GenAI Observability with OpenTelemetry and Grafana

25 Set 2025
David Pereira

David Pereira

Software Architect at CloudCockpit

> “Student for life! Always seeking to improve his skills and diving deeper into cloud-native technologies”

https://github.com/BOLT04

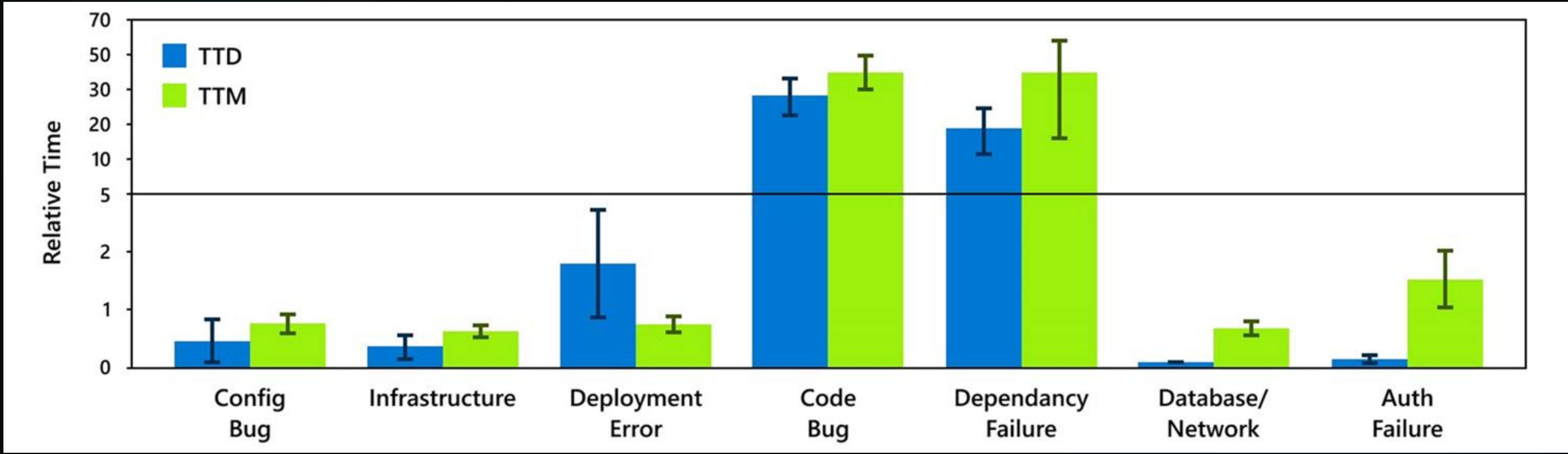
https://www.linkedin.com/in/jose-david-pereira/

@bolt2938

Agenda

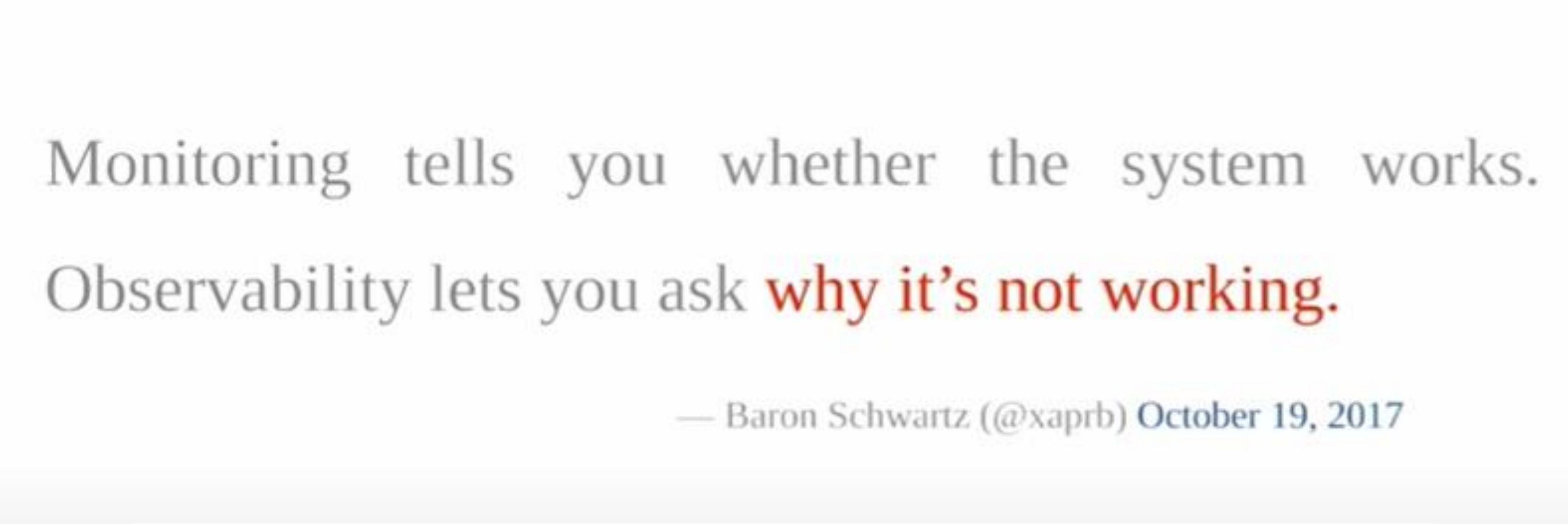
1. Observability Intro
2. GenAI Observabiliy
3. Metrics – What we want
4. Metrics – Self-hosted models
5. GenAI Semantic Conventions
6. Demo
7. GenAI Observability Roadmap
8. Conclusion
9. Resources
10. Q&A

Observability Intro



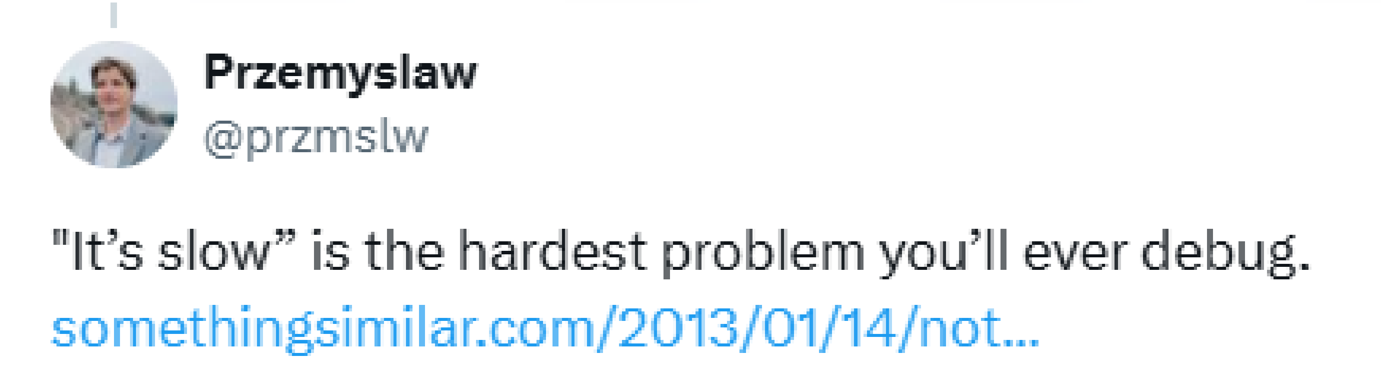
Source: https://www.microsoft.com/en-us/research/articles/towards-highly-reliable-services-with-aiops/

Observability Intro



Monitoring tells you whether the system works. Observability lets you ask why it’s not working.

— Baron Schwartz (@xaprb) October 19, 2017



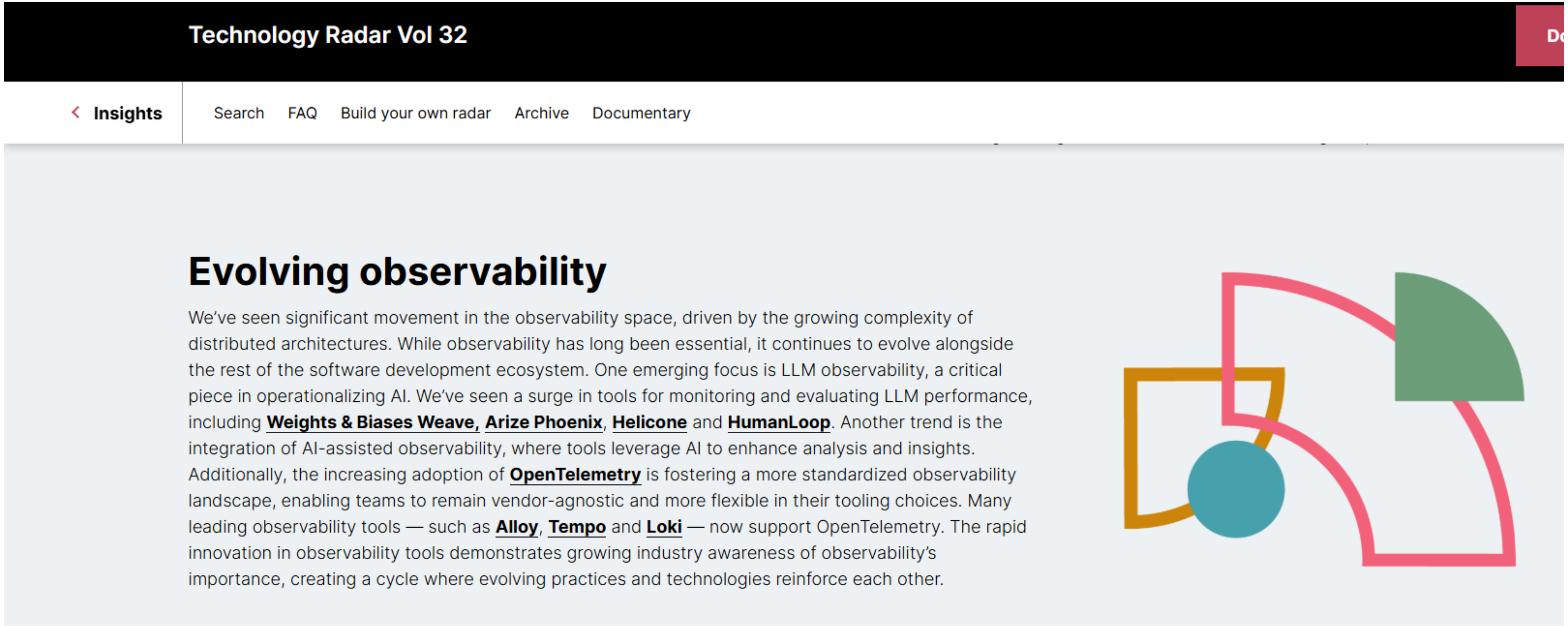
Przemyslaw
@przmslw

"It’s slow” is the hardest problem you’ll ever debug. somethingsimilar.com/2013/01/14/not...

Sources:
https://www.somethingsimilar.com/2013/01/14/notes-on-distributed-systems-for-young-bloods/
https://www.case-podcast.org/54-theo-schlossnagle-software-engineering
https://www.p99conf.io/session/how-to-measure-latency/

GenAI Observability

arize Langfuse



Evolving observability

We've seen significant movement in the observability space, driven by the growing complexity of distributed architectures. While observability has long been essential, it continues to evolve alongside the rest of the software development ecosystem. One emerging focus is LLM observability, a critical piece in operationalizing AI. We've seen a surge in tools for monitoring and evaluating LLM performance, including **Weights & Biases Weave, Arize Phoenix**, **Helicone** and **HumanLoop**. Another trend is the integration of AI-assisted observability, where tools leverage AI to enhance analysis and insights. Additionally, the increasing adoption of **OpenTelemetry** is fostering a more standardized observability landscape, enabling teams to remain vendor-agnostic and more flexible in their tooling choices. Many leading observability tools — such as **Alloy**, **Tempo** and **Loki** — now support OpenTelemetry. The rapid innovation in observability tools demonstrates growing industry awareness of observability's importance, creating a cycle where evolving practices and technologies reinforce each other.

https://www.thoughtworks.com/radar

GenAI Observability - Agents

- **Action/Input**: The trigger for the agent to start working.
- **Memory**: Retain context and learnings (e.g. in a DB).
- **Planning**: The reasoning phase where the agent thinks and comes up with a plan (reasoning models).
- **Tools:** Leverage external tools like databases and APIs for real-world tasks (plugins, MCP, etc)

https://devblogs.microsoft.com/all-things-azure/agentic-philosophers/

GenAI Observability

- Observing Agentic AI systems (e.g. traces)
- Troubleshoot problems with SRE agents or other LLMs
- There is a need for instrumentation for:

agentic frameworks; vector DBs and LLMs

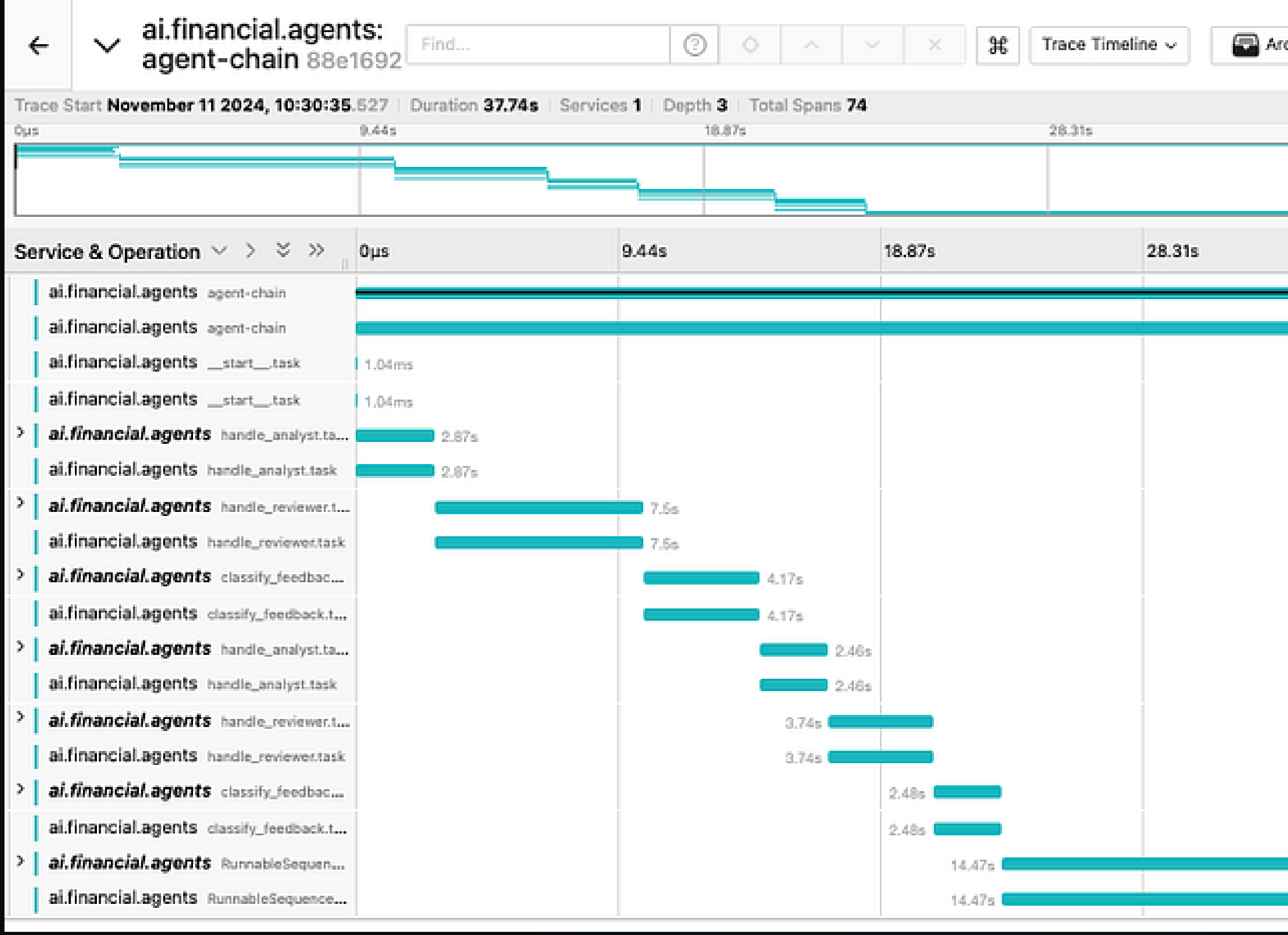
OpenTelemetry

Humanloop

OpenLIT

OpenLLMetry

OPENINFERENCE



ai.financial.agents: agent-chain 88e1692

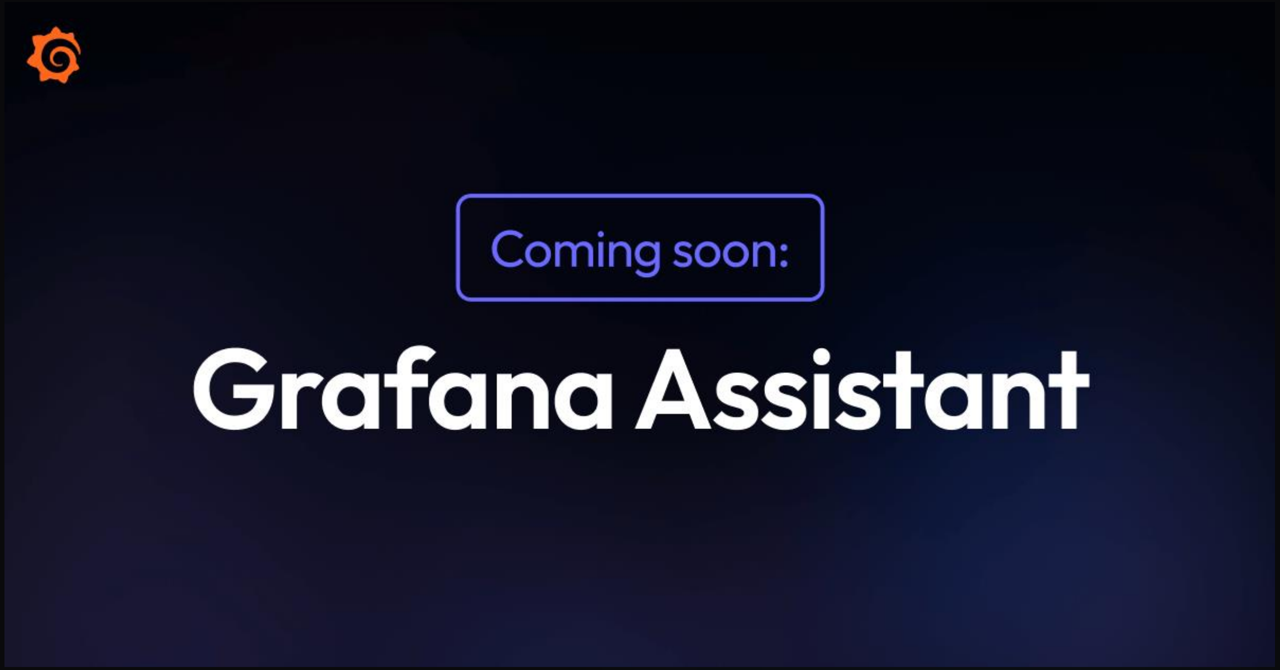
Trace Start **November 11 2024, 10:30:35**.527 | Duration **37.74s** | Services **1** | Depth **3** | Total Spans **74**

Service & Operation	
ai.financial.agents agent-chain	
ai.financial.agents agent-chain	
ai.financial.agents __start__.task	1.04ms
ai.financial.agents __start__.task	1.04ms
ai.financial.agents handle_analyst.ta...	2.87s
ai.financial.agents handle_analyst.task	2.87s
ai.financial.agents handle_reviewer.t...	7.5s
ai.financial.agents handle_reviewer.task	7.5s
ai.financial.agents classify_feedbac...	4.17s
ai.financial.agents classify_feedback.t...	4.17s
ai.financial.agents handle_analyst.ta...	2.46s
ai.financial.agents handle_analyst.task	2.46s
ai.financial.agents handle_reviewer.t...	3.74s
ai.financial.agents handle_reviewer.task	3.74s
ai.financial.agents classify_feedbac...	2.48s
ai.financial.agents classify_feedback.t...	2.48s
ai.financial.agents RunnableSequen...	14.47s
ai.financial.agents RunnableSequence...	14.47s

GenAI Observability

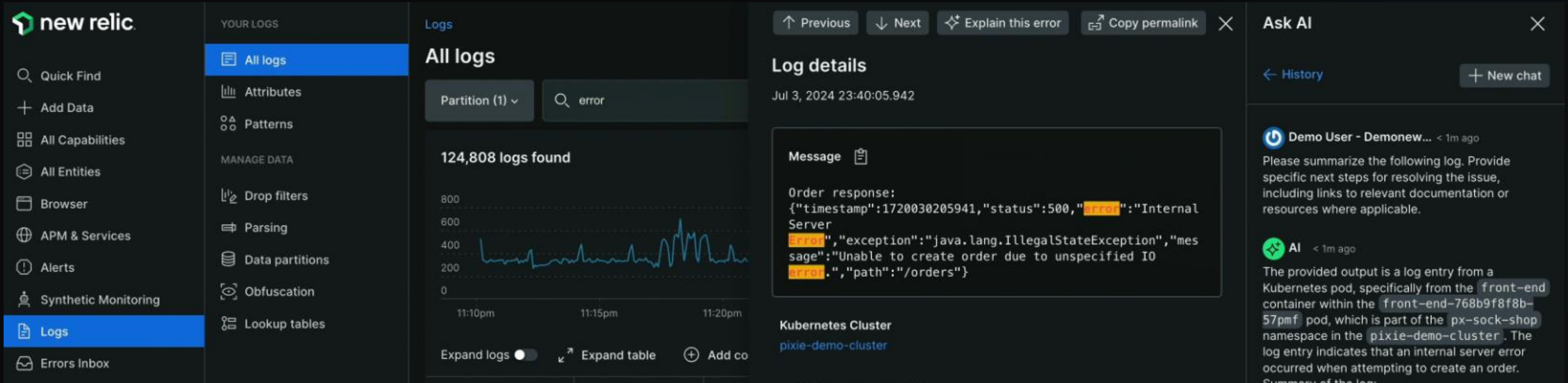
Azure SRE agent

Run apps resiliently in production



Coming soon:

Grafana Assistant



Log details

Jul 3, 2024 23:40:05.942

Message

```
Order response:
{"timestamp":1720030205941,"status":500,"error":"Internal
Server
Error","exception":"java.lang.IllegalStateException","mes
sage":"Unable to create order due to unspecified IO
error.","path":"/orders"}
```

Kubernetes Cluster
pixie-demo-cluster

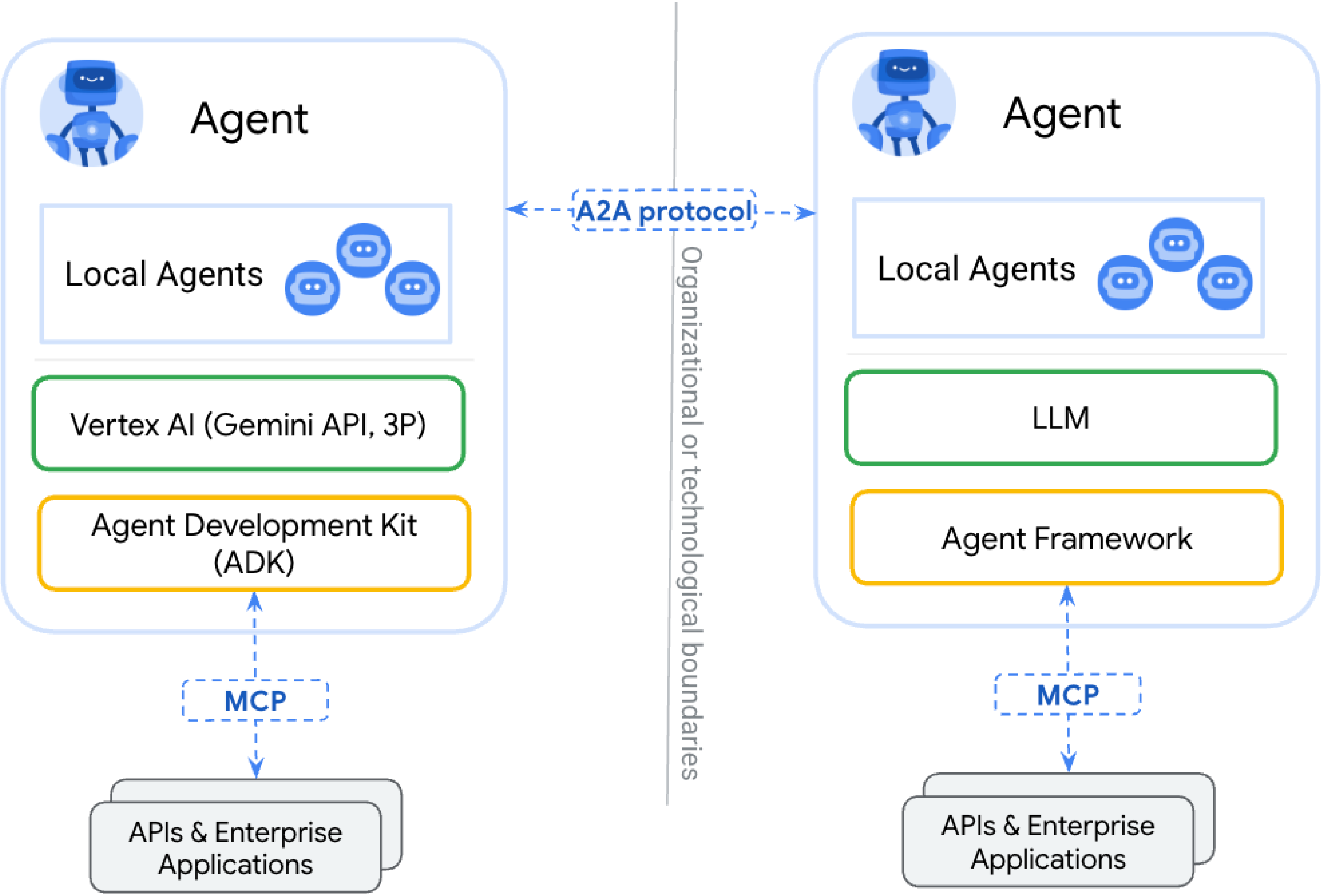
Demo User - Demonew... < 1m ago

Please summarize the following log. Provide specific next steps for resolving the issue, including links to relevant documentation or resources where applicable.

AI < 1m ago

The provided output is a log entry from a Kubernetes pod, specifically from the `front-end` container within the `front-end-768b9f8f8b-57pmf` pod, which is part of the `px-sock-shop` namespace in the `pixie-demo-cluster`. The log entry indicates that an internal server error occurred when attempting to create an order.

Summary of the log:


Agent
Local Agents
Vertex AI (Gemini API, 3P)
Agent Development Kit (ADK)
MCP
APIs & Enterprise Applications
A2A protocol
Organizational or technological boundaries
Agent
Local Agents
LLM
Agent Framework
MCP
APIs & Enterprise Applications

Metrics – What we want

Token Usage + Cost

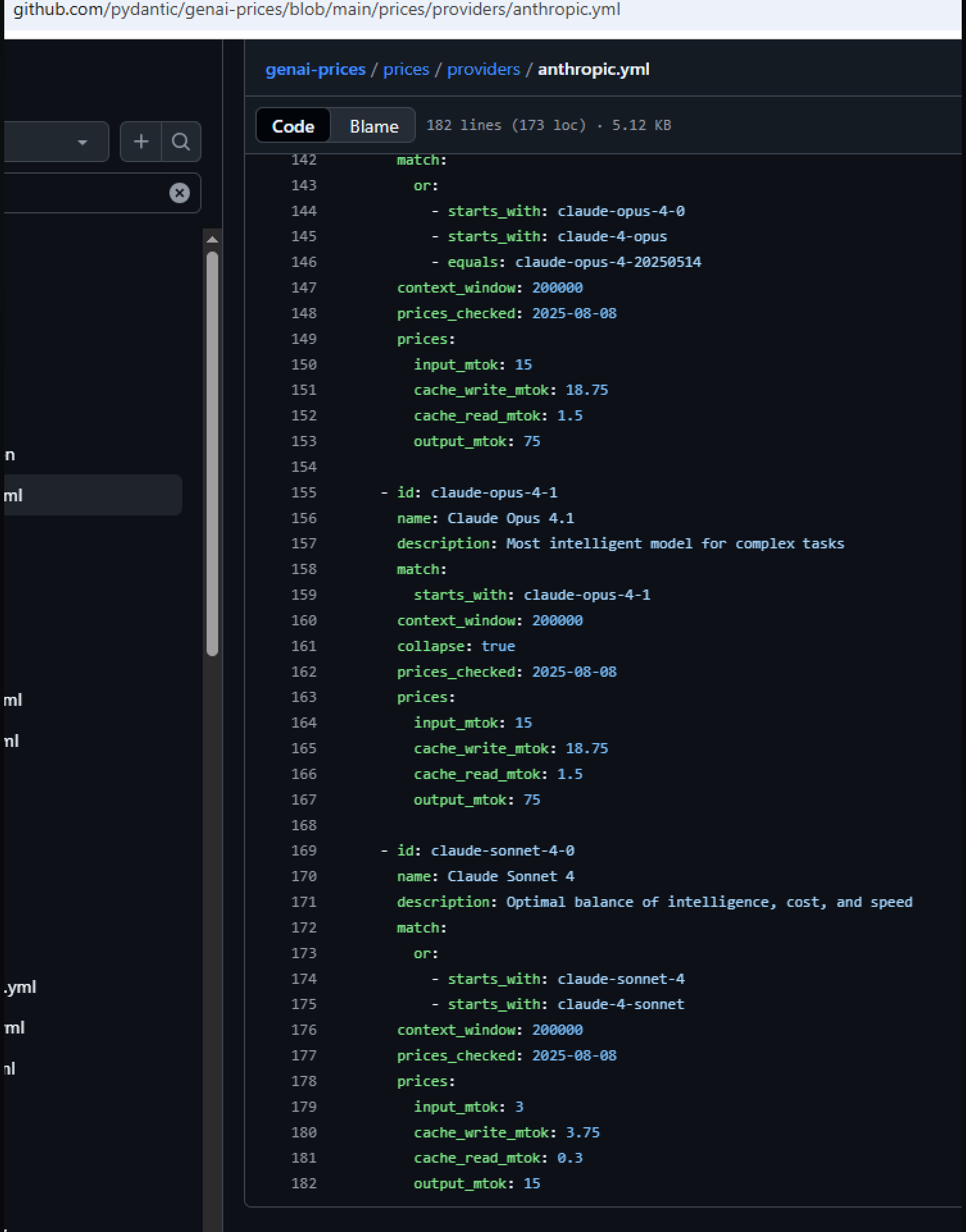
Latency

Prompt and LLM Response

Toxicity and Safety

Hallucinations

Tool calls



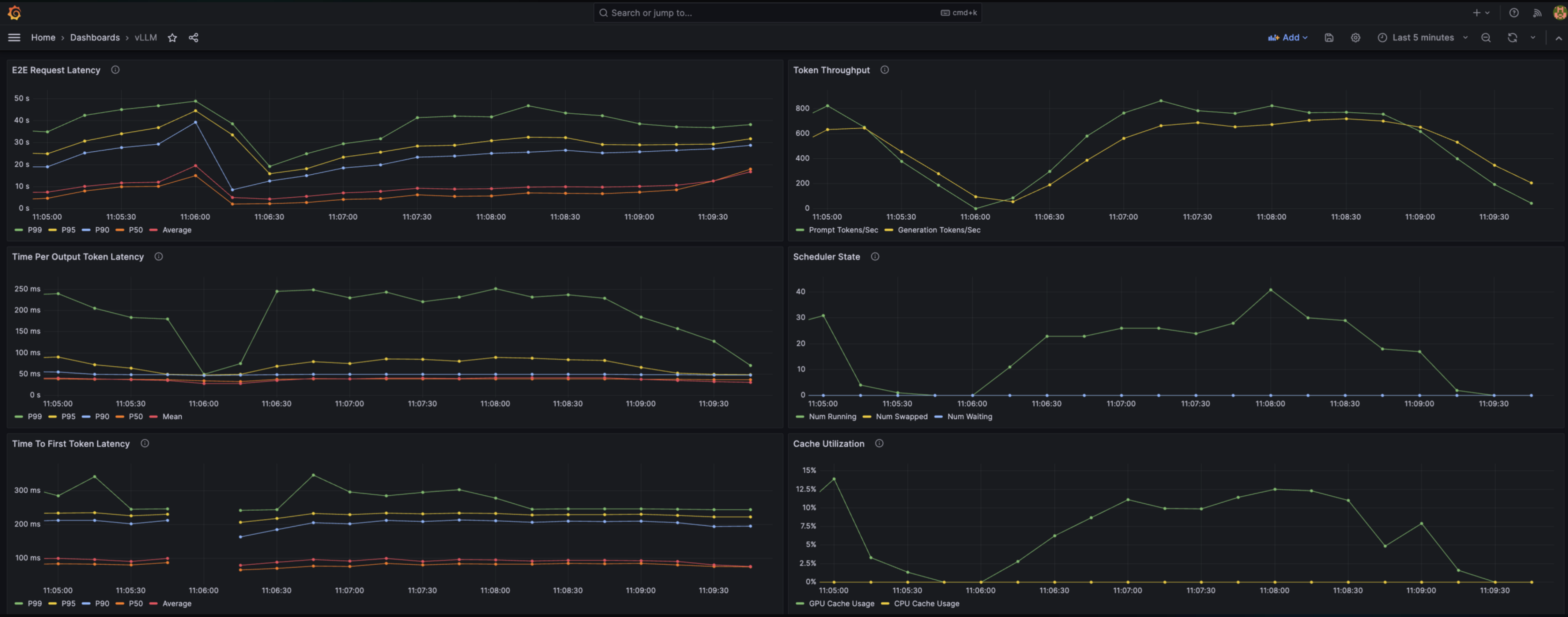
github.com/pydantic/genai-prices/blob/main/prices/providers/anthropic.yml

genai-prices / prices / providers / anthropic.yml

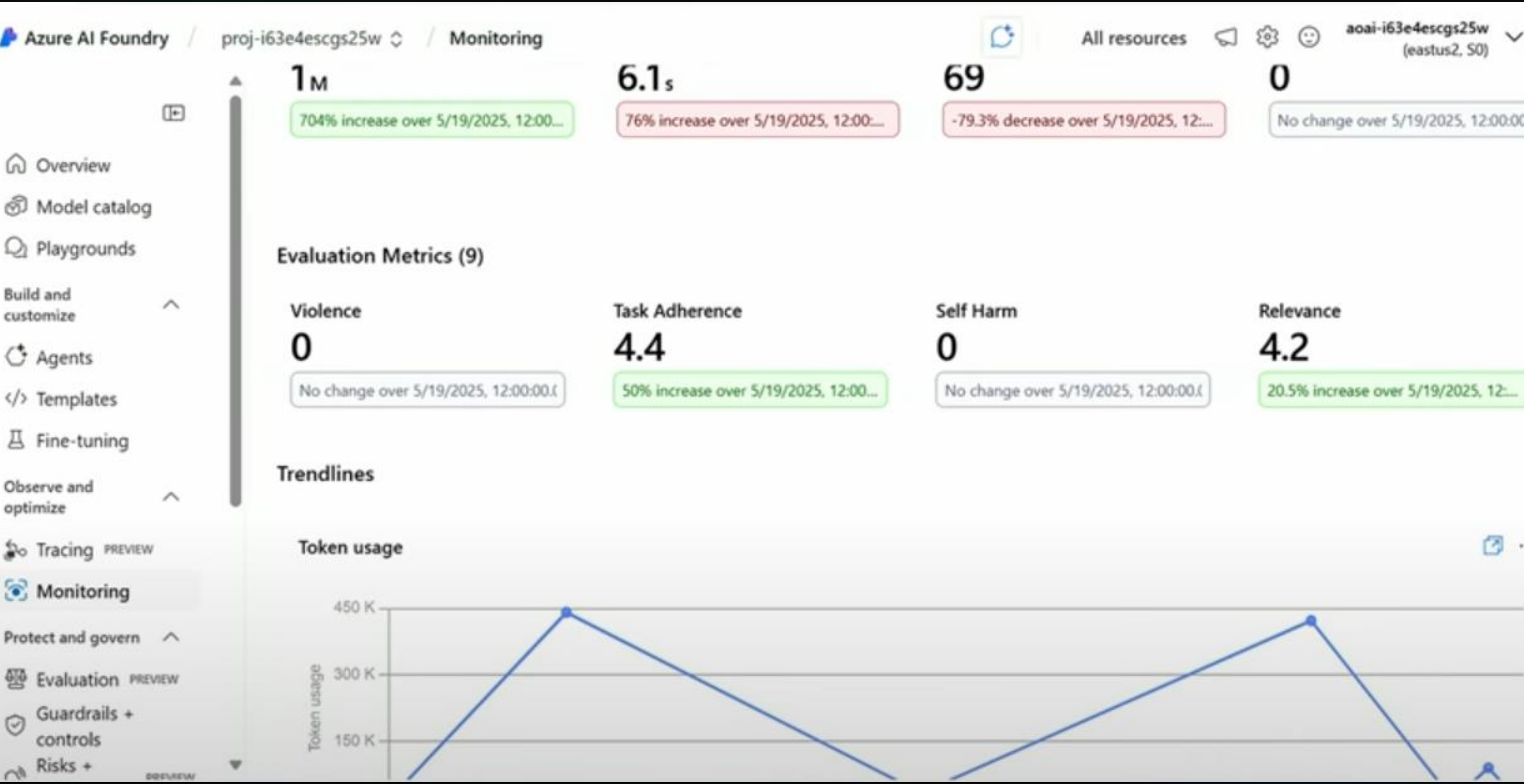
182 lines (173 loc) · 5.12 KB

```yaml
      match:
        or:
          - starts_with: claude-opus-4-0
          - starts_with: claude-4-opus
          - equals: claude-opus-4-20250514
      context_window: 200000
      prices_checked: 2025-08-08
      prices:
        input_mtok: 15
        cache_write_mtok: 18.75
        cache_read_mtok: 1.5
        output_mtok: 75

    - id: claude-opus-4-1
      name: Claude Opus 4.1
      description: Most intelligent model for complex tasks
      match:
        starts_with: claude-opus-4-1
      context_window: 200000
      collapse: true
      prices_checked: 2025-08-08
      prices:
        input_mtok: 15
        cache_write_mtok: 18.75
        cache_read_mtok: 1.5
        output_mtok: 75

    - id: claude-sonnet-4-0
      name: Claude Sonnet 4
      description: Optimal balance of intelligence, cost, and speed
      match:
        or:
          - starts_with: claude-sonnet-4
          - starts_with: claude-4-sonnet
      context_window: 200000
      prices_checked: 2025-08-08
      prices:
        input_mtok: 3
        cache_write_mtok: 3.75
        cache_read_mtok: 0.3
        output_mtok: 15
```

Metrics – Self-hosted models



Source: *https://docs.vllm.ai/en/stable/examples/online_serving/prometheus_grafana.html#import-dashboard*



1M
704% increase over 5/19/2025, 12:00...

6.1s
76% increase over 5/19/2025, 12:00:...

69
-79.3% decrease over 5/19/2025, 12:...

0
No change over 5/19/2025, 12:00:00

Evaluation Metrics (9)

Violence
0
No change over 5/19/2025, 12:00:00.(

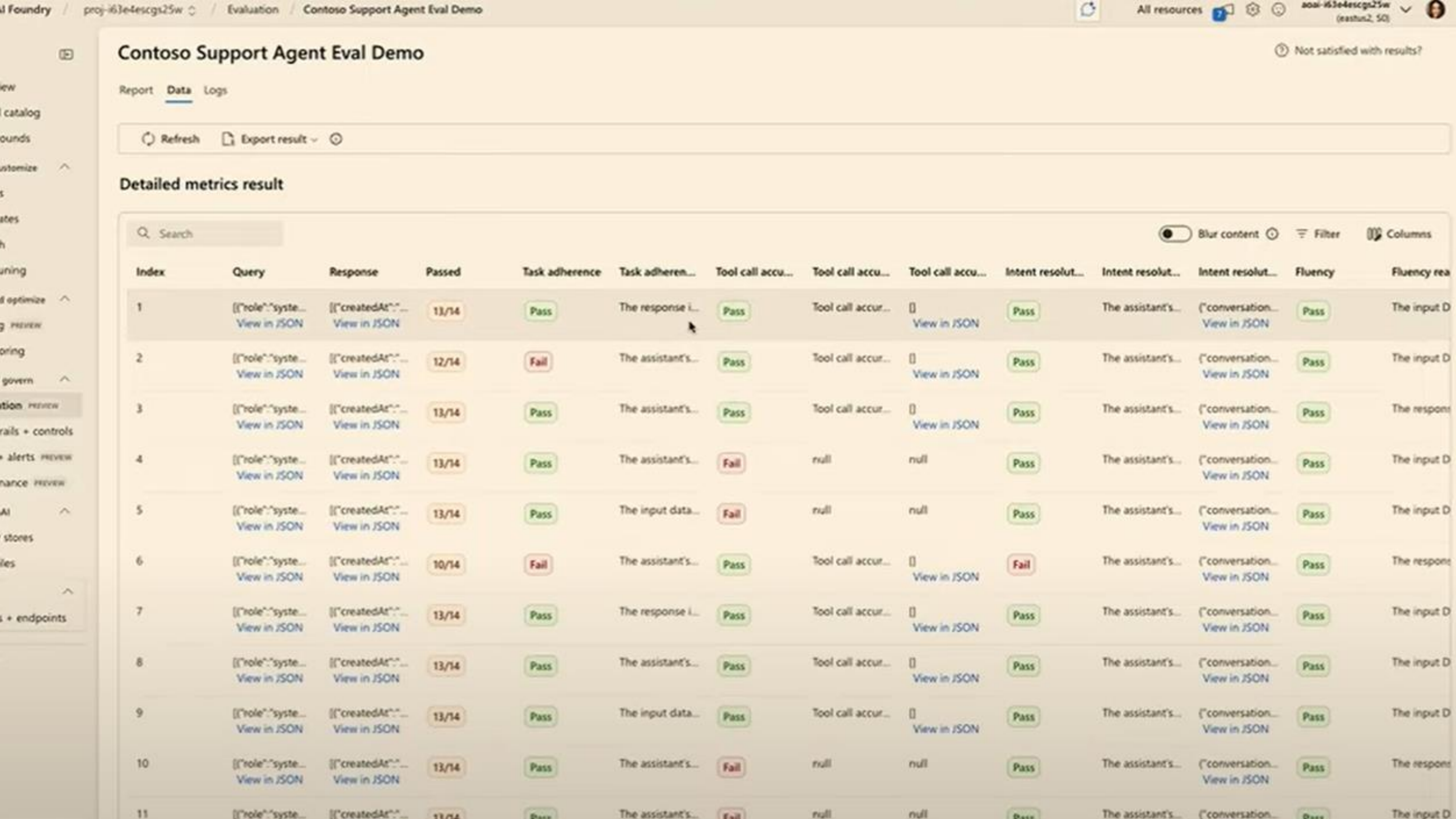
Task Adherence
4.4
50% increase over 5/19/2025, 12:00...

Self Harm
0
No change over 5/19/2025, 12:00:00.(

Relevance
4.2
20.5% increase over 5/19/2025, 12:...

Trendlines

Token usage



Contoso Support Agent Eval Demo

Not satisfied with results?

Report | Data | Logs

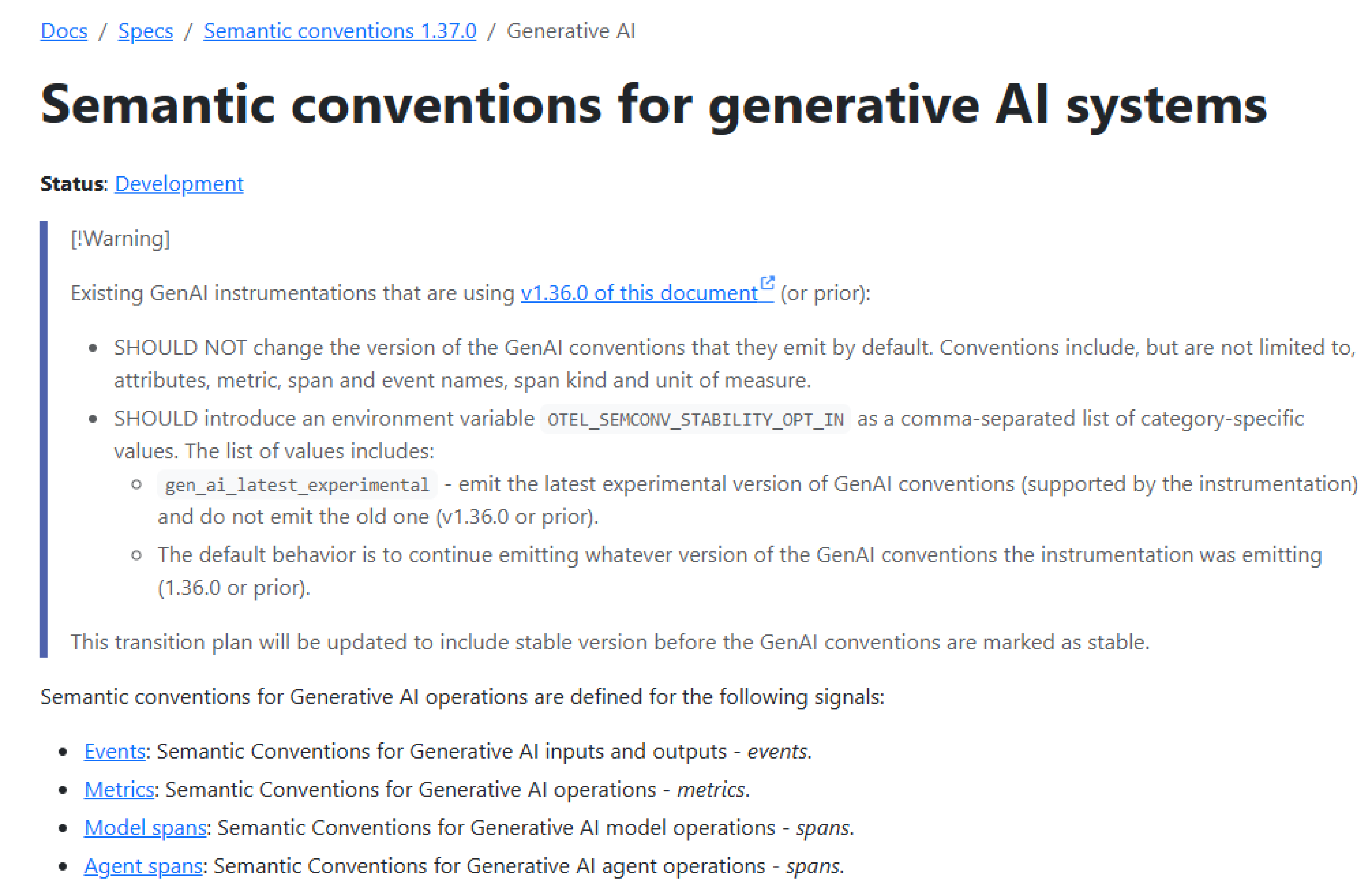
Refresh | Export result

Detailed metrics result

Search | Blur content | Filter | Columns

Index	Query	Response	Passed	Task adherence	Task adheren...	Tool call accu...	Tool call accu...	Tool call accu...	Intent resolut...	Intent resolut...	Intent resolut...	Fluency	Fluency rea
1	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The response i...	Pass	Tool call accur...	[] View in JSON	Pass	The assistant's...	{"conversation... View in JSON	Pass	The input D
2	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	12/14	Fail	The assistant's...	Pass	Tool call accur...	[] View in JSON	Pass	The assistant's...	{"conversation... View in JSON	Pass	The input D
3	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The assistant's...	Pass	Tool call accur...	[] View in JSON	Pass	The assistant's...	{"conversation... View in JSON	Pass	The respons
4	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The assistant's...	Fail	null	null	Pass	The assistant's...	{"conversation... View in JSON	Pass	The input D
5	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The input data...	Fail	null	null	Pass	The assistant's...	{"conversation... View in JSON	Pass	The input D
6	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	10/14	Fail	The assistant's...	Pass	Tool call accur...	[] View in JSON	Fail	The assistant's...	{"conversation... View in JSON	Pass	The respons
7	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The response i...	Pass	Tool call accur...	[] View in JSON	Pass	The assistant's...	{"conversation... View in JSON	Pass	The input D
8	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The assistant's...	Pass	Tool call accur...	[] View in JSON	Pass	The assistant's...	{"conversation... View in JSON	Pass	The input D
9	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The input data...	Pass	Tool call accur...	[] View in JSON	Pass	The assistant's...	{"conversation... View in JSON	Pass	The input D
10	[{"role":"syste... View in JSON	[{"createdAt":"... View in JSON	13/14	Pass	The assistant's...	Fail	null	null	Pass	The assistant's...	{"conversation... View in JSON	Pass	The respons
11	[{"role":"syste...	[{"createdAt":"...	13/14	Pass	The assistant's...	Fail	null	null	Pass	The assistant's...	{"conversation...	Pass	The input D

GenAI Semantic Conventions



Docs / Specs / Semantic conventions 1.37.0 / Generative AI

Semantic conventions for generative AI systems

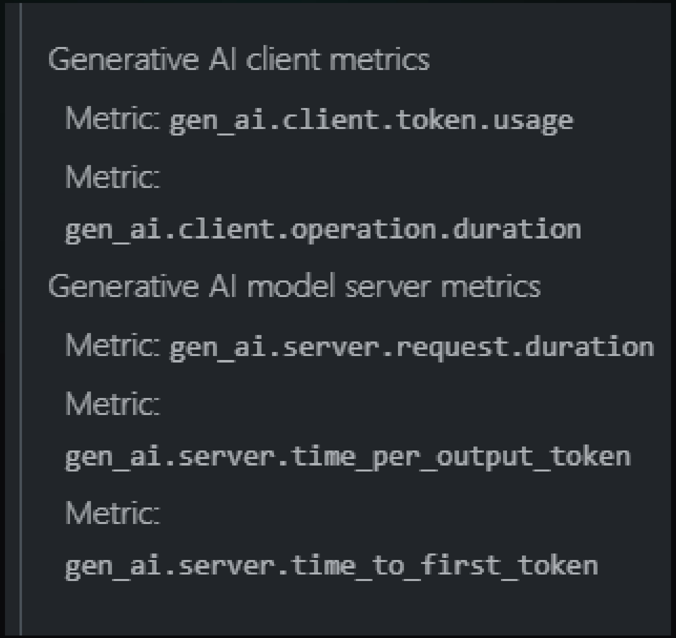
Status: Development

> [!Warning]
>
> Existing GenAI instrumentations that are using v1.36.0 of this document (or prior):
>
> - SHOULD NOT change the version of the GenAI conventions that they emit by default. Conventions include, but are not limited to, attributes, metric, span and event names, span kind and unit of measure.
> - SHOULD introduce an environment variable `OTEL_SEMCONV_STABILITY_OPT_IN` as a comma-separated list of category-specific values. The list of values includes:
>   - `gen_ai_latest_experimental` - emit the latest experimental version of GenAI conventions (supported by the instrumentation) and do not emit the old one (v1.36.0 or prior).
>   - The default behavior is to continue emitting whatever version of the GenAI conventions the instrumentation was emitting (1.36.0 or prior).
>
> This transition plan will be updated to include stable version before the GenAI conventions are marked as stable.

Semantic conventions for Generative AI operations are defined for the following signals:

- Events: Semantic Conventions for Generative AI inputs and outputs - *events*.
- Metrics: Semantic Conventions for Generative AI operations - *metrics*.
- Model spans: Semantic Conventions for Generative AI model operations - *spans*.
- Agent spans: Semantic Conventions for Generative AI agent operations - *spans*.

GenAI Semantic Conventions

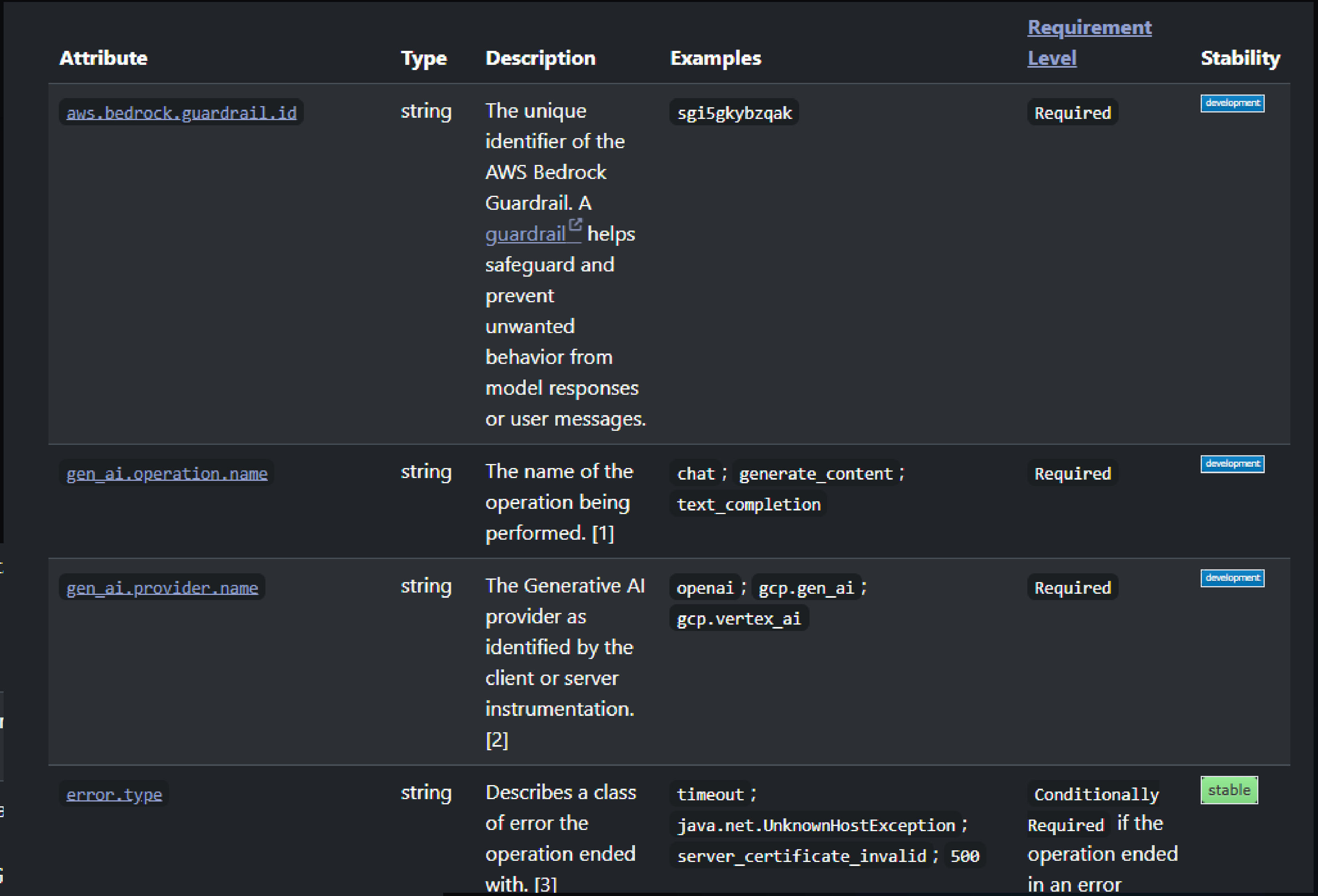


Generative AI client metrics

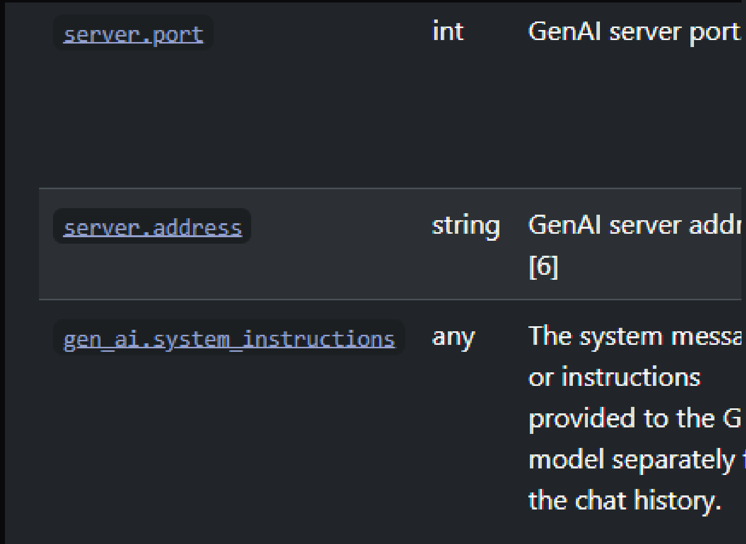
- Metric: `gen_ai.client.token.usage`
- Metric: `gen_ai.client.operation.duration`

Generative AI model server metrics

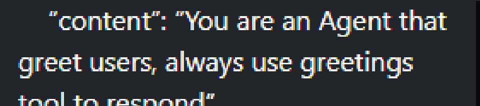
- Metric: `gen_ai.server.request.duration`
- Metric: `gen_ai.server.time_per_output_token`
- Metric: `gen_ai.server.time_to_first_token`



Attribute	Type	Description	Examples	Requirement Level	Stability
`aws.bedrock.guardrail.id`	string	The unique identifier of the AWS Bedrock Guardrail. A guardrail helps safeguard and prevent unwanted behavior from model responses or user messages.	`sgi5gkybzqak`	`Required`	development
`gen_ai.operation.name`	string	The name of the operation being performed. [1]	`chat`; `generate_content`; `text_completion`	`Required`	development
`gen_ai.provider.name`	string	The Generative AI provider as identified by the client or server instrumentation. [2]	`openai`; `gcp.gen_ai`; `gcp.vertex_ai`	`Required`	development
`error.type`	string	Describes a class of error the operation ended with. [3]	`timeout`; `java.net.UnknownHostException`; `server_certificate_invalid`; `500`	`Conditionally Required` if the operation ended in an error	stable

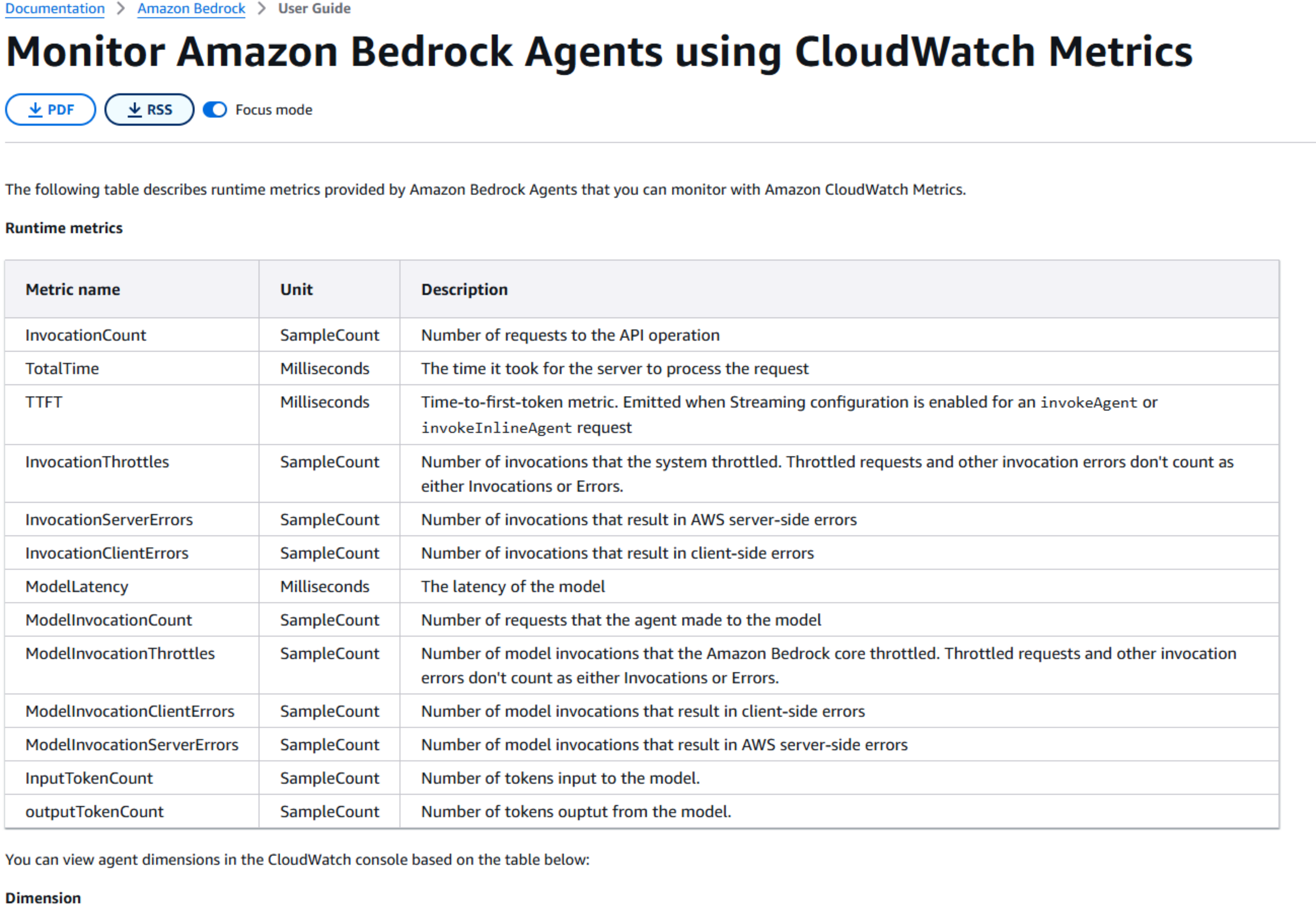


`server.port`	int	GenAI server port
`server.address`	string	GenAI server addr [6]
`gen_ai.system_instructions`	any	The system messa or instructions provided to the G model separately from the chat history.



“content”: “You are an Agent that greet users, always use greetings tool to respond”

GenAI Semantic Conventions – Vendor specific



Monitor Amazon Bedrock Agents using CloudWatch Metrics

The following table describes runtime metrics provided by Amazon Bedrock Agents that you can monitor with Amazon CloudWatch Metrics.

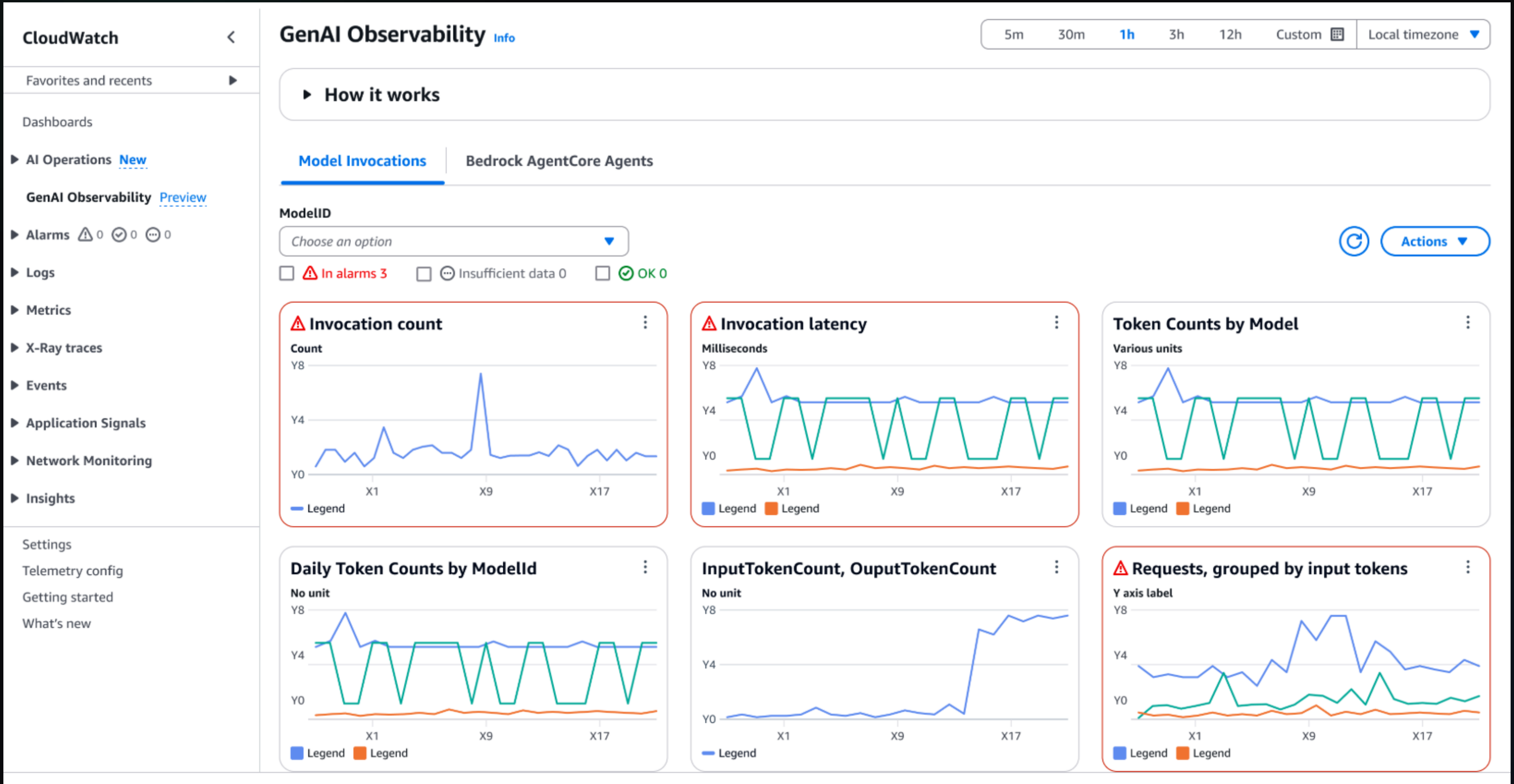
Runtime metrics

Metric name	Unit	Description
InvocationCount	SampleCount	Number of requests to the API operation
TotalTime	Milliseconds	The time it took for the server to process the request
TTFT	Milliseconds	Time-to-first-token metric. Emitted when Streaming configuration is enabled for an `invokeAgent` or `invokeInlineAgent` request
InvocationThrottles	SampleCount	Number of invocations that the system throttled. Throttled requests and other invocation errors don't count as either Invocations or Errors.
InvocationServerErrors	SampleCount	Number of invocations that result in AWS server-side errors
InvocationClientErrors	SampleCount	Number of invocations that result in client-side errors
ModelLatency	Milliseconds	The latency of the model
ModelInvocationCount	SampleCount	Number of requests that the agent made to the model
ModelInvocationThrottles	SampleCount	Number of model invocations that the Amazon Bedrock core throttled. Throttled requests and other invocation errors don't count as either Invocations or Errors.
ModelInvocationClientErrors	SampleCount	Number of model invocations that result in client-side errors
ModelInvocationServerErrors	SampleCount	Number of model invocations that result in AWS server-side errors
InputTokenCount	SampleCount	Number of tokens input to the model.
outputTokenCount	SampleCount	Number of tokens ouptut from the model.

You can view agent dimensions in the CloudWatch console based on the table below:

Dimension

GenAI Semantic Conventions – Vendor specific



CloudWatch

Favorites and recents ▶

Dashboards

▶ AI Operations New

GenAI Observability Preview

▶ Alarms ⚠ 0 ✓ 0 ⊖ 0

▶ Logs

▶ Metrics

▶ X-Ray traces

▶ Events

▶ Application Signals

▶ Network Monitoring

▶ Insights

Settings

Telemetry config

Getting started

What's new

GenAI Observability Info

5m 30m 1h 3h 12h Custom | Local timezone ▼

▶ How it works

Model Invocations | Bedrock AgentCore Agents

ModelID

Choose an option ▼

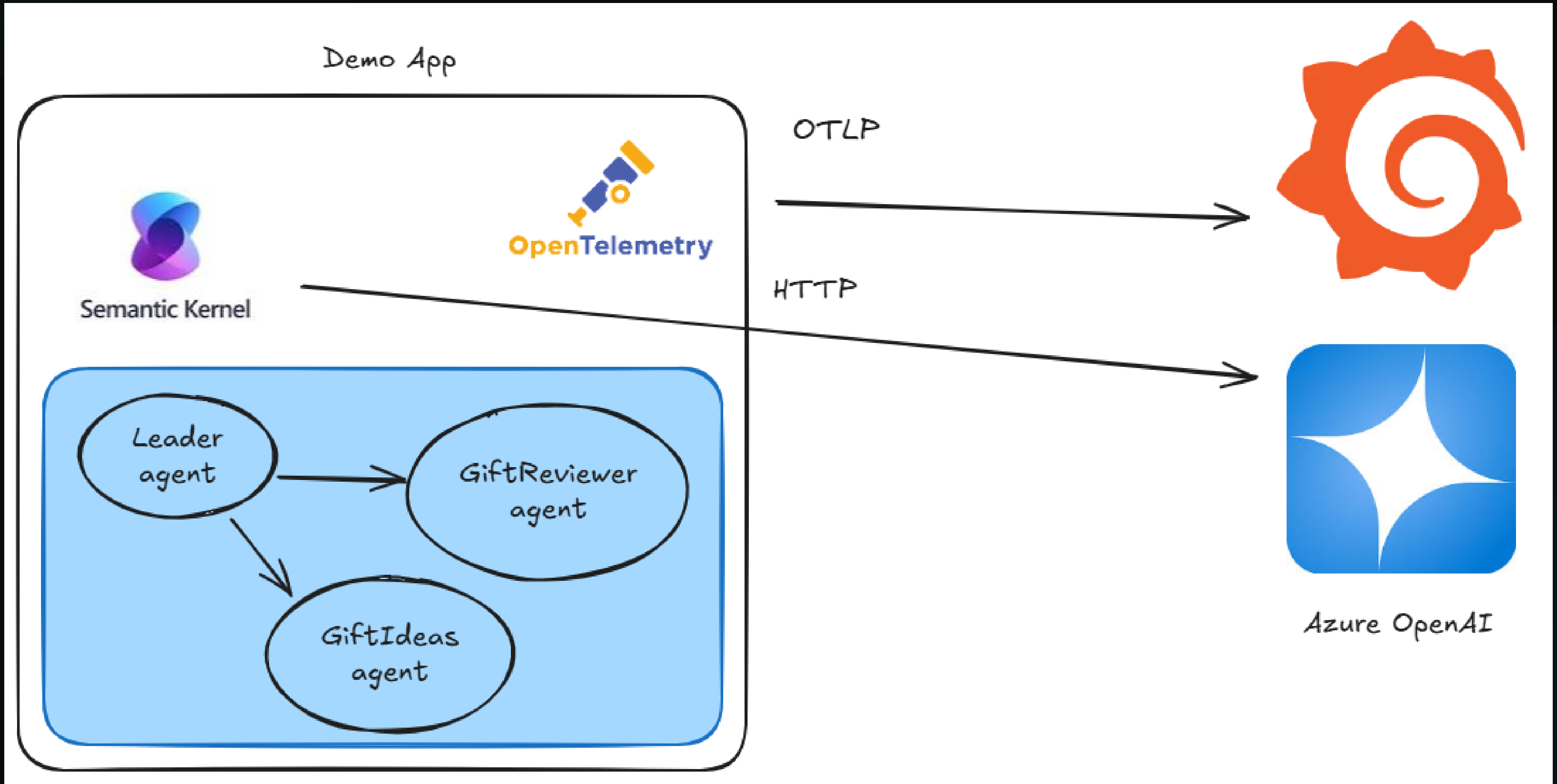
Actions ▼

□ ⚠ In alarms 3 □ ⊖ Insufficient data 0 □ ✓ OK 0

Demo

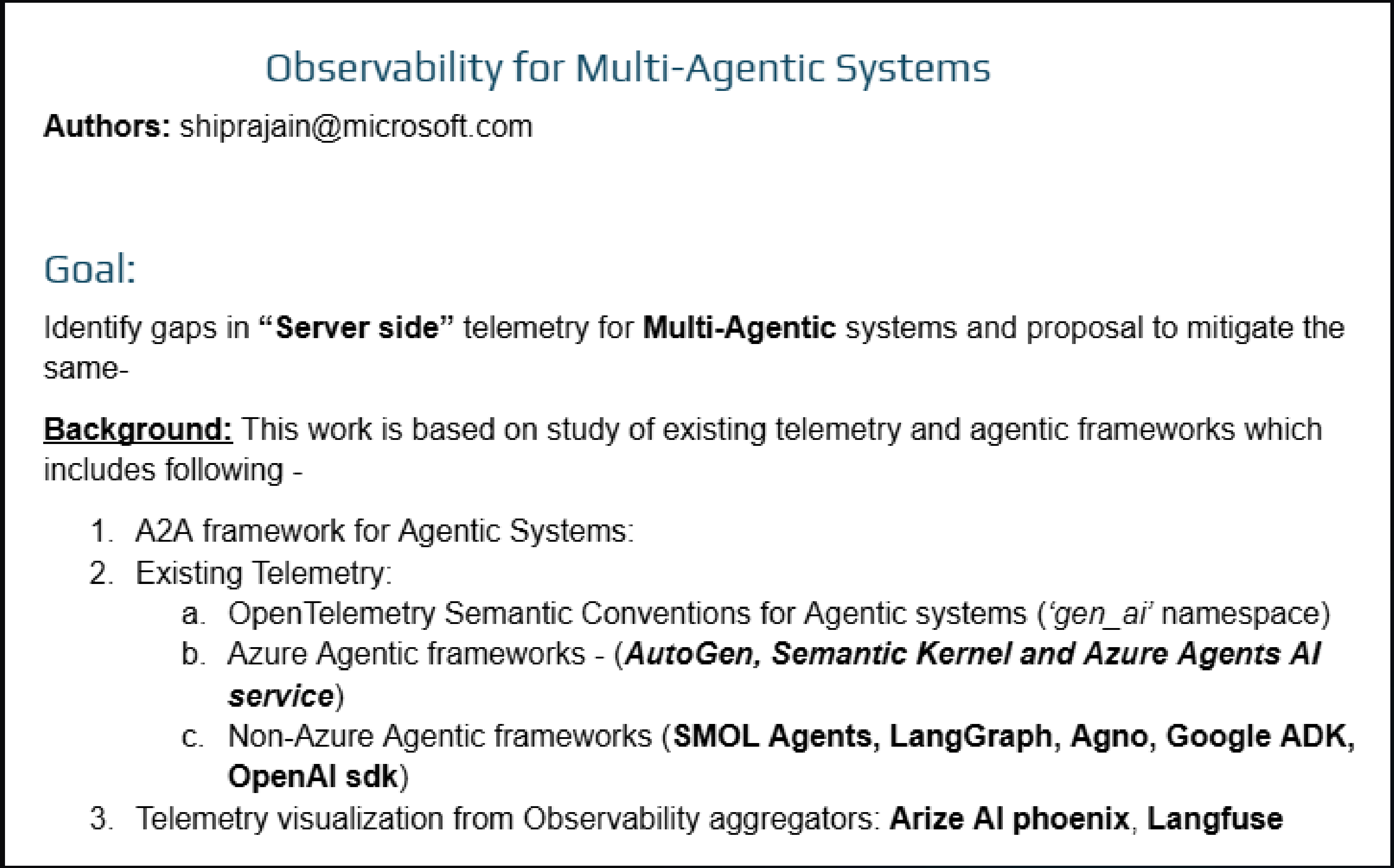
Demo

- Simple chat completion
- Shop agents: A simulated e-commerce checkout driven by an agentic workflow using orchestration.
- Grafana traces, metrics and logs
- Grafana AI Observability integration + GPU monitoring



GenAI Observability Roadmap

- GenAI Sub Projects
- Refactor chat history in attributes
- Cost optimization and monitoring
- Refactoring existing semantic conventions
- Feature: evaluation metrics (https://ai.pydantic.dev/evals/)
- OpenLLMetry donation
- Feature: Multi-Agents (Cisco, Microsoft, etc)
- Feature: MCP



Observability for Multi-Agentic Systems

Authors: shiprajain@microsoft.com

Goal:

Identify gaps in **"Server side"** telemetry for **Multi-Agentic** systems and proposal to mitigate the same-

Background: This work is based on study of existing telemetry and agentic frameworks which includes following -

1. A2A framework for Agentic Systems:
2. Existing Telemetry:
   a. OpenTelemetry Semantic Conventions for Agentic systems (*'gen_ai'* namespace)
   b. Azure Agentic frameworks - (***AutoGen, Semantic Kernel and Azure Agents AI service***)
   c. Non-Azure Agentic frameworks (**SMOL Agents, LangGraph, Agno, Google ADK, OpenAI sdk**)
3. Telemetry visualization from Observability aggregators: **Arize AI phoenix**, **Langfuse**

Conclusion

- OpenTelemetry conventions VS rate of GenAI innovation and speed of change is different
- Rise of other OSS projects like OpenLit and OpenLLMetry
- Semantic Conventions are still very important to have vendor-agnostic solutions and only one instrumentation solution

Resources

Talks:

- Prometheus: PromCon 2024 - Inside a PromQL Query: Understanding the Mechanics
- Modern Platform Engineering: 9 Secrets of Generative Teams - Liz Fong-Jones
- How Prometheus Revolutionized Monitoring at SoundCloud - Björn Rabenstein
- Context Propagation makes OpenTelemetry awesome
- How OpenTelemetry Helps Generative AI - Phillip Carter, Honeycomb
- Keynote: Into the Black Box: Observability in the Age of LLMs - Christine Yen

Links:

- O11y wiki GitHub repo
- Grafana observability report
- Awesome Observability GitHub repo
- AWS observability best practices guide
- Google's SRE book
- About RED and USE method
- Traces Instrumentation best practices in .NET
- Let's use OpenTelemetry with Spring
- AI Agent Observability - Evolving Standards and Best Practices

GitHub:

- Semantic Kernel Observability demo
- Semantic Kernel MultiAgent demo
- Langfuse OpenLit Integration via OpenTelemetry

Resources – Slides and Repo

https://github.com/BOLT04/grafana-observability-demo

https://github.com/BOLT04

https://www.linkedin.com/in/jose-david-pereira/

@bolt2938

Q&A

Thanks

David Pereira