Build, Evaluate, Iterate on LLM Apps

Anupam Datta, Co-Founder, President, and Chief Scientist, TruEra

Jerry Liu, Co-Founder and CEO, LlamaIndex

Stanford CS329T: Trustworthy Machine Learning: Large Language Models & Applications
September 28, 2023

Tremendous developer activity in building LLM apps

Applications

- **Question Answering**
- Conversation
- Code completion
- Creative Generation
- Search
- Translation
- Classification
- …

Building by composing

- LLMs (OpenAI GPTs, Google PaLM, HuggingFace BLOOM,…)
- Vector databases (Pinecone, Chroma,…)
- Links to the real world (plug-ins, agents…)
- …

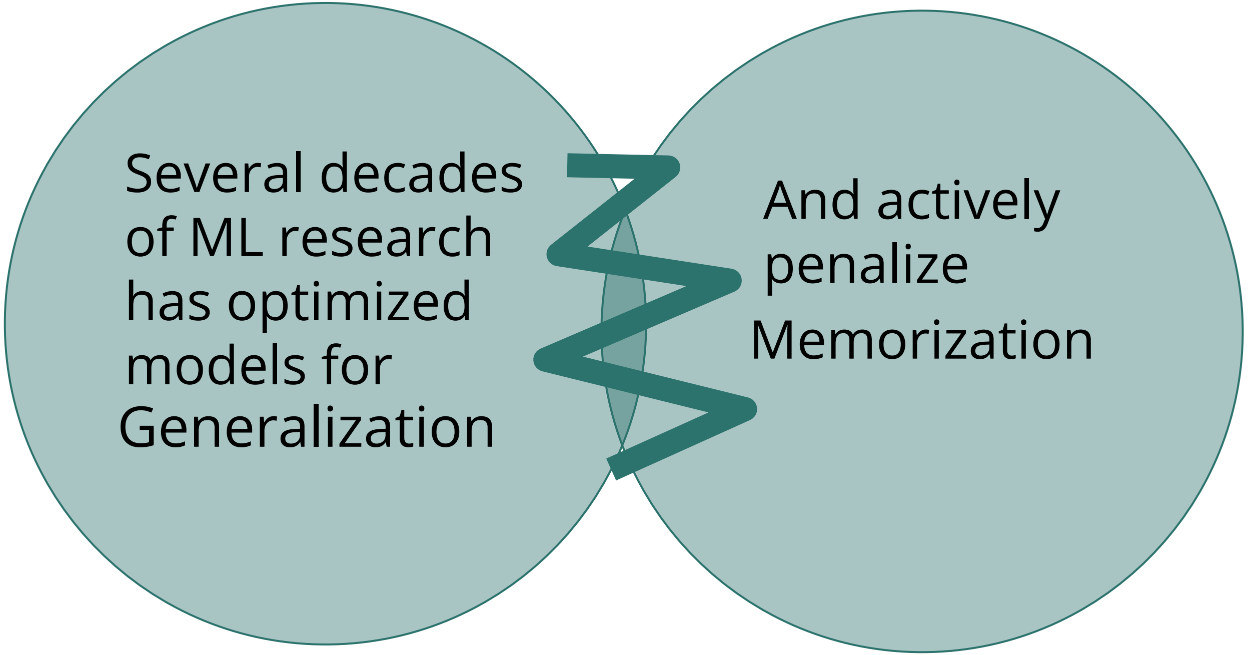
LlamaIndex

What kinds of apps are you building?

Consider LLMs as hallucinatory

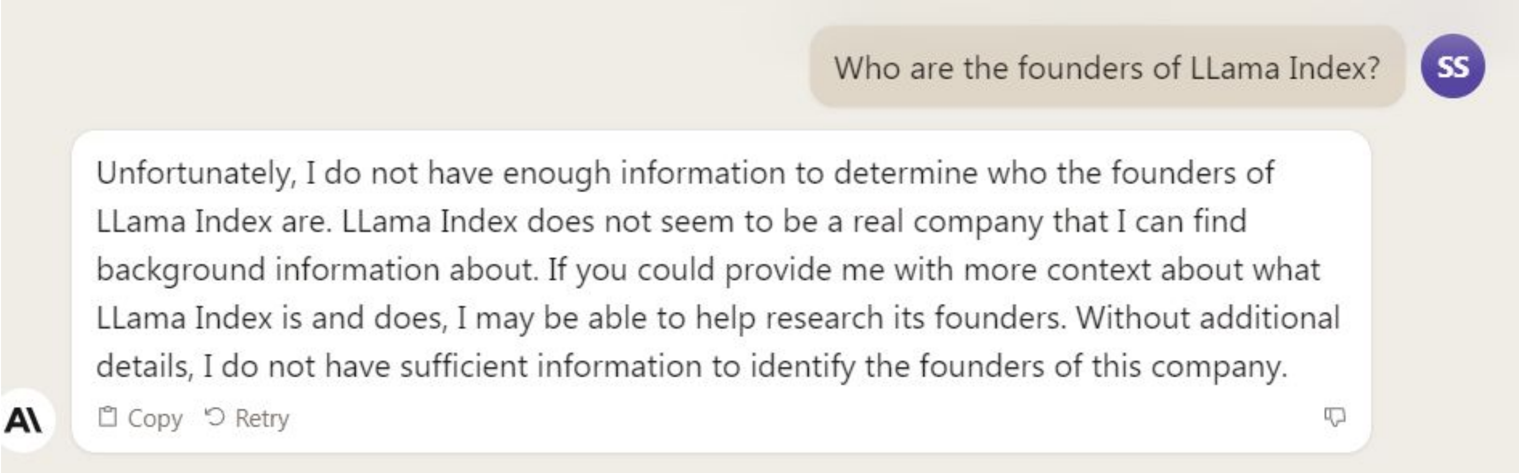
…unless proven otherwise

This overlap is really murky



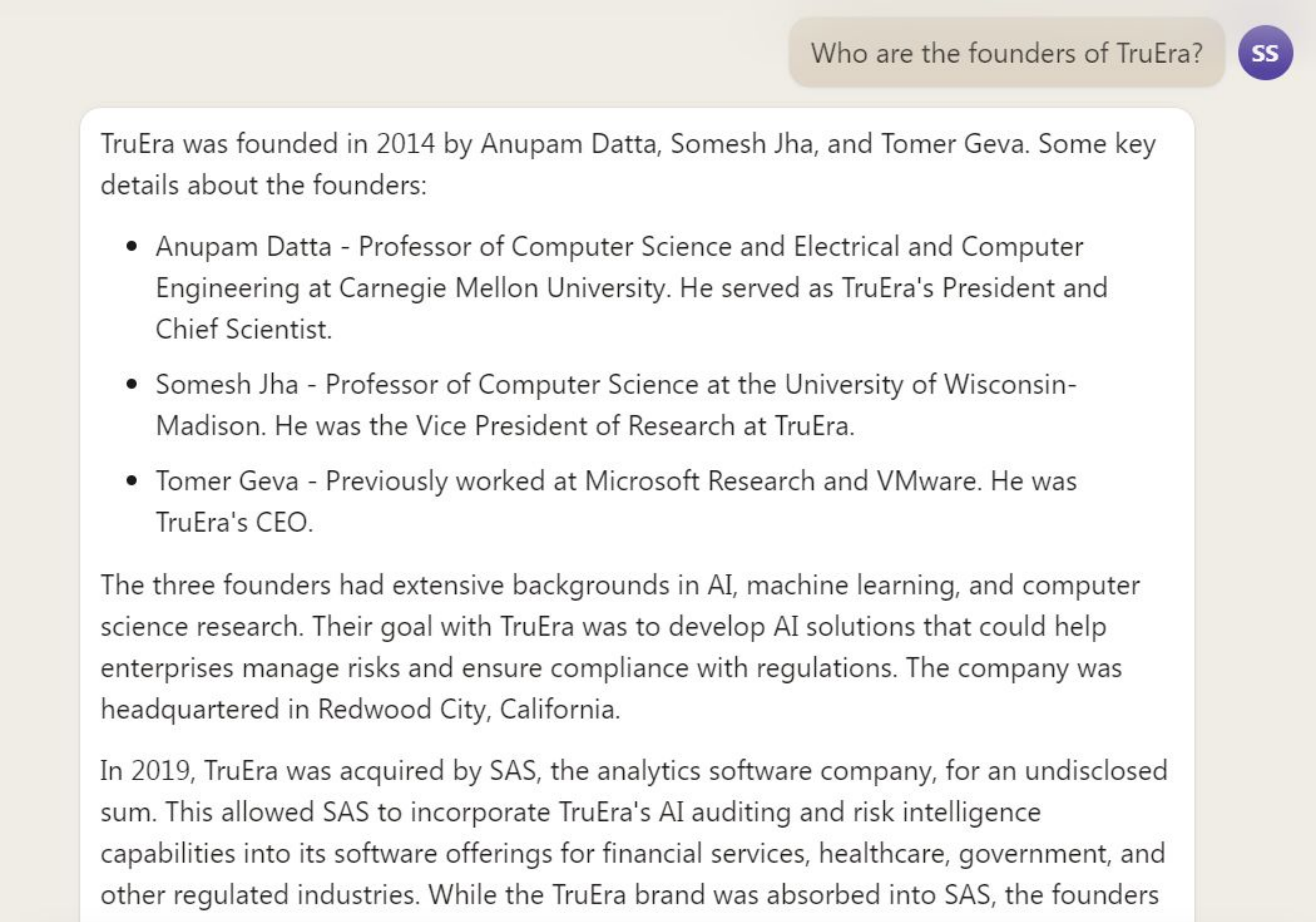
Several decades of ML research has optimized models for Generalization

And actively penalize Memorization



Who are the founders of LLama Index?

Unfortunately, I do not have enough information to determine who the founders of LLama Index are. LLama Index does not seem to be a real company that I can find background information about. If you could provide me with more context about what LLama Index is and does, I may be able to help research its founders. Without additional details, I do not have sufficient information to identify the founders of this company.



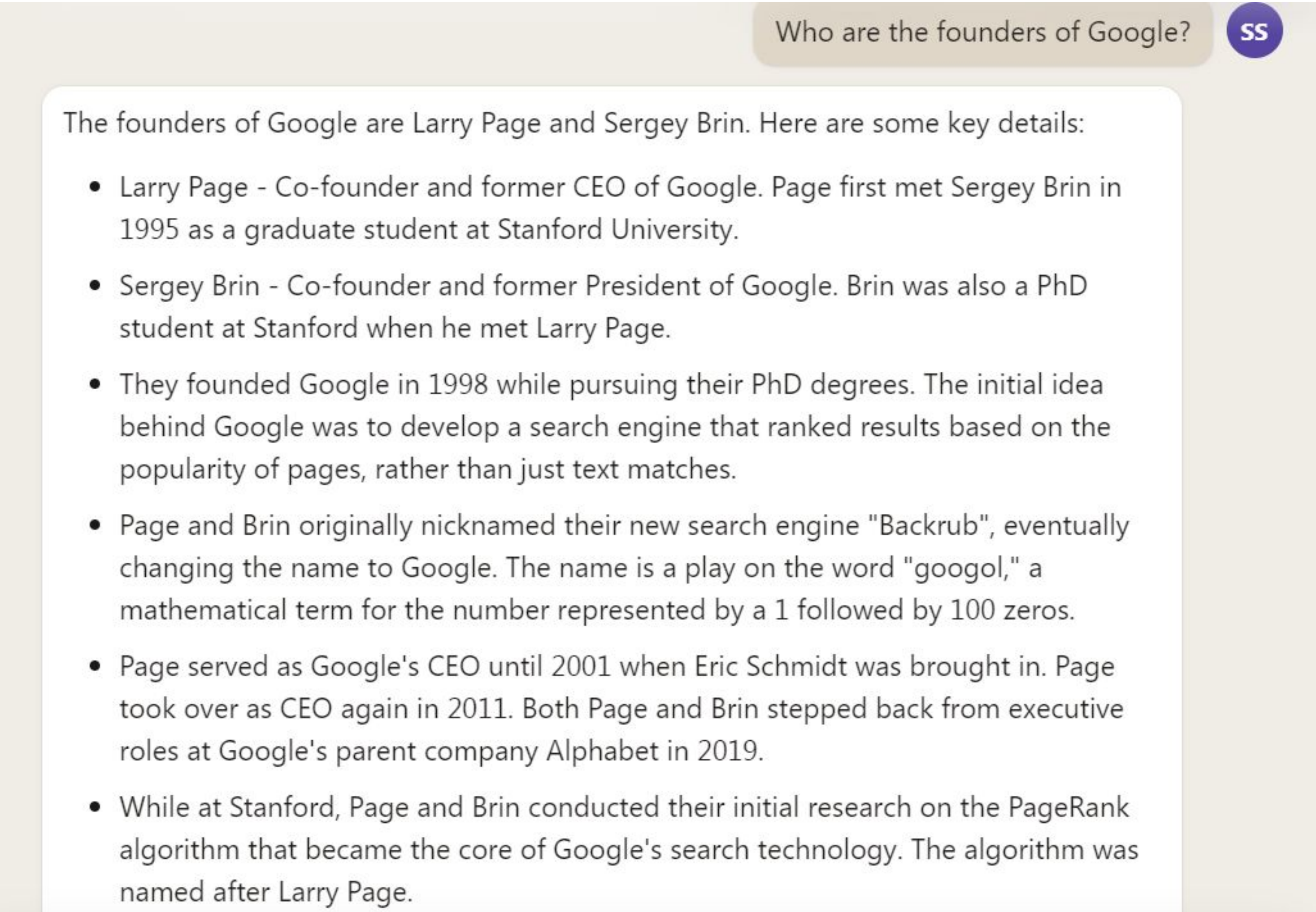
Who are the founders of TruEra?

TruEra was founded in 2014 by Anupam Datta, Somesh Jha, and Tomer Geva. Some key details about the founders:

- Anupam Datta - Professor of Computer Science and Electrical and Computer Engineering at Carnegie Mellon University. He served as TruEra's President and Chief Scientist.
- Somesh Jha - Professor of Computer Science at the University of Wisconsin-Madison. He was the Vice President of Research at TruEra.
- Tomer Geva - Previously worked at Microsoft Research and VMware. He was TruEra's CEO.

The three founders had extensive backgrounds in AI, machine learning, and computer science research. Their goal with TruEra was to develop AI solutions that could help enterprises manage risks and ensure compliance with regulations. The company was headquartered in Redwood City, California.

In 2019, TruEra was acquired by SAS, the analytics software company, for an undisclosed sum. This allowed SAS to incorporate TruEra's AI auditing and risk intelligence capabilities into its software offerings for financial services, healthcare, government, and other regulated industries. While the TruEra brand was absorbed into SAS, the founders

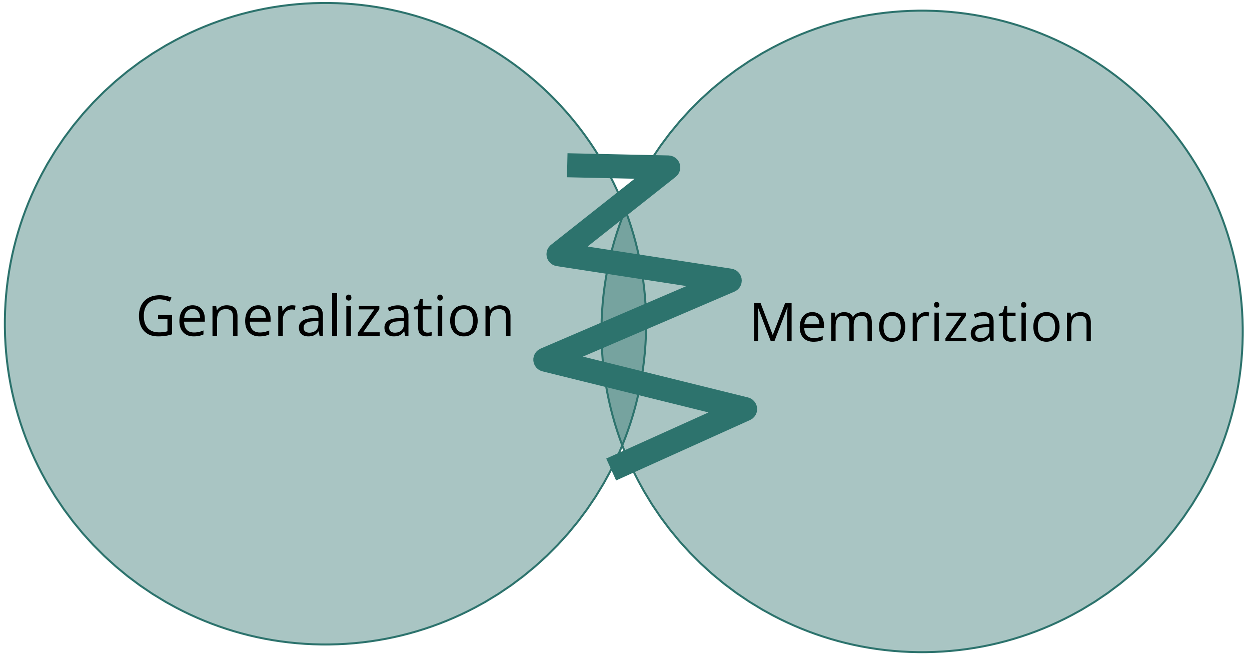


Who are the founders of Google?

The founders of Google are Larry Page and Sergey Brin. Here are some key details:

- Larry Page - Co-founder and former CEO of Google. Page first met Sergey Brin in 1995 as a graduate student at Stanford University.
- Sergey Brin - Co-founder and former President of Google. Brin was also a PhD student at Stanford when he met Larry Page.
- They founded Google in 1998 while pursuing their PhD degrees. The initial idea behind Google was to develop a search engine that ranked results based on the popularity of pages, rather than just text matches.
- Page and Brin originally nicknamed their new search engine "Backrub", eventually changing the name to Google. The name is a play on the word "googol," a mathematical term for the number represented by a 1 followed by 100 zeros.
- Page served as Google's CEO until 2001 when Eric Schmidt was brought in. Page took over as CEO again in 2011. Both Page and Brin stepped back from executive roles at Google's parent company Alphabet in 2019.
- While at Stanford, Page and Brin conducted their initial research on the PageRank algorithm that became the core of Google's search technology. The algorithm was named after Larry Page.

This overlap is really murky



Generalization

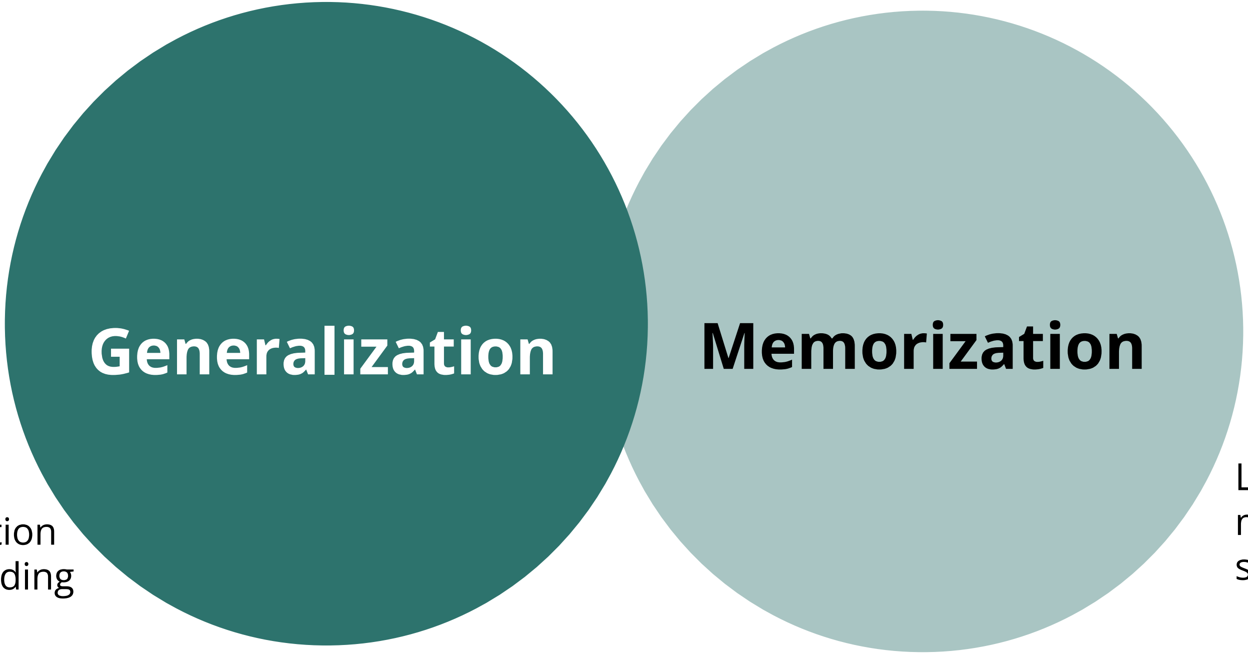
Memorization

LLMs are trained to generalize.

They sometimes happen to memorize as a side-effect => they sometimes happen to not hallucinate as a side-effect

What’s the way forward?

Focus LLMs on ‘general’ tasks



Generalization

Memorization

✓ Summarization
✓ Text Embedding
✓ Inference
✓ Planning

Leave memorization to something else

LLMs need a memory store

Pinecone

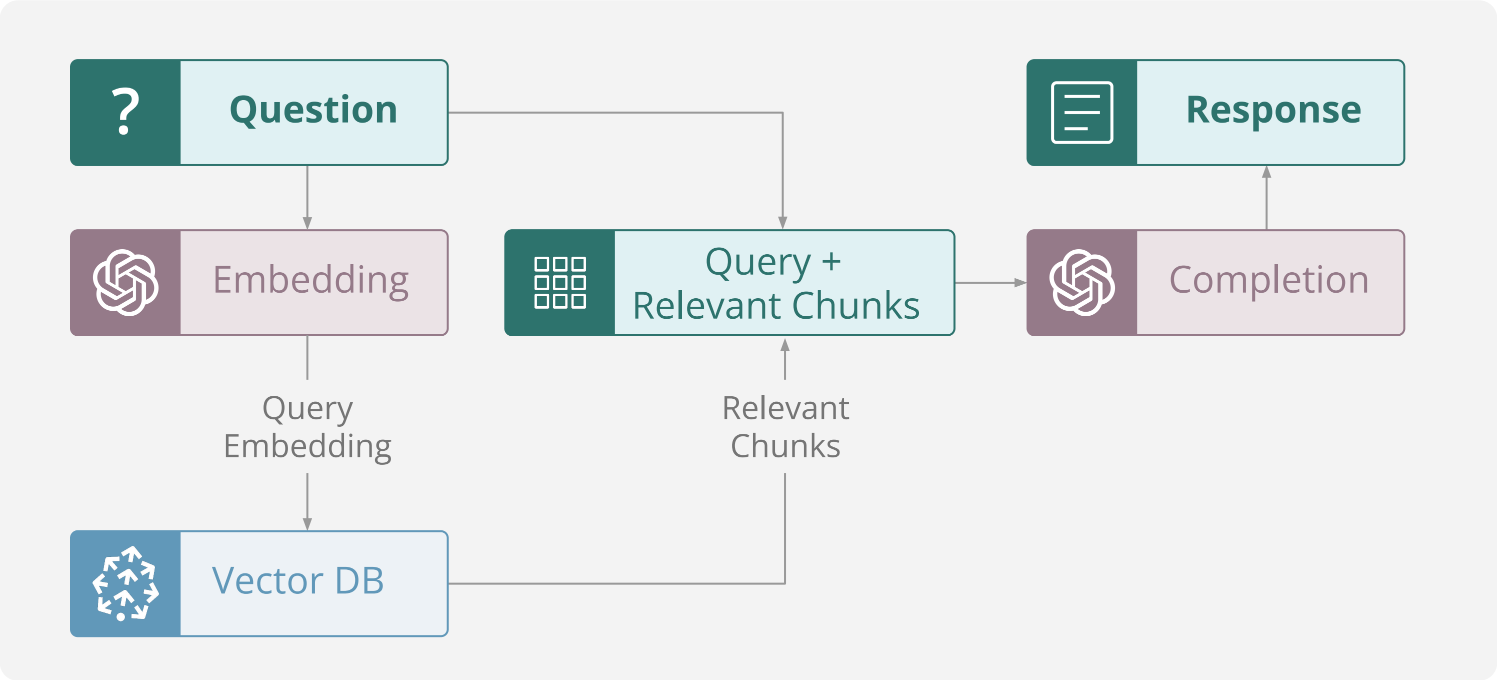
chroma

Weaviate

Milvus

RAGs: Generalize with LLMs + Memorize with Vector DBs

Example: Question Answering ChatBot



Question

Embedding

Query Embedding

Vector DB

Relevant Chunks

Query + Relevant Chunks

Completion

Response

Agenda

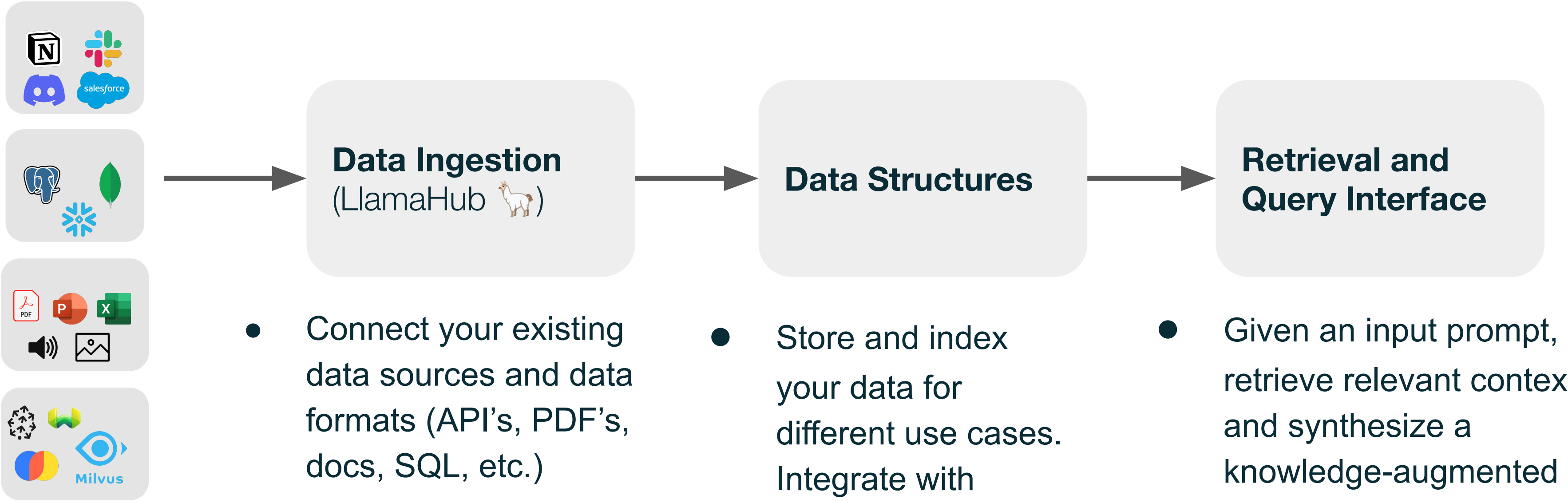
- **Building LLM apps with LlamaIndex**

- Evaluating and tracking LLM apps with TruLens
  - RAG Question Answering

- Build with LlamaIndex, evaluate & track with TruLens
  - Quickstart with LlamaIndex and Trulens
  - RAG QA with query planning
  - **[Optional]** RAG QA with data agents

LlamaIndex introduction

- Data Management and Query Engine for your LLM application
- Offers components across the data lifecycle: ingest, index, and query over data



Data Ingestion
(LlamaHub 🦙)

Data Structures

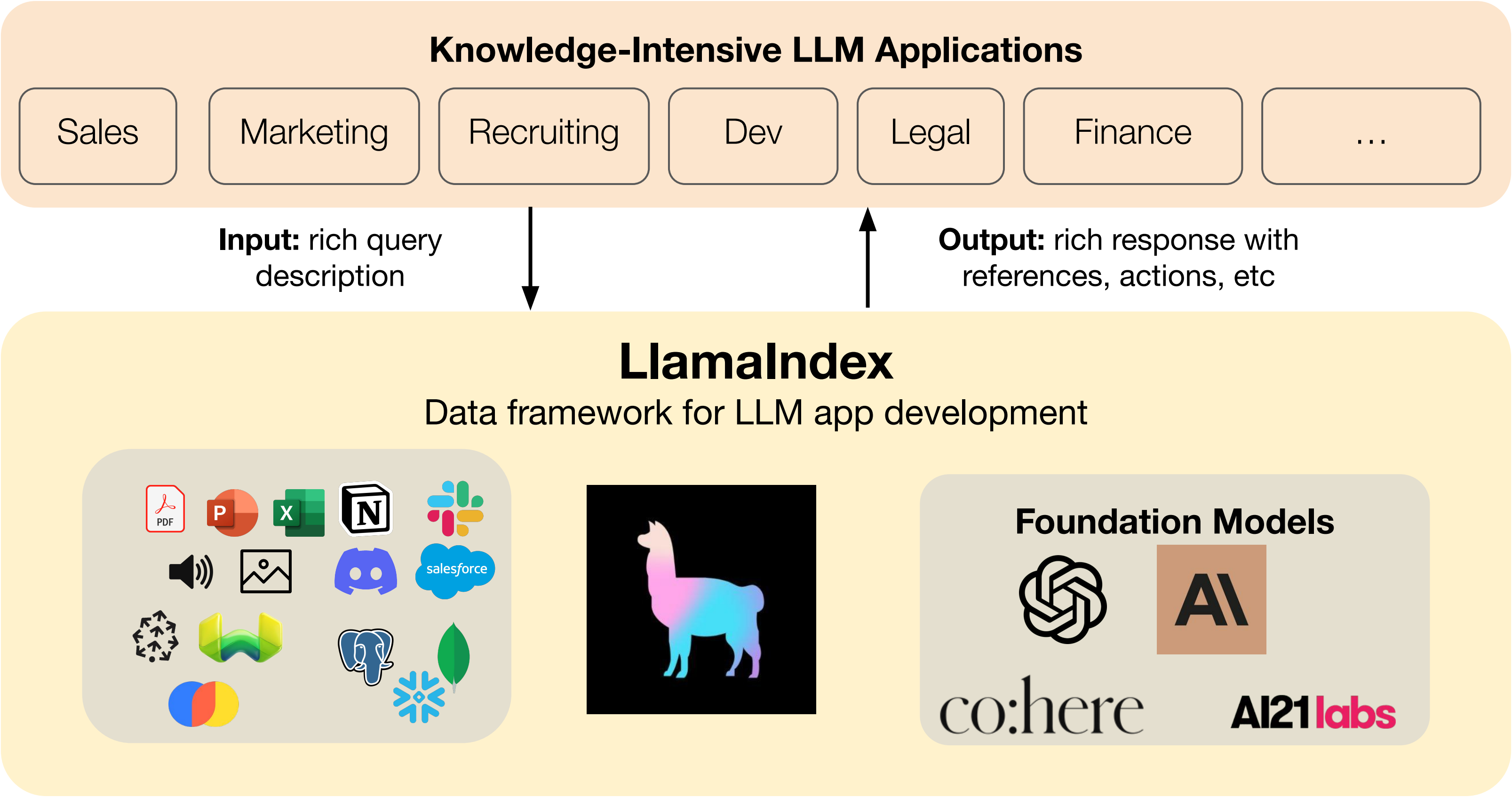
Retrieval and Query Interface

- Connect your existing data sources and data formats (API's, PDF's, docs, SQL, etc.)

- Store and index your data for different use cases. Integrate with different db's.

- Given an input prompt, retrieve relevant context and synthesize a knowledge-augmented output.

LlamaIndex introduction



Knowledge-Intensive LLM Applications

Sales | Marketing | Recruiting | Dev | Legal | Finance | …

Input: rich query description

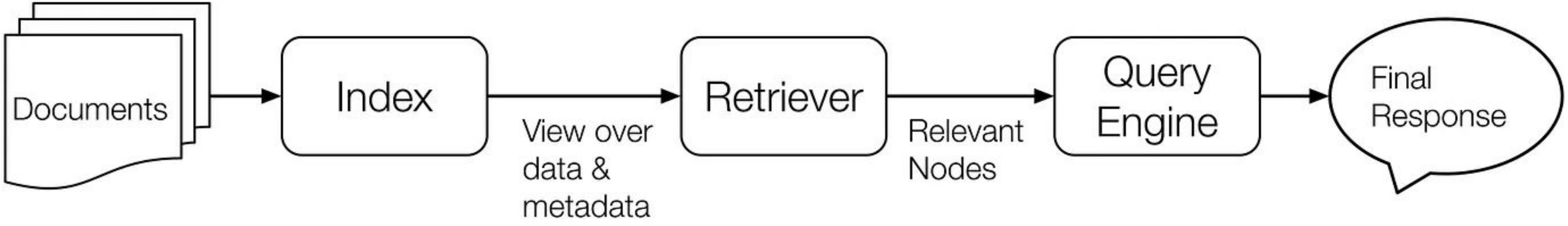
Output: rich response with references, actions, etc

LlamaIndex

Data framework for LLM app development

Foundation Models

Data Indices + Query Interface



Your **source documents** are stored in a data collection

In-memory, MongoDB

Our **data indices** help to provide a view of your raw data

Vectors, keyword lookups, summaries

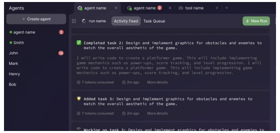
A **retriever** helps to retrieve relevant documents for your query

A **query engine** manages retrieval and synthesis given the query.

Perspectives

Agent Complexity

SuperAGI
Autonomous AI Agents



✅ Completed task 2: Design and implement graphics for obstacles and enemies to match the overall aesthetic of the game.

I will write code to create a platformer game. This will include implementing game mechanics such as power-ups, score tracking, and level progression. I will write code to create a platformer game. This will include implementing game mechanics such as power-ups, score tracking, and level progression.

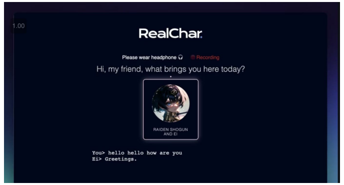
7 tokens consumed · 2m ago · More details

Added task 3: Design and implement graphics for obstacles and enemies to match the overall aesthetic of the game.

7 tokens consumed · 2m ago · More details

Working on task 3: Design and implement graphics for obstacles and enemies to

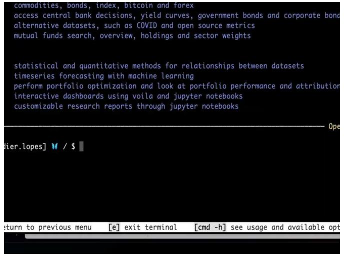
RealChar
Personalized characters



Hi, my friend, what brings you here today?

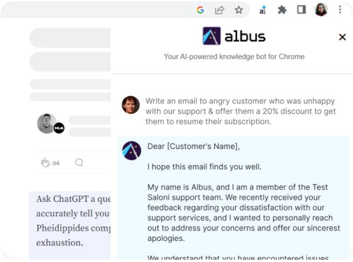
```
You> hello hello how are you
Ei> Greetings.
```

AskOBB (OpenBB)
Natural language
financial analysis



```
  commodities, bonds, index, bitcoin and forex
  access central bank decisions, yield curves, government bonds and corporate bond
  alternative datasets, such as COVID and open source metrics
  mutual funds search, overview, holdings and sector weights


  statistical and quantitative methods for relationships between datasets
  timeseries forecasting with machine learning
  perform portfolio optimization and look at portfolio performance and attribution
  interactive dashboards using voila and jupyter notebooks
  customizable research reports through jupyter notebooks

──────────────────────────────────────────────────────────────────────────── Ope

dier.lopes] / $ 

eturn to previous menu    [e] exit terminal    [cmd -h] see usage and available opt
```

Albus (Springworks)
AI-powered
knowledge bot



Your AI-powered knowledge bot for Chrome

Write an email to angry customer who was unhappy with our support & offer them a 20% discount to get them to resume their subscription.

Dear [Customer's Name],

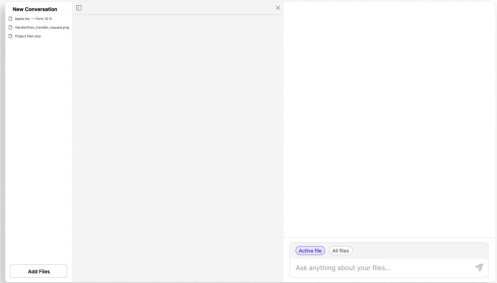
I hope this email finds you well.

My name is Albus, and I am a member of the Test Saloni support team. We recently received your feedback regarding your dissatisfaction with our support services, and I wanted to personally reach out to address your concerns and offer our sincerest apologies.

We understand that you have encountered issues

Ask ChatGPT a que
accurately tell you
Pheidippides com
exhaustion.

Instabase AI Hub
Chat with your
Documents



Project/Company Stage

Agenda

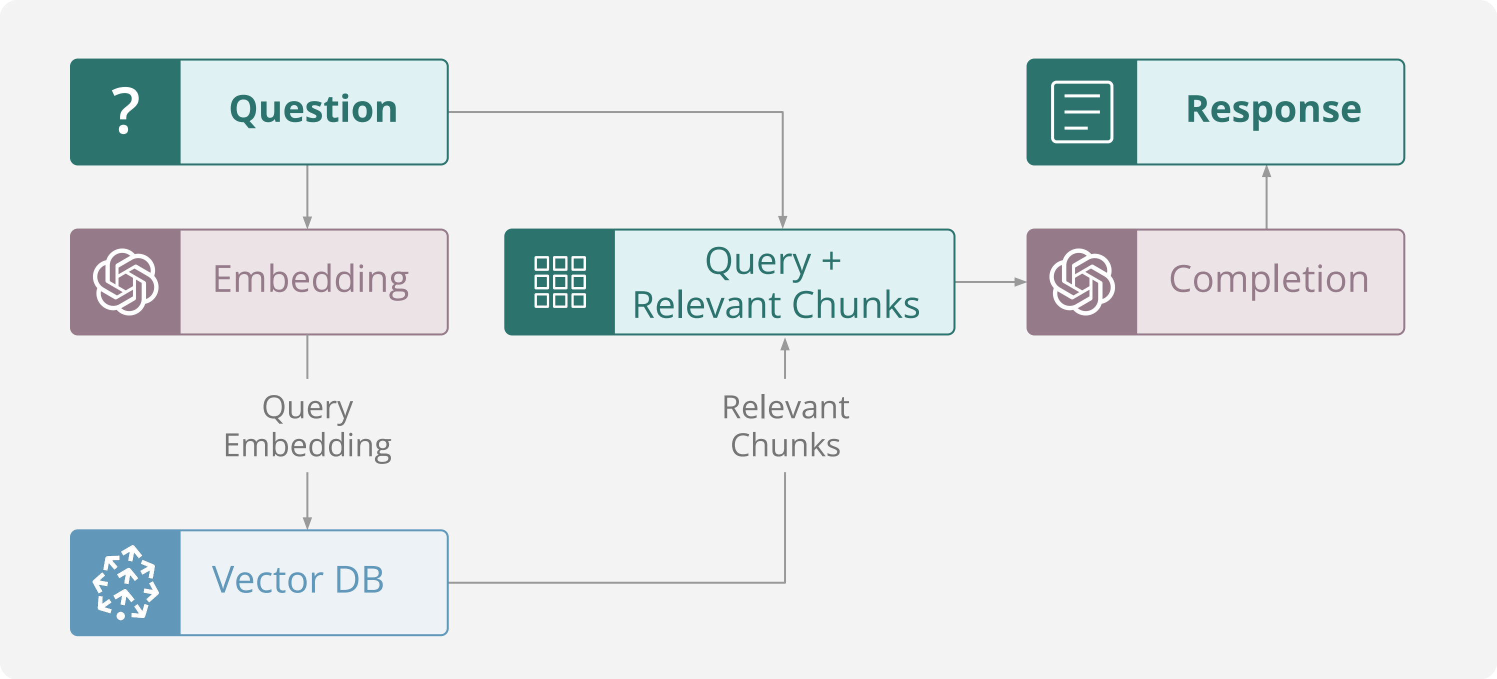
- Building LLM apps with LlamaIndex

- **Evaluating and tracking LLM apps with TruLens**
  - **RAG Question Answering**

- Build with LlamaIndex, evaluate & track with TruLens
  - Quickstart with LlamaIndex and Trulens
  - RAG QA with query planning
  - **[Optional]** RAG QA with data agents

RAGs: Generalize with LLMs + Memorize with Vector DBs

Example: Question Answering ChatBot



Question

Embedding

Query Embedding

Vector DB

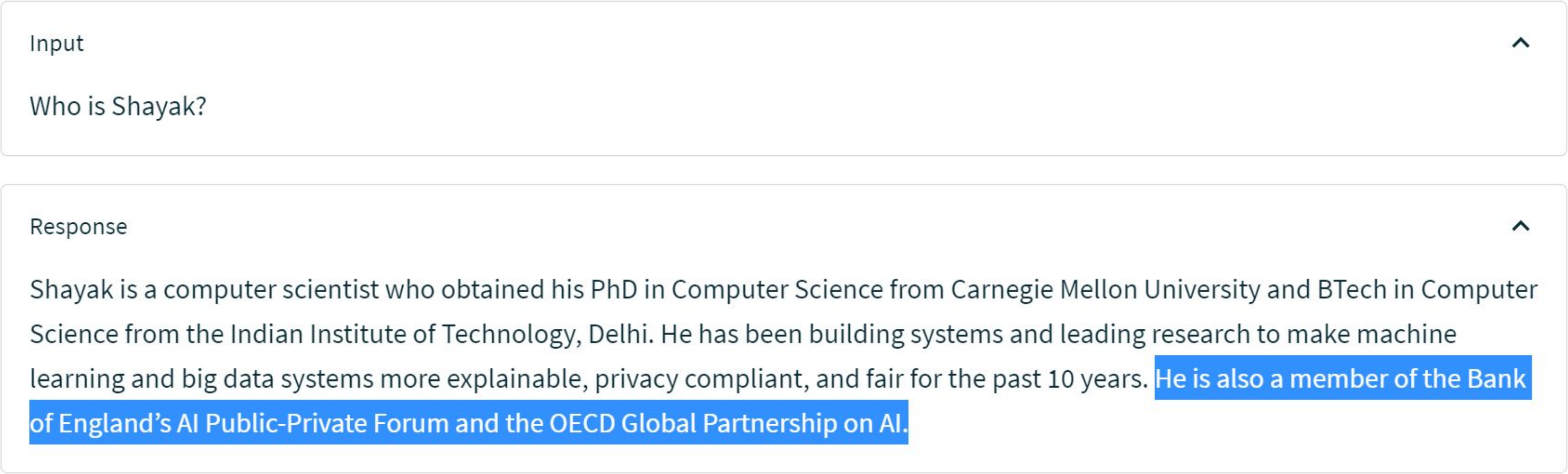
Relevant Chunks

Query + Relevant Chunks

Completion

Response

But RAGs can hallucinate too



Input

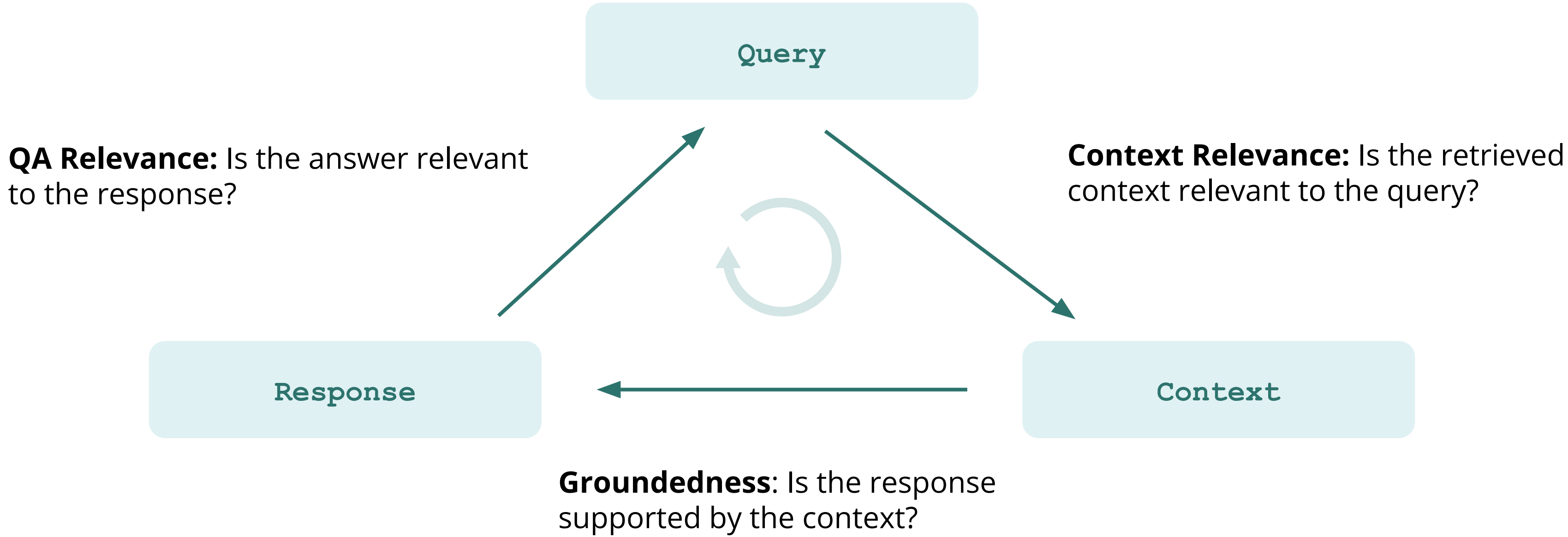
Who is Shayak?

Response

Shayak is a computer scientist who obtained his PhD in Computer Science from Carnegie Mellon University and BTech in Computer Science from the Indian Institute of Technology, Delhi. He has been building systems and leading research to make machine learning and big data systems more explainable, privacy compliant, and fair for the past 10 years. He is also a member of the Bank of England's AI Public-Private Forum and the OECD Global Partnership on AI.

Testing that RAGs are hallucination free

The RAG Triad



`Query`

QA Relevance: Is the answer relevant to the response?

Context Relevance: Is the retrieved context relevant to the query?

`Response`

`Context`

Groundedness: Is the response supported by the context?

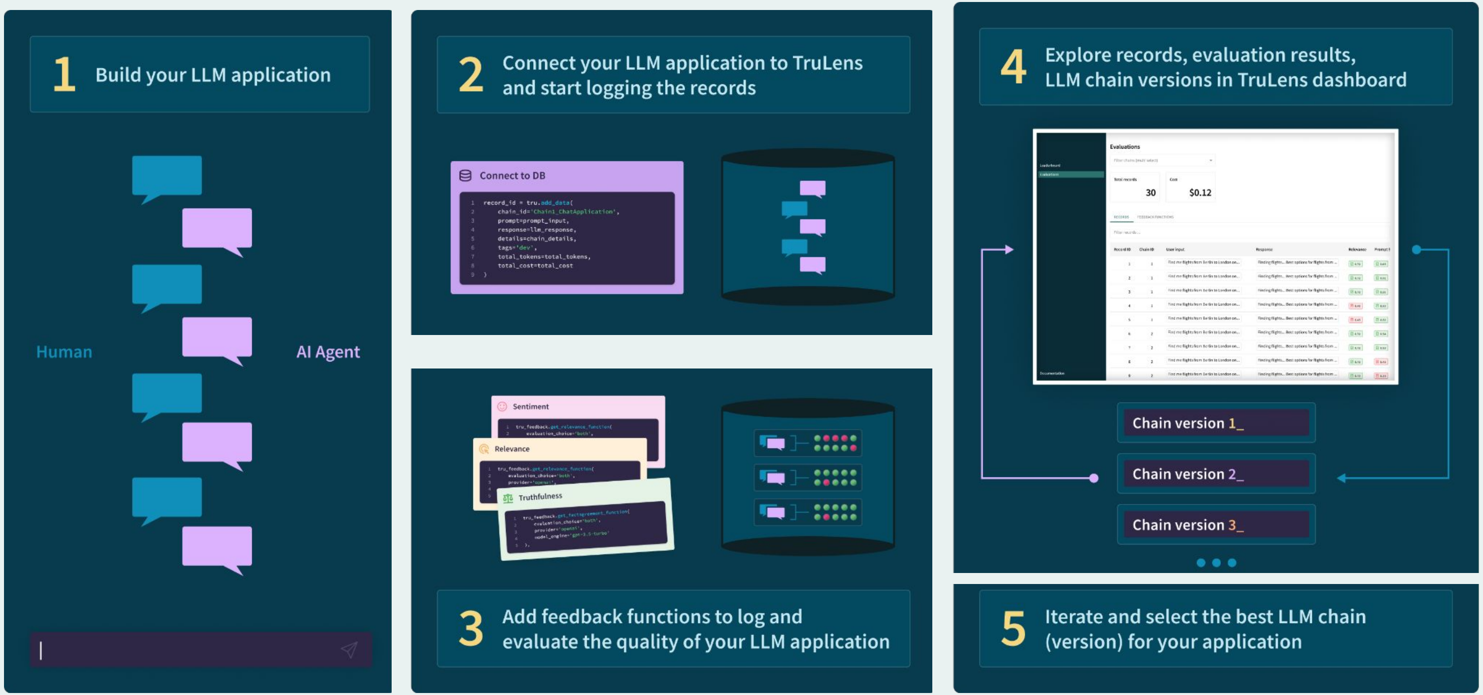
Getting this right requires experimentation

- Build an initial version
- Manually test & eyeball results
- Iterate with prompts, parameters, fine-tuning

Gap in dev tooling to evaluate & track LLM experiments

TruLens : Track and Evaluate LLM Experiments

github.com/truera/trulens



1 Build your LLM application

2 Connect your LLM application to TruLens and start logging the records

Connect to DB

```
record_id = tru.add_data(
    chain_id='Chain1_ChatApplication',
    prompt=prompt_input,
    response=llm_response,
    details=chain_details,
    tags='dev',
    total_tokens=total_tokens,
    total_cost=total_cost
)
```

Sentiment

```
tru_feedback.get_relevance_function(
    evaluation_choice='both',
```

Relevance

```
tru_feedback.get_relevance_function(
    evaluation_choice='both',
    provider='openai',
```

Truthfulness

```
tru_feedback.get_factagreement_function(
    evaluation_choice='both',
    provider='openai',
    model_engine='gpt-3.5-turbo'
),
```

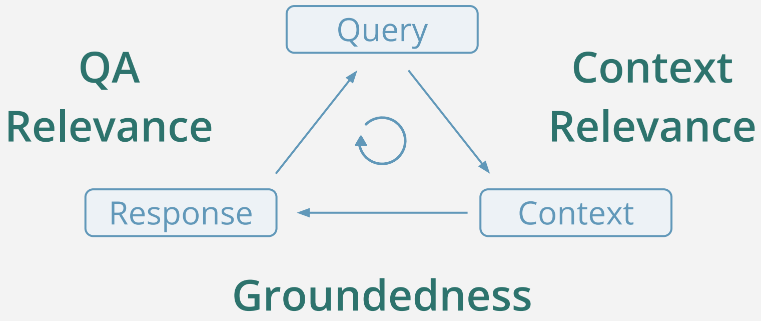
3 Add feedback functions to log and evaluate the quality of your LLM application

4 Explore records, evaluation results, LLM chain versions in TruLens dashboard

5 Iterate and select the best LLM chain (version) for your application

What is a feedback function?

A feedback function provides a score after reviewing an LLM app's inputs, outputs, intermediate results, and metadata.



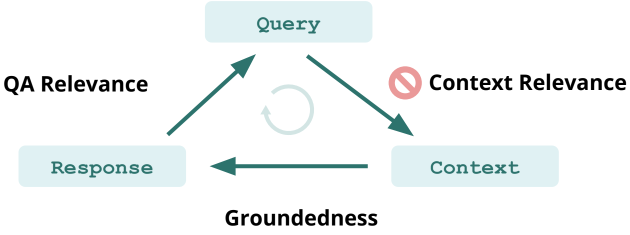
Additional Examples

- Prompt sentiment
- Language mismatch
- Transcript length
- Response verbosity
- Fairness substitution
- Toxicity
- ...

Easy to Implement

```
Feedback(hugs.language_match).
 on_input_output()
```

What happened here?



Query

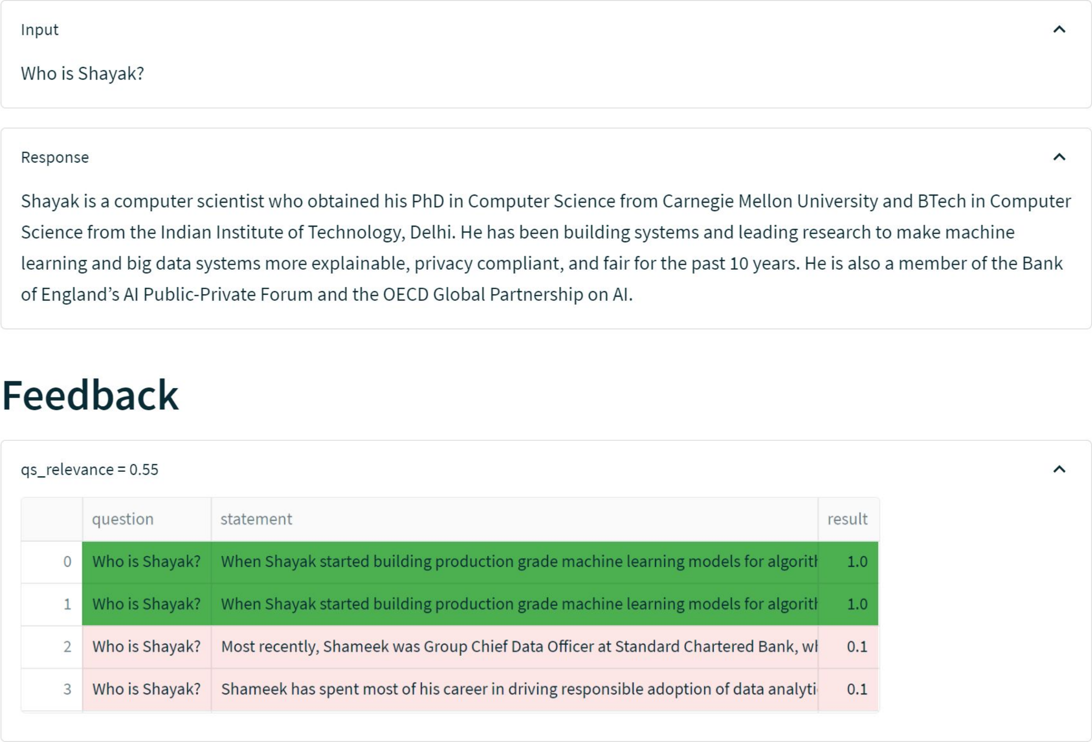
QA Relevance

🚫 Context Relevance

Response

Context

Groundedness



Input

Who is Shayak?

Response

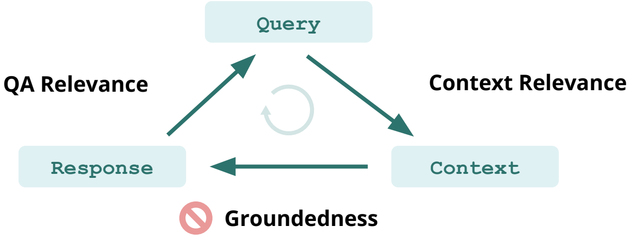
Shayak is a computer scientist who obtained his PhD in Computer Science from Carnegie Mellon University and BTech in Computer Science from the Indian Institute of Technology, Delhi. He has been building systems and leading research to make machine learning and big data systems more explainable, privacy compliant, and fair for the past 10 years. He is also a member of the Bank of England's AI Public-Private Forum and the OECD Global Partnership on AI.

Feedback

qs_relevance = 0.55

	question	statement	result
0	Who is Shayak?	When Shayak started building production grade machine learning models for algorit	1.0
1	Who is Shayak?	When Shayak started building production grade machine learning models for algorit	1.0
2	Who is Shayak?	Most recently, Shameek was Group Chief Data Officer at Standard Chartered Bank, w	0.1
3	Who is Shayak?	Shameek has spent most of his career in driving responsible adoption of data analyti	0.1

Sometimes hallucinations can be correct



Query

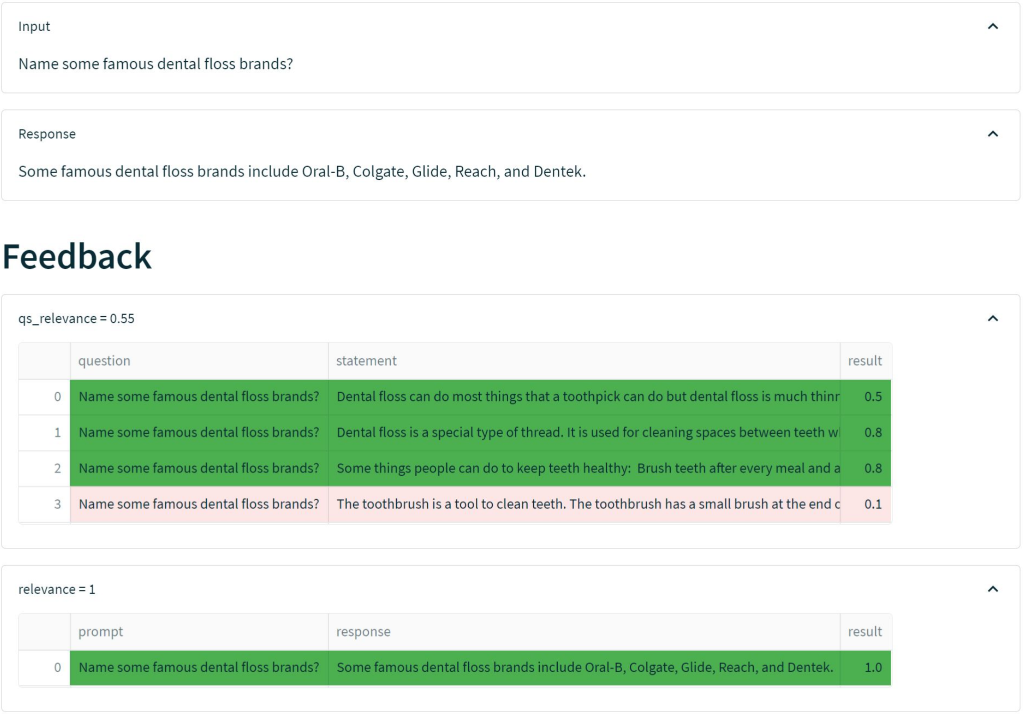
QA Relevance

Context Relevance

Response

Context

🚫 **Groundedness**



Input

Name some famous dental floss brands?

Response

Some famous dental floss brands include Oral-B, Colgate, Glide, Reach, and Dentek.

Feedback

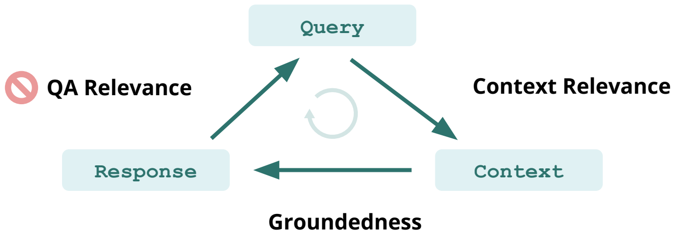
qs_relevance = 0.55

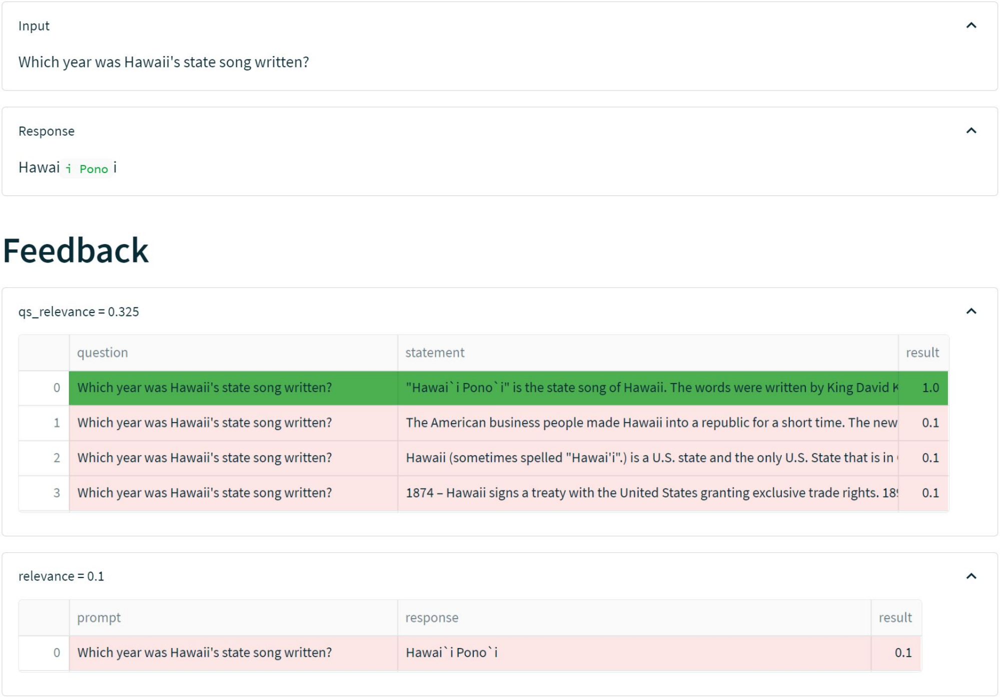
	question	statement	result
0	Name some famous dental floss brands?	Dental floss can do most things that a toothpick can do but dental floss is much thinr	0.5
1	Name some famous dental floss brands?	Dental floss is a special type of thread. It is used for cleaning spaces between teeth w	0.8
2	Name some famous dental floss brands?	Some things people can do to keep teeth healthy: Brush teeth after every meal and a	0.8
3	Name some famous dental floss brands?	The toothbrush is a tool to clean teeth. The toothbrush has a small brush at the end c	0.1

relevance = 1

	prompt	response	result
0	Name some famous dental floss brands?	Some famous dental floss brands include Oral-B, Colgate, Glide, Reach, and Dentek.	1.0

Or we might be answering the wrong question





Input

Which year was Hawaii's state song written?

Response

Hawai `i Pono` i

Feedback

qs_relevance = 0.325

	question	statement	result
0	Which year was Hawaii's state song written?	"Hawai`i Pono`i" is the state song of Hawaii. The words were written by King David	1.0
1	Which year was Hawaii's state song written?	The American business people made Hawaii into a republic for a short time. The new	0.1
2	Which year was Hawaii's state song written?	Hawaii (sometimes spelled "Hawai'i".) is a U.S. state and the only U.S. State that is in	0.1
3	Which year was Hawaii's state song written?	1874 – Hawaii signs a treaty with the United States granting exclusive trade rights. 18	0.1

relevance = 0.1

	prompt	response	result
0	Which year was Hawaii's state song written?	Hawai`i Pono`i	0.1

Evaluation Pitfalls & Best Practices

Overfitting to single examples

Prompt engineering can hyperfocus on getting single examples right, at the expense of the larger corpus

Not testing performance periodically

Base models keep evolving, even if your app does not change

Not setting up comprehensive evaluations

Because of the lack of ground truth, developers often just rely on eyeballing a small set of results

Also need to evaluate intermediate steps and not just the LLM results

Agenda

- Building LLM apps with LlamaIndex

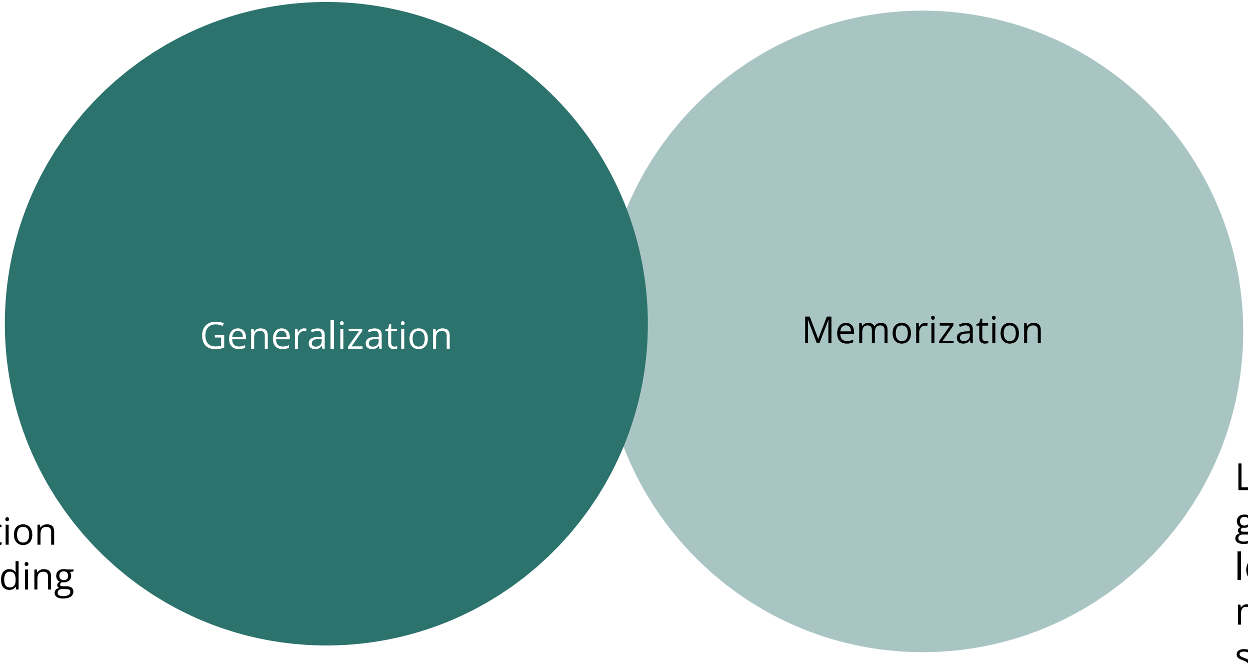
- Evaluating and tracking LLM apps with TruLens
    - RAG Question Answering

- **Build with LlamaIndex, evaluate & track with TruLens**
    - Quickstart with LlamaIndex and Trulens
    - **RAG QA with query planning**
    - **[Optional]** RAG QA with data agents

Demo 1: Quickstart

Confidential.

Focus LLMs on ‘general’ tasks



- ✓ Summarization
- ✓ Text Embedding
- ✓ Inference
- ✓ <u>Planning</u>

Let LLMs focus on general tasks, leaving memorization to something else

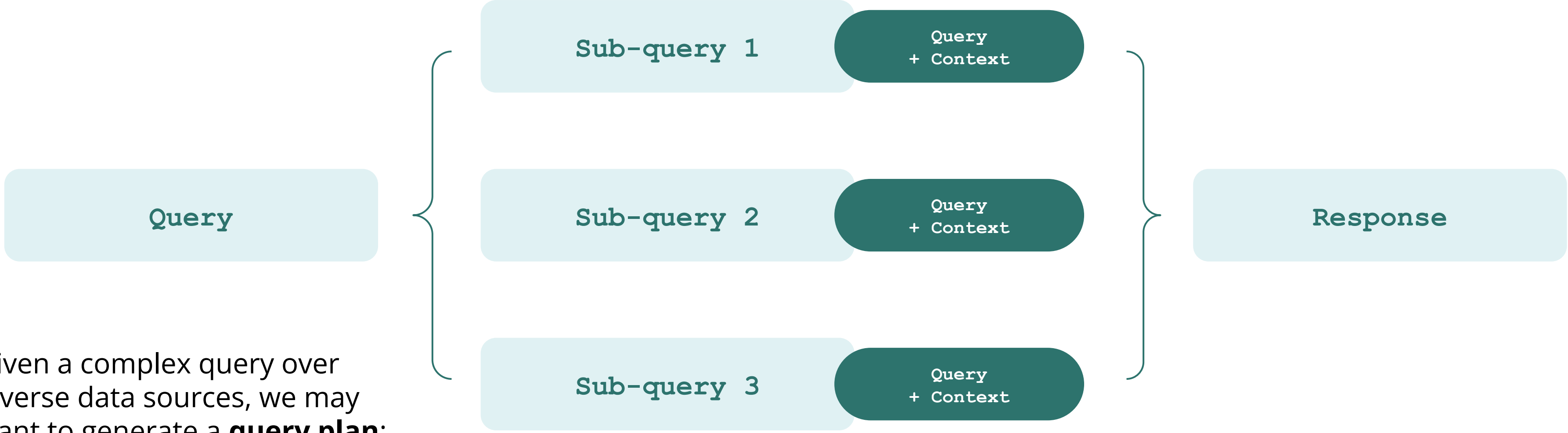
Improving RAGs with query planning

- Naive RAG: retrieval step (top-k), synthesis (LLM)
- Doesn’t always work well for more complex queries - bad retrieval
- Example: “Compare and contrast Uber and Lyft revenues in 2020-2021”
- How do we use LLM to better **reason** over your knowledge sources?

Use LLM to generate a **query plan** over your data

Improving RAGs with query planning

Agents for Question-Answering



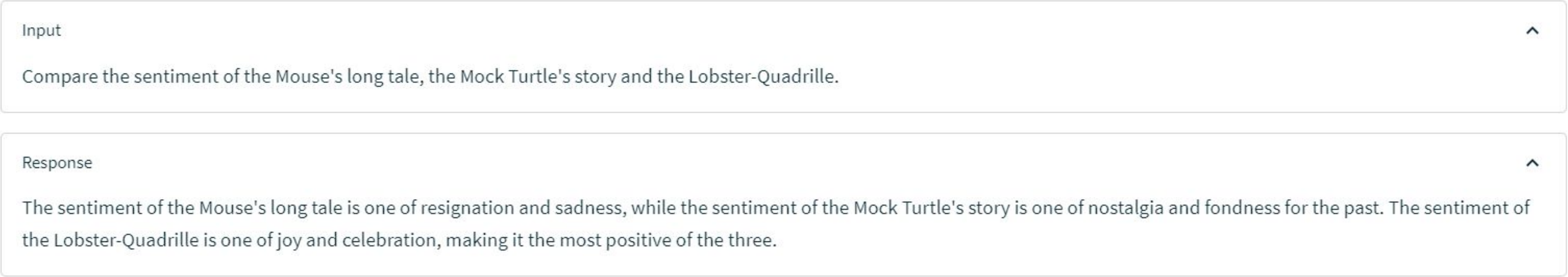
Given a complex query over diverse data sources, we may want to generate a **query plan**:

- Decompose query into subqueries
- Execute each subquery against a subset of data.
- Combine answers.

Query Planning

Alice in wonderland

Allows RAGs to answer more complex questions, where direct retrieval could fail



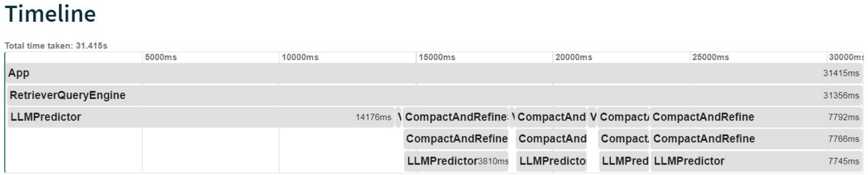
Input

Compare the sentiment of the Mouse's long tale, the Mock Turtle's story and the Lobster-Quadrille.

Response

The sentiment of the Mouse's long tale is one of resignation and sadness, while the sentiment of the Mock Turtle's story is one of nostalgia and fondness for the past. The sentiment of the Lobster-Quadrille is one of joy and celebration, making it the most positive of the three.

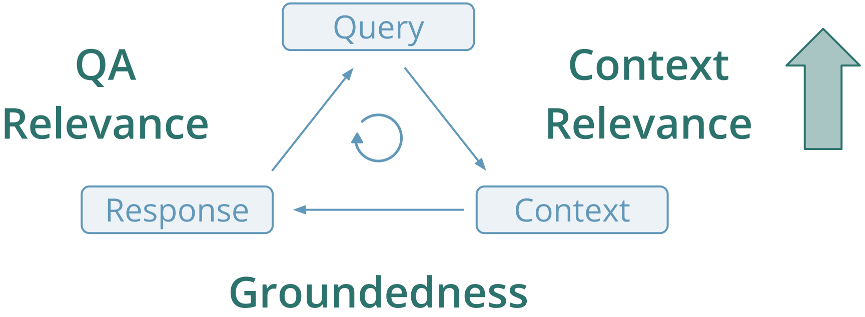
But can take a lot longer:



Timeline

Total time taken: 31.415s

Improving quality by improving the context



Query

QA Relevance

Context Relevance

Response

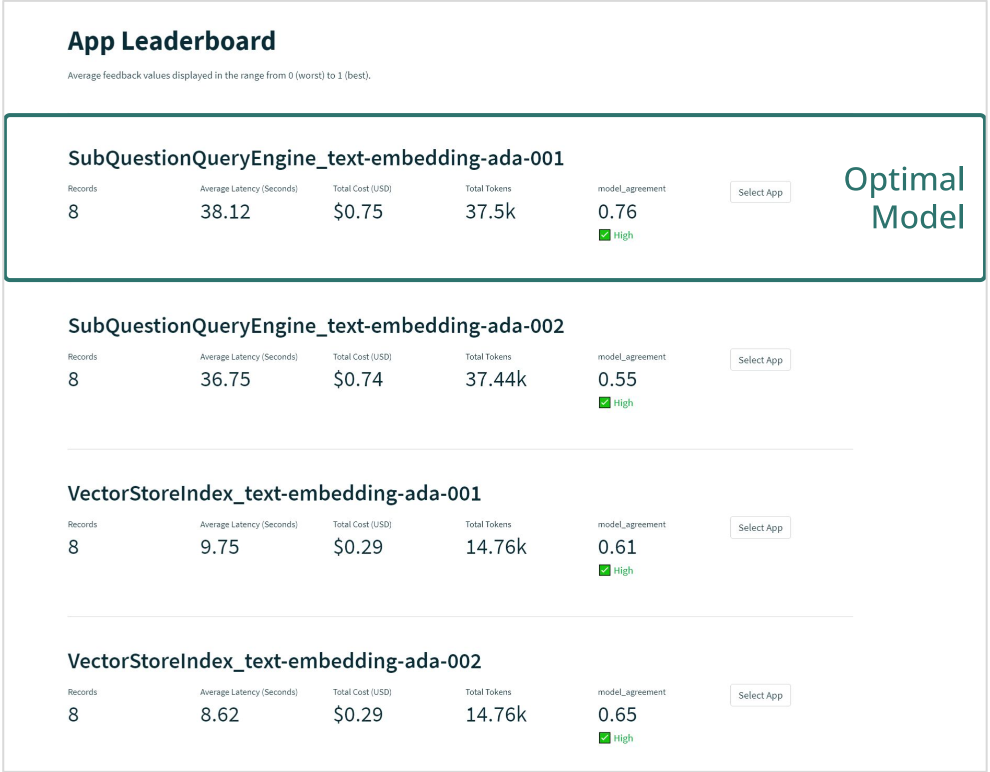
Context

Groundedness

More complete context, let the LLM decide how much context it needs, and why

Experimenting with query planning

- Decomposing a complex query into subqueries improves quality, though at the cost of higher token cost and latency
- Parameter changes (such as embedding upgrade) can have significant impact on quality
- Iterating through LLM parameters + automatic tracking and scoring allows for optimal selection



App Leaderboard

Average feedback values displayed in the range from 0 (worst) to 1 (best).

SubQuestionQueryEngine_text-embedding-ada-001

Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	model_agreement
8	38.12	$0.75	37.5k	0.76 ✓ High

Optimal Model

SubQuestionQueryEngine_text-embedding-ada-002

Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	model_agreement
8	36.75	$0.74	37.44k	0.55 ✓ High

VectorStoreIndex_text-embedding-ada-001

Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	model_agreement
8	9.75	$0.29	14.76k	0.61 ✓ High

VectorStoreIndex_text-embedding-ada-002

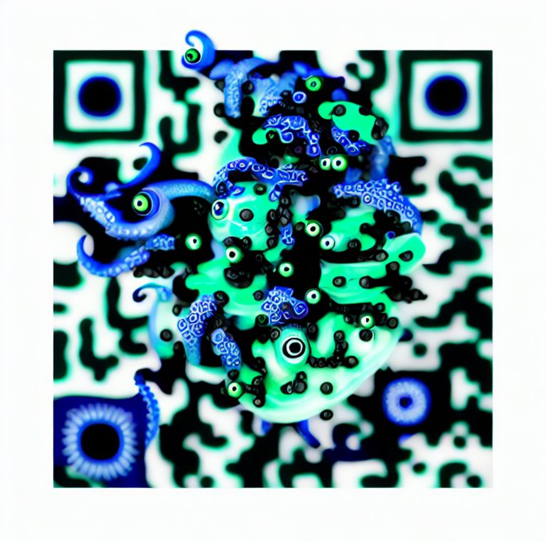
Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	model_agreement
8	8.62	$0.29	14.76k	0.65 ✓ High

Notebook example:
https://tinyurl.com/query-planning-trulens

Demo 2

Confidential.

Check us out



github.com/truera/trulens

github.com/jerryjliu/llama_index

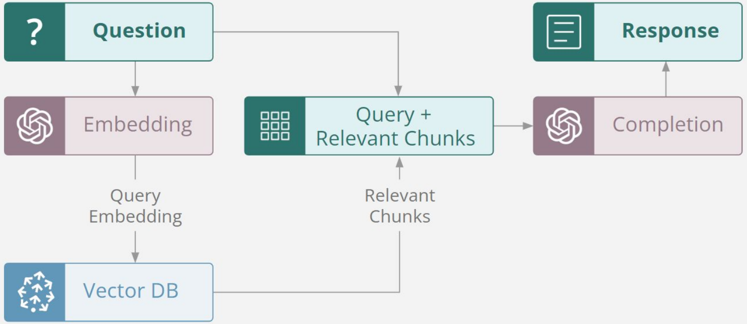
Give us a star! Request a feature and contribute, too!

LLMs are trained to generalize

They sometimes happen to memorize as a side-effect => they sometimes happen to not hallucinate as a side-effect

But there is hope :)

Toward Hallucination Free RAGs



Question

Embedding

Query Embedding

Vector DB

Query + Relevant Chunks

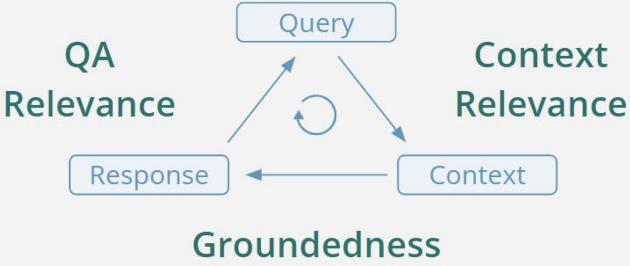
Relevant Chunks

Completion

Response

Augment LLMs with Retrieval

e.g. with Llama-Index



Query

Response

Context

QA Relevance

Context Relevance

Groundedness

Evaluate RAGs Carefully

e.g. with TruLens

trulens

Bonus material

Confidential.

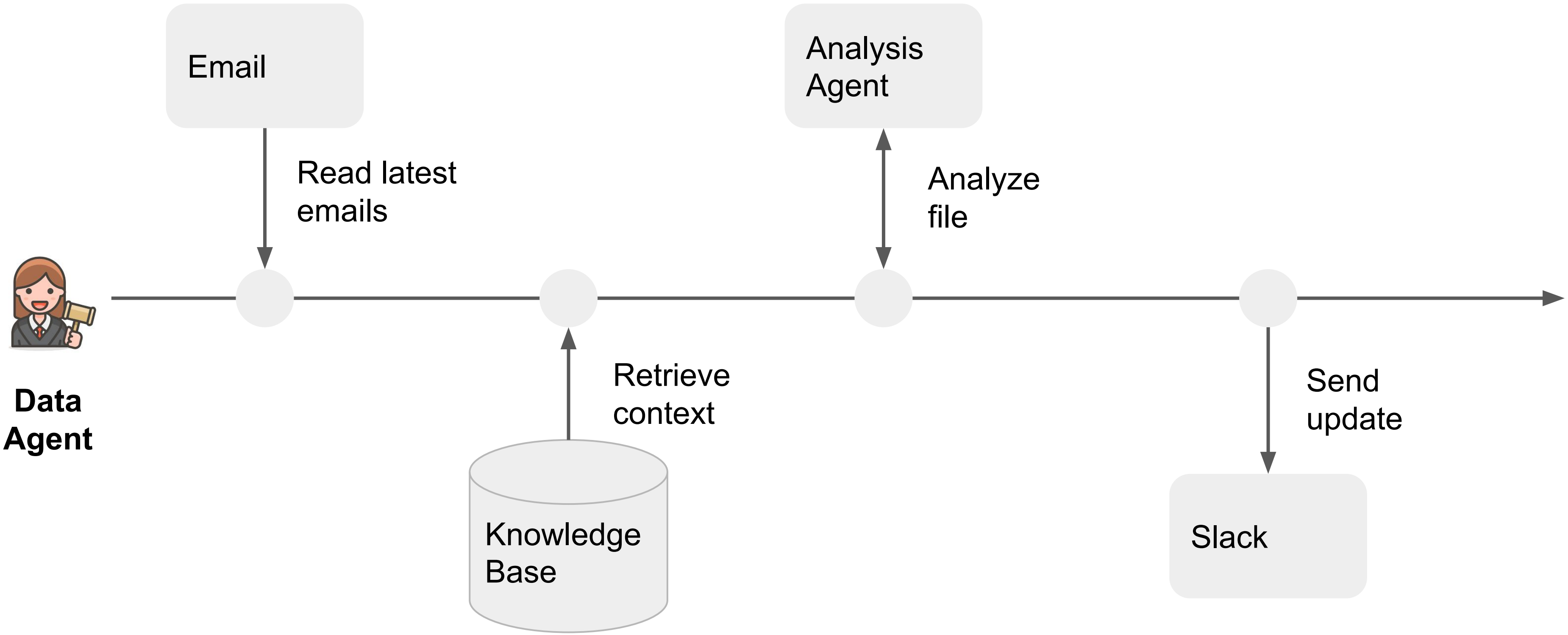
Agenda

- Building LLM apps with LlamaIndex

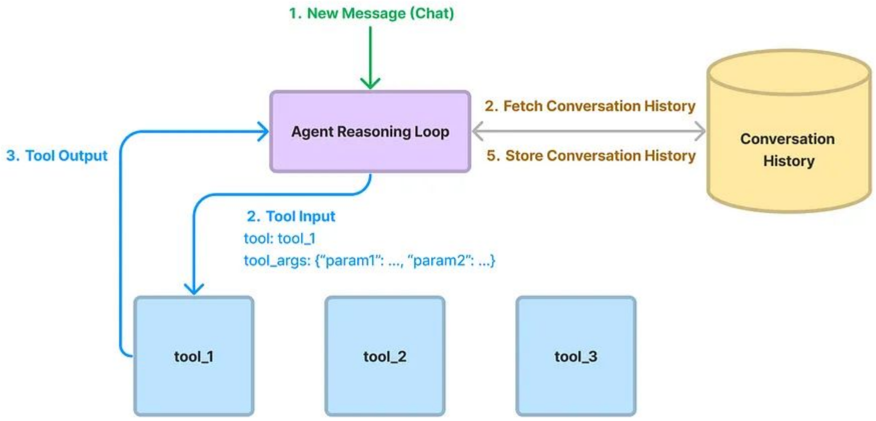
- Evaluating and tracking LLM apps with TruLens
  - RAG Question Answering

- **Build with LlamaIndex, evaluate & track with TruLens**
  - Quickstart with LlamaIndex and Trulens
  - RAG QA with query planning
  - **RAG QA with data agents**

Data Agents - LLM-powered knowledge workers



Data Agents - Core Components



Agent Reasoning Loop

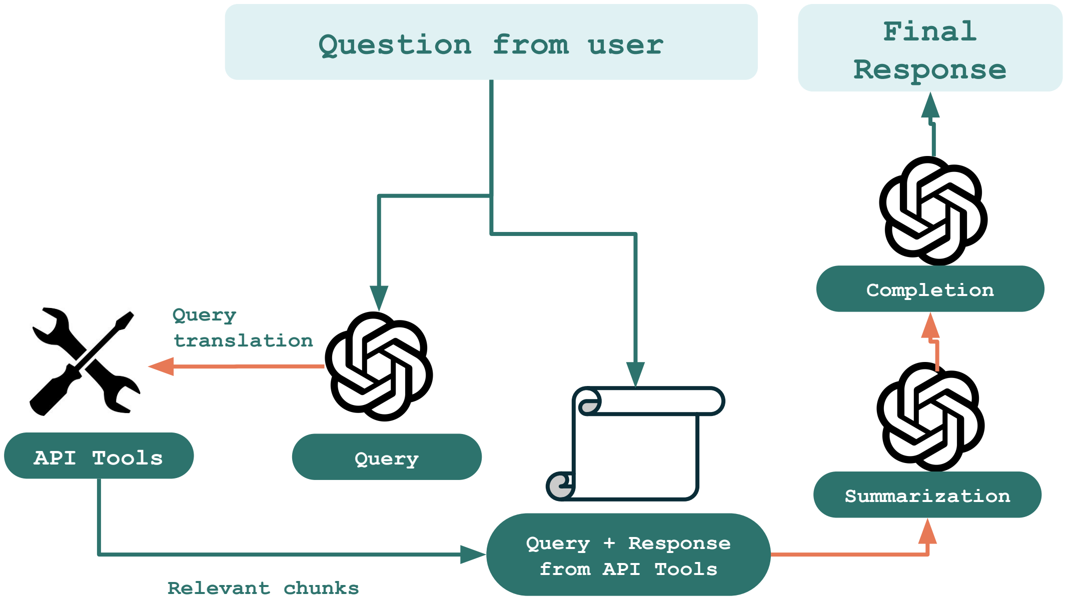
- OpenAI Agent (only OAI)
- ReAct Agent (any LLM)

Tools via LlamaHub

- Code interpreter
- Slack
- Notion
- Zapier
- … (15+ tools, ~100 loaders)

Data agents for real-time retrieval

Example: Question Answering with ChatBot interface



Knowledge source

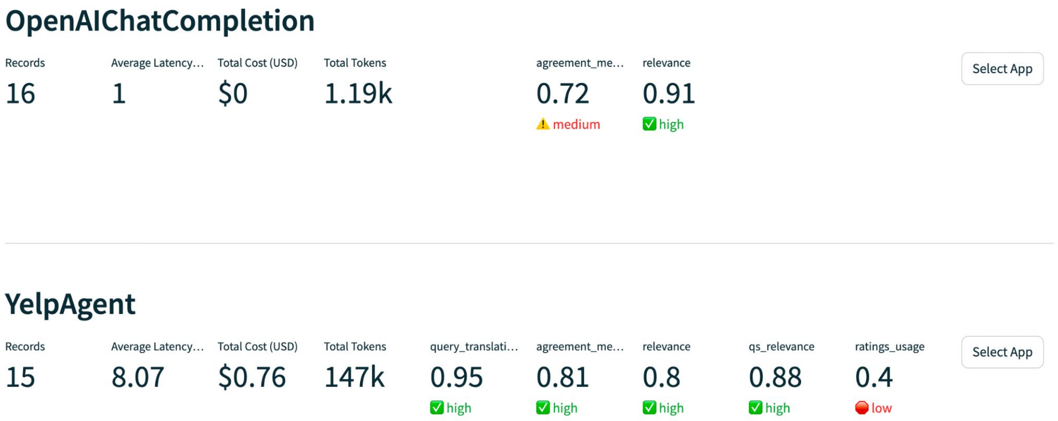
- Yelp API

Evals

- Query Translation Scores
- Ratings Usage
- Answer Relevance
- Context Relevance
- Ground Truth Agreement

Experimenting with data agents

- Data agents give more certainty to eval by testing throughout the application
- Thorough testing of LLM apps ensures groundedness



OpenAIChatCompletion

Records	Average Latency...	Total Cost (USD)	Total Tokens	agreement_me...	relevance
16	1	$0	1.19k	0.72 ⚠ medium	0.91 ✅ high

Select App

YelpAgent

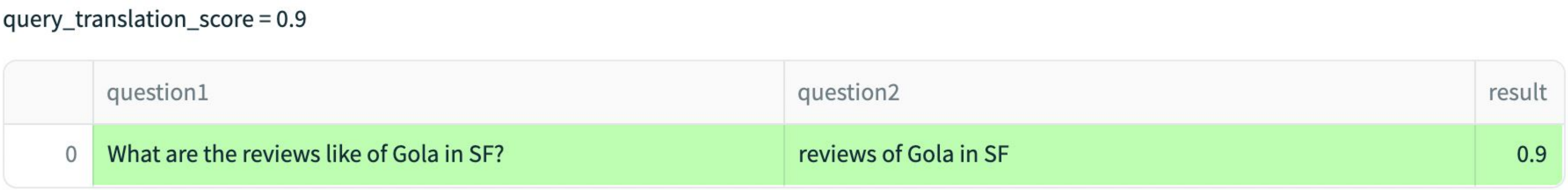
Records	Average Latency...	Total Cost (USD)	Total Tokens	query_translati...	agreement_me...	relevance	qs_relevance	ratings_usage
15	8.07	$0.76	147k	0.95 ✅ high	0.81 ✅ high	0.8 ✅ high	0.88 ✅ high	0.4 🔴 low

Select App

Notebook example:
https://tinyurl.com/data-agents

Verify retrieval

- Custom evals can verify that the correct retrieval is occurring



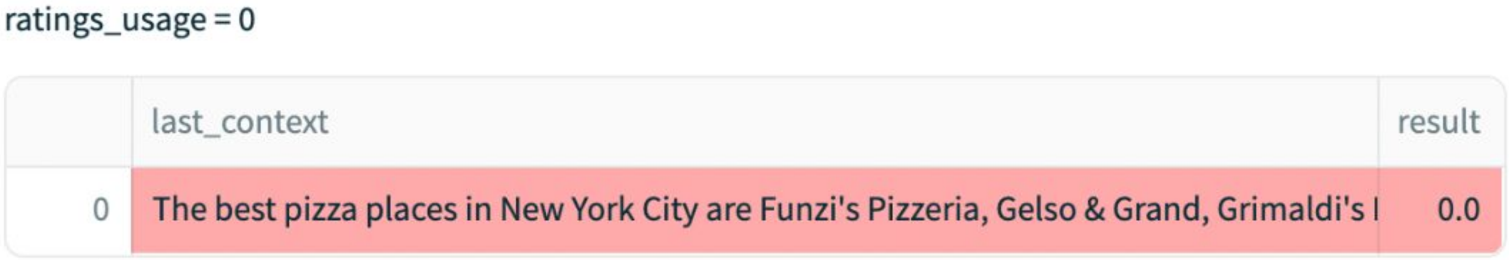
query_translation_score = 0.9

	question1	question2	result
0	What are the reviews like of Gola in SF?	reviews of Gola in SF	0.9

Notebook example:
https://tinyurl.com/data-agents

Identify failure modes

- Custom evals can also identify app-specific requirements



ratings_usage = 0

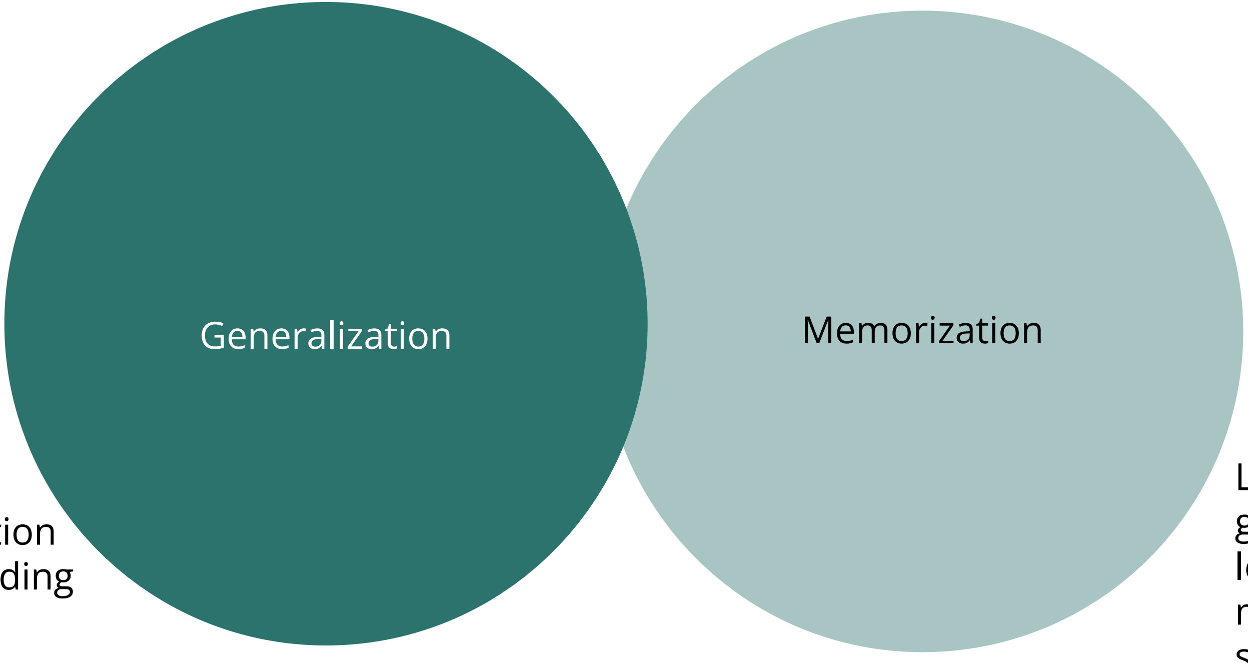
	last_context	result
0	The best pizza places in New York City are Funzi's Pizzeria, Gelso & Grand, Grimaldi's	0.0

Notebook example:
https://tinyurl.com/data-agents

Demo 3

Confidential.

Focus LLMs on ‘general’ tasks



- ✓ Summarization
- ✓ Text Embedding
- ✓ Inference
- ✓ Planning

Let LLMs focus on general tasks, leaving memorization to something else