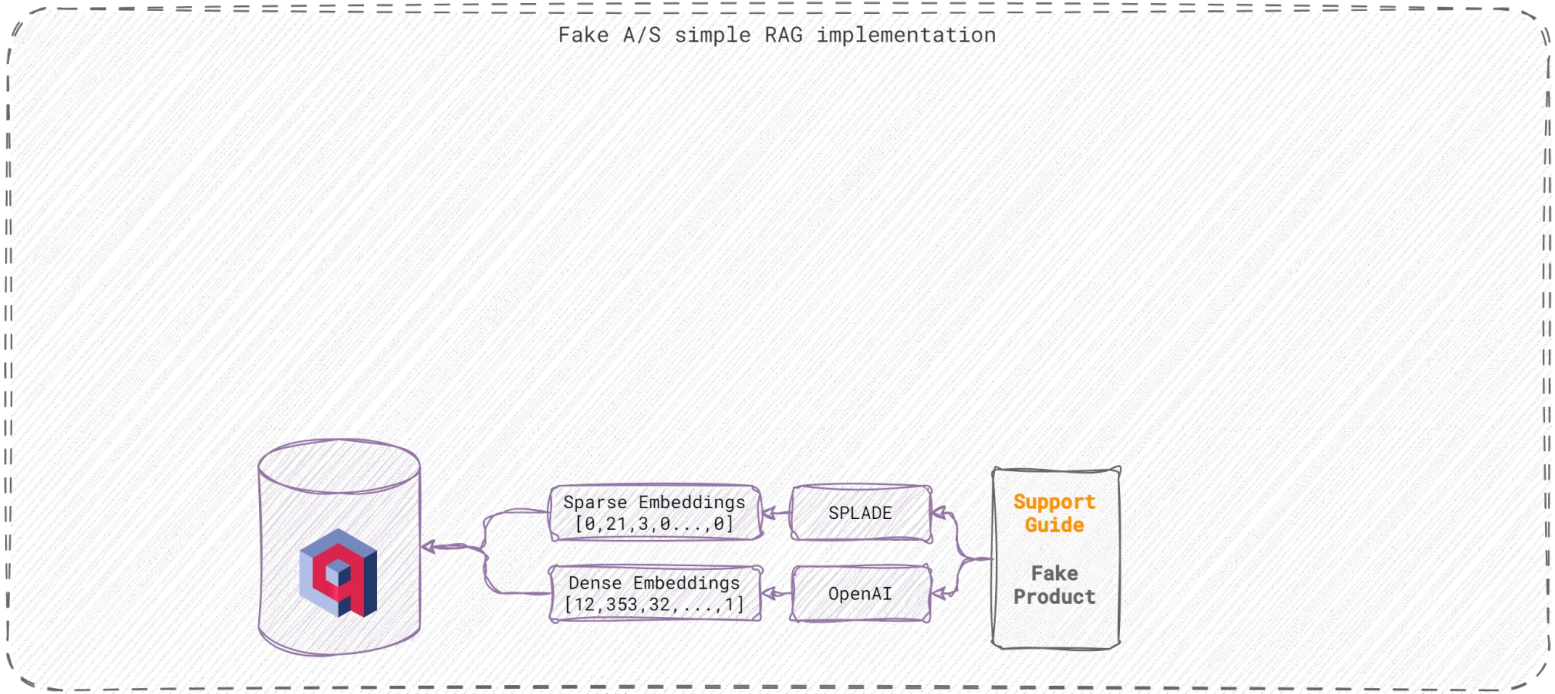
Vector Search using embedding models like OpenAI

Keyword Search using sparse vectors for capturing keywords (BM25, SPLADE)

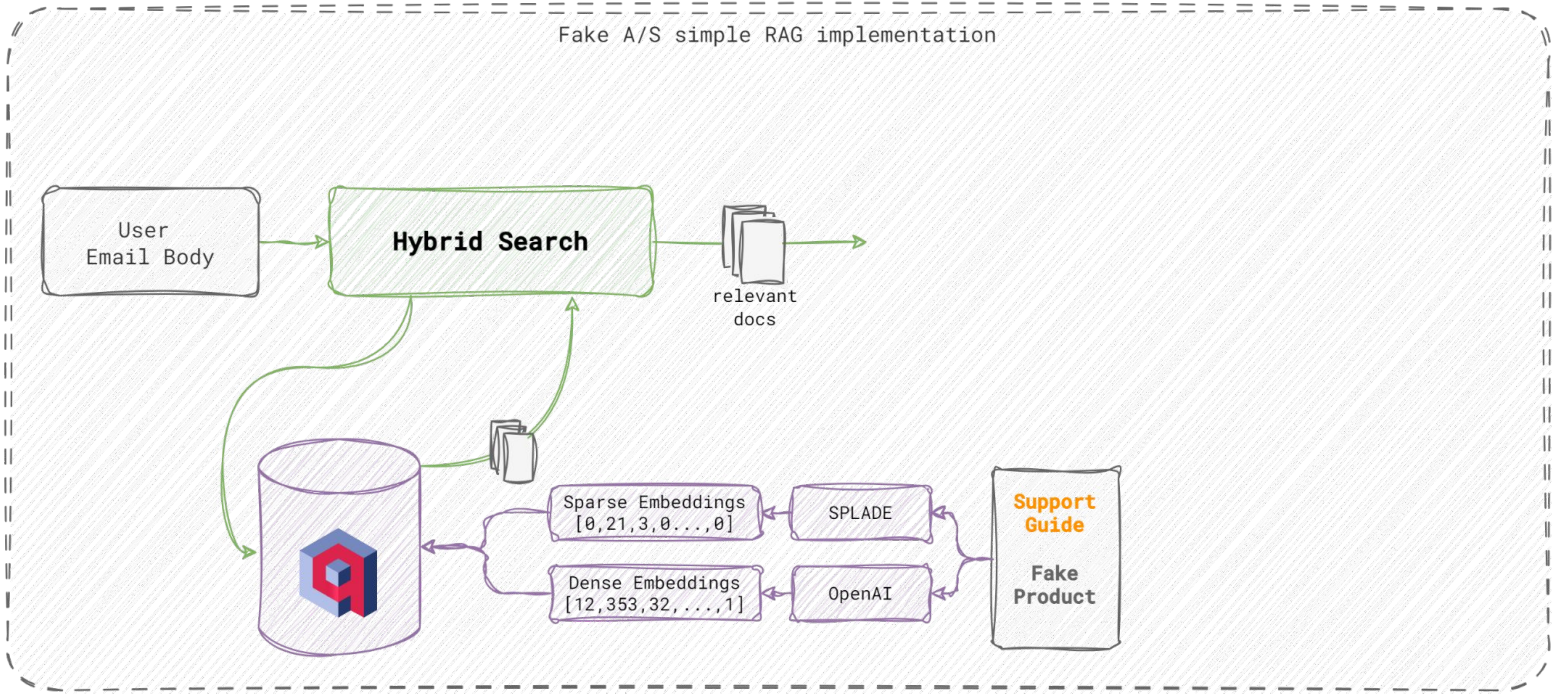


$$\text{hybrid\_score} = (1-\alpha) \times \text{sparse\_score} + \alpha \times \text{dense\_score}$$

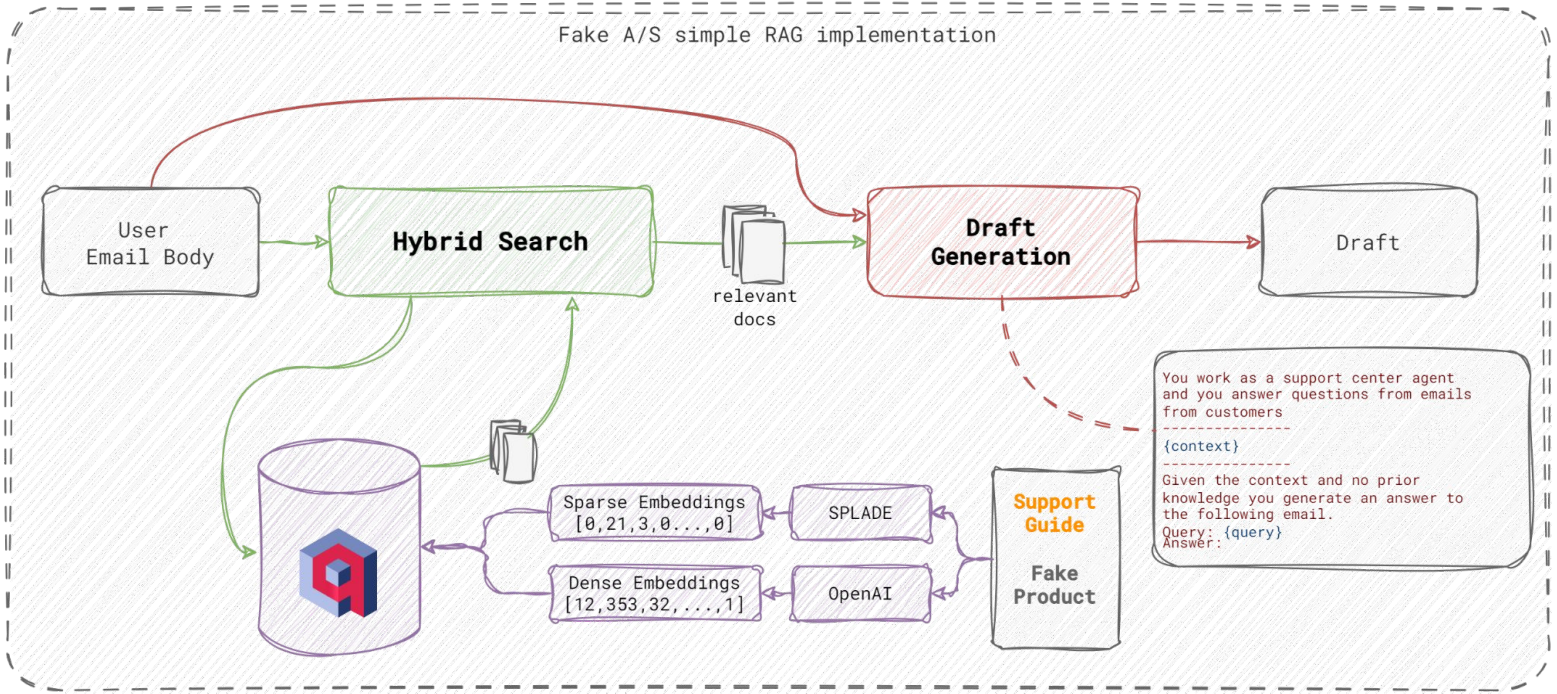
Applied to use case .



Applied to use case .



Applied to use case .



Agenda.

Motivation and use-case

First iteration - The simple case

Second iteration - Multiple categories of documentation

Third iteration - Unstructured documentation

Fourth iteration - Dynamic context and actions

Second iteration .

Handles **multiple** products:
Fake Product, Fake Product 2.0, Fake Widget
Fake App

Tickets may contain **multiple questions**
and are **not presorted**

Questions and answers may be **similar between products, but slightly different**

Documentation .

Support Guide - Fake Product

```
## Q: What are the pricing plans?
A: The Fake Product offers flexible pricing plans. The pricing options
include:

- **Basic Plan**: $39 per month. Ideal for small businesses and startups.
Includes essential features and support.

...
```

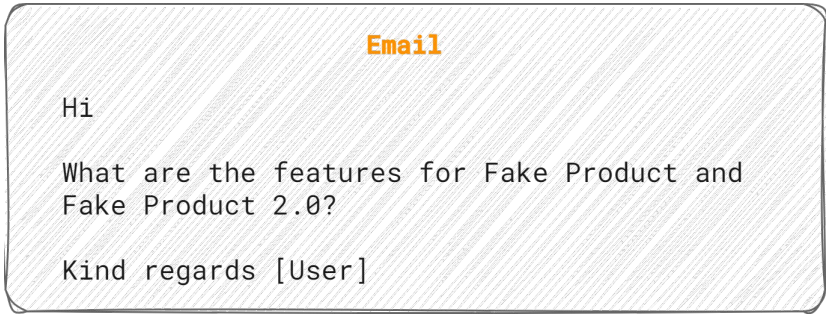
Support Guide - Fake Product 2.0

```
## Q: What are the pricing plans?
A: The Fake Product offers flexible pricing plans. The pricing options
include:

- **Basic Plan**: $49 per month. Ideal for small businesses and startups.
Includes essential features and support.

...
```

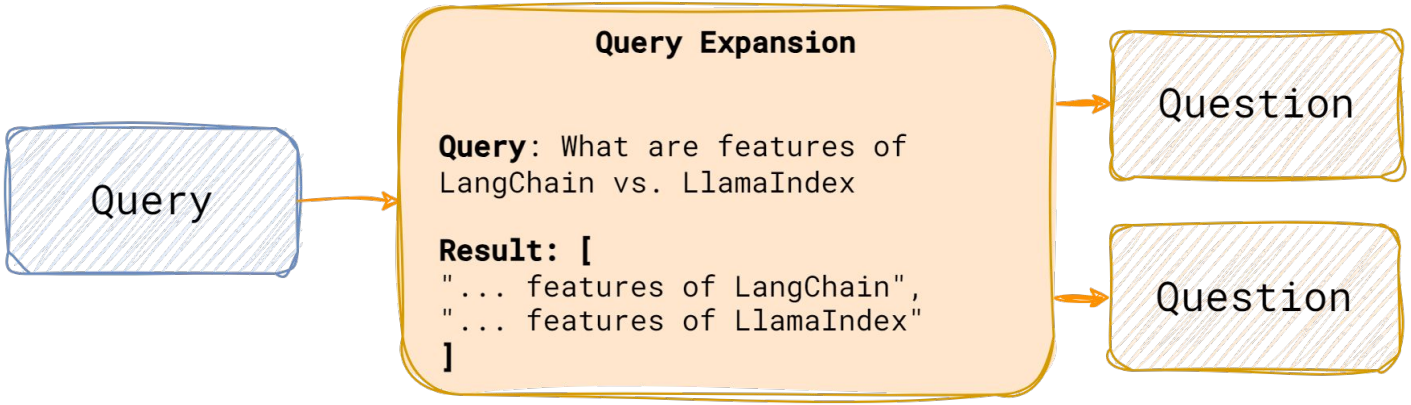
Tickets .

Tickets may contain multiple questions about different products

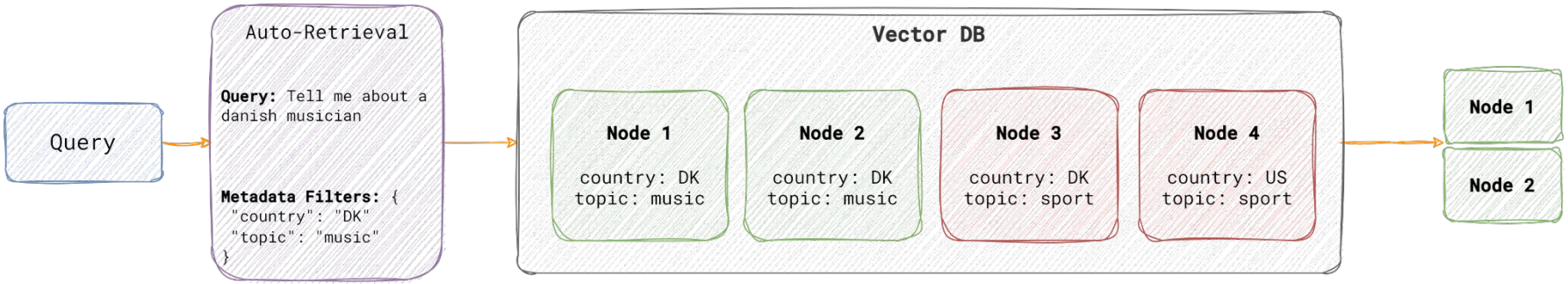
Tickets are not sorted into categories



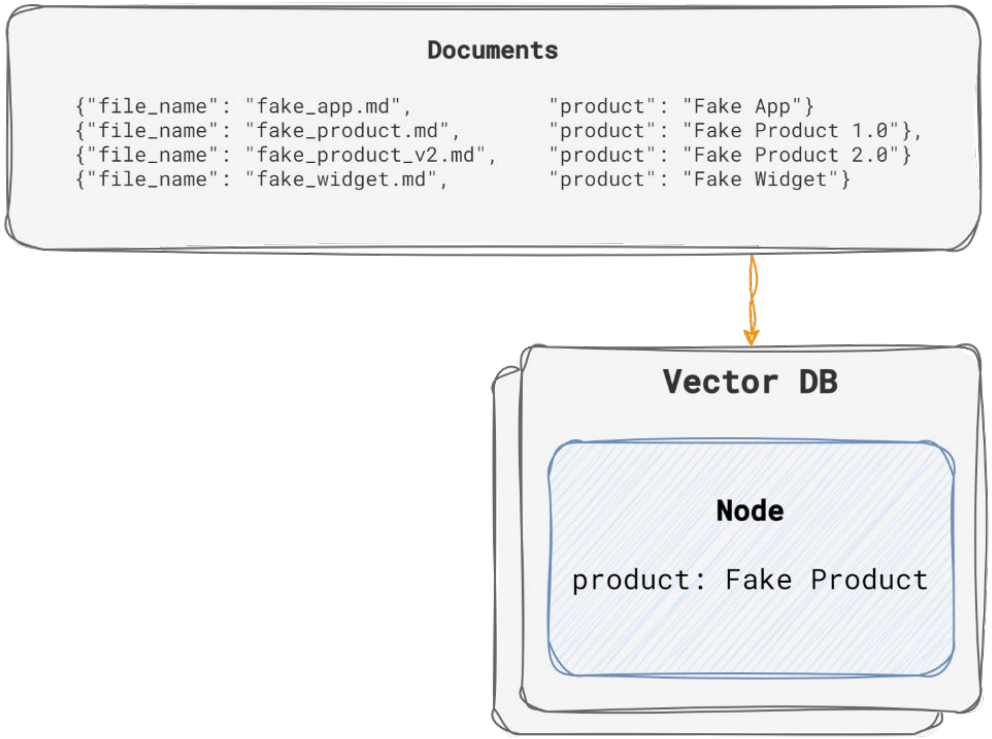
Query Expansion - Divide and Conquer .



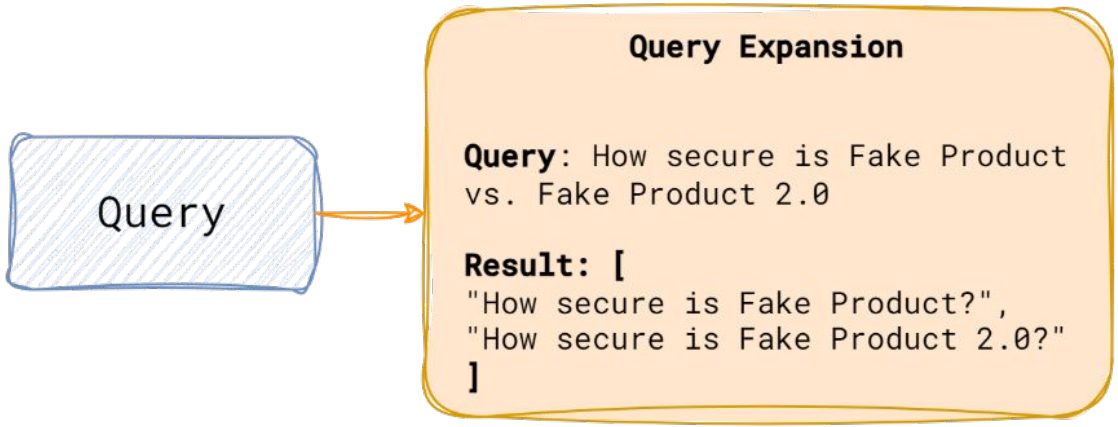
Metadata filters and Auto Retrieval



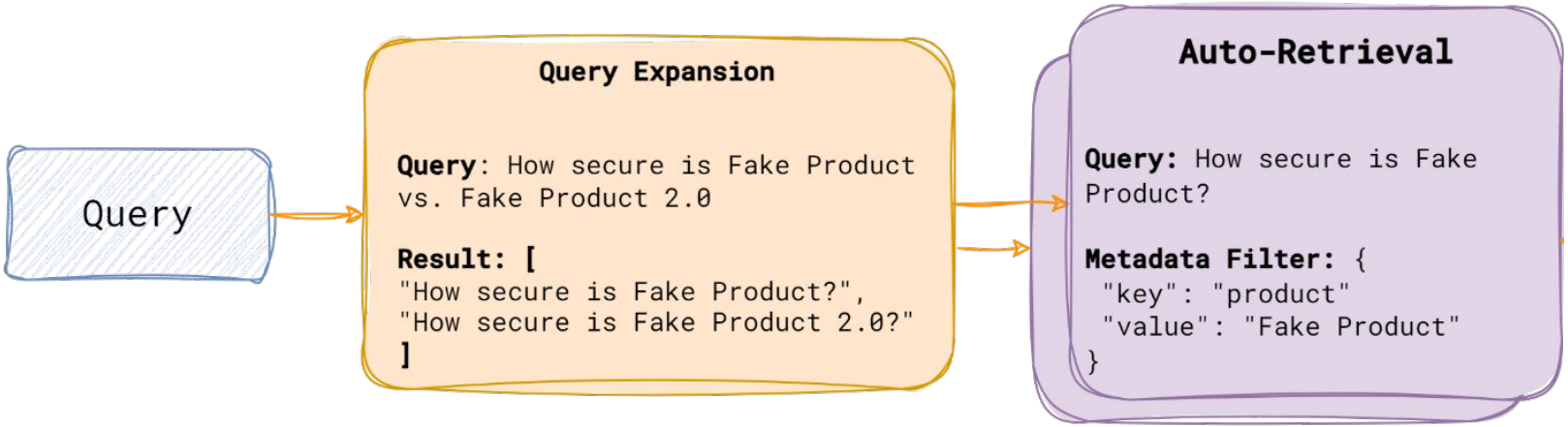
Applied to use-case - Indexing



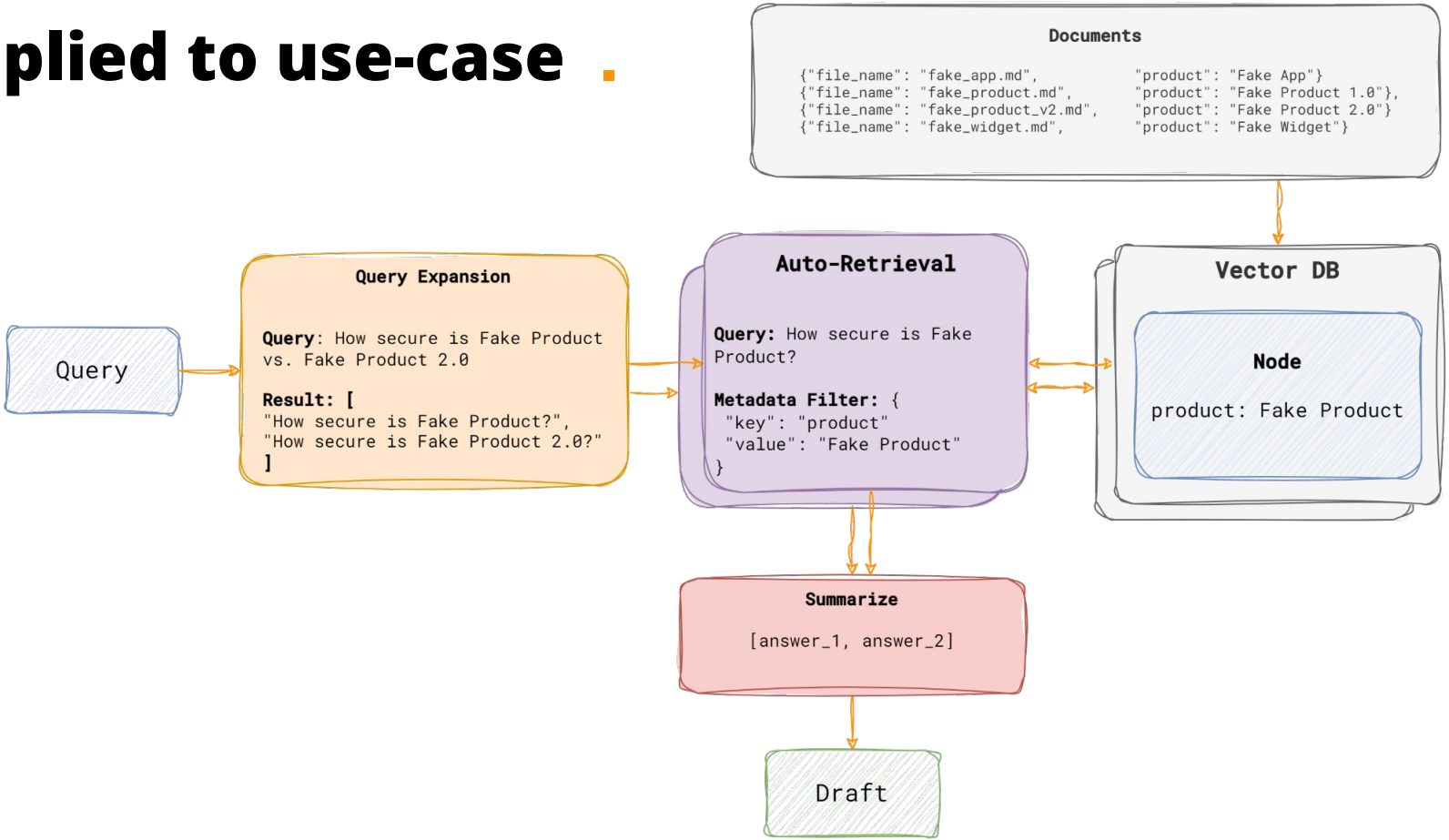
Applied to use-case - Retrieval .



Applied to use-case - Retrieval .



Applied to use-case .



Agenda.

Motivation and use-case

First iteration - The simple case

Second iteration - Multiple categories of documentation

Third iteration - Unstructured documentation

Fourth iteration - Dynamic context and actions

Third iteration .

Handles a single product: *Fake Product*

Tickets may contain **multiple questions** and are not presorted

Documentation is **not organized** into questions and answers

Documentation .

Not questions and answers

Questions are not present

Sections may answer multiple different questions

Support Guide

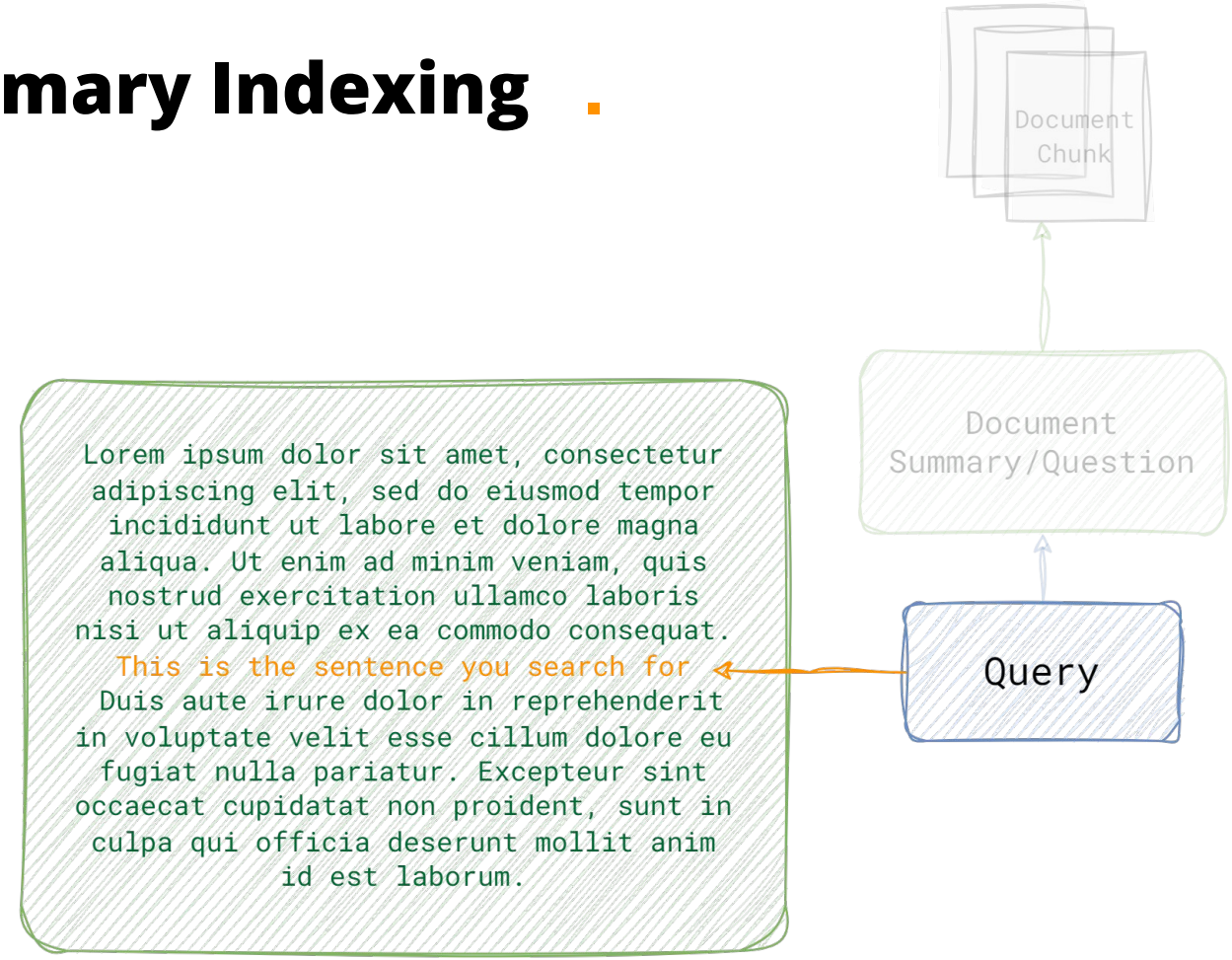
```
# Comprehensive Supporter Guide for Fake Product

Hello, amazing Support Team!

Welcome to our expanded guide for handling email tickets about the
Fake Product...

...

## 1. General Information

When customers ask about the Fake Product, you can say:

"Thank you for your interest in the Fake Product! We're excited to
tell you more about this game-changing solution..."

Key points to highlight:
- Innovative solution for many different industries
- ...

For a deep dive into what makes the Fake Product special, check
out our detailed product page at [fake-product-url]. We're
constantly updating this information, so it's a great resource to
bookmark!"

...
```

Question/Summary Indexing

Decouple documents for Retrieval vs. documents used for Generation

Question/Summary Indexing .

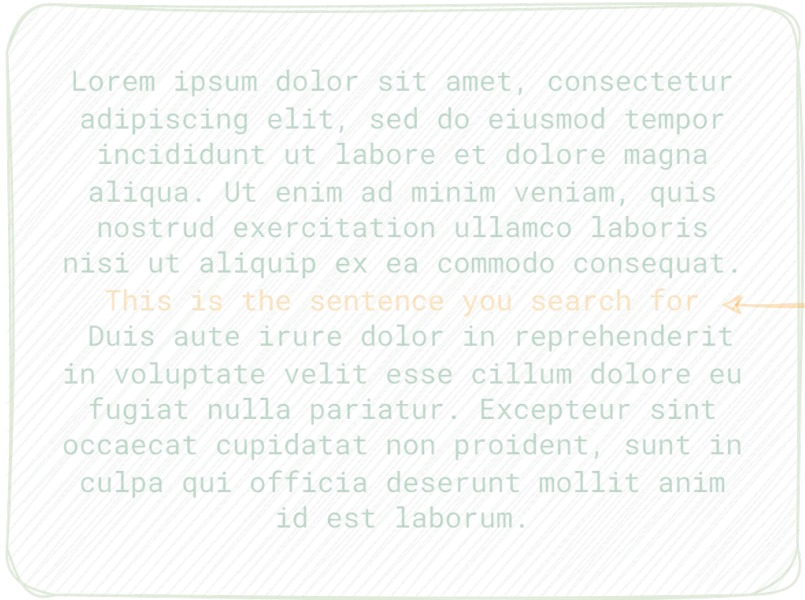
- Sentence Window

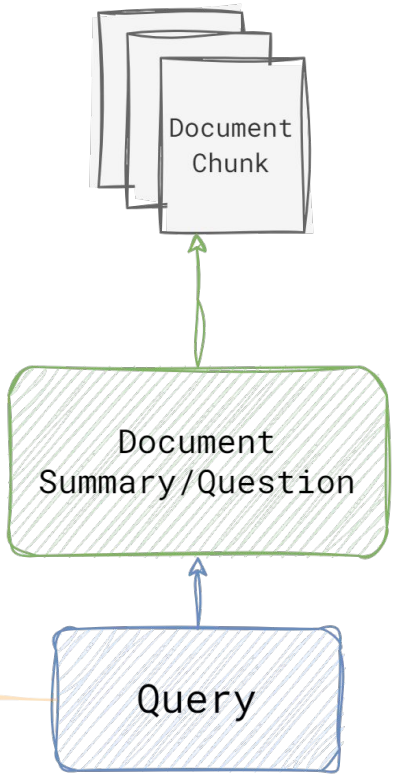


Question/Summary Indexing .

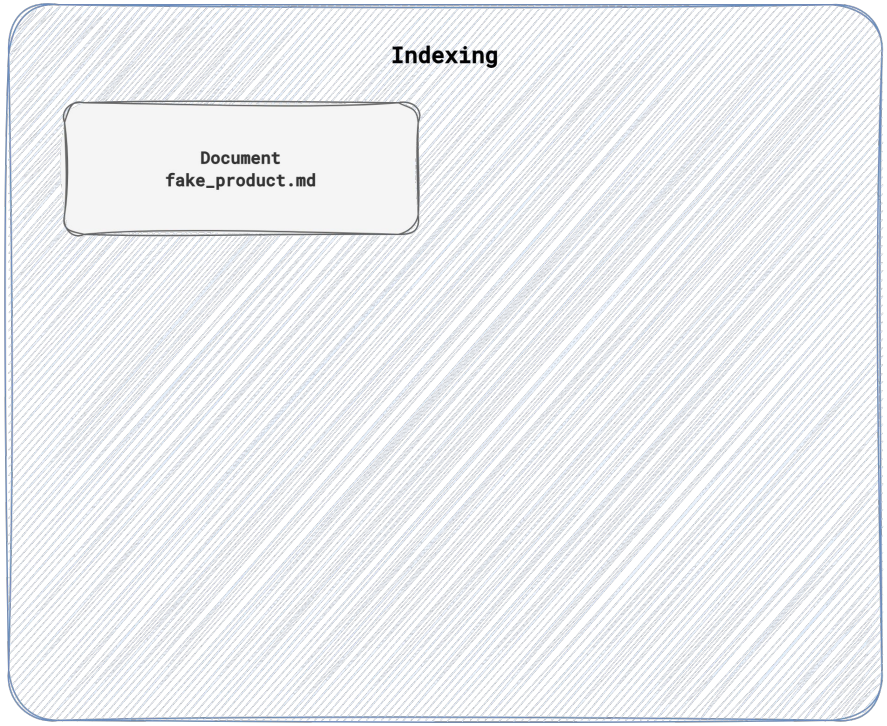
- Sentence Window

- Summary Index
- Question Index

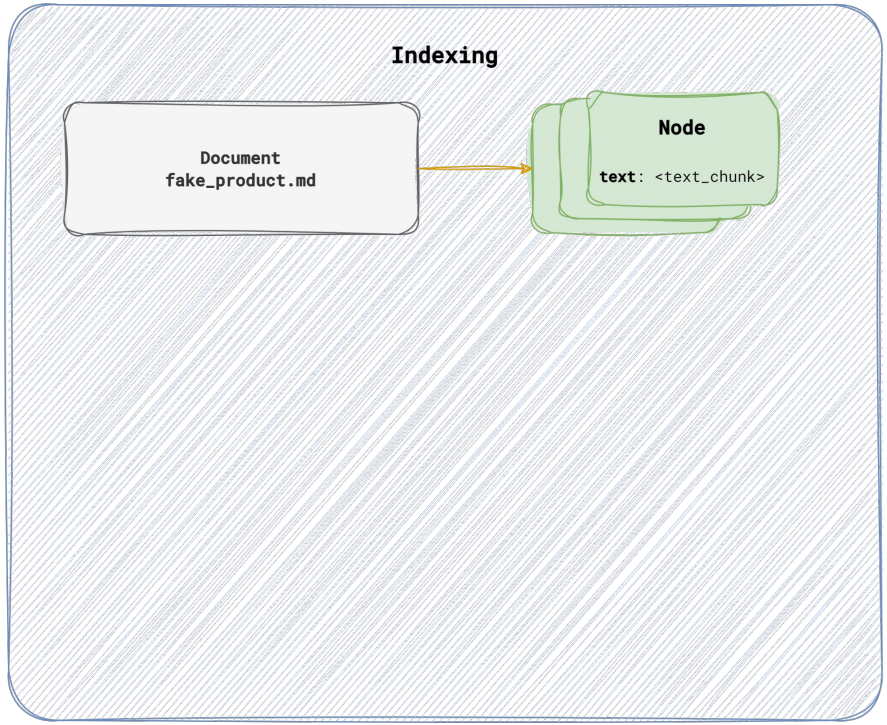




Applied to use-case - Question Indexing



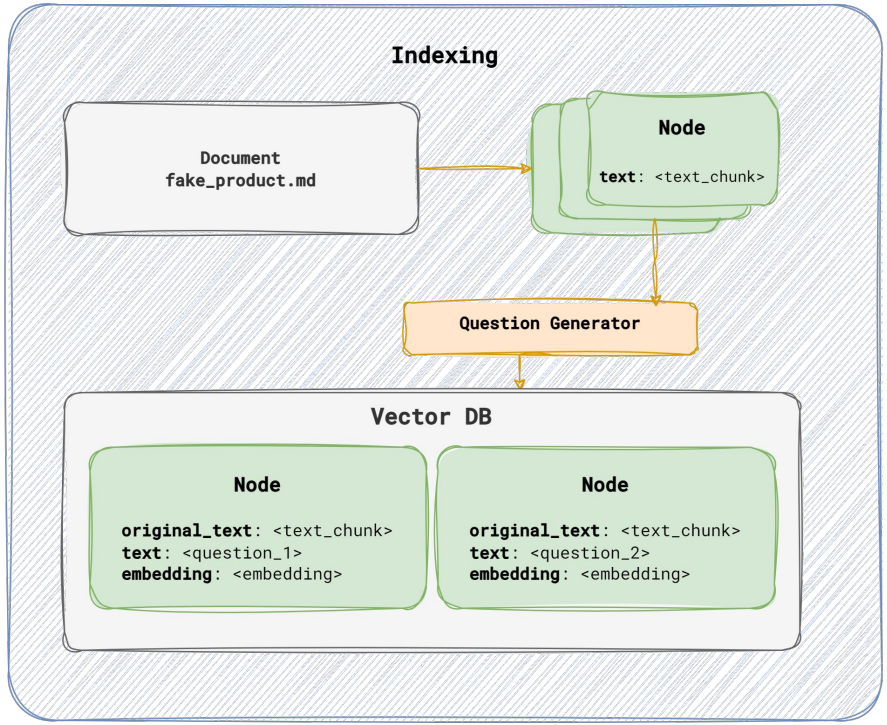
Applied to use-case - Question Indexing



Indexing

Fetch the chunks of text split using the sections in markdown

Applied to use-case - Question Indexing



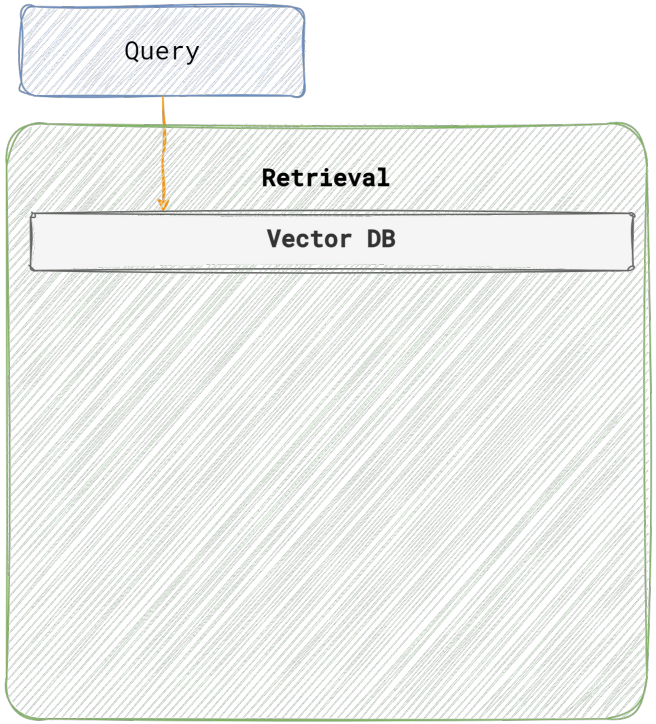
Indexing

Generate questions that are answered by the given chunk of text

Embed each question with a reference to the original text

Applied to use-case - Question Indexing

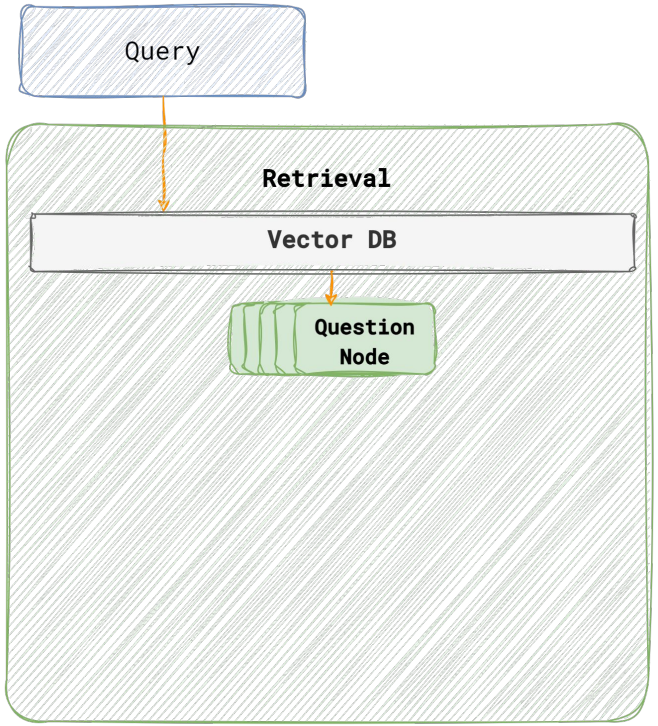
Retrieval



Applied to use-case - Question Indexing

Retrieval

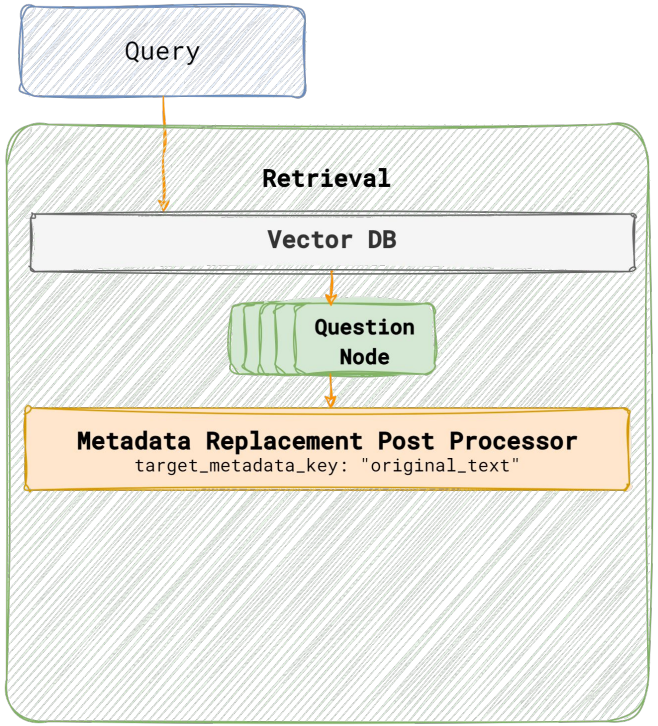
Use vector search to fetch questions



Applied to use-case - Question Indexing

Retrieval

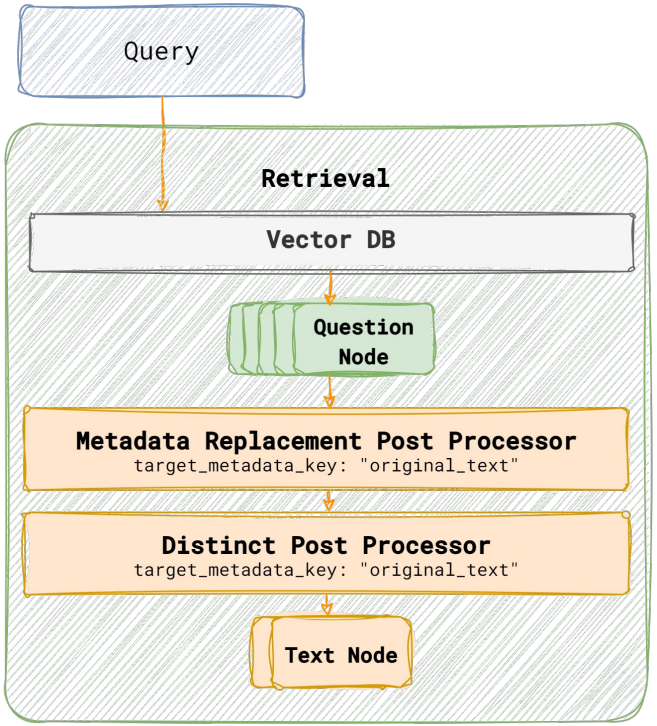
Replace the text with the original text



Applied to use-case - Question Indexing

Retrieval

Fetch each distinct text as multiple questions might be relevant



Agenda.

- Motivation and use-case
- First iteration - The simple case
- Second iteration - Multiple categories of documentation
- Third iteration - Unstructured documentation
- Fourth iteration - Dynamic context and actions

Fourth iteration .

Handles a single product: *Fake Product*

Tickets may contain **multiple questions**

Answers may involve **context that is specific** to the user

Fake A/S don't want the AI to handle **angry customers**

Dynamic Context .

Email

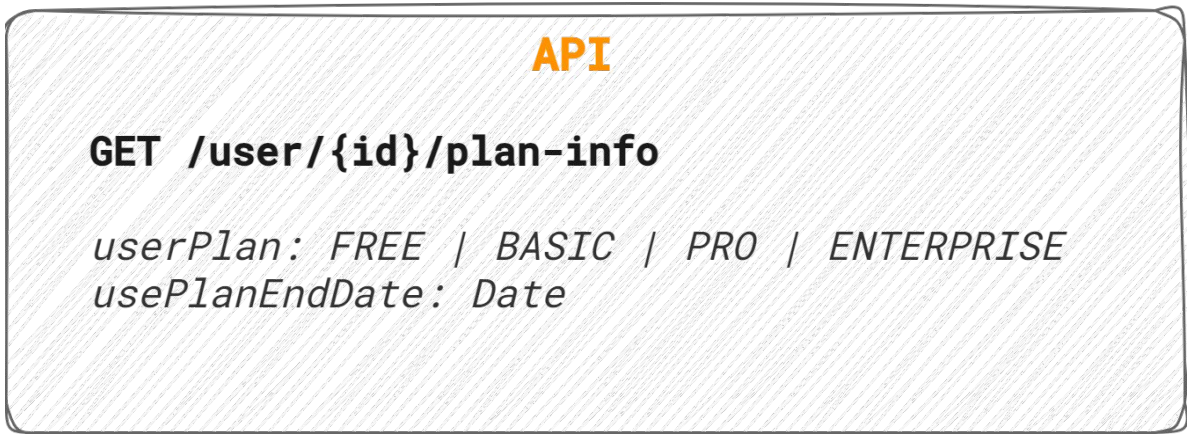
From: user123

Hi

When does my subscription end and what are the main features of my plan?

Kind regards [User]

API.


API
GET /user/{id}/plan-info
userPlan: FREE | BASIC | PRO | ENTERPRISE
usePlanEndDate: Date

Query Routing .

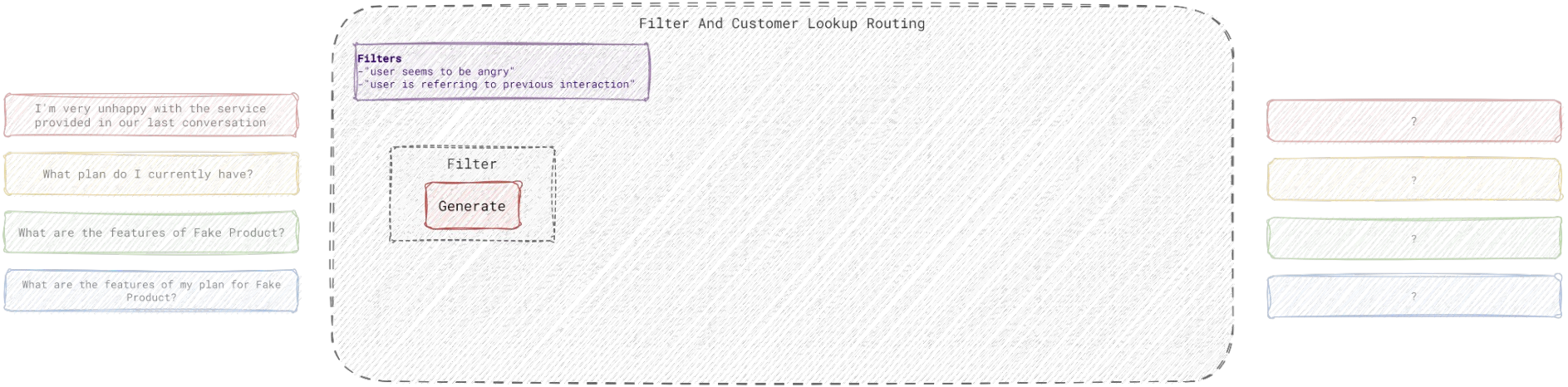
Make your RAG pipeline more **adaptable** to different situations

Add **dynamic context** to the answer when needed

Applied to use-case - Query Routing

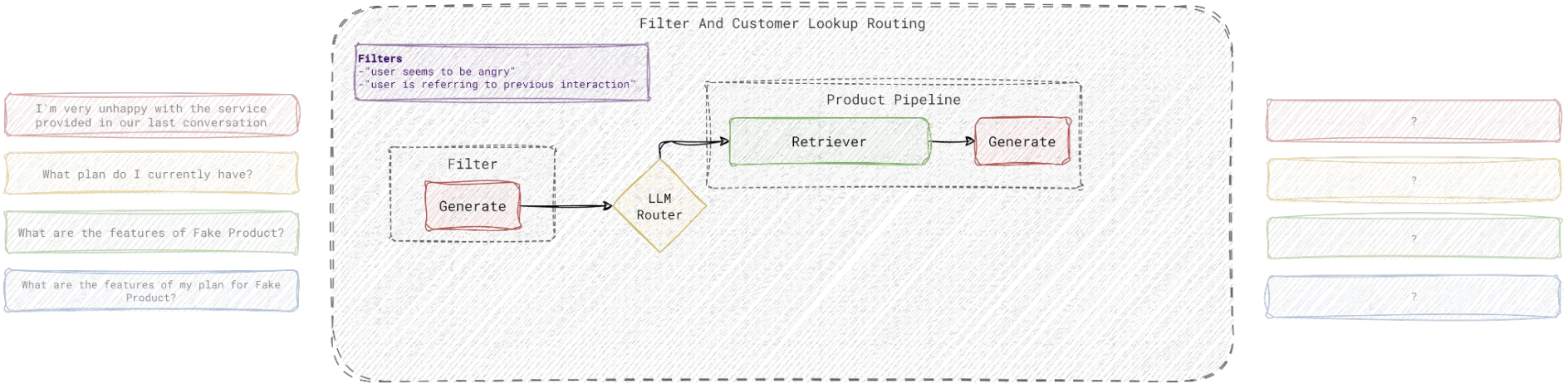
"User seems to be angry"

"User is referring to a previous interaction"



Applied to use-case - Query Routing

"Useful for answering questions about products"



Applied to use-case - Query Routing

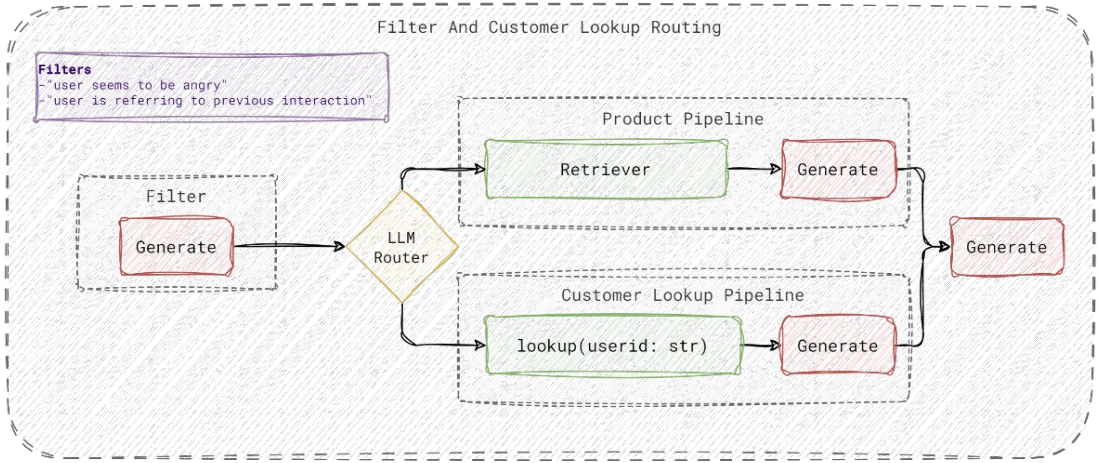
“Useful for answering questions about specific customers with their userId”

I'm very unhappy with the service provided in our last conversation

What plan do I currently have?

What are the features of Fake Product?

What are the features of my plan for Fake Product?



?

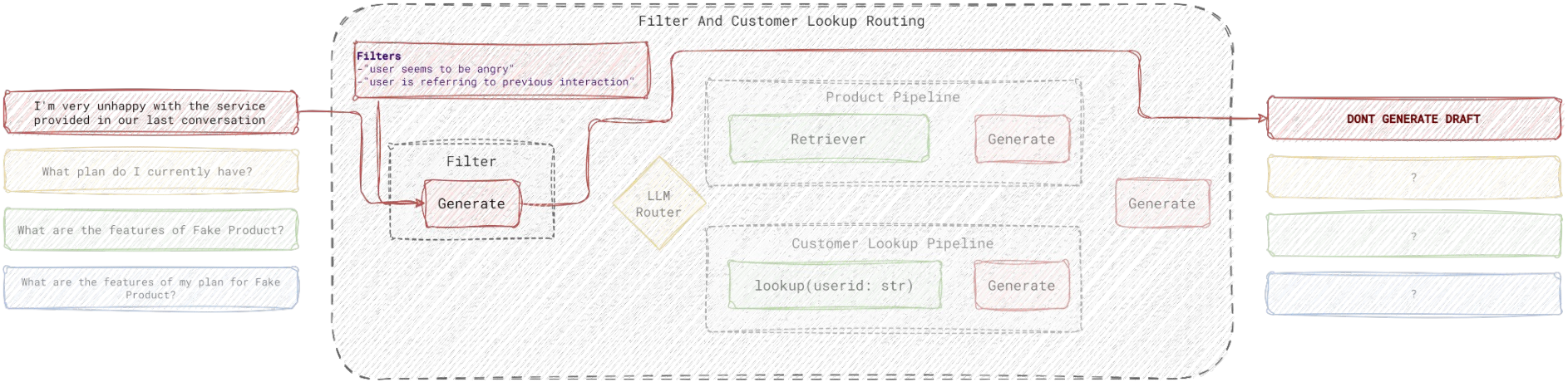
?

?

?

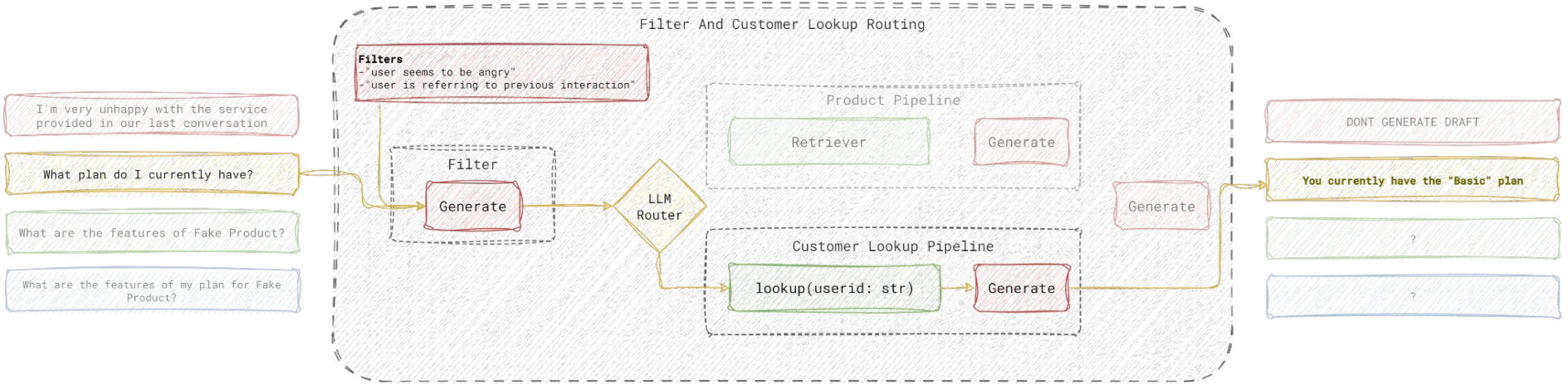
Applied to use-case - Query Routing

"I'm very unhappy about your service!"



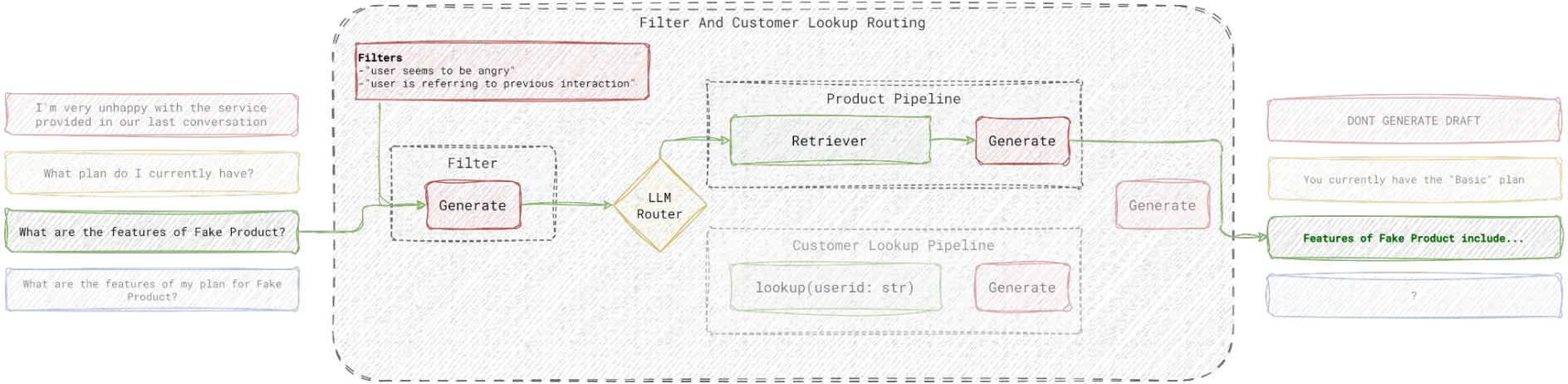
Applied to use-case - Query Routing

"**user123**: *What is my current plan?*"



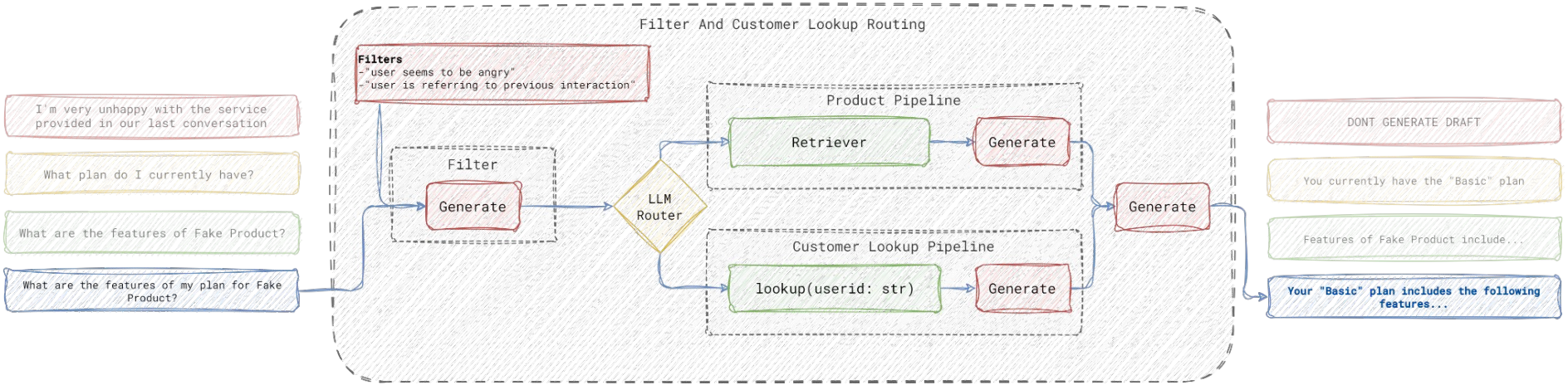
Applied to use-case - Query Routing

"What are the features of Fake Product?"



Applied to use-case - Query Routing

"What are the features of Fake Product given my current plan?"



Take-aways .

Start simple - Base-architecture

Divide and Conquer - Query Expansion, Metadata filters

Decouple retrieval and generation - Question Indexing

Make it Adaptable - Query Routing

Take-aways .

Many more techniques!

GraphRAG, HyDE, Reranking

Fine-tuning LLMs, SelfRAG

...

goto;

Rate this session in the GOTO Guide app
and claim your reward