

Dashboards & Dragons

Crafting SLOs to tame AI platform chaos at scale

SREcon25 EMEA

October 7, 2025

Alexa Griffith - Senior Software Engineer
AI Platforms Engineering @ Bloomberg

Sal Furino
Customer Reliability Engineer @Bloomberg

TechAtBloomberg.com

Engineering

Bloomberg



Alexa Griffith

Senior Software Engineer
AI Platforms Engineering

X: @lexal0u
/in/alex-griffith/

Podcast: Alexa's Input (AI)

Website: <https://alexagriffith.com/>



Sal Furino

Customer Reliability Engineer
Reliability Solutions

@sfurino
/in/salvatore-furino/

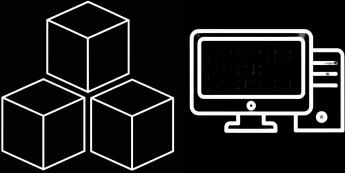
TechAtBloomberg.com

Bloomberg

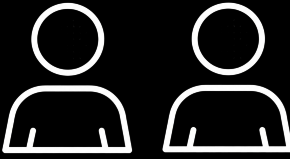
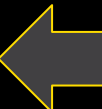
Behold, The Realm of Bloomberg

- Provides financial software tools, applications, news to financial companies
 - Enables analytics, equity trading platform, data services, etc.
- Moves ~400B-650B+ pieces of financial data daily
- Runs and manages one of the world's largest private networks
- >9,000 engineers
- Publishes ~2M+ news stories daily

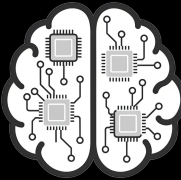
The Guild of AI Platforms at Bloomberg



Infrastructure
Optimized for AI



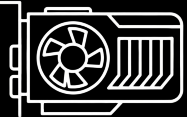
AI Developers



AI Services & Tools



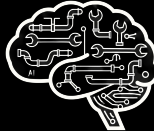
Jupyter Notebooks



Distributed Training & HPC



Managed Serving



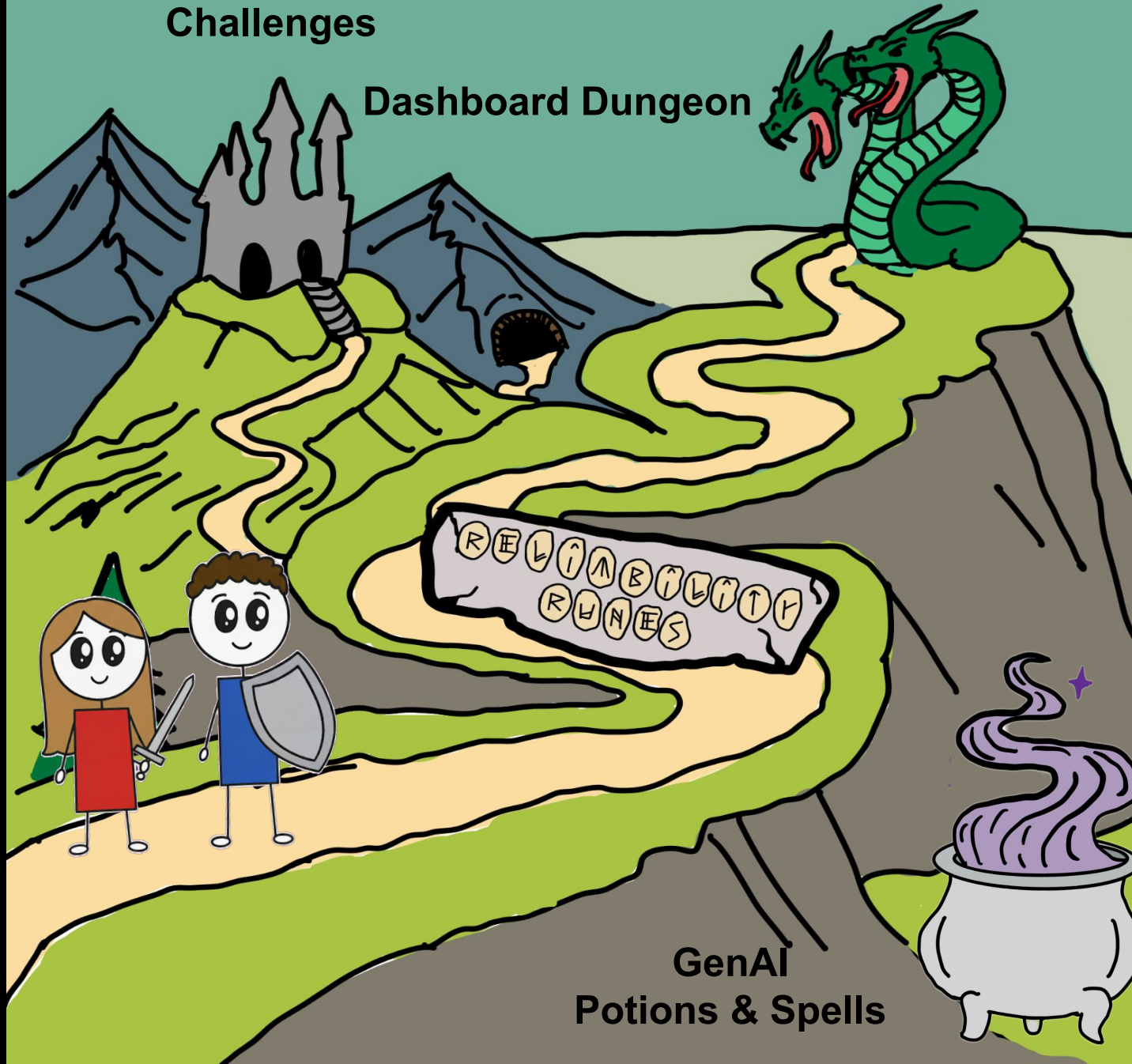
AI Pipelines and Tools



Fortress of Platform
Challenges

GenAI Hydra

Dashboard Dungeon



GenAI
Potions & Spells

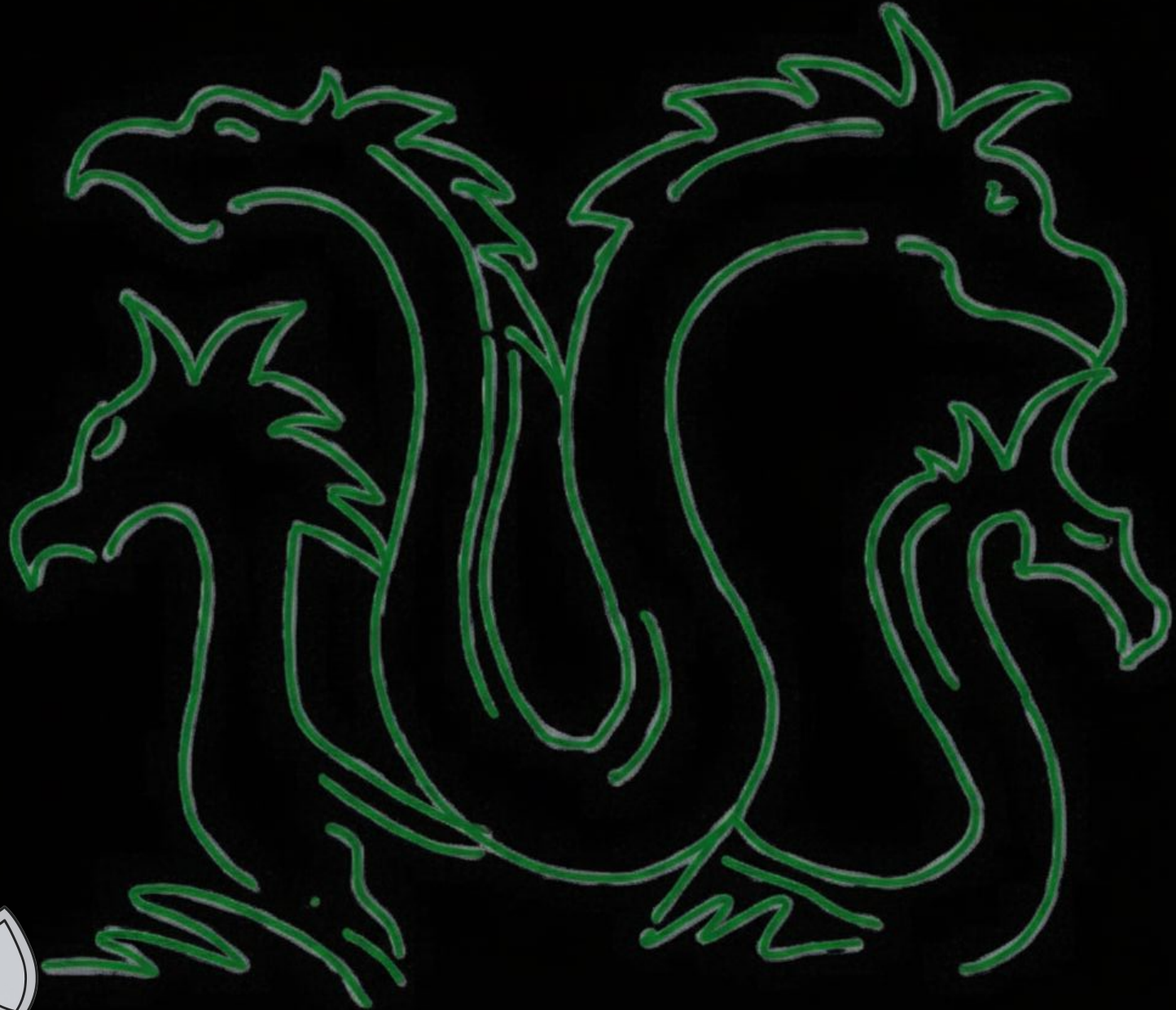
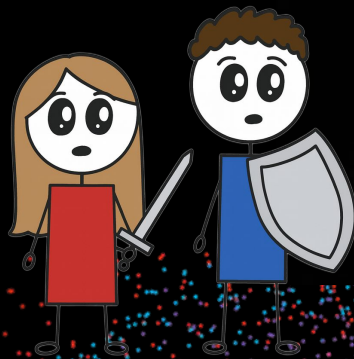
The Call to Adventure: User Story

As a GenAI application developer, I should be able to easily compare the performance and reliability of different models to make an informed decision as to what models (ideally narrowed down to a few models) are best for my use case(s) and application(s).

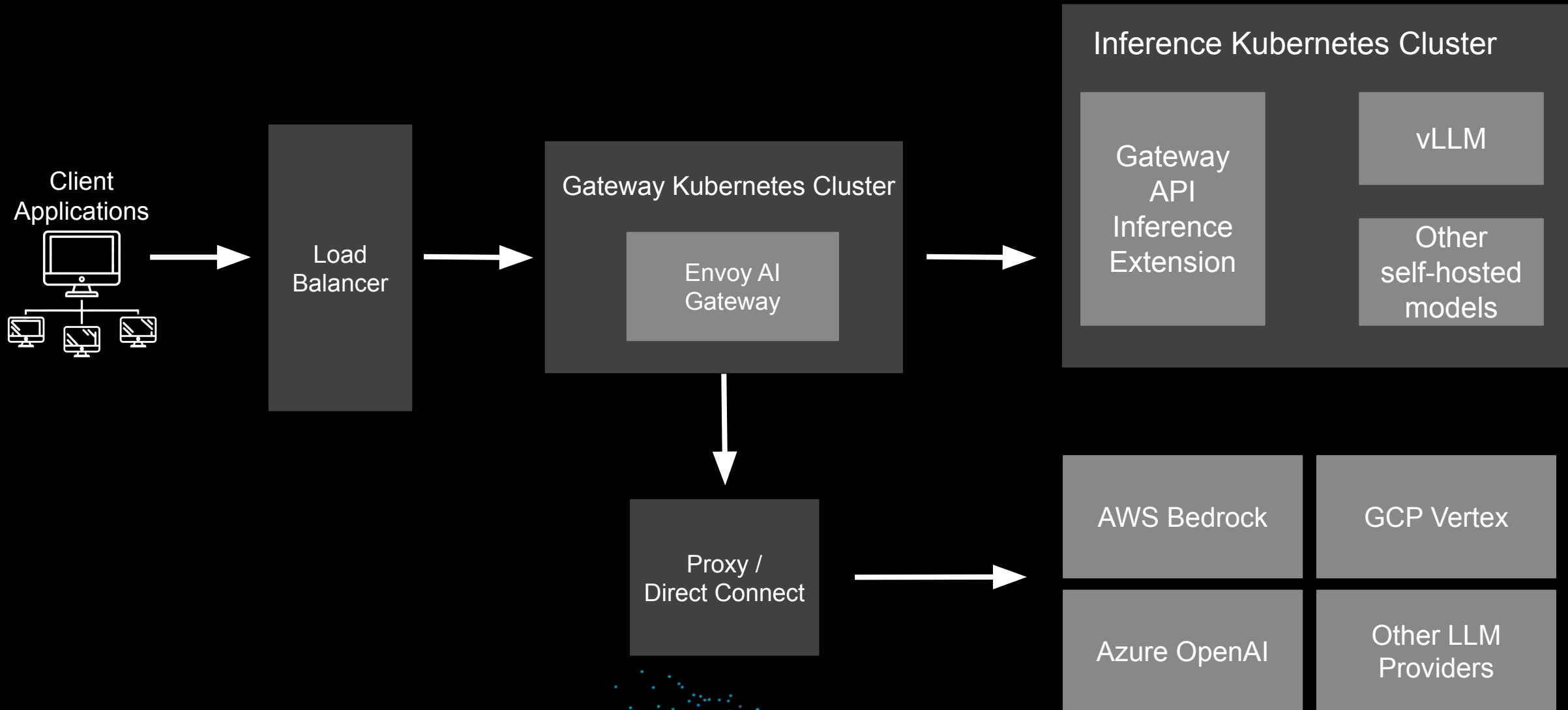
GenAI Hydra

Key issues/expectations for LLMs:

- **Non-deterministic token usage:** Variability in token generation impacts cost and performance
- **Correctness of LLM output:** Is the information accurate and relevant?

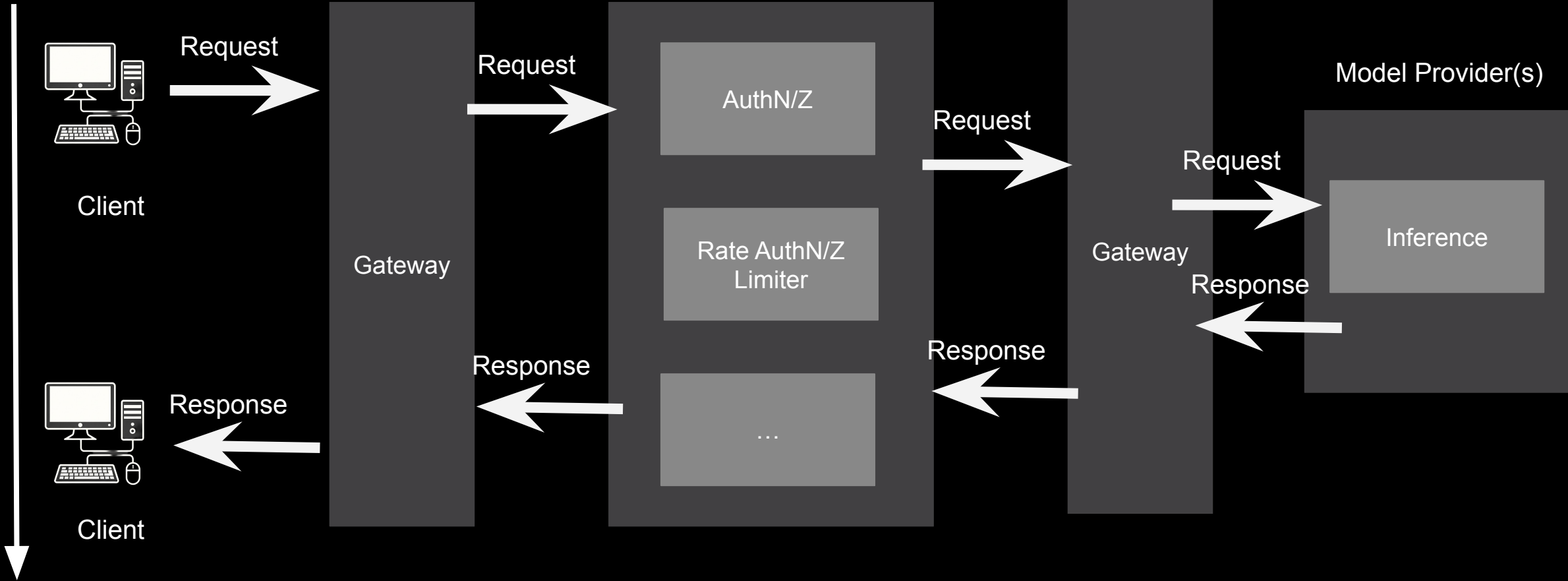


Mapping Our AI Platform Realm



The Request's Quest

Time



Tiers of Enchantment

Client-Facing SLOs (Quality of Service)

- Define user experience standards (latency, reliability)
- Priority Levels
 - High (user-facing, realtime)
 - Medium (internal, real-time)
 - Low (batch, non-urgent)

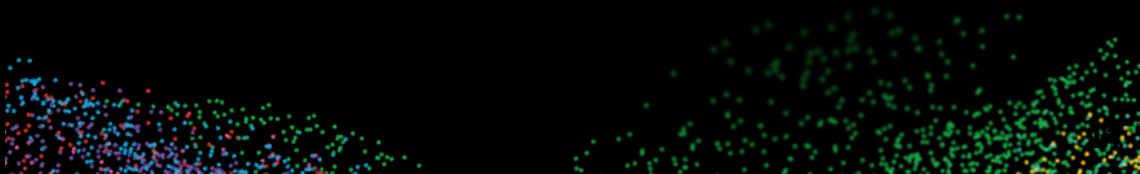
Platform-Level Metrics

- Track internal infrastructure health
- Guide scaling and resource planning

Model-Level Metrics

- Performance of models (self-hosted/vendor)
- Metrics: reliability, latency, cost-efficiency

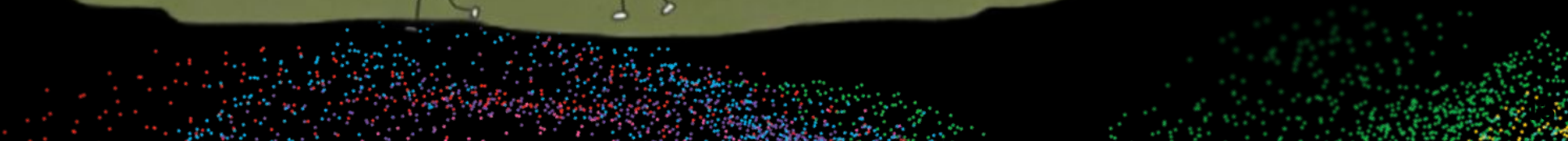
Base Ingredients for Reliability



Base Ingredients for Reliability



Illuminating the Hydra's Lair



Scribing Spells: AI-Native Metrics

Latency metrics for Generative AI:

TTFT (Time to First Token):

- Time until the first piece of the model's response is received

ITL (Inter-Token Latency):

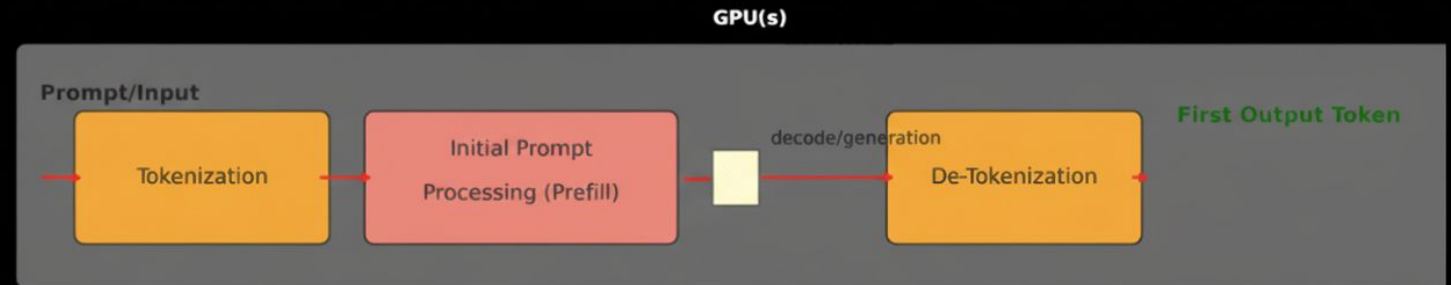
- Time between subsequent tokens in a streaming response

Total Latency (E2E Latency):

- The overall time from when the request is sent until the final token is received

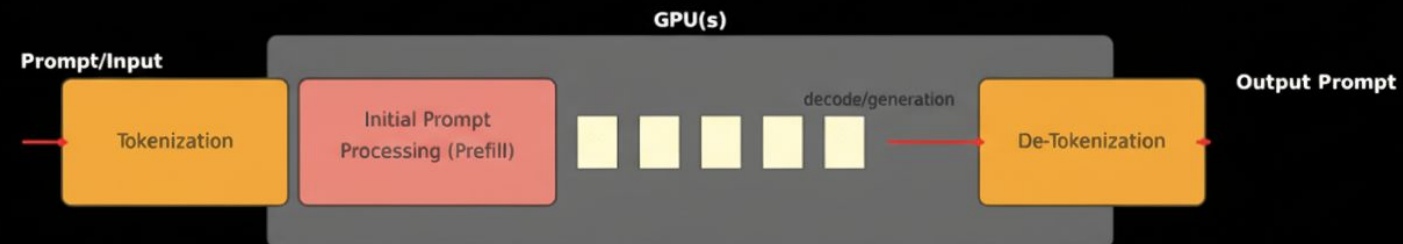
TGT (Token Generation Time):

- The time it takes to stream all tokens after the first one
- $TGT = E2EL - TTFT$

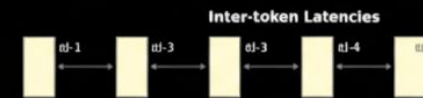


Time to First Token (TTFT)

Signifies time it takes to process the prompt and generate the first token



Inter-token Latencies - Should be constant over all token generations.
Constant time implies efficient generation



The Resource Battlefield



SLOs: Ancient Runes of Reliability

Service Level Indicator (SLI)

Metric generated by a query

Objective

The desired performance level for the SLI

Target Value

How often the Objective must be met

Time Window

The duration over which the target value is measured

SLOs: Ancient Runes of Reliability



The Four Arcane Components: SLOs

- Service Level Indicator (SLI)
 - Time To First Token - user perceived “snappiness”
- Objective
 - $\leq 500\text{ms}$
- Target value
 - 99.9%
- Time window
 - Rolling 1 day window
- Error budget
 - 0.1% of traffic over any day can take longer than 500ms

TechAtBloomberg.com

© 2024 Bloomberg Finance L.P. All rights reserved.

Bloomberg

Engineering

The Crystal Ball of Reliability

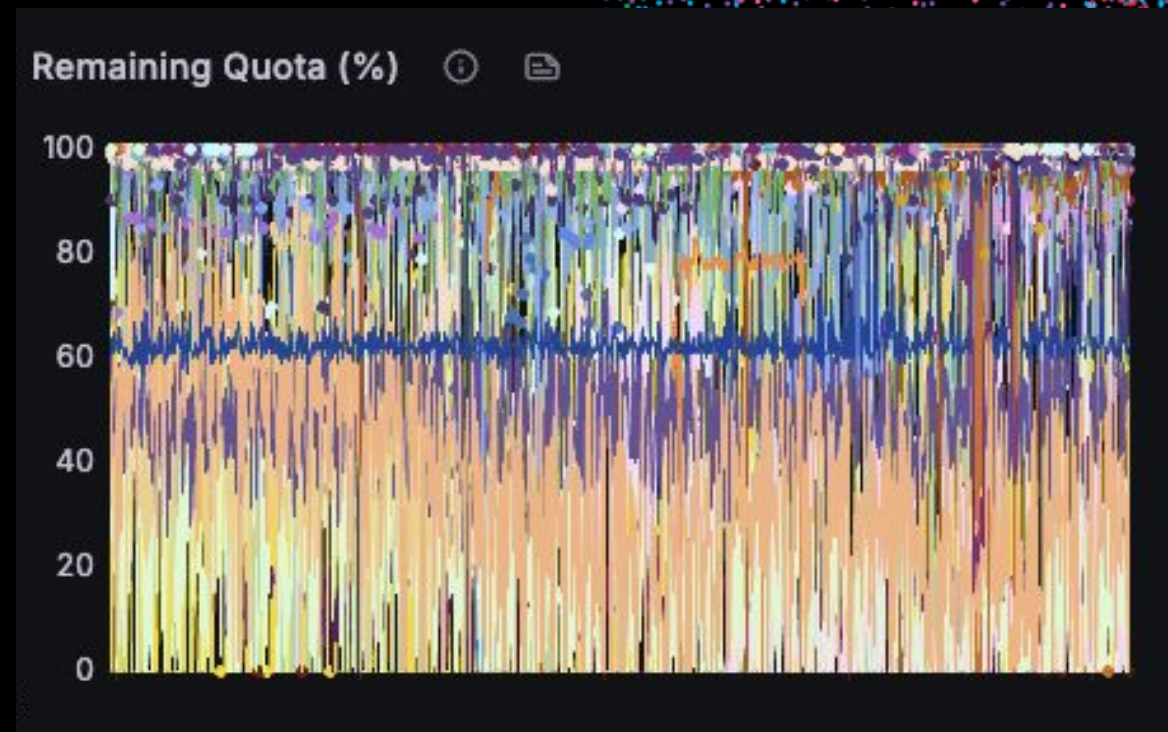
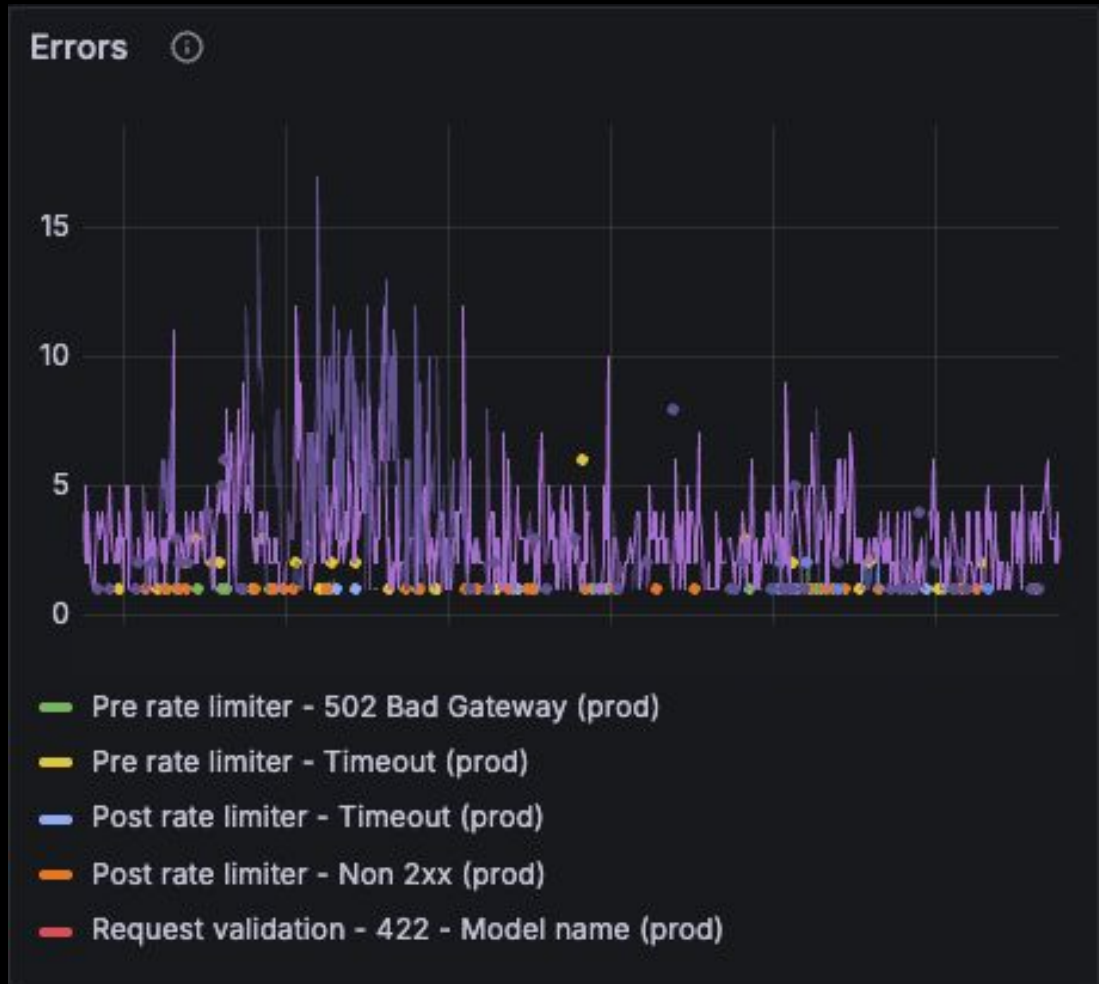




Disclaimer

All of the dashboards used in this presentation are for the purpose of demonstration. They don't indicate the actual of state of systems in any organization.

Natural 1 Examples



Welcome to the Dashboard Dungeon

SLO Target

99.9%

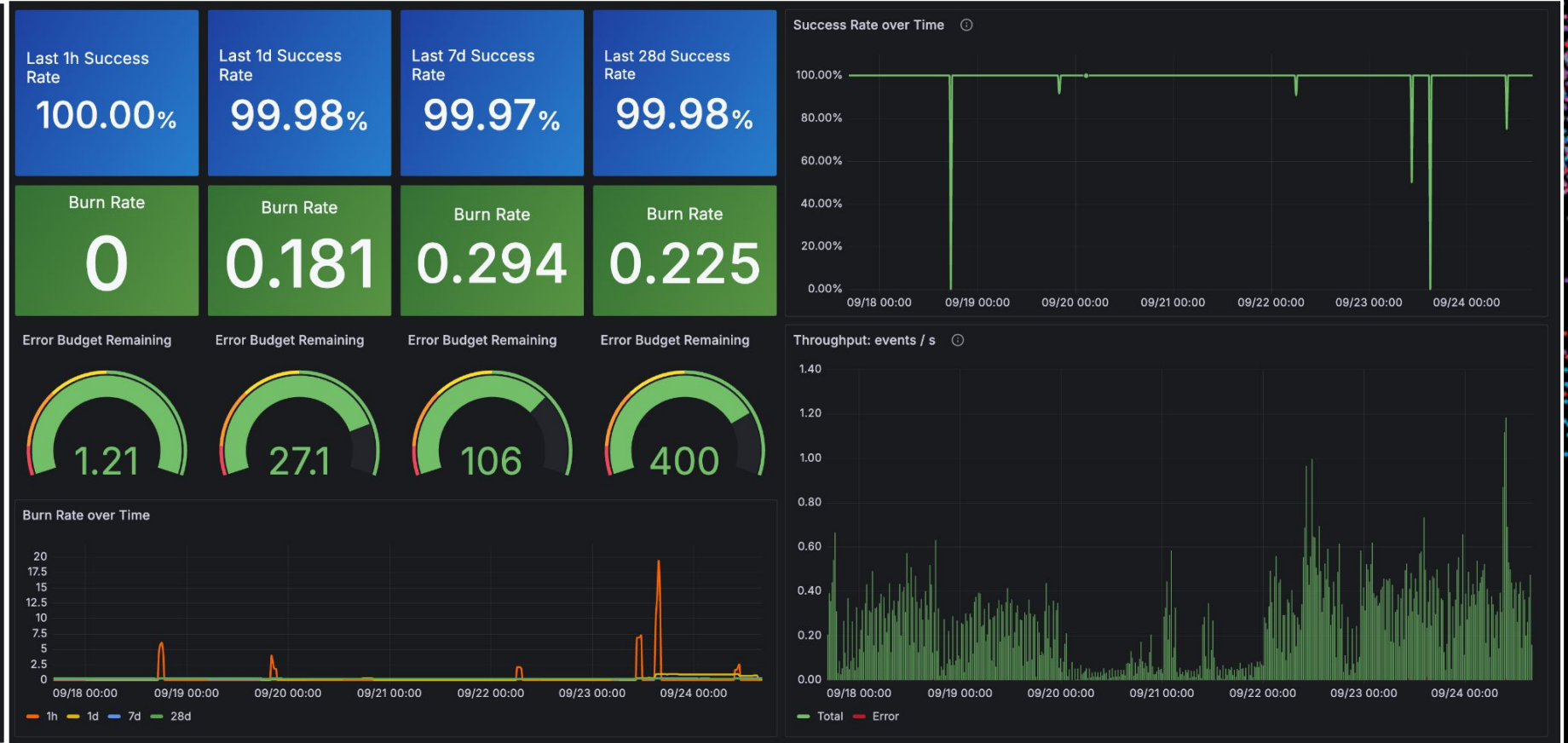
- **Type** - Errors
- **Specification** - Success rate of requests should be > 99.9%
- **Implementation** - The % of API calls that result in non-5XX HTTP codes

User Story

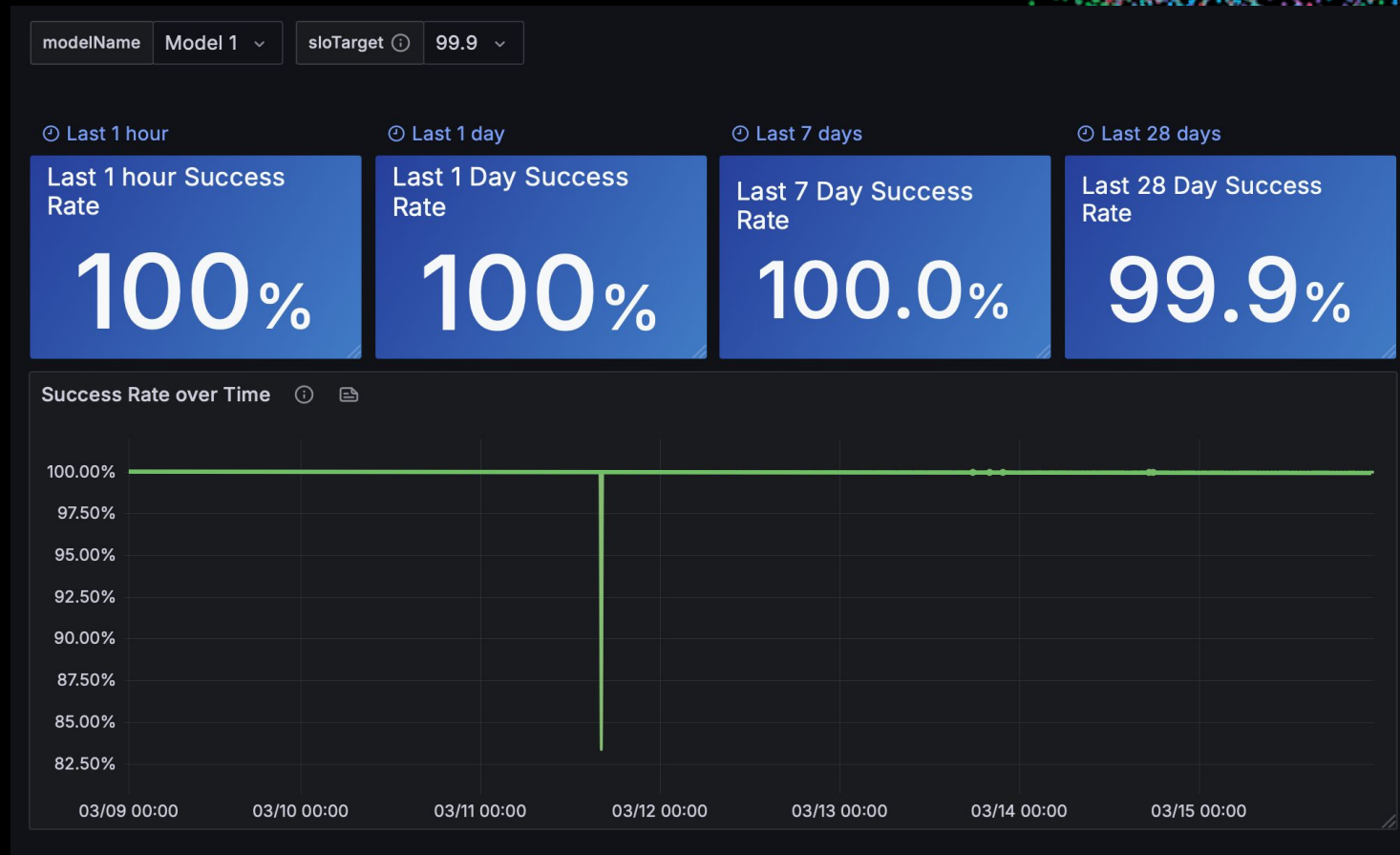
As a user, I want to get a successful response from this LLM Inference service for each request

Links

- [Service Error Log](#)
- [SLO Configuration File](#)
- [Good Time Series](#)
- [Total Time Series](#)



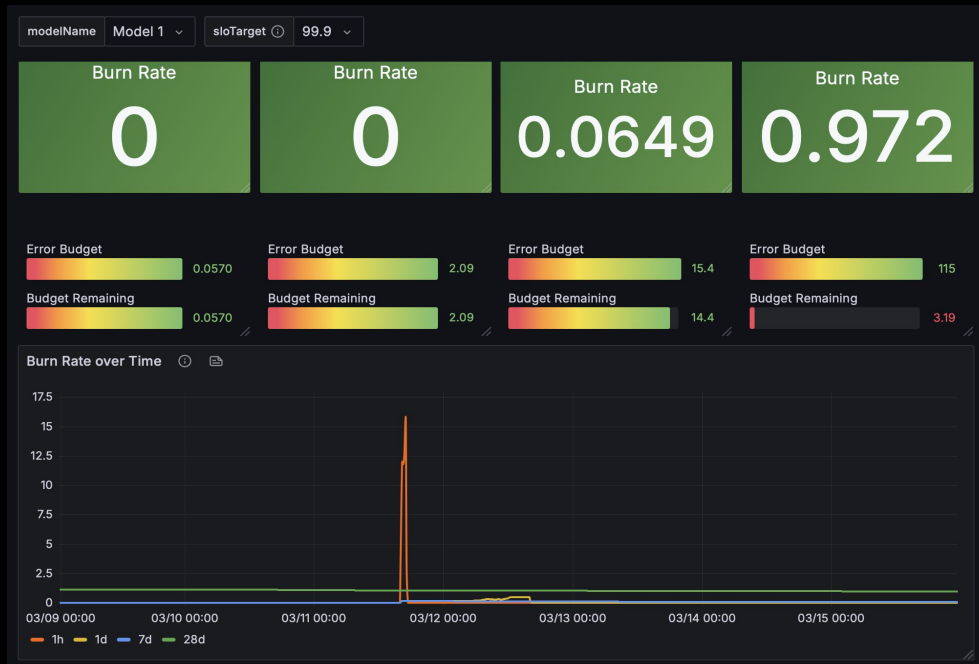
Did you make your saving throw?



TechAtBloomberg.com

$$\text{Success Rate} = \frac{\text{Good Events}}{\text{Valid Events}} \times 100$$

Did the Fireball hit?



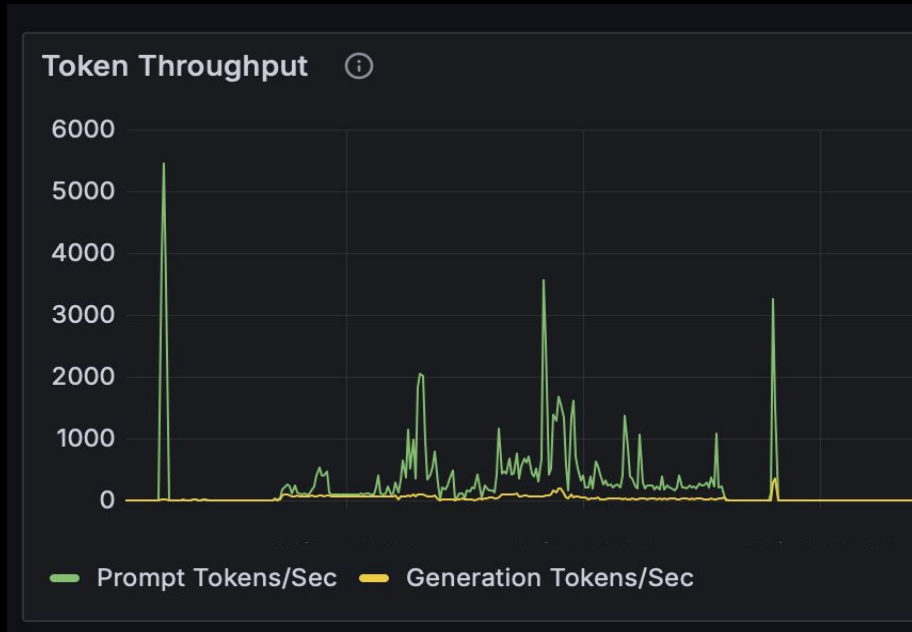
$$\text{Burn Rate} = \frac{\text{Errors observed in the time window}}{\text{Error budget for the time window}}$$

Demonstrates how burn rate across multiple time windows helps distinguish transient incidents from persistent issues

Arrows in the Air



The Speed of Tokens



When comparing the inference speed of LLMs, how long should the prompts and completions be?

Typical use cases benchmarks:

- **Single-turn chat:** ~3:1 prompt:generation ratio
- **Multi-turn chat:** ~10:1 prompt:generation ratio
- **Extensive summarization:** ~30:1 prompt:generation ratio

Casting Speed: Streaming Latency

Streaming Latency

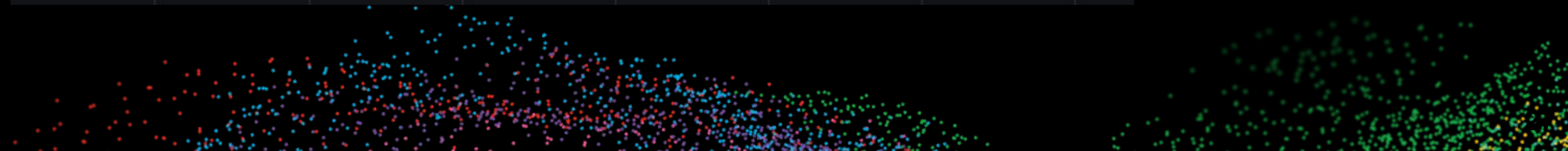
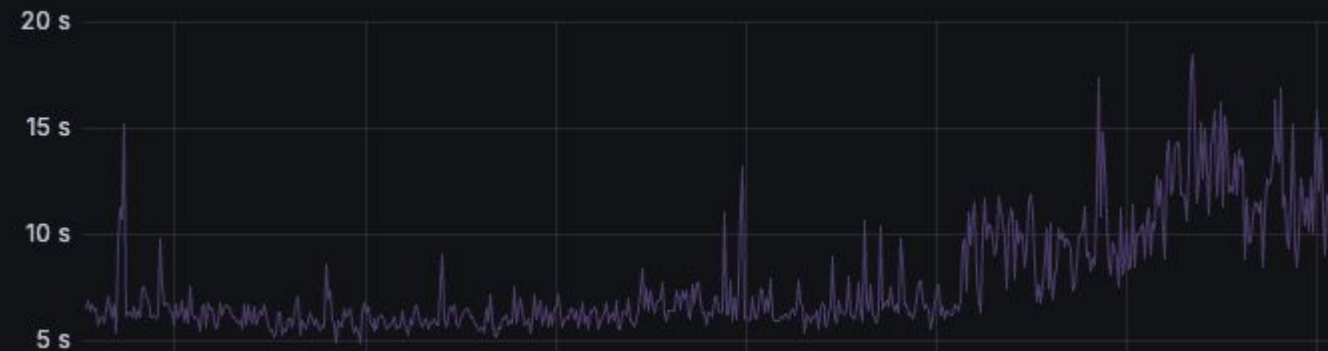
TTFT Per Model



ITL Per Model





Generate Latency



Latency Labyrinth: End-to-End Performance

Latency

Average p90 of Latency

Generate Latency  

p90 Latency - now

6.50 s

p90 Latency - 1d

8.42 s

p90 Latency - 7d

4.98 s

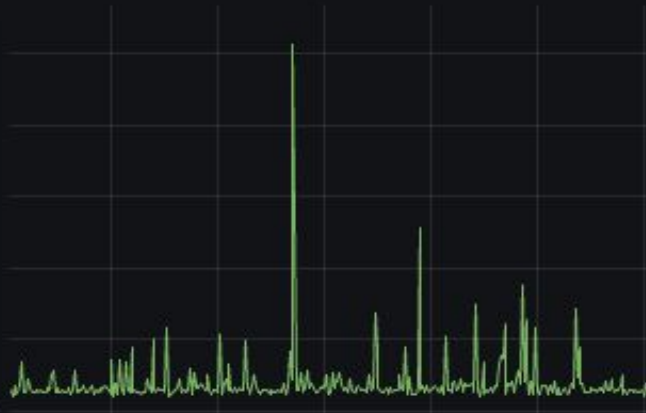


Preprocess Latency MS

Generate Latency MS

Post Process Latency MS

0 ms



0 ms



0 ms



Whiffs, Fumbles, and Critical Misses

Total Errors

Total count of Errors (Not a rate)

Total 5XX Count ⓘ ⓘ

5XX Errors - Now

4280

5XX Errors - 1d

1688

5XX Errors -7d

405

Total 4XX Count ⓘ ⓘ

4XX Errors - Now

18951

4XX Errors - 1d

20834

4XX Errors -7d

16804

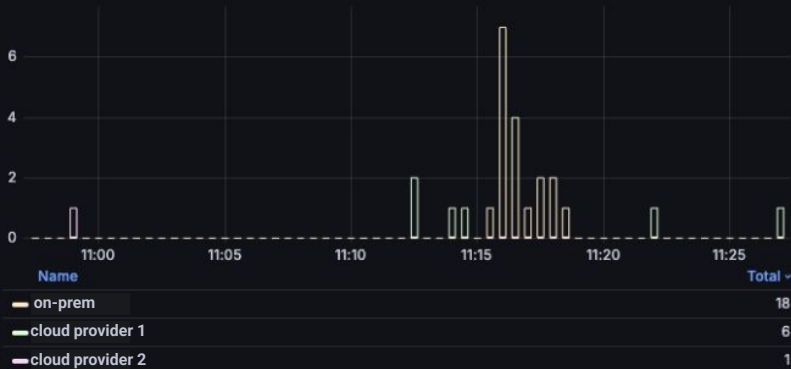
The Spell Backfires: 5XX Errors

5xx Errors

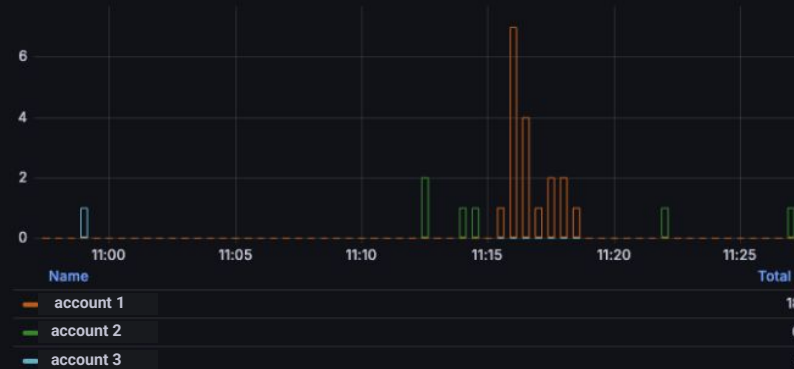
External dashboard links

- [Cloud Provider 1](#)
- [Cloud Provider 2](#)
- [Cloud Provider 3](#)

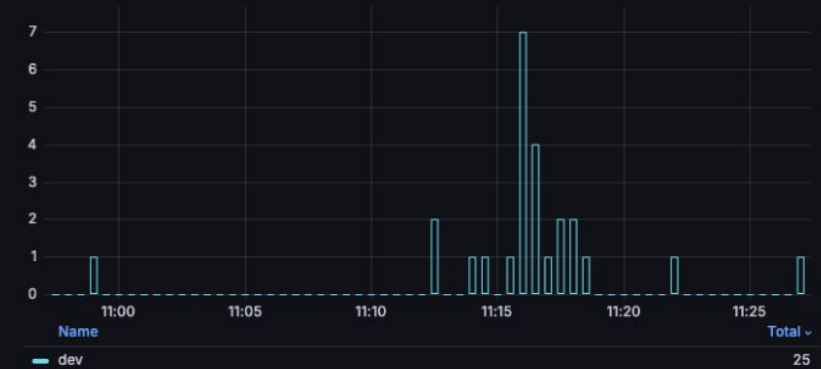
Cloud Providers with 5XX errors ⓘ



Accounts with 5XX errors ⓘ



Environment 5XX errors ⓘ

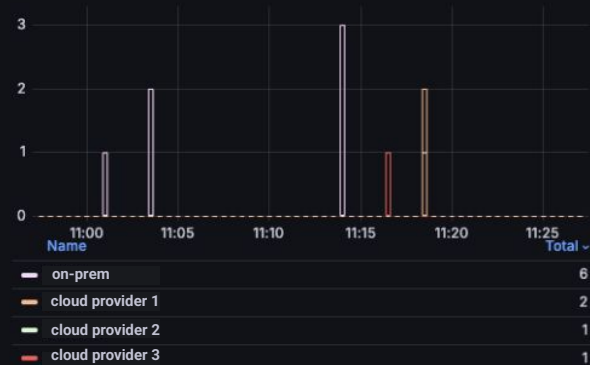


Wrong Key, Wrong Door: 4XX Errors

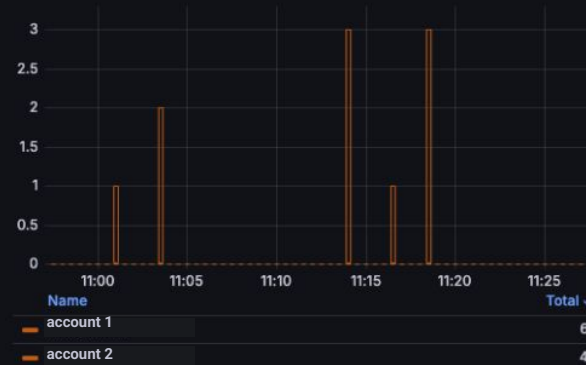
Common 4XX Errors with Description

Code	Name	Description
400	Bad Request	The request could not be understood or was missing required parameters.
401	Unauthorized	Authentication failed or user does not have permissions for the desired action.
403	Forbidden	Authentication succeeded but the authenticated user does not have access.
404	Not Found	The requested resource could not be found.
409	Conflict	A request conflict with the current state of the server (e.g., duplicate data).
422	Unprocessable Entity	The request was well-formed but could not be processed due to semantic errors.
429	Too Many Requests	The user has sent too many requests in a given amount of time (rate limit).

Cloud Providers with 4XX errors



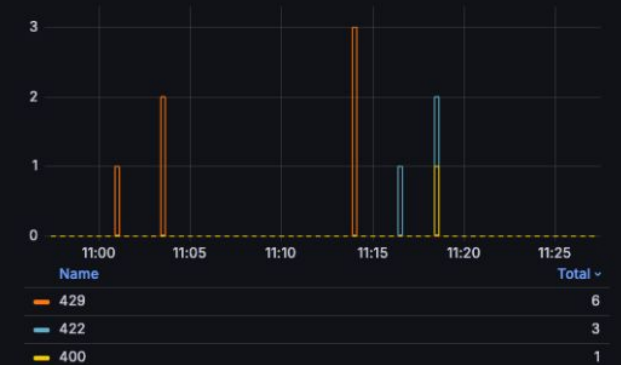
Accounts with 4XX errors



Environments with 4XX errors



Count of each 4XX Error Type

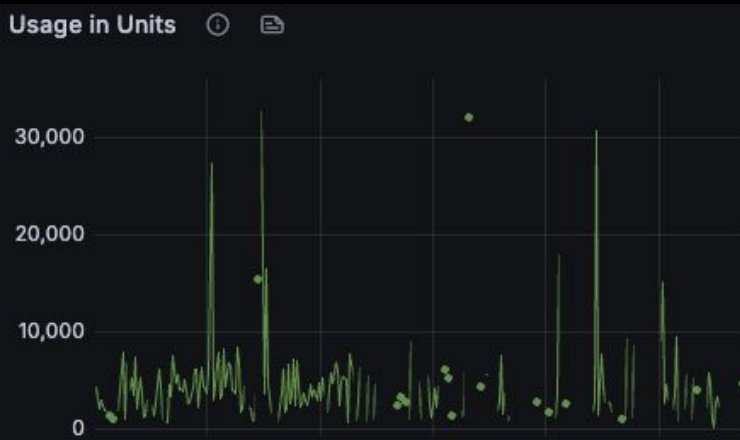


4XX Error Traces - Filters not working

Trace ID	Start time	Service	Name	Duration												
exampletrace1	2026-09-05 11:33:35	exampleservice1	POST /v1/completions	561 ms												
<table border="1"> <thead> <tr> <th>Span ID</th> <th>Start time</th> <th>error.status_code</th> <th>service.name</th> <th>status</th> <th>Duration</th> </tr> </thead> <tbody> <tr> <td>examplespan1</td> <td>2026-09-05 10:33:06</td> <td>400</td> <td>exampleservicename1</td> <td>error</td> <td>539 ms</td> </tr> </tbody> </table>					Span ID	Start time	error.status_code	service.name	status	Duration	examplespan1	2026-09-05 10:33:06	400	exampleservicename1	error	539 ms
Span ID	Start time	error.status_code	service.name	status	Duration											
examplespan1	2026-09-05 10:33:06	400	exampleservicename1	error	539 ms											
exampletrace2	2026-09-05 10:31:49	exampleservice2	POST /v1/completions	33 ms												
exampletrace3	2026-09-04 12:18:35	exampleservice1	POST /v1/completions	44 ms												
exampletrace4	2026-09-04 11:10:20	exampleservice3	POST /v1/chat/completions	12.1 s												

The Expedition's Rations: Capacity & Usage

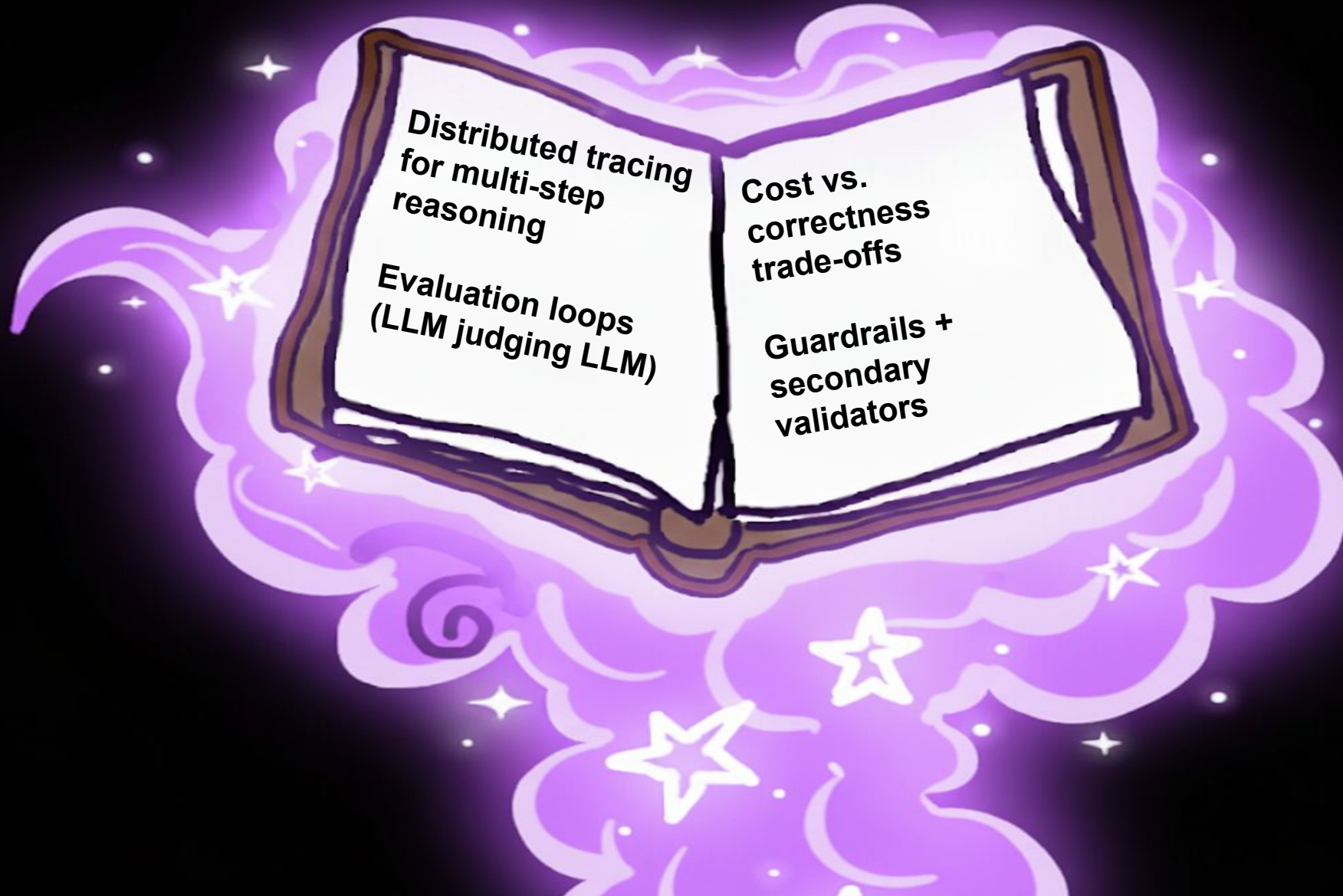
Usage and Quota



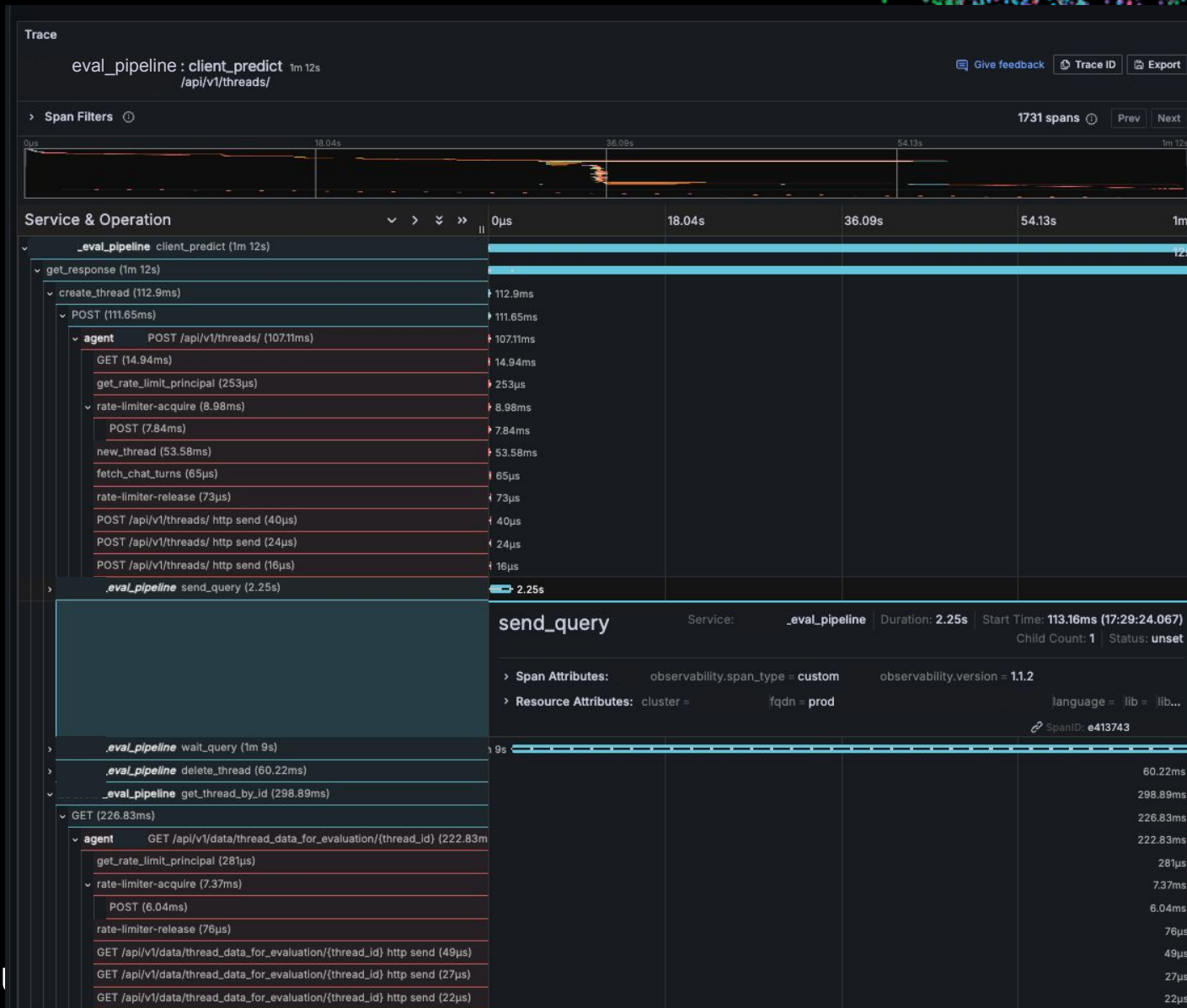
The Expedition's Rations: Capacity & Usage



More Spells: Agentic Observability



Illuminating the Shadows with Tracing



Traces give

can't provide

Agentic Observability

← Back to Search

asdkjekljkfkladsjfa3jhdkfah5a

Resources **agentsmith (683)** **ssr-retriever (210)**

Duration 72.17 s
Trace Start 02-Oct-2025

Trace Details | LLM Usage

Collapse All Errors Logs Tool Spans LLM Spans Custom Spans

Resource / Operation	Span Type	Duration	Timeline
<input checked="" type="checkbox"/> agentsmith get_llm_response	Custom	9.54 s	
<input checked="" type="checkbox"/> agentsmith get_llm_response	Custom	4.06 s	
<input checked="" type="checkbox"/> agentsmith get_llm_response	Custom	3.25 s	
<input checked="" type="checkbox"/> agentsmith get_llm_response	Custom	5.07 s	
<input checked="" type="checkbox"/> agentsmith get_llm_response	Custom	4.84 s	
<input checked="" type="checkbox"/> agentsmith get_llm_response	Custom	14.28 s	

agentsmith
get_llm_response

Tags | **Logs**

example logs
examples logs

Evaluation Check: Appraising the Treasure

← Back to All Runs

app_eval

Compare ▾

Export All ▾

Run ID	16a1b378	Start Time	30-Sep-2025	Config	View
Application		End Time	30-Sep-2025	Dataset	travel-dataset
Experiment	demo_l&l	Duration	45s	Runtime Info	-

Expand Rows

Input	Output	Expectation	concise	fun	get-current-time-routing-accur...	recommend-activity_rou
<i>Enter Input</i>	<i>Enter Output</i>	<i>Enter Expectation</i>	<i>Enter Value</i>	<i>Enter Value</i>	<i>Enter Value</i>	<i>Enter Value</i>
Ouagadougou	You should go to the jazz club.	<pre>{ "reference_output": null, "metadata": { "expected_tool_calls": ["get-current-time", "recommend-activity"] } }</pre>	5	6	1	
New York	You should go to the jazz club.	<pre>{ "reference_output": null, "metadata": { "expected_tool_calls": ["get-current-time", "recommend-activity"] } }</pre>	4	8	1	
the moon	Based on our discussion and the tool results, I recommend that you take a walk in the park on the moon. Please note that this answer may be incomplete, and more information could be requested if needed, as the moon does not have a traditional park or	<pre>{ "reference_output": null, "metadata": { "expected_tool_calls": [] } }</pre>	1			

View Metric: concise

Metric 1 of 4

Value
5

Explanation

The travel plan is very brief and directly states the activity. It does not contain any unnecessary information or repetitive phrases.

Evaluator

TravelPlanEvaluator

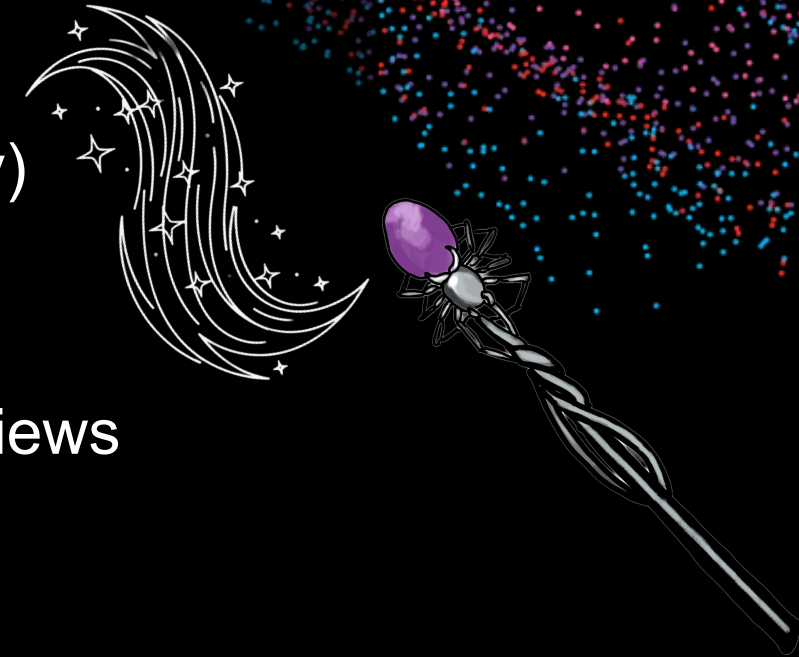
Previous

Next

Close

Conjure your Alerts with the Right SLOs

- **% Remaining** → *How much error budget do we have left?*
 - **Burn Rate** → How much error budget is actively being used?
 - **Time till Exhausted** → How much time do we have until the budget is gone!
-
- Define environment-specific thresholds (Prod vs. Dev)
 - Set meaningful alert conditions, not just noisy ones
 - Prioritize alerts to avoid fatigue and escalation gaps
 - Tune iteratively using incident feedback and SLO reviews

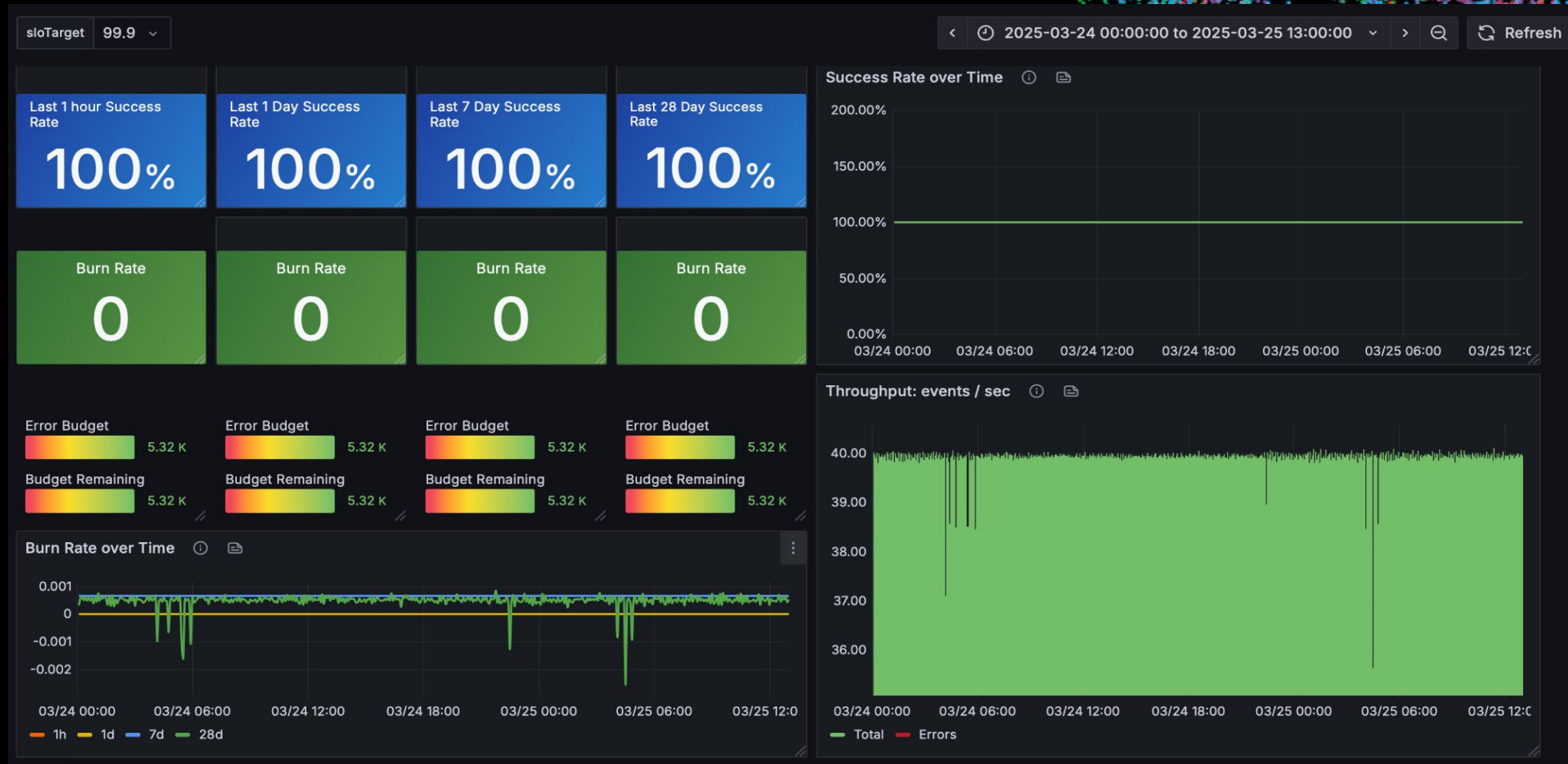


Redefining SLOs

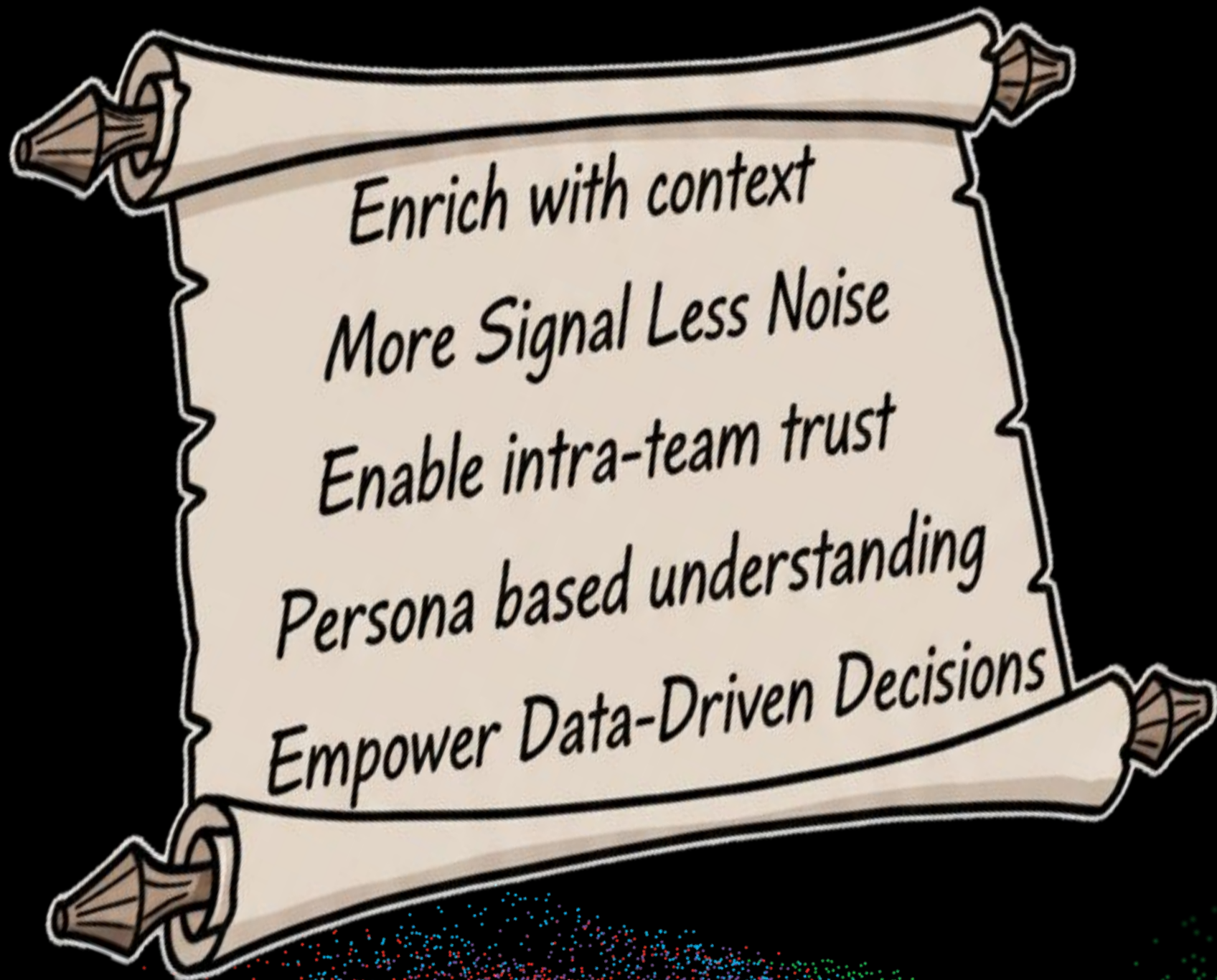


Rapid consumption of your budget is like a dragon on the rampage!

Redefining SLOs



Dreamland ✨ Too good to be true?



Enrich with context

More Signal Less Noise

Enable intra-team trust

Persona based understanding

Empower Data-Driven Decisions

Are You a Hero Looking for Your Next Quest?

We are hiring: [bloomberg.com/engineering](https://www.bloomberg.com/engineering)

- Gen AI Platform Team Lead - AI Engineering (London)
- Senior Software Engineer - Data Technologies Compute Platform (London)
- Senior Software Engineer - AI Hardware (NYC)
- Senior ML Ops Engineer - AI Engineering (NYC)
- Senior Software Engineer - Bare Metal as a Service (NYC)
- Senior Software Engineer - Public Cloud IAM and Org Management (NYC)
- Senior Software Engineer - Public Cloud Pipelines (NYC)
- Senior Software Engineer - Public Cloud Visibility (NYC)
- Senior Java Engineer - Search Infrastructure (NYC)
- Technical Product Manager, GenAI Developer Platform - CTO Office (NYC)
- Technical Product Manager, LLM Platforms - CTO Office (NYC)



TechAtBloomberg.com

© 2025 Bloomberg Finance L.P. All rights reserved.

Bloomberg

Engineering

The Campaign Continues...

Thank you!
How can we help you
on your quest?

