LLM Observability

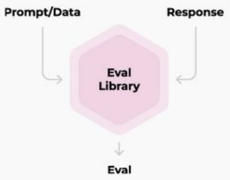
Amber Roberts, ML Growth Lead @ Arize AI

arize

LLM Observability Pillars

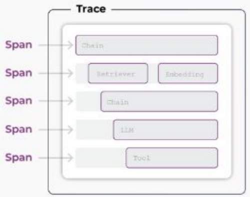
Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM



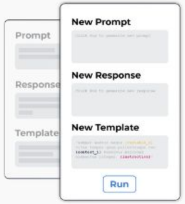
Traces & Spans

Visibility into where the agentic workflow broke



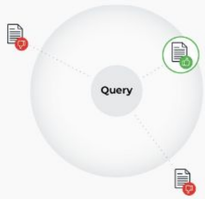
Prompt Engineering

Iterating on prompt templates for improved results



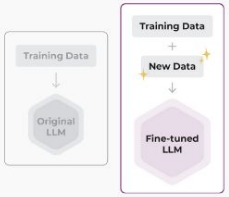
Search & Retrieval

Locate and improve retrieved context



Fine-tuning

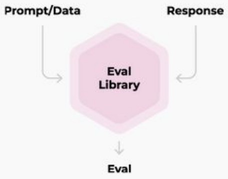
Re-train LLM on use case / company data



LLM Observability Pillars

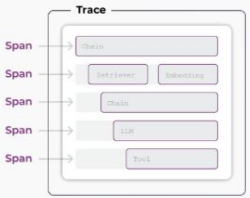
Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM



Traces & Spans

Visibility into where the agentic workflow broke



Prompt Engineering

Iterating on prompt templates for improved results

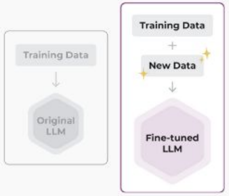
Prompt
Response
Template
New Prompt
New Response
New Template
Run

Search & Retrieval

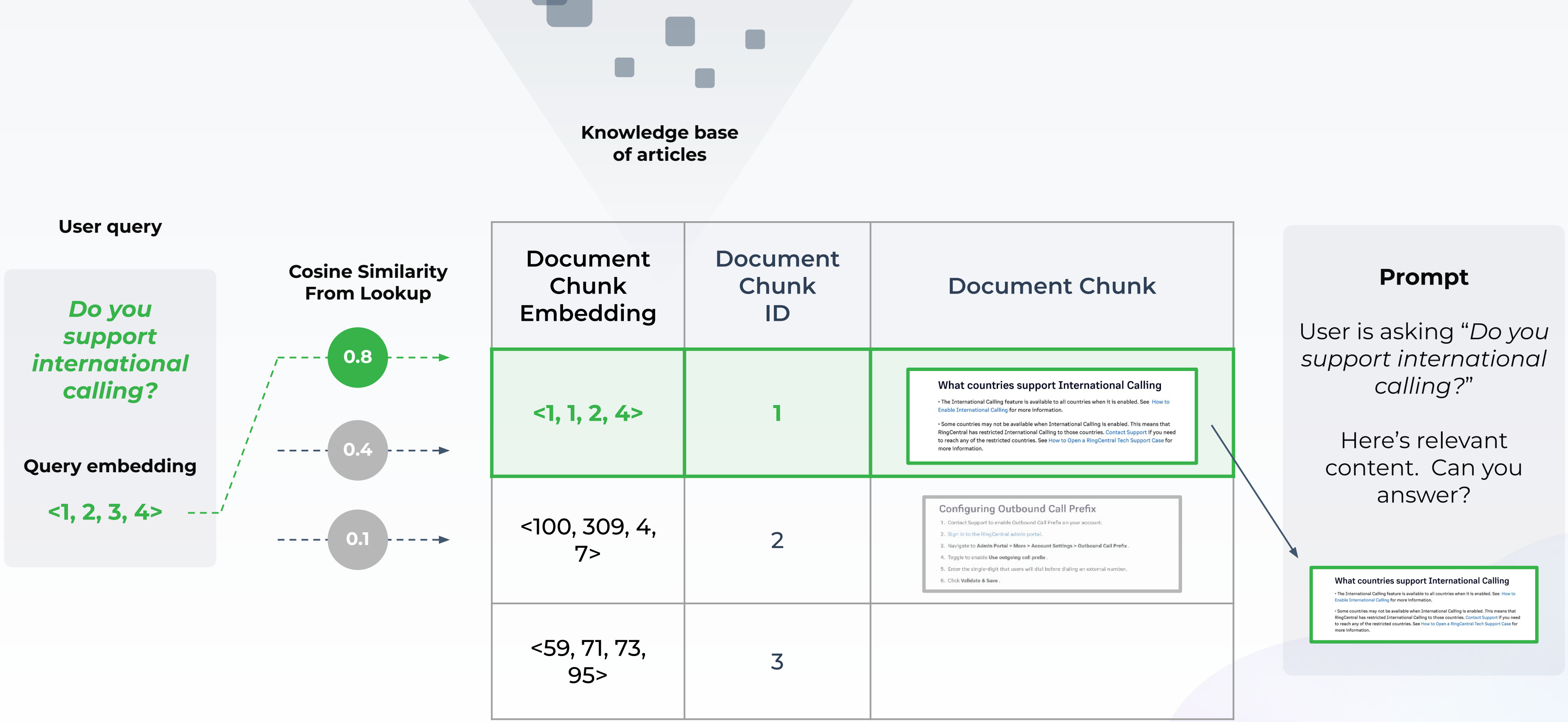
Locate and improve retrieved context

Fine-tuning

Re-train LLM on use case / company data



How does RAG (Retrieval Augmented Generation) work?



Knowledge base of articles

User query

Do you support international calling?

Query embedding

<1, 2, 3, 4>

Cosine Similarity From Lookup

Document Chunk Embedding	Document Chunk ID	Document Chunk
<1, 1, 2, 4>	1	**What countries support International Calling** • The International Calling feature is available to all countries when it is enabled. See How to Enable International Calling for more information. • Some countries may not be available when International Calling is enabled. This means that RingCentral has restricted International Calling to those countries. Contact Support if you need to reach any of the restricted countries. See How to Open a RingCentral Tech Support Case for more information.
<100, 309, 4, 7>	2	Configuring Outbound Call Prefix 1. Contact Support to enable Outbound Call Prefix on your account. 2. Sign in to the RingCentral admin portal. 3. Navigate to **Admin Portal > More > Account Settings > Outbound Call Prefix** . 4. Toggle to enable **Use outgoing call prefix** . 5. Enter the single-digit that users will dial before dialing an external number. 6. Click **Validate & Save** .
<59, 71, 73, 95>	3	

Prompt

User is asking "*Do you support international calling?*"

Here's relevant content. Can you answer?

What countries support International Calling

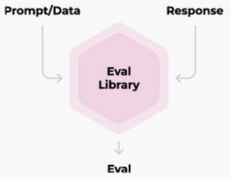
• The International Calling feature is available to all countries when it is enabled. See How to Enable International Calling for more information.

• Some countries may not be available when International Calling is enabled. This means that RingCentral has restricted International Calling to those countries. Contact Support if you need to reach any of the restricted countries. See How to Open a RingCentral Tech Support Case for more information.

LLM Observability Pillars

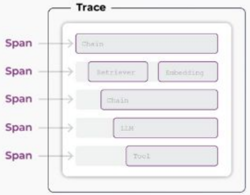
Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM



Traces & Spans

Visibility into where the agentic workflow broke



Prompt Engineering

Iterating on prompt templates for improved results

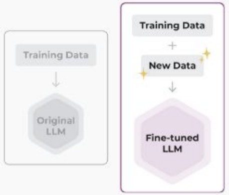
Prompt
Response
Template
New Prompt
New Response
New Template
Run

Search & Retrieval

Locate and improve retrieved context

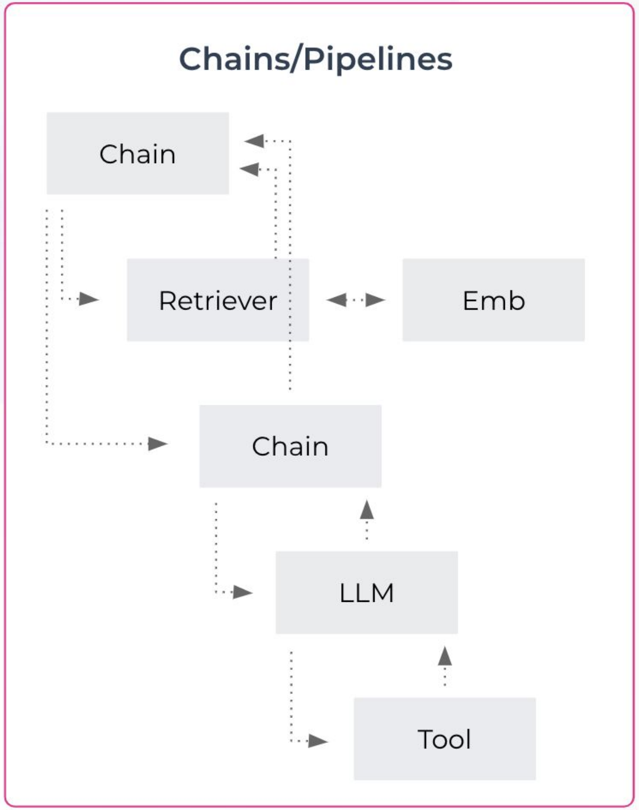
Fine-tuning

Re-train LLM on use case / company data



Traces and Spans

- LLM usage is documented in a callback system by a trace.
- In this trace a span can refer to any unit of execution
- You may annotate a span with a specific name or a general term like a chain.



Span Types

LLM
Call to a LLM for completion or chat

Chain
Link between application steps

Tool
API or Function invoked on behalf of an LLM

Agent
Root of a set of LLM and Tool invocations

Embedding
Encoding of unstructured data

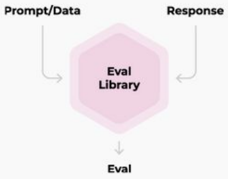
Retriever
Query for context from a data store

Reranker
Relevance-based re-ordering of documents

LLM Observability Pillars

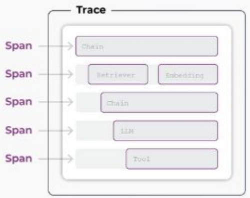
Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM



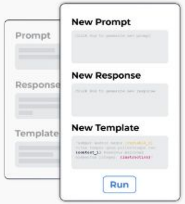
Traces & Spans

Visibility into where the agentic workflow broke



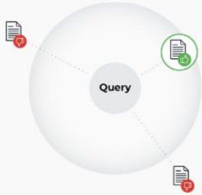
Prompt Engineering

Iterating on prompt templates for improved results



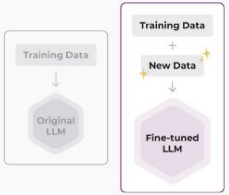
Search & Retrieval

Locate and improve retrieved context



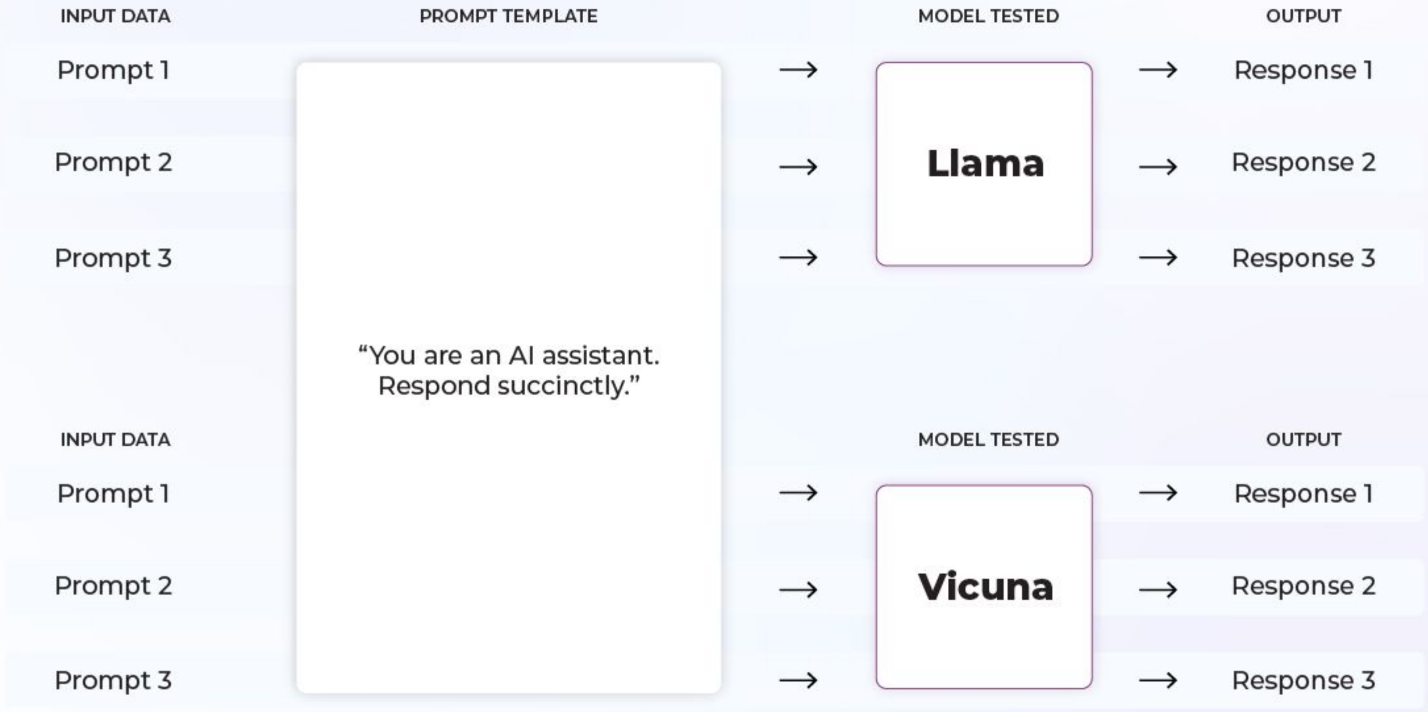
Fine-tuning

Re-train LLM on use case / company data



Model vs. System Evals

LLM Model Evals



INPUT DATA

Prompt 1

Prompt 2

Prompt 3

PROMPT TEMPLATE

“You are an AI assistant. Respond succinctly.”

MODEL TESTED

Llama

OUTPUT

Response 1

Response 2

Response 3

INPUT DATA

Prompt 1

Prompt 2

Prompt 3

MODEL TESTED

Vicuna

OUTPUT

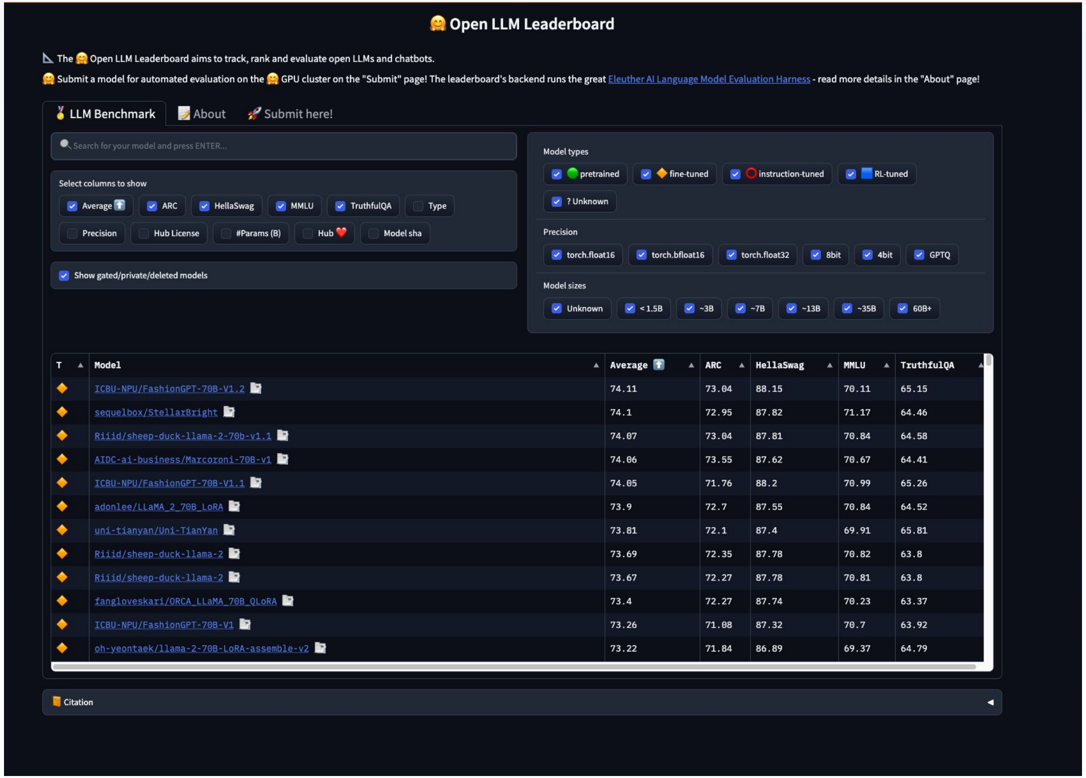
Response 1

Response 2

Response 3

OpenAI Eval library

- The OpenAI Eval library and LLM leaderboards look at how well the various LLMs stack up against each other.
- Using LLM evaluation metrics like:
    - HellaSwag – how well an LLM can complete a sentence
    - TruthfulQA - truthfulness of responses
    - MMLU - how well the LLM can multitask.



🤗 Open LLM Leaderboard

📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🤗 Submit a model for automated evaluation on the 🤗 GPU cluster on the "Submit" page! The leaderboard's backend runs the great Eleuther AI Language Model Evaluation Harness - read more details in the "About" page!

Select columns to show

☑ Average ⬆️ ☑ ARC ☑ HellaSwag ☑ MMLU ☑ TruthfulQA □ Type

□ Precision □ Hub License □ #Params (B) □ Hub ❤️ □ Model sha

☑ Show gated/private/deleted models

Model types

☑ 🟢 pretrained ☑ 🔶 fine-tuned ☑ ⭕ instruction-tuned ☑ 🟦 RL-tuned

☑ ? Unknown

Precision

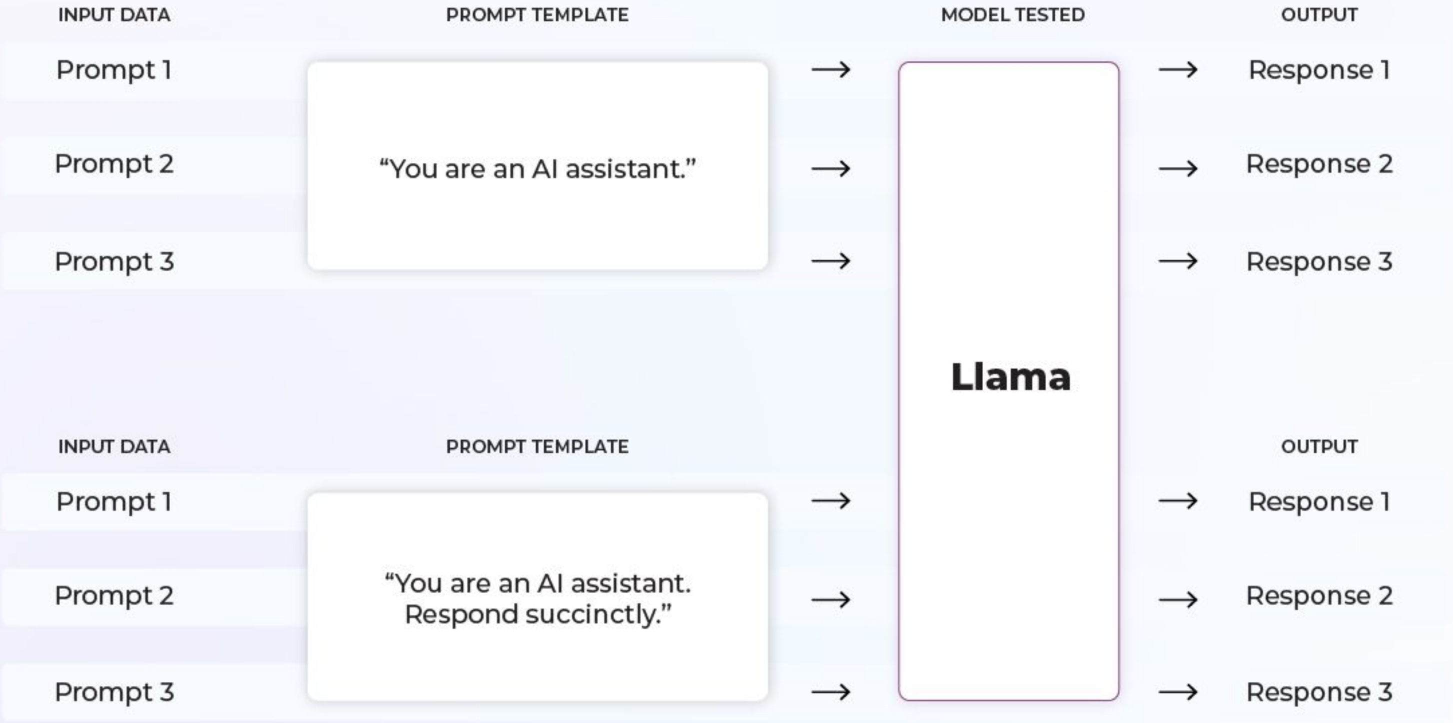
☑ torch.float16 ☑ torch.bfloat16 ☑ torch.float32 ☑ 8bit ☑ 4bit ☑ GPTQ

Model sizes

☑ Unknown ☑ < 1.5B ☑ ~3B ☑ ~7B ☑ ~13B ☑ ~35B ☑ 60B+

T	Model	Average ⬆️	ARC	HellaSwag	MMLU	TruthfulQA
🔶	ICBU-NPU/FashionGPT-70B-V1.2	74.11	73.04	88.15	70.11	65.15
🔶	sequelbox/StellarBright	74.1	72.95	87.82	71.17	64.46
🔶	Riiid/sheep-duck-llama-2-70b-v1.1	74.07	73.04	87.81	70.84	64.58
🔶	AIDC-ai-business/Marcoroni-70B-v1	74.06	73.55	87.62	70.67	64.41
🔶	ICBU-NPU/FashionGPT-70B-V1.1	74.05	71.76	88.2	70.99	65.26
🔶	adonlee/LLaMA_2_70B_LoRA	73.9	72.7	87.55	70.84	64.52
🔶	uni-tianyan/Uni-TianYan	73.81	72.1	87.4	69.91	65.81
🔶	Riiid/sheep-duck-llama-2	73.69	72.35	87.78	70.82	63.8
🔶	Riiid/sheep-duck-llama-2	73.67	72.27	87.78	70.81	63.8
🔶	fangloveskari/ORCA_LLaMA_70B_QLoRA	73.4	72.27	87.74	70.23	63.37
🔶	ICBU-NPU/FashionGPT-70B-V1	73.26	71.08	87.32	70.7	63.92
🔶	oh-yeontaek/llama-2-70B-LoRA-assemble-v2	73.22	71.84	86.89	69.37	64.79

LLM System Evals



INPUT DATA

Prompt 1

Prompt 2

Prompt 3

PROMPT TEMPLATE

"You are an AI assistant."

MODEL TESTED

Llama

OUTPUT

Response 1

Response 2

Response 3

INPUT DATA

Prompt 1

Prompt 2

Prompt 3

PROMPT TEMPLATE

"You are an AI assistant.
Respond succinctly."

OUTPUT

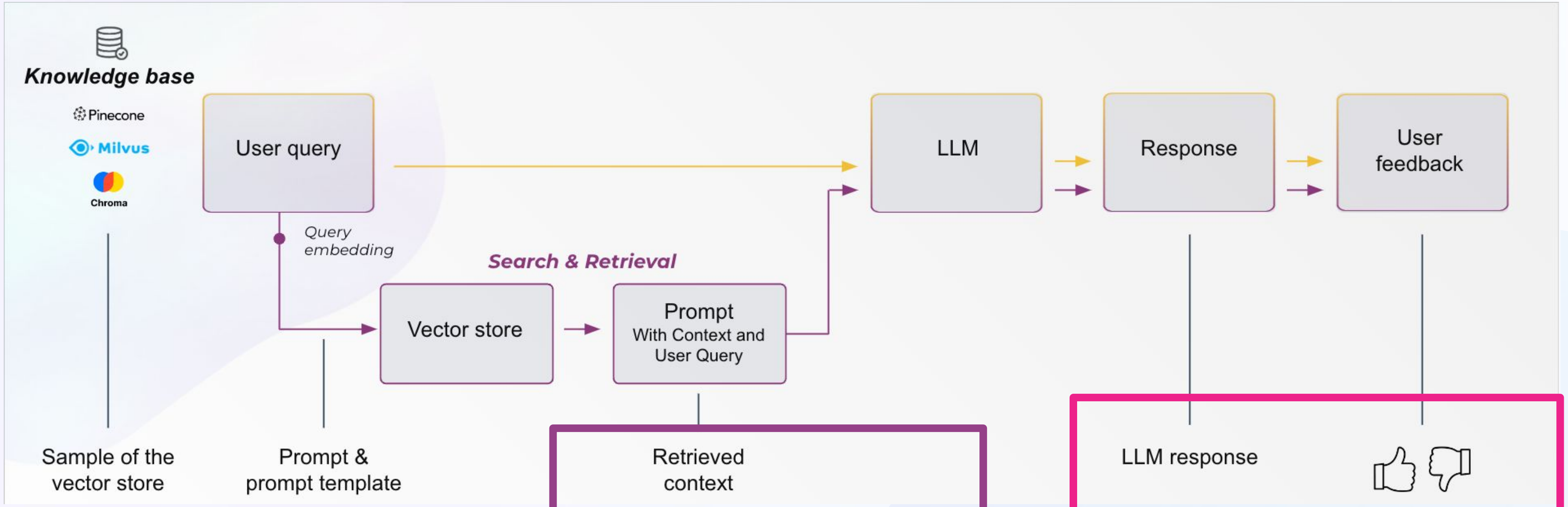
Response 1

Response 2

Response 3

Response vs. Retrieval Evals

Retrieval Evals vs Response Evals



Knowledge base

Pinecone

Milvus

Chroma

User query

Query embedding

Search & Retrieval

Vector store

Prompt
With Context and User Query

LLM

Response

User feedback

Sample of the vector store

Prompt & prompt template

Retrieved context

Precision @4 = 80%
NDCG = 75%
Hit = True

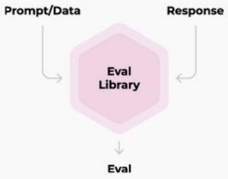
LLM response

Hallucinations = False
Correctness = True

Putting it all together

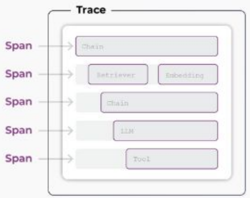
Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM



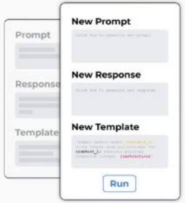
Traces & Spans

Visibility into where the agentic workflow broke



Prompt Engineering

Iterating on prompt templates for improved results

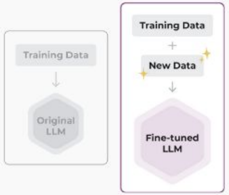


Search & Retrieval

Locate and improve retrieved context

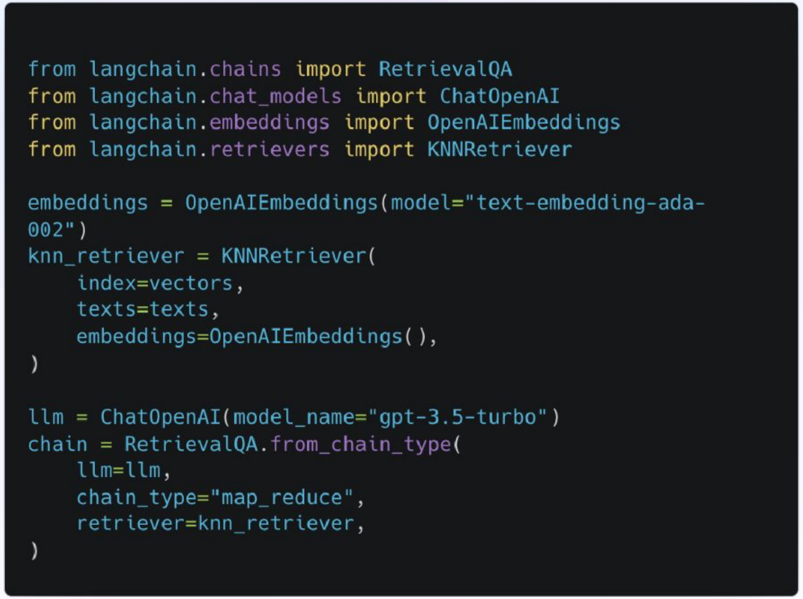
Fine-tuning

Re-train LLM on use case / company data



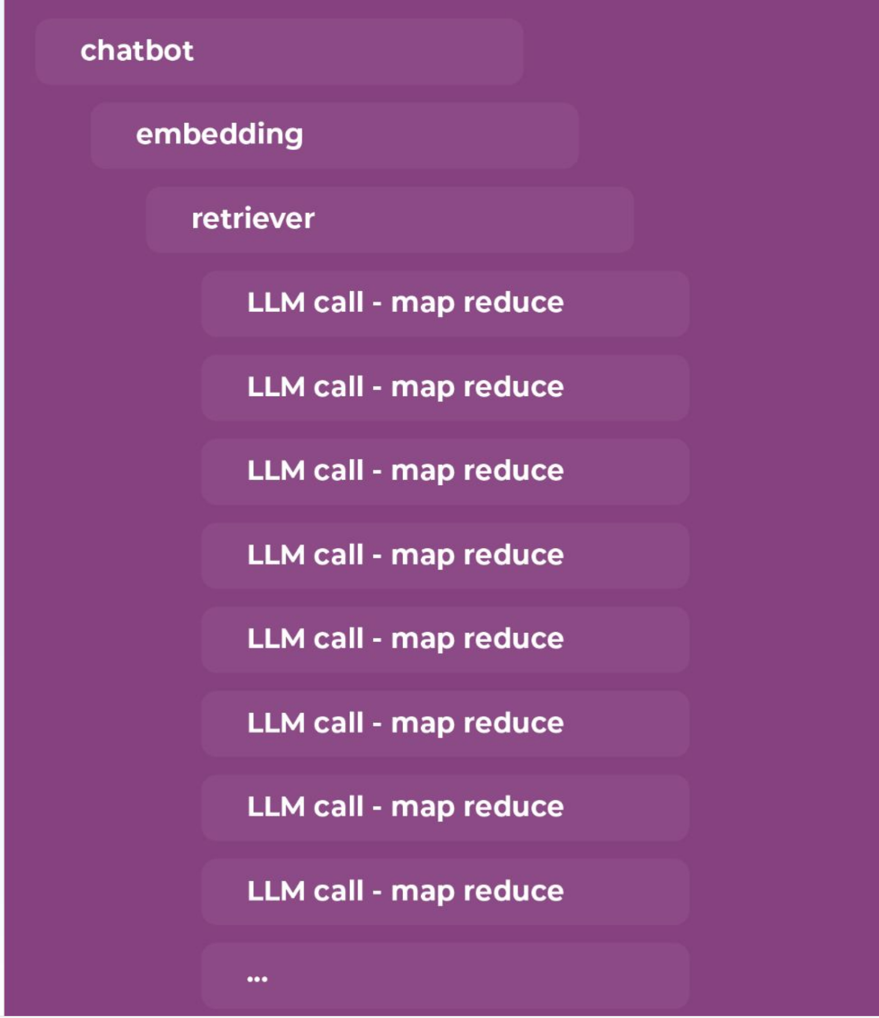
Arize Phoenix Overview

4–5 lines of code ⟶ 10+ LLM calls



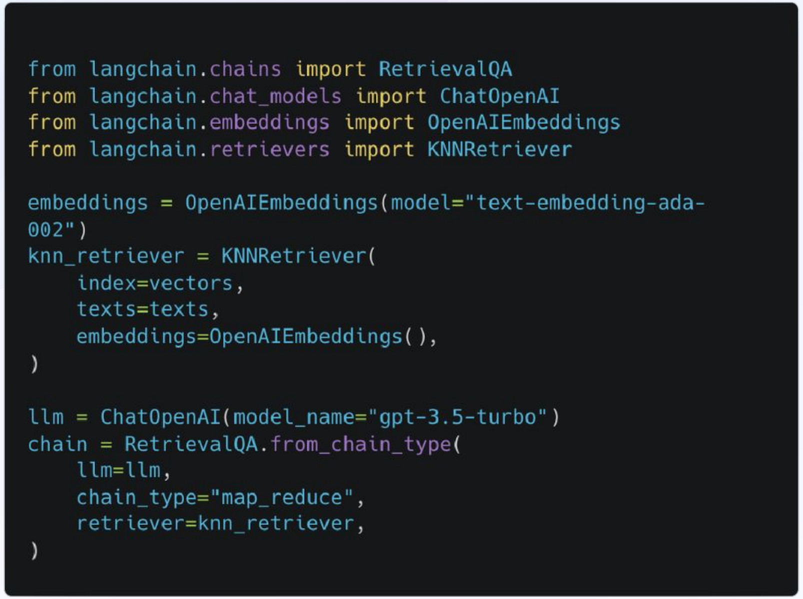
```
from langchain.chains import RetrievalQA
from langchain.chat_models import ChatOpenAI
from langchain.embeddings import OpenAIEmbeddings
from langchain.retrievers import KNNRetriever

embeddings = OpenAIEmbeddings(model="text-embedding-ada-
002")
knn_retriever = KNNRetriever(
    index=vectors,
    texts=texts,
    embeddings=OpenAIEmbeddings(),
)

llm = ChatOpenAI(model_name="gpt-3.5-turbo")
chain = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="map_reduce",
    retriever=knn_retriever,
)
```

Making sense of a large number of distributed system calls is what Phoenix is designed to do.



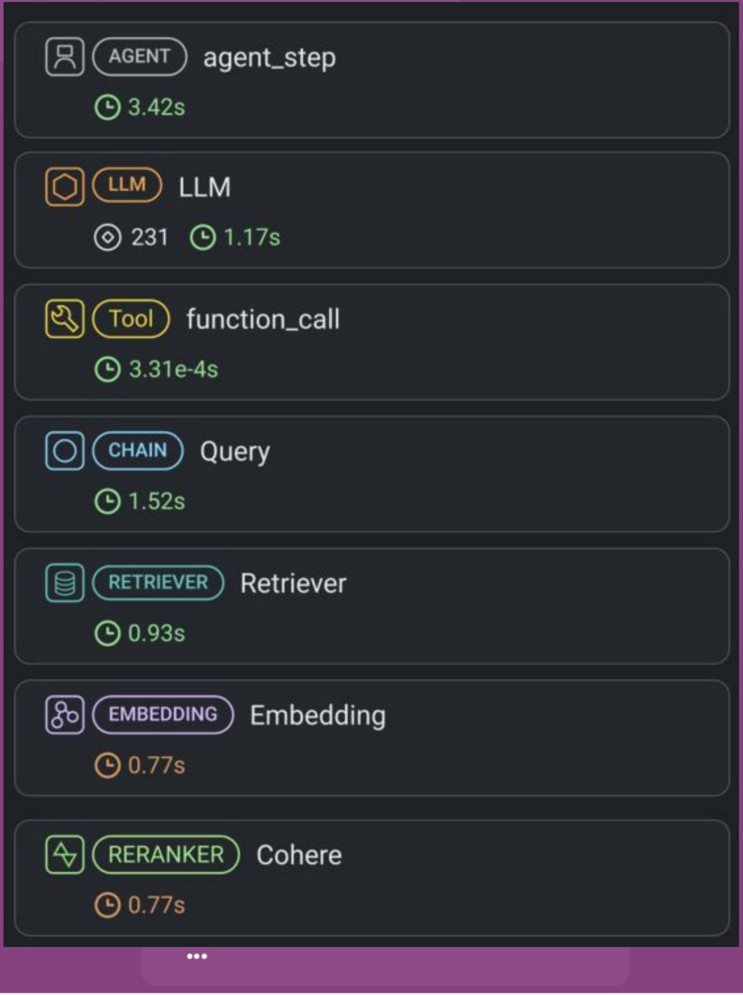
Arize Phoenix Overview

4–5 lines of code ⟶ 10+ LLM calls



```
from langchain.chains import RetrievalQA
from langchain.chat_models import ChatOpenAI
from langchain.embeddings import OpenAIEmbeddings
from langchain.retrievers import KNNRetriever

embeddings = OpenAIEmbeddings(model="text-embedding-ada-
002")
knn_retriever = KNNRetriever(
    index=vectors,
    texts=texts,
    embeddings=OpenAIEmbeddings(),
)

llm = ChatOpenAI(model_name="gpt-3.5-turbo")
chain = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="map_reduce",
    retriever=knn_retriever,
)
```

Making sense of a large number of distributed system calls is what Phoenix is designed to do.



AGENT agent_step
3.42s

LLM LLM
231 1.17s

Tool function_call
3.31e-4s

CHAIN Query
1.52s

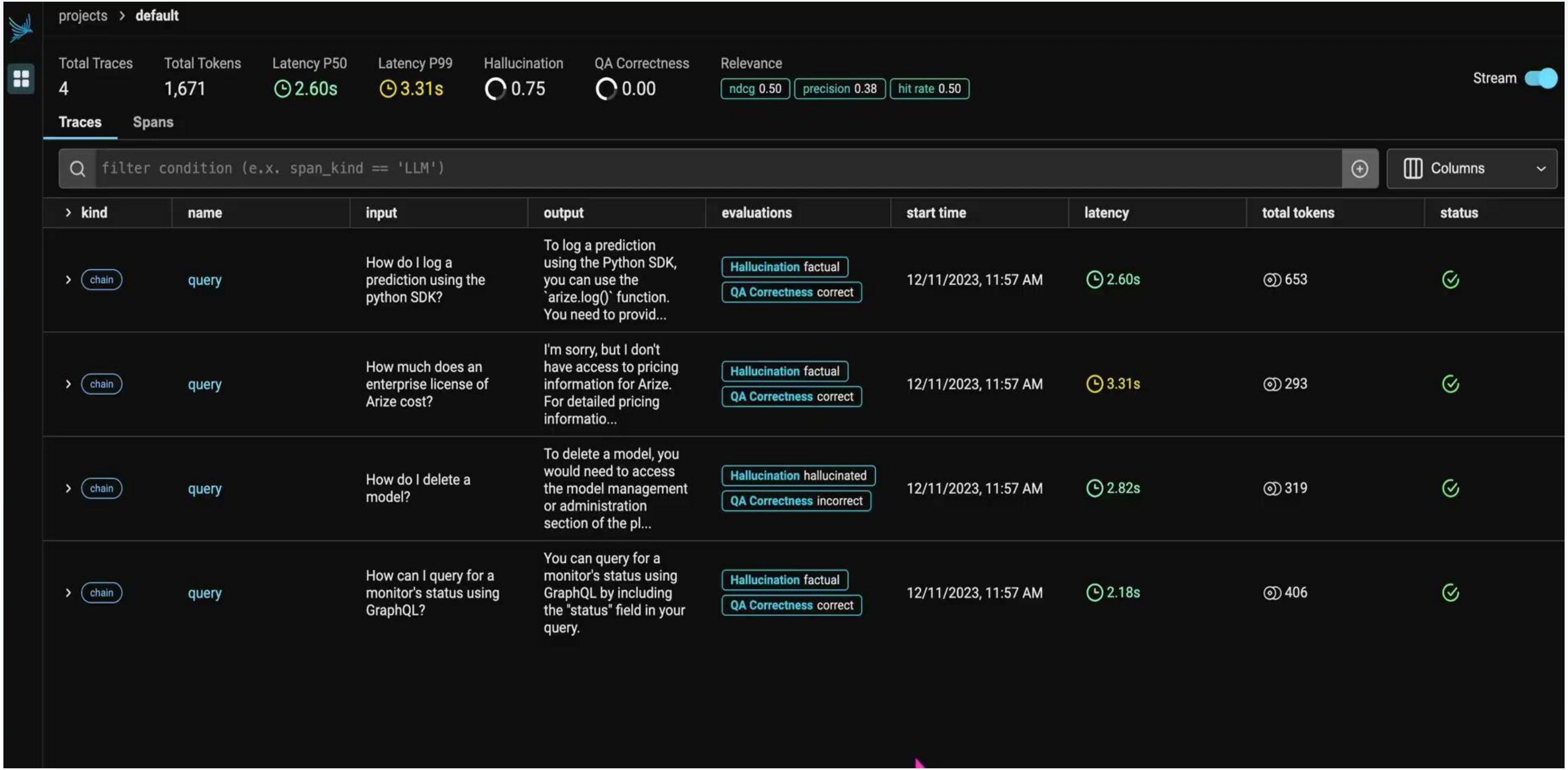
RETRIEVER Retriever
0.93s

EMBEDDING Embedding
0.77s

RERANKER Cohere
0.77s

...

Phoenix Demo



projects > **default**

Total Traces 4 | Total Tokens 1,671 | Latency P50 2.60s | Latency P99 3.31s | Hallucination 0.75 | QA Correctness 0.00 | Relevance ndcg 0.50 precision 0.38 hit rate 0.50

Stream

Traces Spans

```
filter condition (e.x. span_kind == 'LLM')
```

Columns

kind	name	input	output	evaluations	start time	latency	total tokens	status
chain	query	How do I log a prediction using the python SDK?	To log a prediction using the Python SDK, you can use the `arize.log()` function. You need to provid...	Hallucination factual QA Correctness correct	12/11/2023, 11:57 AM	2.60s	653	✓
chain	query	How much does an enterprise license of Arize cost?	I'm sorry, but I don't have access to pricing information for Arize. For detailed pricing informatio...	Hallucination factual QA Correctness correct	12/11/2023, 11:57 AM	3.31s	293	✓
chain	query	How do I delete a model?	To delete a model, you would need to access the model management or administration section of the pl...	Hallucination hallucinated QA Correctness incorrect	12/11/2023, 11:57 AM	2.82s	319	✓
chain	query	How can I query for a monitor's status using GraphQL?	You can query for a monitor's status using GraphQL by including the "status" field in your query.	Hallucination factual QA Correctness correct	12/11/2023, 11:57 AM	2.18s	406	✓

Thank you.

Sign up for a free account at:

arize.com/join

Check out Phoenix, our OSS tool, at:

phoenix.arize.com