Generative Conversational AI

TECHNOLOGIES

Diego Gosmar

OPEN VOICE INTEROPERABILITY

Diego Gosmar © 2024

Diego Gosmar

Chief AI Officer — xcally Human Singularity AI

Start-up Mentor & Advisor — i3P

Ambassador — OPEN VOICE TRUSTMARK, OPEN VOICE INTEROPERABILITY

Diego Gosmar

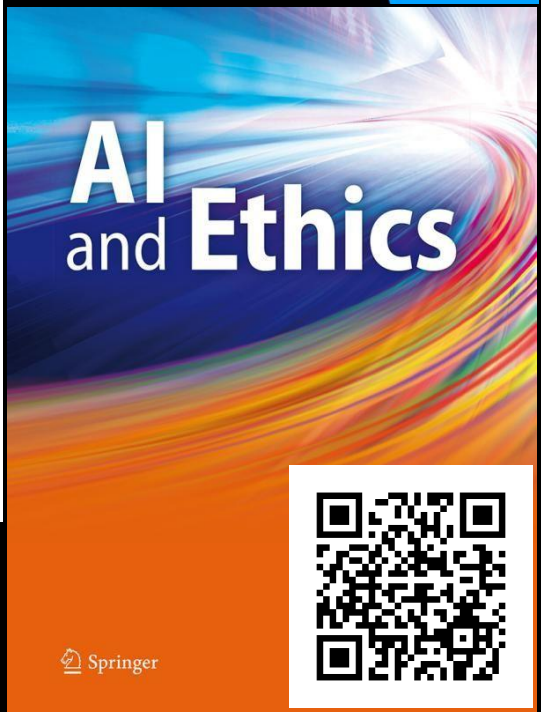
FOLLOWING

Chief AI Officer, XCALLY.ai
Verified email at ieee.org - Homepage

Artificial Intelligence Conversational AI Conversation Analysis

Google Scholar

TITLE	CITED BY	YEAR
Conversational hyperconvergence: an onlife evolution model for conversational AI agency D Gosmar AI and Ethics, 1-15	1	2024
Conversational AI Multi-Agent Interoperability, Universal Open APIs for Agentic Natural Language Multimodal Communications D Gosmar, DA Dahl, E Coin arXiv preprint arXiv:2407.19438		2024
Insight AI Risk Detection Model-Vulnerable People Emotional Situation Support D Gosmar, E Peretto, O Coleman Proceedings of the 28th International Conference on Evaluation and …		2024



- AI intro
- Generative AI and Conversational AI intro
- LLM and RAG intro
- AI new models, self-supervised pretrained and fine-tuning
- LLM Key concepts and Embedding
- Demo exercises
- RAG
- Local examples
- Demo exercises
- RAG Framework example
- Demo exercises
- RAG with Open AI
- Demo exercises
- LLM, RAG management and monitoring
- OVON use case (smart library)

Conversational AI
What is it?

Discriminative AI vs Generative AI

Diego Gosmar

- **Discriminative AI** is capable of classifying data and making predictions based on predefined models, data patterns and historical data.

- **Generative AI** is capable of generating new information and content from provided datasets.

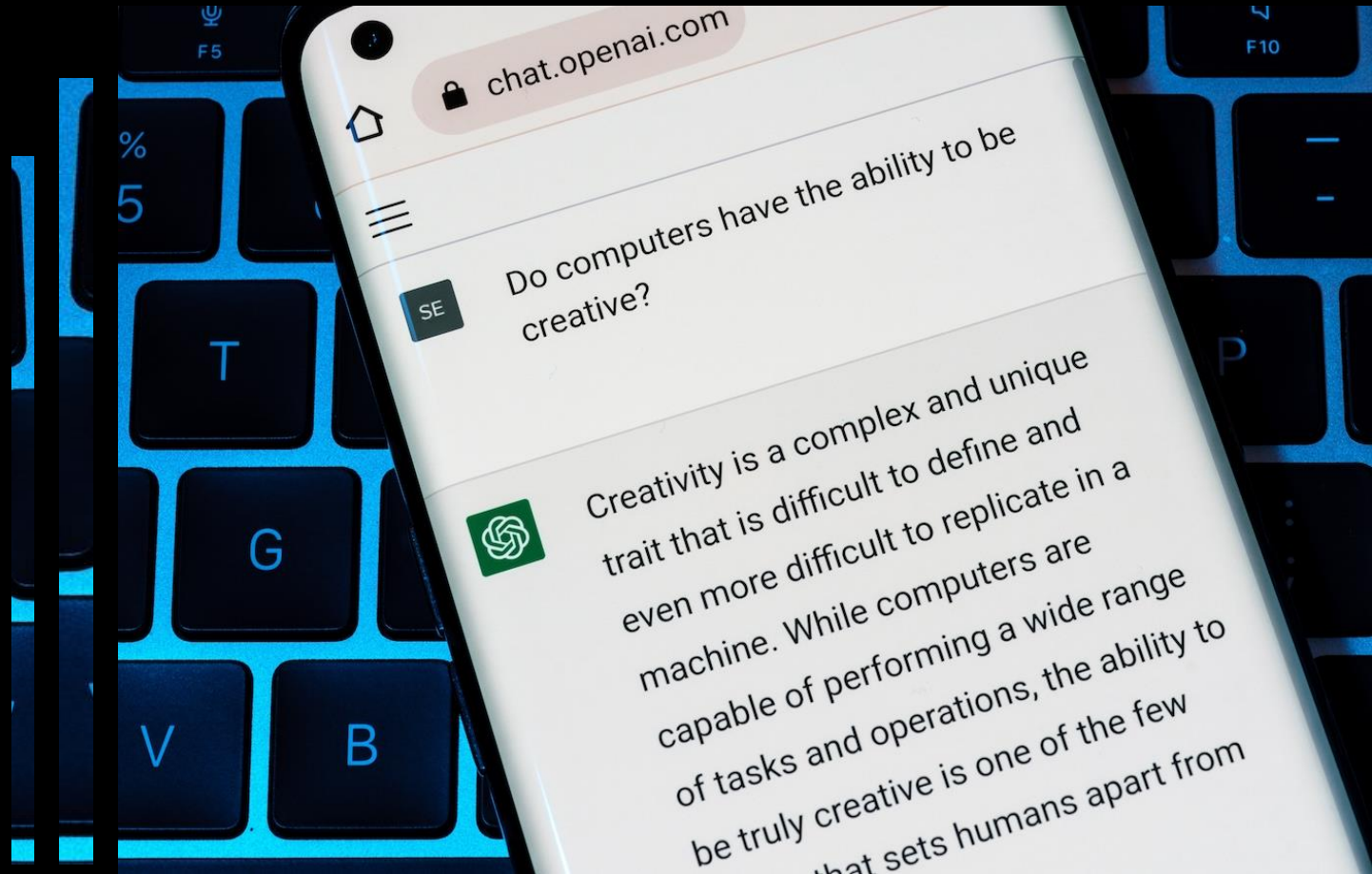
Generative AI

Machine Learning models/AI solutions **trained on a large amount of DATA in order to generate NEW DATA themselves**

Often **Multimodal**, with different kind of input/output:

i.e. text, voice, images, video

LMM



Generative AI

INPUT DATA
(i.e. text prompt, context, role, data sources)

OUTPUT DATA
(i.e. articulated text, images, sounds, contents…)

Diego Gosmar

Conversational AI

Evolution Timeline

Intents
→ Responses

2017

Discriminative AI

Dialogflow, Lex,…

Generative AI
→ Prompt Engineering

2022

LLM, LMM
(GPT,…)

Generative AI

Retrieval Augmented
Generation (RAG)

2024

LMM + Vector DB + API
(LangChain, LlamaIndex,…)

t

Conversation Naturalness

CONTEXT AWARENESS

RAG (Retrieval Augmented Generation)

Fine Tuning (Open-Weight LLM)

LAM (Large Action Models)

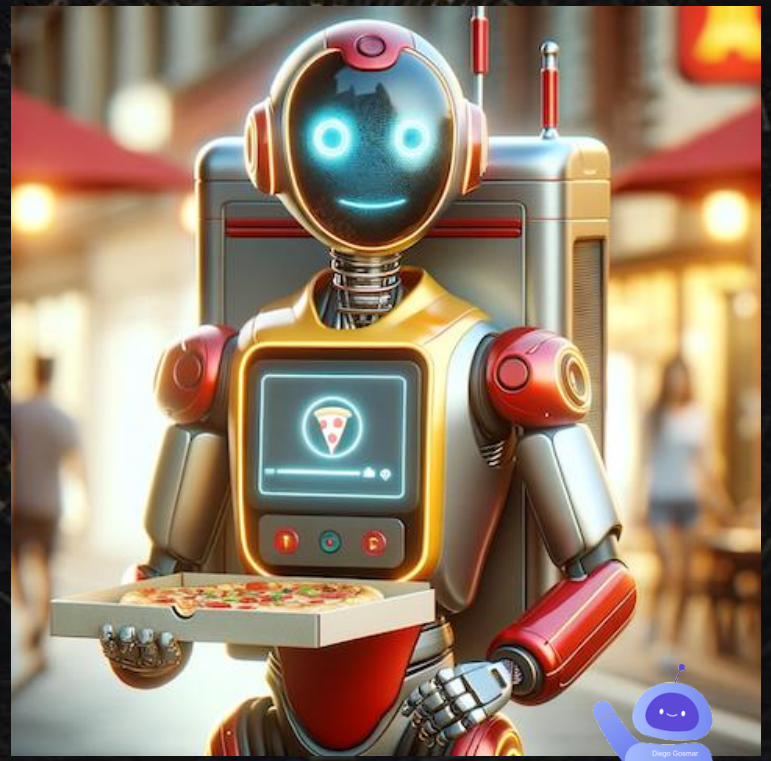
RAG, Fine Tuning and LAM

CUTOM AGENTS for everything!

HEALTHCARE

FINANCE

FOOD ORDERING



Diego Gosmar

RAG

Retrieval-Augmented Generation

Prompt:

what's the amount related to our latest higher invoices in 2024?

Context:

Invoice #1, amount: 1200 $
Invoice #2, amount: 750 $
…
Invoice #N, amount: 2350 $

LLM

Role:

You are an assistant agent specialized in billing information

Output:

Sure, the following ones are the highest 2023 invoces:
Invoice #5: 1850 $
Invoice #43: 2470 $
Etc… etc…

RAG (Retrieval Augmented Generation)

Fine Tuning (Open-Weight LLM)

LAM (Large Action Models)

Function-Calling capabilities

Fine-Tuning EXERCISES

Diego Gosmar

Homeworks

https://huggingface.co/models

i.e. EleutherAI/gpt-neo-125m


LLM
LMM

Language Model

A language model is a probabilistic model of a natural language

Text generation: (predict next words, utterances, sentences…)

Text Summarization

Translation

…

LMM: beyond text!

LLM: training and tuning

KEY concepts

- **Model Architecture**: This refers to the specific implementation and mathematical structure of the model. Most top-performing LLMs today use a variation of the "decoder-only" **Transformer architecture**.

- **Training Dataset**: The dataset comprises examples and documents used to train the model. These documents can be in natural languages (like English or Chinese), programming languages (like Python), or any structured text (like tables or equations). The training process involves the model learning patterns from these documents.

Exploratory

LLM: training and tuning

KEY concepts

- **Tokenizer**: This is a tool that converts text from the training dataset into numbers, which the model can process. It breaks down text into smaller units called **tokens**, which could be words, sub-words, or characters. The tokenizer's vocabulary size varies, indicating the number of different tokens it recognizes: typically between **32k and 200k**

*The size of an LLM dataset is often measured as the number of tokens it contains once split in a sequence of these individual, "atomistic" units: these days it ranges from **several hundred billion tokens** to **several trillion tokens**!*

LLM: training and tuning

KEY concepts

- **Training Process**: Requires significant computing power and skilled personnel. It involves setting up the model architecture and running the training algorithm with the chosen **hyperparameters** on the training dataset. The outcome is a set of learned model **weights**, which are used for making predictions on new inputs.

- **Model Weights**: the **weights** can then be used for **inference**, i.e., for prediction on new inputs, for instance, to generate text.
  Pretrained LLMs can also be specialized or adapted for a specific task after pretraining, particularly **when the weights are openly released**. They are then used as a starting point for use cases and applications through a process called **fine-tuning**.

LLM: training and tuning

KEY concepts

- **Pretraining and Fine-Tuning**: Pretrained LLMs can be further specialized for specific tasks through fine-tuning, which involves **additional training on a more specialized dataset**.
  This process is **less resource-intensive** than training a model from scratch and allows for the **adaptation of open-source pretrained models for various applications**, even with **limited computing resources**.

i.e. HuggingFace H4 Zephyr 7b beta model

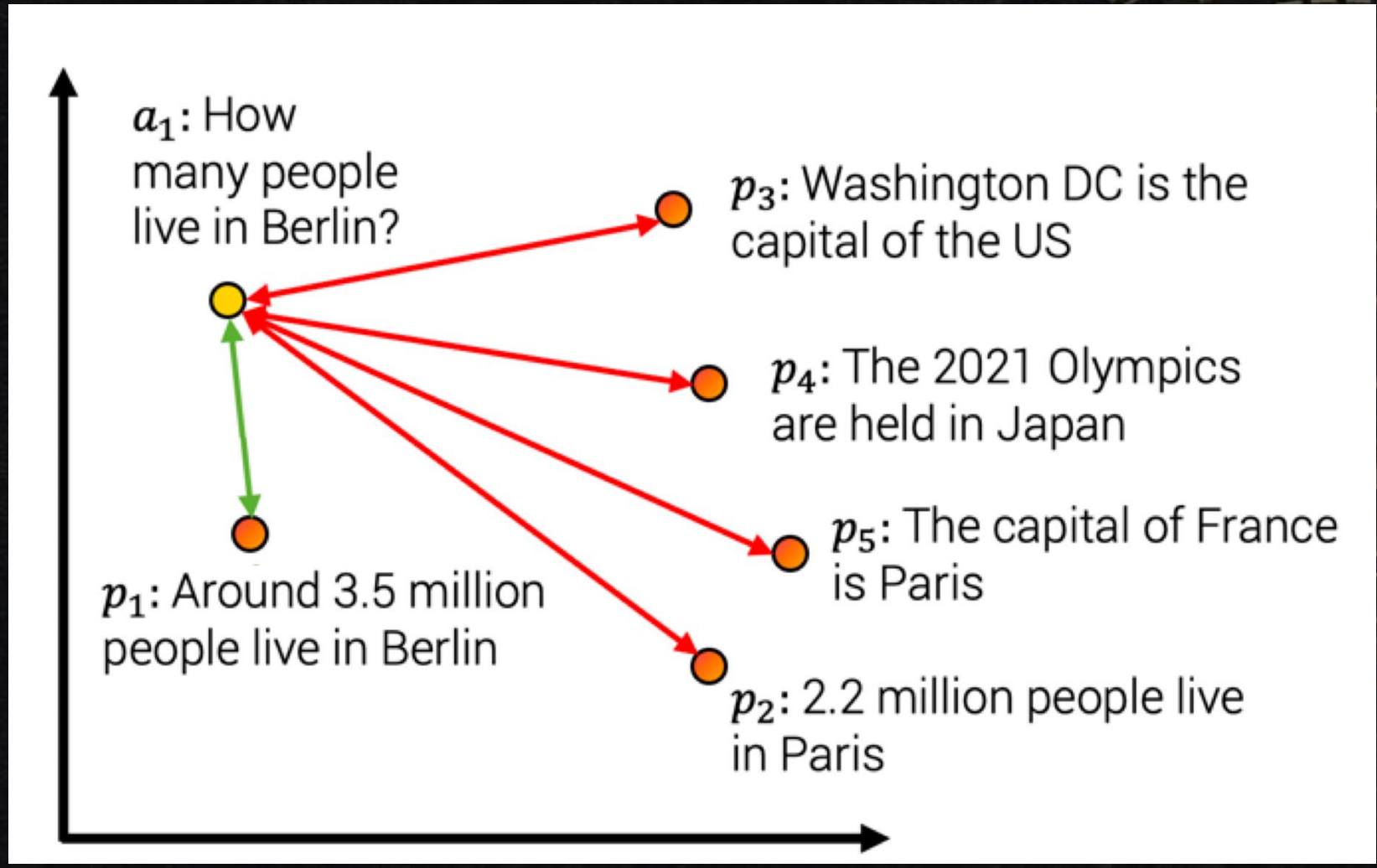
7.24 billion parameters

Fine-tuned version of mistralai/Mistral-7B-v0.1

Datasets: stingning/ultrachat

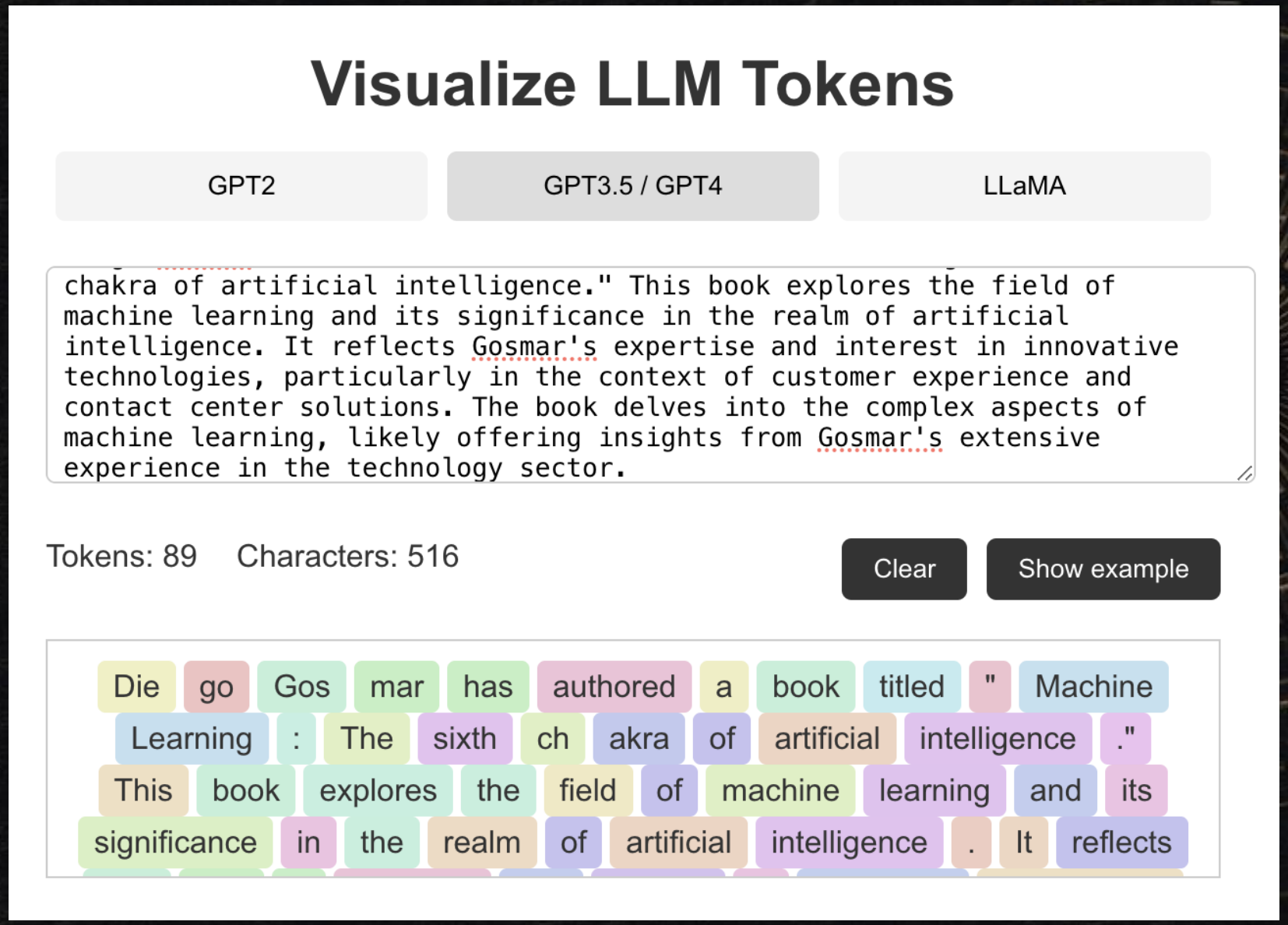
774K synthetic dialogues generated by ChatGPT

Embedding


a_1: How many people live in Berlin?
p_3: Washington DC is the capital of the US
p_4: The 2021 Olympics are held in Japan
p_5: The capital of France is Paris
p_1: Around 3.5 million people live in Berlin
p_2: 2.2 million people live in Paris

EXERCISES

LLM: training and tuning



Tokenizer VS Embedding

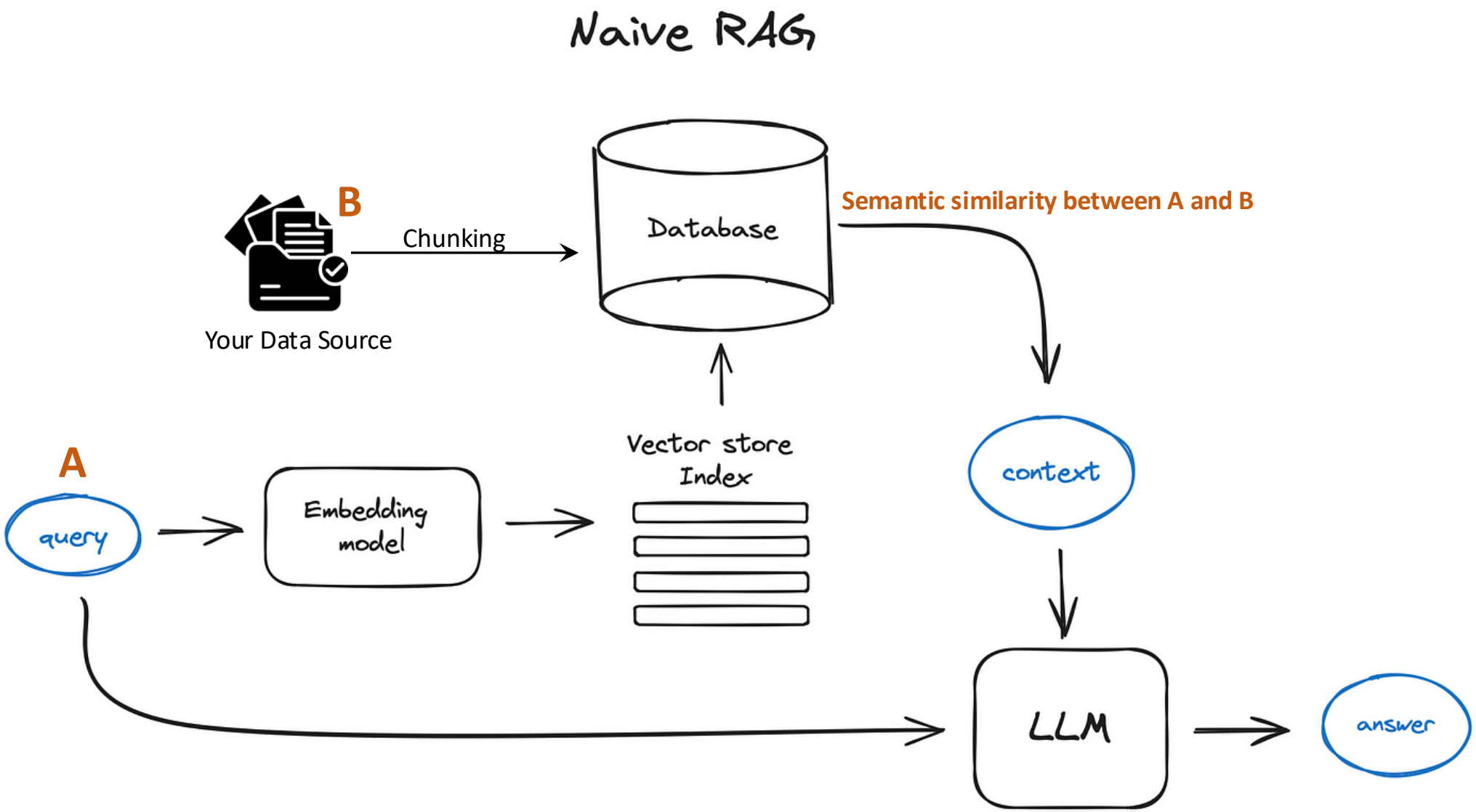
Tokenizer → Dictionary → Syntactic domain

Syllabus, Word, part of text mapping rules to numbers!

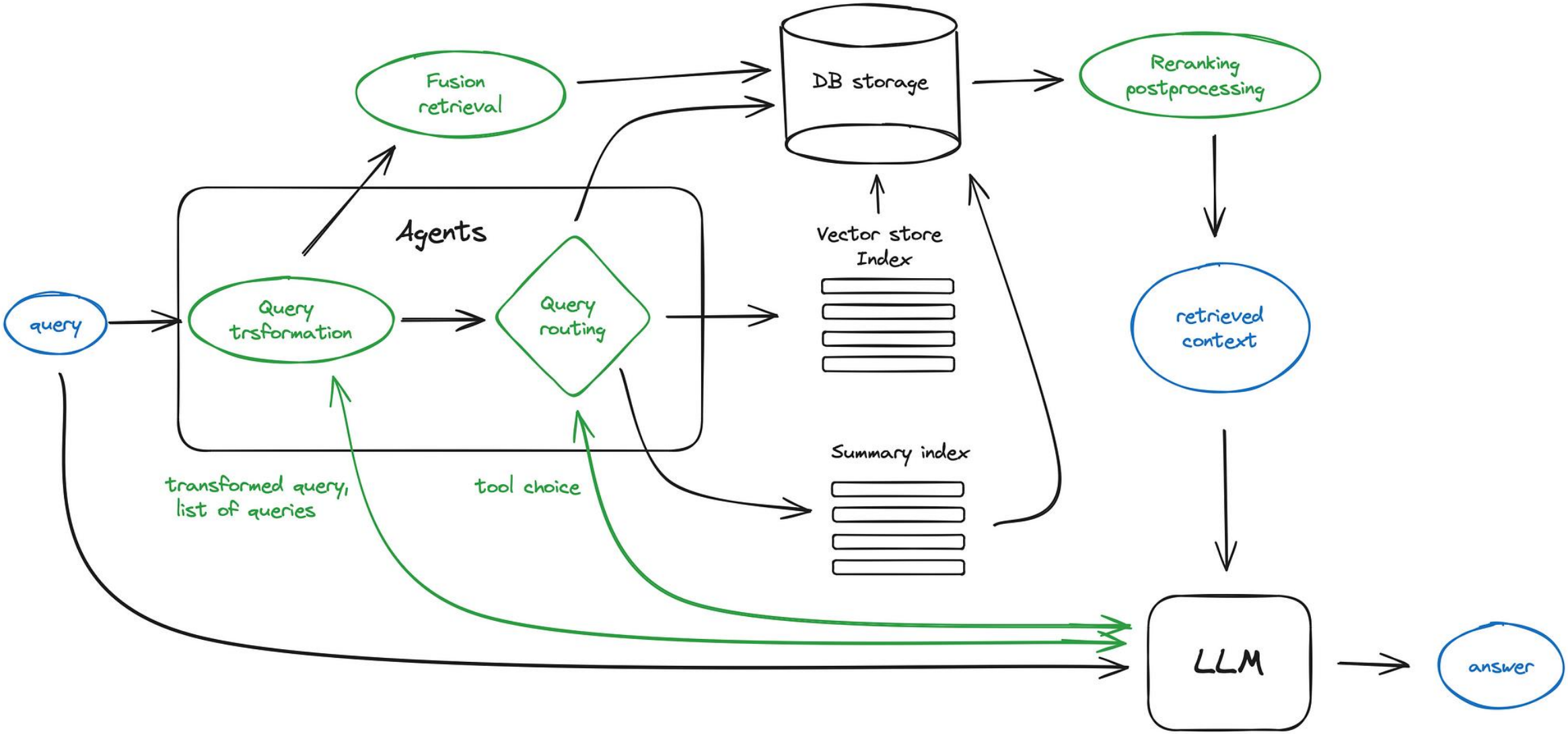
Embedding → Vectorization → Semantic domain

Text meaning, correlations, context,…

RAG


Naive RAG
B
Chunking
Database
Semantic similarity between A and B
Your Data Source
Vector store
Index
A
query
Embedding
model
context
LLM
answer

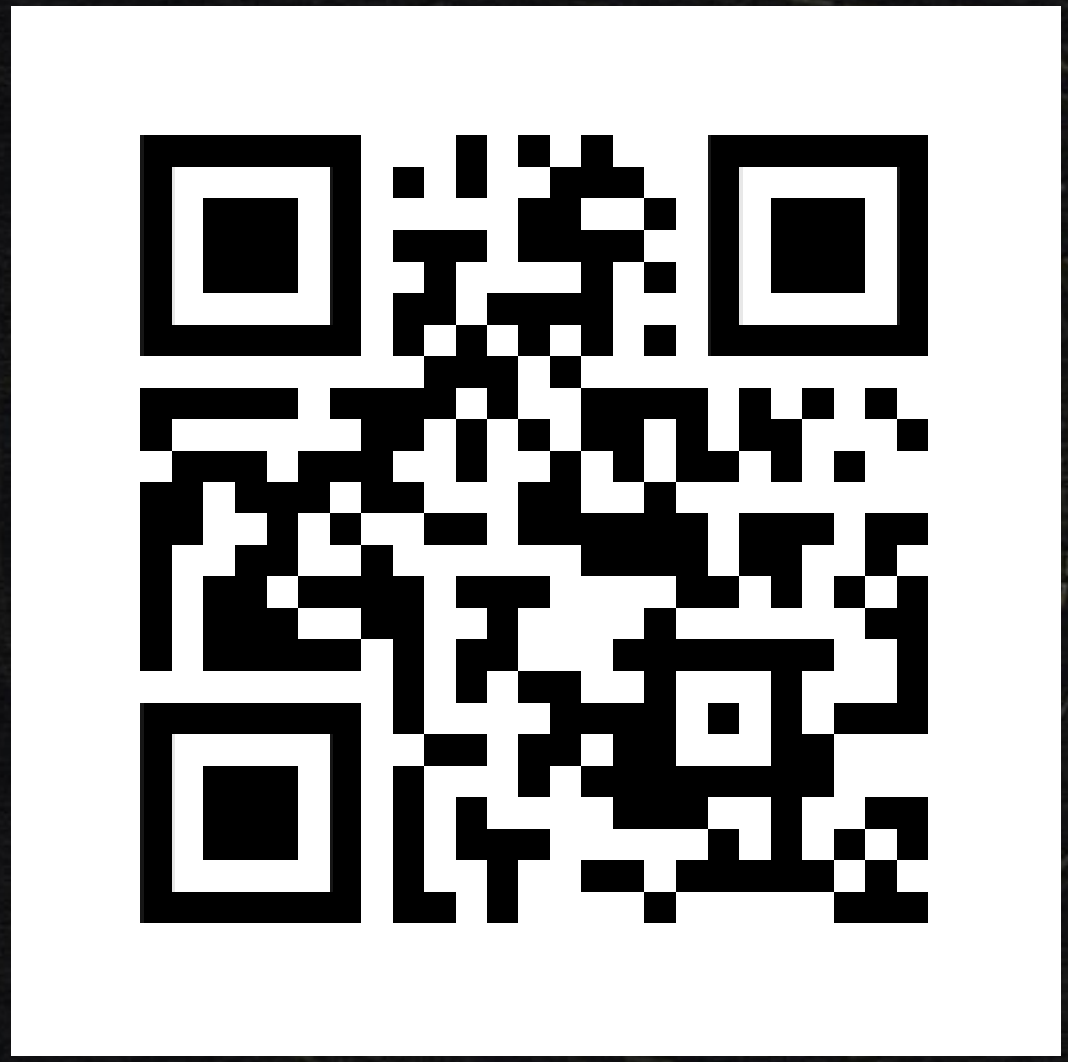
Advanced RAG



EXERCISES

ollama setup

https://ollama.com



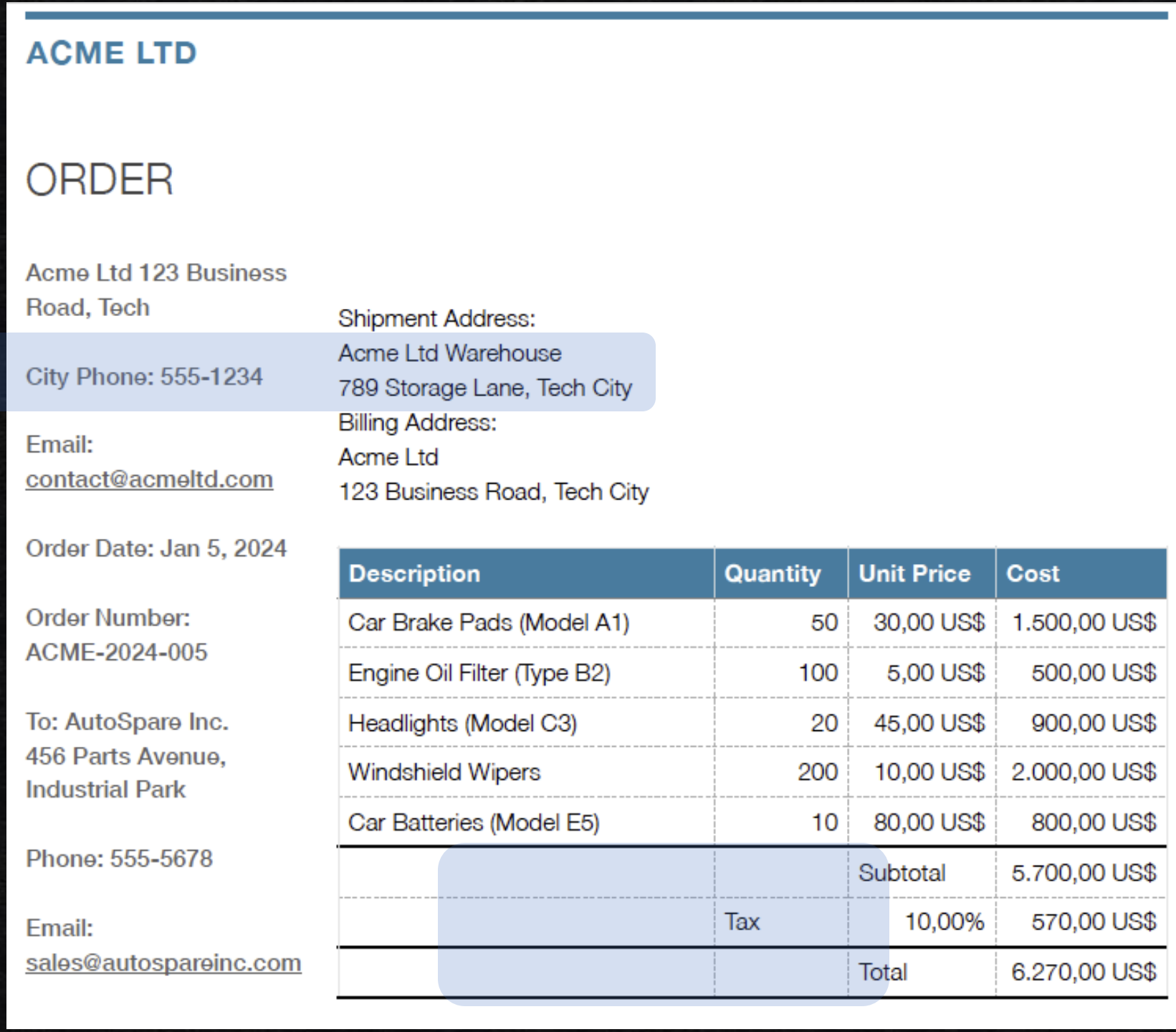
ollama cfg

- Run a model
- Pull a model
- Customize a Role for a Model
- ollama API

SEE NOTES

Diego Gosmar

RAG: ollama web app demo-kit



ACME LTD

ORDER

Acme Ltd 123 Business Road, Tech

City Phone: 555-1234

Email: contact@acmeltd.com

Order Date: Jan 5, 2024

Order Number: ACME-2024-005

To: AutoSpare Inc.
456 Parts Avenue,
Industrial Park

Phone: 555-5678

Email: sales@autospareinc.com

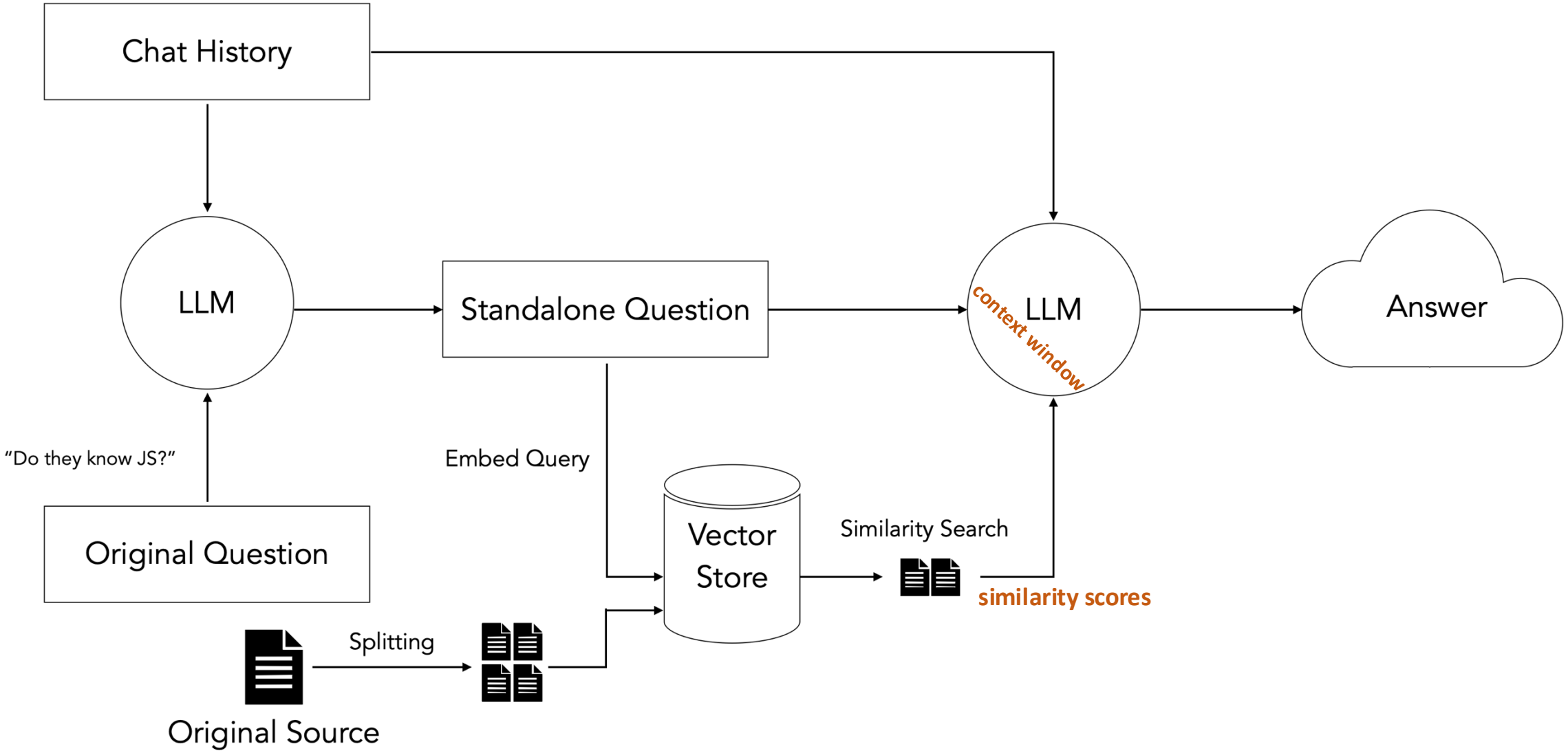
Shipment Address:
Acme Ltd Warehouse
789 Storage Lane, Tech City

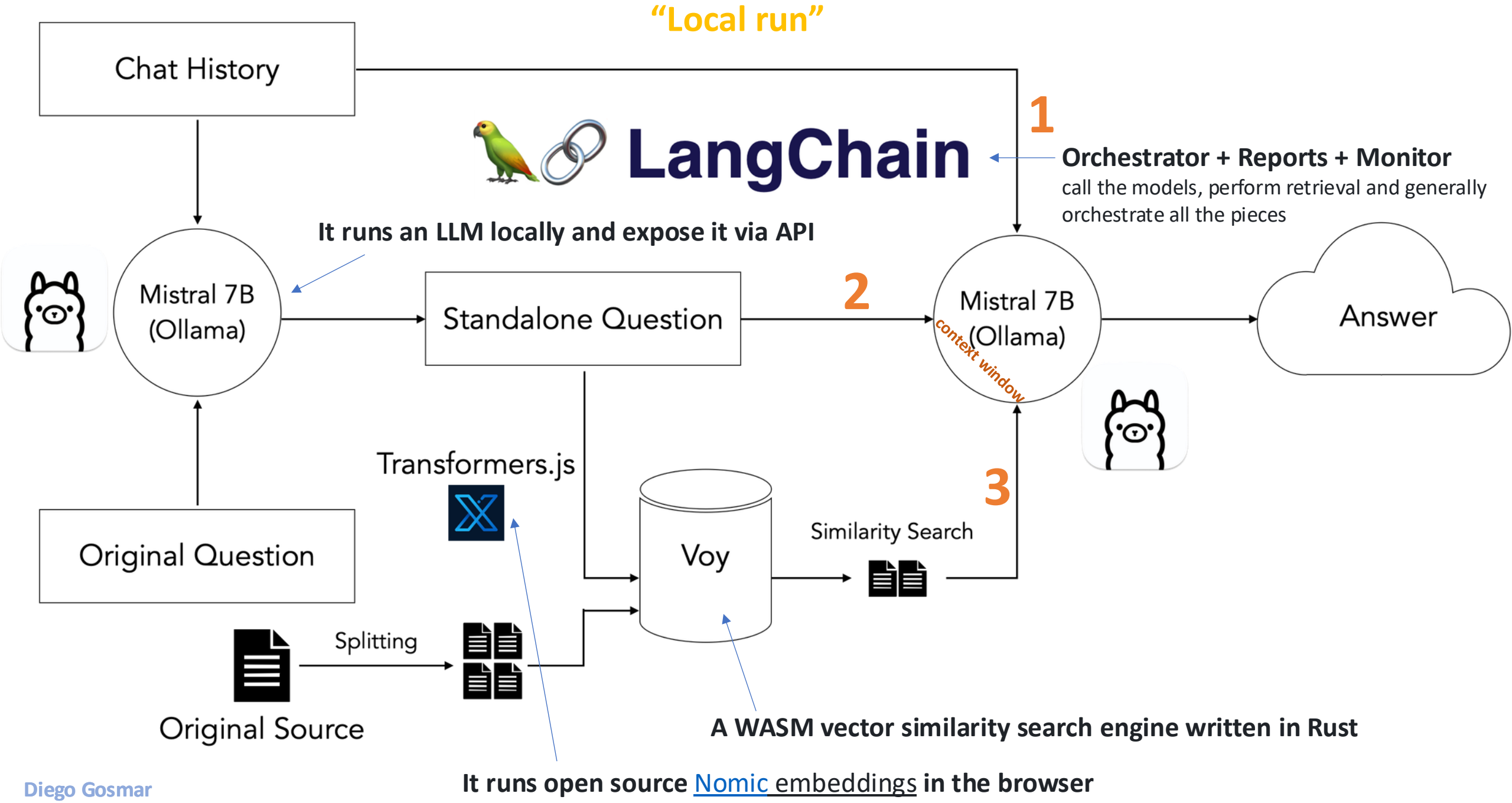
Billing Address:
Acme Ltd
123 Business Road, Tech City

Description	Quantity	Unit Price	Cost
Car Brake Pads (Model A1)	50	30,00 US$	1.500,00 US$
Engine Oil Filter (Type B2)	100	5,00 US$	500,00 US$
Headlights (Model C3)	20	45,00 US$	900,00 US$
Windshield Wipers	200	10,00 US$	2.000,00 US$
Car Batteries (Model E5)	10	80,00 US$	800,00 US$
		Subtotal	5.700,00 US$
	Tax	10,00%	570,00 US$
		Total	6.270,00 US$

Diego Gosmar

Architecture




"Local run"
Chat History
LangChain
1
Orchestrator + Reports + Monitor
call the models, perform retrieval and generally orchestrate all the pieces
It runs an LLM locally and expose it via API
Mistral 7B (Ollama)
Standalone Question
2
Mistral 7B (Ollama)
context window
Answer
Transformers.js
3
Voy
Similarity Search
Original Question
Splitting
Original Source
A WASM vector similarity search engine written in Rust
It runs open source Nomic embeddings in the browser

RAG: ollama web app demo-kit

Local fork and setup:

clone https://github.com/jacoblee93/fully-local-pdf-chatbot.git
yarn
npm run-script dev

OLLAMA_ORIGINS=http://localhost:3000 OLLAMA_HOST=127.0.0.1:11435 ollama serve
OLLAMA_HOST=127.0.0.1:11435 ollama pull mistral

Access via: http://localhost:3000/

RAG: ollama web app demo-kit

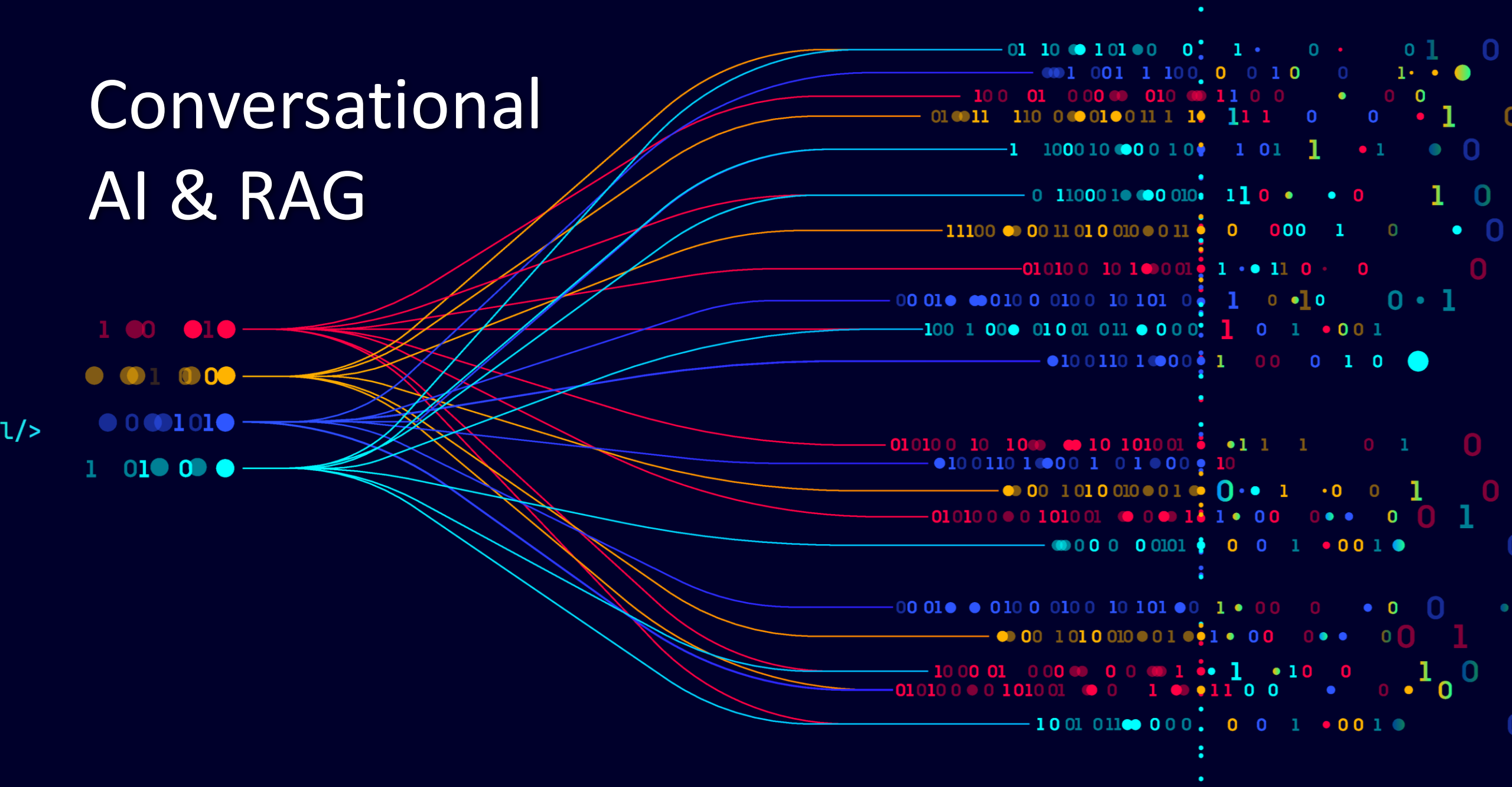
Here are some links for the various pieces used in the app:

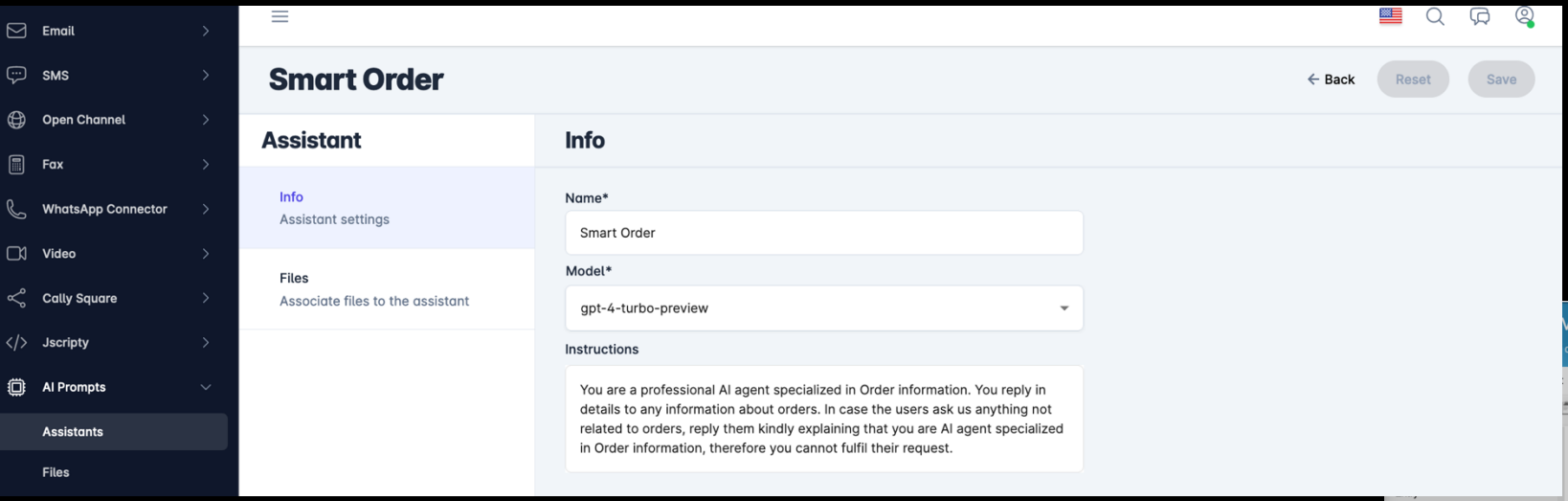
- Demo app: https://webml-demo.vercel.app/
- Demo app GitHub repo: https://github.com/jacoblee93/fully-local-pdf-chatbot
- Voy: https://github.com/tantaraio/voy
- Ollama: https://github.com/jmorganca/ollama/
- LangChain.js: https://js.langchain.com/
- Transformers.js: https://huggingface.co/docs/transformers.js/index

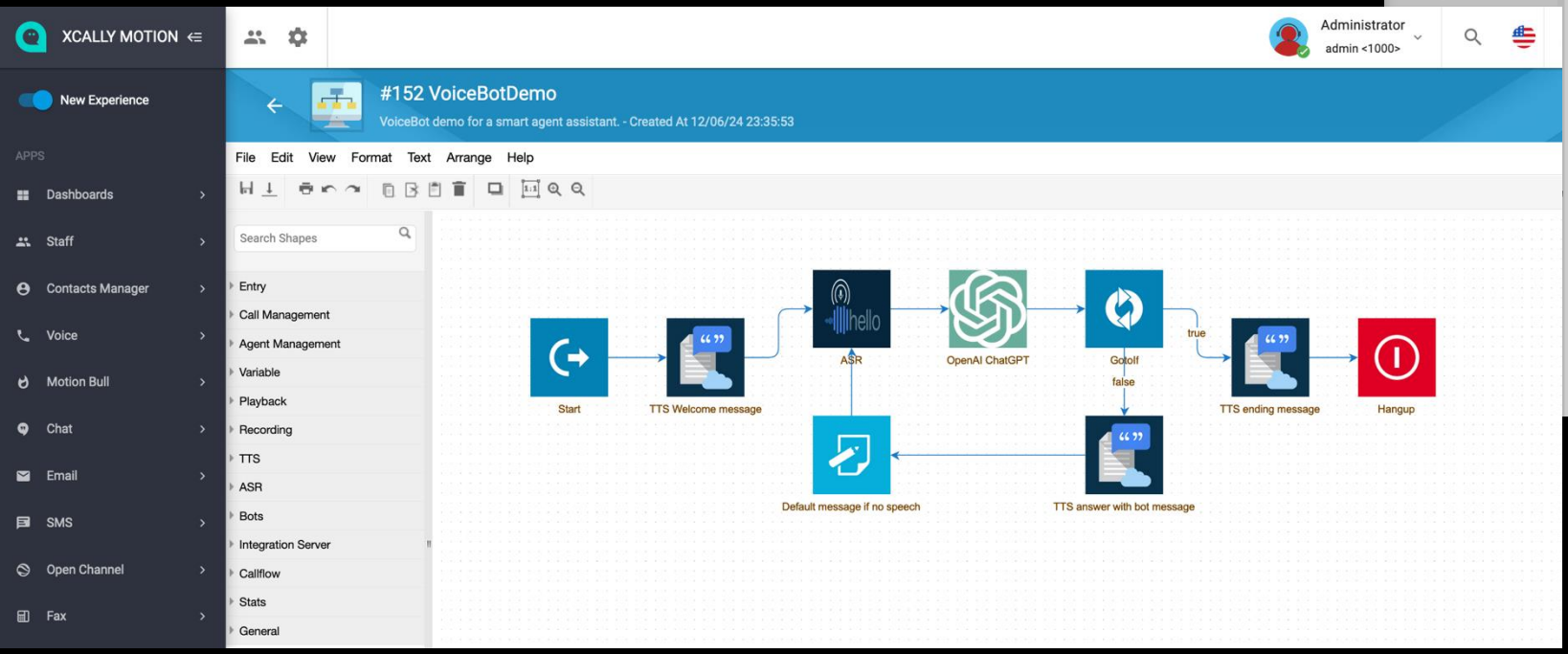
Exploratory

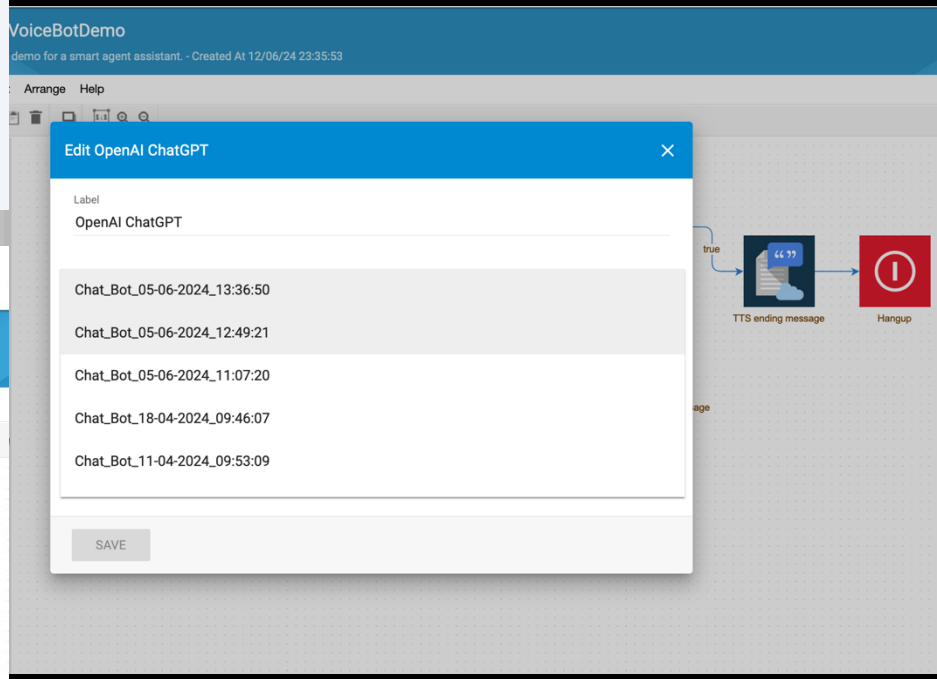
XCALLY demo

Conversational AI & RAG



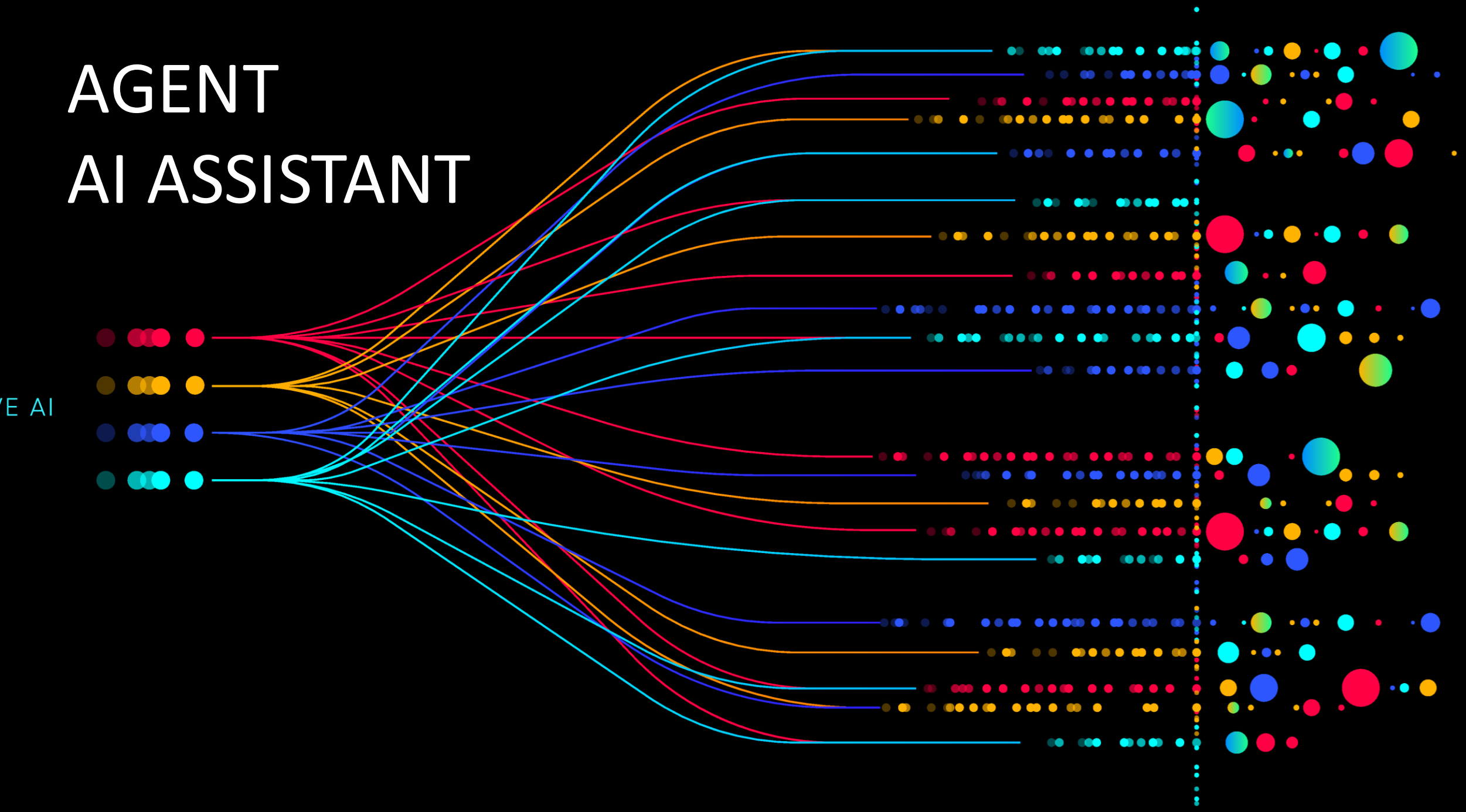






Diego Gosmar

AGENT
AI ASSISTANT



00:07:58

TechAcademy Demo
tech.academy <1006>

TCHAC - john.doe@email.com

Currently in use by

#30 - TCHAC Open 00:00:52

john.doe@email.com - 24/06/24 09:58:09

New chat request
from: http://localhost:7002
name: John Doe
email: john.doe@email.com

john.doe@email.com - 24/06/24 09:58:34

Can you tell me something about the error code 3

Get help from XCALLY Assistant

ode 3ZE0VZ1K?

SEND ANSWER

The error code 3ZE0VZ1K indicates that the appliance has detected that the heating element is calcified. To resolve this issue, you should descale the appliance and operate it with the water softening system.【4:0†source】.

CONTACT INTERACTION CUSTOMEI

Name
john.doe@email.com

Email
john.doe@email.com

Phone

Mobile

Fax

MORE

Insert your text message here

Diego Gosmar

What more?

chain = prompt | llm | output_parser

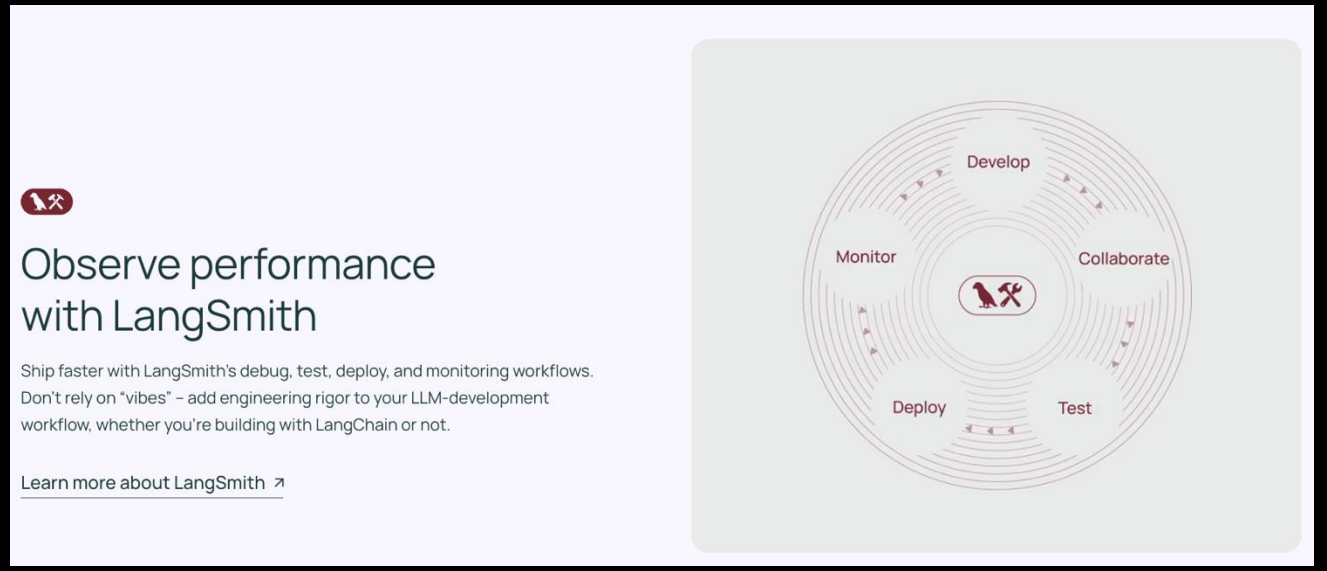
LangChain

LlamaIndex

GPT4All

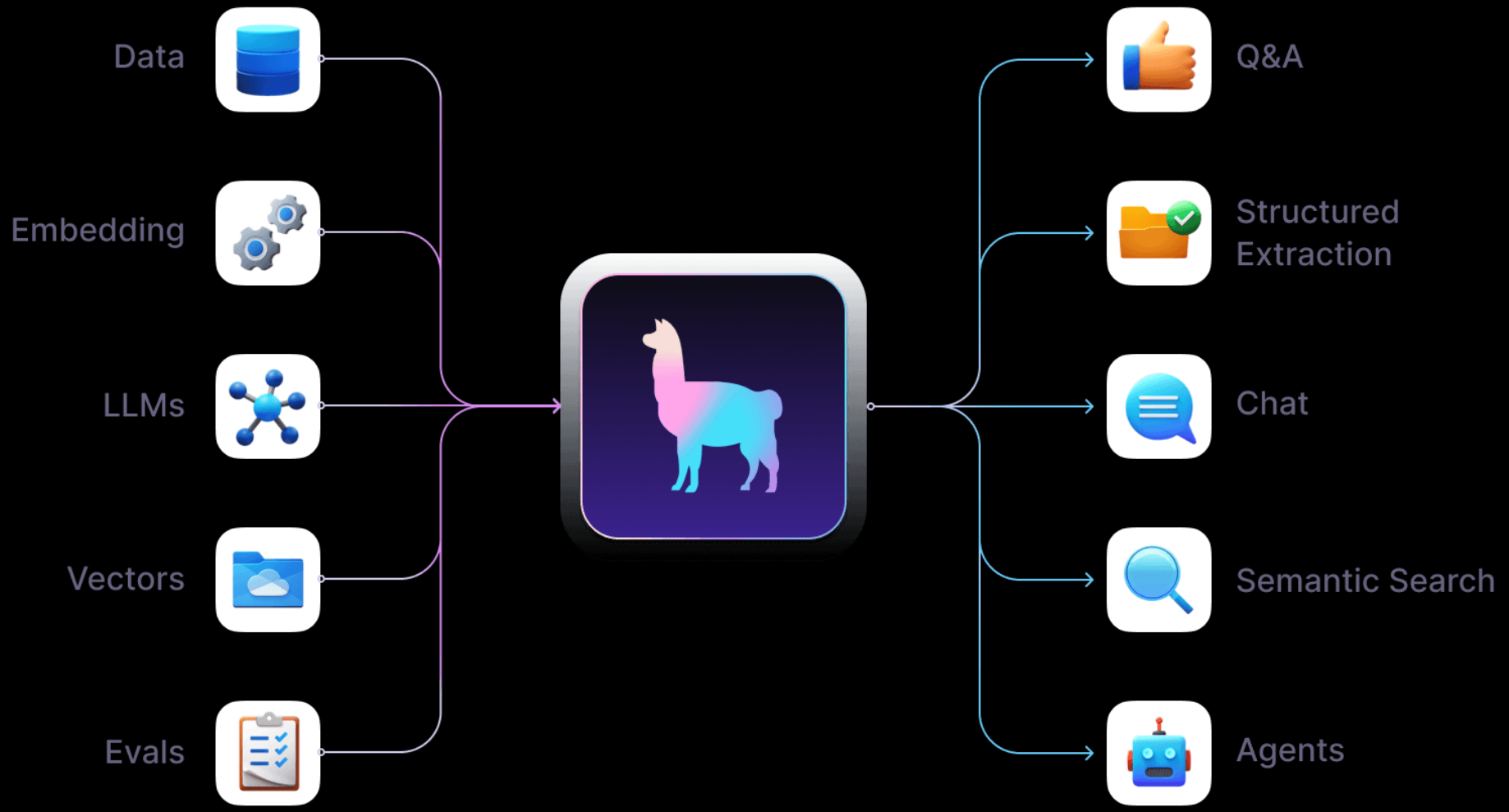
A free-to-use, locally running, privacy-aware chatbot. **No GPU or internet required.**

graph databases VS vector databases



Diego Gosmar

LLamaIndex: data framework for building LLM applications



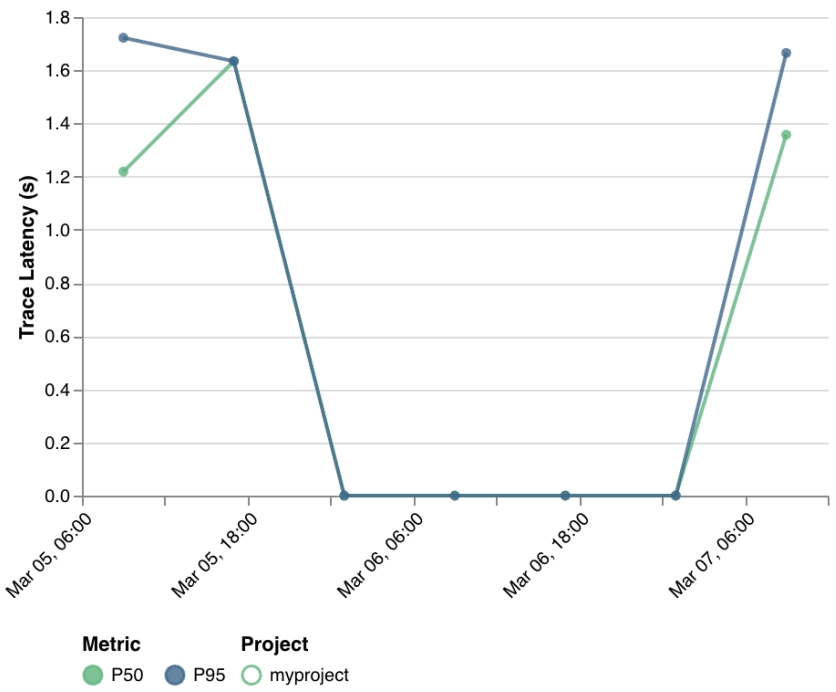
Personal > Projects > myproject --> One API Key and project x Manager Customer

myproject

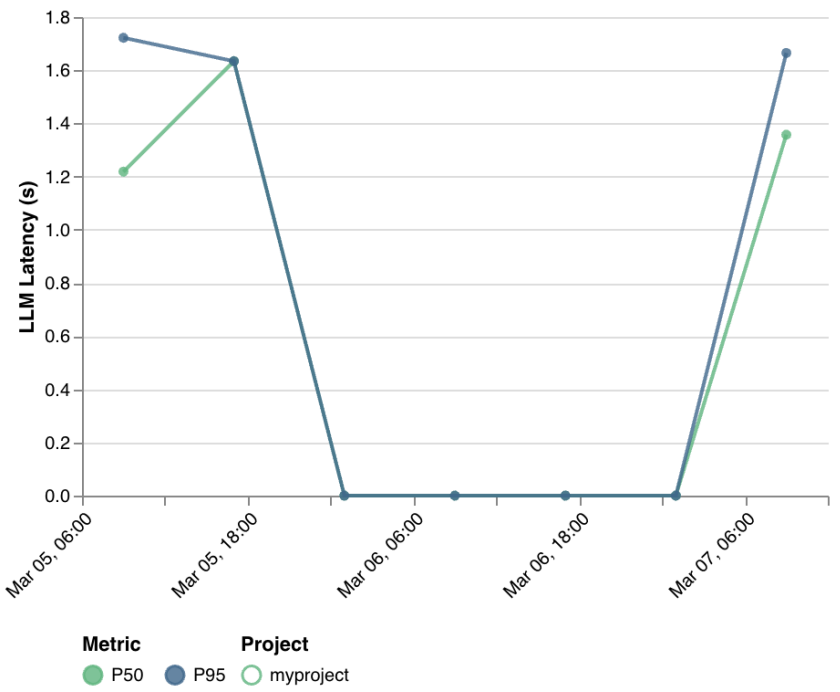
Edit

Traces | LLM Calls | All Runs | Monitor | Setup

Trace Latency (s)



LLM Latency (s)



Stats

Last 7 days

RUN COUNT
17

TOTAL TOKENS
1,534 / $0.001203

MEDIAN TOKENS
102

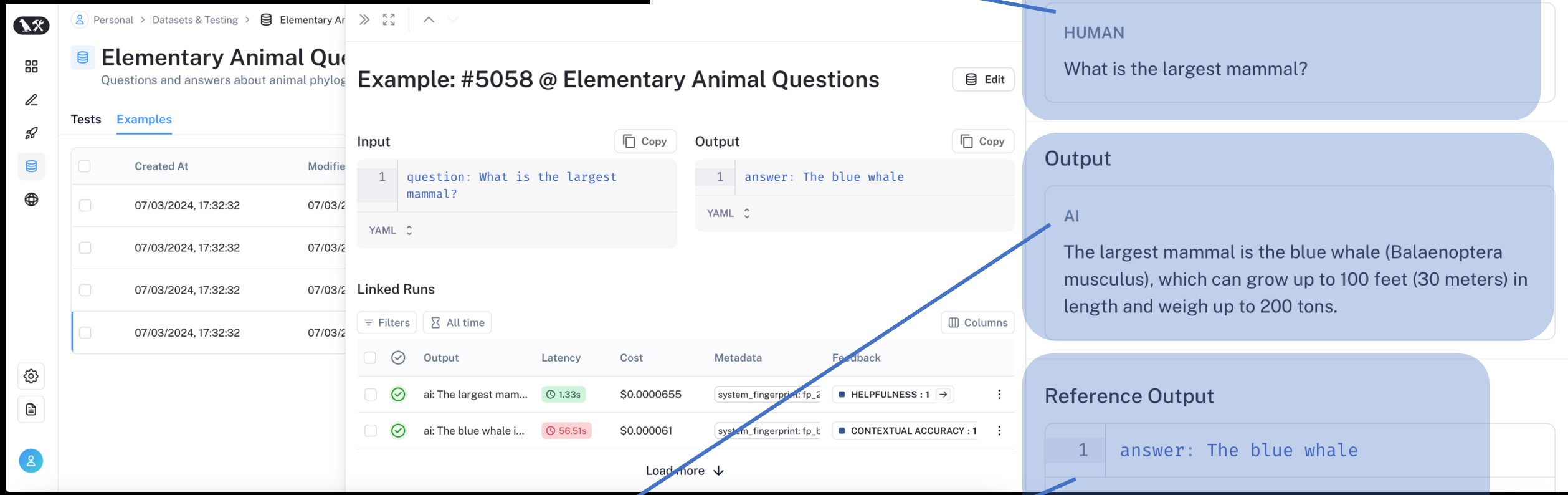
ERROR RATE
0%

% STREAMING
0%

LATENCY
P50: 1.38s P99: 1.78s

Diego Gosmar

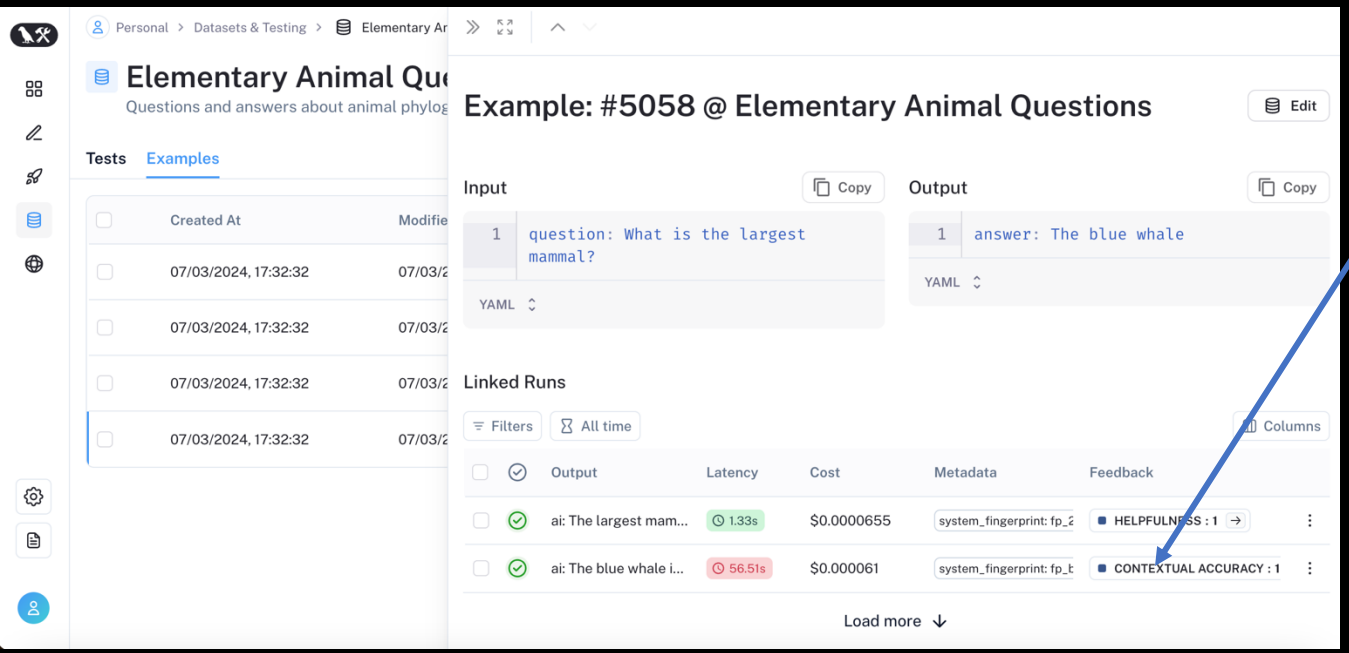
User query from the imported dataset



Simulated LLM answer from the Evaluator!

Labeled answer from imported dataset

Diego Gosmar

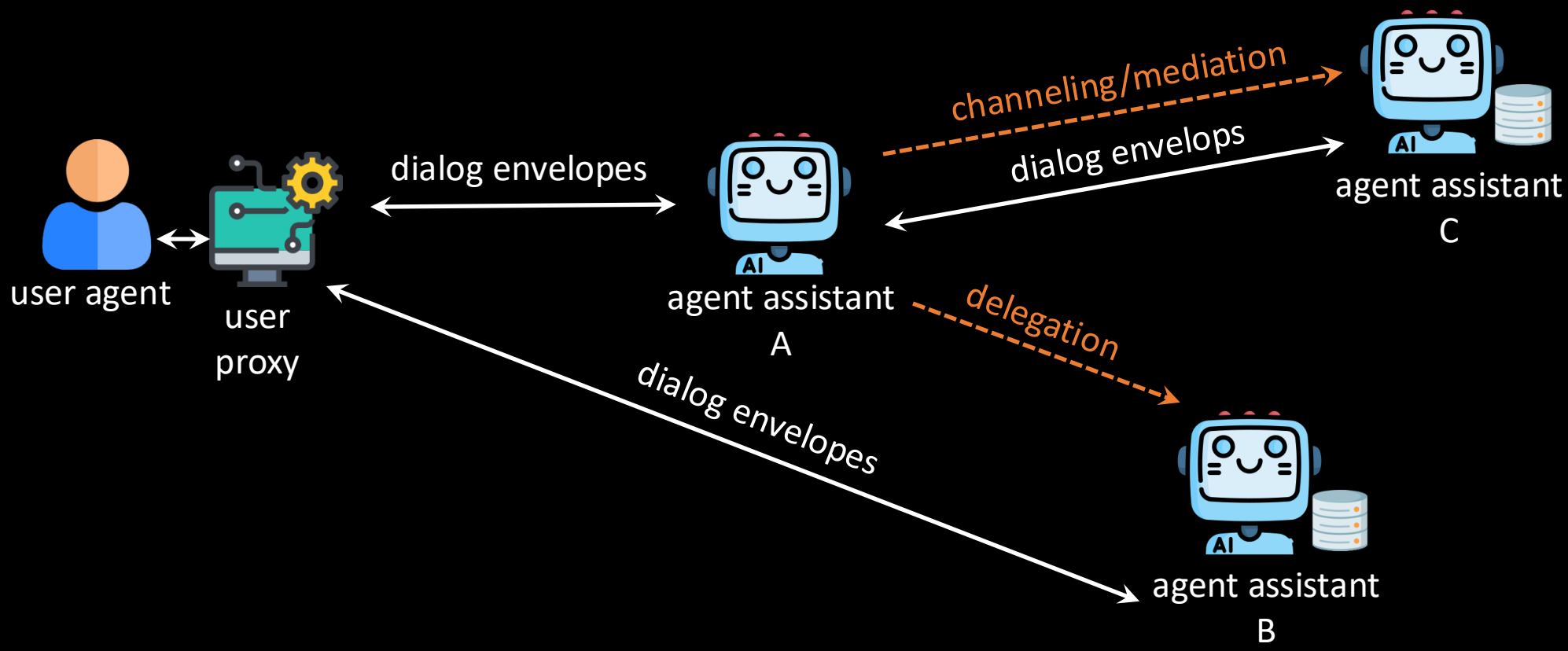


YAML feedback:
key: COT Contextual Accuracy
score: 1
value: CORRECT
comment: |-
The answer correctly identifies the blue whale as the largest mammal, which is the correct answer according to the context. The additional information about the size and weight of the blue whale is also correct and does not conflict with the context or the question.
GRADE: CORRECT
evaluator_info:
__run:
run_id: 91f6b719-2ea2-436c-a012-8a29f7a091ca

OPEN VOICE
INTEROPERABILITY

Dialog Agent Message Envelopes

Universal JSON structure



Agent A: 1st level dialog agent assistant

Agent B/C: usually with specific knowledge, and resource connections (i.e. CRM, specific KB etc…)

OPEN VOICE INTEROPERABILITY

Diego Gosmar

The Open Voice Network
Open Voice Interoperability Initiative - LF AI & Data Foundation
Architecture Work Group
3 July 2024
Draft Version 0.9.2 Status: Published

Editor-in-Chief: David Attwater
Contributors: Emmett Coin, Deborah Dahl, Jim Larson, Allan Wylie, Rainer Türner and Diego Gosmar

https://github.com/open-voice-interoperability/docs

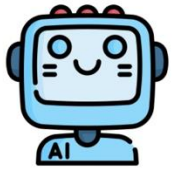
Conversational AI Multi-Agent Interoperability, Universal Open APIs for Agentic Natural Language Multimodal Communications

Diego Gosmar, Deborah A. Dahl, Emmett Coin

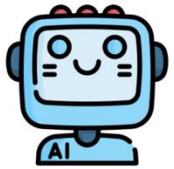
https://arxiv.org/abs/2407.19438

OPEN VOICE INTEROPERABILITY

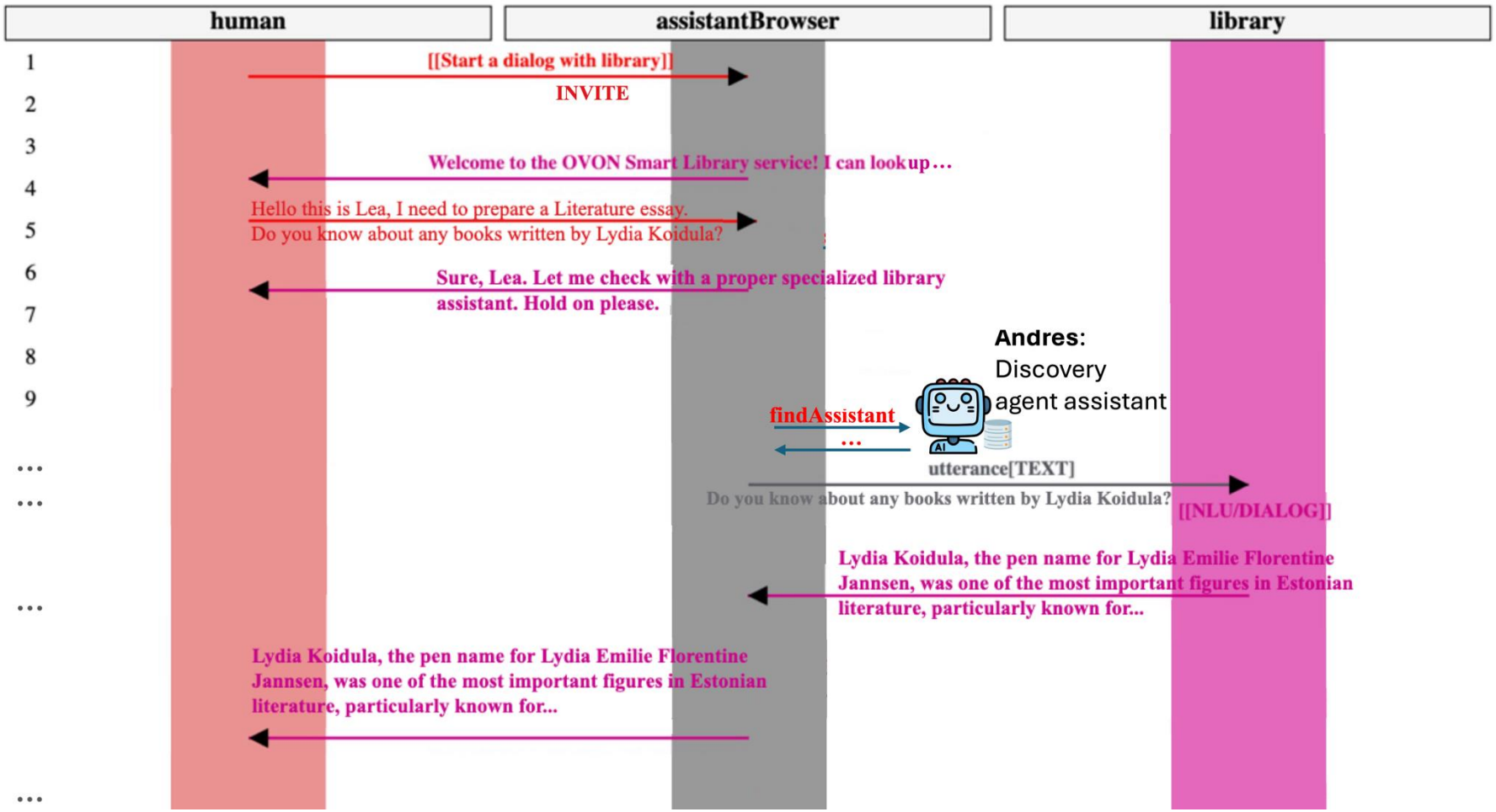
Lea: user agent (i.e. Estonian Citizen)



Juri: Frontend agent assistant



Kaja: Smart Library agent assistant



Example
OVON REQUEST
UTTERANCE & WHISPER

To be sent as part of the body of an HTTPS POST request to the endpoint: https://ovon.xcally.com/smartlibrary

Replace values with your desired questions about books:

utterance: main request
whisper: additional infos you'd like to send/request

OPEN VOICE INTEROPERABILITY

Diego Gosmar

Formatted according to the:
Conversational Envelope and Dialog Event specs


```
"ovon": {
"schema": {
"version": ”0.9.0",
...
{
"eventType": "invite",
"parameters": {
"to": {
"url": "https://ovon.xcally.com"
}
...
{
"eventType": "utterance",
"parameters": {
...
"mimeType": "text/plain",
"tokens": [ { "value": "Do you know about any books written by Lydia Koidula?" } ]
}
...
{
"eventType": "whisper",
"parameters": {
...
"features": {
"text": {
"mimeType": "text/plain",
"tokens": [ { "value": "Lydia Koidula, the pen name for Lydia Emilie Florentine Jannsen, was one of the most important figures in Estonian literature, particularly known for..." } ]
```

“It’s a Wrap-up”

Thank you!
https://www.linkedin.com/in/diegogosmar