Retrieval-Augmented Everything: RAG in Practice



Good for Chat



Good for Coding



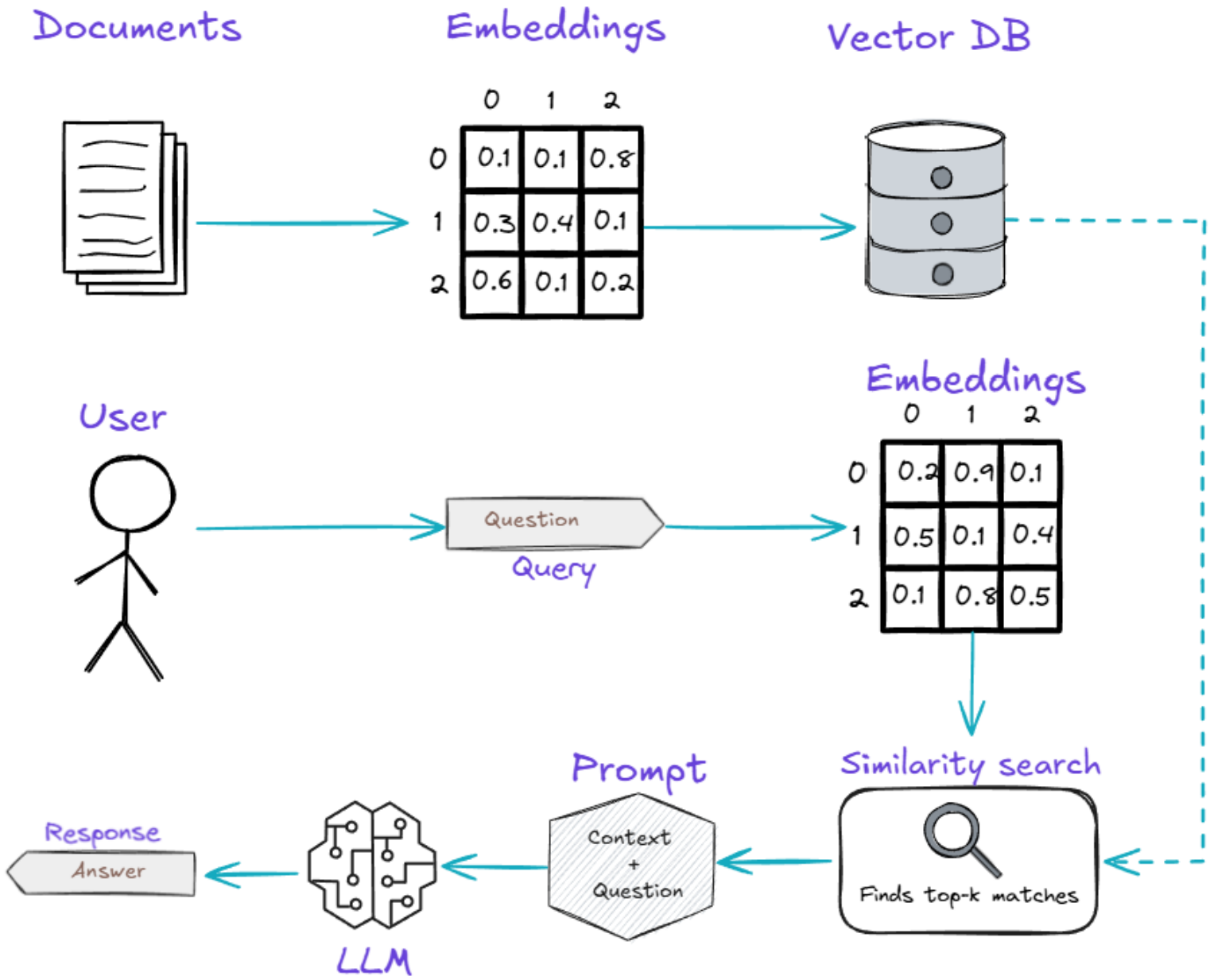
Good Morning

Grounding Strategies: Reducing LLM Hallucination

- Retrieval-Augmented Generation (RAG)
- Tool use / function calling
- Calibration and fine-tuning
- …

Typical RAG Architecture

DOI:10.1007/s12599-025-00945-3

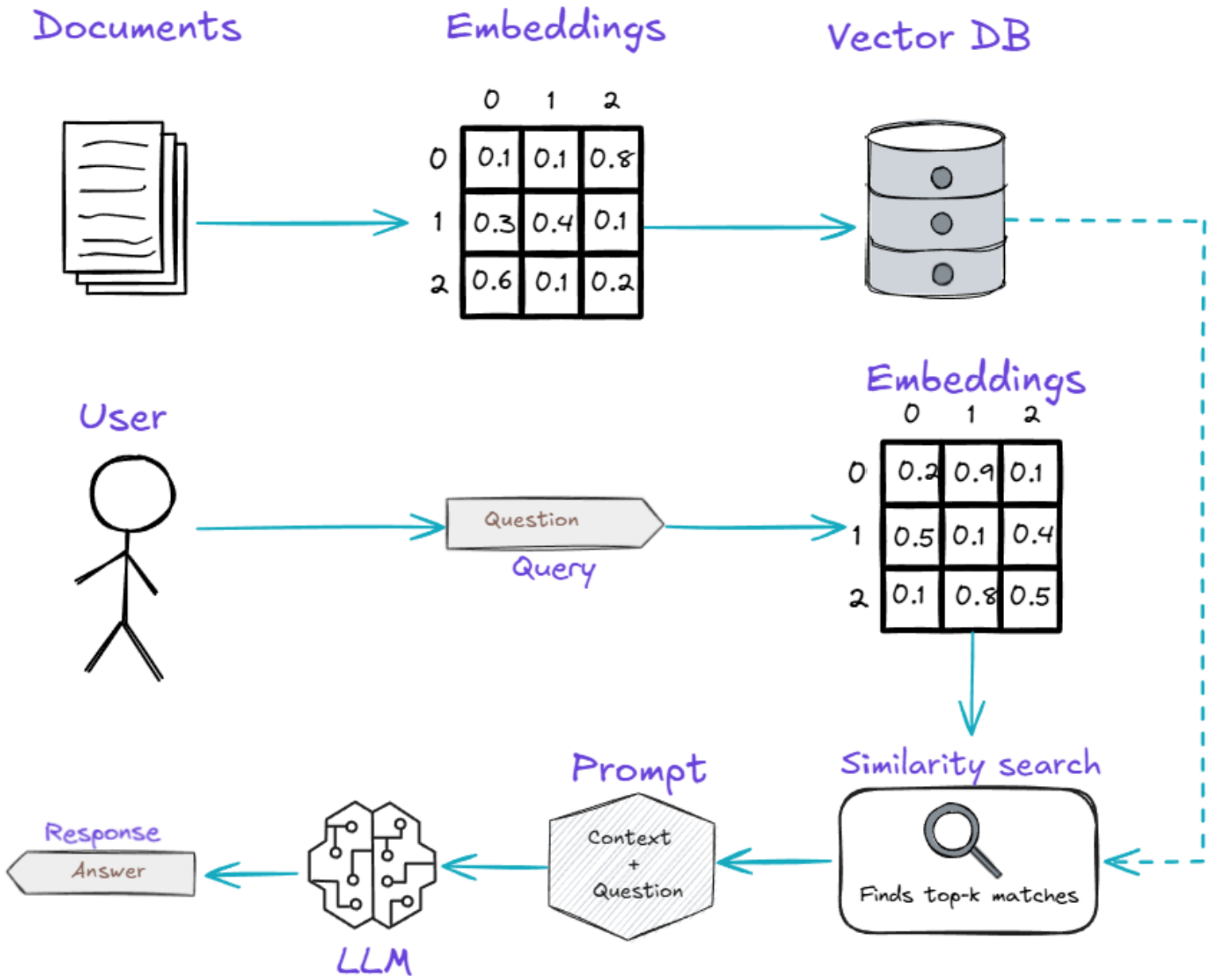


Simplified embeddings matrix

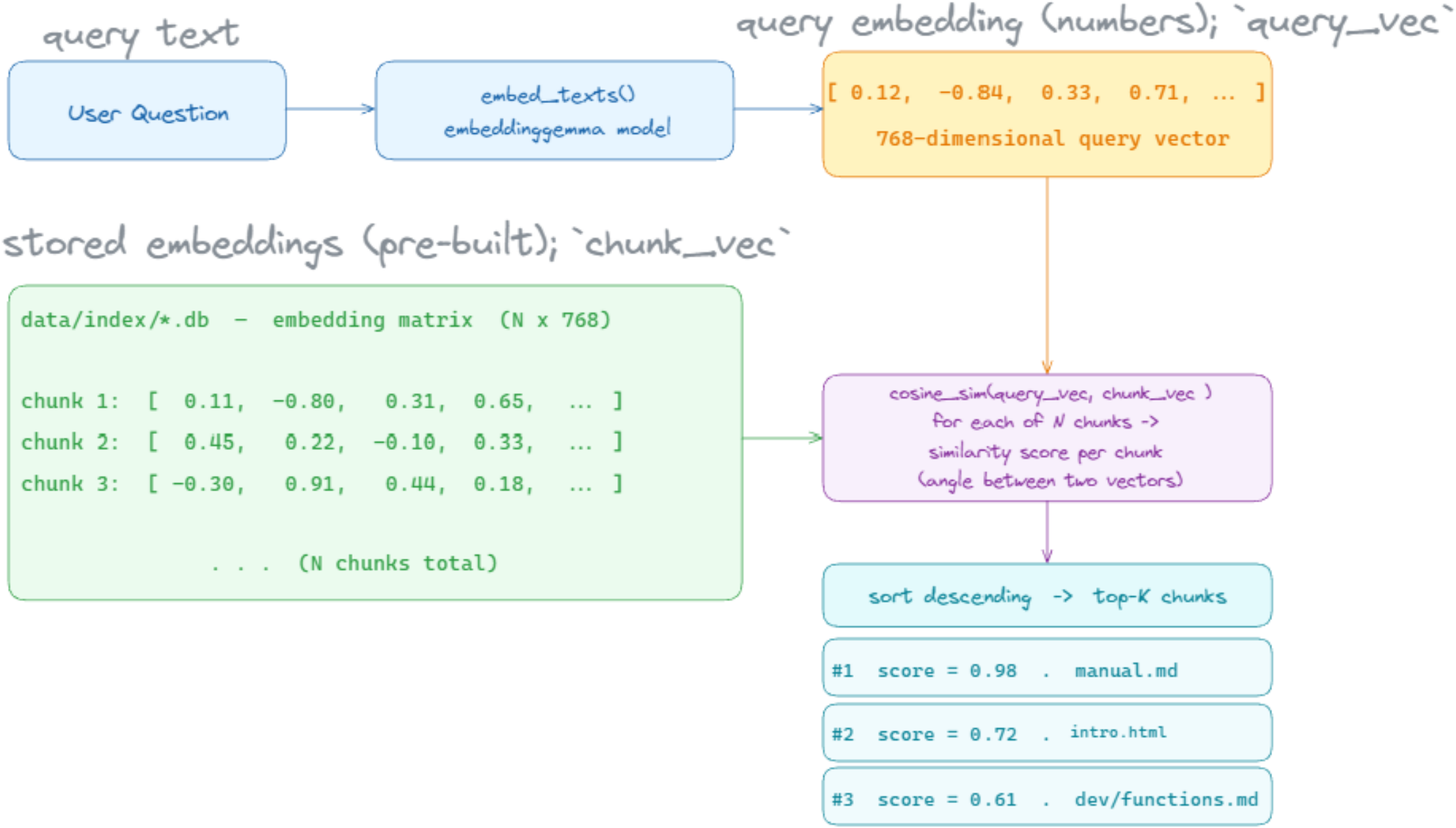
Text	Trait 1: Is it a treatment?	Trait 2: Is it a recruitment?	Trait 3: Is it a futility rule?
"Experimental arm"	0.97	0.03	0.01
"Control arm"	0.95	-0.3	0.02
"Accrual rate"	0.02	0.96	0.05
"Stop for futility"	0.02	-0.6	0.98

Typical RAG Architecture

DOI:10.1007/s12599-025-00945-3

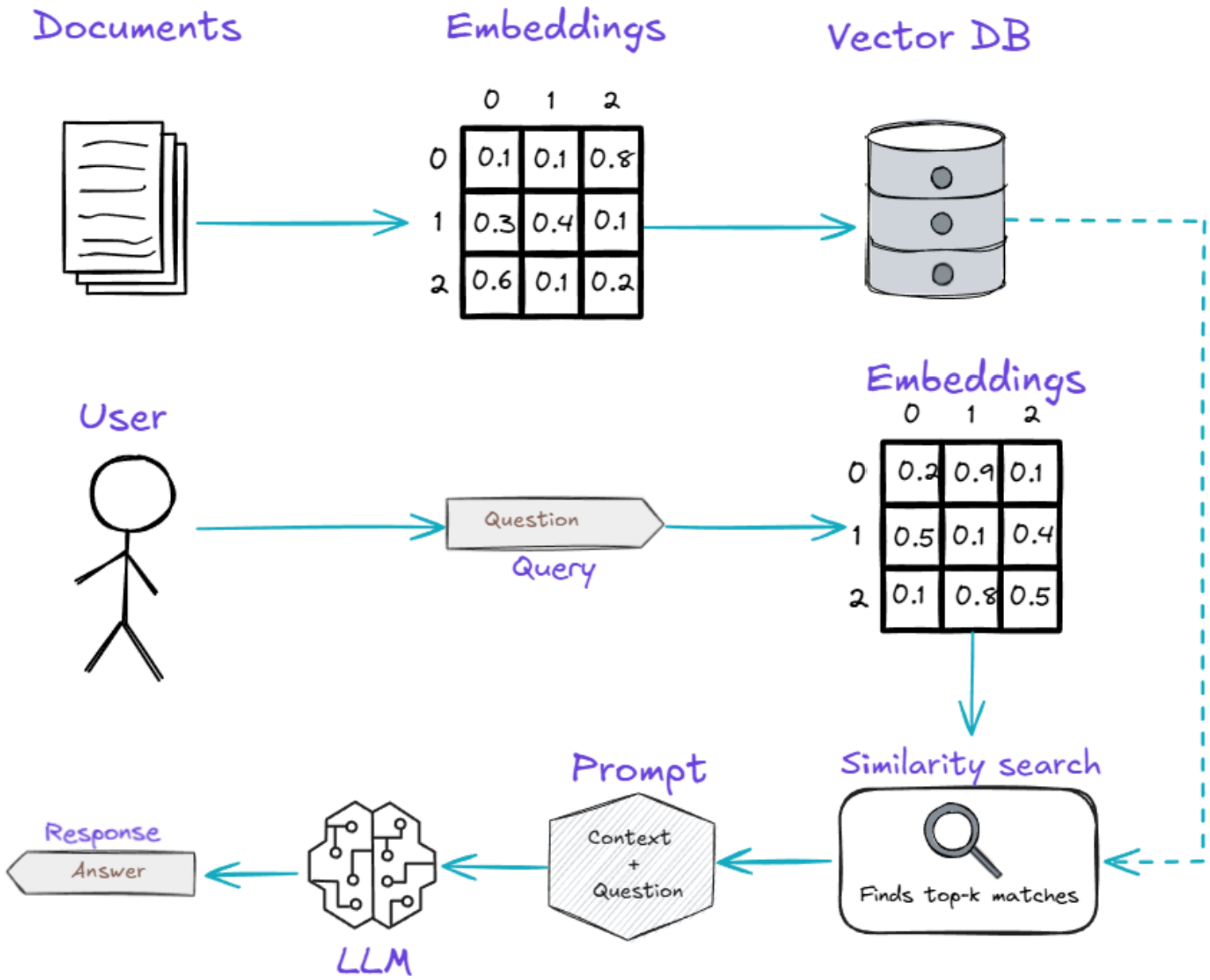


Embedding-based Retrieval: how a user question finds relevant chunks

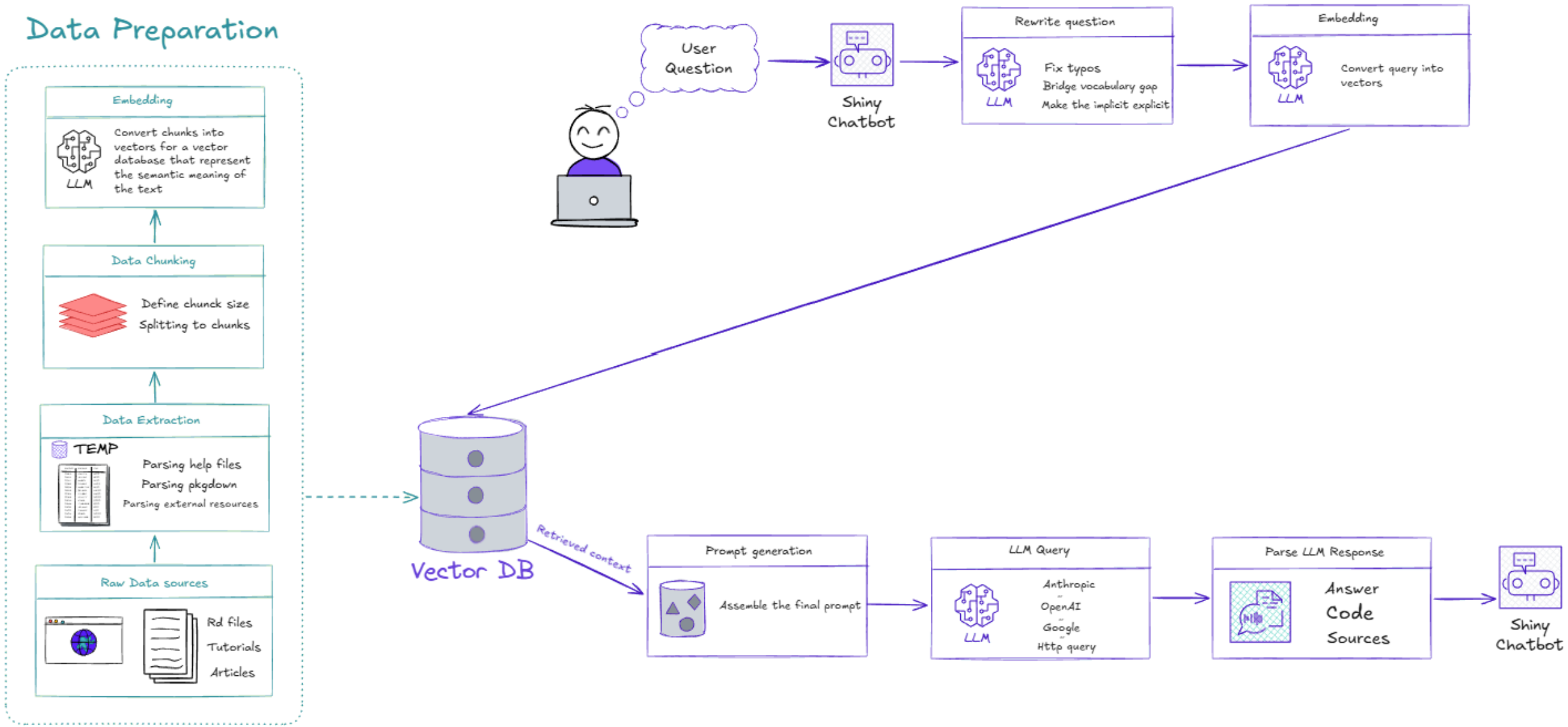


Typical RAG Architecture

DOI:10.1007/s12599-025-00945-3



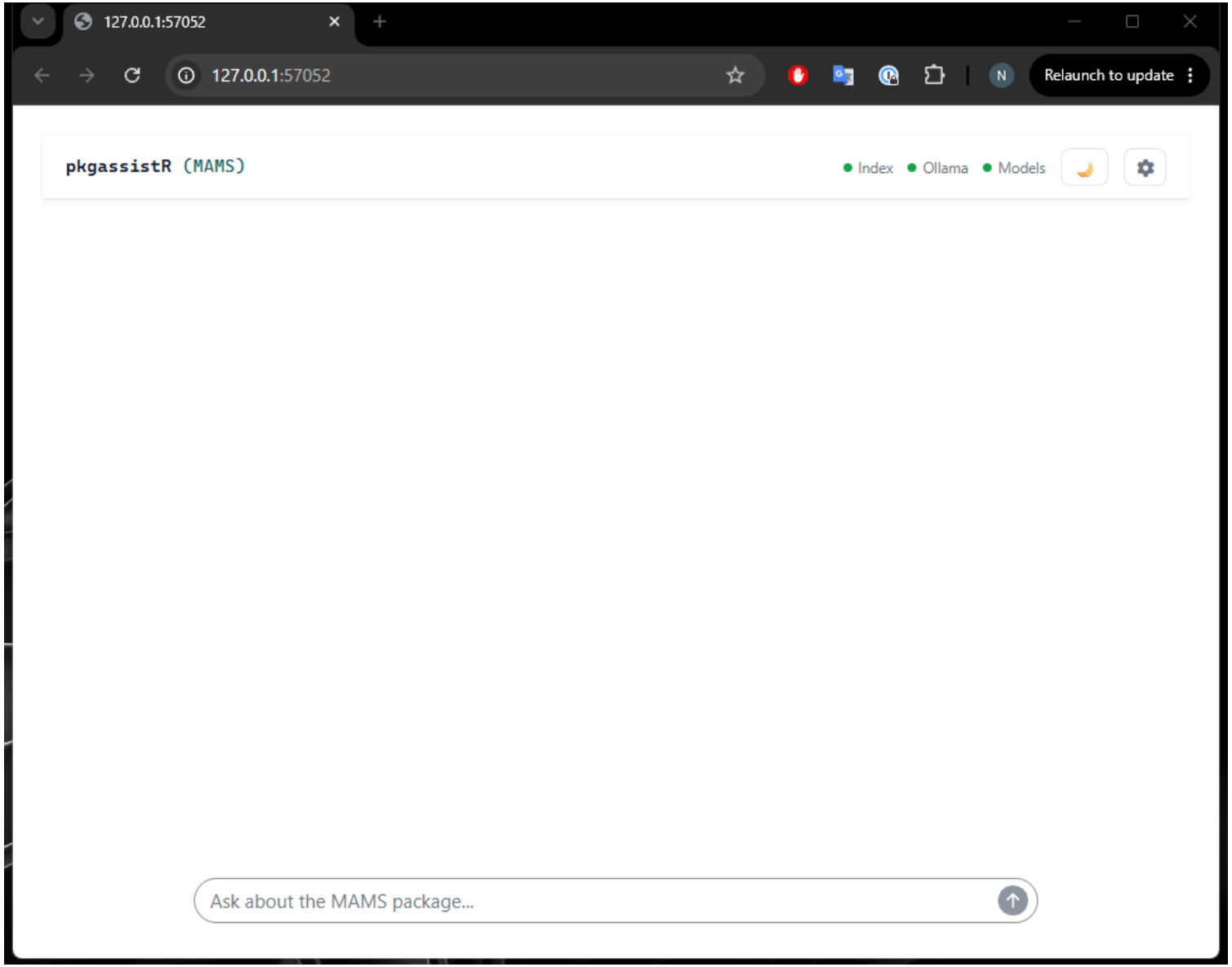
RAG: a case study

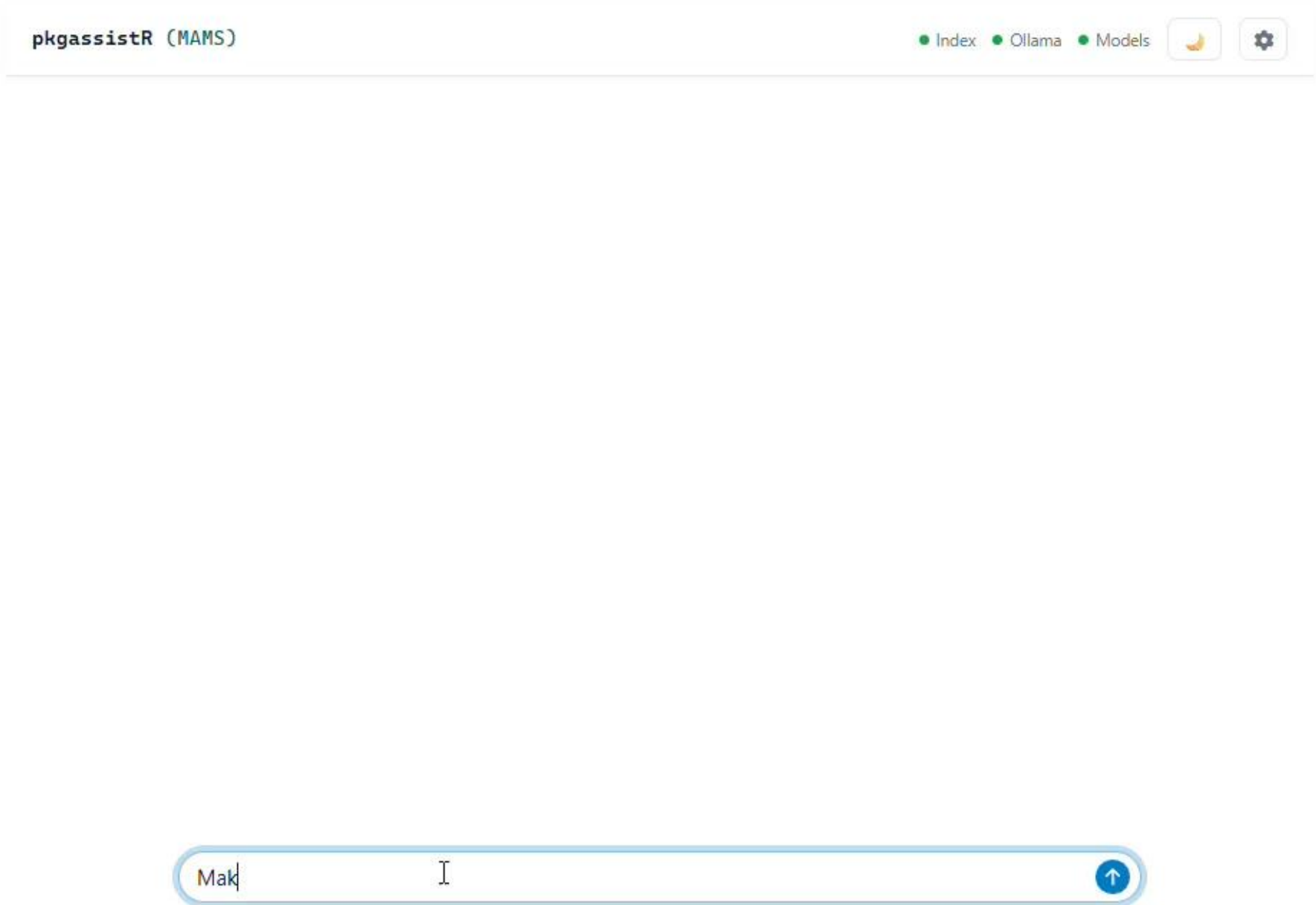

Data Preparation
Embedding
LLM
Convert chunks into vectors for a vector database that represent the semantic meaning of the text
Data Chunking
Define chunck size
Splitting to chunks
Data Extraction
TEMP
Parsing help files
Parsing pkgdown
Parsing external resources
Raw Data sources
Rd files
Tutorials
Articles
Vector DB
User Question
Shiny Chatbot
Rewrite question
LLM
Fix typos
Bridge vocabulary gap
Make the implicit explicit
Embedding
LLM
Convert query into vectors
Retrieved context
Prompt generation
Assemble the final prompt
LLM Query
LLM
Anthropic
OpenAI
Google
Http query
Parse LLM Response
Answer
Code
Sources
Shiny Chatbot

The Recipe

	Ingredient	Package / Tool
Data preparation	Corpus source	CRAN (package tarball)
	Parsing sources	base R + tidyverse
	Chunking	custom logic (base R + tidyverse)
	Embeddings	rollama (local Ollama for embedding model)
	Storage	plain RDS (base R)
Retrieval and Response	Retrieval scoring	dplyr, purrr (embeddings similarity)
	Prompt assembly	custom R script
	LLM chat/query	rollama (local Ollama) httr2 (Ollama Cloud direct) ellmer (OpenAI / Anthropic / Google)
	UI	shiny, shinychat (chat bubbles UI)

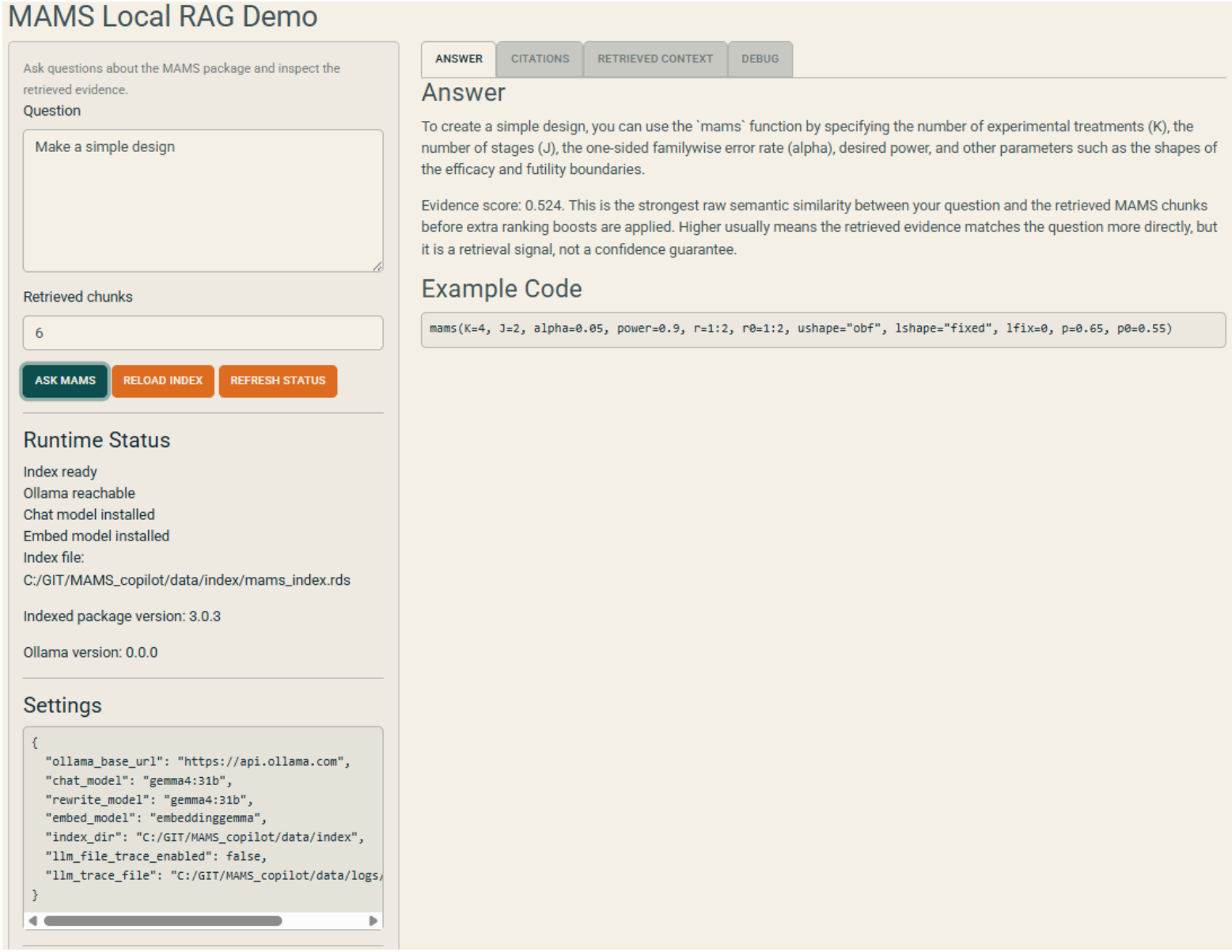
The chat interface




pkgassistR (MAMS)
Index
Ollama
Models
Mak

Backend

Example query and generated code



Retrieved chunks and citations

MAMS Local RAG Demo

Ask questions about the MAMS package and inspect the retrieved evidence.

Question

Make a simple design

Retrieved chunks

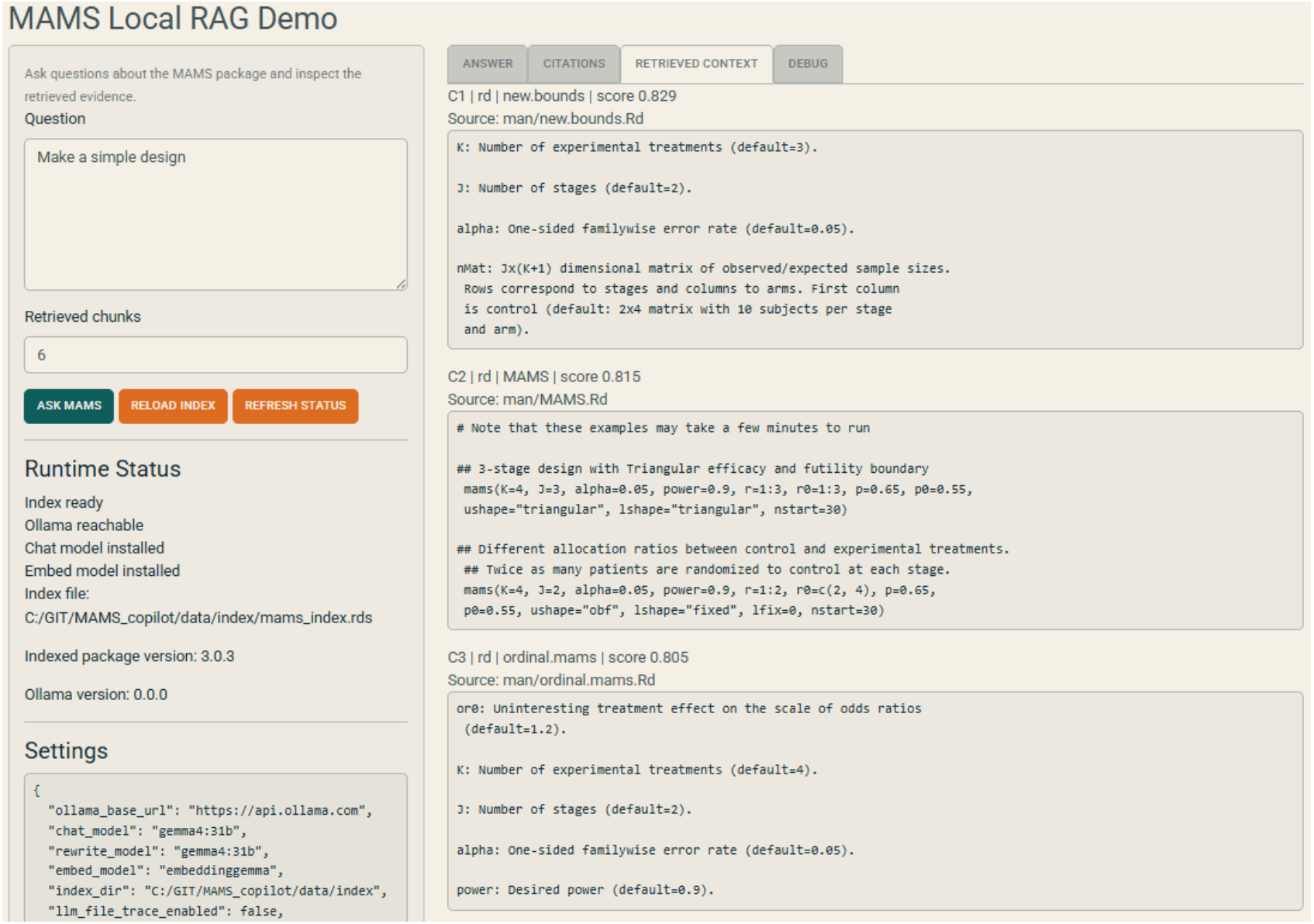
6

ASK MAMS | RELOAD INDEX | REFRESH STATUS

ANSWER | CITATIONS | RETRIEVED CONTEXT | DEBUG

citation	type	topic	function_name	source_file	score
C1	rd	new.bounds	new.bounds	man/new.bounds.Rd	0.83
C2	rd	MAMS	MAMS	man/MAMS.Rd	0.81
C3	rd	ordinal.mams	ordinal.mams	man/ordinal.mams.Rd	0.80
C4	rd	ordinal.mams	ordinal.mams	man/ordinal.mams.Rd	0.80
C5	rd	tite.mams	tite.mams	man/tite.mams.Rd	0.80
C6	rd	mams.sim	mams.sim	man/mams.sim.Rd	0.79

Runtime status and settings



Interpreting the model's output

Interpretation Summary

No explicit groups/arms count detected.

No explicit stage count detected.

Query rewrite enabled -> using `gemma4:31b` when available.

Retrieval Query

Original question

```
Make a simple design
```

LLM rewrite

```
Generate a simple experimental design with K treatments and J stages.
```

Model Output

Chat model: gemma4:31b

Embedding model: embeddinggemma

File trace: disabled

```
ANSWER:
A simple design can be created using the `mams` function by specifying the number of experimental treatments (K), the number of stages (J), the one-sided familywise error rate (alpha), and the desired p

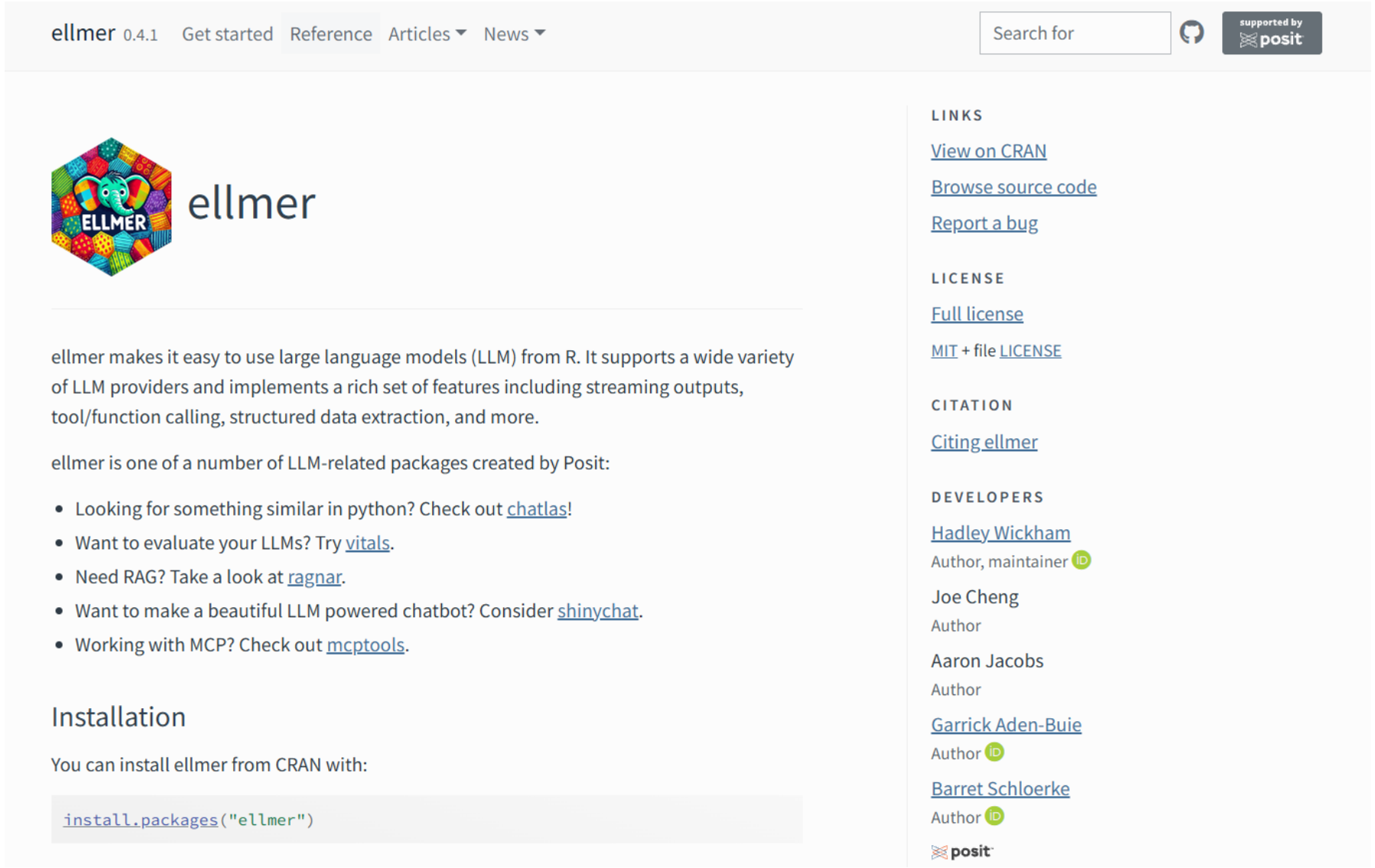
CODE:
mams(K=3, J=2, alpha=0.05, power=0.9, r=1:2, r0=1:2, p=0.65, p0=0.55)

FUNCTIONS:
mams
```

Key takeaway(s):

- RAG is not perfect.
- RAG could be combined with the other grounding strategies.
- Better documentation for packages makes them more suitable for RAG.
- RAG can be used on almost everything:
  - Knowledge base of previous research.
  - DB with articles.
  - DB for consultancy projects.
  - ….

ellmer: R interface to LLM providers



Thank you
Q&A