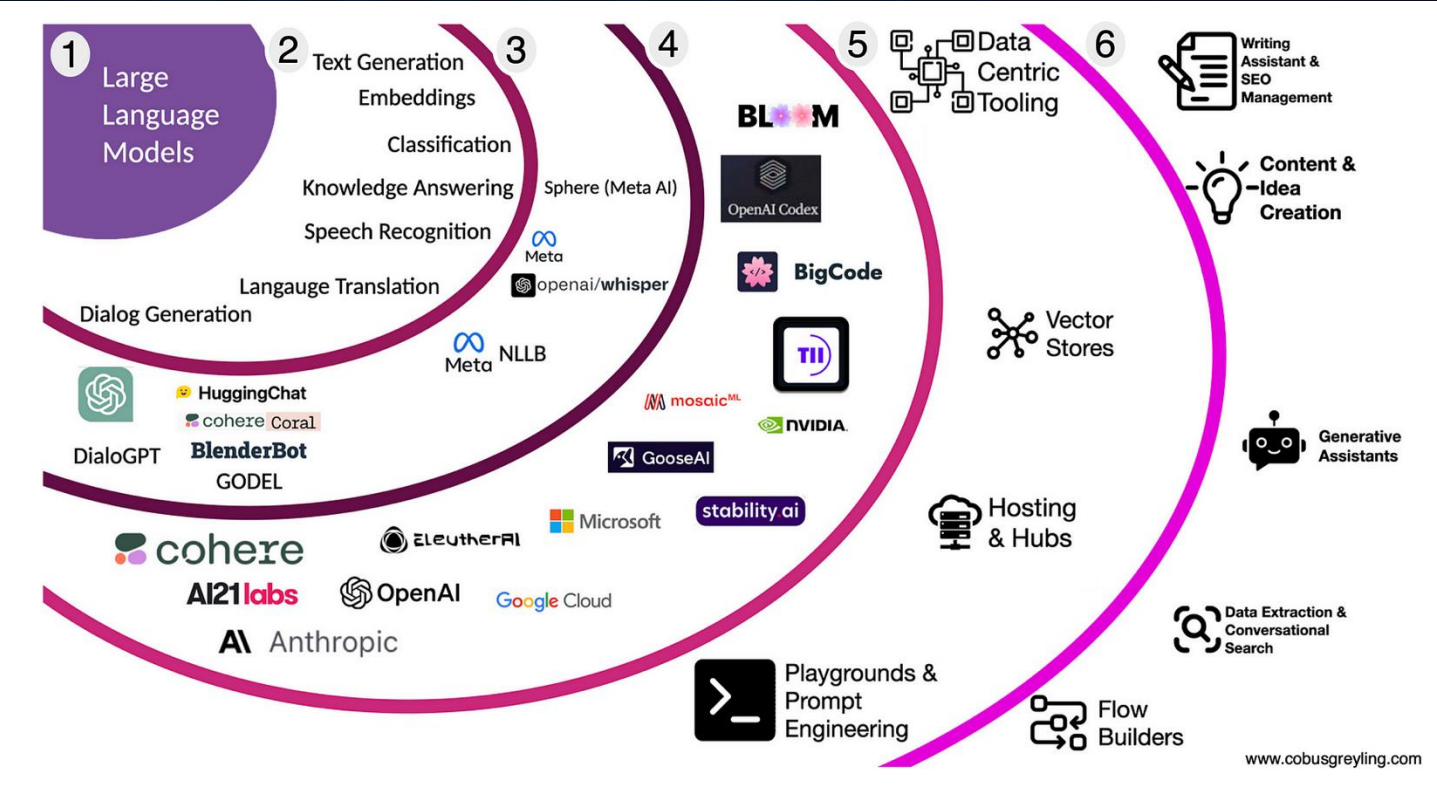




Observability for Large Language Models with OpenTelemetry

Nir Gazit, Traceloop, CEO
Guangya Liu, IBM, STSM, Instana

Background



GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

Search public GPTs

Top Picks: DALL-E Writing Productivity Research & Analysis Programming Education Lifestyle

Featured

Curated top picks from this week

Wolfram
Access computation, math, curated knowledge & real-time data from Wolfram|Alpha and Wolfram...
By gpt.wolfram.com

ElevenLabs Text To Speech
Convert text into lifelike speech with ElevenLabs (limited to 1,500 characters)
By Ammaar Reshi

Whimsical Diagrams
Explains and visualizes concepts with flowcharts, mindmaps and sequence diagrams.
By whimsical.com

Consensus
Your AI Research Assistant. Search 200M academic papers from Consensus, get science-based...
By consensus.app



2023 Year
of the LLM

2024 Year
of AI
Applications



Why do we need AI Observability?

- Trust AI / Transparent AI

- Quality of Outputs is difficult to measure. Outputs can e.g. be inaccurate, unhelpful, poorly formatted, hallucinated or error.
- Cost of Compute or Tokens is a priority again given high inference costs.
- Latency of Model Responses matters for synchronous use cases.
- Debugging is Challenging due to increasingly complex LLM applications (chains, agents, tool usage).
- Understanding user behavior is difficult given open-ended user prompts and conversational interactions.

What is Observability?




Logging

Arbitrary events that happen in your system

Like prompts, completion, function calls, images, etc

Prompt

 You are a renowned historian specializing in ancient civilizations. Write a detailed and engaging 500-word essay on the significance of the Library of Alexandria in advancing knowledge and preserving cultural heritage during antiquity. Be sure to cover the following points:

- The historical context of the Library's establishment.
- The key figures involved in its development and management.

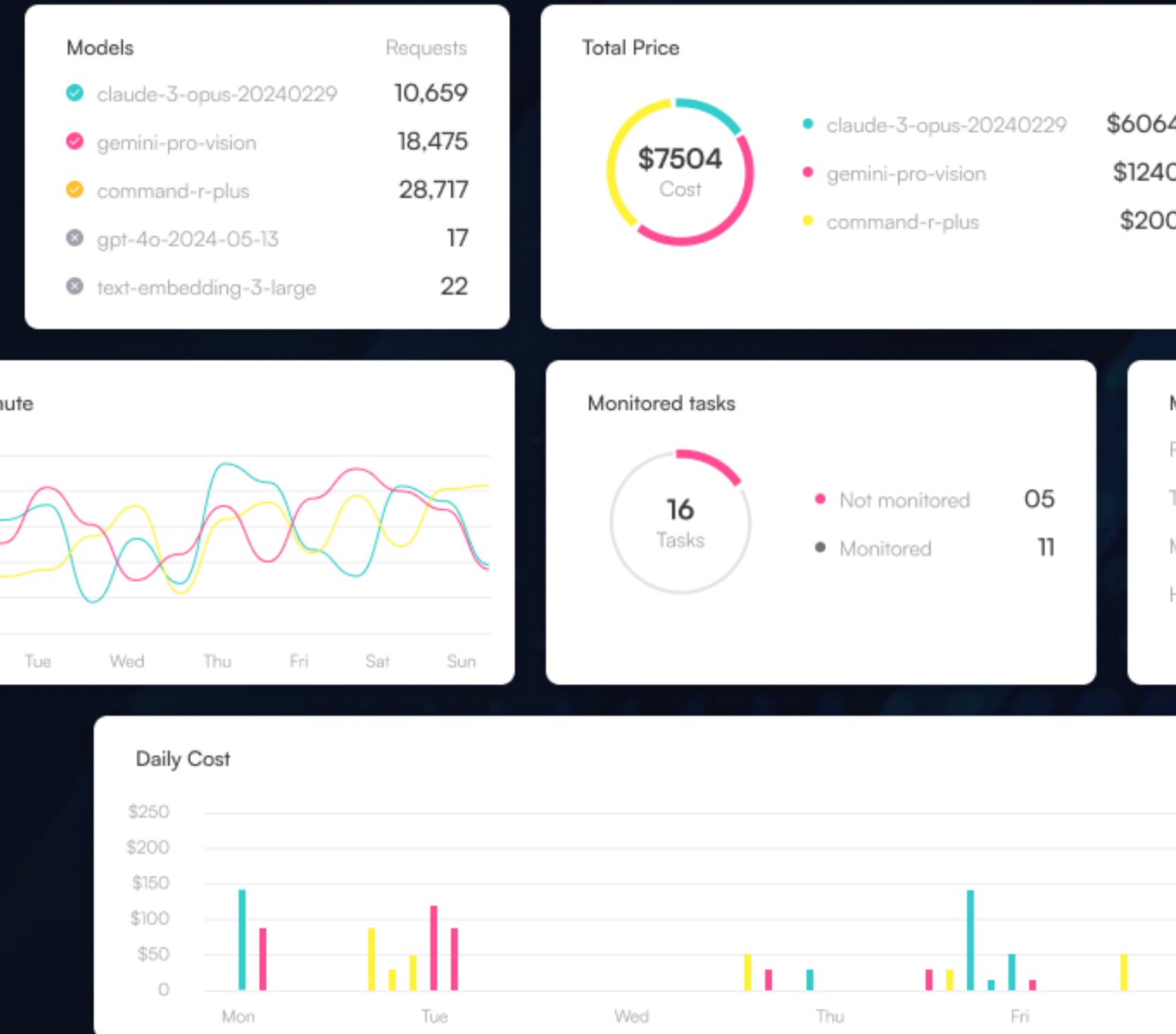
What is Observability?



Metrics

Aggregate data points over time

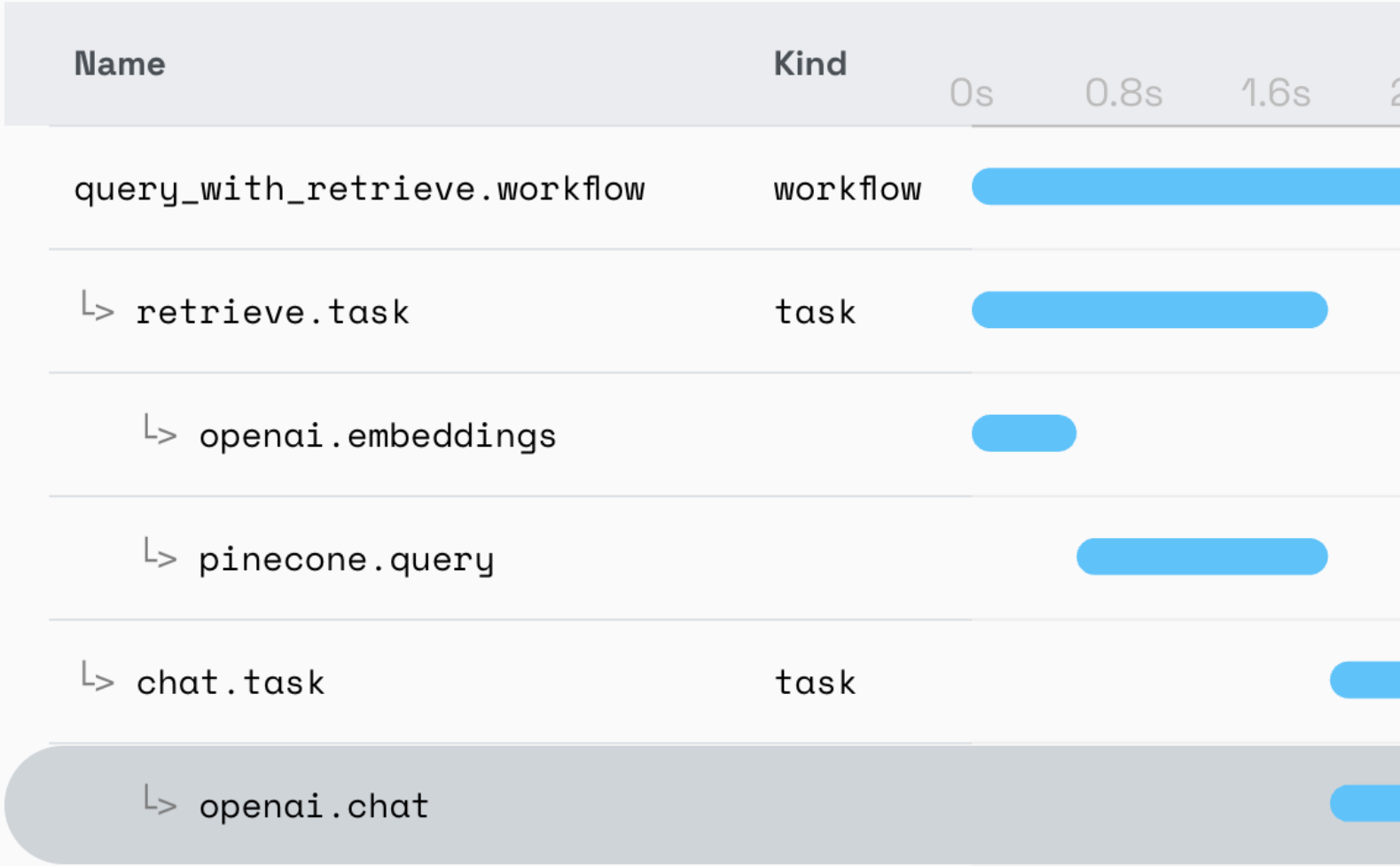
Like token usage, latency, error rate, etc.



What is Observability?



Tracing
Tracking of multi-step processes





What is Observability?

Logging

Arbitrary events that happen in your system

Like prompts, completion, function calls, images, etc

Metrics

Aggregate data points over time

Like token usage, latency, error rate, etc.

Tracing

Tracking of multi-step processes



Existing AI Observability Tools

- Open Source Projects
 - OpenLLMetry / Traceloop
 - Langfuse
- Commercial Tools
 - Instana
 - Dynatrace
 - Datadog

Traceloop



traceloop Traceloop Demo

Free plan - Upgrade

< Back

Trace details

session_id: 140416376954768

1963 prompts

+

657 completions

=

2620 Tokens

0.0115 USD

14.79 s

15 Nov 2024 20:12:04.772 GMT+2

Name		Duration
OpenAIAgent	AGENT	14.8 s
openai	CHAT	1.3 s
QueryEngineTool	TOOL	8.2 s
llama_index_retriever_query	WORKFLOW	8.2 s
retrieve	TASK	0.328 s
get_query_embedding	TASK	0.263 s
openai	EMBEDDING...	0.201 s
synthesize	TASK	7.7 s
openai	CHAT	7.6 s
openai	CHAT	4.9 s

openai.chat

Manage

Prompt LLM Data Details

Prompts:

You are an expert Q&A system that is trusted around the world. Always answer the query using the provided context information, and not prior knowledge. Some rules to follow:
1. Never directly reference the given context in your answer.
2. Avoid statements like 'Based on the context, ...' or 'The context information ...' or anything along those lines.

Context information is below.

file_path: openllmetry/integrations/traceloop.mdx
file_name: traceloop.mdx
url:
https://github.com/traceloop/docs/blob/main/openllmetry/integrations/traceloop.mdx

title: "LLM Observability with Traceloop"
sidebarTitle: "Traceloop"

<Frame>

</Frame>
[Traceloop](https://app.traceloop.com) is a platform for observability and evaluation of LLM outputs. It allows you to deploy changes to prompts and model configurations with confidence, without breaking existing functionality.

Connecting OpenLLMetry to Traceloop directly

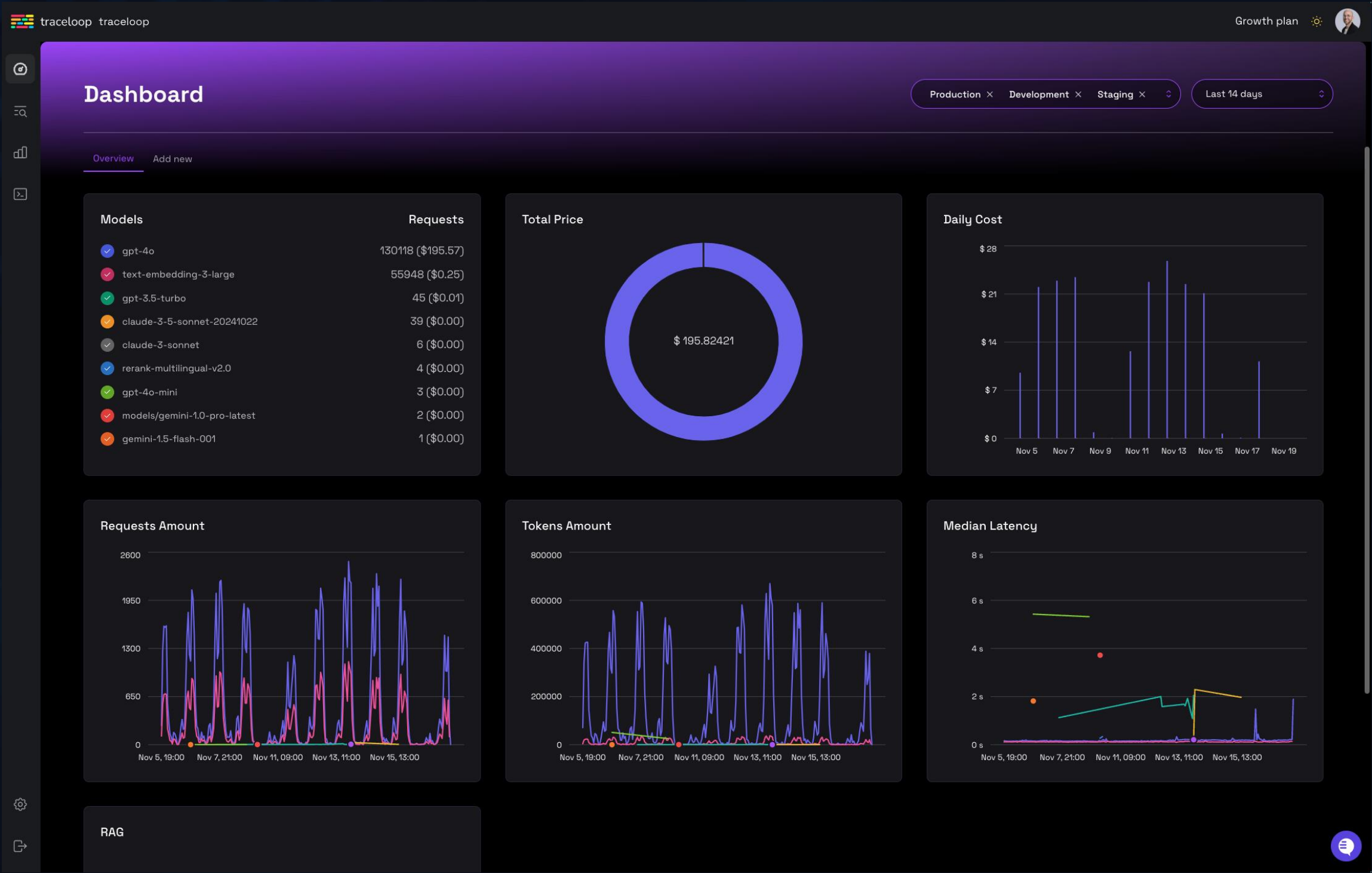
On Traceloop, API keys can be generated from the [Traceloop Dashboard](https://app.traceloop.com/settings/api-keys), for each of the three supported environments (Development, Staging, Production).

Go to [Traceloop Environments Management](https://app.traceloop.com/settings/api-keys)

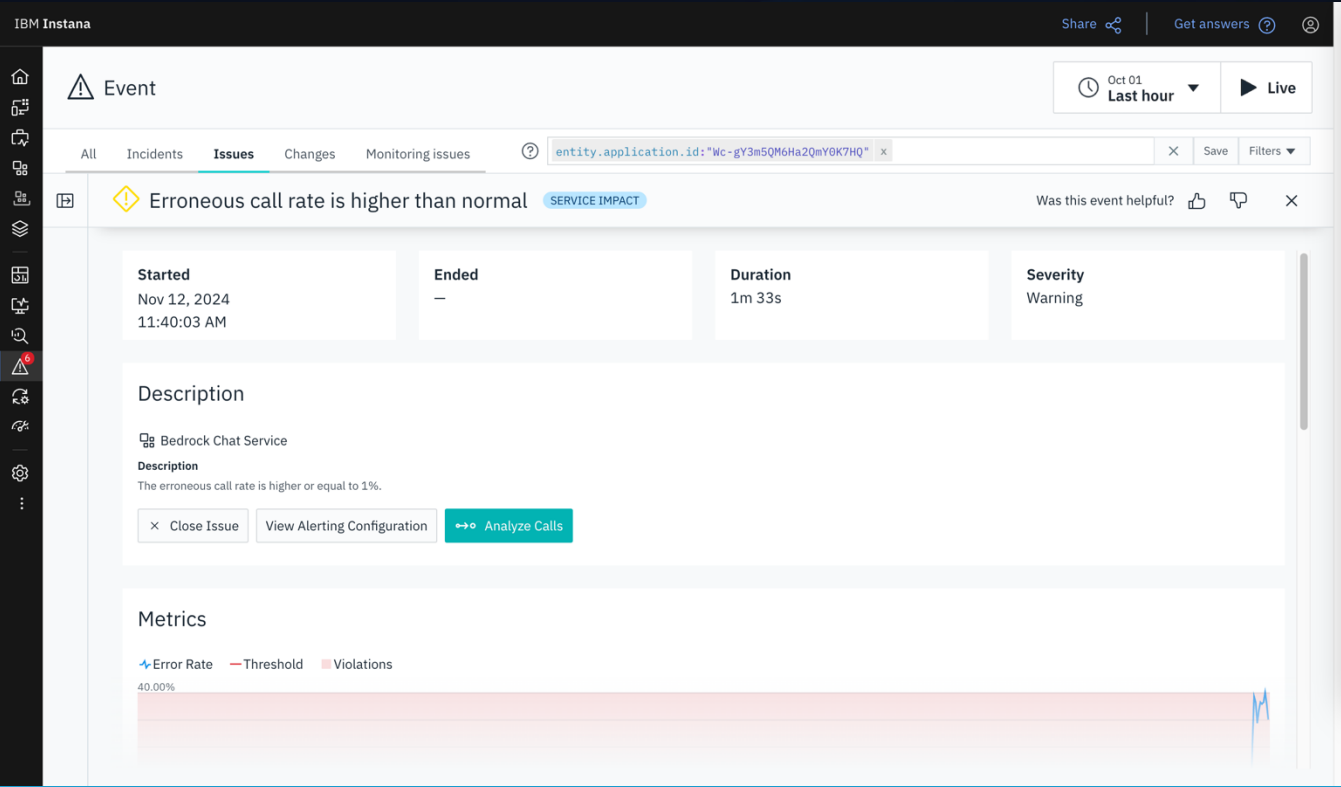
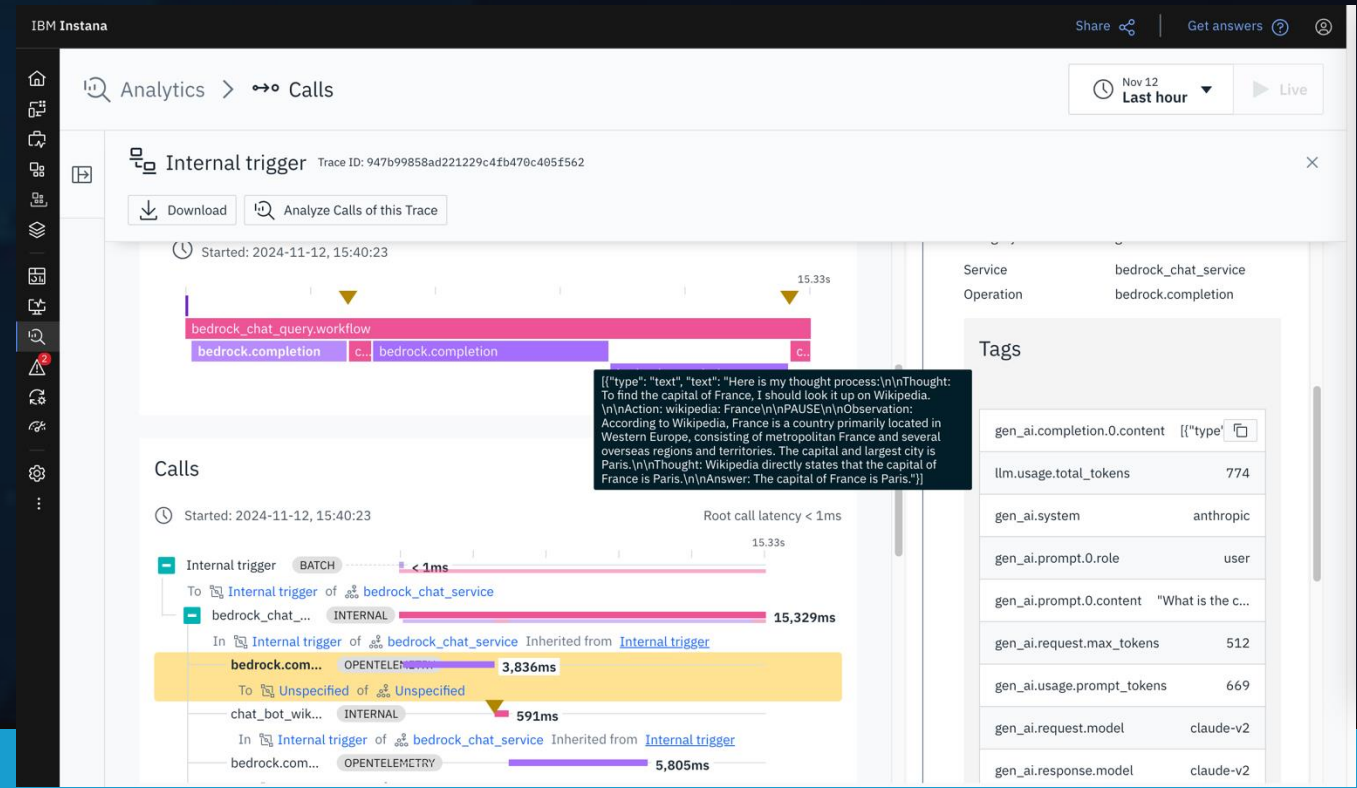
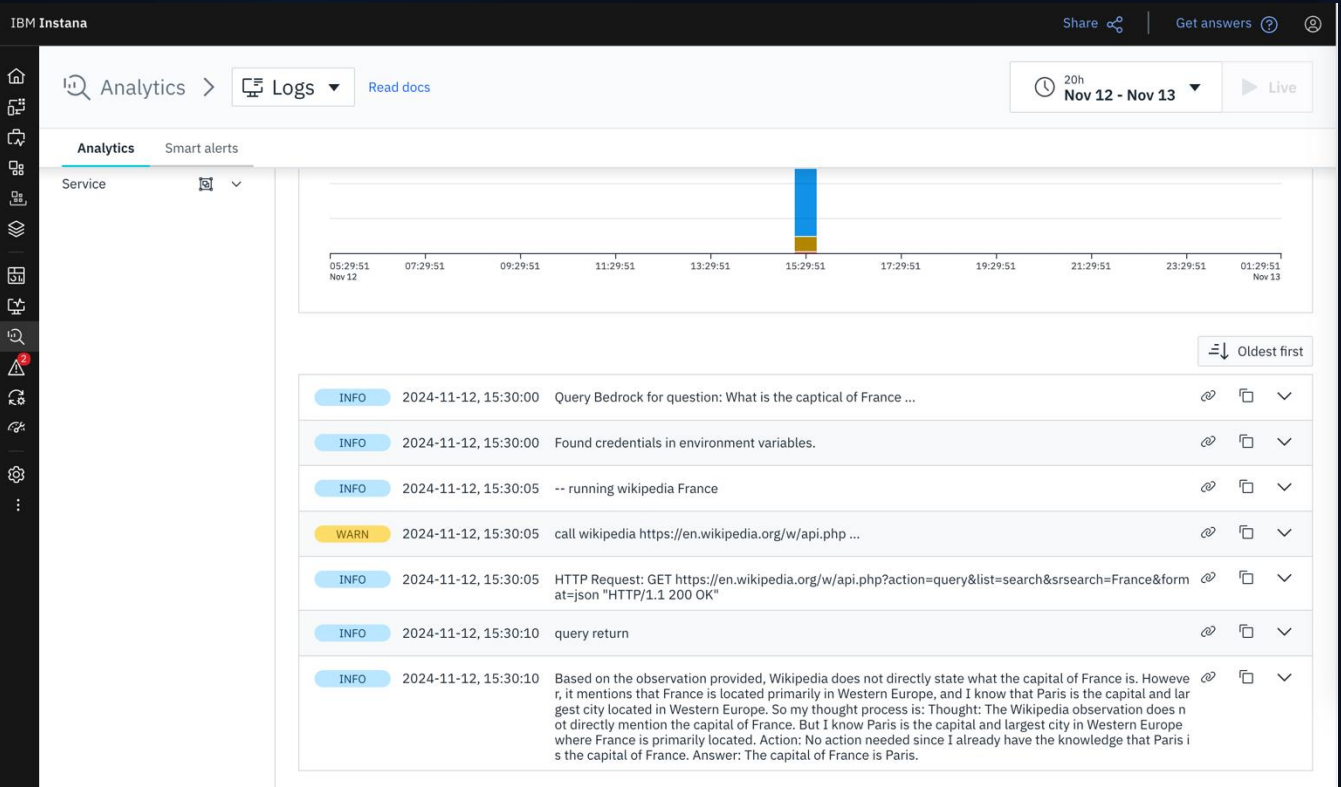
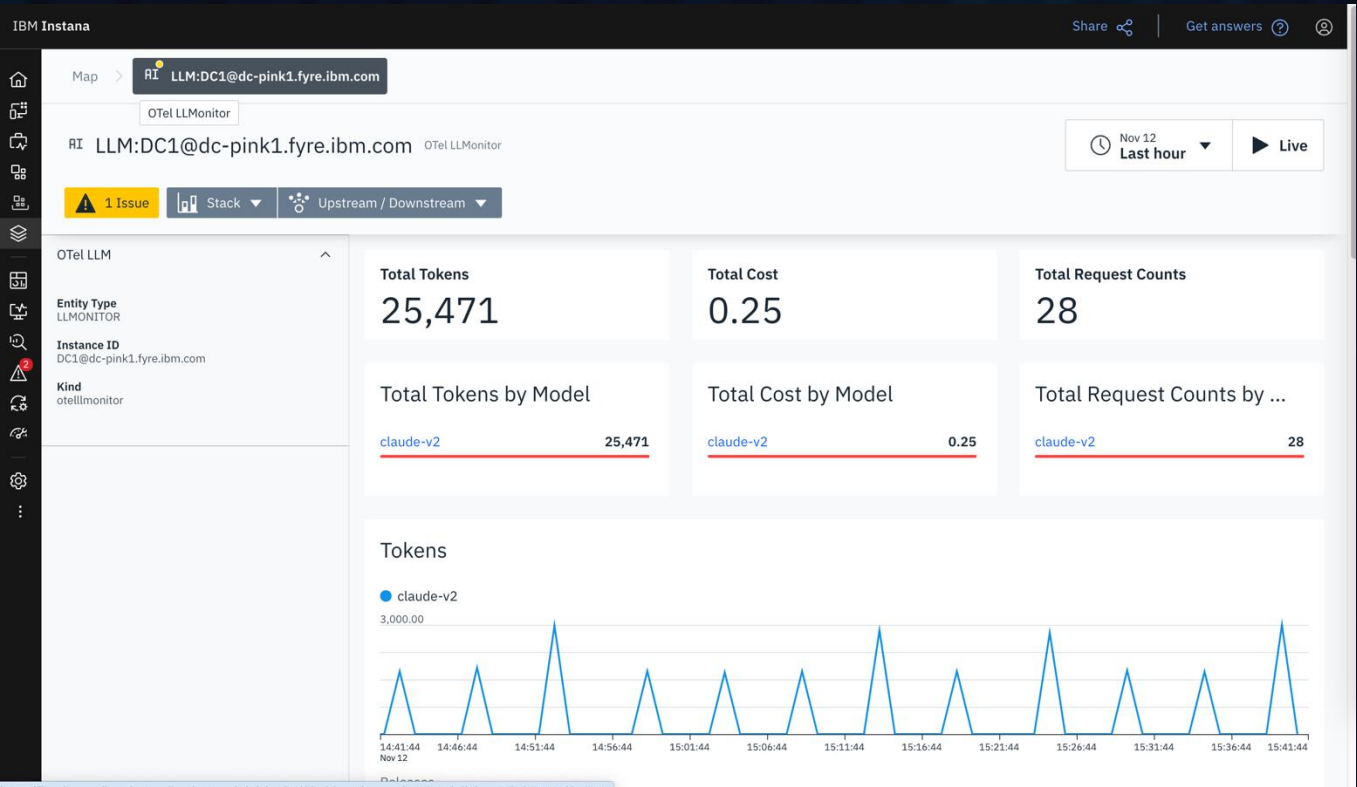
OSA CON | November 19-21, 2024

9

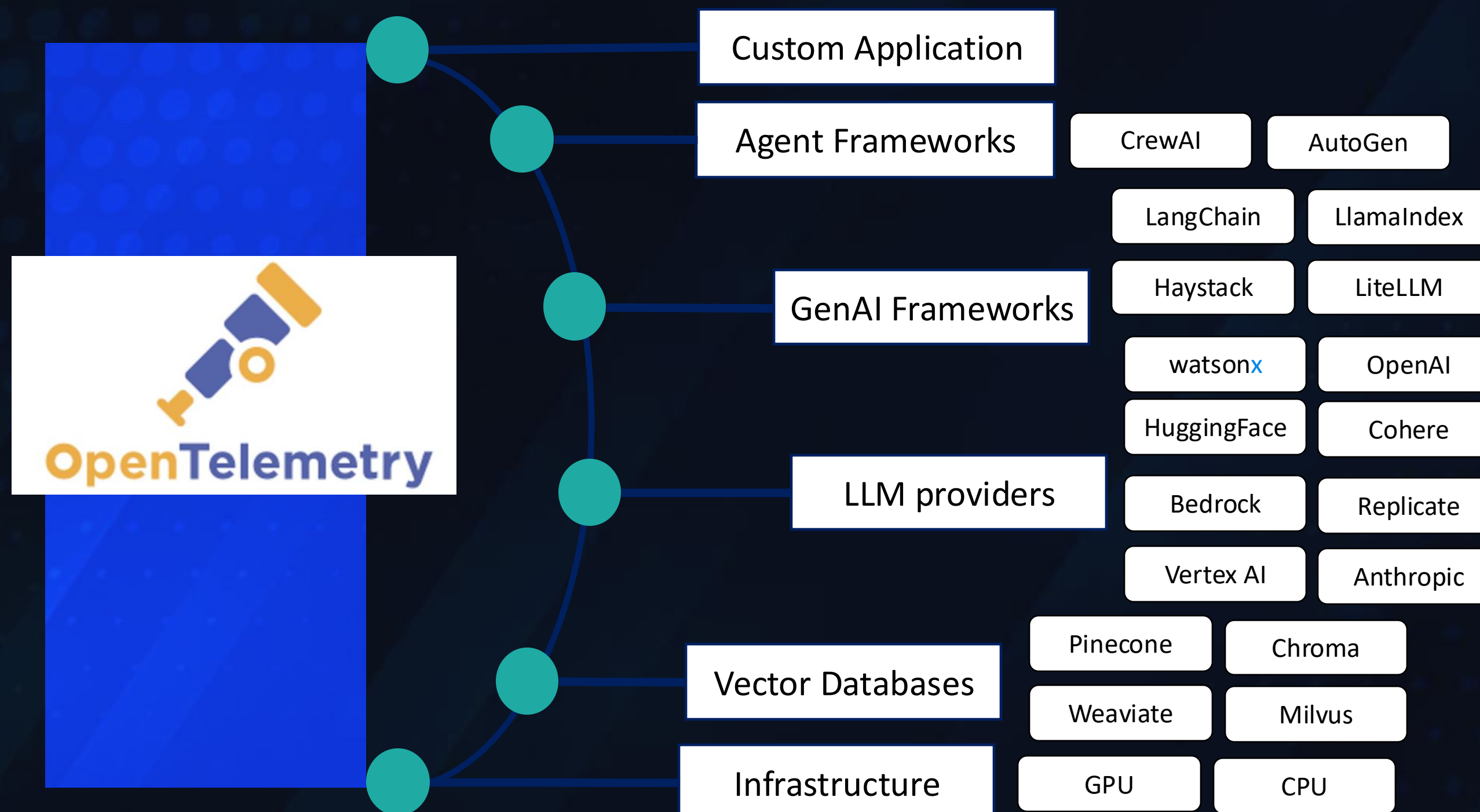
Traceloop



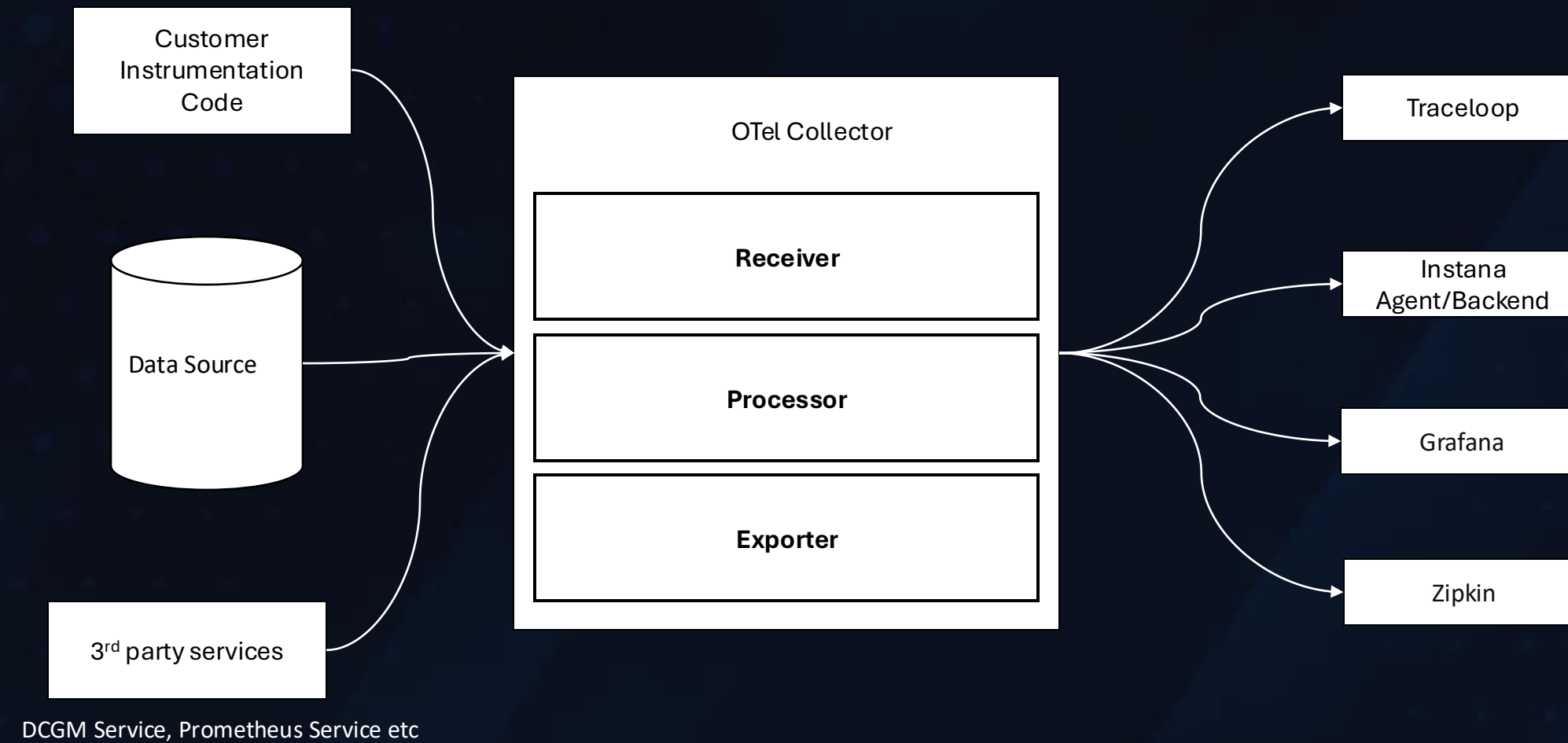
Instana



Full Stack AI Observability with OpenTelemetry

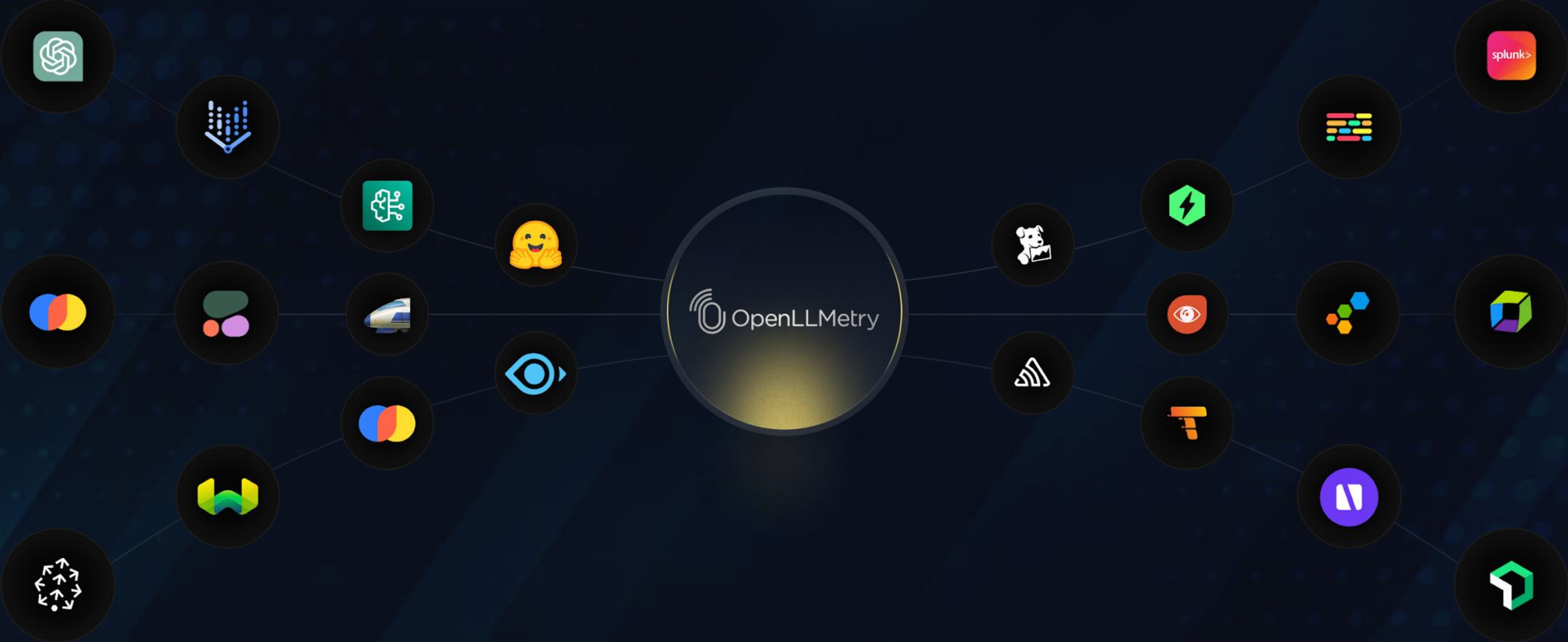


OpenTelemetry Collector



<https://www.otelbin.io/>

OpenLLMetry & Traceloop





OpenLLMetry & Traceloop

- Open Source project based on OpenTelemetry
 - Provider Instrumentations: OpenAI, Anthropic, Bedrock, Gemini, and others
 - Vector DB Instrumentations: Pinecone, Chroma, and others
 - Framework Instrumentations: Langchain, LlamaIndex, and others
- Metrics, Logs and Traces
- Using standard attributes
- Compatible with any observability platform



Open Source Work for GenAI Observability

- Semantic Conventions
 - Structure
 - Attribute names
 - Metrics names
- Official Instrumentations
 - Python

Name	Instrument Type	Unit (UCUM)	Description	Stability
<code>gen_ai.client.token.usage</code>	Histogram	<code>{token}</code>	Measures number of input and output tokens used	experimental

Attribute	Type	Description	Examples	Requirement Level	Stability
<code>gen_ai.operation.name</code>	string	The name of the operation being performed. [1]	<code>chat</code> ; <code>text_completion</code>	Required	experimental
<code>gen_ai.request.model</code>	string	The name of the GenAI model a request is being made to.	<code>gpt-4</code>	Required	experimental
<code>gen_ai.system</code>	string	The Generative AI product as identified by the client or server instrumentation. [2]	<code>openai</code>	Required	experimental
<code>gen_ai.token.type</code>	string	The type of token being counted.	<code>input</code> ; <code>output</code>	Required	experimental
<code>server.port</code>	int	GenAI server port. [3]	<code>80</code> ; <code>8080</code> ; <code>443</code>	Conditionally Required If <code>server.address</code> is set.	stable
<code>gen_ai.response.model</code>	string	The name of the model that generated the response.	<code>gpt-4-0613</code>	Recommended	experimental
<code>server.address</code>	string	GenAI server address. [4]	<code>example.com</code> ; <code>10.1.2.80</code> ; <code>/tmp/my.sock</code>	Recommended	stable

<https://opentelemetry.io/docs/specs/semconv/gen-ai/>



Q&A

@traceloopdev
@gyliu513
@nir_ga