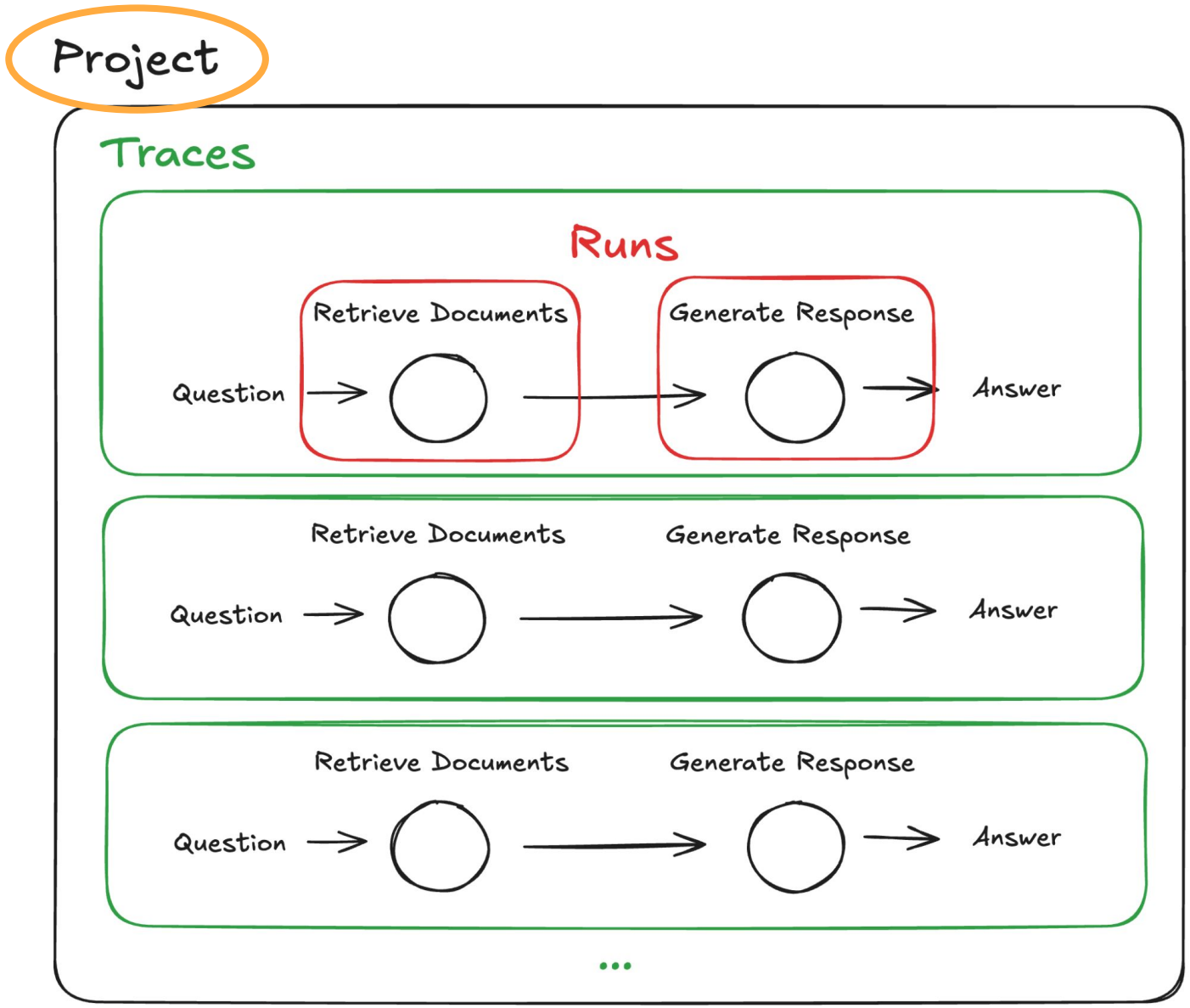
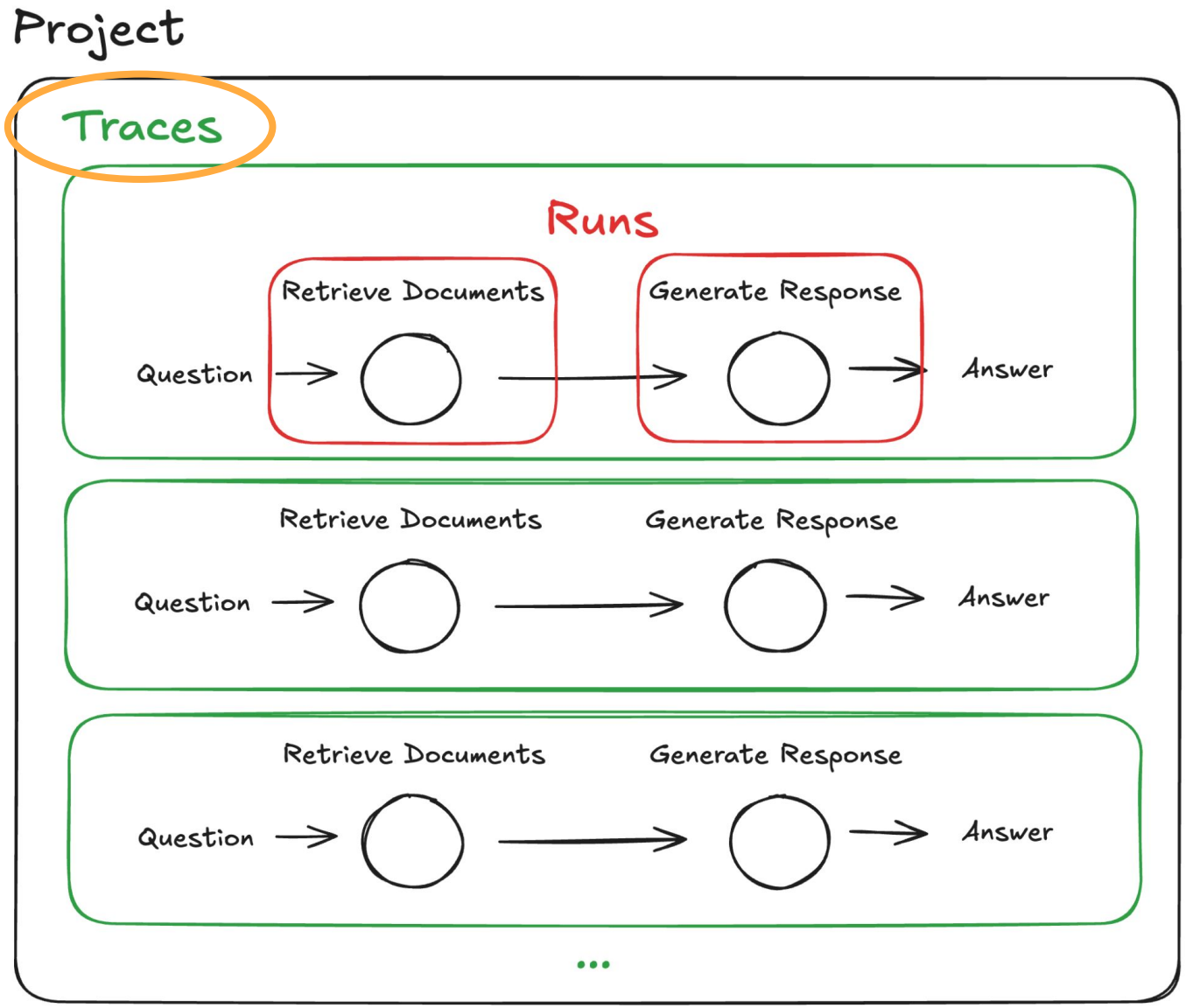
LangChain Academy

Tracing Basics

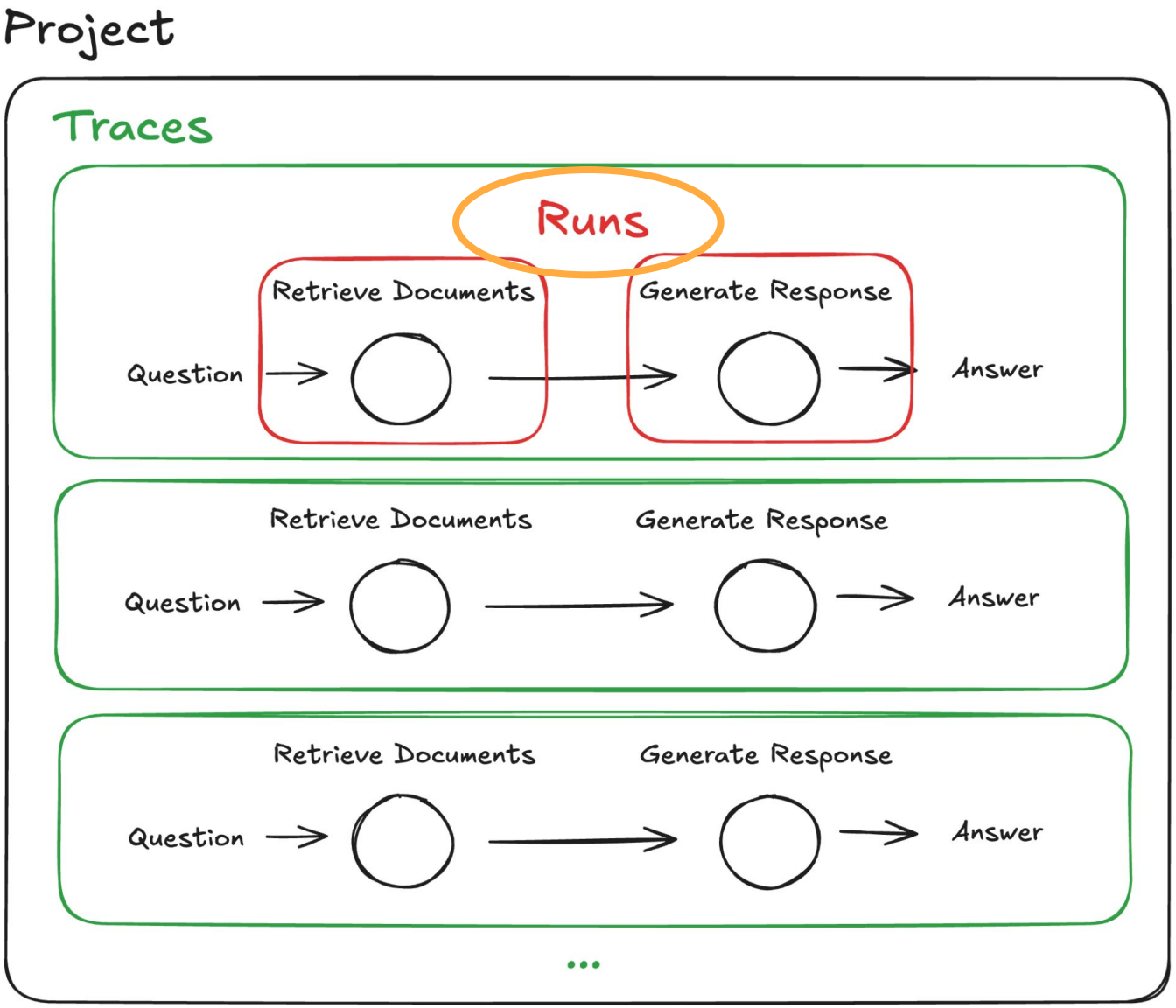
Tracing Concepts in LangSmith



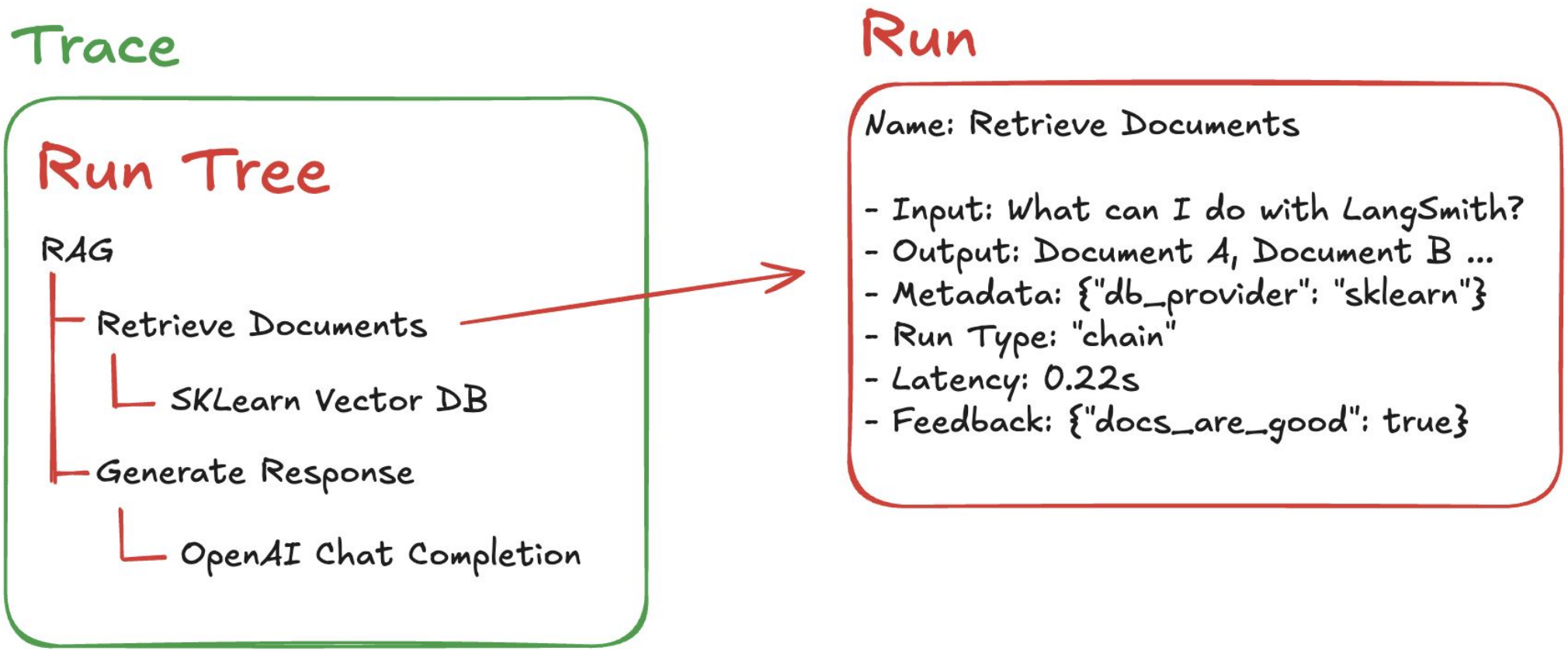
Tracing Concepts in LangSmith

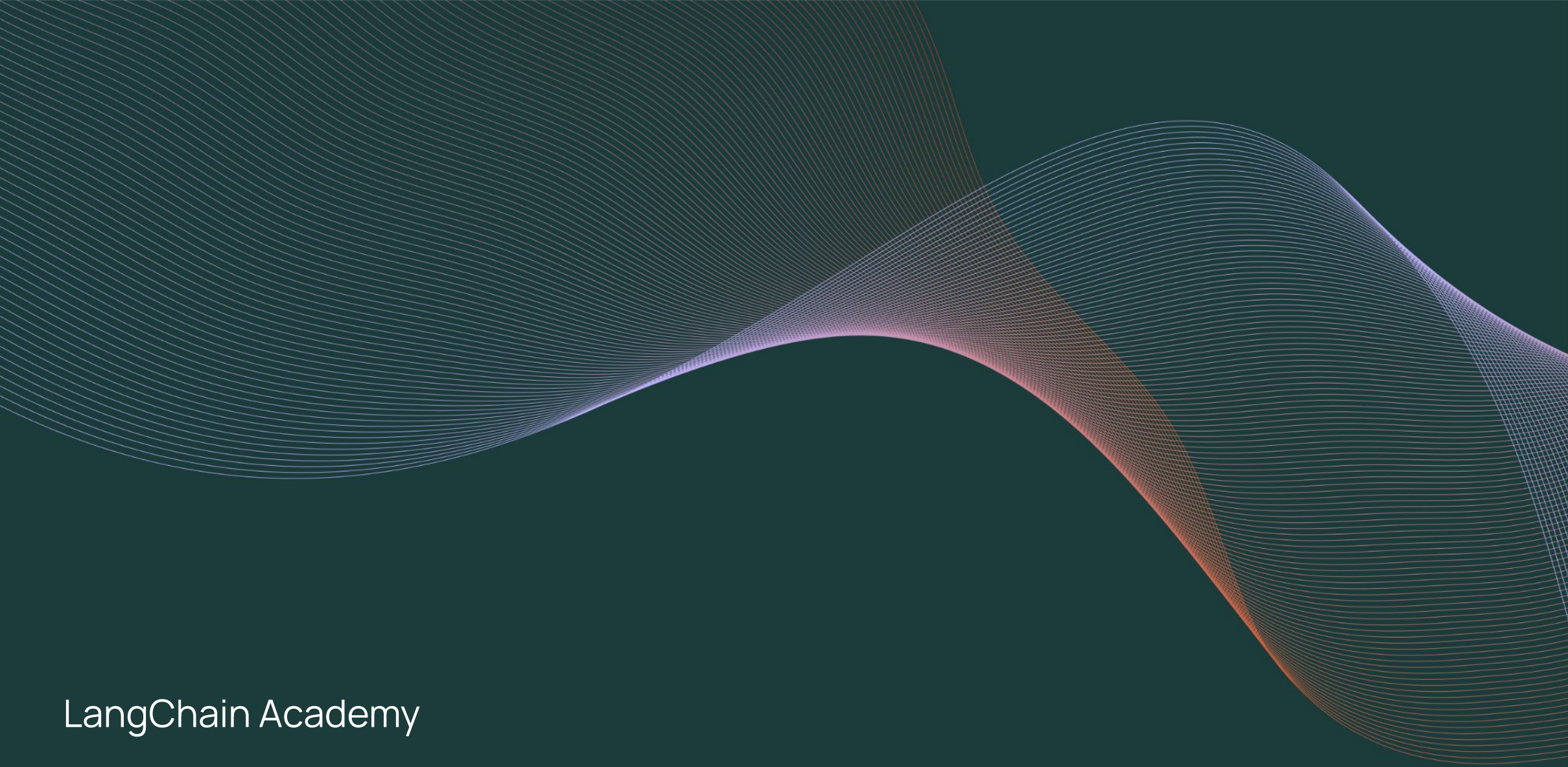


Tracing Concepts in LangSmith



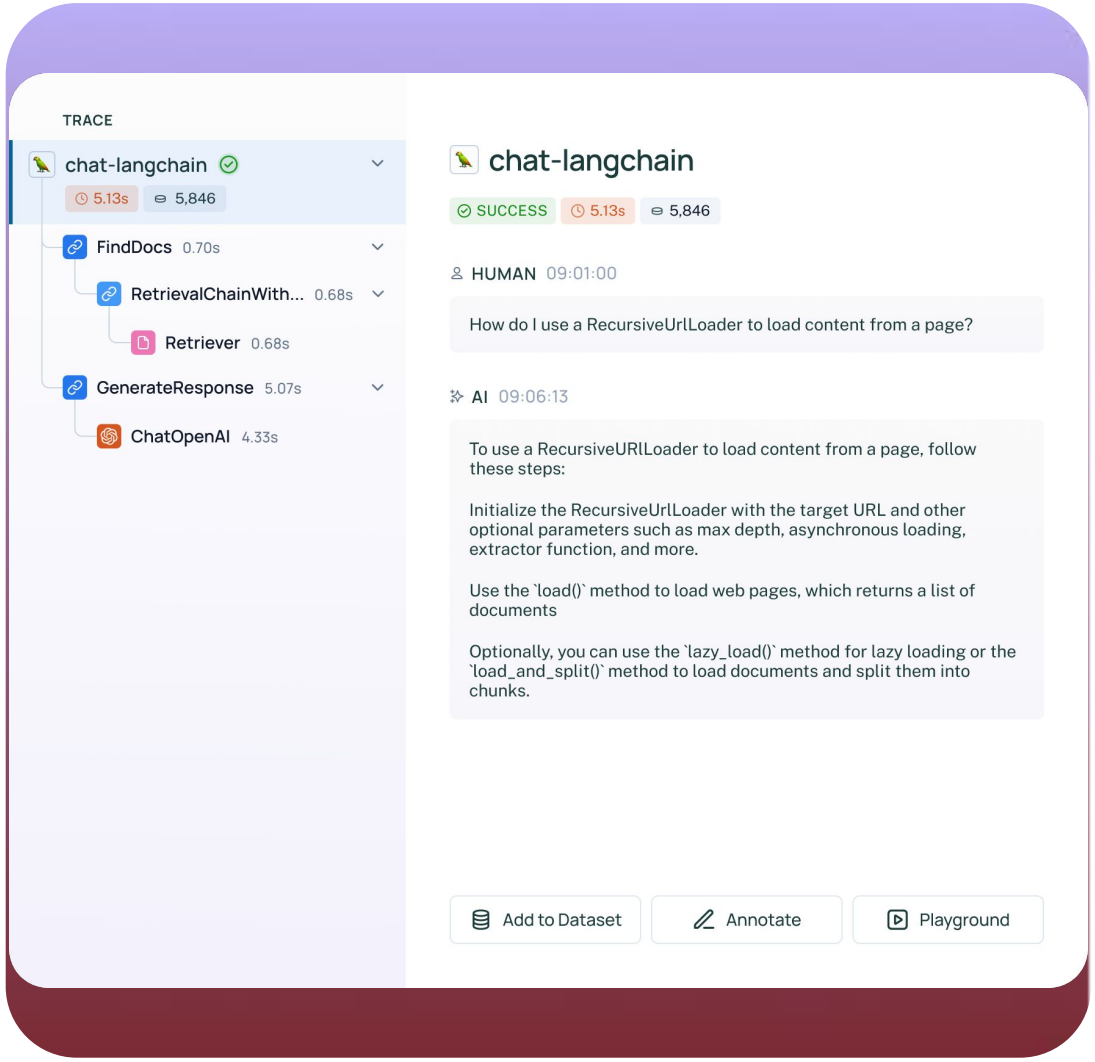
Tracing Concepts in LangSmith




LangChain Academy
Types of Runs

Logs vs. Tracing in LangSmith

```
((), {'supervisor': {'next': 'researcher'}})
----
(('researcher:ee1ffa79-59fc-f3f5-5da0-f2c473080a21',), {'agent': {'me
----
(('researcher:ee1ffa79-59fc-f3f5-5da0-f2c473080a21',), {'tools': {'me
----
(('researcher:ee1ffa79-59fc-f3f5-5da0-f2c473080a21',), {'agent': {'me
----
((), {'researcher': {'messages': [HumanMessage(content='The current p
----
((), {'supervisor': {'next': 'coder'}})
----
(('coder:b4f947a0-983f-d52d-631e-b6c6fb40e38e',), {'agent': {'message
----
(('coder:b4f947a0-983f-d52d-631e-b6c6fb40e38e',), {'tools': {'message
----
(('coder:b4f947a0-983f-d52d-631e-b6c6fb40e38e',), {'agent': {'message
----
((), {'coder': {'messages': [HumanMessage(content='The population of
----
((), {'supervisor': {'next': '__end__'}})
----
```



Types of Runs in LangSmith

LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

Types of Runs in LangSmith

LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

-  **LLM:** Invokes an LLM

Types of Runs in LangSmith

LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

-  **LLM:** Invokes an LLM
-  **Retriever:** Retrieves documents from databases or other sources

Types of Runs in LangSmith

LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

- **LLM:** Invokes an LLM
- **Retriever:** Retrieves documents from databases or other sources
- **Tool:** Executes actions with function calls

Types of Runs in LangSmith

LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

- **LLM:** Invokes an LLM
- **Retriever:** Retrieves documents from databases or other sources
- **Tool:** Executes actions with function calls
- **Chain:** Default type; combines multiple Runs into a larger process

Types of Runs in LangSmith

LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

- **LLM:** Invokes an LLM
- **Retriever:** Retrieves documents from databases or other sources
- **Tool:** Executes actions with function calls
- **Chain:** Default type; combines multiple Runs into a larger process
- **Prompt:** Hydrates a prompt to be used with an LLM

Types of Runs in LangSmith

LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

- **LLM:** Invokes an LLM
- **Retriever:** Retrieves documents from databases or other sources
- **Tool:** Executes actions with function calls
- **Chain:** Default type; combines multiple Runs into a larger process
- **Prompt:** Hydrates a prompt to be used with an LLM
- **Parser:** Extracts structured data

LangChain Academy

Alternative Ways to Trace

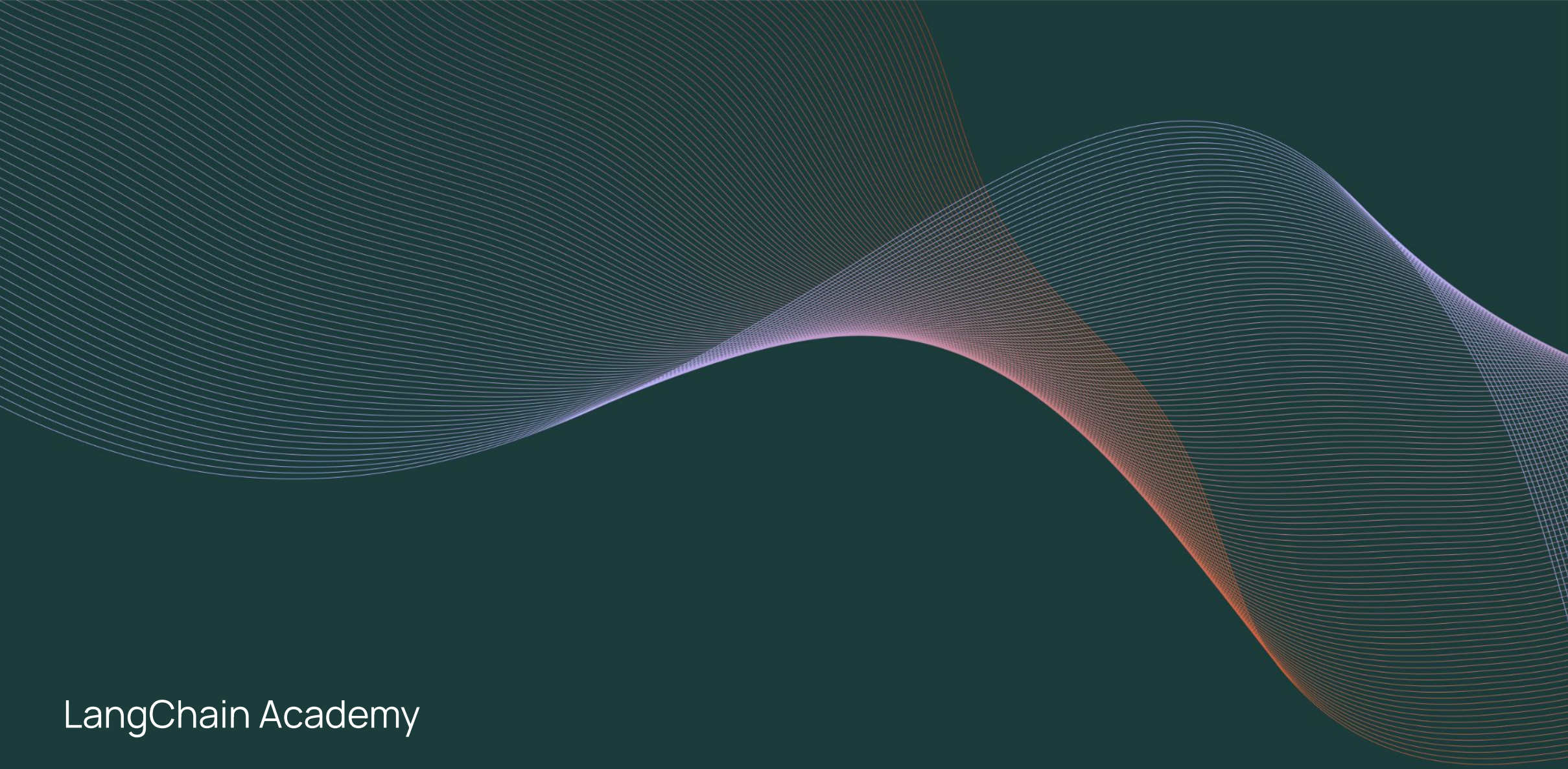
Method	When-to-use
@traceable	- Default way to set up tracing - Manages RunTree, inputs, and outputs for you

Method	When-to-use
@traceable	- Default way to set up tracing - Manages RunTree, inputs, and outputs for you
LangChain/LangGraph	- Get tracing out of the box!

Method	When-to-use
@traceable	- Default way to set up tracing - Manages RunTree, inputs, and outputs for you
LangChain/LangGraph	- Get tracing out of the box!
with trace():	- Want more control over what inputs and outputs get logged - Cannot use decorators or wrappers (not tracing a function)

Method	When-to-use
@traceable	- Default way to set up tracing - Manages RunTree, inputs, and outputs for you
LangChain/LangGraph	- Get tracing out of the box!
with trace():	- Want more control over what inputs and outputs get logged - Cannot use decorators or wrappers (not tracing a function)
wrap_openai()	- Want to use the OpenAI SDK directly

Method	When-to-use
@traceable	- Default way to set up tracing - Manages RunTree, inputs, and outputs for you
LangChain/LangGraph	- Get tracing out of the box!
with trace():	- Want more control over what inputs and outputs get logged - Cannot use decorators or wrappers (not tracing a function)
wrap_openai()	- Want to use the OpenAI SDK directly
RunTree	- Want more low level control over tracing configuration - Need to know run_id (for adding user feedback or logging to another service)

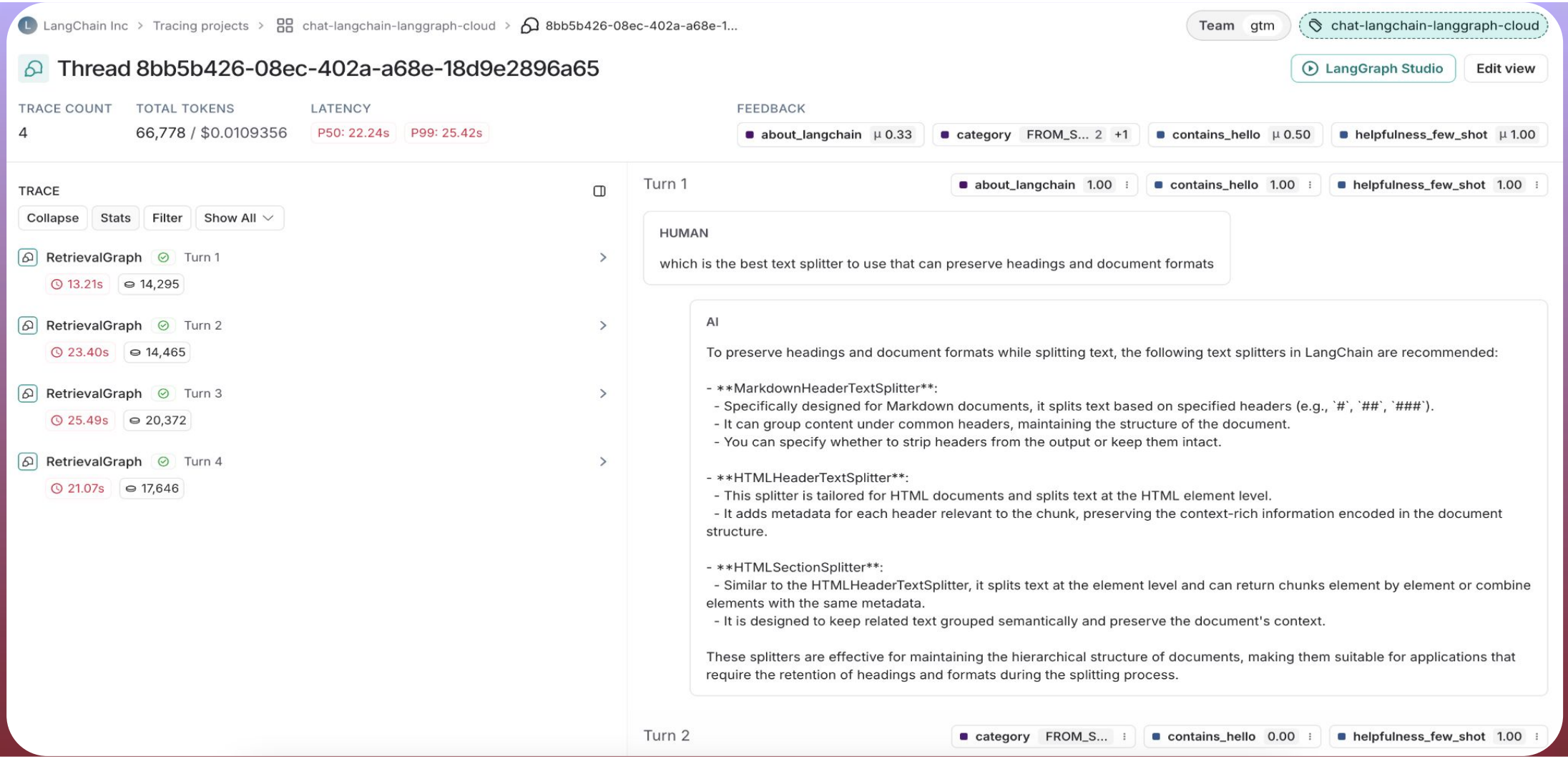

LangChain Academy
Conversational Threads

Types of Runs in LangSmith

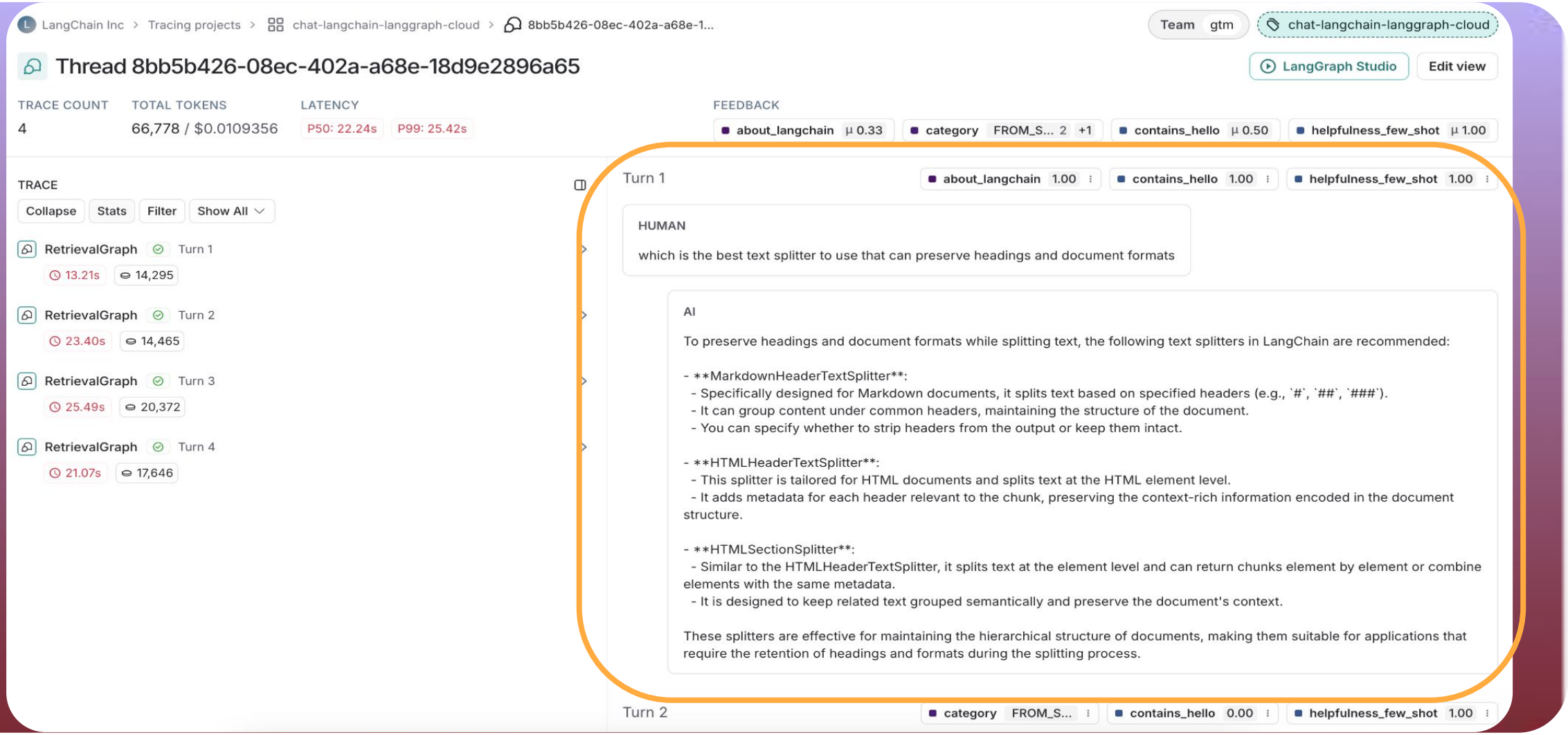
LangSmith supports many different types of Runs - you can specify what type your Run is in the **@traceable** decorator. The types of runs are:

- **LLM:** Invokes an LLM
- **Retriever:** Retrieves documents from databases or other sources
- **Tool:** Executes actions with function calls
- **Chain:** Default type; combines multiple Runs into a larger process

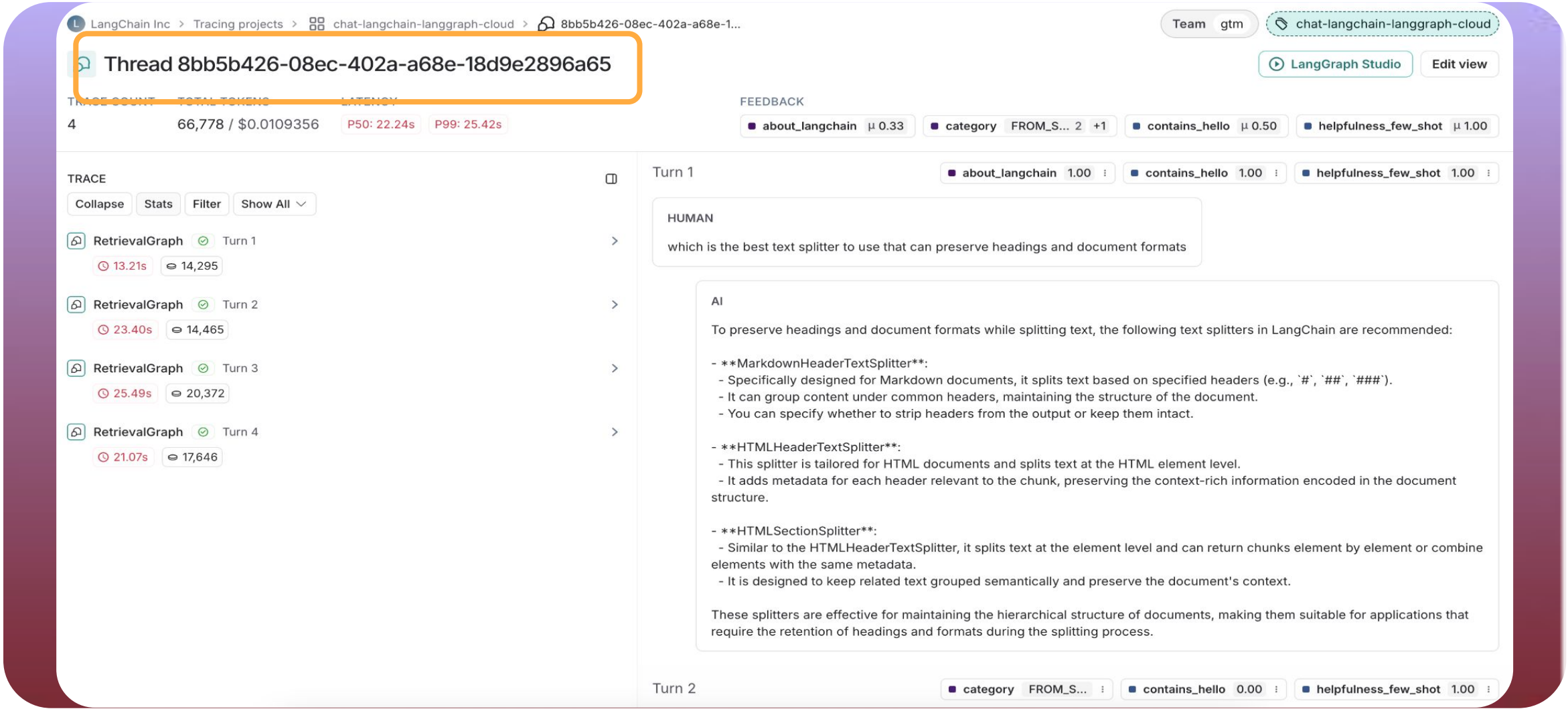
Thread in LangSmith



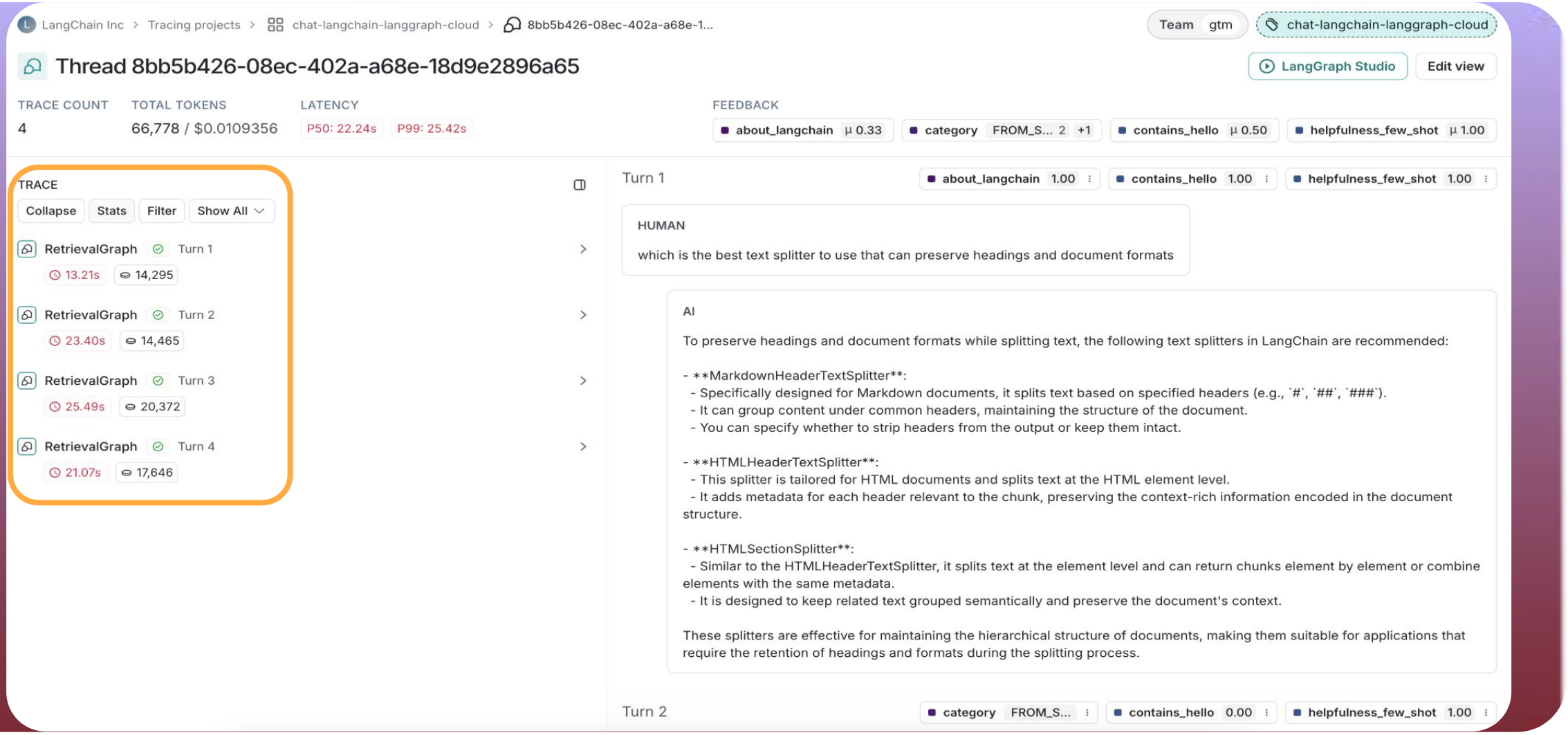
Thread in LangSmith



Thread in LangSmith



Thread in LangSmith

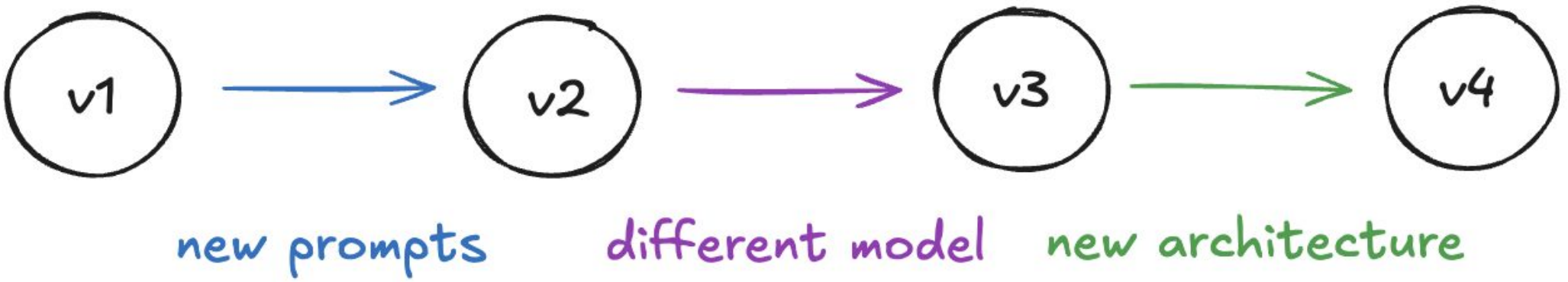


LangChain Academy

Datasets

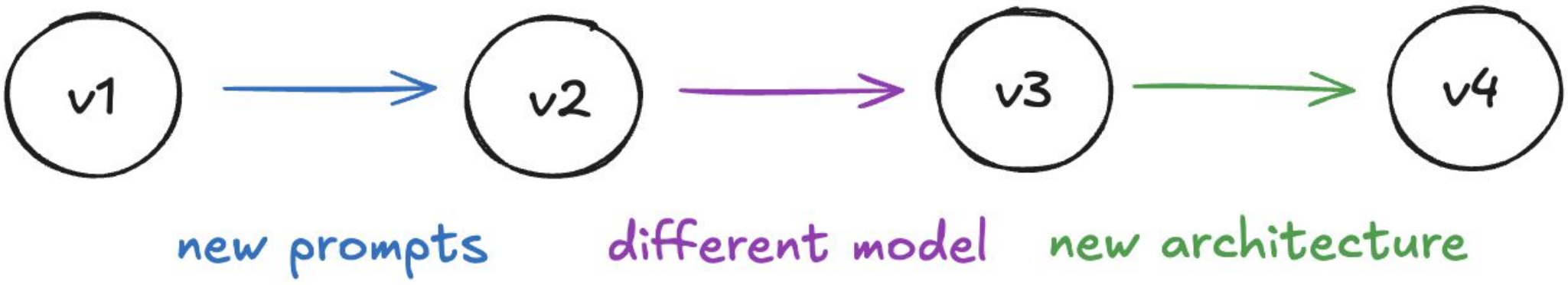
Datasets

How do we know that our application is getting better, not worse, over time?



Datasets

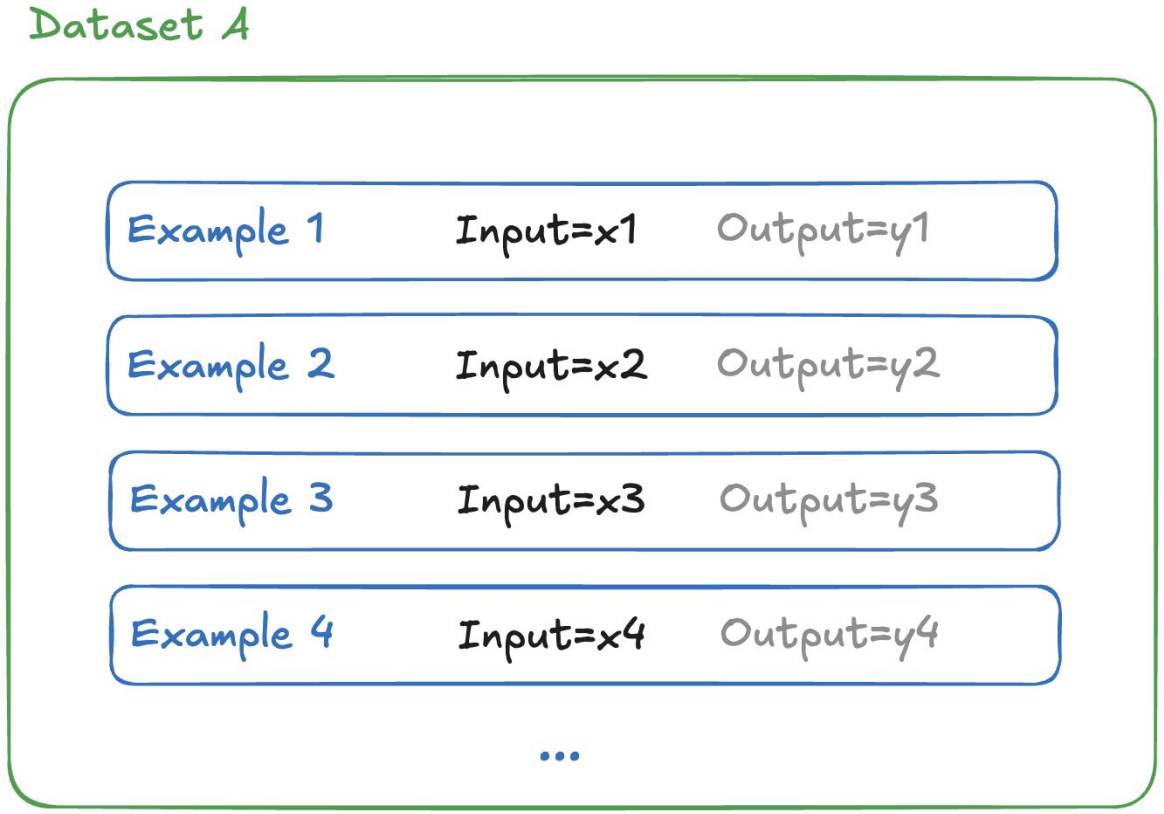
How do we know that our application is getting better, not worse, over time?



We need to do more than run a few “gut check” tests

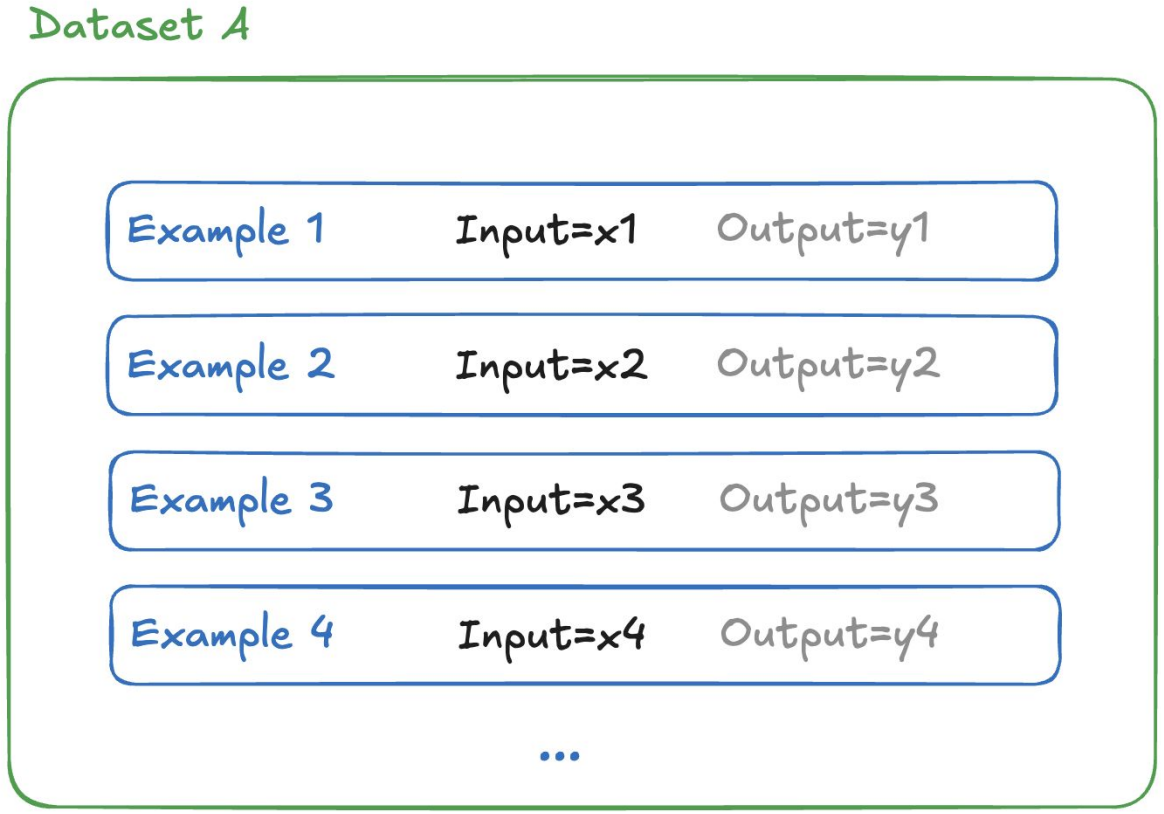
Datasets

In order to test and evaluate our application, we need **datasets**



Datasets

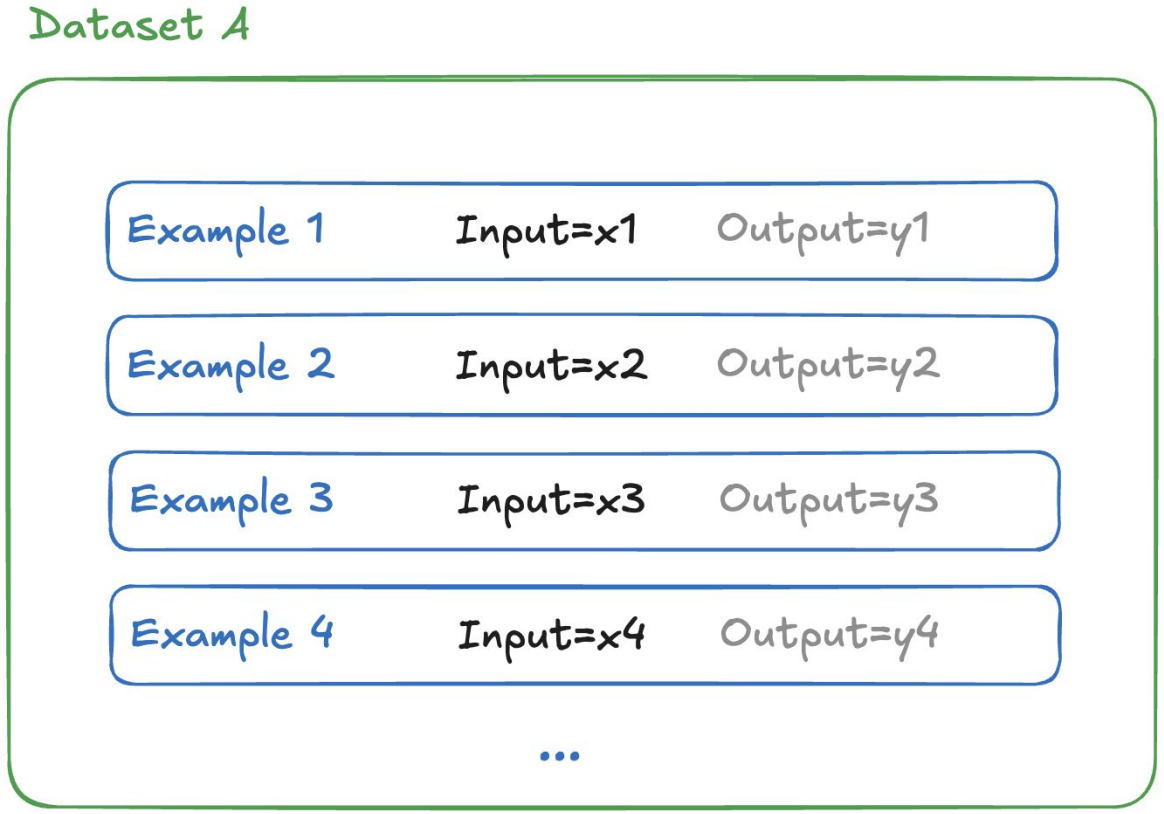
In order to test and evaluate our application, we need **datasets**



Datasets are fundamentally a **list of examples**

Datasets

In order to test and evaluate our application, we need **datasets**



Datasets are fundamentally a **list of examples**

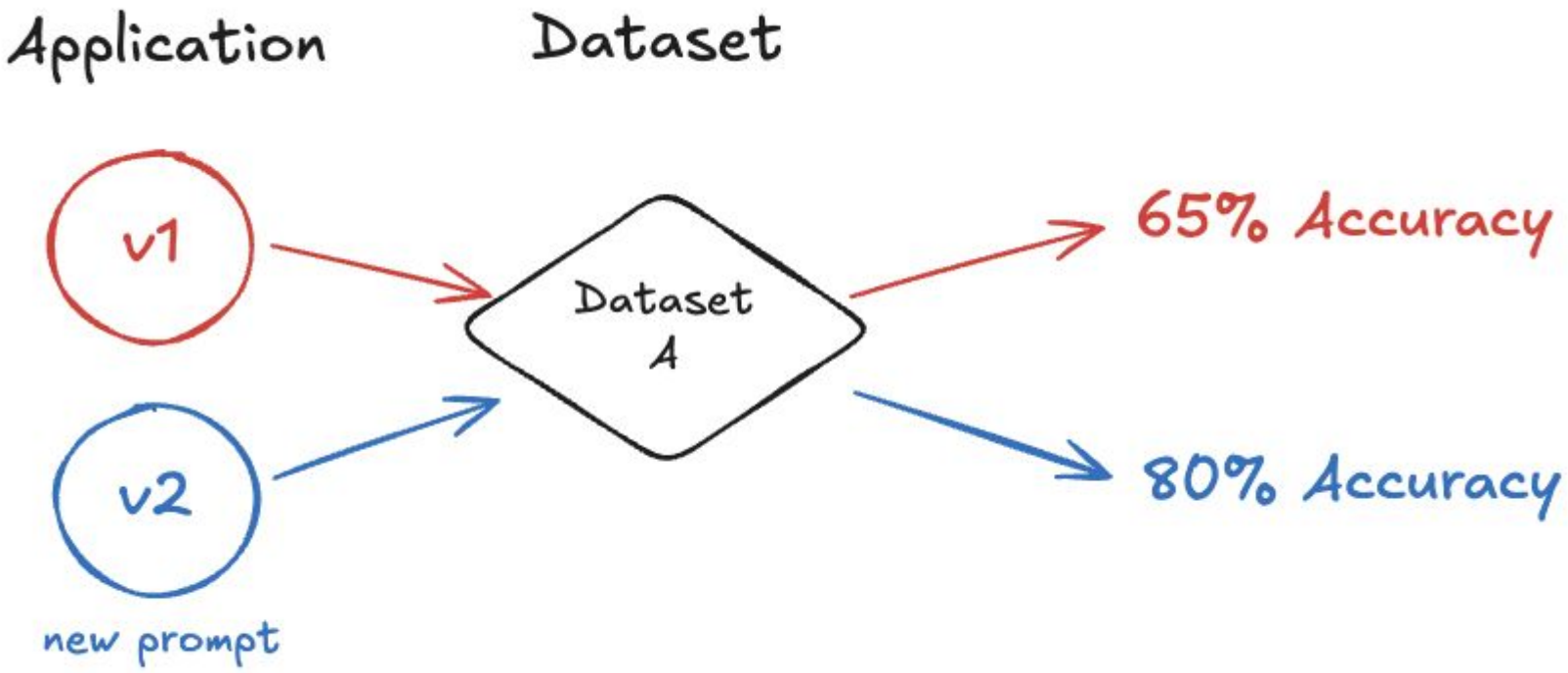
Examples contain an **input**, and an optional **output**

LangChain Academy

Evaluators

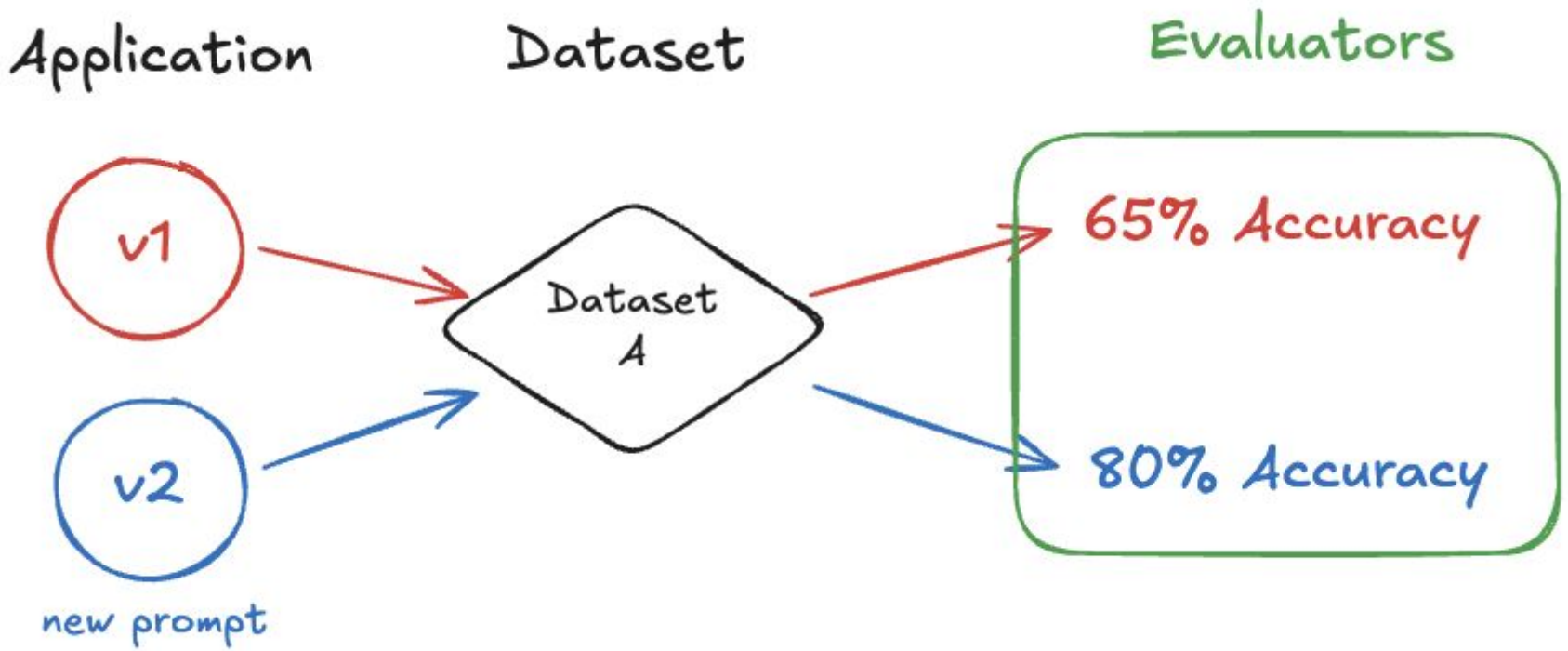
Evaluators

How do we measure performance against our datasets?



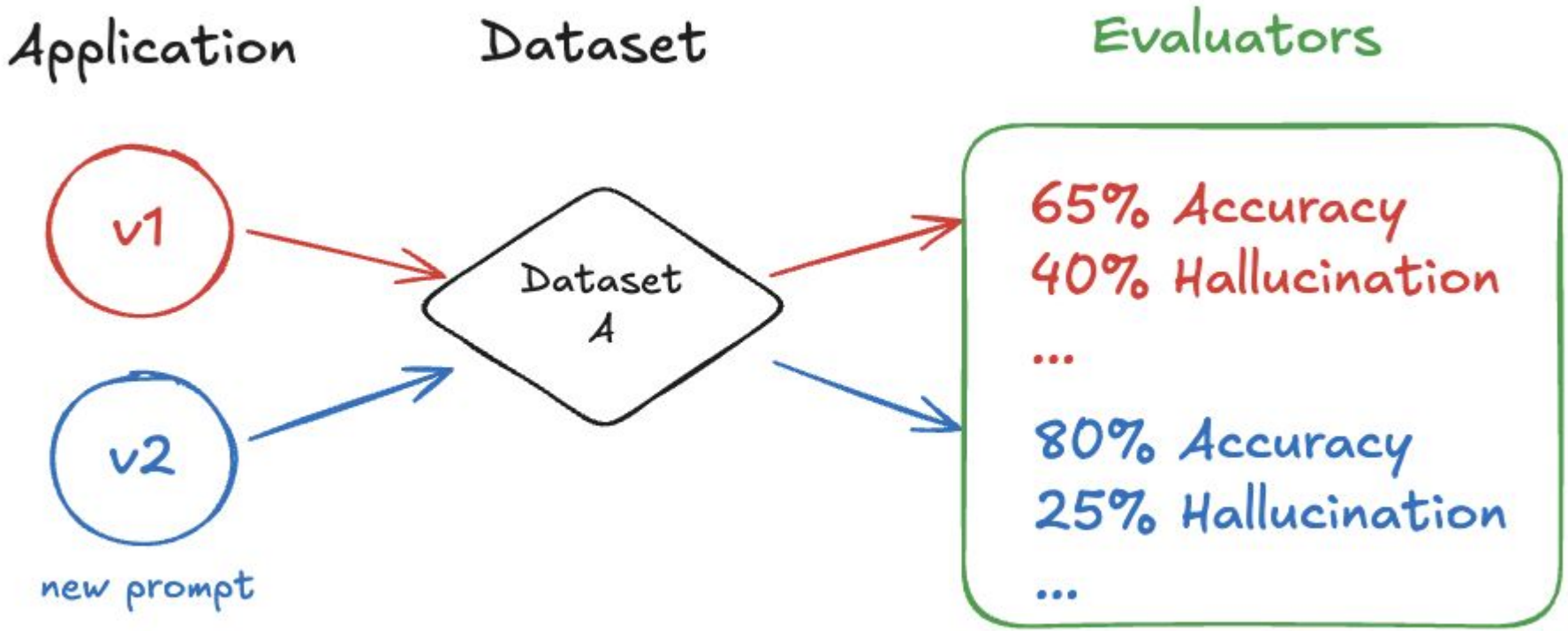
Evaluators

How do we measure performance against our datasets?



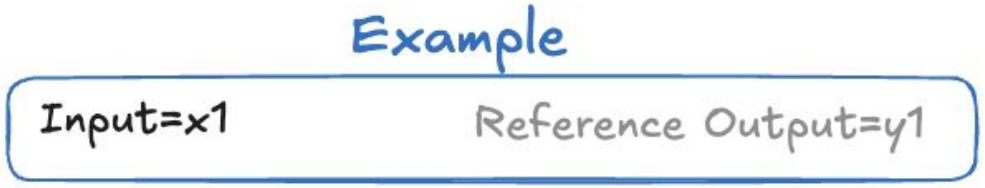
Evaluators

You can define evaluators for multiple metrics (ex. accuracy, hallucination)



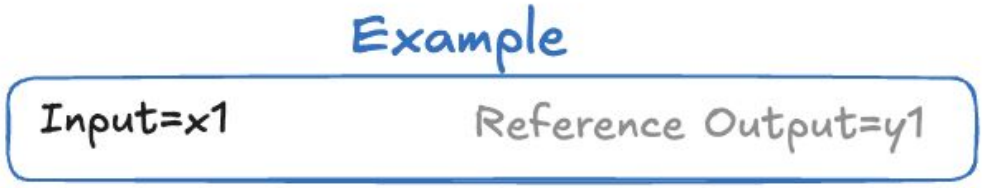
Evaluators

Evaluators compare the dataset **example** against the output **run** from your app



Evaluators

Evaluators compare the dataset **example** against the output **run** from your app



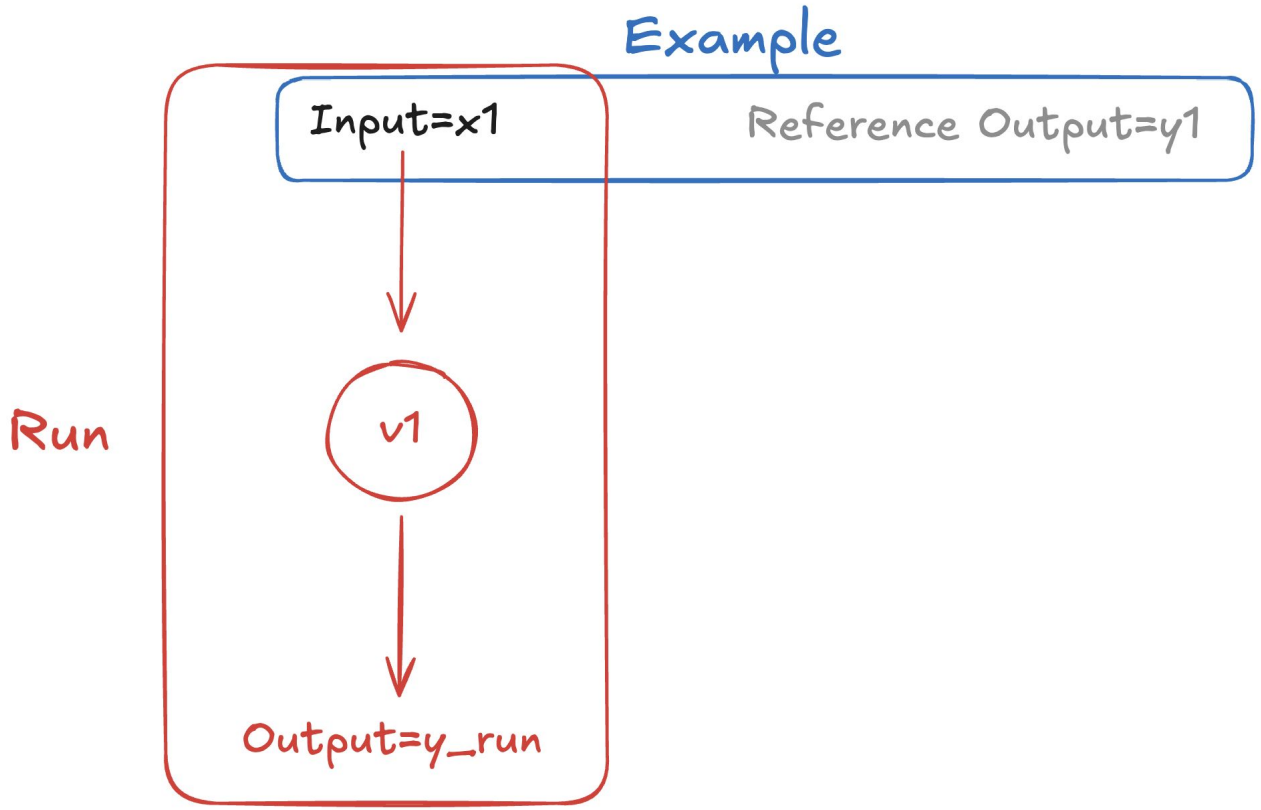
Evaluators

Evaluators compare the dataset **example** against the output **run** from your app

- From a Trace
- Manually Added
- AI Generated
- etc.

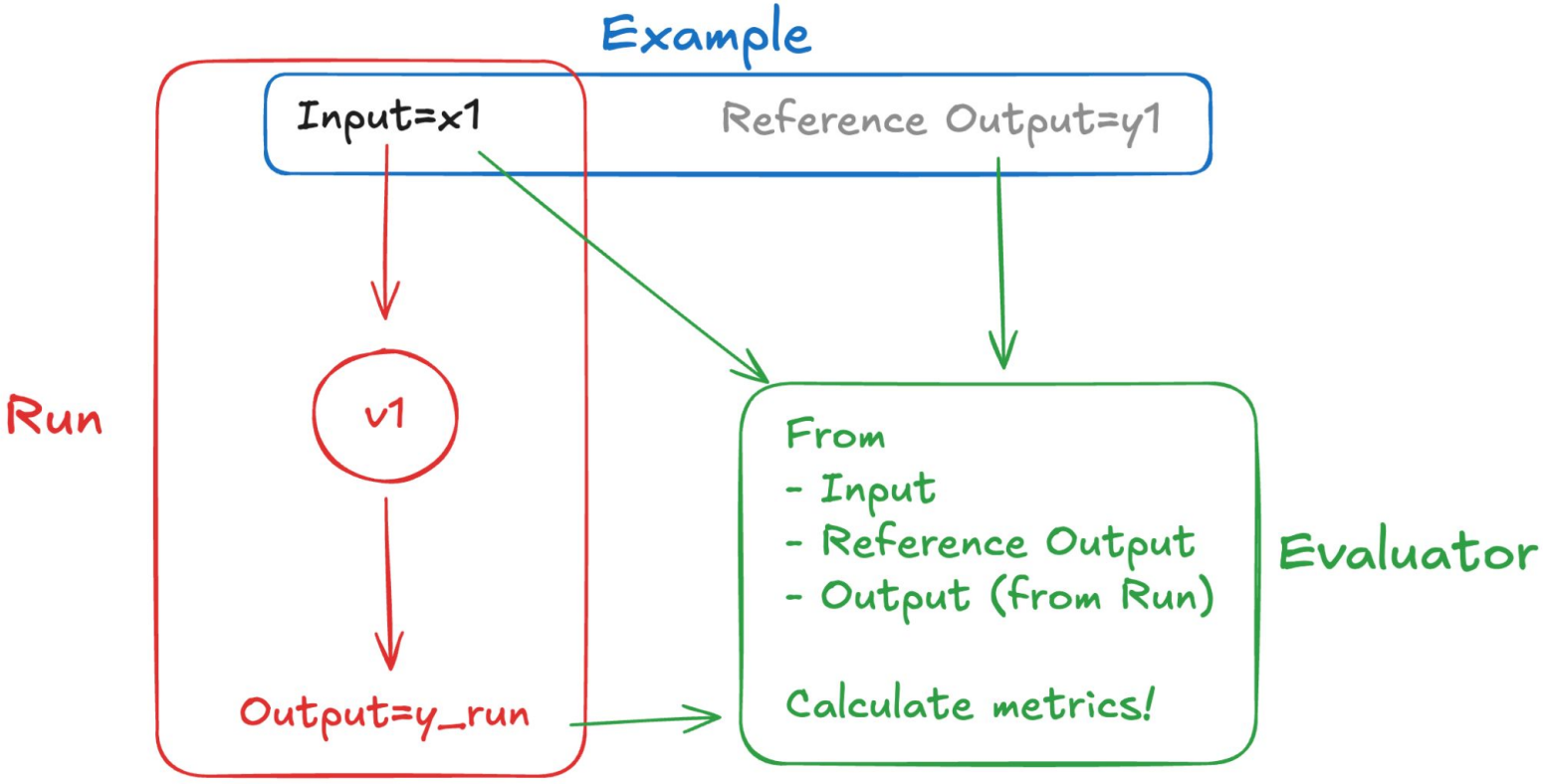
Evaluators

Evaluators compare the dataset **example** against the output **run** from your app



Evaluators

Evaluators compare the dataset **example** against the output **run** from your app



Takeaways

- Evaluators calculate metrics based on a Run and an Example

Takeaways

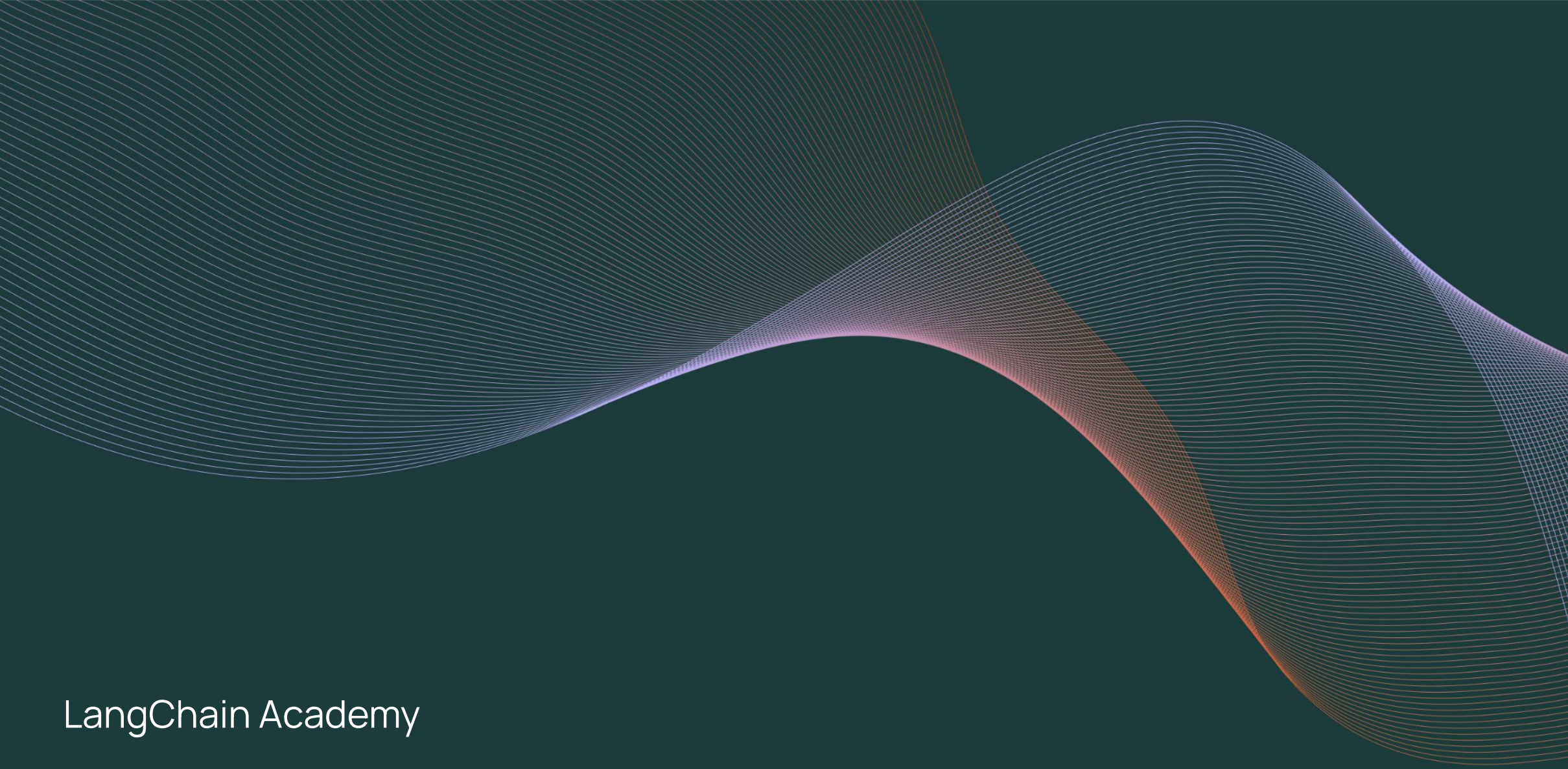
- Evaluators calculate metrics based on a Run and an Example
  - Specifically, from the Input, Reference output, and Run Output

Takeaways

- Evaluators calculate metrics based on a Run and an Example
  - Specifically, from the Input, Reference output, and Run Output
- You can define Evaluators directly in your local code

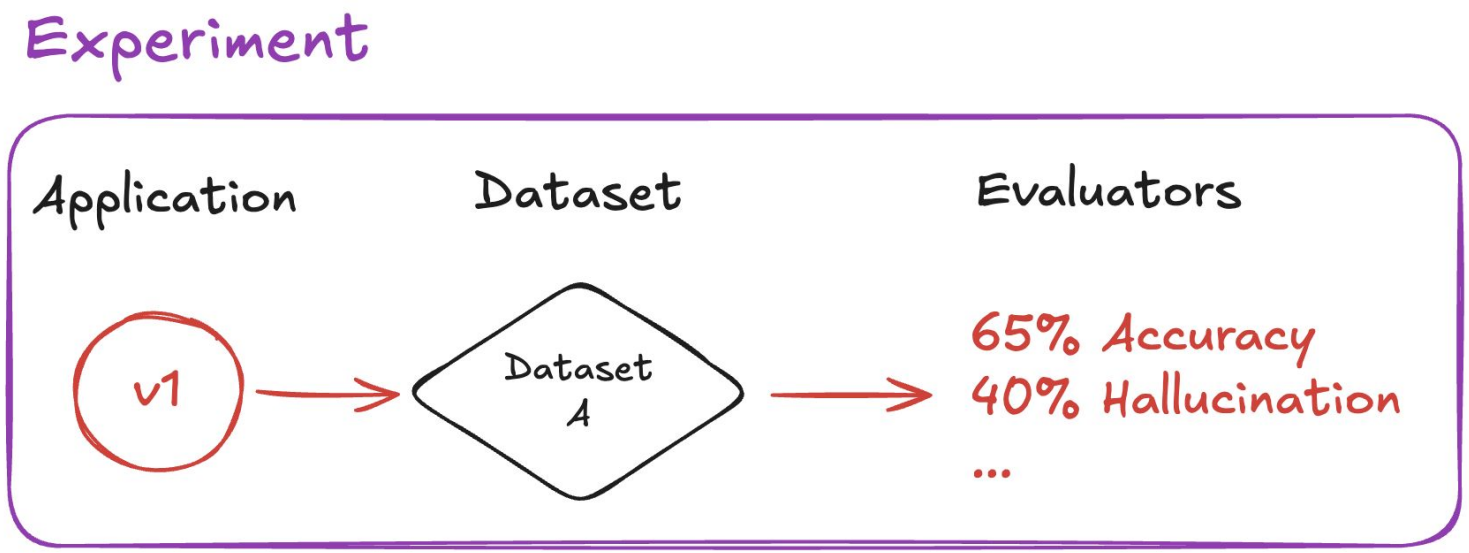
Takeaways

- Evaluators calculate metrics based on a Run and an Example
  - Specifically, from the Input, Reference output, and Run Output
- You can define Evaluators directly in your local code
- You can ALSO define Evaluators in the LangSmith UI
  - LLM-as-judge evaluators
  - Custom code evaluators


LangChain Academy
Experiments

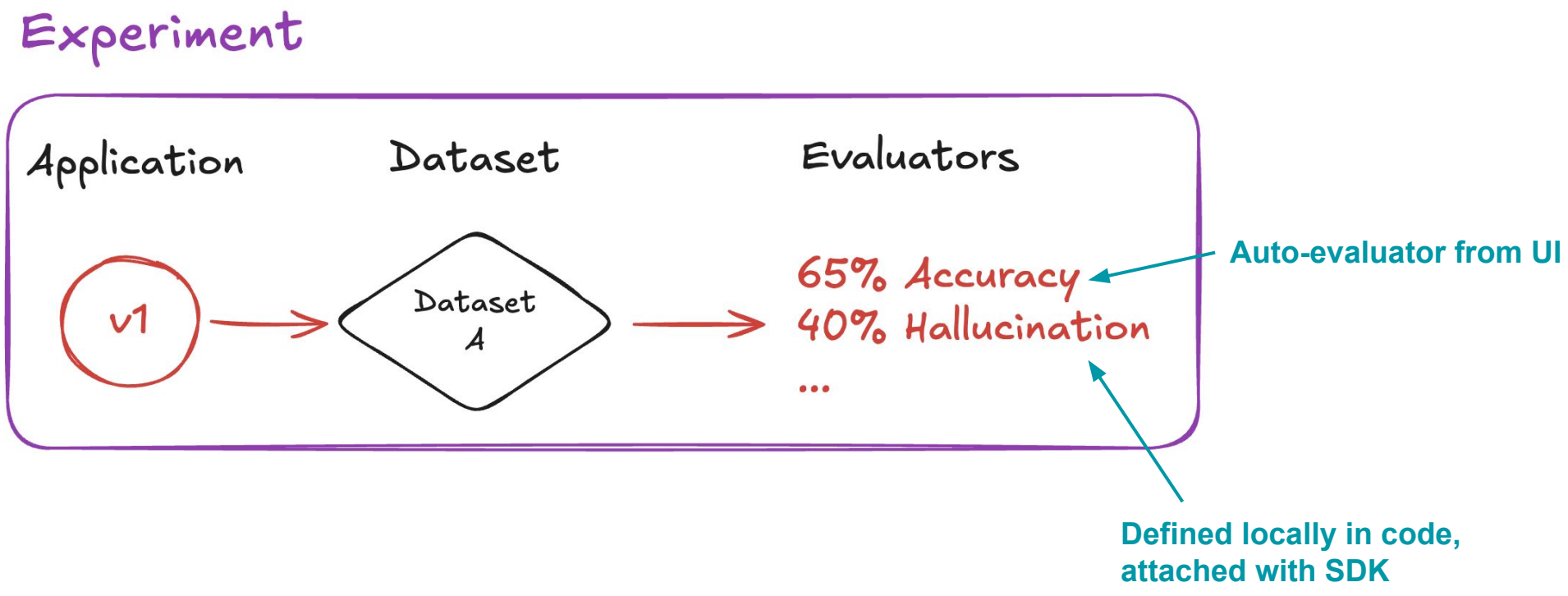
Experiments

Experiment: Running your **application** over a **dataset**, and **evaluating** performance



Experiments

You can attach **evaluators** to your **experiment** in the UI, or locally with the SDK



Takeaways

- Experiment: Running your **application** over a **dataset**, and **evaluating** its performance

Takeaways

- Experiment: Running your **application** over a **dataset**, and **evaluating** its performance
- Experiments can be run using the LangSmith SDK with **evaluate()**

Takeaways

- Experiment: Running your **application** over a **dataset**, and **evaluating** its performance
- Experiments can be run using the LangSmith SDK with **evaluate()**
- Experiments can be run over an entire dataset OR
  - A specific version
  - A split
  - Specific examples

Takeaways

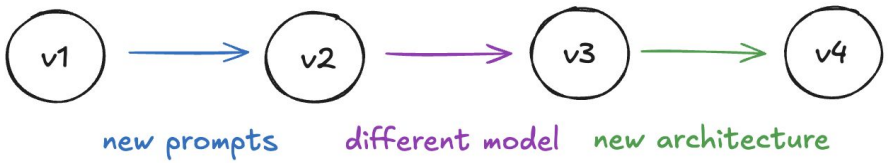
- Experiment: Running your **application** over a **dataset**, and **evaluating** its performance
- Experiments can be run using the LangSmith SDK with **evaluate()**
- Experiments can be run over an entire dataset OR
  - A specific version
  - A split
  - Specific examples
- Experiments can be run with other parameters
  - Repetitions
  - Concurrent Threads
  - Metadata

LangChain Academy

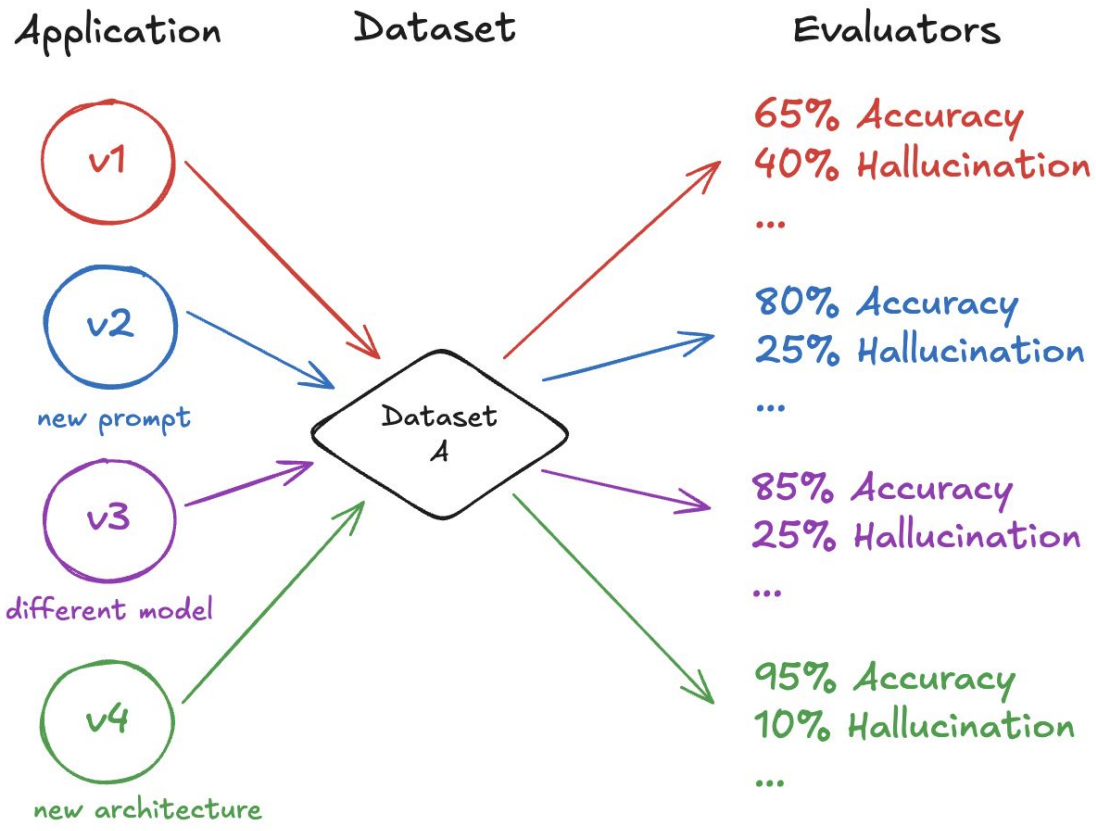
Analyzing Experiment Results

Analyzing Experiment Results

Without Testing



With Testing and Evaluation



Takeaways

- Experiments are useful for seeing trends in your application performance as you improve it over time

Takeaways

- Experiments are useful for seeing trends in your application performance as you improve it over time
- You can deep dive into a single experiment and look into how each individual run performed on the dataset example

Takeaways

- Experiments are useful for seeing trends in your application performance as you improve it over time
- You can deep dive into a single experiment and look into how each individual run performed on the dataset example
- You can compare multiple experiments side-by-side and see how they scored on your evaluator metrics

Takeaways

- Experiments are useful for seeing trends in your application performance as you improve it over time
- You can deep dive into a single experiment and look into how each individual run performed on the dataset example
- You can compare multiple experiments side-by-side and see how they scored on your evaluator metrics
- Experiments give you hard empirical data to push changes to production with confidence

LangChain Academy

Playground

Hard-coded Prompts → Prompt Templates

Hard-coded Prompt

You are a chat model who only speaks Spanish.

Prompt Template

You are a chat model who only speaks {language}

language = "English"

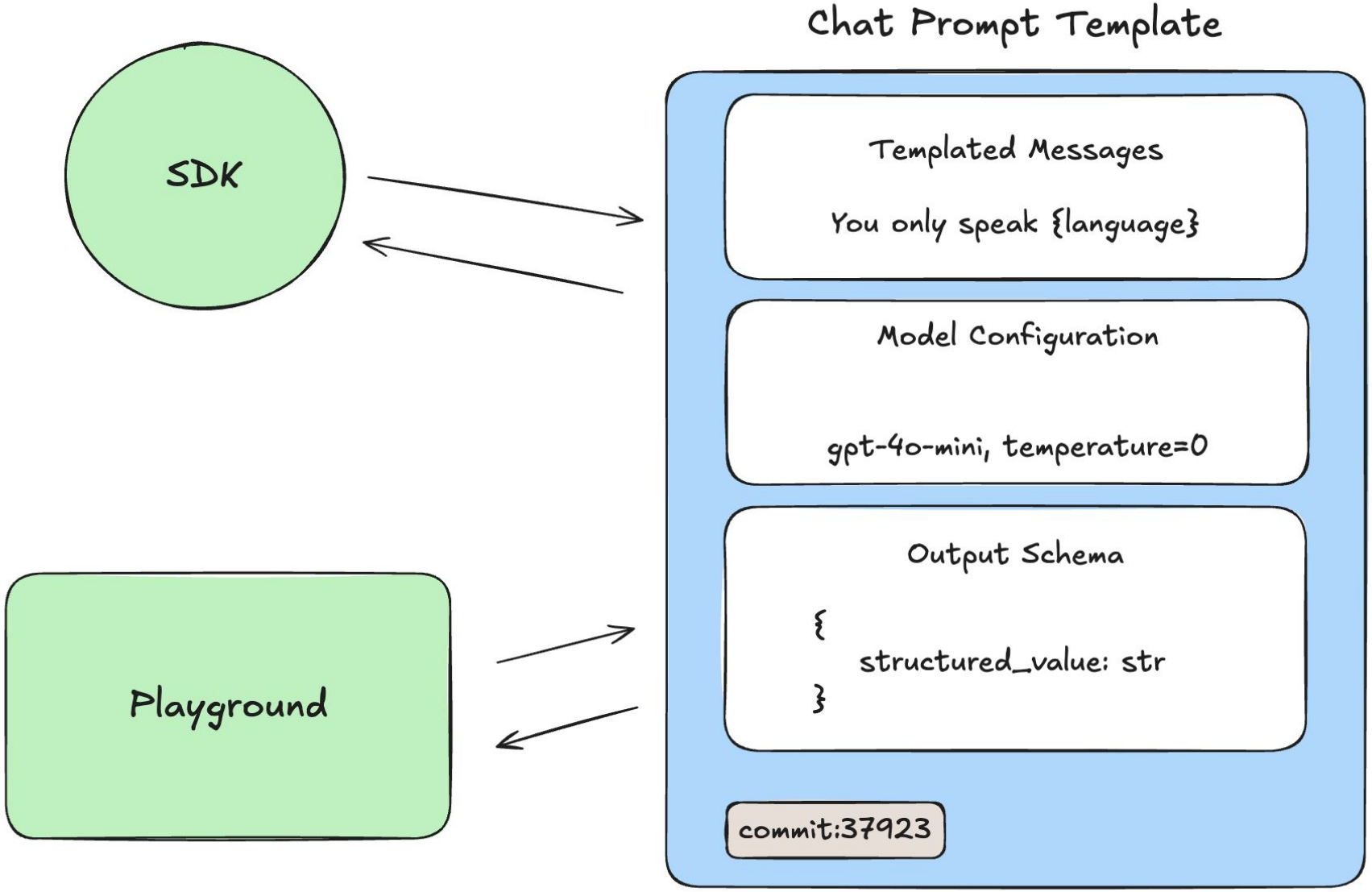
language = "Chinese"

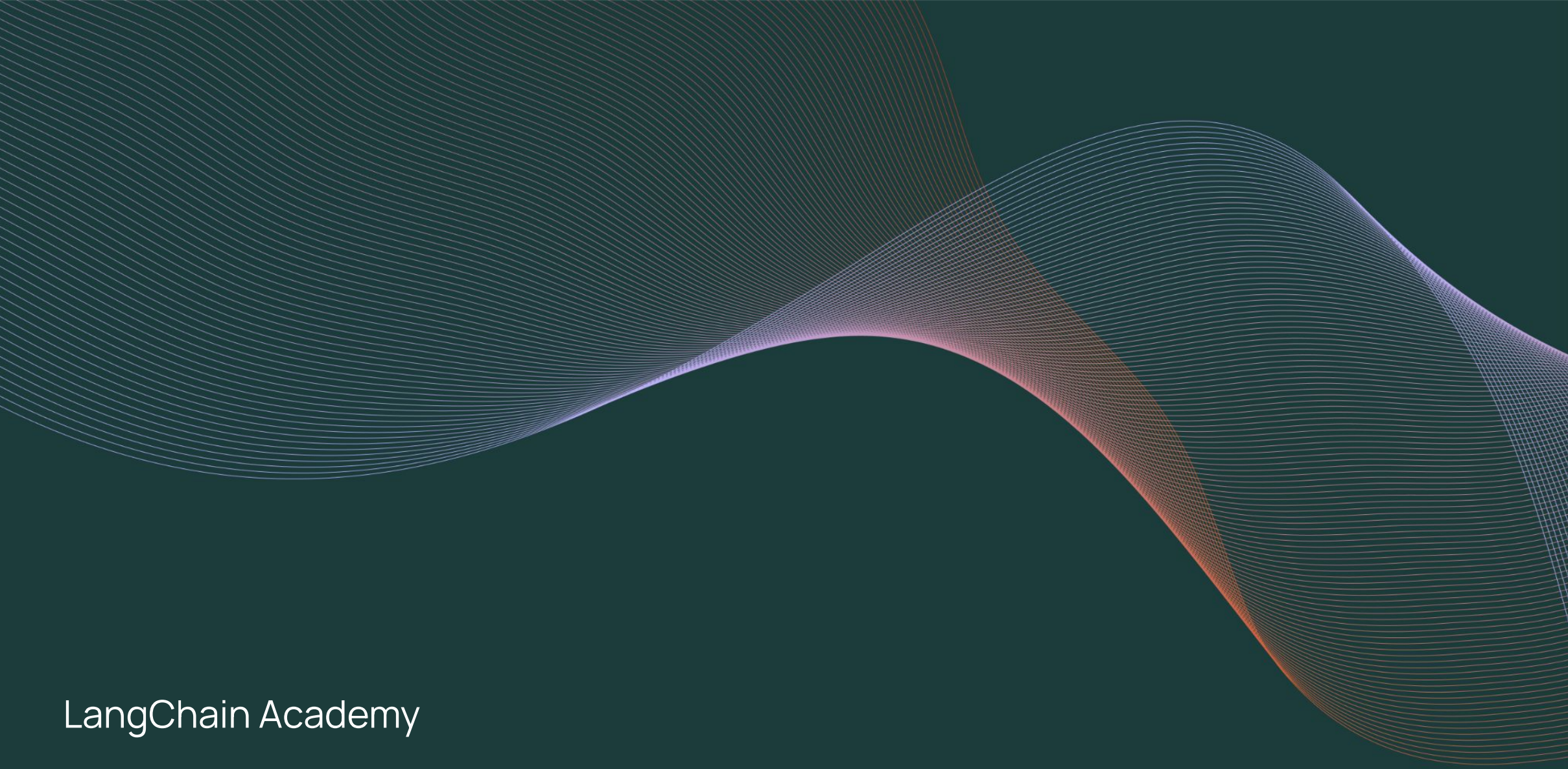
language = "Spanish"

LangChain Academy

Prompt Hub

Prompt Hub: Prompt Templates





LangChain Academy

Playground and Prompt Hub in Practice

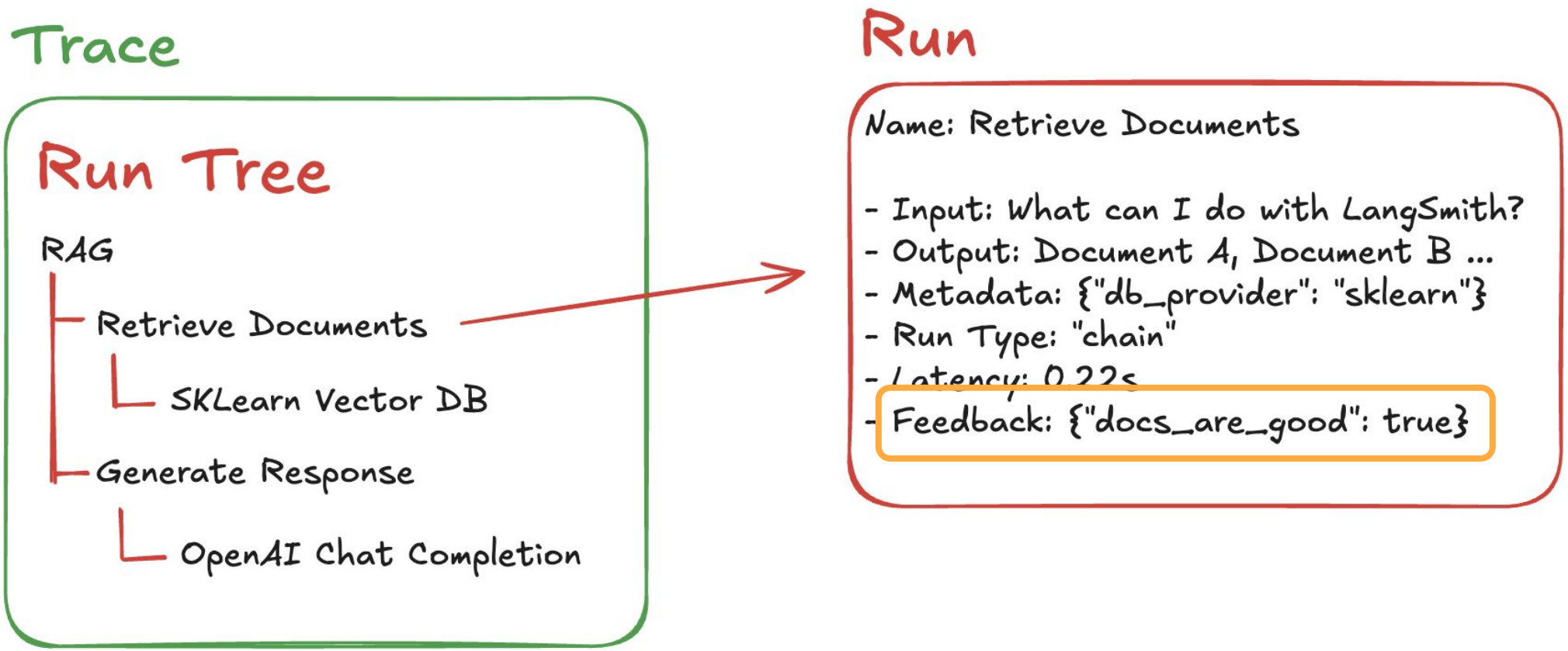
LangChain Academy

User Feedback

User Feedback

It’s very important to monitor how your users feel about your application’s performance

User Feedback



Types of Feedback

Categorical

- is_correct (yes, no)
- type_of_input (integer, number, string)
- is_helpful (yes, no)
- did_succeed (succeeded, failed)

Continuous

- correctness 7 (1 - 10)
- similarity 3 (1 - 5)
- helpfulness 3 (1 - 10)

Types of Feedback

Categorical

- is_correct (yes, no)
- type_of_input (integer, number, string)
- is_helpful (yes, no)
- did_succeed (succeeded, failed)

ex.

```
{
    "key": "is_correct"
    "value": "yes"
}
```

Continuous

- correctness 7 (1 - 10)
- similarity 3 (1 - 5)
- helpfulness 3 (1 - 10)

Types of Feedback

Categorical

- is_correct (yes, no)
- type_of_input (integer, number, string)
- is_helpful (yes, no)
- did_succeed (succeeded, failed)

ex.

```
{
    "key": "is_correct"
    "value": "yes"
}
```

Continuous

- correctness 7 (1 - 10)
- similarity 3 (1 - 5)
- helpfulness 3 (1 - 10)

ex.

```
{
    "key": "correctness"
    "score": 7
}
```

Feedback Tags

Feedback Tags

Create and organize feedback tags for this tenant

New tag

3184 feedback tags

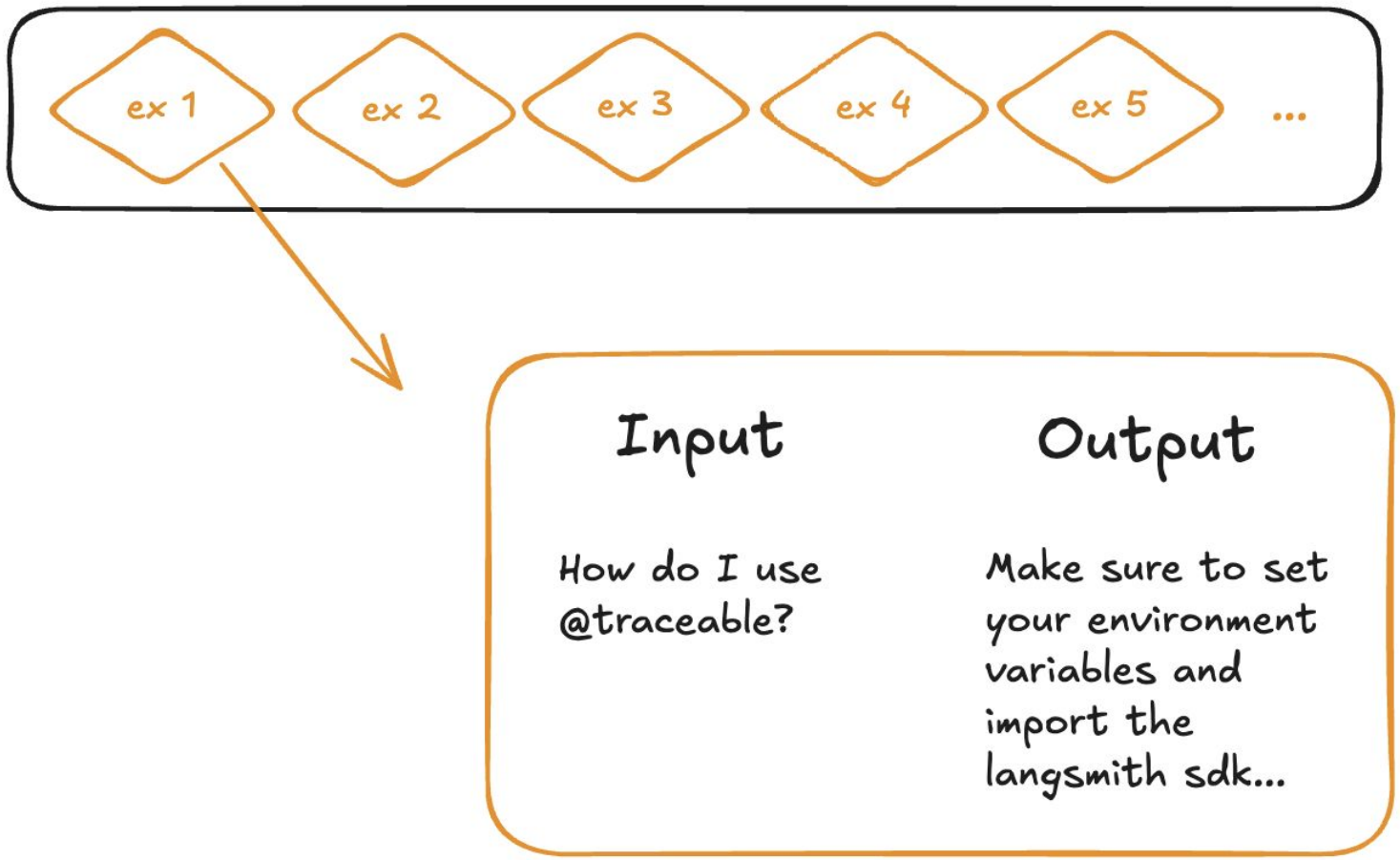
Feedback	Type	Values
faithfulness_score	Continuous	
answer_relevancy_score	Continuous	
context_ relevancy_score	Continuous	
test_entity_examples	Continuous	
test_employer_org_bias	Continuous	
test_person_profile_bias	Continuous	

LangChain Academy

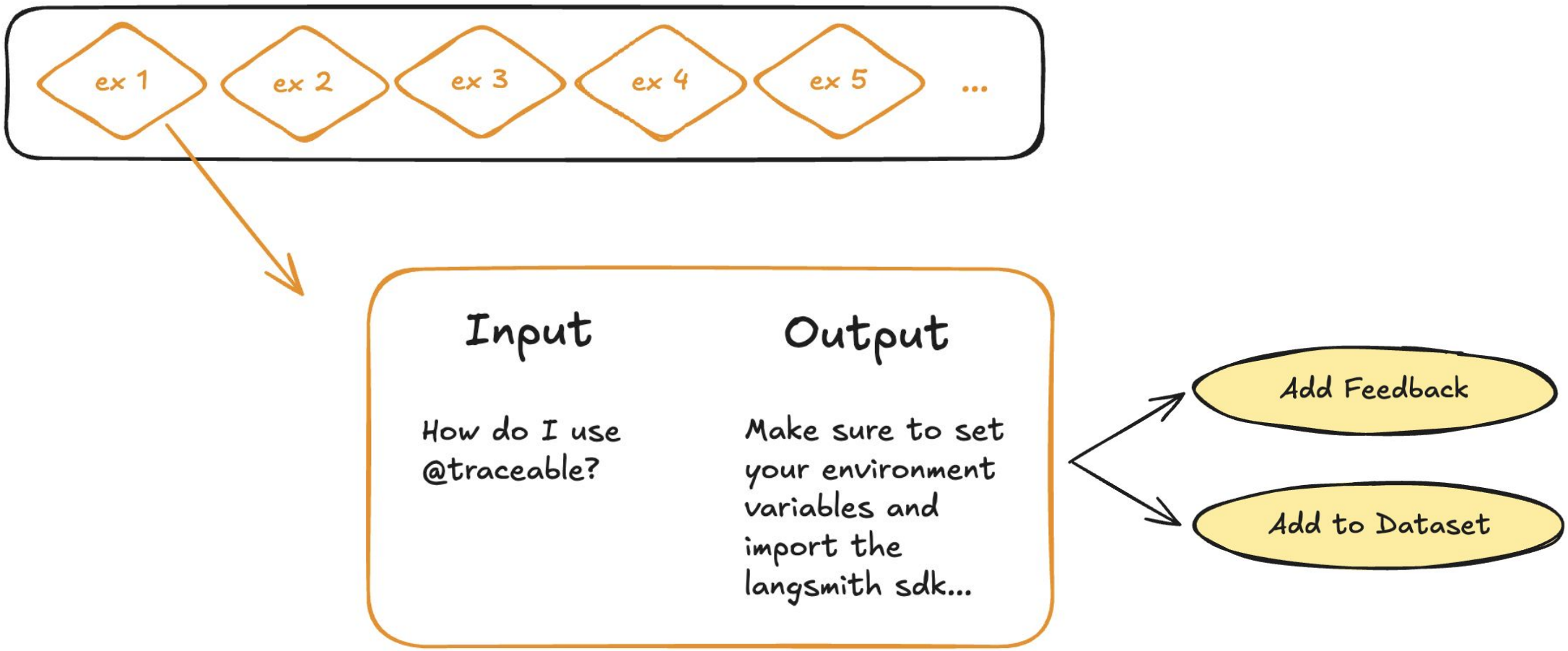
Annotation Queues

Annotation Queues

Annotation Queues are a user friendly way to quickly cycle through and annotate data



Annotation Queues



LangChain Academy

Filtering

Common Filters to Create

Here are some common filters that we see!

Common Filters to Create

Here are some common filters that we see!

- **Status = “error”** (all runs with errors)

Common Filters to Create

Here are some common filters that we see!

- **Status = “error”** (all runs with errors)
- **Feedback score < X** (received bad feedback)

Common Filters to Create

Here are some common filters that we see!

- **Status = “error”** (all runs with errors)
- **Feedback score < X** (received bad feedback)
- **run_type = “LLM” AND latency > X** (slow LLM runs)

Common Filters to Create

Here are some common filters that we see!

- **Status = “error”** (all runs with errors)
- **Feedback score < X** (received bad feedback)
- **run_type = “LLM” AND latency > X** (slow LLM runs)
- **Metadata[“ls_model_name”] = X** (create a filter for a metadata field)

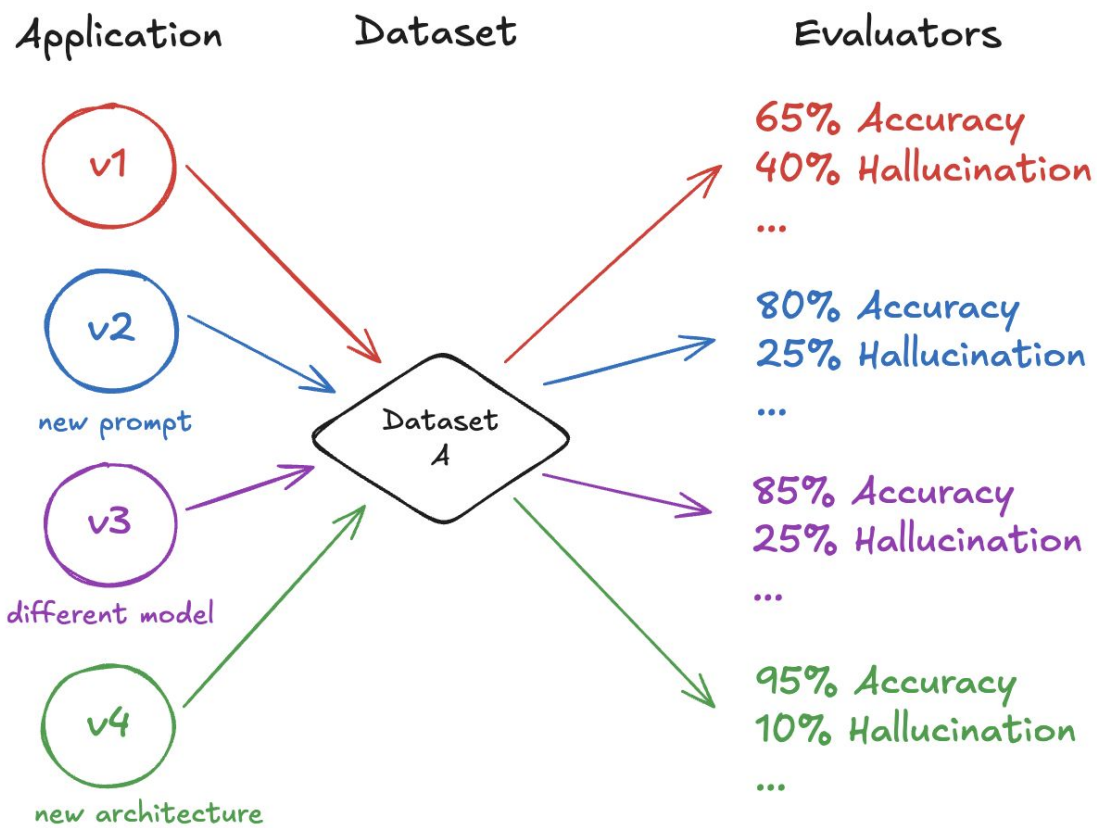
LangChain Academy

Online Evaluation

Offline and Online Evaluation

Recap: Offline Evaluation involves testing different versions of your Application against a dataset

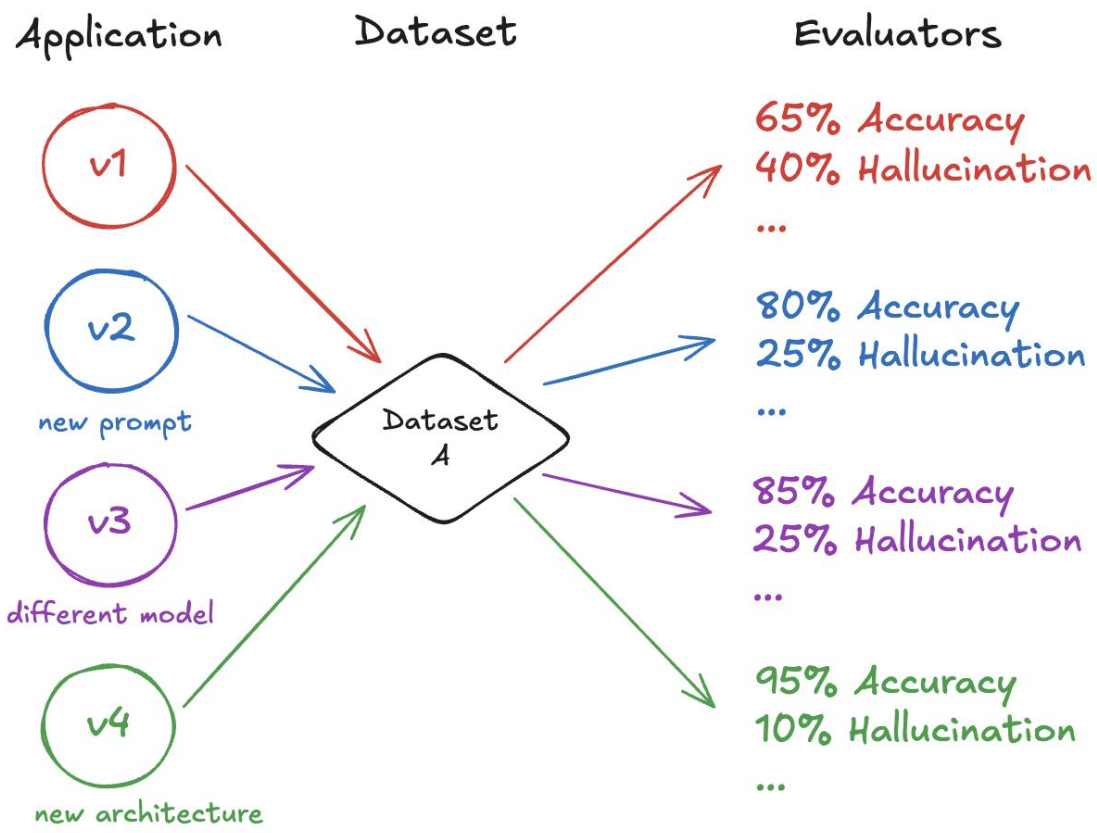
Offline Evaluation



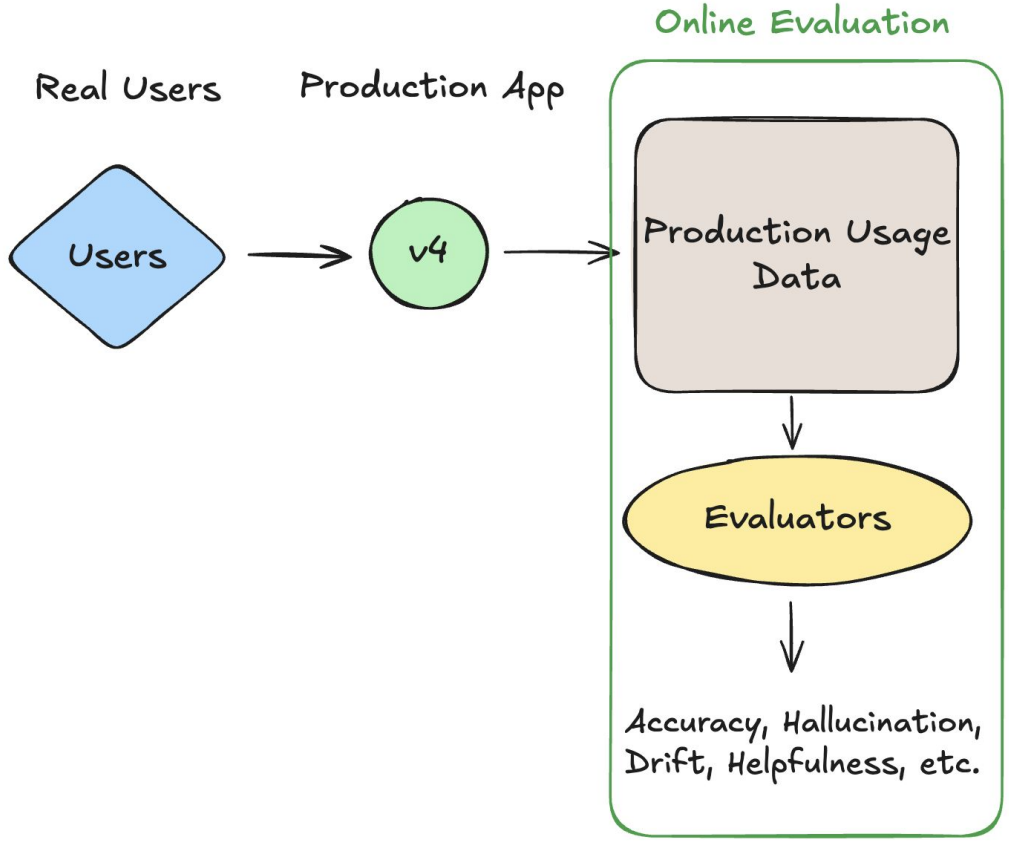
Offline and Online Evaluation

Online Evaluation involves evaluating real production performance with end users

Offline Evaluation



Online Evaluation

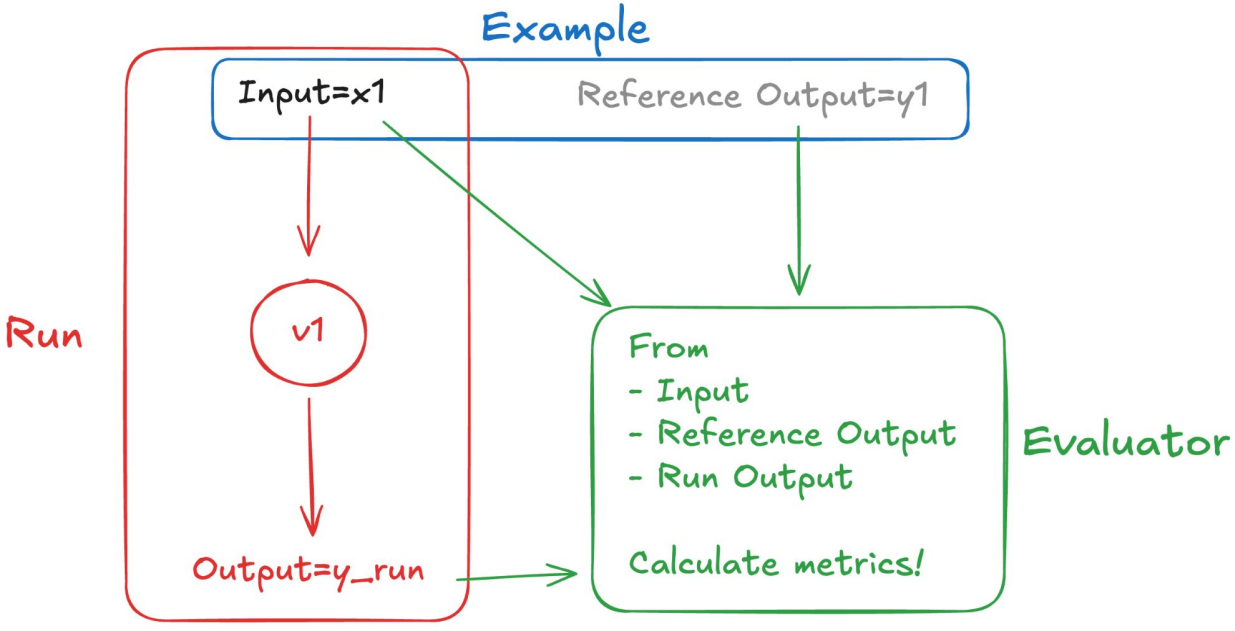


Offline and Online Evaluation

Offline Evaluators can make use Reference Output

Offline Evaluator

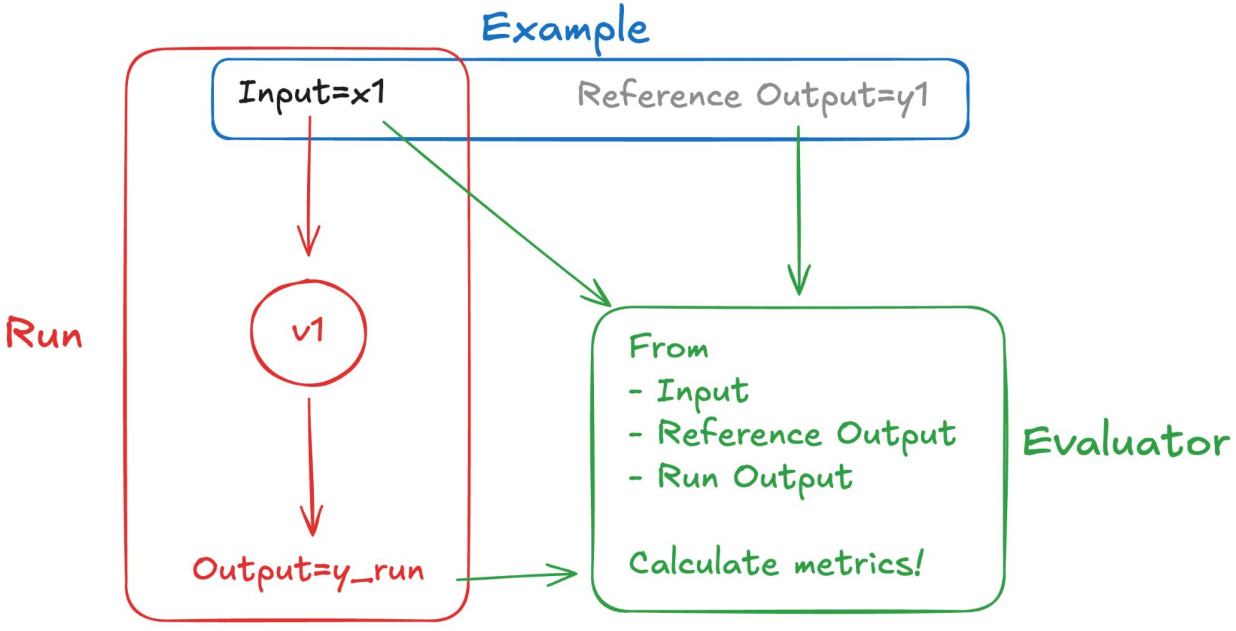
Online Evaluator



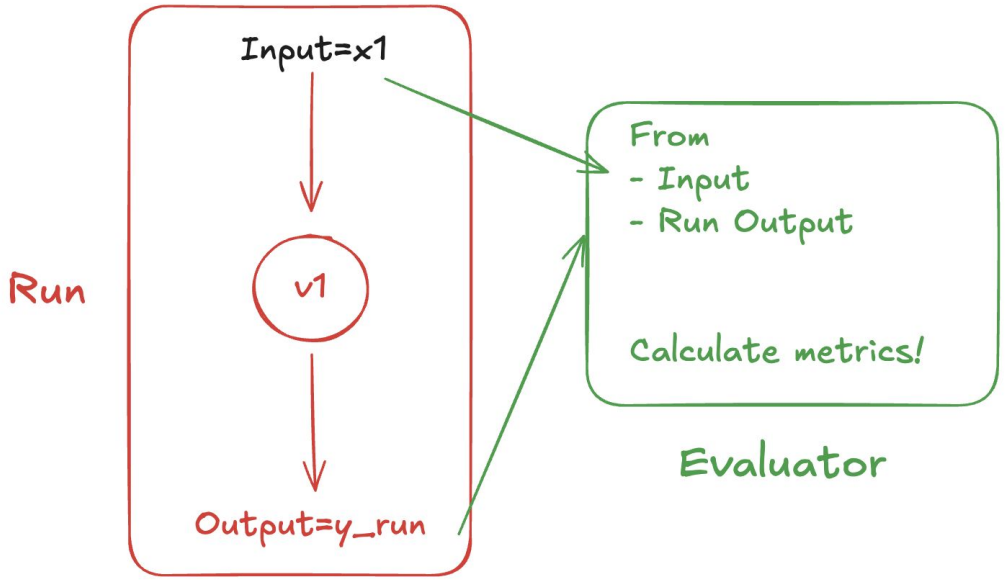
Offline and Online Evaluation

Online Evaluators only have access to the Input, and the output of the production Run

Offline Evaluator



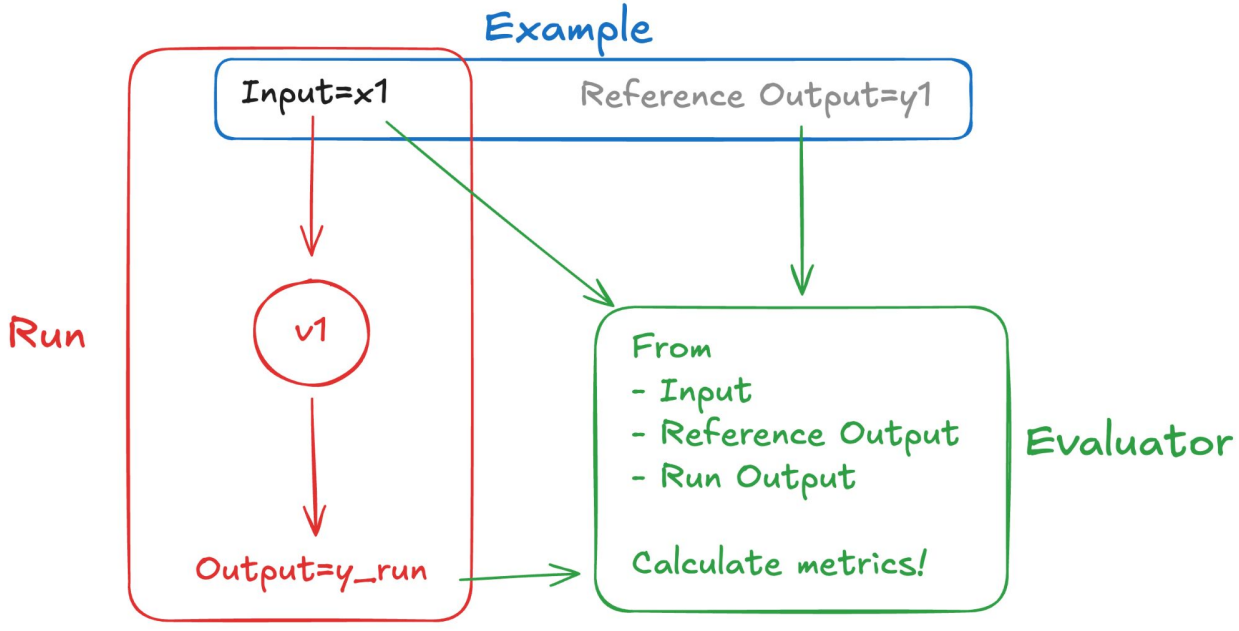
Online Evaluator



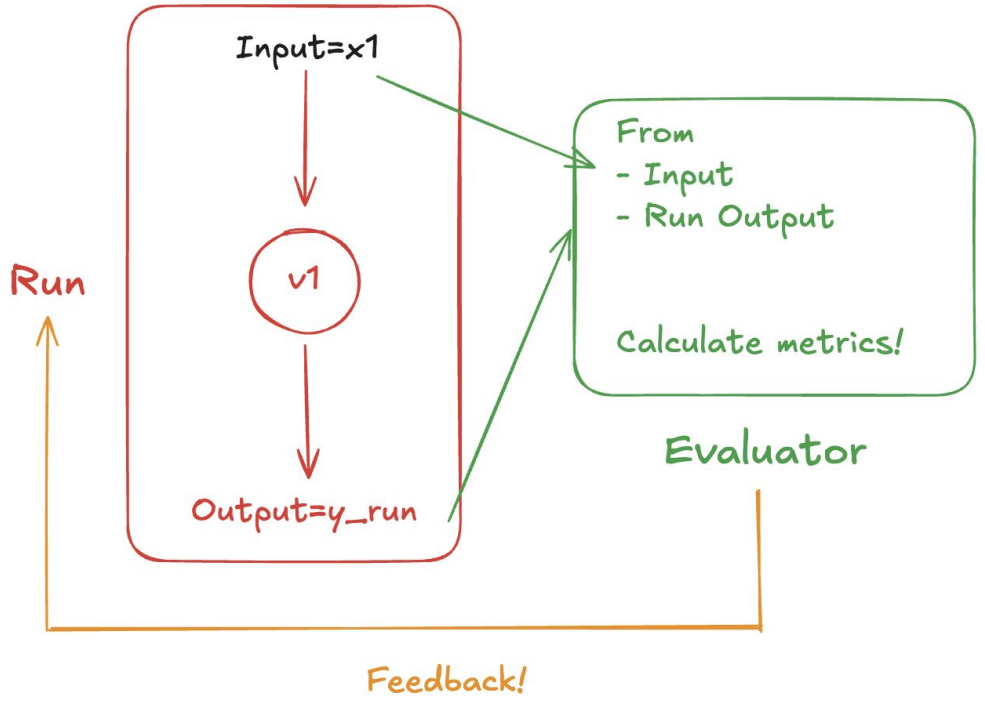
Offline and Online Evaluation

Online Evaluators only have access to the Input, and the output of the production Run

Offline Evaluator



Online Evaluator



LangChain Academy

Online Evaluation

Popular Examples of Online Evaluation

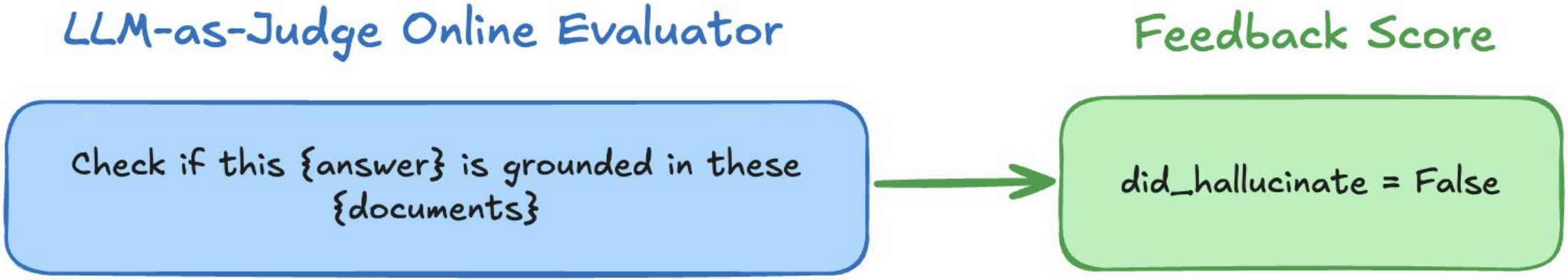
For document retrieval tasks, check for hallucinations in the output

Popular Examples of Online Evaluation

For document retrieval tasks, check for hallucinations in the output

Popular Examples of Online Evaluation

For document retrieval tasks, check for hallucinations in the output

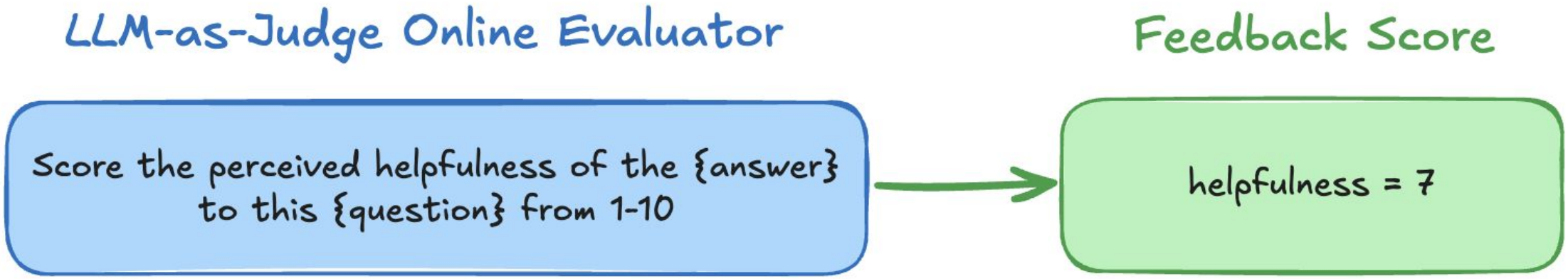


Popular Examples of Online Evaluation

Score the perceived helpfulness of an answer to a user

Popular Examples of Online Evaluation

Score the perceived helpfulness of an answer to a user



Popular Examples of Online Evaluation

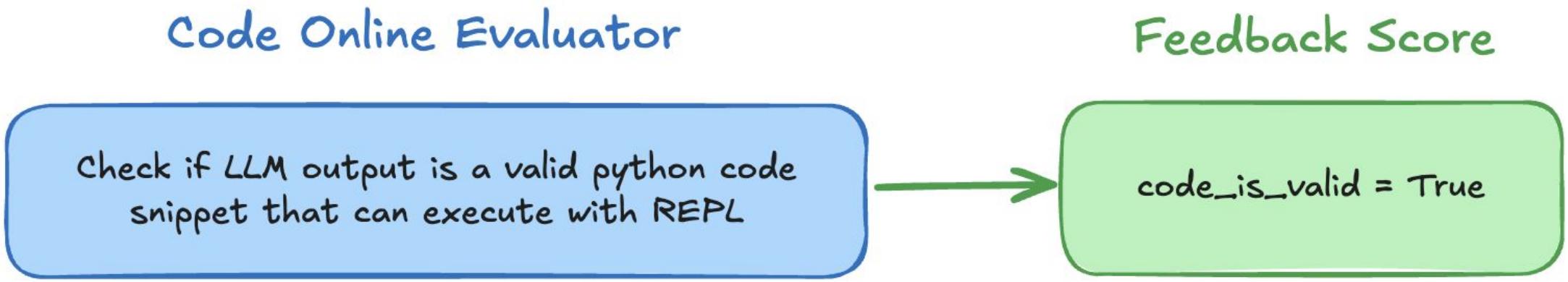
For a Coding Assistant, check to see if the outputted code actually compiles and executes

Code Online Evaluator

Check if LLM output is a valid python code snippet that can execute with REPL

Popular Examples of Online Evaluation

For a Coding Assistant, check to see if the outputted code actually compiles and executes



Popular Examples of Online Evaluation

Check that the output matches a certain structure
ex. Email Assistant should sign emails a certain way

Code Online Evaluator

For an Email Assistant, regex match that we correctly signed the email "Best wishes, Nick"

Popular Examples of Online Evaluation

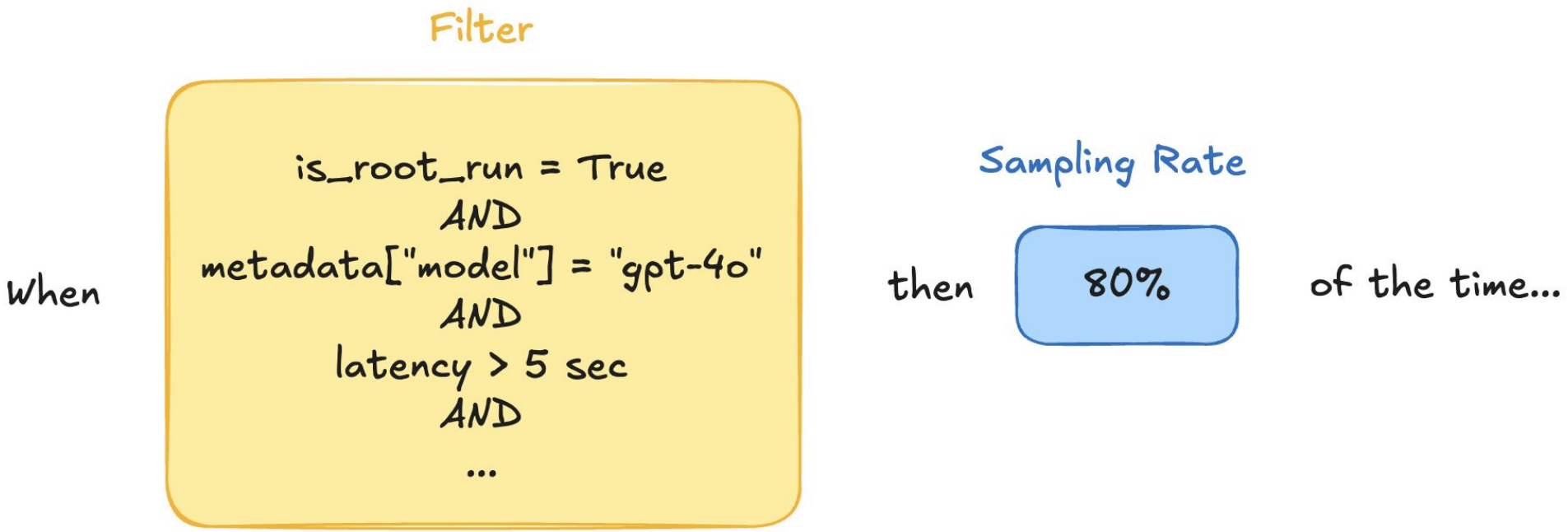
Check that the output matches a certain structure
ex. Email Assistant should sign emails a certain way

LangChain Academy

Automations

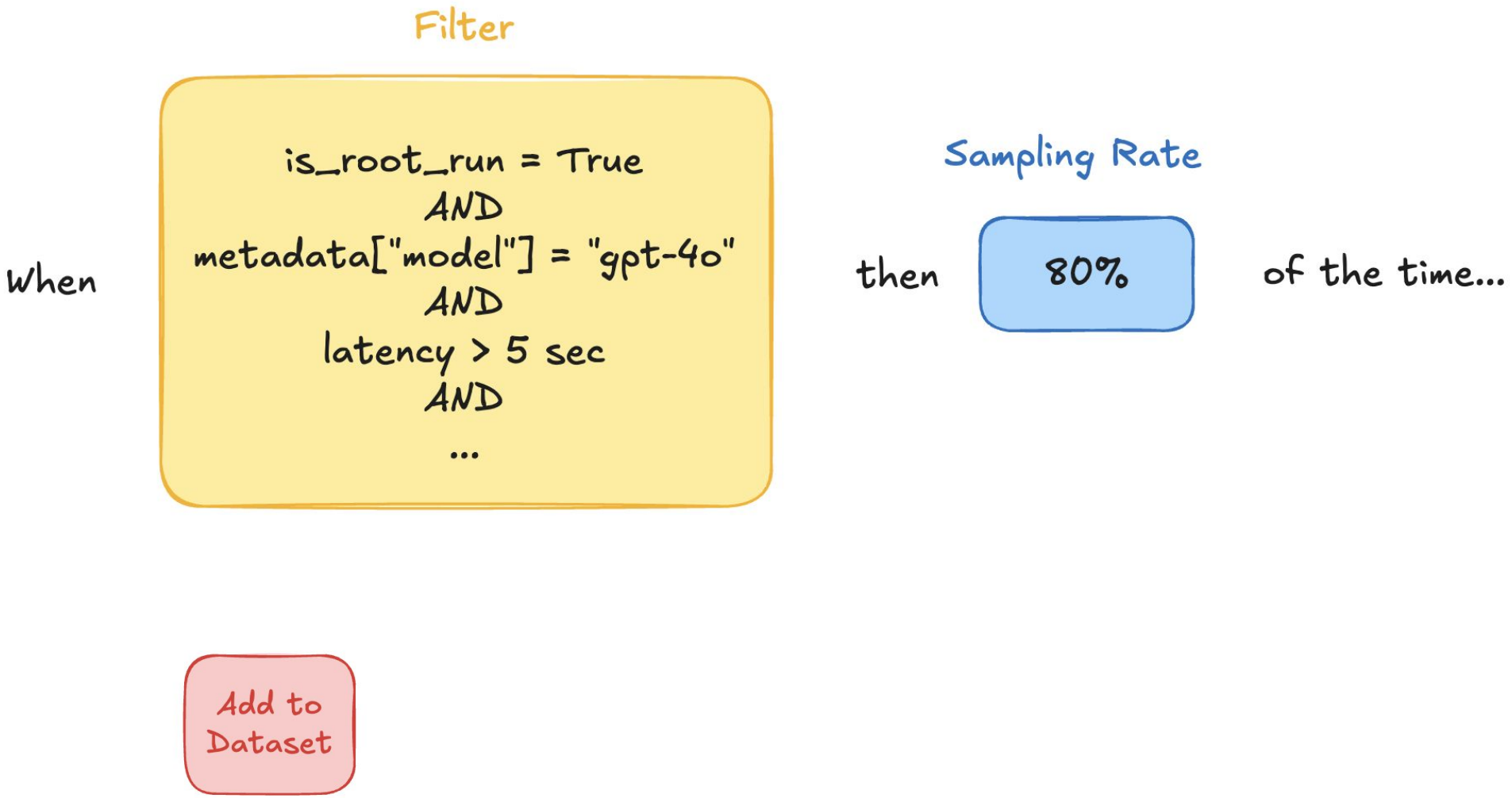
Automations

LangSmith allows you to create Rules that automatically trigger actions when in production



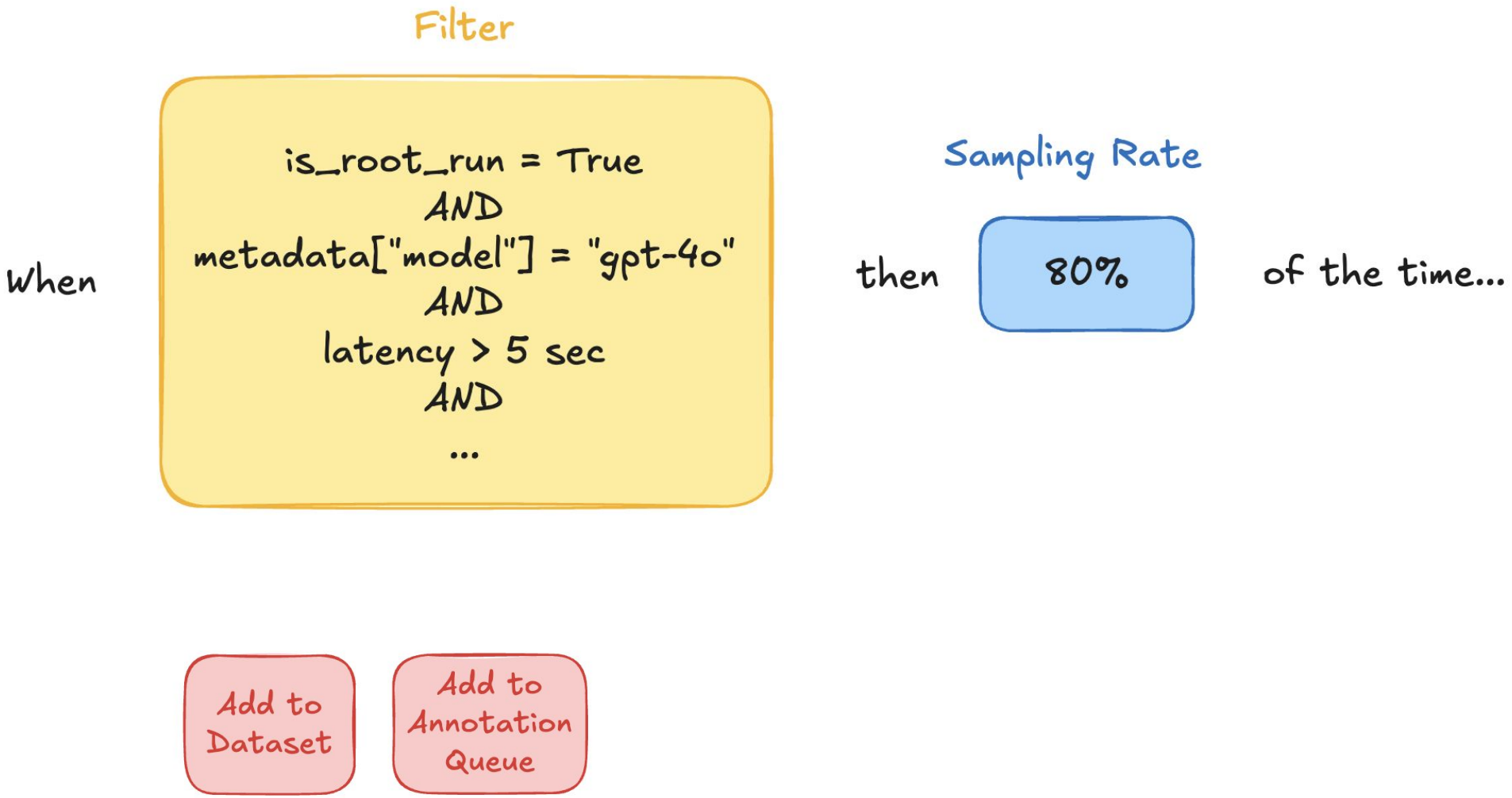
Automations

LangSmith allows you to create Rules that automatically trigger actions when in production



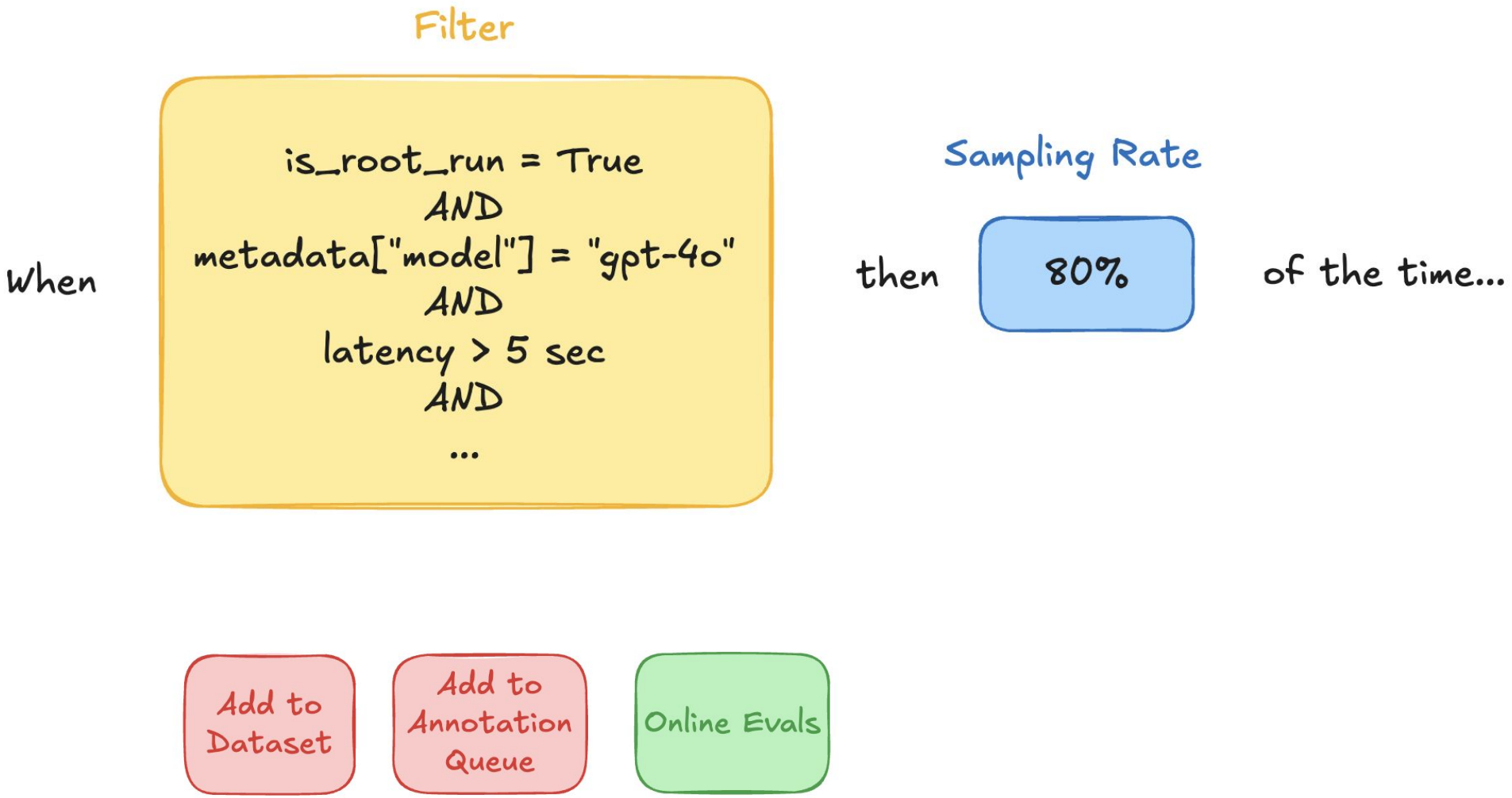
Automations

LangSmith allows you to create Rules that automatically trigger actions when in production



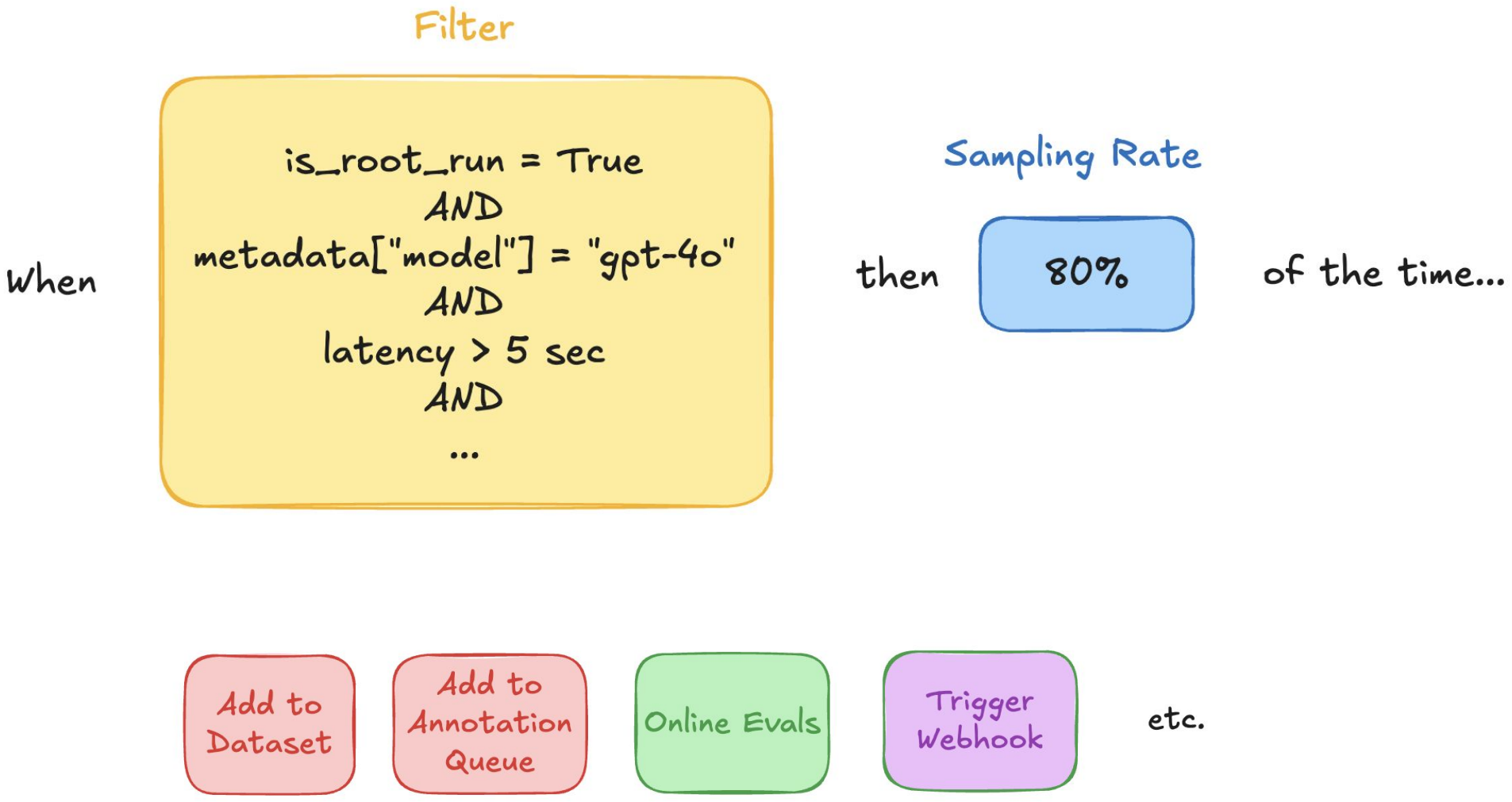
Automations

LangSmith allows you to create Rules that automatically trigger actions when in production



Automations

LangSmith allows you to create Rules that automatically trigger actions when in production

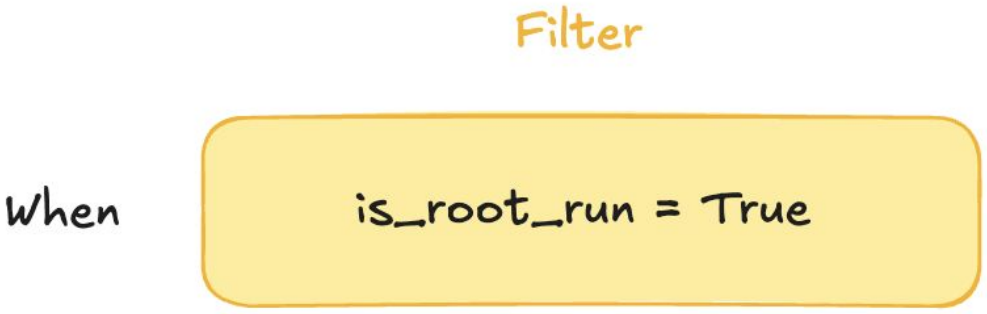


LangChain Academy

Automations

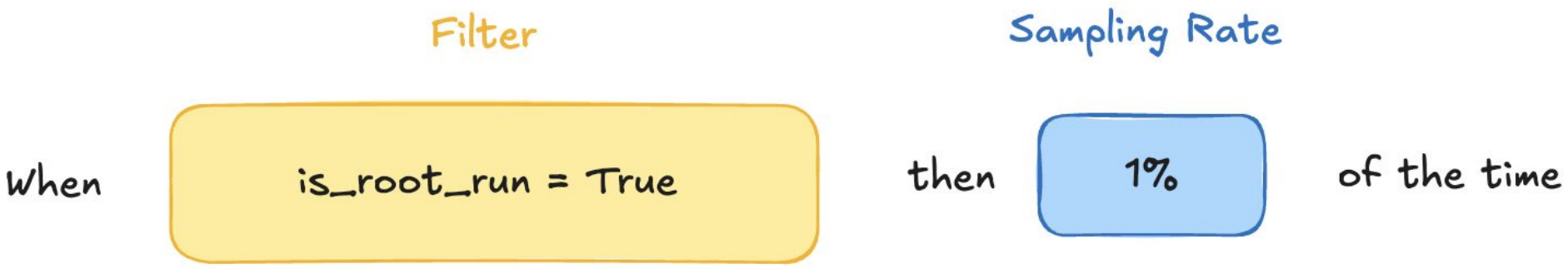
Popular Examples of Automations

Sample with low frequency from all traces, and add to an annotation queue



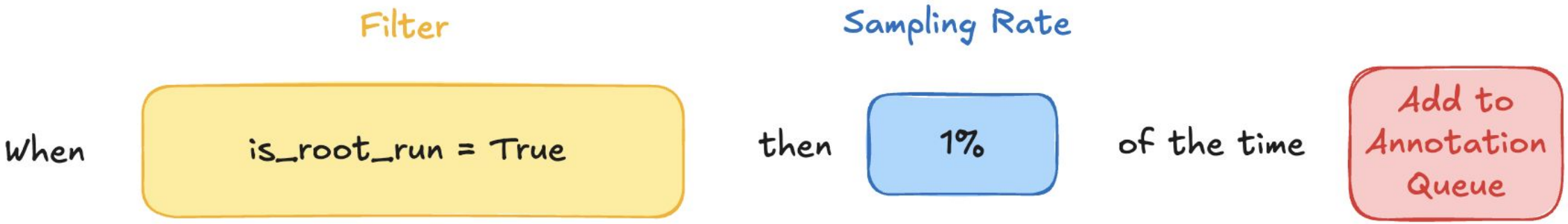
Popular Examples of Automations

Sample with low frequency from all traces, and add to an annotation queue



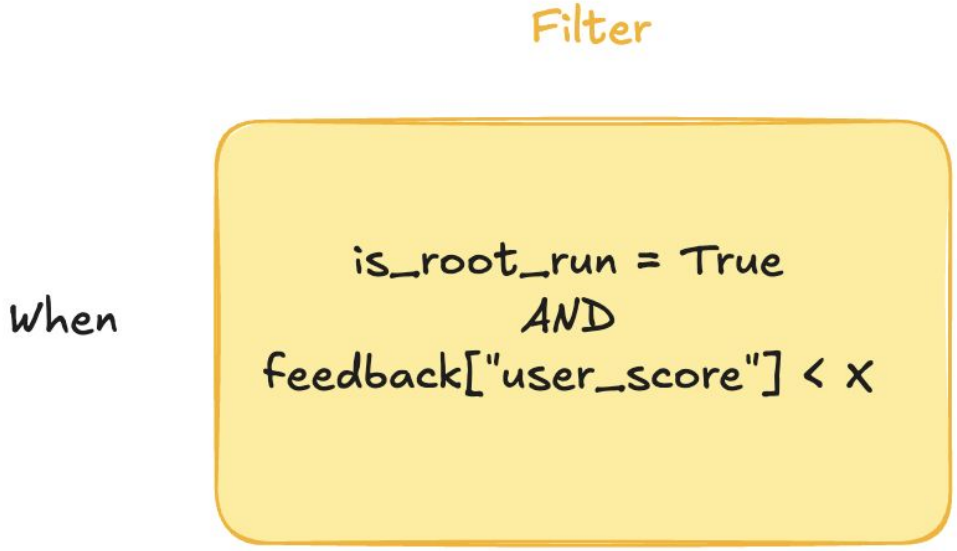
Popular Examples of Automations

Sample with low frequency from all traces, and add to an annotation queue



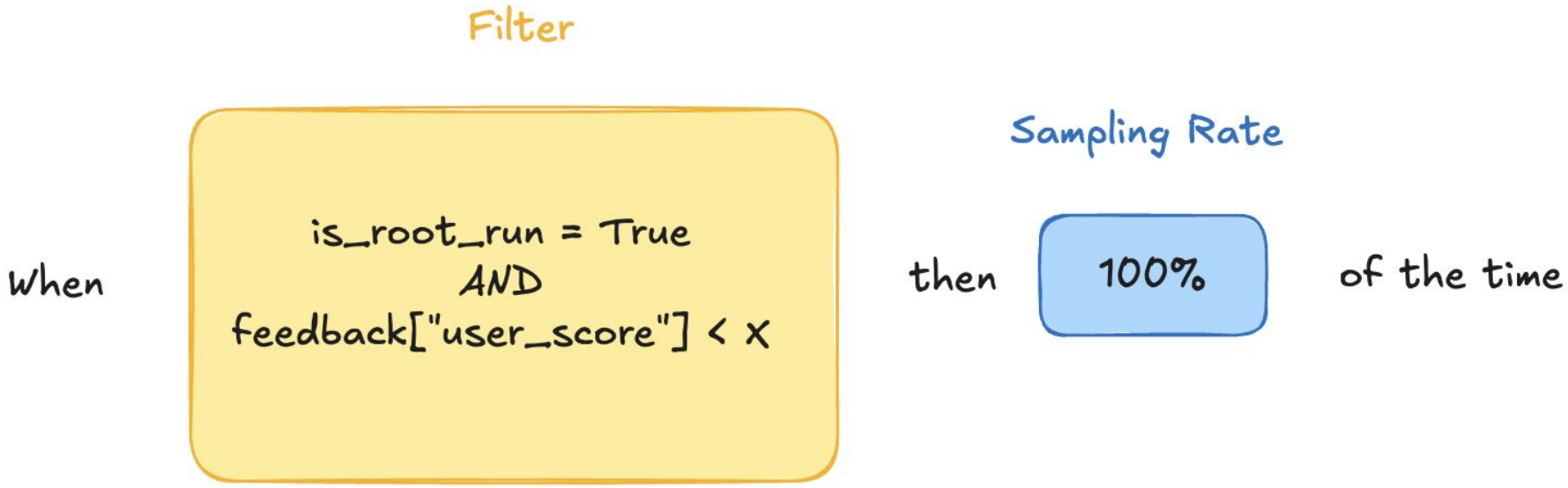
Popular Examples of Automations

Add all traces (or runs) with **negative** feedback to an annotation queue



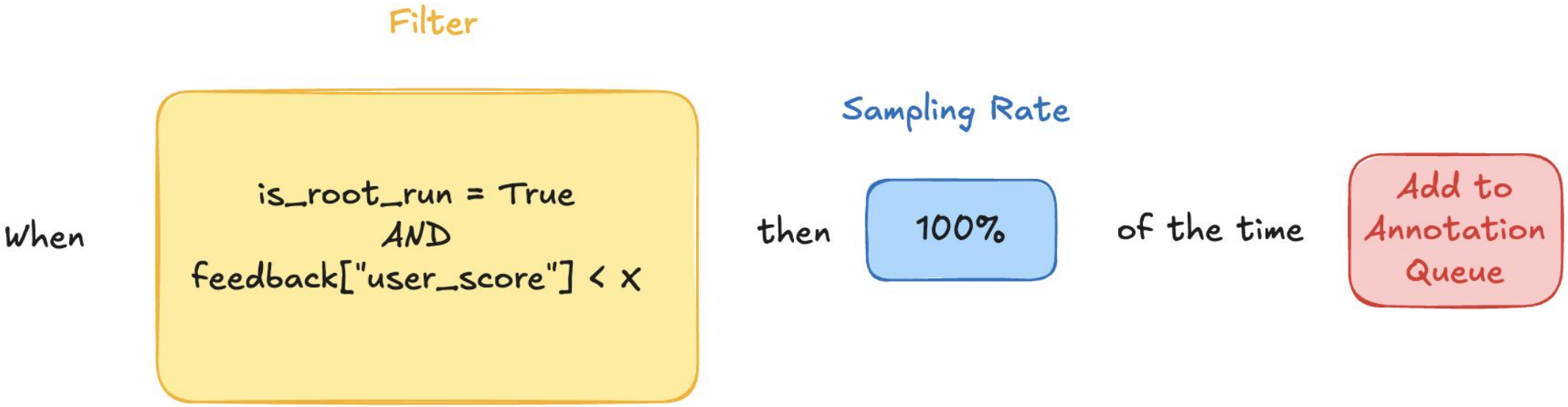
Popular Examples of Automations

Add all traces (or runs) with **negative** feedback to an annotation queue



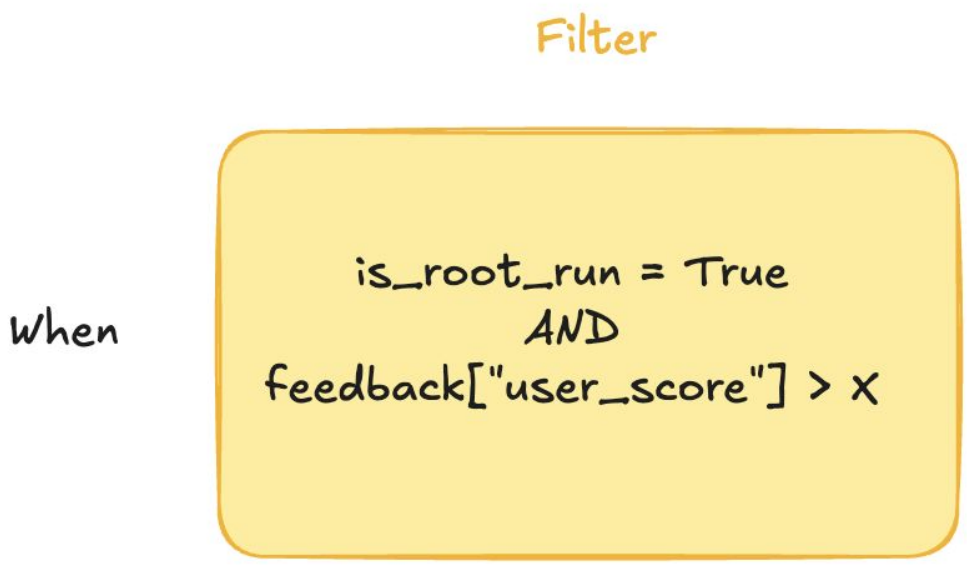
Popular Examples of Automations

Add all traces (or runs) with **negative** feedback to an annotation queue



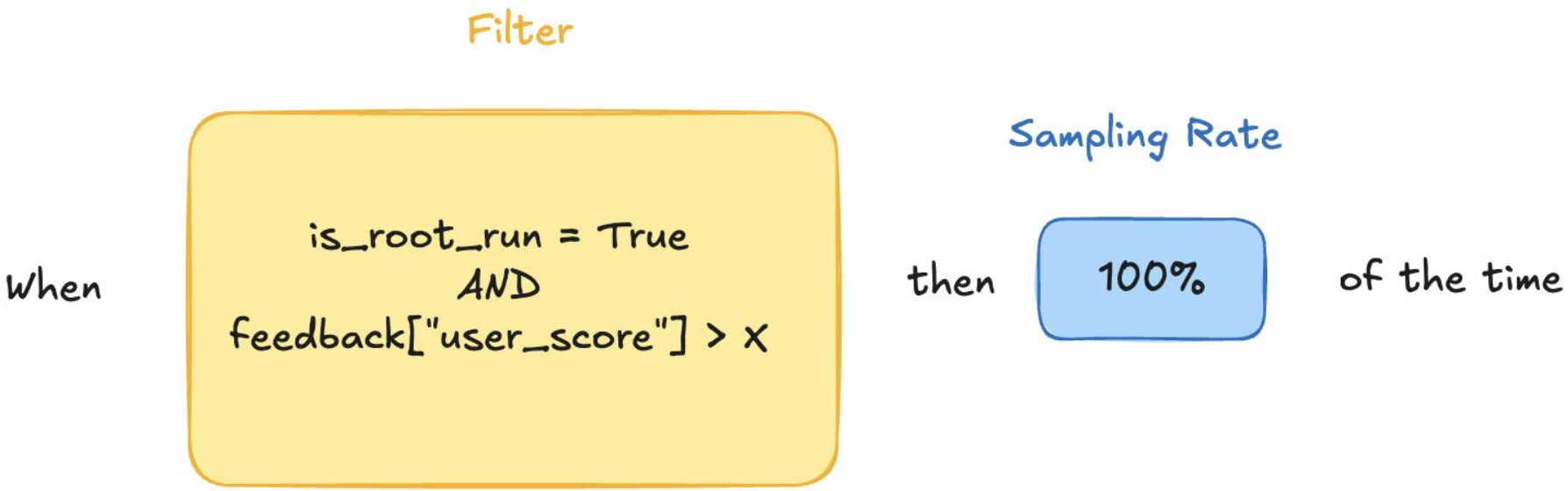
Popular Examples of Automations

Add all traces (or runs) with **positive** feedback to a golden dataset



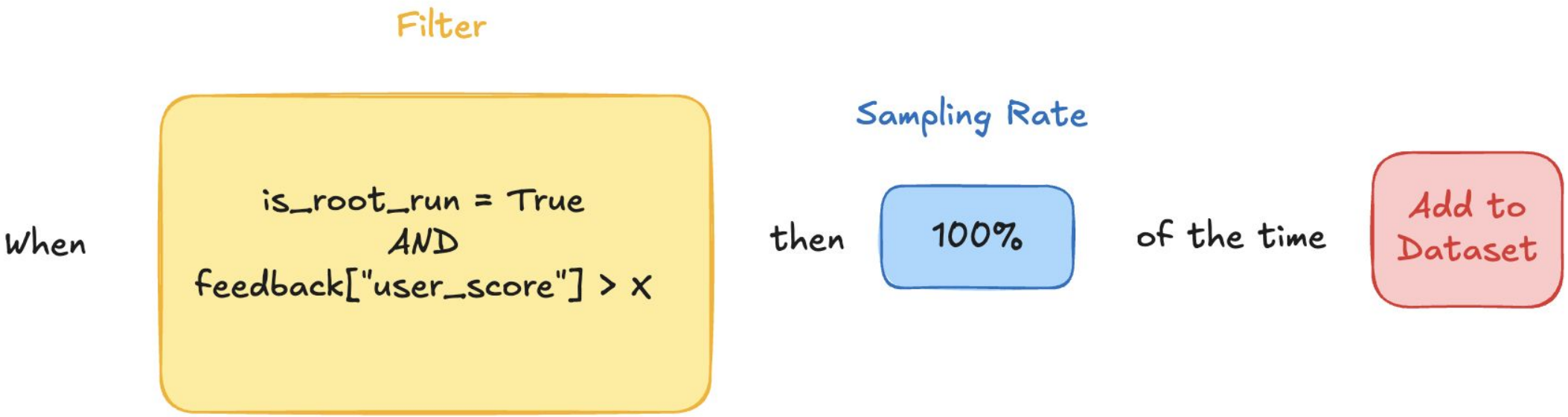
Popular Examples of Automations

Add all traces (or runs) with **positive** feedback to a golden dataset



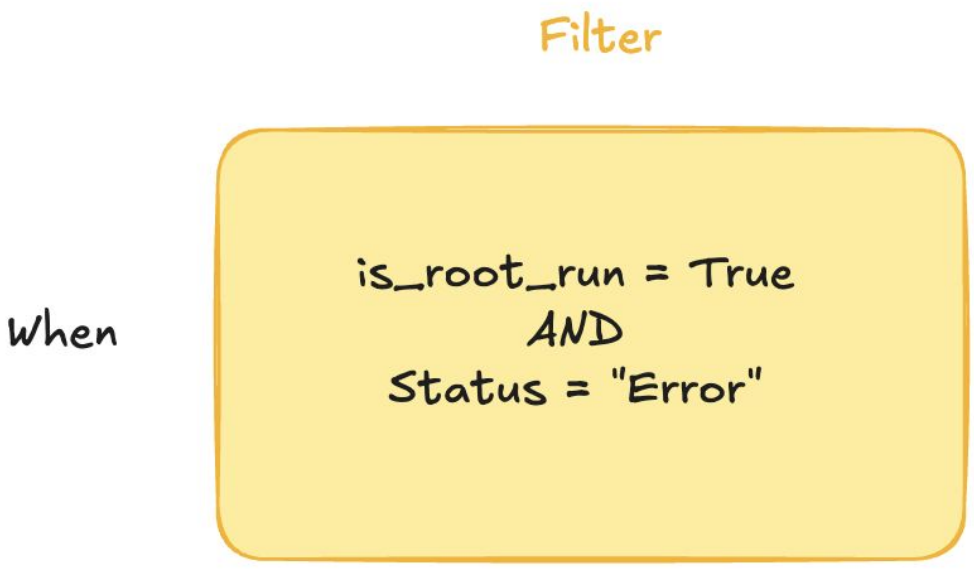
Popular Examples of Automations

Add all traces (or runs) with **positive** feedback to a golden dataset



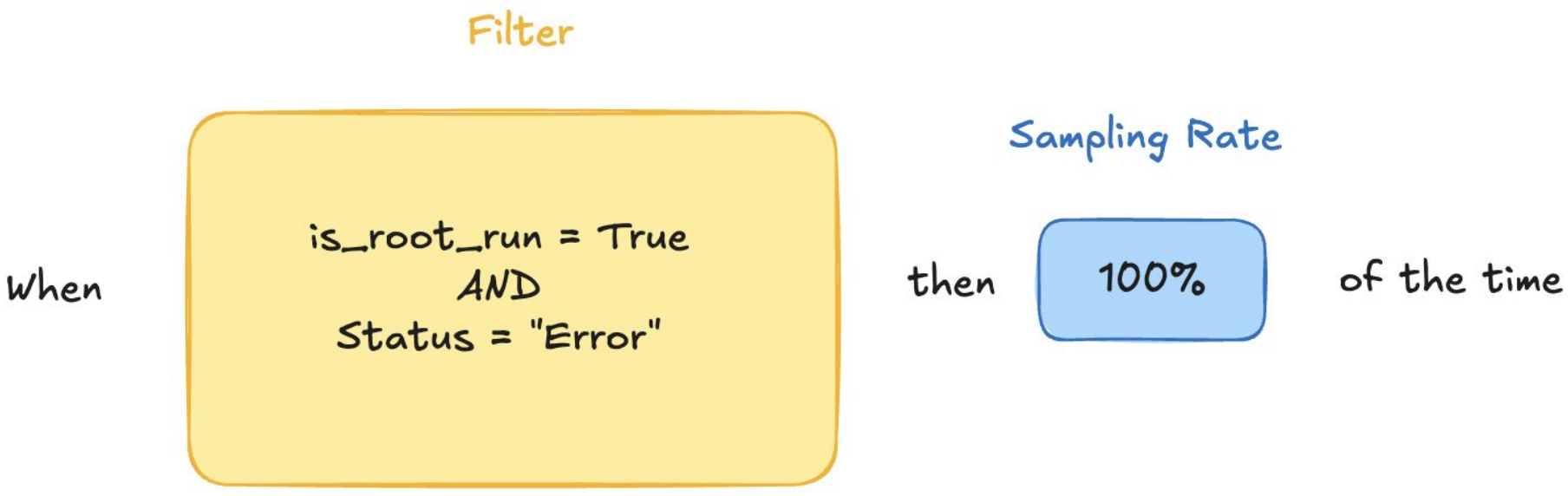
Popular Examples of Automations

For all traces (or runs) with an **error**, alert PagerDuty



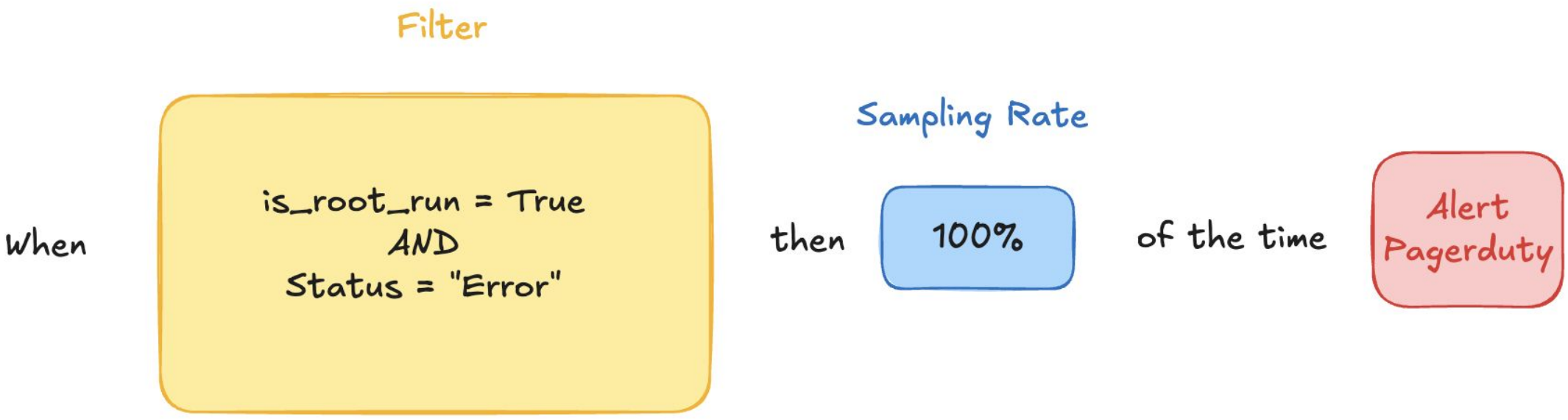
Popular Examples of Automations

For all traces (or runs) with an **error**, alert PagerDuty



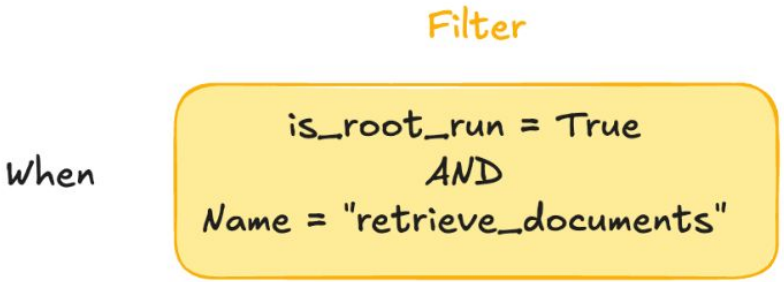
Popular Examples of Automations

For all traces (or runs) with an **error**, alert PagerDuty



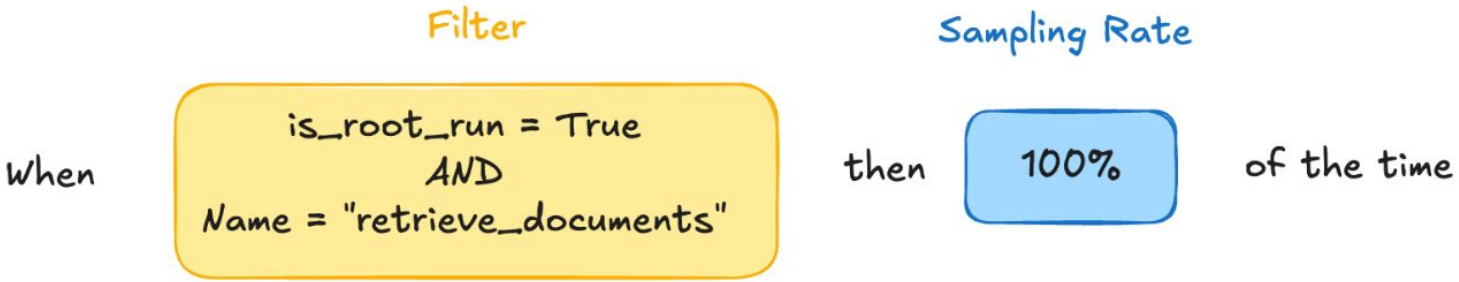
Popular Examples of Automations

Chain Online Evaluator Feedback, with Adding to Dataset



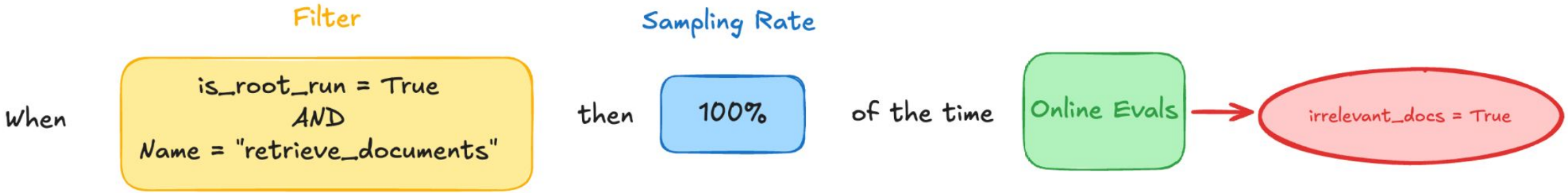
Popular Examples of Automations

Chain Online Evaluator Feedback, with Adding to Dataset



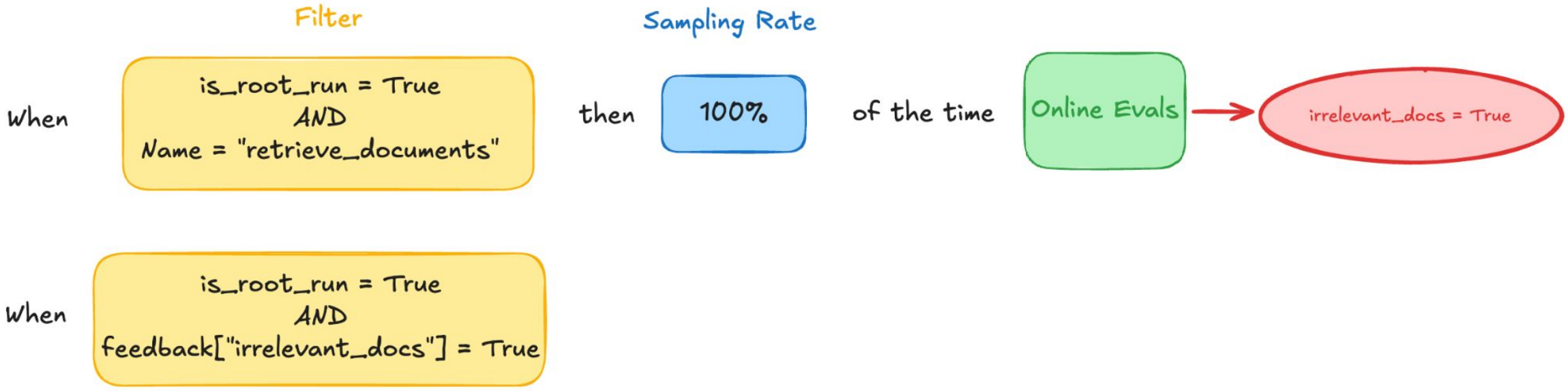
Popular Examples of Automations

Chain Online Evaluator Feedback, with Adding to Dataset



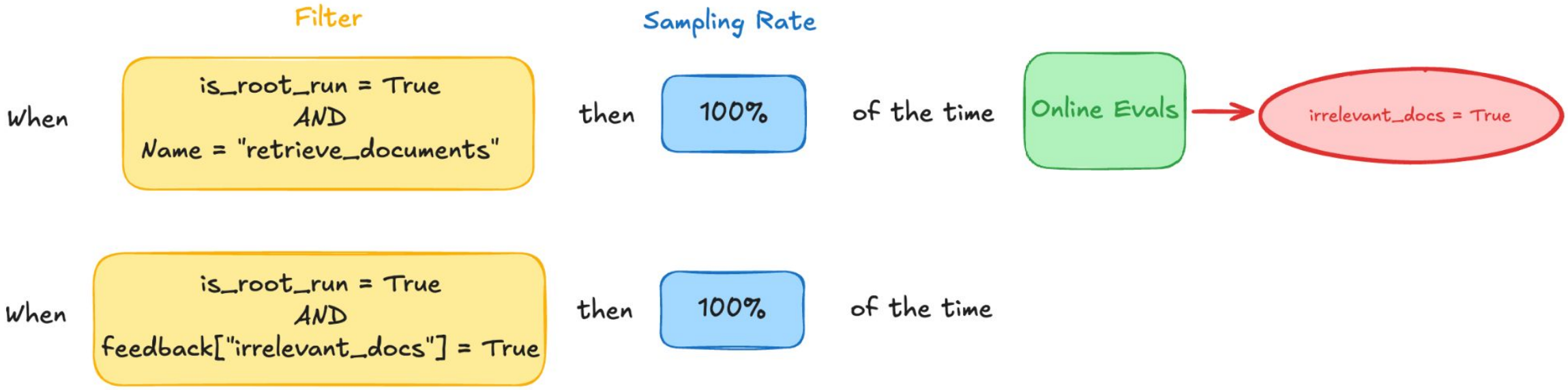
Popular Examples of Automations

Chain Online Evaluator Feedback, with Adding to Dataset



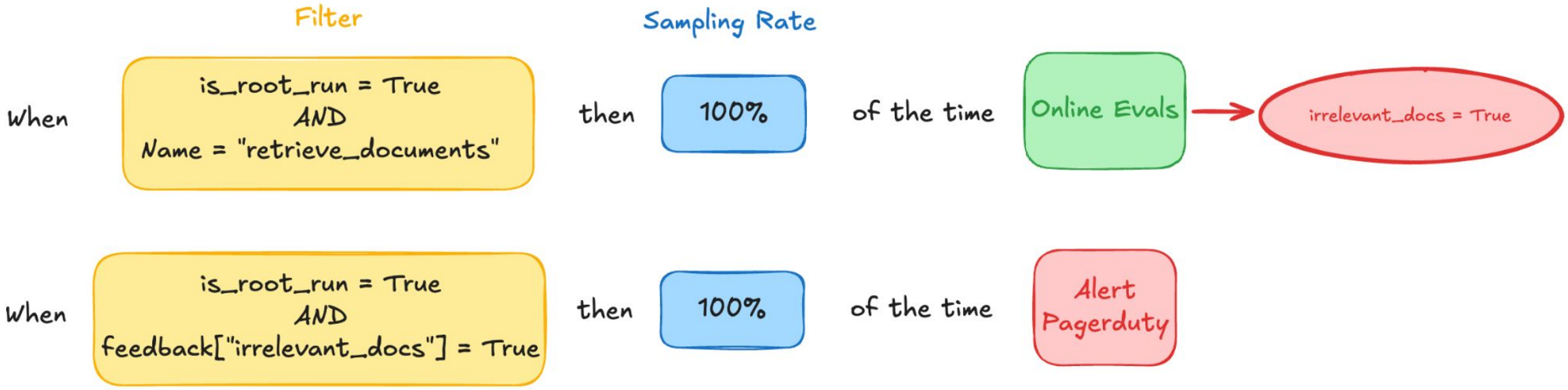
Popular Examples of Automations

Chain Online Evaluator Feedback, with Adding to Dataset



Popular Examples of Automations

Chain Online Evaluator Feedback, with Adding to Dataset



LangChain Academy

Monitoring View