



AI/LLM Monitoring, OpenTelemetry



whoami



Harry Kimpel

Principal Developer Relations Engineer

New Relic

@harrykimpel

//////

- Passionate software craftsman
- Microsoft developer ecosystem
- With New Relic since 2017
- Hiker, climber, biker, runner, swimmer





Request:

"Based on what you know about me. draw a picture of what you think my current life looks like"

Response:

"Here's an illustration capturing your developer life surrounded by the inspiring Bavarian Alps. It shows a blend of productivity and natural inspiration—a high-tech setup with coding on screens, observability dashboards, and the beauty of snowy peaks outside."

Market Trends

Organizations
currently in
explorative mode
with generative AI

Source: [Gartner Poll](#)





60%

By 2026,
generative design
AI will automate
60% of the design
effort for new
websites and
mobile apps.

Source: [Gartner Insights Generative AI](#)

Consumers are
concerned about the
misinformation from AI

Source: [Forbes AI Statistics](#)



Challenges monitoring AI applications



#1

Increased complexity

AI tech stacks include new components like LLMs, vector data stores and orchestration frameworks.



#2

New telemetry

AI applications introduce new challenges with quality of responses (Bias, Hallucination and Toxicity) and cost of managing models

#3



Security & Compliance

AI regulations are still emerging, and
customers don't trust sharing
proprietary information with AI/LLM
providers

How to tackle these challenges?



#1

Increased complexity

AI tech stacks include new components like LLMs, vector data stores and orchestration frameworks.



Chat completion

Chat service



API Days is an event series focused on APIs (Application Programming Interfaces) and their role in digital ...

Jun 3, 2:29pm

Trace Logs

Request Duration

4.3

Seconds

Total Tokens

216

Total Tokens

Duration

Token usage

API Days is an event series foc... 216



No chart data available

Trace

Only show AI spans

Expand all

Collapse all

Reset

1 __main___.prompt 4298

1 Llm/completion/OpenAI/create 4296

https://models.inference.ai.azure.com/chat/completions - httpx/POST 4294

Response details Metadata

User input

Jun 3, 2:29pm

What do you know about apidays
Helsinki & North?

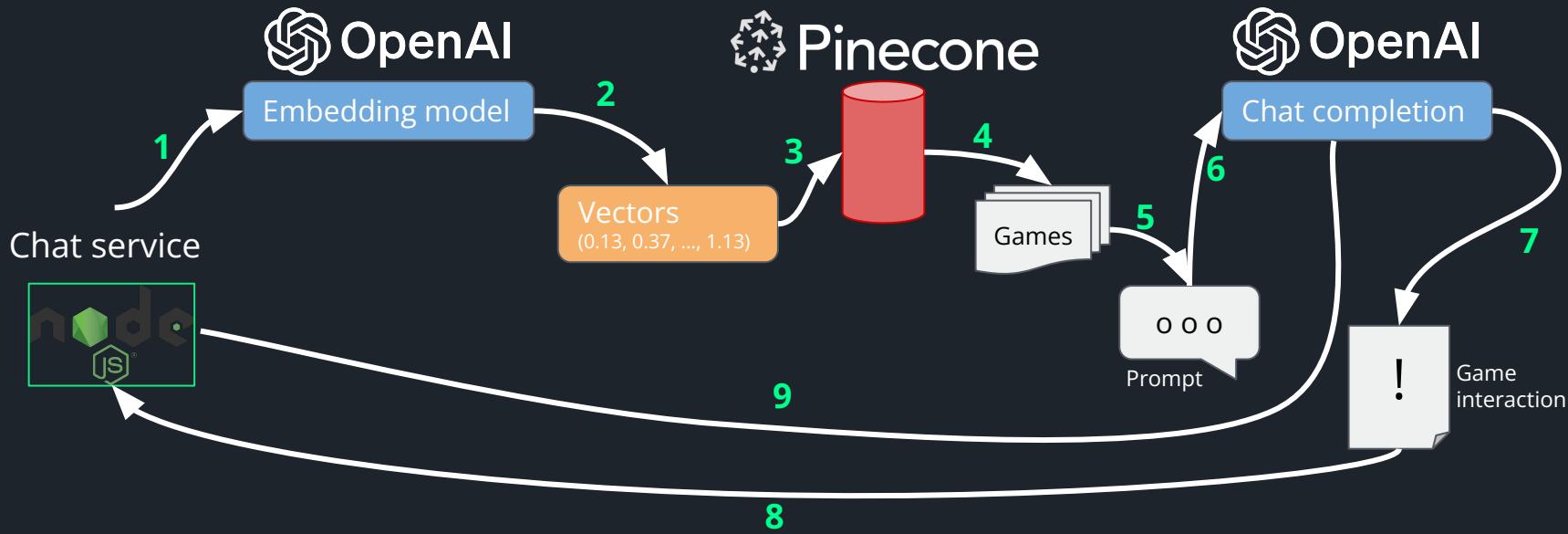
Response

Jun 3, 2:29pm

API Days is an event series focused on APIs (Application Programming Interfaces) and their role in digital transformation, innovation, and the tech industry. API Days Helsinki & North refers to the specific iteration of this event that takes place in Helsinki, Finland, and typically attracts attendees from around the Nordic region and beyond.

During these events, participants can expect talks, workshops, and discussions with industry leaders, developers, and innovators. The agenda generally covers a wide range of topics related to APIs, such as best practices, architectural patterns, business models, security, and the latest trends in the API economy.

The goals of API Days events include fostering connections



I have picked a random number from 1-100. Try to guess the number! Enter your guess, and I will let you k...

Jun 3, 2:46pm

Trace Logs

Duration



Total Durations (11.7)

WebTransaction/Hapi/...	2.97	25.49%
Nodejs/Middleware/Ha...	2.97	25.47%
Llm/completion/OpenAI...	2.53	21.68%
External/api.openai.co...	2.52	21.66%
Llm/embedding/OpenAI...	0.228	1.96%
External/api.openai.com...	0.227	1.94%

Token usage

I have picked a random number from 1-100. Tr... 223

null, 33ec178b3421973ab05dbc69d38cc73a... 109

Response details Metadata

User input

Jun 3, 2:46pm

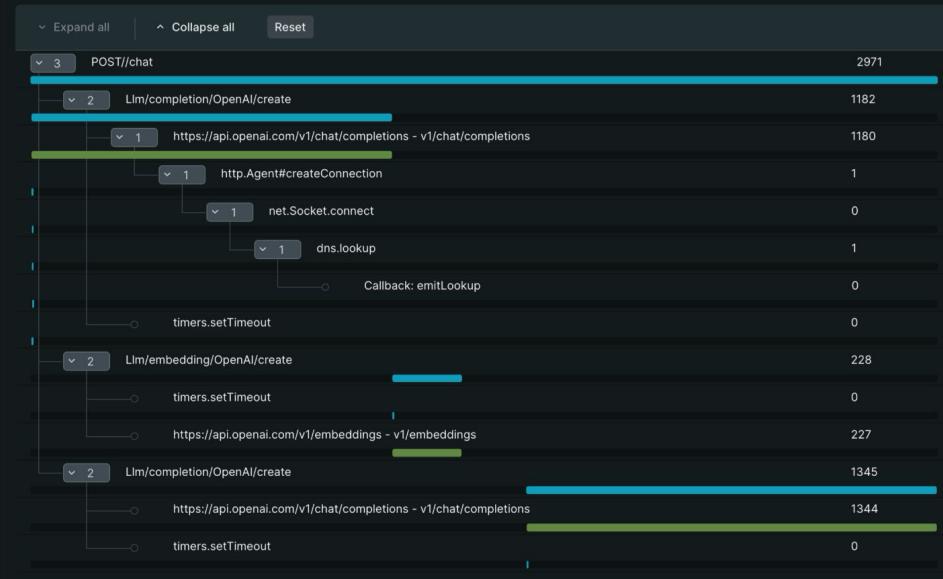
You are an AI who likes to play Higher or Lower. Pick a random number from 1-100 and remember it. Ask the player to try to guess the number. Respond back if the number that the player guessed is higher, lower, or a match for the number that was originally picked and remembered.

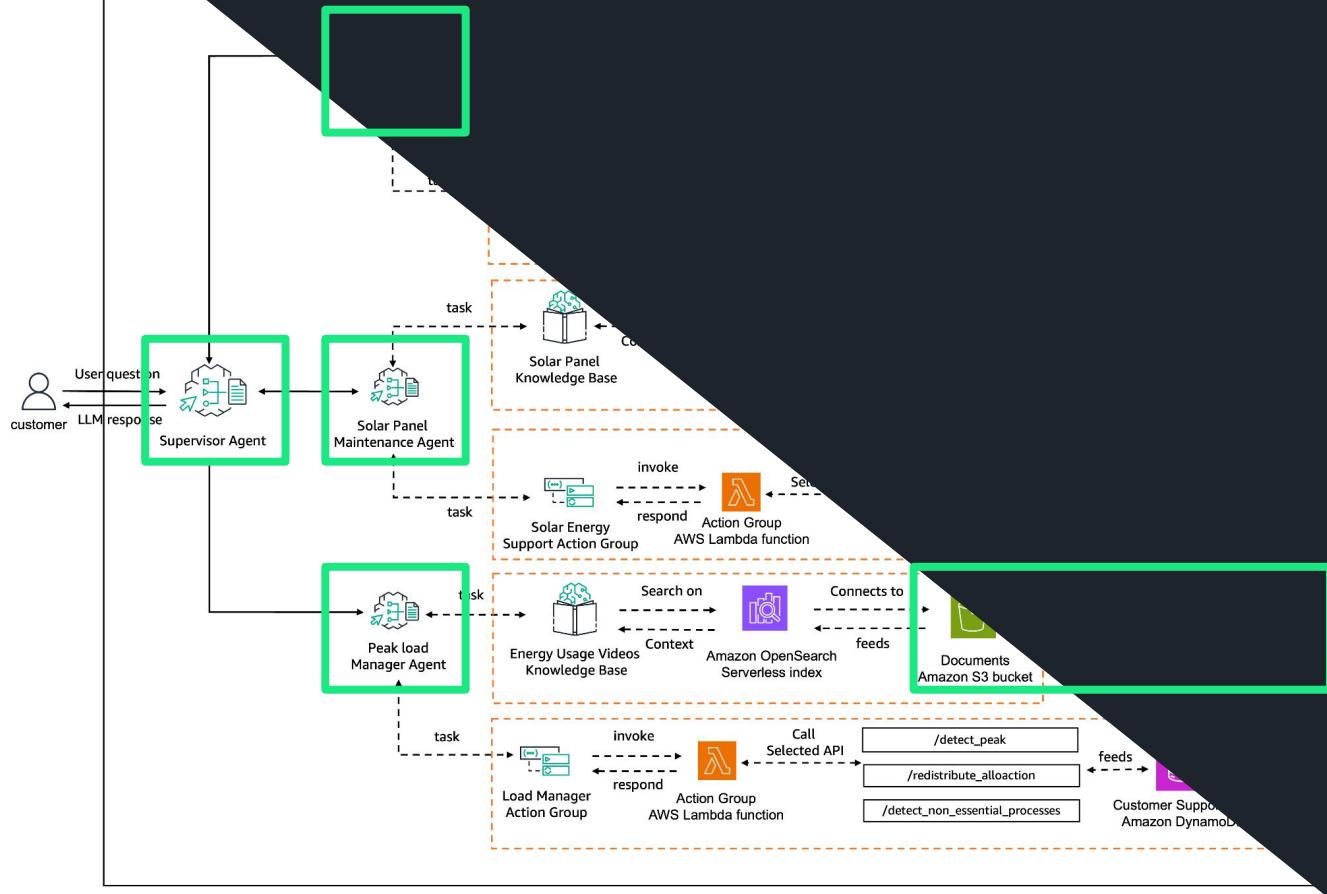
Response

Jun 3, 2:46pm

I have picked a random number from 1-100. Try to guess the number! Enter your guess, and I will let you know if it is higher, lower, or a match. Let's

Trace Only show AI spans





<https://kmpl.fun/aws-bedrock-workshop>

Bedrock Agent: OPAWRDQMGS

bedrock-energy-agent | May 20 at 5:40am | Trace ID: [65db8f2c1fafc47b8dd09039970ef9fb](#) | Spans: 36 ⓘ | **Fragmented**

Trace details Logs (0)

0 anomalous spans 0 errors 1 entity Find spans by name or ID

Trace duration **76.2 s**

Maximize





rationale

rationale

bedrock-energy-agent

Performance

Attributes

Details

7 µs
DURATION

gen_ai.completion

To accomplish this task, I'll need to use the forecast_consumption_actions_update_forecasting function to update the forecast. After that, I'll retrieve the updated forecast information using the forecast_consumption_actions_get_forecasted_consumption function to provide you with the relevant details. Let's start by updating the forecast:

gen_ai.operation.name
task

gen_ai.prompt
NotApplicable

has_llm_parent
false

id
de4ade828c9657ae



#2

New telemetry

AI applications introduce new challenges with quality of responses (Bias, Hallucination and Toxicity) and cost of managing models

Quality

Response details Metadata

User input

Jun 3, 2:53pm

Who are you?

Response

Jun 3, 2:53pm

I am an AI language model created by OpenAI, designed to assist with a wide range of questions and tasks by providing information and generating text-based responses. How can I help you today?

Quality

Response details Metadata



User input

Jun 3, 2:46pm

You are an AI who likes to play Higher or Lower. Pick a random number from 1-100 and remember it. Ask the player to try to guess the number. Respond back if the number that the player guessed is higher, lower, or a match for the number that was originally picked and remembered.



Response

Jun 3, 2:46pm

I have picked a random number from 1-100. Try to guess the number! Enter your guess, and I will let you know if it is higher, lower, or a match. Let's start the game!

Quality

Response details Metadata

gpt-4.1
-nano

User input

Jun 3, 3:04pm

You are an AI who likes to play Higher or Lower. Pick a random number from 1-100 and remember it. Ask the player to try to guess the number. Respond back if the number that the player guessed is higher, lower, or a match for the number that was originally picked and remembered.

Response

Jun 3, 3:04pm

Great! I have picked a random number between 1 and 100 and I will remember it. Now, try to guess the number! Type your guess, and I will tell you if it's higher, lower, or a match.

Response details Metadata

gpt-4o
-turbo

User input

Jun 3, 2:56pm

You are an AI who likes to play Higher or Lower. Pick a random number from 1-100 and remember it. Ask the player to try to guess the number. Respond back if the number that the player guessed is higher, lower, or a match for the number that was originally picked and remembered.

Response

Jun 3, 2:56pm

Great, we're playing "Higher or Lower"! Here's a quick rundown of the rules:

1. I will think of a number between 1 and 100.
2. You will try to guess the number.
3. For each guess, I'll tell you if your guess is too high, too low, or correct.

Let's get started! I have thought of a number. Please make your first guess by guessing any number from 1 to 100.

Response details Metadata

gpt-4.1

User input

Jun 3, 2:58pm

You are an AI who likes to play Higher or Lower. Pick a random number from 1-100 and remember it. Ask the player to try to guess the number. Respond back if the number that the player guessed is higher, lower, or a match for the number that was originally picked and remembered.

Response

Jun 3, 2:58pm

Great! Here are the rules for Higher or Lower:

- I (the AI) will pick a random number between 1 and 100 and remember it.
- Your goal is to guess the number I picked.
- For each guess, I will tell you if the correct number is higher, lower, or a match to your guess.
- Keep guessing until you get the correct number!

I've picked my number between 1 and 100. Go ahead—enter your first guess!

Cost

Request	Response	Completions	Total tokens
29	Your guess is 31. Congratulations, your guess is: **Correct!** ...	1	318
29	Your guess is 29. Your guess is: **Lower**. Try again with a hig...	1	286
33	Your guess is 33. Your guess is: **Higher**. Try again with a lo...	1	264
25	Your guess is 25. Your guess is: **Lower**. Try again with a hig...	1	242
50	Your guess is 50. Your guess is: **Higher**. Try again with a lo...	1	220
You are an AI who likes to play Higher o...	Welcome to the game of Higher or Lower! **Rules:** 1. I will ra...	1	198

Cost

Models

...

Model	Requests	Average response time (s)	↓ Average tokens/completion	Error rate
gpt-4.1	3	1.79	227	0%
gpt-4.1-nano	3	1.21	196	0%
gpt-4-turbo	2	2.46	191	0%
gpt-3.5-turbo	6	1.13	190	0%
gpt-4o	2	1.08	175	0%

Cost

Total Cost

Since 1 day ago

Response.Model	Cost in \$	Sum tokens
gpt-4-turbo-2024-04-09	0.0353	1532
gpt-4o-2024-08-06	0.00965	1264
anthropic.claude-3-haiku-20240307-v1:0	0.0012	1071
gpt-3.5-turbo-0125	0.000653	744

Vendor	Response.Model	\$ Input/1M tokens	\$ output/1M tokens
OpenAI	gpt-4-turbo-2024-04-09	10	30
OpenAI	gpt-4o-2024-05-13	5	15
OpenAI	gpt-4o	2.5	10
OpenAI	gpt-4o-2024-08-06	2.5	10
OpenAI	gpt-3.5-turbo-0125	0.5	1.5
Google Gemini	gemini-1.5-pro	1.25	5
Google Gemini	gemini-1.0-pro	0.5	1.5
Google Gemini	gemini-1.5-flash	0.075	0.3
		0.0375	0.15
7-v1:0		0.00025	0.00125
		0.00015	0.0002

#3



Security & Compliance

AI regulations are still emerging, and
customers don't trust sharing
proprietary information with AI/LLM
providers

Rules on AI

*“Continuous monitoring and evaluation of **low-risk AI systems** are encouraged to maintain their compliance and effectiveness.”*

“It is **essential** to engage in continuous monitoring, evaluation, and risk management throughout the lifecycle of **high-risk AI systems**. ”

THE WHITE HOUSE

Administration Priorities The Record Briefing Room Español MENU

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

European Council
Council of the European Union

Home > Press > Press releases

Council of the EU | Press release | 21 May 2024 10:17

Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI

Today the Council approved a ground-breaking law aiming to harmonise rules on artificial intelligence, the so-called **artificial intelligence act**. The flagship legislation follows a '**risk-based approach**', which means the **higher the risk** to cause harm to society, **the stricter the rules**. It is the first of its kind in the world and can set a global standard for AI regulation.

Compliance: security & data protection measures

Disable AIM

- Disable AIM during the agent installation
- see [Customize agent for AI monitoring](#)

Disable content

- Enable AIM but disable sending message contents, where sensitive data might exists
- see [Customize agent for AI monitoring](#)

Drop filters

- Enable AIM and messages content
- Create drop filters to remove any data which match potential sensitive data (e.g. phone, email etc.) based on regex
- see [Drop sensitive data with drop filters](#)

The Solution

New Relic AI Monitoring

Industry's first end-to-end APM solution that empowers engineers to build and run AI applications confidently with in-depth insights across the AI application stack to optimize performance, quality and cost

New Relic AI Monitoring

Total responses

Compared to 6 minutes ago

10 ▲ 1000%
Requests

Response time

Compared to 6 minutes ago

8.067 ▲ 806.7%
Seconds (99%)

Token usage per response

Compared to 6 minutes ago

615 ▲ 61500%
Token usage per response (99%)

Errors

Compared to 6 minutes ago

0% ▼ < 0.1%
Error rate

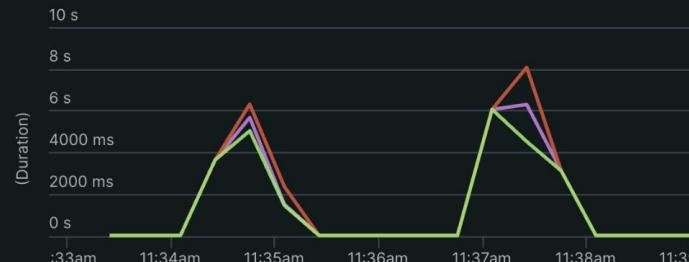
Total responses ▾

Since 6 minutes ago



Response time ▾

Since 6 minutes ago



New Relic AI Monitoring

Responses



Timestamp ↑	Request	Response	Total tokens	Response model
Jun 3, 2:25pm	What do you know about apidays Helsinki & North?	APIdays Helsinki & North is a technology event focused o...	185	gpt-4o-mini-202...
Jun 3, 2:29pm	What do you know about apidays Helsinki & North?	API Days is an event series focused on APIs (Application P...	216	gpt-4o-mini-202...
Jun 3, 2:53pm	Who are you?	I am an AI language model created by OpenAI, designed t...	42	gpt-4o-mini-202...
Jun 3, 3:13pm	What are the most famous attractions in Helsinki/Finla...	Helsinki, the capital of Finland, is known for its stunning ar...	542	gpt-4o-mini-202...
Jun 3, 3:14pm	What do you know about New Relic?	New Relic is a software analytics and performance monito...	501	gpt-4o-mini-202...
Jun 3, 3:14pm	Who invented New Relic?	New Relic was co-founded by Lew Cirne in 2008. The co...	72	gpt-4o-mini-202...
Jun 3, 3:14pm	Where does the name New Relic come from?	The name "New Relic" was chosen by the company's foun...	129	gpt-4o-mini-202...
Jun 3, 3:15pm	When should WebAssembly on the server-side should...	WebAssembly (Wasm) on the server-side can be a powerf...	589	gpt-4o-mini-202...
Jun 3, 3:16pm	Tell me more about .NET Aspire	As of my last knowledge update in October 2021, "ASP.NE...	357	gpt-4o-mini-202...
Jun 3, 3:16pm	What are the unique selling points of Microsoft Azure?	Microsoft Azure offers a range of unique selling points (U...	616	gpt-4o-mini-202...

New Relic AI Monitoring

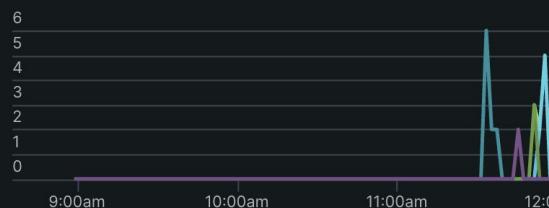
Models

...

Model	Requests	Average response time (s)	Average tokens/completi...	Error rate
gpt-4.1-mini	10	4.08	330	0%
gpt-4o-mini	7	6.59	299	28.57%
mistral-small-2503	3	2.7	374	0%
gpt-4.1-nano	2	2.49	389	0%

Requests

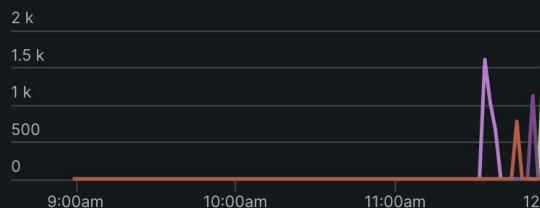
...



- Requests
- gpt-4.1-mini
- gpt-4o-mini

Tokens used

...



- Tokens
- gpt-4.1-mini-2025-04-14
- gpt-4o-mini-2024-07-18

New Relic AI Monitoring

Select models to compare

Compare the cost and performance of your models. Choose your entity, your model and pick the relevant time range and compare with another model on the same or different entity and on the same or different time range to compare their average performance and cost

Service = All

gpt-4.1-mini ▾

🕒 Since 3 hours ago (GMT+2) ▾

Service = All

mistral-small-2503 ▾

🕒 Since 3 hours ago (GMT+2) ▾

See the comparison

Cost

Tokens per completion

329.500 tokens per completion -12%

3.295k total tokens

Tokens per completion

373.667 tokens per completion

1.121k total tokens

Prompt tokens

10.500 tokens per prompt

105 prompt tokens

Prompt tokens

8.333 tokens per prompt -21%

25 prompt tokens

Response tokens

319 tokens per response -13%

3.19k response tokens

Response tokens

365.333 tokens per response

1.096k response tokens

Performance

Throughput

0.056 model calls per minute +233%

Throughput

0.017 model calls per minute

Response time

4.081s

Response time

2.702 s -34%

Error rate

0 %

0 %

OpenTelemetry

<https://github.com/open-telemetry/community/blob/main/projects/gen-ai.md>

The screenshot shows the GitHub repository for OpenTelemetry's community projects. The file 'gen-ai.md' is displayed, containing information about Generative AI Observability. Key sections include:

- Important**: A note stating "This project extends LLM Semantic Conventions project".
- Description**: A paragraph explaining that Generative AI observability is evolving fast and focused on development time and verbose telemetry.
- A list of GenAI-focused observability projects: Arize Phoenix, Langfuse, LangSmith, Langtrace, OpenLit, Traceloop, and many others.
- A note that some projects are based on OpenTelemetry, while others are "inspired by" it.
- A section on the scope of GenAI observability, mentioning instrumentations of various clients, frameworks, response evaluations, and model (server-side) instrumentations.

<https://github.com/Scale3-Labs/langtrace>

The screenshot shows the GitHub repository for Langtrace, which provides open-source observability for LLM applications. It includes:

- README.md**: A file detailing the project's purpose and status.
- Langtrace Dashboard**: A screenshot of the dashboard interface, showing various monitoring and analysis tools.
- Open Source Observability for LLM Applications**: A summary section with project statistics and deployment options.

Project statistics shown in the README:

- License: AGPL-3.0
- Status: Active
- Pull requests: 1 open
- TypeScript SDK: v6.3.5
- Downloads: 588/month
- Python SDK: v3.8.15
- Downloads/month: 24k
- Deploy on Railway
- Downloads: 308k

What else?

LLM Evaluation

AI Monitoring

LLM Evaluation Preview

Account: 4472875 - OpenTelemetry



Since 30 minutes ago (GMT+2)

Thresholds

The threshold defines the sensitivity of the measurement. Lower thresholds mean higher sensitivity. Higher thresholds mean edge cases will be flagged for bad quality.

Threshold values must be between 0 and 1.

Identity Hate

0.5

Insult

0.2

Negativity

0.1

Obscenity

0.1

Threat

0.5

Toxicity

0.5

Toxicity (severe)

0.1

Enable keyword scanning

We are looking for specific keywords within the model responses to highlight the existence of these keywords which highlight improper model responses.

Sampling Rate

0.1

Update

Summary

Summary

Trends

- Exploration
- Enterprise adoption
- Concerns

Challenges

- Increased complexity
- New telemetry
- Security & Compliance

Observability

- New Relic AI Monitoring
- OpenTelemetry



Harry Kimpel
Principal Developer Relations Engineer
New Relic
@harrykimpel





THANK YOU



Harry Kimpel

Principal Developer Relations Engineer

New Relic
@harrykimpel

