arize

Evaluating Agents and Assistants

Jason Lopatecki | Cofounder & CEO, Arize AI

2024 - “The Year of Agents”

Agent Tooling

MultiOn

crewai

LlamaIndex

LangChain

Tool Calling Capabilities

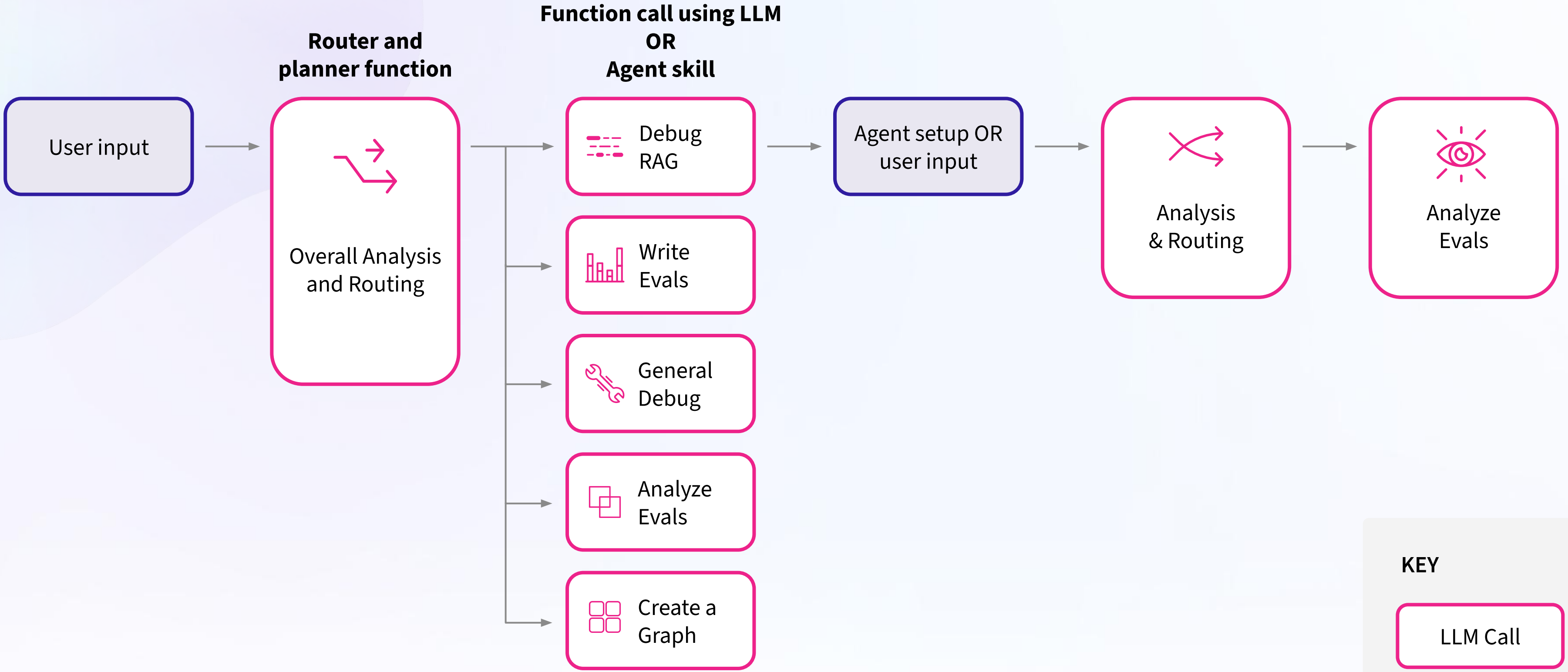
Vertex.ai

OpenAI

ANTHROP\C

MISTRAL
AI_

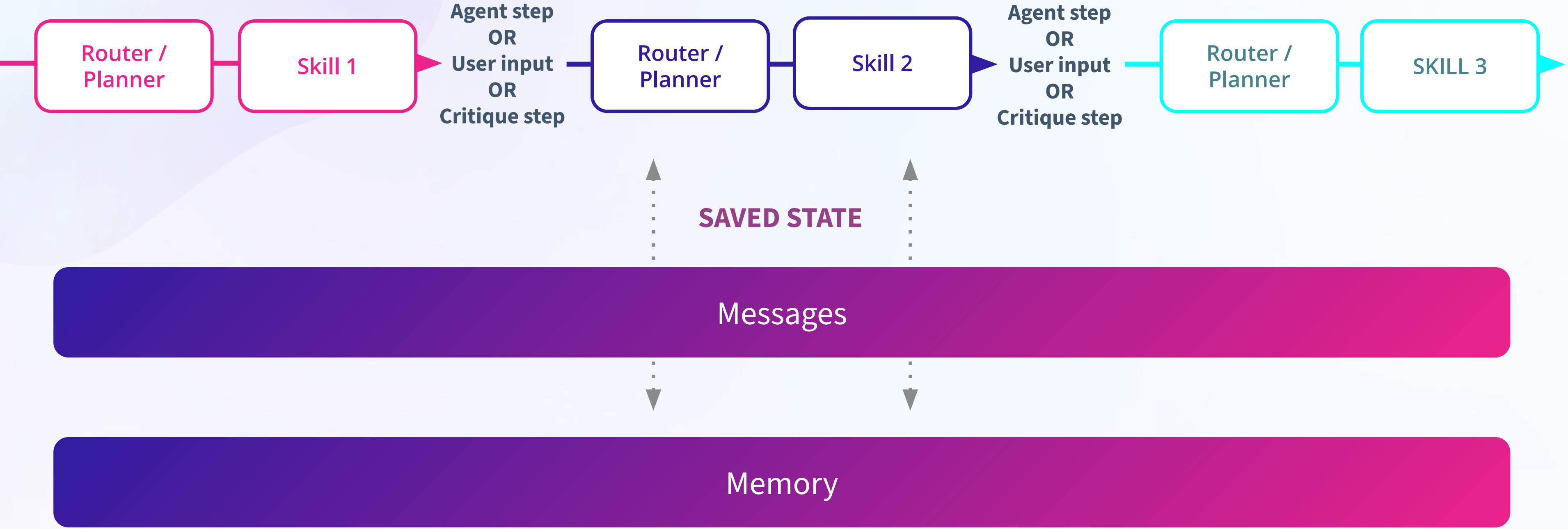
What We See in the Ecosystem


Router and planner function
Function call using LLM OR Agent skill
User input
Overall Analysis and Routing
Debug RAG
Write Evals
General Debug
Analyze Evals
Create a Graph
Agent setup OR user input
Analysis & Routing
Analyze Evals

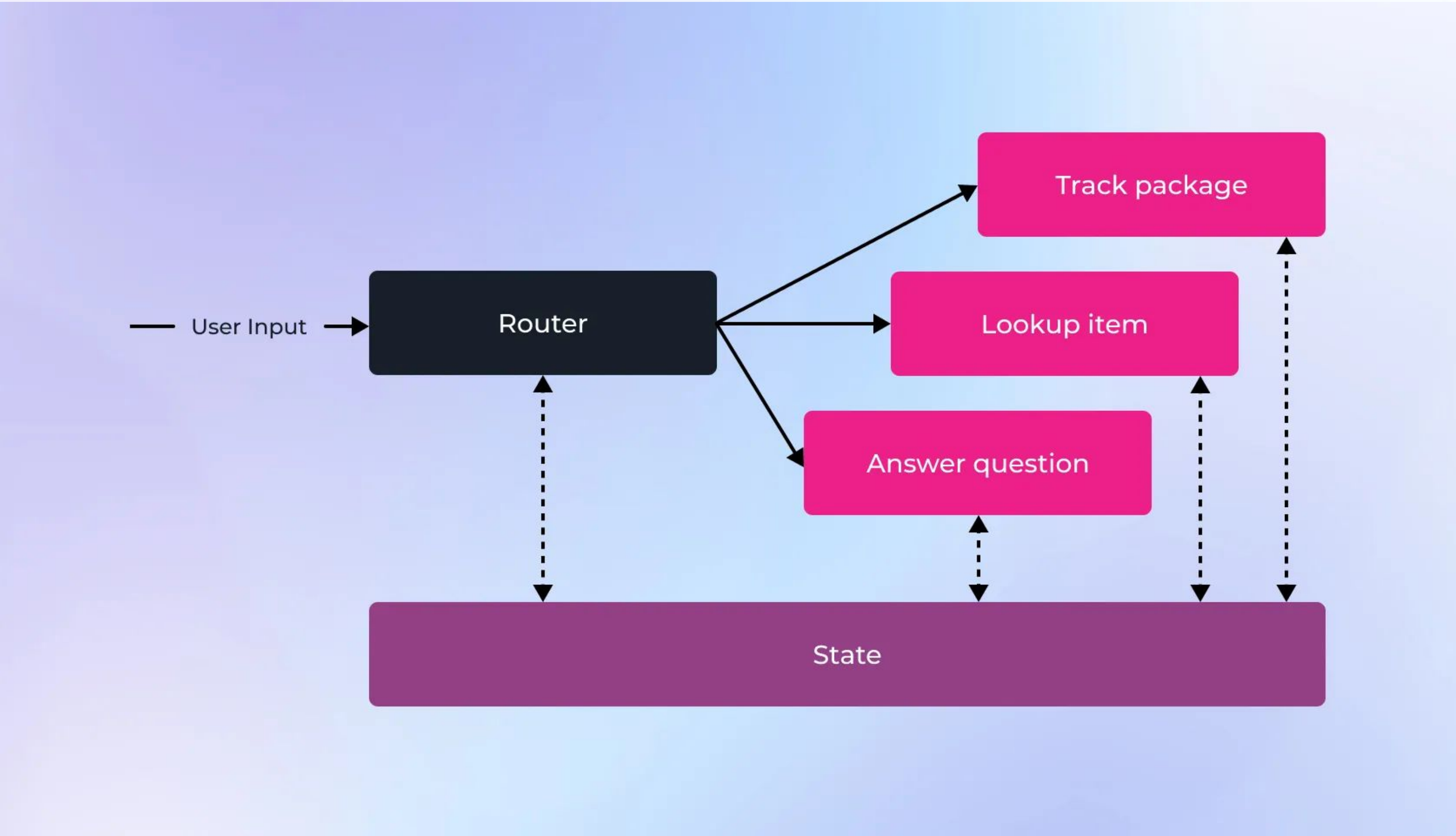
KEY

LLM Call

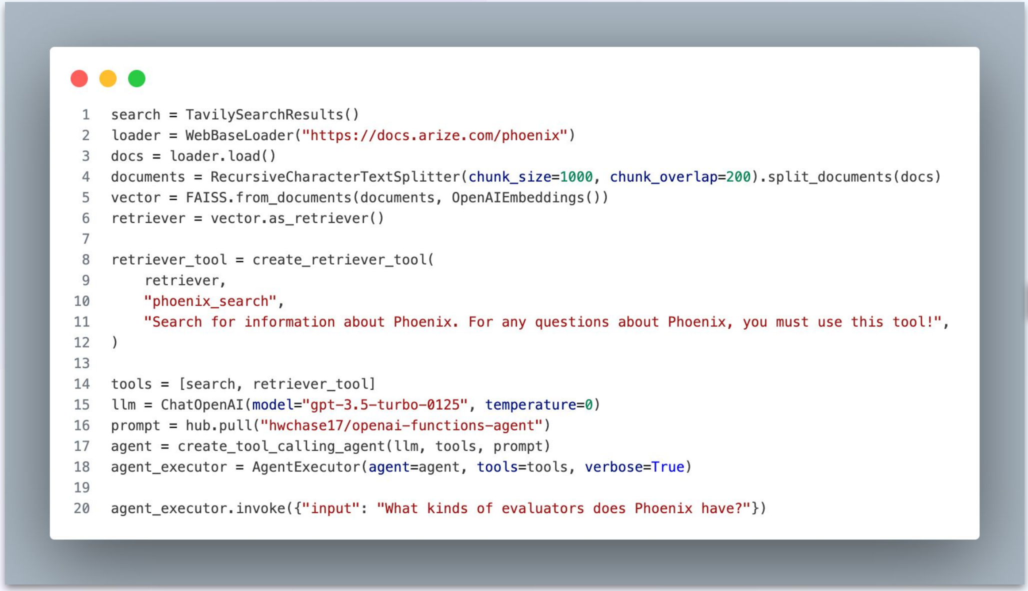
Architecture: Agent Process



Mixing Branches of LLM Calls with Tools & State



Difficulties of Evaluating Agents

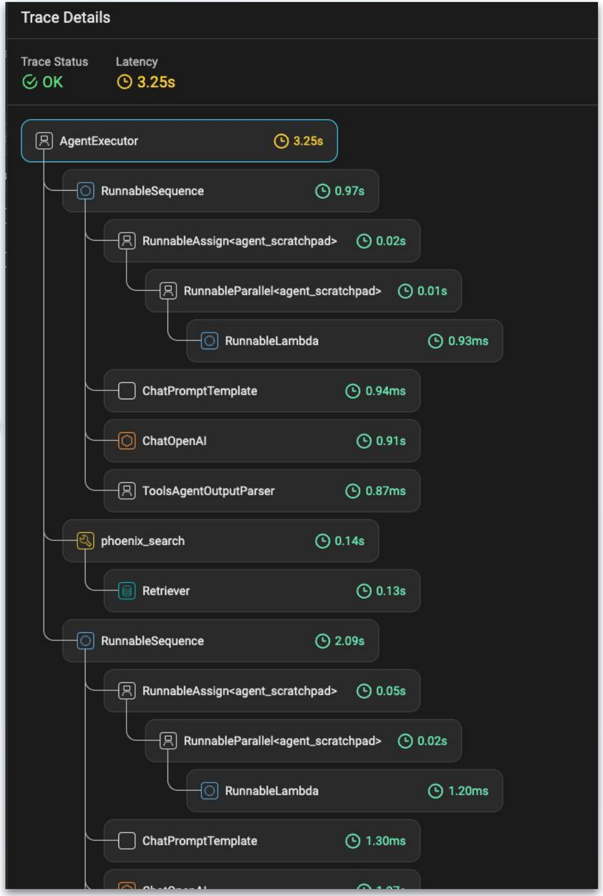


```
search = TavilySearchResults()
loader = WebBaseLoader("https://docs.arize.com/phoenix")
docs = loader.load()
documents = RecursiveCharacterTextSplitter(chunk_size=1000, chunk_overlap=200).split_documents(docs)
vector = FAISS.from_documents(documents, OpenAIEmbeddings())
retriever = vector.as_retriever()

retriever_tool = create_retriever_tool(
    retriever,
    "phoenix_search",
    "Search for information about Phoenix. For any questions about Phoenix, you must use this tool!",
)

tools = [search, retriever_tool]
llm = ChatOpenAI(model="gpt-3.5-turbo-0125", temperature=0)
prompt = hub.pull("hwchase17/openai-functions-agent")
agent = create_tool_calling_agent(llm, tools, prompt)
agent_executor = AgentExecutor(agent=agent, tools=tools, verbose=True)

agent_executor.invoke({"input": "What kinds of evaluators does Phoenix have?"})
```

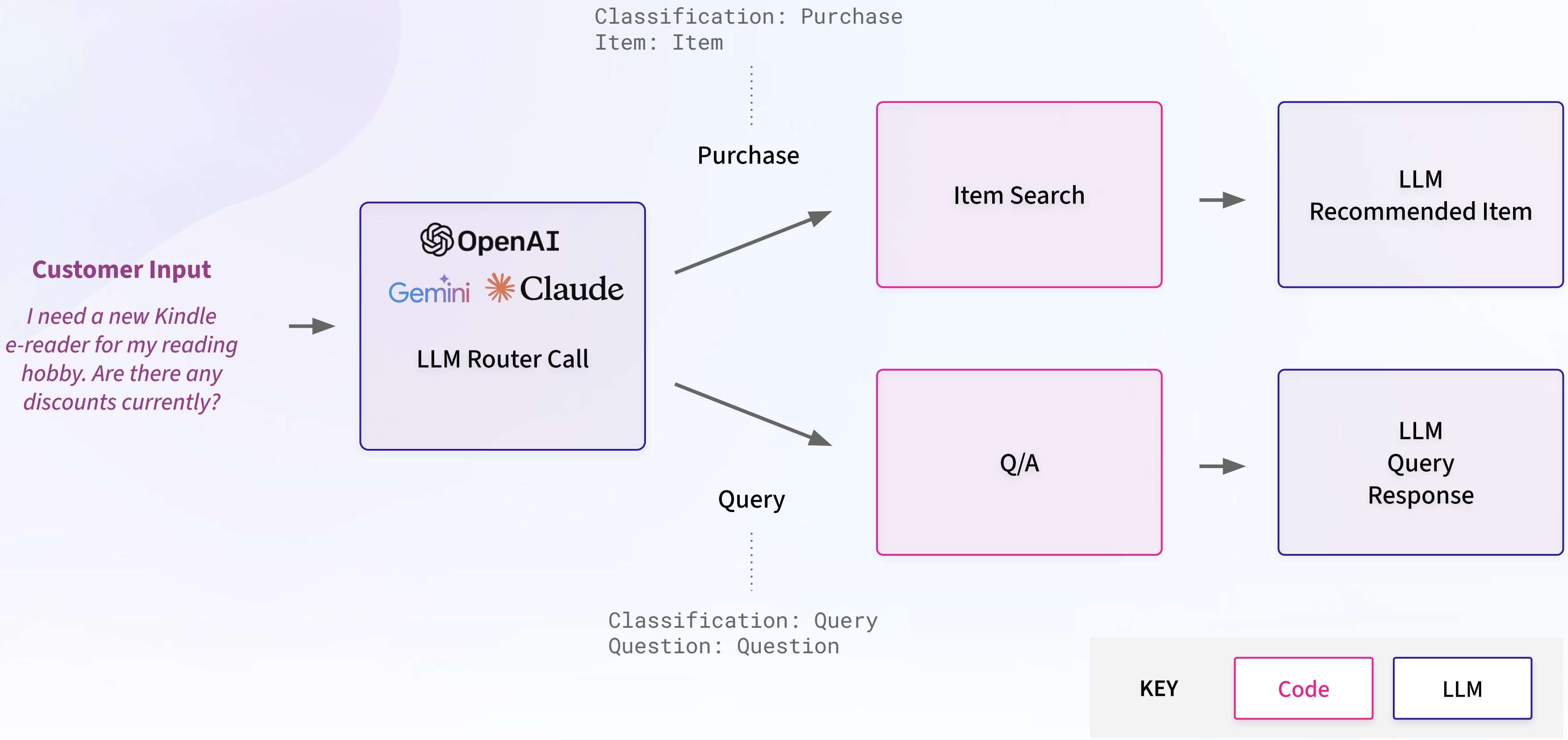


Trace Details

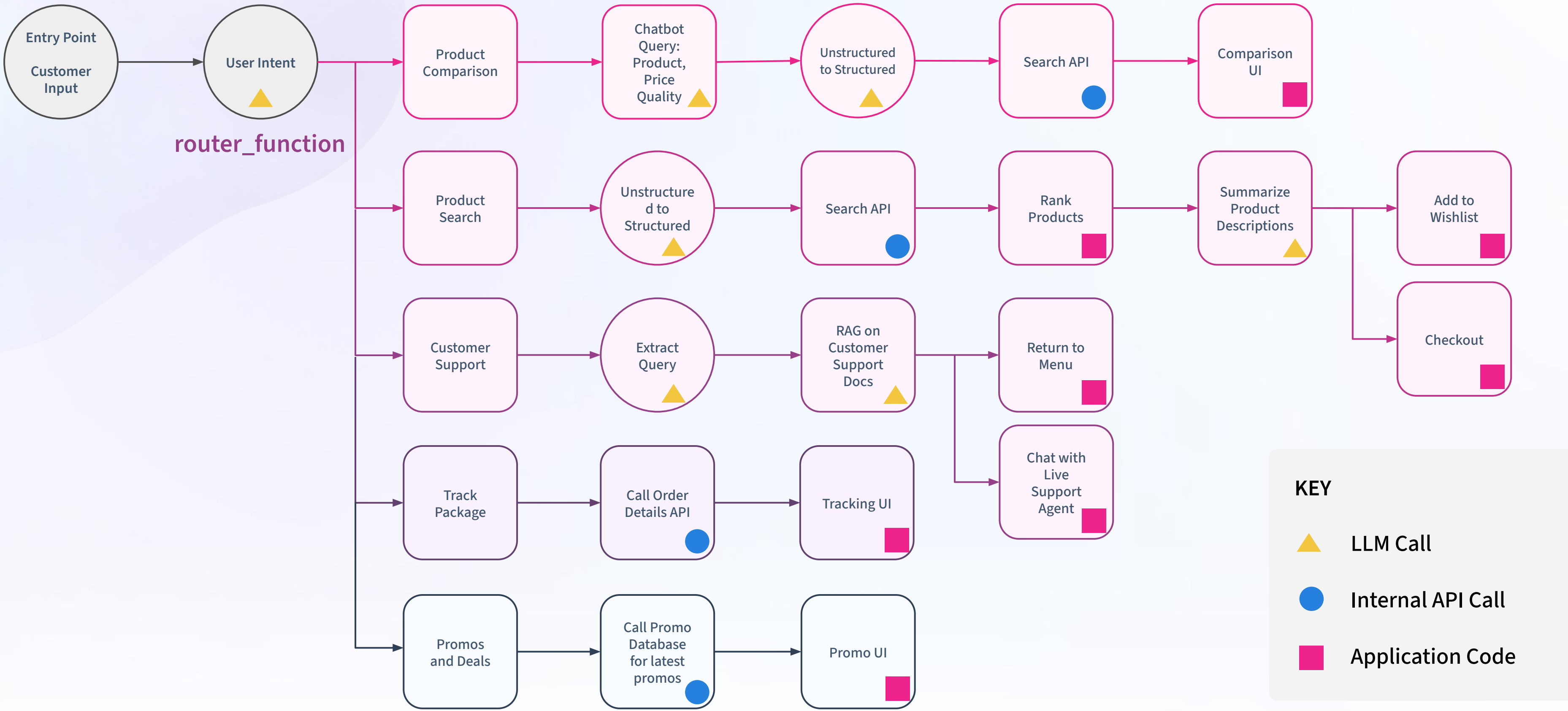
Trace Status: OK
Latency: 3.25s

```
AgentExecutor  3.25s
  RunnableSequence  0.97s
    RunnableAssign<agent_scratchpad>  0.02s
      RunnableParallel<agent_scratchpad>  0.01s
        RunnableLambda  0.93ms
    ChatPromptTemplate  0.94ms
    ChatOpenAI  0.91s
    ToolsAgentOutputParser  0.87ms
  phoenix_search  0.14s
    Retriever  0.13s
  RunnableSequence  2.09s
    RunnableAssign<agent_scratchpad>  0.05s
      RunnableParallel<agent_scratchpad>  0.02s
        RunnableLambda  1.20ms
    ChatPromptTemplate  1.30ms
```

Example Router: Chat to Purchase App

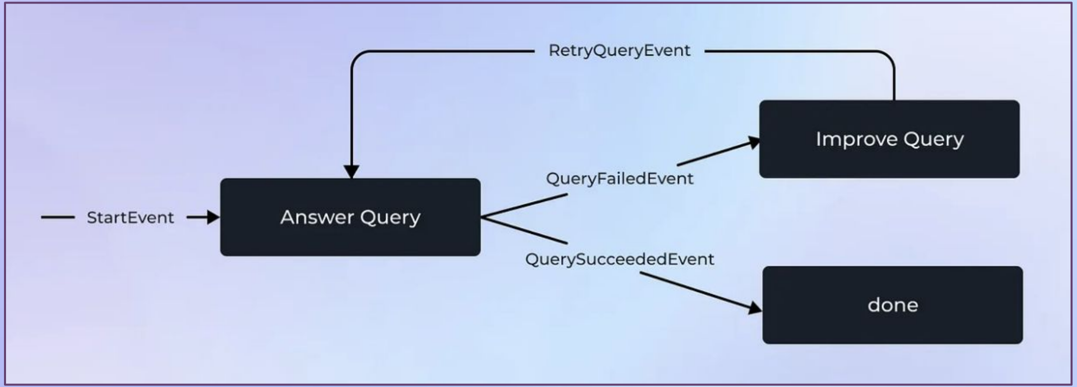


Under the Hood: Chat-to-purchase Router

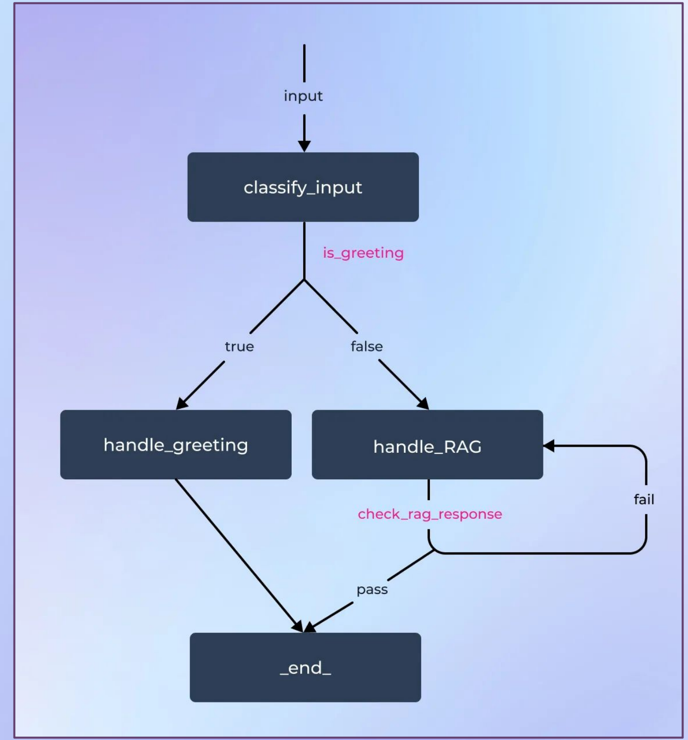


Agent Frameworks

LlamaIndex Workflows



LangGraph



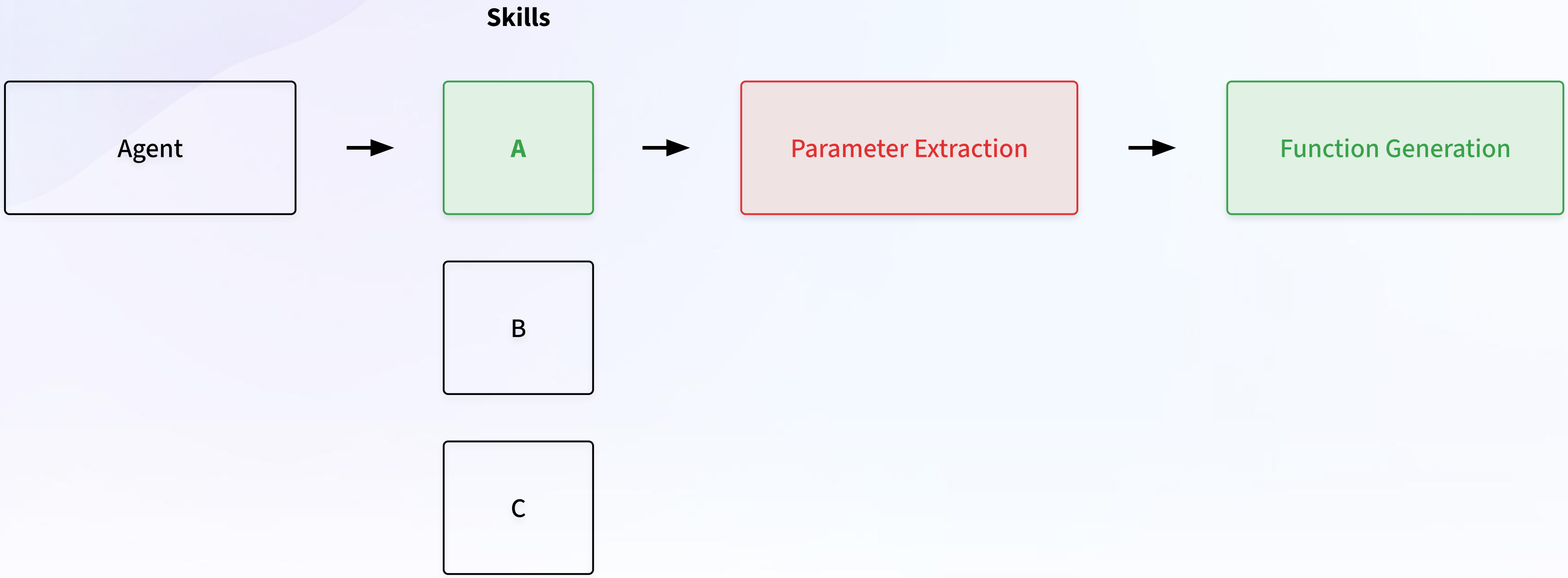
“Working” ≠ Performing Well

So how do you evaluate these agents?

0. (Add observability to your application!)
1. Build a set of test cases
2. Breakdown individual agent steps
3. Create evaluators for each step
4. Experiment and iterate

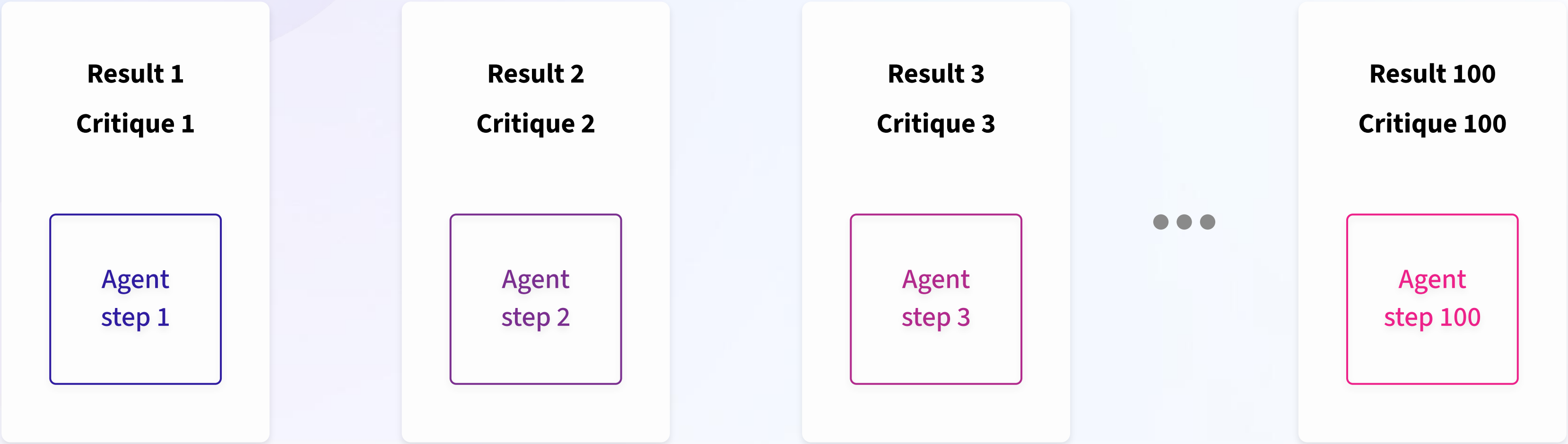
Breakdown of “Steps”

Am I using the right skill correctly?



Am I Converging?

LLM as a judge eval: Agent convergence



Result 1

Critique 1

Agent step 1

Result 2

Critique 2

Agent step 2

Result 3

Critique 3

Agent step 3

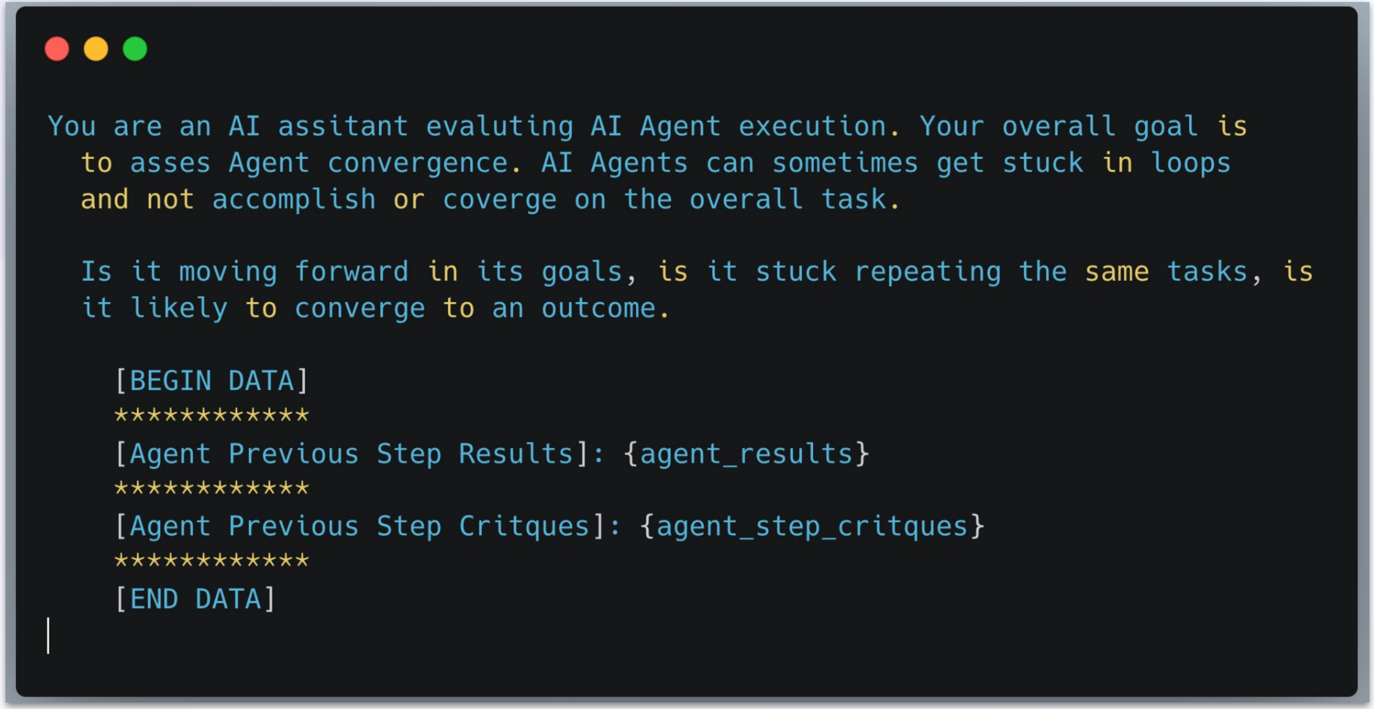
●●●

Result 100

Critique 100

Agent step 100

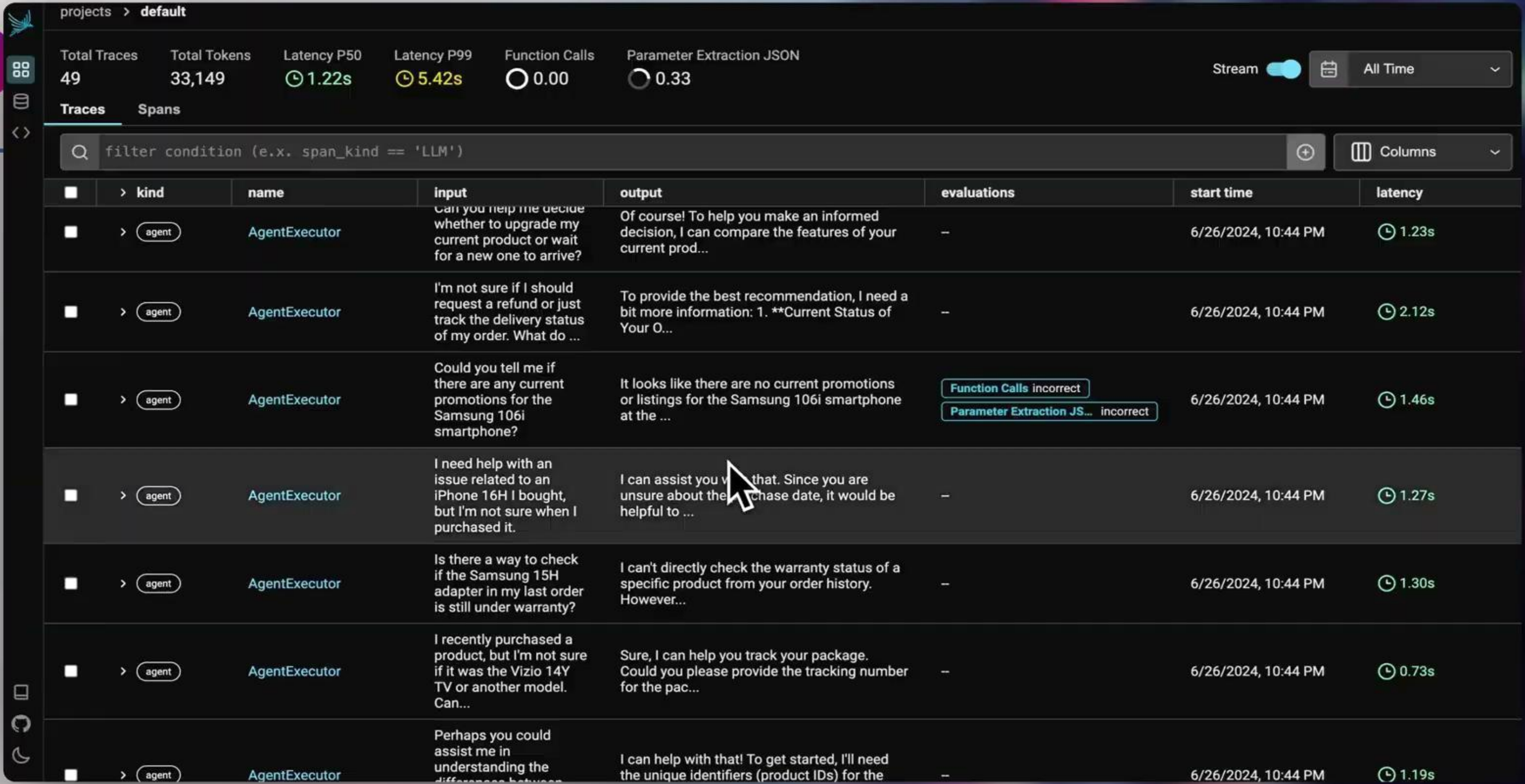
Teaser: Convergence Evals



```
You are an AI assitant evaluting AI Agent execution. Your overall goal is
  to asses Agent convergence. AI Agents can sometimes get stuck in loops
  and not accomplish or coverge on the overall task.

  Is it moving forward in its goals, is it stuck repeating the same tasks, is
  it likely to converge to an outcome.

    [BEGIN DATA]
    ************
    [Agent Previous Step Results]: {agent_results}
    ************
    [Agent Previous Step Critques]: {agent_step_critques}
    ************
    [END DATA]
```



projects > **default**

Total Traces	Total Tokens	Latency P50	Latency P99	Function Calls	Parameter Extraction JSON
49	33,149	1.22s	5.42s	0.00	0.33

Stream · All Time

Traces · Spans

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

Columns

□	kind	name	input	output	evaluations	start time	latency
□	agent	AgentExecutor	Can you help me decide whether to upgrade my current product or wait for a new one to arrive?	Of course! To help you make an informed decision, I can compare the features of your current prod...	–	6/26/2024, 10:44 PM	1.23s
□	agent	AgentExecutor	I'm not sure if I should request a refund or just track the delivery status of my order. What do ...	To provide the best recommendation, I need a bit more information: 1. **Current Status of Your O...	–	6/26/2024, 10:44 PM	2.12s
□	agent	AgentExecutor	Could you tell me if there are any current promotions for the Samsung 106i smartphone?	It looks like there are no current promotions or listings for the Samsung 106i smartphone at the ...	Function Calls incorrect Parameter Extraction JS... incorrect	6/26/2024, 10:44 PM	1.46s
□	agent	AgentExecutor	I need help with an issue related to an iPhone 16H I bought, but I'm not sure when I purchased it.	I can assist you w that. Since you are unsure about the chase date, it would be helpful to ...	–	6/26/2024, 10:44 PM	1.27s
□	agent	AgentExecutor	Is there a way to check if the Samsung 15H adapter in my last order is still under warranty?	I can't directly check the warranty status of a specific product from your order history. However...	–	6/26/2024, 10:44 PM	1.30s
□	agent	AgentExecutor	I recently purchased a product, but I'm not sure if it was the Vizio 14Y TV or another model. Can...	Sure, I can help you track your package. Could you please provide the tracking number for the pac...	–	6/26/2024, 10:44 PM	0.73s
□	agent	AgentExecutor	Perhaps you could assist me in understanding the differences between	I can help with that! To get started, I'll need the unique identifiers (product IDs) for the	–	6/26/2024, 10:44 PM	1.19s

Q/A

✨ Star us on Github ✨

github.com/Arize-ai/phoenix