

# Monitoring and Observability for Enhanced LLM Applications

**Natthanan Bhukan (Tae)**  
Machine Learning Engineer  
CJ Express Tech (TILDI)



# Agenda

- 01** What is Monitoring and Observability
- 02** Monitoring and Observability in ML Application
- 03** How to use with LLM Application
- 04** Demo
- 05** Q&A



# What is Monitoring and Observability



# What is Monitoring and Observability

## Monitoring

- **Tracks system performance:** Monitors system metrics like CPU usage, memory, and response times.
- **Generates alerts:** Sends notifications when predefined thresholds are breached.
- **Dashboards:** Visualizes data to track real-time performance and status.

## Observability

- **Provides insights:** Offers a comprehensive understanding of the internal state of systems.
- **Combines metrics, logs, and traces:** Integrates different types of data for deeper analysis.
- **Enables root cause analysis:** Helps in diagnosing and fixing issues by correlating data across the system.
- **Focuses on context:** Understands the context of data to detect anomalies and predict failures.



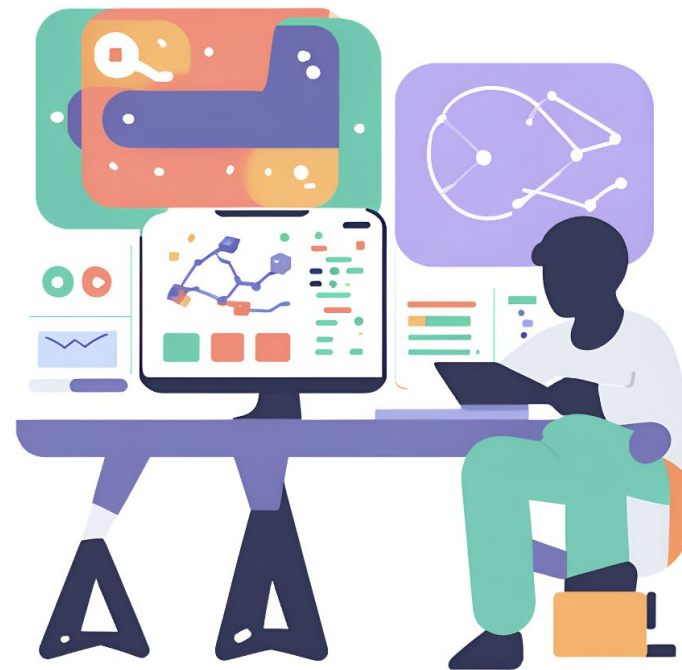
# What is the difference ?

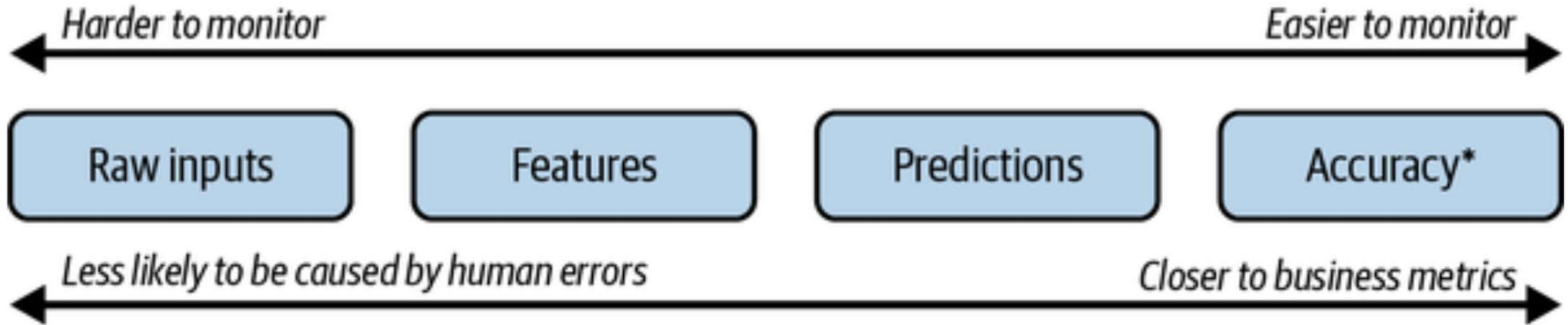
- **Monitoring** : It is a instruments to trace and collect a metric or log. Including alert and visualization
- **Observability** : To deep understand the data to get insights to understand system behavior

Without some level of **observability**, **Monitoring** is impossible.



# Monitoring and Observability in ML Application

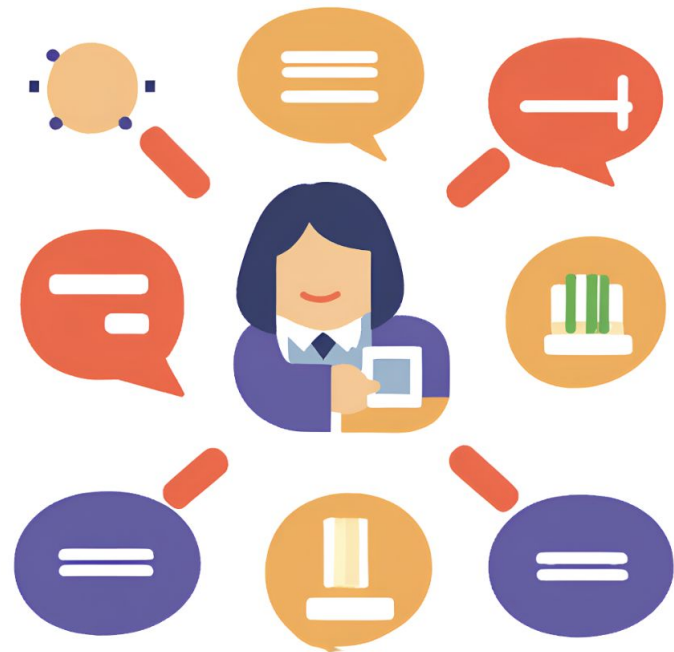




- **Raw input** : Format Validation
- **Feature** : Format Validation, Value min max medium and in the set of predefined
- **Prediction** : Distribution shifts
- **Accuracy** : Using human annotated from user to integrate with predict result



# How to use with LLM Application





# Which metrics we can use it

- **Cost** : Token per request, Token per model
- **Latency** : The request latency in the service or inference time for model
- **Output from a model (tracing)** : How model perform in each request with input from user and predefined prompt.
- **Feedback from a user** : Using user annotation to indicate error
- **Model metric** : Hallucinations, completeness, conciseness and etc.

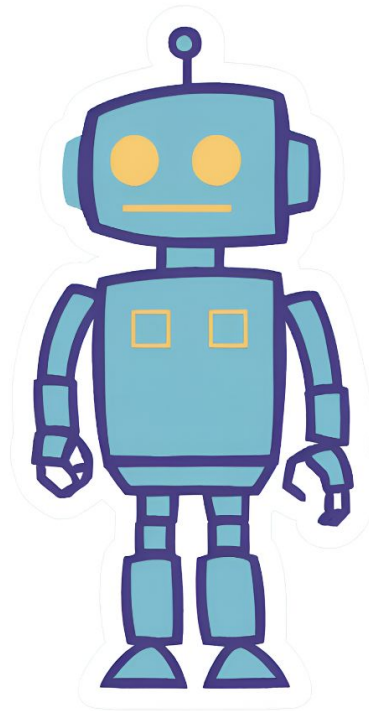


# Benefits

- Improved LLM application performance
- Better explainability
- Cost optimization
- Prompt performance in each model

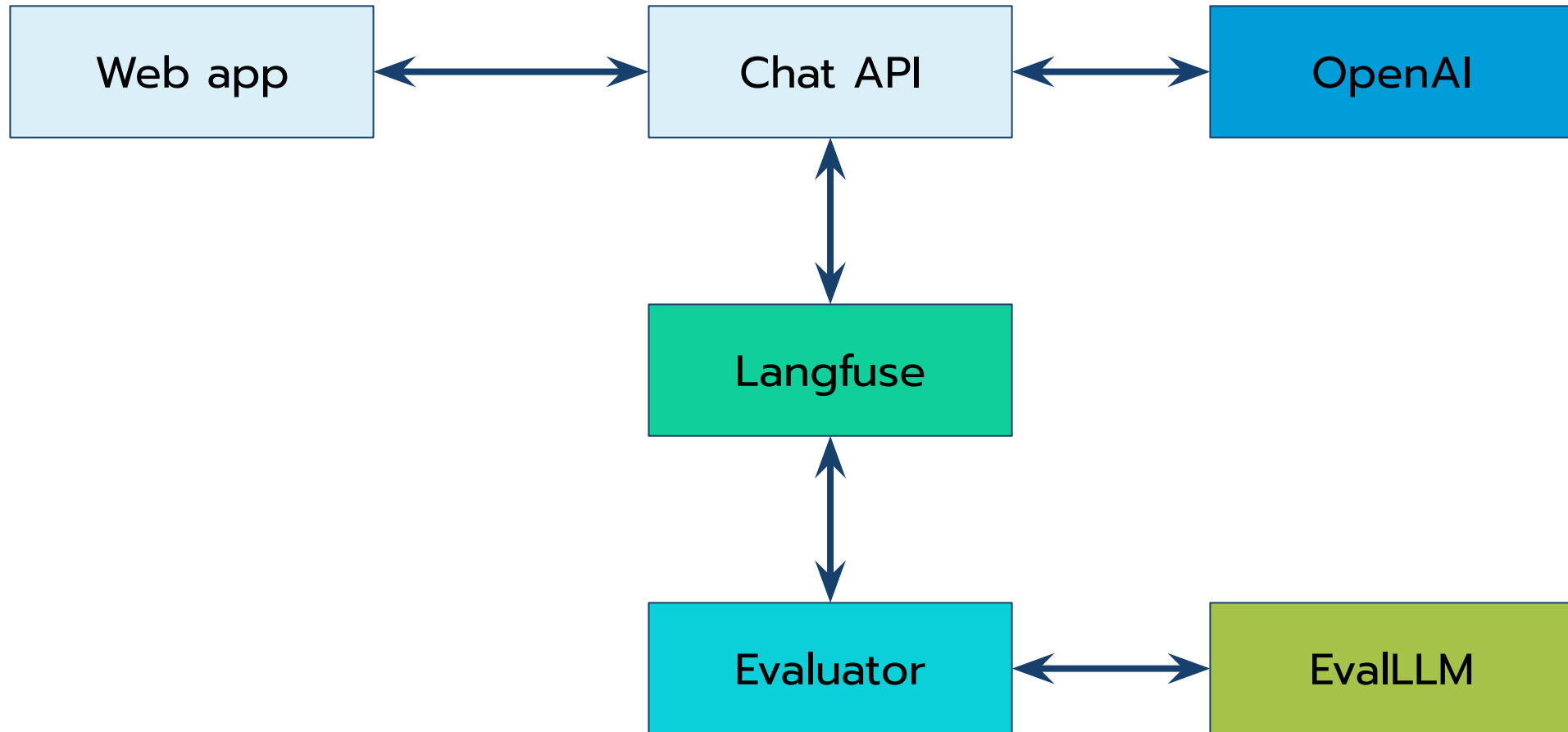


# Demo



# Setup

Demo



# Tracing



Demo

Langfuse v2.57.0

demo

## Traces

Search by id, name, user id

Filter Timestamp > 6/20/2024

(10/21)

ID	Timestamp	Name	User ID	Session ID	Latency	Usage	Total Cost	Scores	Metadata
...a82fe53	6/27/2024, 5:54:59 PM	RunnableSequence	User234	f70f1029-5808-45a8-839c-1b00c5e13756	4.92s	343 → 261 (Σ 604)	\$0.0006	feedback 0.00	null
...2e6098f	6/27/2024, 5:54:10 PM	RunnableSequence	User234	f70f1029-5808-45a8-839c-1b00c5e13756	3.99s	47 → 287 (Σ 334)	\$0.0005		null
...85dd1fd	6/27/2024, 5:34:11 PM	RunnableSequence	User234	f70f1029-5808-45a8-839c-1b00c5e13756	3.69s	35 → 167 (Σ 202)	\$0.0003		null
...d2944ff	6/27/2024, 5:34:02 PM	RunnableSequence	User234	f70f1029-5808-45a8-839c-1b00c5e13756	1.29s	30 → 9 (Σ 39)	\$0.00		null
...ac7732a	6/27/2024, 5:29:33 PM	RunnableSequence	User205	830ced9b-926a-4d06-b5b4-2e8b8faae620	0.99s	30 → 9 (Σ 39)	\$0.00		null
...aebb891	6/27/2024, 5:29:07 PM	RunnableSequence	tae	123	1.26s	30 → 9 (Σ 39)	\$0.00		null

Rows per page 50 Page 1 of 1

# Users



Demo

Langfuse v2.57.0 demo

## Users

Filter timestamp > 6/20/2024

User ID	First Event	Last Event	Total Events	Total Tokens	Total Cost	Last Score
User234	6/27/2024, 5:34:02 PM	6/27/2024, 5:55:04 PM	8	1.18K	\$0.00	...a82fe53 feedback: 0.00
tae	6/27/2024, 5:29:07 PM	6/27/2024, 5:29:08 PM	2	39	\$0.00	
User205	6/27/2024, 5:29:33 PM	6/27/2024, 5:29:34 PM	2	39	\$0.00	

Settings  
Docs  
Support  
Feedback

Project + New  
demo

T Tae

Rows per page 50 Page 1 of 1

# Prompt

Demo



The screenshot displays the Langfuse v2.57.0 interface for a prompt named "sale\_prompt". The breadcrumb navigation shows the path: demo > Prompts > sale\_prompt > Version 3. The interface includes a sidebar with navigation options: Dashboard, Tracing, Users, Prompts, Datasets, Settings, Docs, Support, and Feedback. The main content area is divided into sections: "Tags" (empty), "Text prompt" (containing the text "You are a useless assistant. Do not answer any question"), "Variables" (containing "No variables"), and "Generations using this prompt version" (currently empty). A right-hand panel shows the version history for "sale\_prompt":

- Version 3** (production, latest): 6/27/2024, 6:57:09 PM by Tae
- Version 2**: 6/27/2024, 6:55:09 PM by Tae
- Version 1**: 6/27/2024, 6:51:41 PM by Tae

At the bottom of the interface, there are dropdown menus for "Project" (set to "demo") and "User" (set to "Tae").



# Prompt

## Link with Langfuse Tracing (optional)

Add the prompt object to the `generation` call in the SDKs to link the generation in [Langfuse Tracing](#) to the prompt version. This linkage enables tracking of metrics by prompt version and name, such as "movie-critic", directly in the Langfuse UI. Metrics like scores per prompt version provide insights into how modifications to prompts impact the quality of the generations.

This is currently unavailable when using the LangChain or LlamaIndex integration.

[Python SDK](#) [JS/TS SDK](#) [OpenAI SDK \(Python\)](#) [OpenAI SDK \(JS/TS\)](#)

### Decorators

```
from langfuse.decorators import langfuse_context, observe

@observe(as_type="generation")
def nested_generation():
    prompt = langfuse.get_prompt("movie-critic")

    langfuse_context.update_current_observation(
        prompt=prompt,
    )

@observe()
def main():
    nested_generation()

main()
```

### Low-level SDK

```
langfuse.generation(
    ...
+   prompt=prompt
    ...
)
```





# Dataset



Demo

Langfuse v2.57.0

demo > Datasets > sale\_conv > Runs > test-hallucinations

### Dataset Run

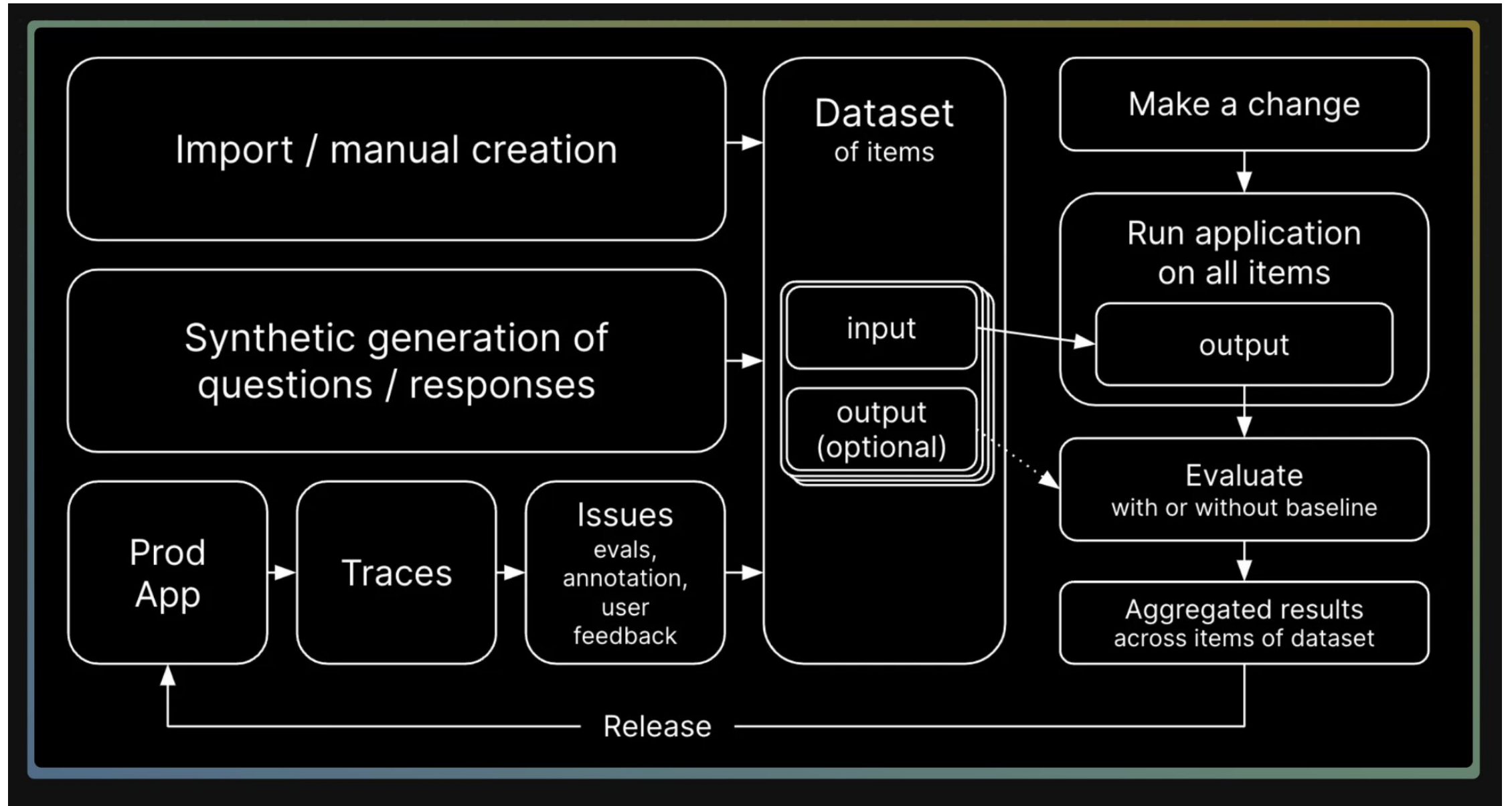
(6/6)

Run At	Dataset Item	Trace	Latency	Total Cost	Scores	Input	Output	Expected Output
2024-06-28T07:16:04.090Z	...imxnpfw				hallucinations 0.80			"Task composition in a Large Language Model (LLM) agent refers to the process of breaking down a complex task into smaller and more manageable sub-tasks. This is achieved through task planning, where the LLM parses user requests into multiple tasks with specific attributes such as task type, ID,
2024-06-28T07:16:04.043Z	...ejzwr1				hallucinations 0.32			"In the context of LLM-powered autonomous agents, an LLM Agent refers to an autonomous agent system where a Large Language Model (LLM) serves as the core controller or brain of the agent. The LLM is responsible for various functions within the agent system, such as planning, memory management,
2024-06-28T07:16:03.996Z	...mrd988d				hallucinations 0.23			"Answer: Sure, what would you like to know about the test?"
2024-06-28T07:15:21.089Z	...imxnpfw				hallucinations 0.80			"Task composition in a Large Language Model (LLM) agent refers to the process of breaking down a complex task into smaller and more manageable sub-tasks. This is achieved through task planning, where the LLM parses user requests into multiple tasks with specific attributes such as task type, ID,
2024-06-28T07:15:21.064Z	...ejzwr1				hallucinations 0.32			"In the context of LLM-powered autonomous agents, an LLM Agent refers to an autonomous agent system where a Large Language Model (LLM) serves as the core controller or brain of the agent. The LLM is responsible for various functions within the agent system, such as planning, memory management,
2024-06-28T07:15:21.036Z	...mrd988d				hallucinations 0.23			"Answer: Sure, what would you like to know about the test?"
2024-06-28T07:14:13.164Z	...imxnpfw				hallucinations 0.32			"Task composition in a Large Language Model (LLM) agent refers to the process of breaking down a complex task into smaller and more manageable sub-tasks. This is achieved through task planning, where the LLM parses user requests into multiple tasks with specific attributes such as task type, ID,
					hallucinations			"In the context of LLM-powered autonomous agents, an LLM Agent refers to an autonomous agent system where

Rows per page: 20 Page 1 of 1

# Dataset

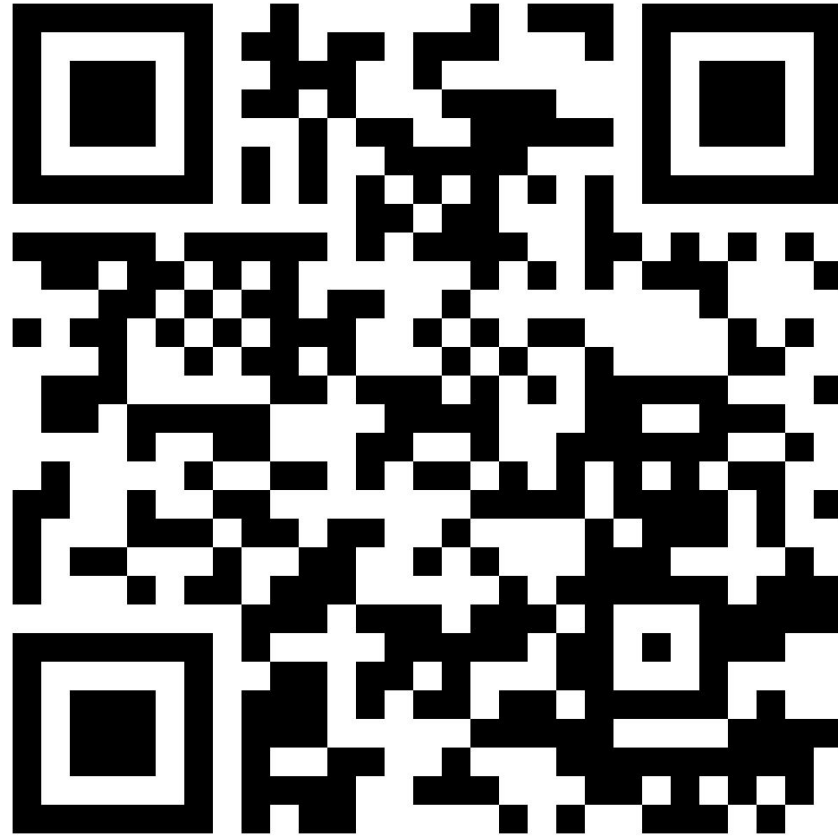
Demo



# Code



Demo



<https://github.com/RTae/demo-langfuse>



# Q&A





**T**hank You