

truera

Evaluating LLM Agents

Anupam Datta



ChatGPT plugins

We've implemented initial support for plugins in ChatGPT. Plugins are language models that help ChatGPT access computations, or



Expedia

Bring your trip plans to life—get there, stay there, find things to see and do.



FiscalNote

Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.



Instacart

Order from your favorite local grocery stores.



KAYAK

Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.



Klarna Shopping

Search and compare prices from thousands of online shops.



Milo Family AI

Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?



OpenTable

Provides restaurant recommendations, with a direct link to book.



Shop

Search for millions of products from the world's greatest brands.



Speak

Learn how to say anything in another language with Speak, your AI-powered language tutor.



Wolfram

Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language.



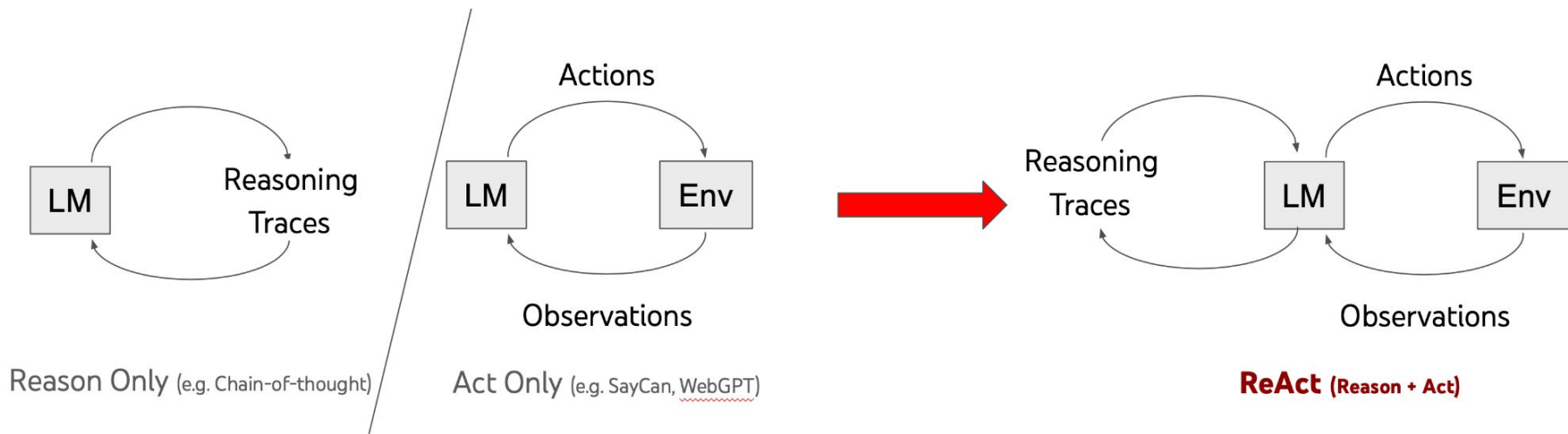
Zapier

Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

[ChatGPT plugins](#)

ReAct: Reasoning + Acting with LLMs

truera



Source: <https://react-lm.github.io/>

ReAct: HotpotQA example

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Standard

Answer: iPod



Reason only

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch



Act only

Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced...
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ...
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software ...
Act 4: Finish[yes]



ReAct

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search Front Row (software) .

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software ...

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

Act 4: Finish[keyboard function keys]



ReAct: LangChain Agent Example (1)

First, let's install and import the necessary libraries:

```
%%capture
# update or install the necessary libraries
!pip install --upgrade openai
!pip install --upgrade langchain
!pip install --upgrade python-dotenv
!pip install google-search-results

# import libraries
import openai
import os
from langchain.llms import OpenAI
from langchain.agents import load_tools
from langchain.agents import initialize_agent
from dotenv import load_dotenv
load_dotenv()

# load API keys; you will need to obtain these if you haven't yet
os.environ["OPENAI_API_KEY"] = os.getenv("OPENAI_API_KEY")
os.environ["SERPER_API_KEY"] = os.getenv("SERPER_API_KEY")
```

<https://www.promptingguide.ai/techniques/react>

ReAct: LangChain Agent Example (2)

Now we can configure the LLM, the tools we will use, and the agent that allows us to leverage the ReAct framework together with the LLM and tools. Note that we are using a search API for searching external information and LLM as a math tool.

```
llm = OpenAI(model_name="text-davinci-003" ,temperature=0)
tools = load_tools(["google-serper", "llm-math"], llm=llm)
agent = initialize_agent(tools, llm, agent="zero-shot-react-description", verbose=True)
```

Once that's configured, we can now run the agent with the desired query/prompt. Notice that here we are not expected to provide few-shot exemplars as explained in the paper.

```
agent.run("Who is Olivia Wilde's boyfriend? What is his current age raised to the 0.23 power?")
```

<https://www.promptingguide.ai/techniques/react>

ReAct: LangChain Agent Example (3)

The chain execution looks as follows:

```
> Entering new AgentExecutor chain...
I need to find out who Olivia Wilde's boyfriend is and then calculate his age raised to the 0.23 power.
Action: Search
Action Input: "Olivia Wilde boyfriend"
Observation: Olivia Wilde started dating Harry Styles after ending her years-long engagement to Jason Sudeikis - see their relationship timeline.
Thought: I need to find out Harry Styles' age.
Action: Search
Action Input: "Harry Styles age"
Observation: 29 years
Thought: I need to calculate 29 raised to the 0.23 power.
Action: Calculator
Action Input: 29^0.23
Observation: Answer: 2.169459462491557

Thought: I now know the final answer.
Final Answer: Harry Styles, Olivia Wilde's boyfriend, is 29 years old and his age raised to the 0.23 power is 2.169459462491557.

> Finished chain.
```

The output we get is as follows:

```
"Harry Styles, Olivia Wilde's boyfriend, is 29 years old and his age raised to the 0.23 power is 2.169459462491557."
```

<https://www.promptingguide.ai/techniques/react>

Hackathon Overview

Our AI hackathon brought together a diverse group of participants, who collaborated to develop a variety of impressive projects based on:

[Auto-GPT](#) [MetaGPT](#) [Weaviate](#)

5093 **518** **24**

Participants Teams AI Applications

Winners and Finalists

- We are in the process of selecting the finalist teams.
- Your voice matters! Vote on your favorite projects in the section below.
- Join us for the winner announcement stream, which will be streamed live on [Twitch](#).

ChadGPT

Building the most Chad and powerful Agent out there 🤖👉👉👉
ChadGPT

[Auto-GPT](#) [OpenAI](#) [GPT-3.5](#) [GPT-4](#)

evo-ninja

evo.ninja is an agent that can flow between multiple agent personas to solve any task.
Evo-ninja

[OpenAI](#) [GPT-3.5](#) [GPT-4](#)

MLJAR

We develop agent that relies on micro-agents and functions calling.
MLJAR

[Weaviate](#) [Auto-GPT](#) [OpenAI](#)

IdeationAgent

We're dedicated to curating bespoke startup ideas just for you. By harnessing the information from yo...
AntlerAI

[Auto-GPT](#) [ChatGPT](#)

Ghostcoder

An agent with the main focus to generate and edit code.
Ghostcoder

[GPT-4](#) [Auto-GPT](#)

gpt-engineer

We will push the powerful GPT-engineer coding agent to its limits in the code generation challenge.
gpt-engineer

[GPT-4](#)

Limitations

- Wrong tool selection
- Infinite loops
- Hallucinations

Range of AI agents are possible

General Data Agents

- Access to more than one tool
- Can accomplish a wider range of tasks



Specialized Data Agents

- Similar to retrieval from vector store
- But with access to real-time information

Agents that can take action in real world

- Book plane tickets
- Scheduling appointment
- Order doordash
- ...

Range of AI agents are possible

General Data Agents

- Access to more than one tool
- Can accomplish a wider range of tasks



Specialized Data Agents

- Similar to retrieval
- access to real-time information

Agents that can take action in real world

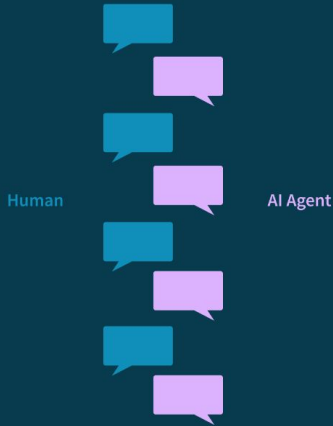
- Book plane tickets
- Create calendar invites
- ...

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

```

1 record_id = truera.record_id
2 chain_id = truera.chain_id
3 prompt_prompt_id, response_llm_response,
4 model_chain_id, tags,
5 total_tokens_total_tokens,
6 trace_context,
7 }

```

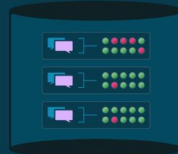


3 Add feedback functions to log and evaluate the quality of your LLM application

```

Sentiment
Relevance
Truthfulness

```



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

Record ID	Chain ID	User Input	Response	Accuracy	Cost
30				\$0.12	

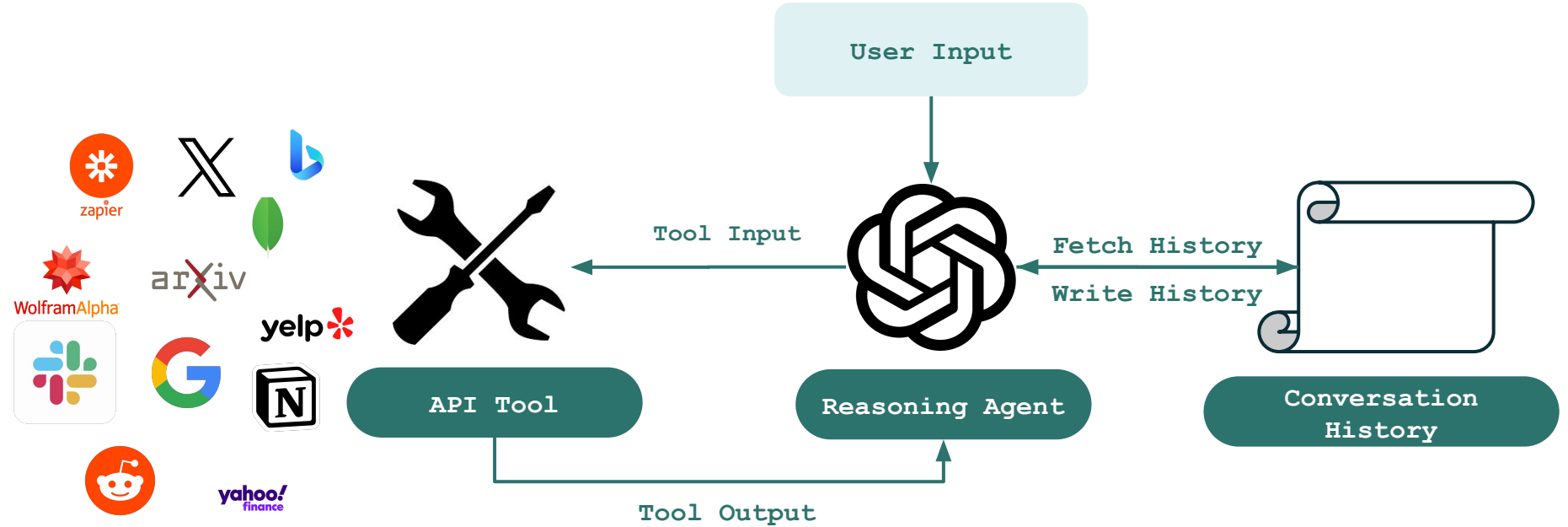
Chain version 1_

Chain version 2_

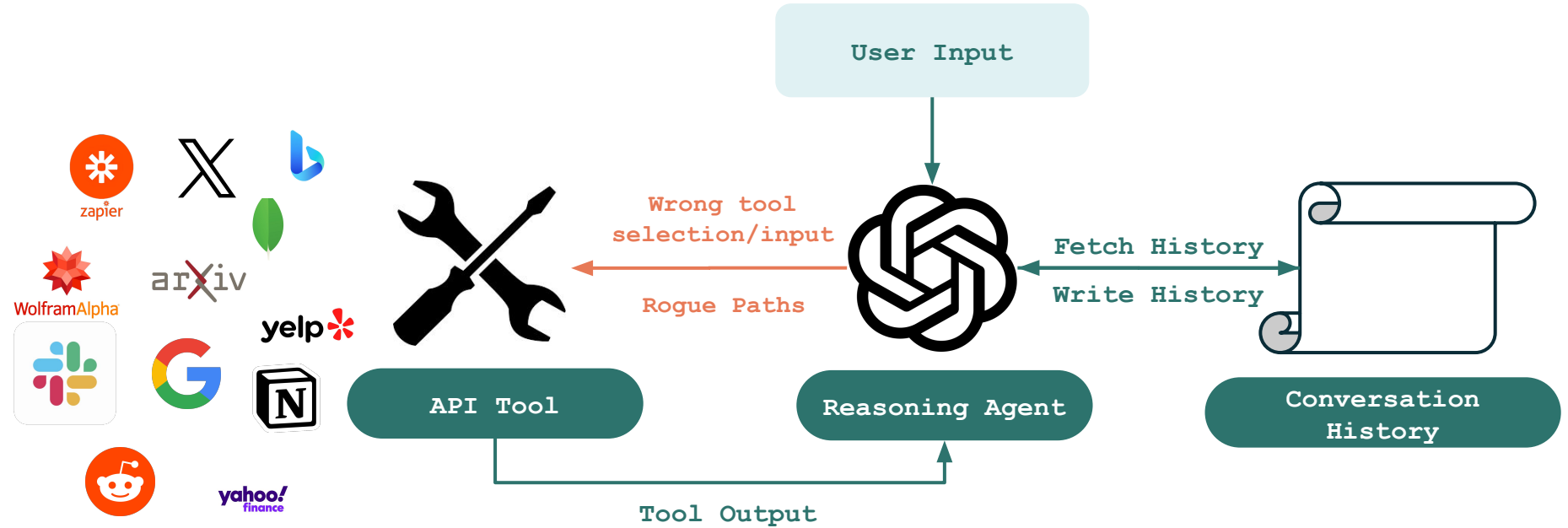
Chain version 3_

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

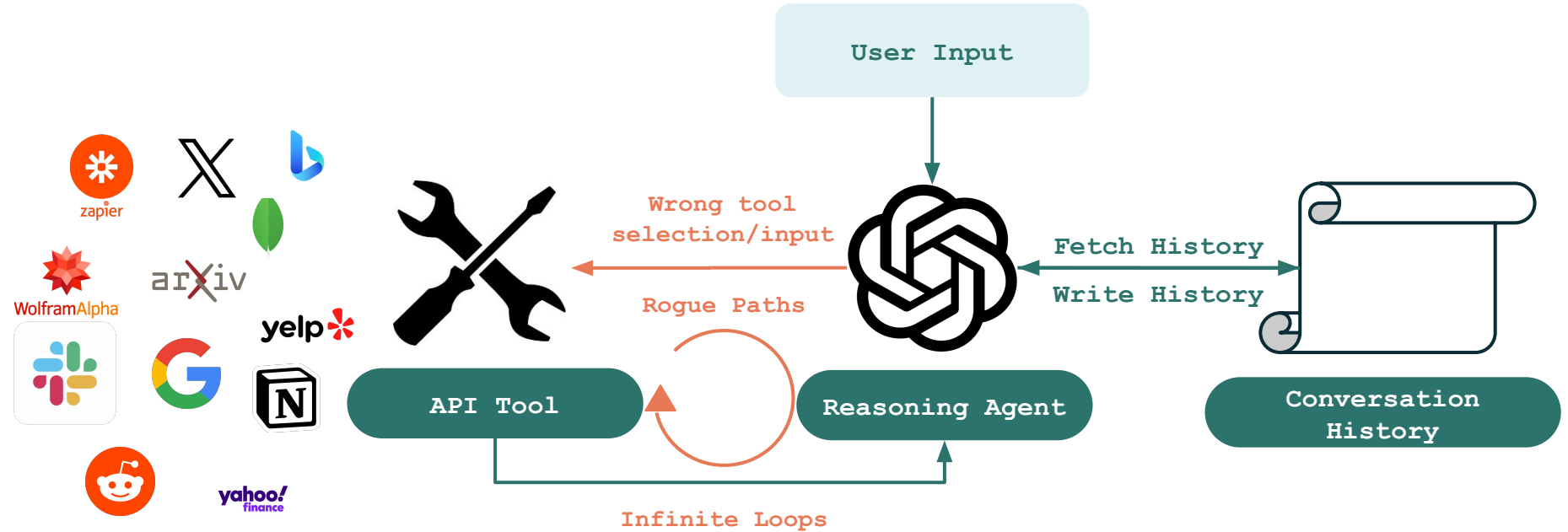
Data agents for real-time retrieval



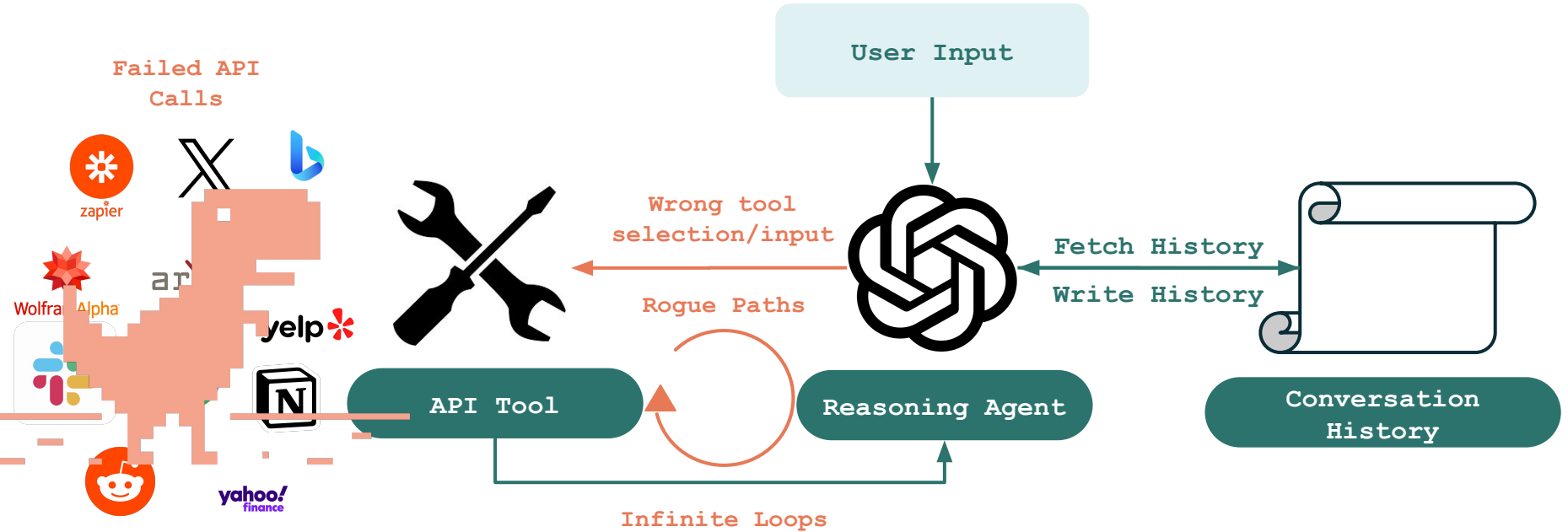
Agent Failure Modes



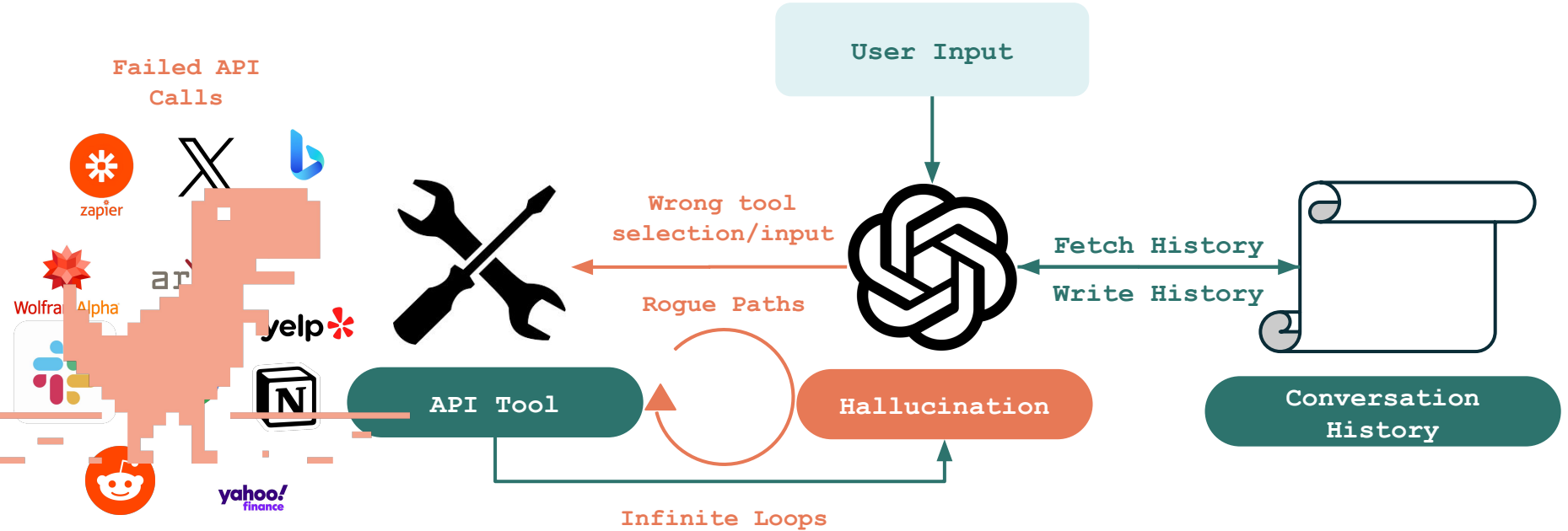
Agent Failure Modes



Agent Failure Modes

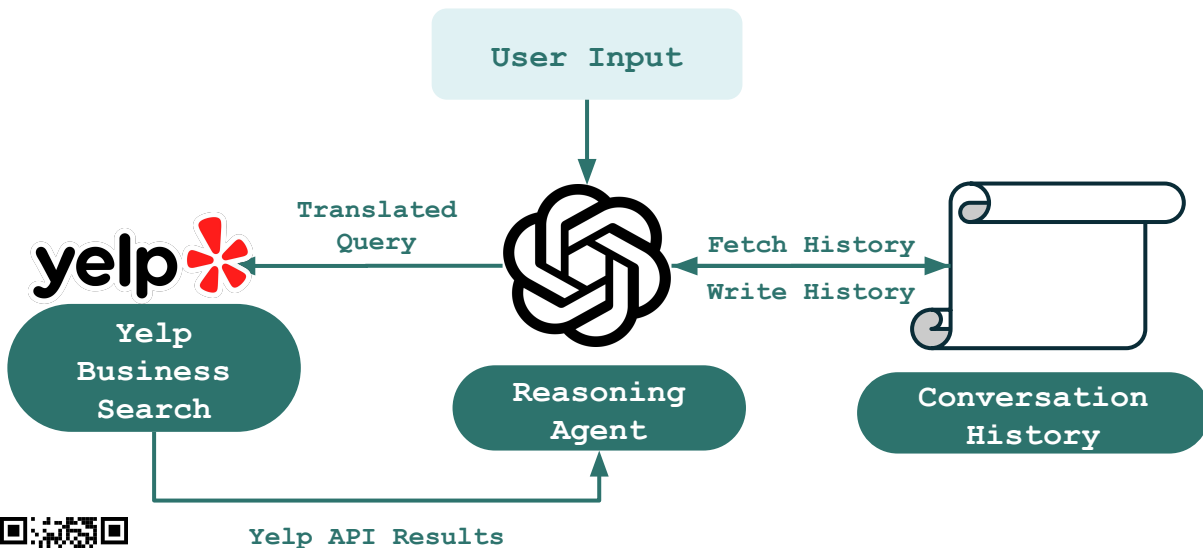


Agent Failure Modes



Data agents for real-time retrieval

Example: Restaurant Information Chatbot



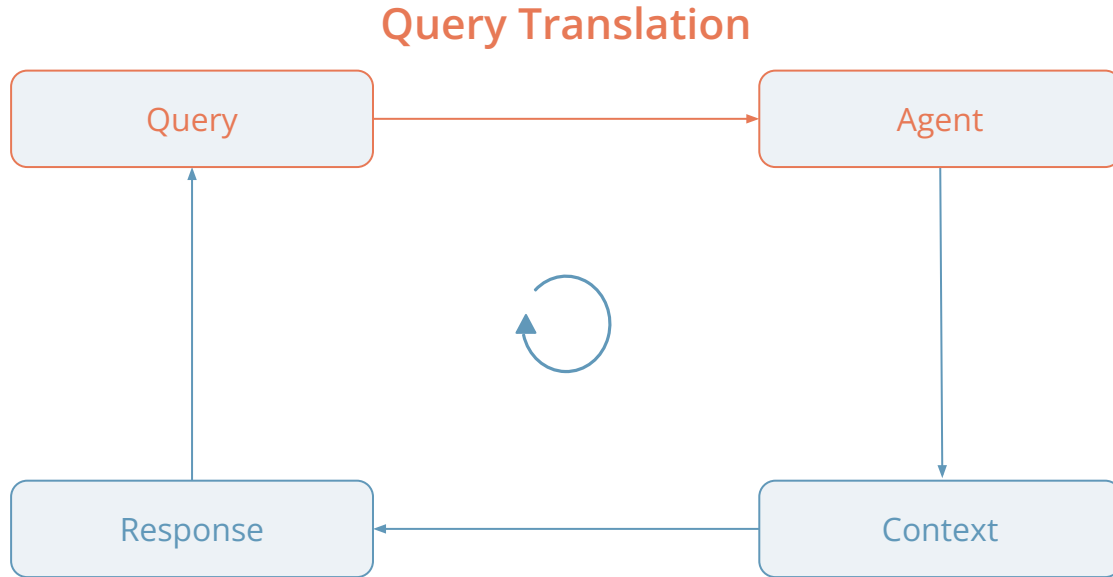
Knowledge source

- Yelp API



Notebook example:
<https://tinyurl.com/agent-eval>

Typical Eval for Non-Hallucination with Agents

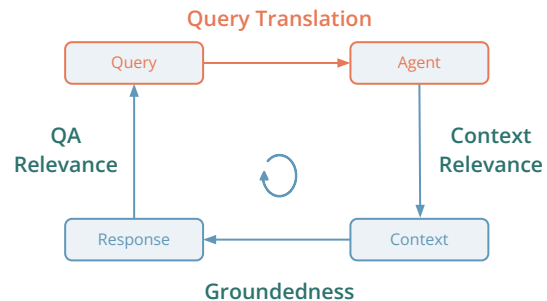


Query Translation

- Custom evals can verify that the correct retrieval is occurring

query_translation_score = 1

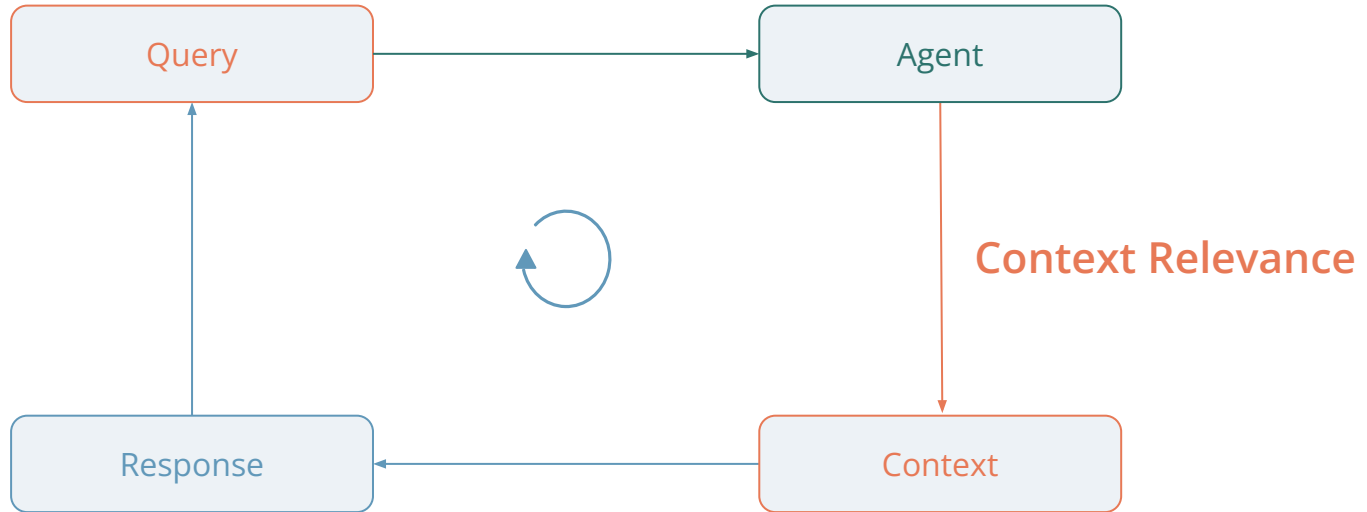
	question1	question2	result
0	What's the address of Gumbo Social in San Francisco?	address of Gumbo Social in San Francisco	1.0



Notebook example:
<https://tinyurl.com/agent-eval>



Typical Eval for Non-Hallucination with Agents

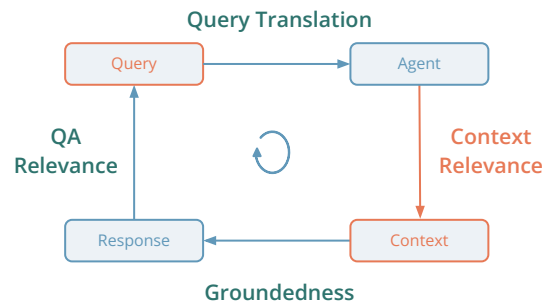


Context Relevance

- Test that retrieved context from agent relevant to user query

qs_relevance = 1

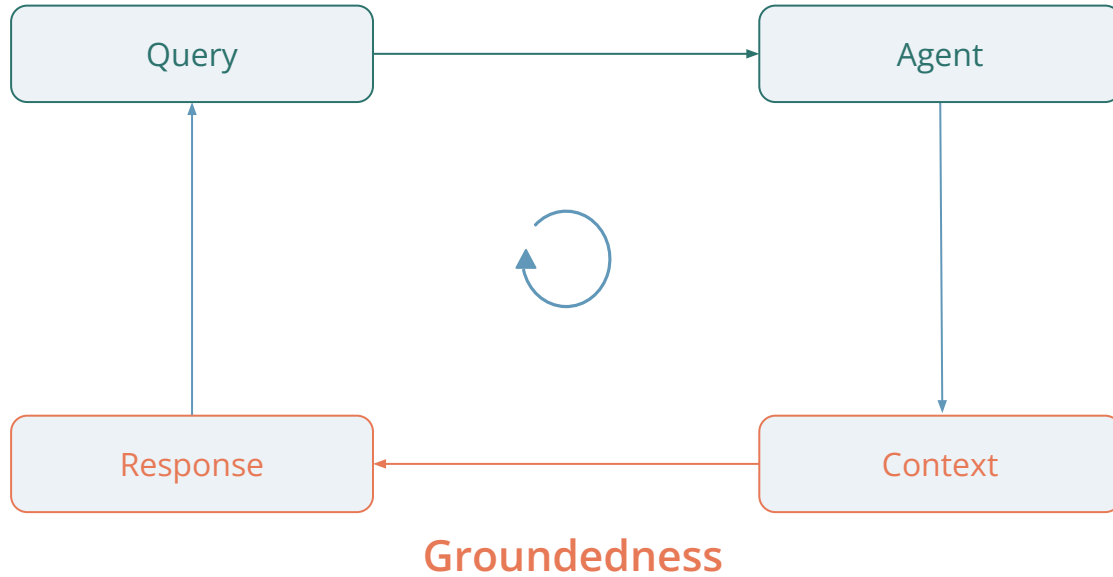
	question	statement	result
0	What's the address of Gumbo Social in San Francisco?	5176 3rd St, San Francisco, CA 94124	1.0



Notebook example:
<https://tinyurl.com/agent-eval>



Typical Eval for Non-Hallucination with Agents

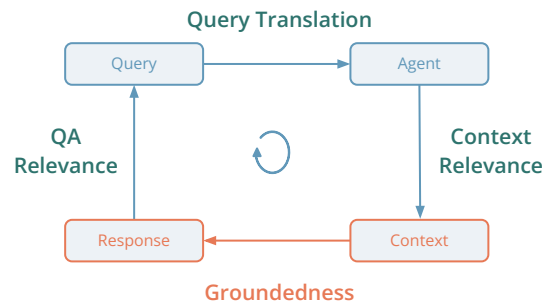


Groundedness

- Test that the app response is supported by the context

groundedness_measure = 1

	source	statement	result	reason
0	5176 3rd St, San Francisco, CA 94124	The address of Gumbo Soc	Statement Sentence: The address of Gumbo Social in San Francisco is 5176 3rd St, San Francisco, CA 94124. Supporting Evidence: 5176 3rd St, San Francisco, CA 94124 Score: 10	



Notebook example:
<https://tinyurl.com/agent-eval>



Groundedness

- Test that the app response is supported by the context

Statement Sentence: The best diner in Toronto is subjective and can vary depending on personal preferences.

Supporting Evidence: It is difficult to determine the best diner in Toronto without prior knowledge. However, based on the context information provided, some of the diners that may be worth considering are Golden Diner Family Restaurant, KOS Cafe and Restaurant, The Diner's Corner, The Lakeview, and Rosedale Diner.

Score: 10

Statement Sentence: However, some popular diners in Toronto include Golden Diner Family Restaurant, KOS Cafe and Restaurant, The Diner's Corner, The Lakeview, and Rosedale Diner.

Supporting Evidence: Golden Diner Family Restaurant, KOS Cafe and Restaurant, The Diner's Corner, The Lakeview, and Rosedale Diner.

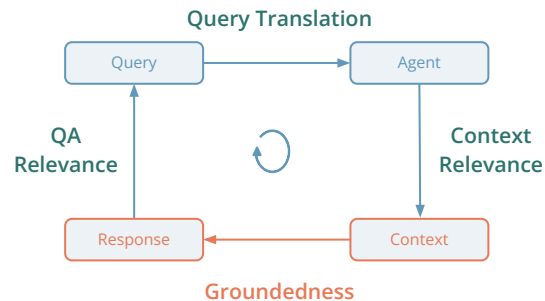
Score: 10

Statement Sentence: I recommend trying them out and deciding for yourself which one you like the best. Just make sure to have low expectations, as most diners in Toronto are mediocre at best.

Supporting Evidence: NOTHING FOUND

Score: 1

Notebook example:
<https://tinyurl.com/agent-eval>



Groundedness

- Test that the app response is supported by the context

Statement Sentence: The best pizza places in New York City are Rubirosa, L'industrie Pizzeria, and Lombardi's Pizza.

Supporting Evidence: Based on the context information, the best pizza places in New York City are Rubirosa, L'industrie Pizzeria, and Lombardi's Pizza.

Score: 10

Statement Sentence: These places are known for their delicious and authentic pizzas.

Supporting Evidence: NOTHING FOUND

Score: 1

Statement Sentence: Rubirosa offers a variety of thin-crust pizzas with fresh toppings, while L'industrie Pizzeria is known for its creative and unique pizza combinations.

Supporting Evidence: NOTHING FOUND

Score: 1

Statement Sentence: Lombardi's Pizza, on the other hand, is a historic pizzeria that has been serving classic New York-style pizzas since 1905.

Supporting Evidence: NOTHING FOUND

Score: 1

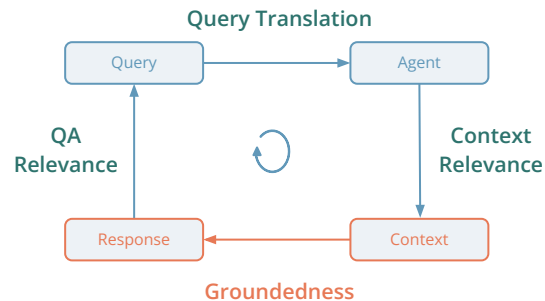
Statement Sentence: So, if you're looking for the best pizza in New York City, these three places are definitely worth a visit. Enjoy your pizza adventure!

Supporting Evidence: NOTHING FOUND

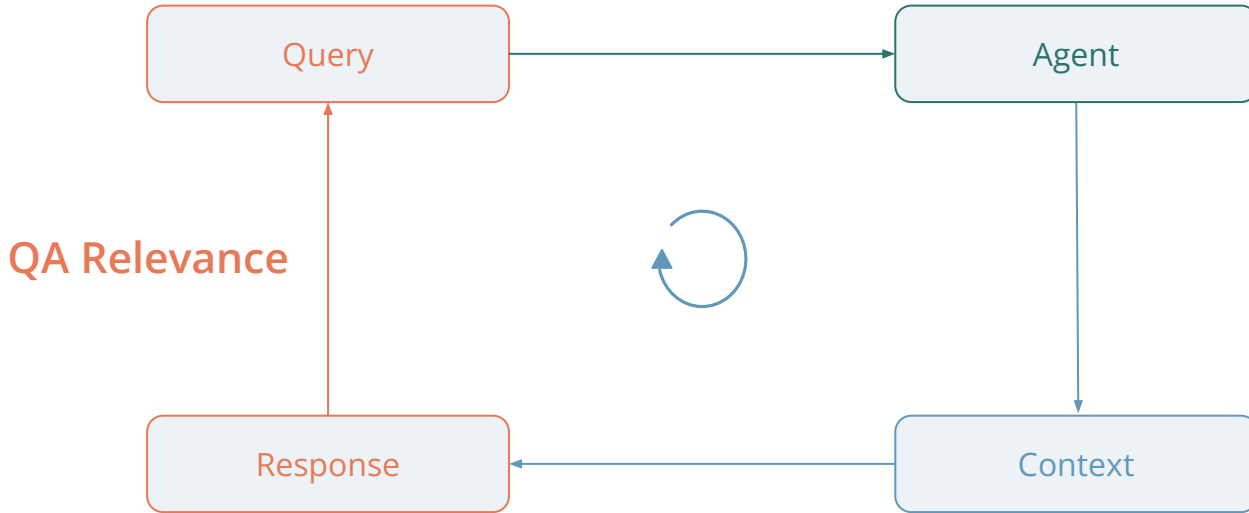
Score: 1

Notebook example:

<https://tinyurl.com/agent-eval>



Typical Eval for Non-Hallucination with Agents



Question Answer Relevance

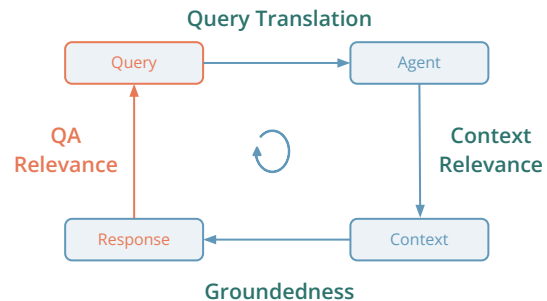
- Is the app's response helpful?

relevance = 1

	prompt	response	result
0	What's the address of Gumbo Social	The address of Gumbo Social in San Francisco is 5176 3rd St, San Francisco, CA 94124. Now, do yourself a favor and go enjoy some gumbo!	1.0



Notebook example:
<https://tinyurl.com/agent-eval>



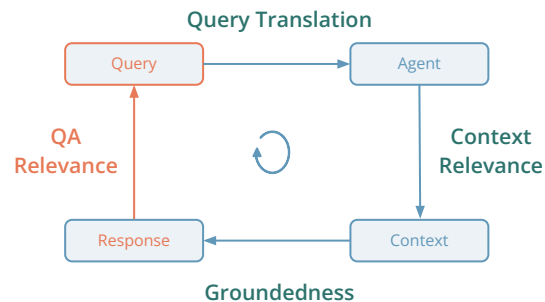
Question Answer Relevance

- Is the app's response helpful?

relevance = 0.7

	prompt	result
0	What's the best diner in Toronto?	0.7

The best diner in Toronto is subjective and can vary depending on personal preferences. However, some popular diners in Toronto include Golden Diner Family Restaurant, KOS Cafe and Restaurant, The Diner's Corner, The Lakeview, and Rosedale Diner. I recommend trying them out and deciding for yourself which one you like the best. Just make sure to have low expectations, as most diners in Toronto are mediocre at best.



Notebook example:
<https://tinyurl.com/agent-eval>

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	Select App
16	1	\$0	1.19k	0.72 ▲ medium	0.91 ✔ high	

YelpAgent

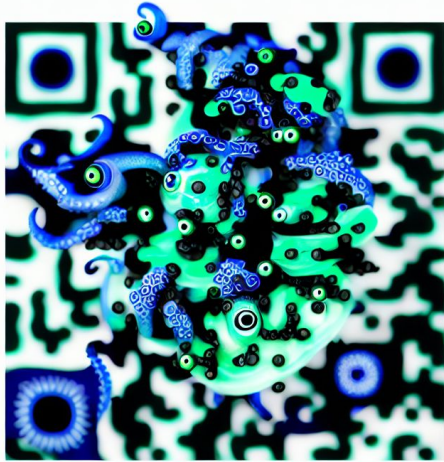
Records	Average Latency...	Total Cost (USD)	Total Tokens	query_translati...	agreement_me...	relevance	qs_relevance	ratings_usage	Select App
15	8.07	\$0.76	147k	0.95 ✔ high	0.81 ✔ high	0.8 ✔ high	0.88 ✔ high	0.4 ● low	

Notebook example:
<https://tinyurl.com/agent-eval>

Get keys: <https://docs.developer.yelp.com/docs/yelp-platform>



Find us in Open Source!



github.com/truera/trulens

Give us a star, request a feature, contribute

Sample LLM agent eval notebooks

truera

- [LlamaIndex + TruLens](#)
- [LangChain + TruLens](#)