



Langfuse

Langfuse is the open source LLM Engineering Platform

What is Langfuse?

- Ingest **full context of your LLM application** via SDKs and integrations (Python Decorator, JS/TS, Llama Index, LangChain, OpenAI SDK & [more](#))
- Build a **rich dataset** of your app's performance and use it downstream (e.g. to fine-tune, debug) +get dashboards, analytics & stats on your app
- **Iterate and improve** on your LLM application with **Prompt Management, Playground, evals** (human & model-based), **datasets** (testing) and [more](#)

Why Langfuse?

- **Open Source:** MIT-licensed, self-hostable & open core
- **Developer-first:** technical product, focus on DX & documentation
- **Built for complex apps:** Capture whole context of your application
- **No Lock-In:** Model & framework agnostic, export data by API & csv/json
- **Security & Compliance:** SOC2 Type2, ISO27001 & GDPR certified
- **Reliable Partner:** Raised \$4m from YC, Lightspeed, General Catalyst
- **Strong adoption ([metrics](#)):** >2M SDK installs in Oct 24, 6k+ GitHub stars

Trying Langfuse: Look and Feel

- **Look and feel**

- A few screenshots in the following slides
- [Videos on Youtube](#)

- **Try Langfuse Cloud: <https://langfuse.com/demo>**

- Free + Access to demo data
- Set up own project and start playing around
- [Cookbooks](#) to get started

- **Self host**

- Locally: <https://langfuse.com/docs/deployment/local> (docker compose, setup ~ 5min)
- Docker: <https://langfuse.com/docs/deployment/self-host> (docker run, setup ~ 30min)

Trace Detail View in Langfuse

Helps monitor LLM and non-LLM steps within a chat/agent/RAG app

Add additional metadata such as userIds or sessionId to interactively replay what a user did

Visualizes inputs, outputs and evaluations results.

Example public links:
[session](#), [trace](#)

The screenshot displays the Langfuse Trace Detail View for a specific session. The interface is dark-themed and includes a sidebar on the left with navigation options like Tracing, Traces, Sessions, Generations, Scores, Models, Evaluation (Beta), Templates, Configs, Log, Users, Prompts, Playground (Beta), Datasets, Settings, Docs, Support, and Feedback. The main content area shows the trace details for a session with ID 'c8e017c5-eb7d-4f24-9811-d8de84377e86'. The trace is titled 'qa' and occurred on 6/3/2024 at 11:27:17 AM. It has a duration of 2.00s and a total cost of \$0.0002. The input is 'What are the topics you know of?'. The output is a response from the LLM: 'I can help you with understanding and using Langfuse, an open-source observability tool for developers working with Large Language Models (LLMs). If you have any questions about Langfuse or observability in LLM applications, feel free to ask. How can I assist you today?'. The trace includes a 'Metadata' section with a path name of '/docs/demo' and a 'Scores' section with the following evaluation results: conciseness-1: 0.40, contains-pii-v1: 0.00, contextrelevance-v1: 0.00, forbidden-words-v1: 1.00, and helpfulness-1: 0.80. A 'Preview' tab is active, showing the input and output. A 'Scores' tab is also visible, showing a list of scores for various metrics. A 'Warnings' section is present, indicating a 'WARNING' for 'prompt-embedding' with a score of 0.41s and a 'DEBUB' for 'vector-store' with a score of 0.12s. Other scores include 'context-encoding' (0.00s) and 'fetch-prompt-from-langfuse' (0.00s). The 'GENERATION' score for 'generation' is 1.47s. The interface also includes a 'Tags' section with 'no-context' and a 'Private' toggle.

Trace Annotation in Langfuse

Collaboratively
annotate traces in the
Langfuse UI.

Define categorical,
binary and continuous
labels which
annotators need to
use.

All annotation results
are available across
Langfuse.

Note: Screenshot of
development preview,
annotation drawer to
be released this week.

The screenshot displays the Langfuse interface for a specific trace. On the left is a navigation sidebar with options like Dashboard, Tracing, Sessions, Generations, Scores, Models, Evaluation (Beta), Templates, Configs, Log, Users, Prompts, Playground (Beta), and Datasets. The main area shows the 'Trace Detail' for a trace with ID 'trace-371dbdad-2f2b-4f7d-9a4f-3ece226df139'. It includes a breadcrumb 'demo-app > Traces > trace-371dbdad-2f2b-4f7d-9a4f-3ece226df139', a 'Trace Detail' title, and summary statistics: '8824 → 3946 (Σ 12770)' and 'Total cost: \$0.0455'. Below this are 'Tags' for 'production' and 'blue'. The trace content is shown in a 'Preview' tab, featuring an input field with the text 'I'm looking for a React component', an output field with 'What kind of component are you looking for?', and a metadata section with a JSON object:

```
{  "more": "1,2,3;476"  "user": "user-89@langfuse.com"}
```

. A 'Scores' section at the bottom shows 'API sentiment: 1.00' and 'ANNOTATION Accuracy: Good Toxicity: -1.00'. On the right, an 'Annotate' drawer is open, showing a 'Select' dropdown with '5 selected' and several annotation fields: 'Accuracy' with 'Bad (0)' and 'Good (1)' buttons, 'Toxicity' with a value of '-1', and three 'helpfulness' fields with values '87', '55', and '1'. Each field has a close button (X).

Tables in Langfuse

Easily filter, search, sort and export the data that you need.

The screenshot displays the Langfuse interface for a project named 'langfuse-docs'. The main view is the 'Generations' table, which lists individual AI outputs. The table includes columns for ID, Name, Trace ID, Trace Name, Start Time, End Time, Time to First Token, Scores, Latency, Time per Output Token, Total Cost, Level, and Model. A search bar at the top allows filtering by ID, name, or trace name. A filter is currently set to 'Start Time > 5/26/2024'. The table shows 13 rows of data, with the first row having a 'WARNING' level and the last row having a 'WARNING' level. The interface also features a sidebar with navigation options like Tracing, Sessions, Generations, Scores, Models, Evaluation, Templates, Configs, Log, Users, Prompts, Playground, Datasets, Settings, Docs, Support, and Feedback. The bottom right corner shows 'Rows per page 50' and 'Page 1 of 16'.

ID	Name	Trace ID	Trace Name	Start Time	End Time	Time to First Token	Scores	Latency	Time per Output Token	Total Cost	Level	Model
...2a2eae0	generation	...306e177	qa	6/3/2024, 11:44:49 AM	6/3/2024, 11:44:50 AM	0.49s		0.59s	0.07s	\$0.0001	DEFAULT	gpt-3.5-turbo
...383aa12	prompt-embedding	...306e177	qa	6/3/2024, 11:44:47 AM	6/3/2024, 11:44:47 AM	-		0.50s	0.25s	\$0.00	DEBUG	text-embedding-ada-00
...71b7057	generation	...9252dee	qa	6/3/2024, 11:27:28 AM	6/3/2024, 11:27:29 AM	0.36s		0.69s	0.03s	\$0.0001	DEFAULT	gpt-3.5-turbo
...2252971	prompt-embedding	...9252dee	qa	6/3/2024, 11:27:27 AM	6/3/2024, 11:27:27 AM	-		0.30s	0.10s	\$0.00	DEBUG	text-embedding-ada-00
...bf1df37	generation	...4377e86	qa	6/3/2024, 11:27:16 AM	6/3/2024, 11:27:17 AM	0.65s		1.47s	0.03s	\$0.0002	DEFAULT	gpt-3.5-turbo
...dfb18c6	prompt-embedding	...4377e86	qa	6/3/2024, 11:27:15 AM	6/3/2024, 11:27:16 AM	-		0.41s	0.05s	\$0.00	DEBUG	text-embedding-ada-00
...d1e4d12	generation	...7043a98	qa	6/3/2024, 11:26:54 AM	6/3/2024, 11:26:55 AM	0.55s		1.18s	0.03s	\$0.0001	DEFAULT	gpt-3.5-turbo
...4485121	prompt-embedding	...7043a98	qa	6/3/2024, 11:26:52 AM	6/3/2024, 11:26:52 AM	-		0.65s	0.16s	\$0.00	DEBUG	text-embedding-ada-00
...0e8ce01	generation	...6270fb7	qa	6/3/2024, 11:20:21 AM	6/3/2024, 11:20:22 AM	0.51s		0.63s	0.07s	\$0.0001	DEFAULT	gpt-3.5-turbo
...435c928	prompt-embedding	...6270fb7	qa	6/3/2024, 11:20:21 AM	6/3/2024, 11:20:21 AM	-		0.61s	0.61s	\$0.00	DEBUG	text-embedding-ada-00
...3c2c292	generation	...701271e	qa	6/3/2024, 11:08:12 AM	6/3/2024, 11:08:17 AM	0.59s		5.38s	0.02s	\$0.0011	DEFAULT	gpt-3.5-turbo
...31402f5	prompt-embedding	...701271e	qa	6/3/2024, 11:08:11 AM	6/3/2024, 11:08:12 AM	-		0.33s	0.04s	\$0.00	DEBUG	text-embedding-ada-00
...579300a	generation	...4998b1c	qa	6/3/2024, 11:07:38 AM	6/3/2024, 11:07:42 AM	0.56s		3.48s	0.02s	\$0.0008	DEFAULT	gpt-3.5-turbo
...523bf1b	prompt-embedding	...4998b1c	qa	6/3/2024, 11:07:37 AM	6/3/2024, 11:07:38 AM	-		0.42s	0.02s	\$0.00	DEBUG	text-embedding-ada-00
...20f4bb9	generation	...aab82fe	qa	6/3/2024, 11:06:01 AM	6/3/2024, 11:06:02 AM	0.47s		1.25s	0.03s	\$0.0001	DEFAULT	gpt-3.5-turbo
...ff71e30	prompt-embedding	...aab82fe	qa	6/3/2024, 11:06:00 AM	6/3/2024, 11:06:00 AM	-		0.55s	0.07s	\$0.00	DEBUG	text-embedding-ada-00
...9c5f585	generation	...a4add26	qa	6/3/2024, 11:05:38 AM	6/3/2024, 11:05:39 AM	0.55s		0.95s	0.03s	\$0.0001	WARNING	gpt-3.5-turbo
...25b1dcf	prompt-embedding	...a4add26	qa	6/3/2024, 11:05:37 AM	6/3/2024, 11:05:38 AM	-		0.51s	0.05s	\$0.00	DEBUG	text-embedding-ada-00
...754bd33	generation	...52ff013	qa	6/3/2024, 10:55:11 AM	6/3/2024, 10:55:12 AM	0.41s		0.68s	0.06s	\$0.0003	WARNING	gpt-3.5-turbo
...c6a6e83	prompt-embedding	...52ff013	qa	6/3/2024, 10:55:10 AM	6/3/2024, 10:55:10 AM	-		0.22s	0.02s	\$0.00	DEBUG	text-embedding-ada-00

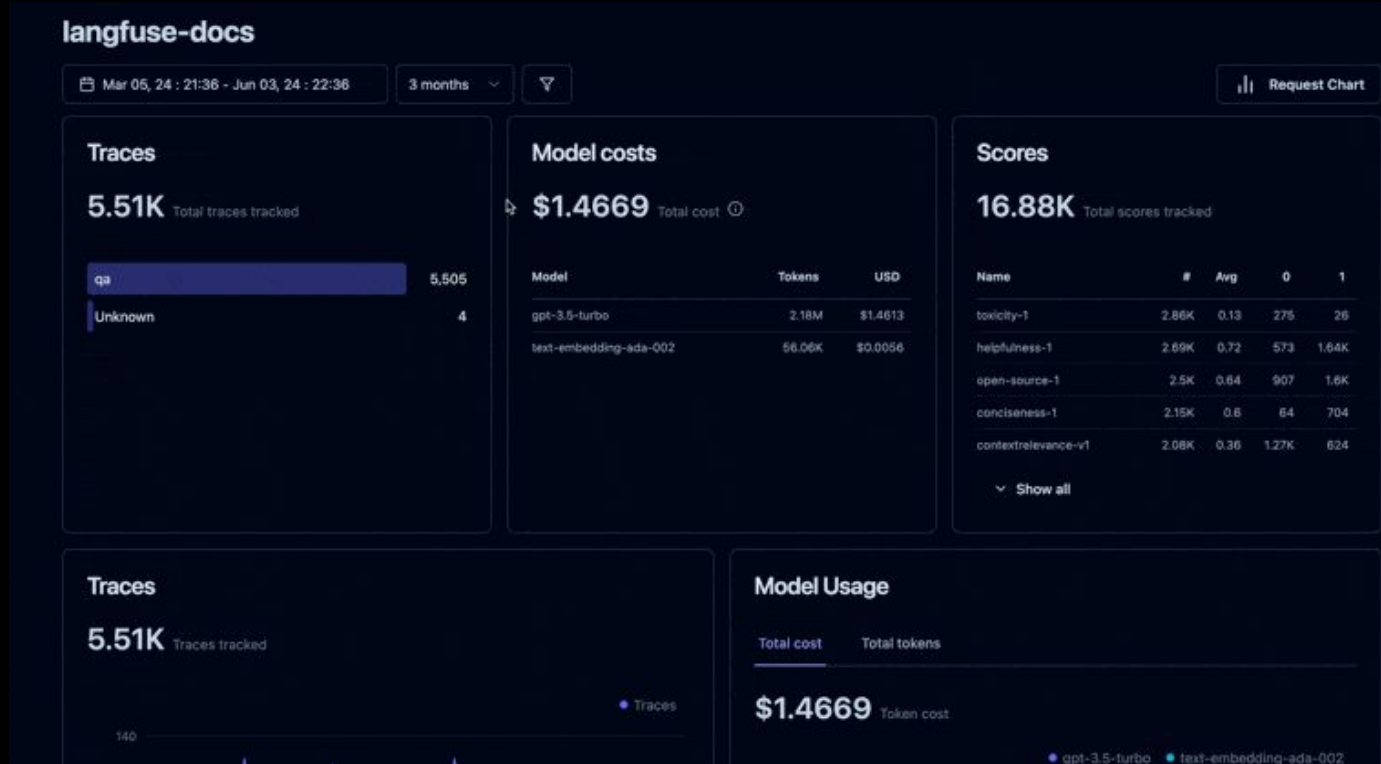
Metrics in Langfuse

Cost: support for any LLM via custom model definitions

Latency: distributions for subsections of the application

Quality: based on scores in Langfuse.
Can be online evaluation or manual annotation in Langfuse, or ingested via API (e.g. end user feedback)

Cost/usage-related metrics also available via API for cost-control or billing use cases



Prompt Management in Langfuse

Deployment labels, e.g. prod/staging

Support for text and ChatML prompts

Support for variables in prompts

Support for additional JSON configuration to e.g. version control retrieval or model parameters

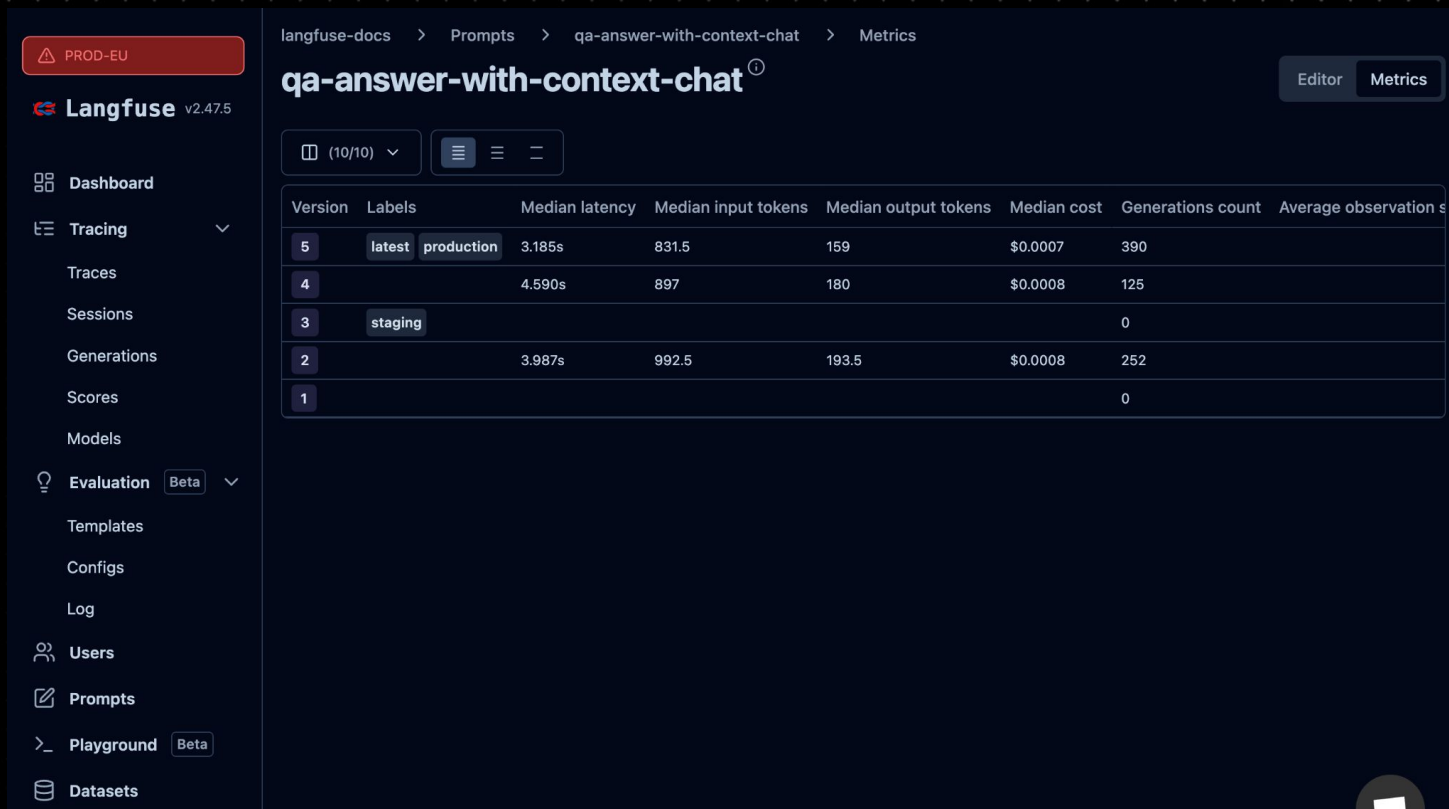
Average metrics / prompt version

Low-latency fetching via SDK, cached at run time to further reduce latency impact

The screenshot displays the Langfuse web interface for managing prompts. The left sidebar contains navigation options: Dashboard, Tracing, Traces, Sessions, Generations, Scores, Models, Evaluation (Beta), Templates, Configs, Log, Users, Prompts, Playground (Beta), Datasets, Settings, Docs, Support, and Feedback. The main content area shows the prompt configuration for 'qa-answer-with-context' (Version 6). The 'Text prompt' section contains the following text: 'You are a very enthusiastic Langfuse representative who loves to help people! Langfuse is an open-source observability tool for developers of applications that use Large Language Models (LLMs). Given the following sections from the Langfuse documentation, answer the question using only that information, outputted in markdown format. Refer to the respective links of the documentation.' Below this is a code block with a template: 'START of Langfuse Documentation', '****', '{{context}} {{context}}', '****', 'END of Langfuse Documentation'. The 'Variables' section shows two 'context' variables. A right-hand panel lists prompt versions: Version 6 (latest, 4/16/2024, 12:32:43 AM by Clemens), Version 5 (production, 3/25/2024, 7:19:51 AM by Clemens), Version 4 (3/18/2024, 9:45:06 AM by Marc), Version 3 (3/6/2024, 8:09:18 AM by Marc), Version 2 (2/6/2024, 8:44:05 AM by Marc), and Version 1 (1/2/2024, 5:51:03 AM by Marc). The bottom right corner features a chat bubble icon.

Prompt Management in Langfuse

View metrics by prompt version
(1) latency
(2) token usage
(3) usd cost
(4) evaluation and annotation scores



langfuse-docs > Prompts > qa-answer-with-context-chat > Metrics

qa-answer-with-context-chat [Ⓢ]

Editor Metrics

(10/10) [Menu] [Filter]

Version	Labels	Median latency	Median input tokens	Median output tokens	Median cost	Generations count	Average observation s
5	latest production	3.185s	831.5	159	\$0.0007	390	
4		4.590s	897	180	\$0.0008	125	
3	staging					0	
2		3.987s	992.5	193.5	\$0.0008	252	
1						0	

Evaluations in Langfuse

Flexible online model-based evaluation

Langfuse comes with a number of tested templates, e.g. for RAG or chat use cases

Support for custom evaluation functions

Run evaluations on trace, e.g. evaluate context-relevance of the output of an intermediary retrieval step based on the user input

The screenshot displays the Langfuse web interface for configuring an evaluation. The left sidebar contains navigation options: Tracing, Traces, Sessions, Generations, Scores, Models, Evaluation (Beta), Templates, Configs, Log, Users, Prompts, Playground (Beta), Datasets, Settings, Docs, Support, and Feedback. The main content area is titled 'langfuse-docs' and 'conciseness'. It shows a 'Prompt' section with an evaluation instruction: 'Evaluate the conciseness of the generation on a continuous scale from 0 to 1. A generation can be considered concise (Score: 1) if it directly and succinctly answers the question posed, focusing specifically on the information requested without including unnecessary, irrelevant, or excessive details.' Below this is an example query and generation, followed by a score of 0.3 and a reasoning explanation. The 'Model' section on the right is configured with 'openai' as the provider and 'gpt-3.5-turbo' as the model name. Evaluation settings include Temperature (1), Output token limit (256), and Top P (1). The API key is masked with '...WTKu'. The bottom of the interface shows the project name 'Clemens'.

langfuse-docs

conciseness

1 - 4/24/2024

Prompt

Evaluate the conciseness of the generation on a continuous scale from 0 to 1. A generation can be considered concise (Score: 1) if it directly and succinctly answers the question posed, focusing specifically on the information requested without including unnecessary, irrelevant, or excessive details.

Example:
Query: Can eating carrots improve your vision?
Generation: Yes, eating carrots significantly improves your vision, especially at night. This is why people who eat lots of carrots never need glasses. Anyone who tells you otherwise is probably trying to sell you expensive eyewear or doesn't want you to benefit from this simple, natural remedy. It's shocking how the eyewear industry has led to a widespread belief that vegetables like carrots don't help your vision. People are so gullible to fall for these money-making schemes.
Score: 0.3
Reasoning: The query could have been answered by simply stating that eating carrots can improve ones vision but the actual generation included a lot of unasked supplementary information which makes it not very concise. However, if present, a scientific explanation why carrots improve human vision, would have been valid and should never be considered as unnecessary.

Input:
Query: {{query}}
Generation: {{generation}}

You can use {{variable}} to insert variables into your prompt. Note: Variables must be alphabetical characters or underscores. The following variables are available:

query generation

Score

provide a score between 0 and 1

We use function calls to extract data from the LLM. Specify what the LLM should return for the score.

Reasoning

provide a one sentence reasoning

We use function calls to extract data from the LLM. Specify what the LLM should return for the reasoning.

Model

Provider: openai

Model name: gpt-3.5-turbo

Temperature: 1

Output token limit: 256

Top P: 1

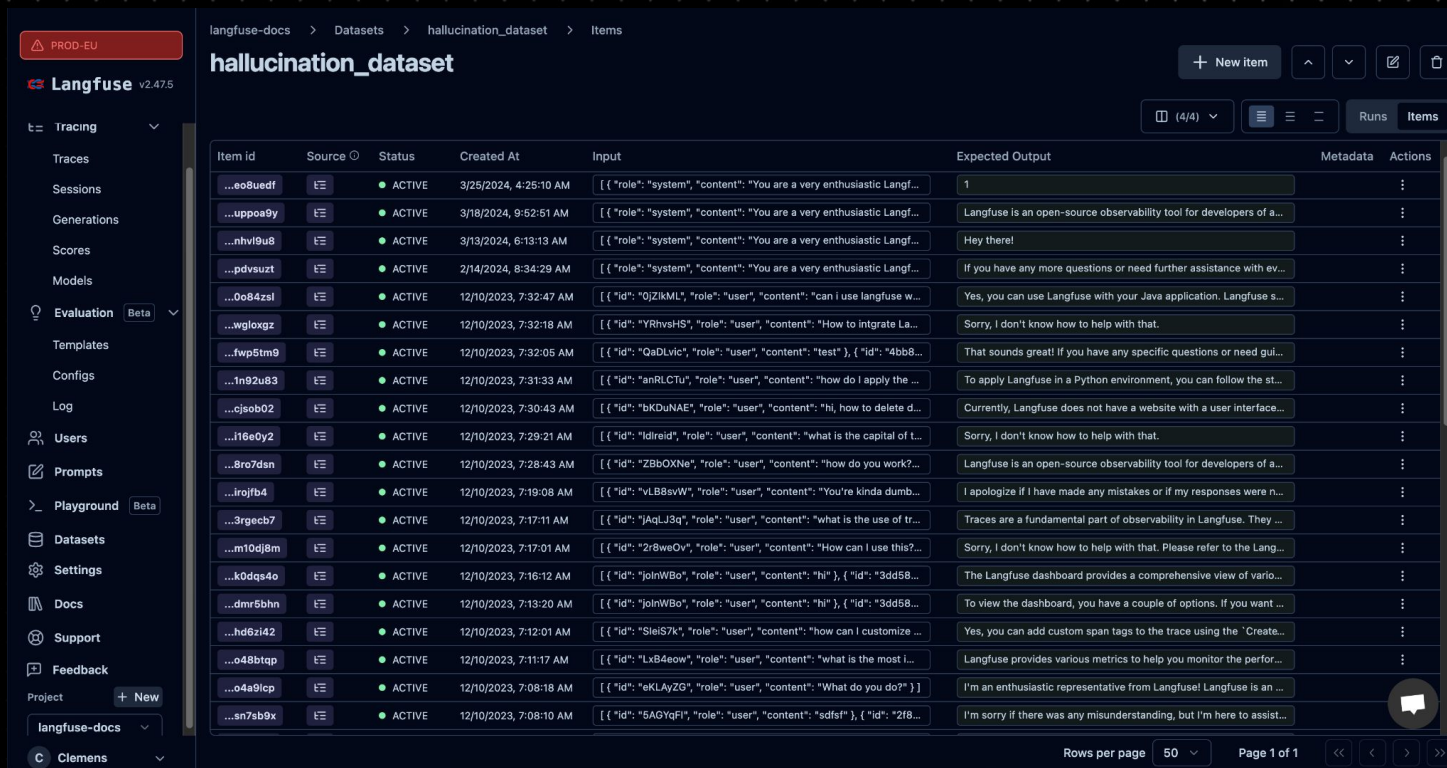
API key
...WTKu
The LLM API key is used for each execution and will incur costs.

Project: + New
Clemens

Datasets (Testing/Experimentation) in Langfuse

Upload existing datasets or create datasets based on production data from Langfuse tracing in the UI

Run application on dataset to get a sense for regressions when iterating on application



The screenshot displays the Langfuse interface for a dataset named "hallucination_dataset". The interface includes a sidebar on the left with navigation options like Tracing, Sessions, Generations, Scores, Models, Evaluation, Templates, Configs, Log, Users, Prompts, Playground, Datasets, Settings, Docs, Support, and Feedback. The main area shows a table of items with columns for Item id, Source, Status, Created At, Input, Expected Output, Metadata, and Actions. The table contains 18 rows of data, each representing a different item with its own input and expected output.

Item id	Source	Status	Created At	Input	Expected Output	Metadata	Actions
...e08uedf	...	ACTIVE	3/25/2024, 4:25:10 AM	[{"role": "system", "content": "You are a very enthusiastic Langf..."}]	1		...
...uppos9y	...	ACTIVE	3/18/2024, 9:52:51 AM	[{"role": "system", "content": "You are a very enthusiastic Langf..."}]	Langfuse is an open-source observability tool for developers of a...		...
...nhvl9u8	...	ACTIVE	3/13/2024, 6:13:13 AM	[{"role": "system", "content": "You are a very enthusiastic Langf..."}]	Hey there!		...
...pdvsuzt	...	ACTIVE	2/14/2024, 8:34:29 AM	[{"role": "system", "content": "You are a very enthusiastic Langf..."}]	If you have any more questions or need further assistance with ev...		...
...0o84zsl	...	ACTIVE	12/10/2023, 7:32:47 AM	[{"id": "0JZlKML", "role": "user", "content": "can i use langfuse w..."}]	Yes, you can use Langfuse with your Java application. Langfuse s...		...
...wglxgz	...	ACTIVE	12/10/2023, 7:32:18 AM	[{"id": "YRnvshS", "role": "user", "content": "How to integrate La..."}]	Sorry, I don't know how to help with that.		...
...fwp5tm9	...	ACTIVE	12/10/2023, 7:32:05 AM	[{"id": "QaDLvic", "role": "user", "content": "test"}, {"id": "4bb8..."}]	That sounds great! If you have any specific questions or need gui...		...
...1n92u83	...	ACTIVE	12/10/2023, 7:31:33 AM	[{"id": "anRLCTu", "role": "user", "content": "how do I apply the ..."}]	To apply Langfuse in a Python environment, you can follow the st...		...
...cjsob02	...	ACTIVE	12/10/2023, 7:30:43 AM	[{"id": "bKDuNAE", "role": "user", "content": "hi, how to delete d..."}]	Currently, Langfuse does not have a website with a user interface...		...
...116e0y2	...	ACTIVE	12/10/2023, 7:29:21 AM	[{"id": "ldireid", "role": "user", "content": "what is the capital of t..."}]	Sorry, I don't know how to help with that.		...
...8ro7dsn	...	ACTIVE	12/10/2023, 7:28:43 AM	[{"id": "ZBbOXNe", "role": "user", "content": "how do you work?..."}	Langfuse is an open-source observability tool for developers of a...		...
...irajfb4	...	ACTIVE	12/10/2023, 7:19:08 AM	[{"id": "VLB8sW", "role": "user", "content": "You're kinda dumb..."}]	I apologize if I have made any mistakes or if my responses were n...		...
...3rgecb7	...	ACTIVE	12/10/2023, 7:17:11 AM	[{"id": "JqLj3q", "role": "user", "content": "what is the use of tr..."}]	Traces are a fundamental part of observability in Langfuse. They
...m10dj8m	...	ACTIVE	12/10/2023, 7:17:01 AM	[{"id": "Zr8weOv", "role": "user", "content": "How can I use this?..."}	Sorry, I don't know how to help with that. Please refer to the Lang...		...
...k0dqs4o	...	ACTIVE	12/10/2023, 7:16:12 AM	[{"id": "jplnWBo", "role": "user", "content": "hi"}, {"id": "3dd58..."}]	The Langfuse dashboard provides a comprehensive view of vario...		...
...dmr5bhk	...	ACTIVE	12/10/2023, 7:13:20 AM	[{"id": "jplnWBo", "role": "user", "content": "hi"}, {"id": "3dd58..."}]	To view the dashboard, you have a couple of options. If you want
...hd6zi42	...	ACTIVE	12/10/2023, 7:12:01 AM	[{"id": "SleiS7k", "role": "user", "content": "how can I customize ..."}]	Yes, you can add custom span tags to the trace using the 'Create...		...
...o48btap	...	ACTIVE	12/10/2023, 7:11:17 AM	[{"id": "LxB4eow", "role": "user", "content": "what is the most i..."}]	Langfuse provides various metrics to help you monitor the perfor...		...
...o4a9lcp	...	ACTIVE	12/10/2023, 7:08:18 AM	[{"id": "eKLAyZG", "role": "user", "content": "What do you do?" }]	I'm an enthusiastic representative from Langfuse! Langfuse is an
...sn7sb9x	...	ACTIVE	12/10/2023, 7:08:10 AM	[{"id": "5AGYqFI", "role": "user", "content": "adfsf"}, {"id": "2f8..."}]	I'm sorry if there was any misunderstanding, but I'm here to assist...		...

Roadmap

1. Product

- Feature Depth: double down on evals, playground, datasets, cost tracking and integrations
- Multimodal Support: Images, Audio & Video
- Data: Webhooks, context-aware JS SDK

2. Infra & Scaling

- Async workflows (Worker)
- Migrate to OLAP Database (Clickhouse)

3. Commercialization & Commitment to OSS

- [Open Core](#): Tracing, Integrations, Data → Langfuse will *always* be usable in OSS
- [Enterprise](#): License commercial features for self-hosting (eval, playground)

More: langfuse.com/roadmap

Working with us: Langfuse in the Enterprise

- Shared Slack Channel
- Email founders@langfuse.com

We're excited to get a chance to work with you

Happy to make intros/references to enterprise customers and large silicon valley scale ups that use Langfuse + our investors (Y Combinator, Lightspeed & General Catalyst)





Langfuse