

# OpenLLMetrics



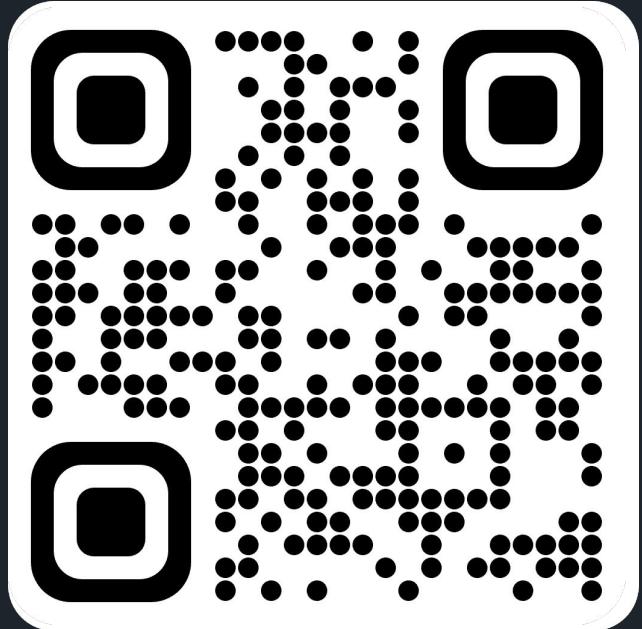
# \$whoami\_



**Zameer Fouzan**

Lead Developer Relations Engineer -  
APJ

- Full Stack Developer 😎
- Open Telemetry Advocate at NR
- AI RAG Picker



A composite image featuring a young girl in the foreground and a burning house in the background. The girl has light brown hair and is looking directly at the viewer with a neutral expression. In the background, a single-story house is engulfed in bright orange and yellow flames, with smoke billowing out of the windows and roof. A yellow fire hose lies on the ground in front of the house, and several firefighters are visible, though they are out of focus. The overall mood is somber and ironic.

**WORKED FINE  
IN DEV**

**OPS PROBLEM NOW**

# Every single company on earth



# LLMs

AI apps are **unreliable** blackboxes

# How old is Anil Kapoor

A screenshot of a GitHub repository page for "osoderholm/langchain-rag-demo". The page shows a green parrot icon and a blue link icon. The repository is described as a "Demonstration of RAG using LangChain". It contains files like ".gitignore", "LICENSE", "README.md", and "ai.py". The repository has 1 branch, 0 tags, 1 commit, 1 fork, and 0 stars. The commit was made by "osoderholm" 8 months ago.

Code Issues Pull requests Actions Projects Security Insights

osoderholm / langchain-rag-demo Public

Code Issues Pull requests Actions Projects Security Insights

osoderholm initial commit Initial commit 8 months ago 1 Commit

.gitignore Initial commit 8 months ago

LICENSE Initial commit 8 months ago

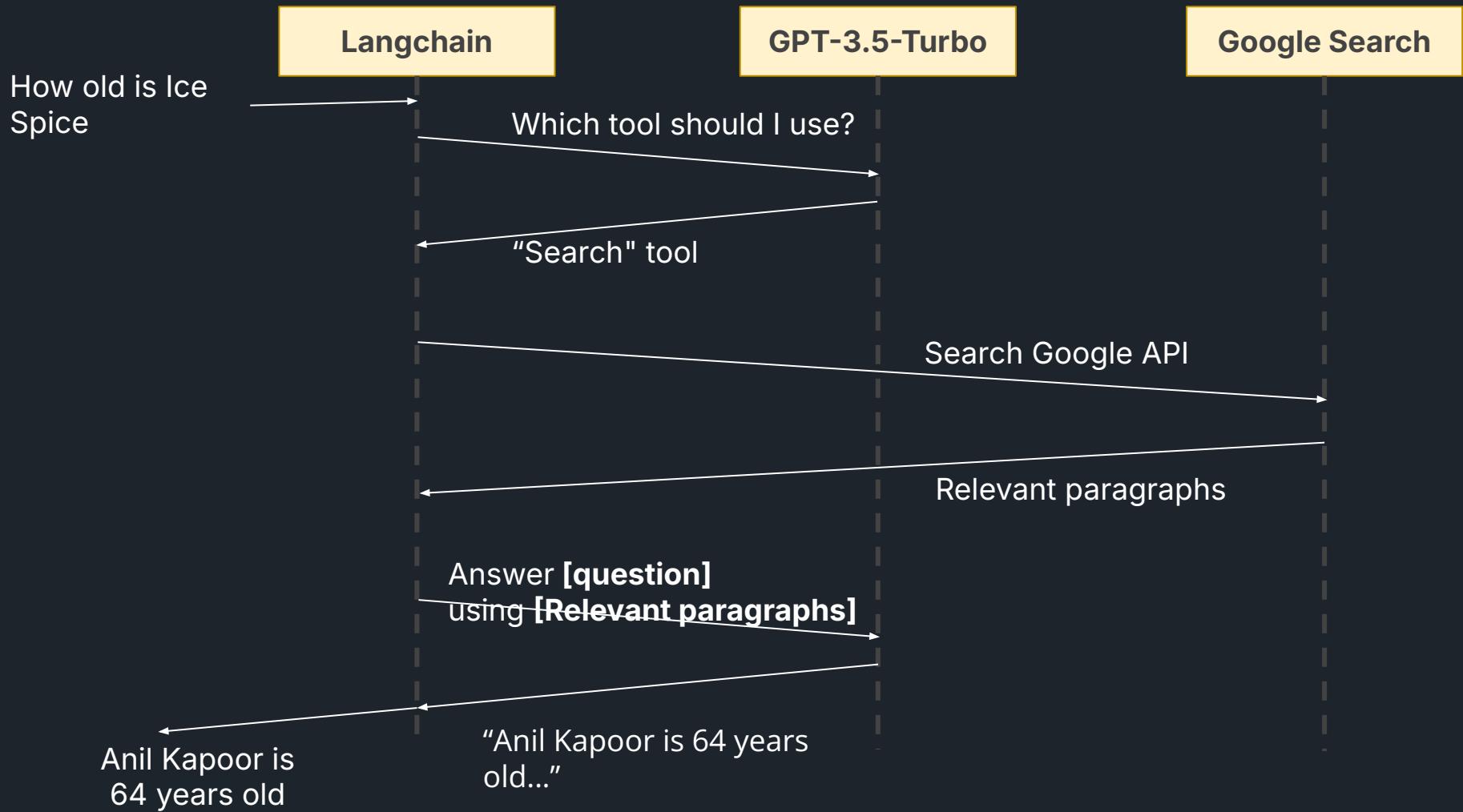
README.md Initial commit 8 months ago

ai.py Initial commit 8 months ago

About

Demonstration of RAG using LangChain

Readme MIT license Activity 0 stars 1 watching 1 fork



INPUT\*

what is anil kapoors age divided by total number of world records held by sachin?



> Entering new AgentExecutor chain...

First, I need to find out Anil Kapoor's age. Then, I need to find out the total number of world records held by Sachin Tendulkar.

Action: Search

Action Input: "Anil Kapoor age"

Observation: 67 years

Thought: Now that I know Anil Kapoor's age, I need to find out the total number of world records held by Sachin Tendulkar.

Action: Search

Action Input: "Total number of world records held by Sachin Tendulkar"

Observation: ['Sachin Tendulkar . Allround: Batting and fielding. record. 5000 runs and 50 fielding dismissals . Batting: Most runs. 1st. Most runs r (15921) . Batting: ...', 'Sachin still holds the record with 51 test centuries. The master blaster has a mindblowing record of scoring 100 centur international ...', 'Sachin Tendulkar Career Stats ; List A, 551, 538, 55, 21999 ; T20s, 96, 96, 11, 2797 ...', 'Tendulkar played 664 international c tches in total, scoring 34,357 runs. ... He also holds the world record for playing the highest number of Test ...', 'Most runs in World Cups. Sachin kar's record in the World Cup is second to none. He played a record six World Cups from 1992 to 2011.', 'As the Master Blaster turns 51 today (April re's a look at 15 most prominent records created by Sachin Tendulkar during his illustrious ...', 'List Of Records Of Sachin Tendulkar - Live • 1st lklekele SL 381/3 AFG(17/50 ov, T:382) 93/5 Afghanistan need 289 from 33 overs.', 'Sachin has been breaking cricket records ever since his debut in 1 is the Highest Run scorer in the world cup (1,796 at an average of 59) He is one ...', 'He scored 15,921 runs in the game's longest format, while hi lly stood at 18, 426 runs and he got out for 10 in his only T20I appearance.', 'His total of 51 centuries in Test matches is a world record for high er ... List of career achievements by Sachin Tendulkar . List of cricketers by ...']

Thought: The search results are not clear about the exact number of world records held by Sachin Tendulkar. I need to refine my search.

Action: Search

Action Input: "How many world records does Sachin Tendulkar hold"

Observation: ['Batting: Most runs ; 1st. Most runs in career (18426) ; 12th. Most runs in an innings (200\*) ; 10th. Most runs in an innings (progress ord holder) (200\*).', 'Hailed as the world's most prolific batsman of all time, he is the all-time highest run-scorer in both ODI and Test cricket w than 18,000 runs and ...', 'Sachin still holds the record with 51 test centuries. The master blaster has a mindblowing record of scoring 100 centur international ...', 'Sachin Tendulkar Career Stats ; List A, 551, 538, 55, 21999 ; T20s, 96, 96, 11, 2797 ...', 'As the Master Blaster turns 51 today 24), here's a look at 15 most prominent records created by Sachin Tendulkar during his illustrious ...', 'He played a record six World Cups from 1 1. He also holds the record of the highest number of runs at the mega event. Sachin has scored ...', 'His total of 51 centuries in Test matches is a record for highest number of centuries by a batsman and his 49 centuries in ODI matches are the second ...', 'Sachin accumulated 18,426 runs in 463 o n average of 44.83. In addition to it, the Virat has now scored 13794 runs in 291 matches at an ...', 'He currently holds a staggering 19 Guinness R ords titles. These are: Most runs scored in One-Day Internationals in a calendar year (male) ...', 'Sashin Tendulkar World Record – Sachin Tendulkar World Cups from 1992 to 2011. He is the highest run scorer in tournament history.']

Thought: The search results indicate that Sachin Tendulkar holds 19 Guinness World Records titles. Now that I have both numbers, I can calculate Anil's age divided by the total number of world records held by Sachin Tendulkar.

Action: Calculator

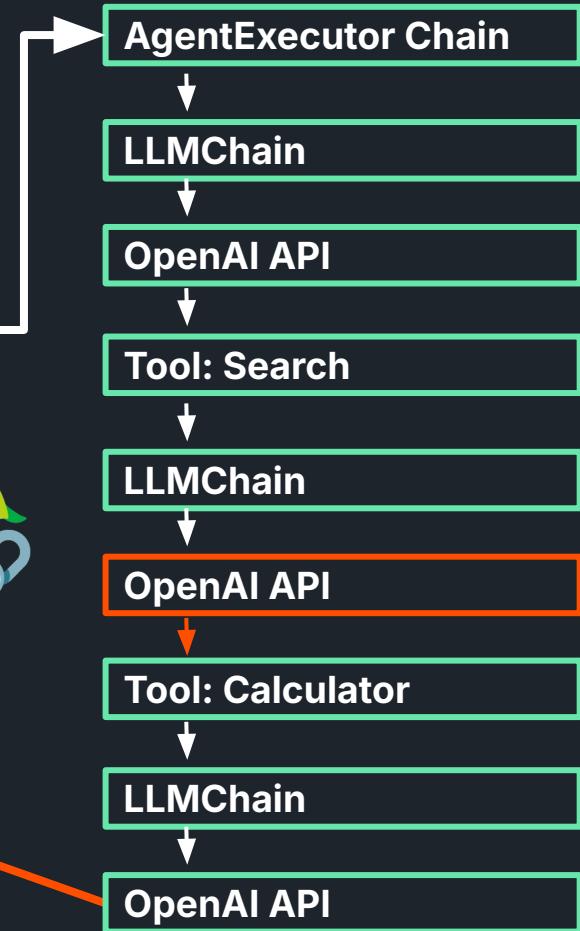
Action Input: 67 / 19

Observation: Answer: 3.526315789473684

introducing

# Tracing

# What is Distributed Tracing?



Ship  
scale

Backed by YC

Get started

Trusted by thou

epsource



Menu

Backed by Y Combinator

# AI With An Immune System

Maitai detects faults in AI output in real-time, takes corrective action, then uses that information to build you more reliable, higher performance AI models.

Get Started

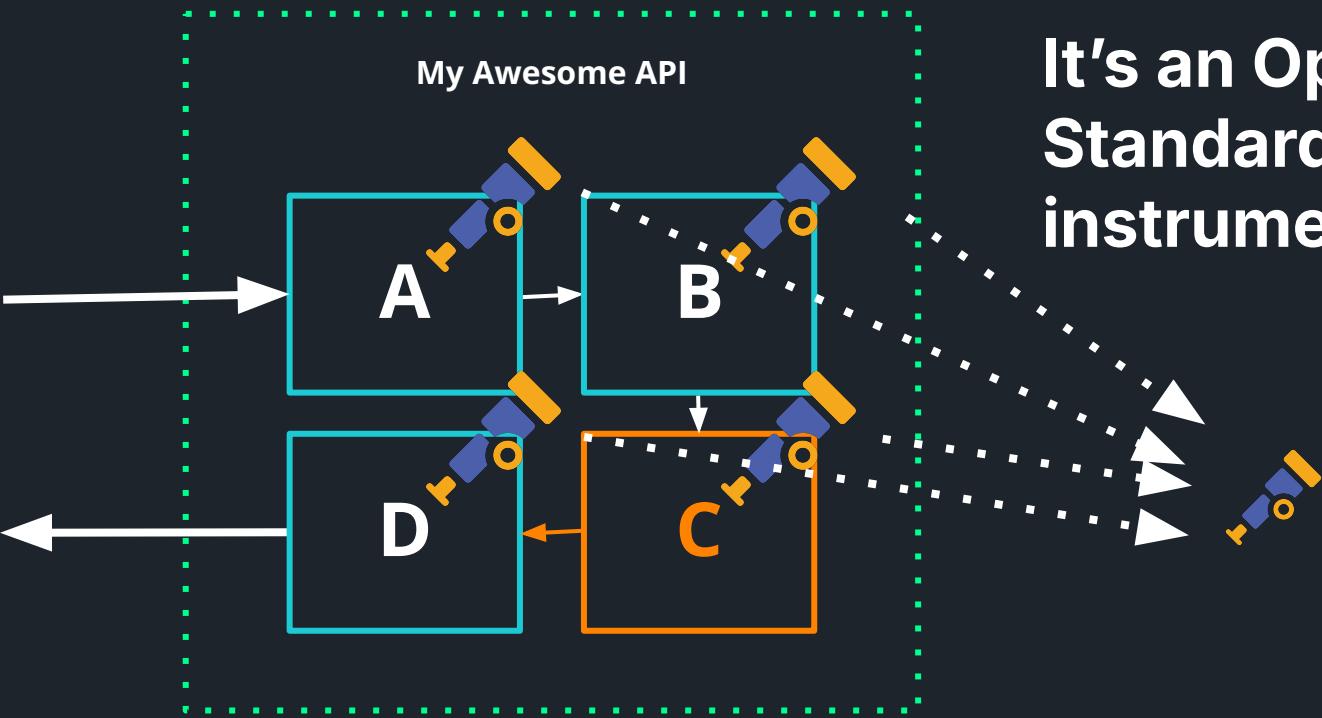




introducing

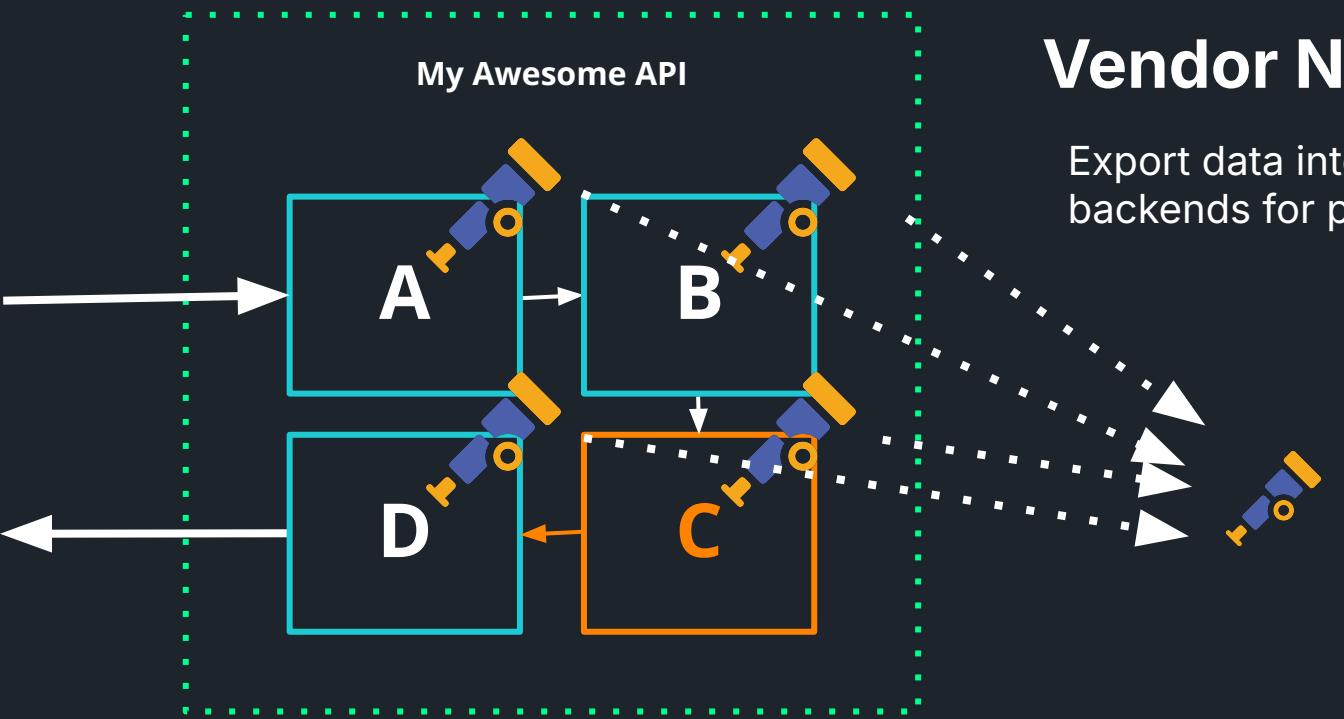
# OpenTelemetry

# What is OpenTelemetry?



**It's an Open Source  
Standard for  
instrumentation**

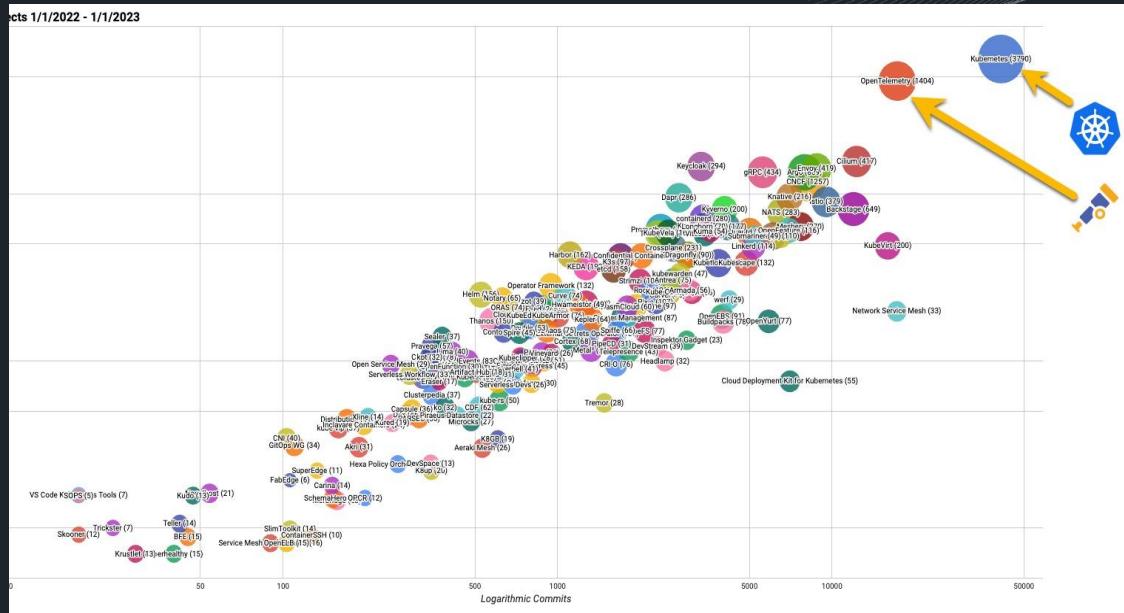
# What is OpenTelemetry?



## Vendor Neutral!

Export data into multiple data  
backends for processing

# The Rise of OpenTelemetry



**OpenTelemetry** is second in CNCF - strong interest in modern **Observability**.

Enhanced support for **Open Standards**, including OpenTelemetry, eBPF, and Grafana - Gartner MQ 2022

# Getting started with OpenTelemetry

## Automatic Instrumentation

100% automatic &  
out-of-the-box  
No custom spans/attributes

Available in select languages

## Manual Instrumentation

Customize spans  
Add additional attributes

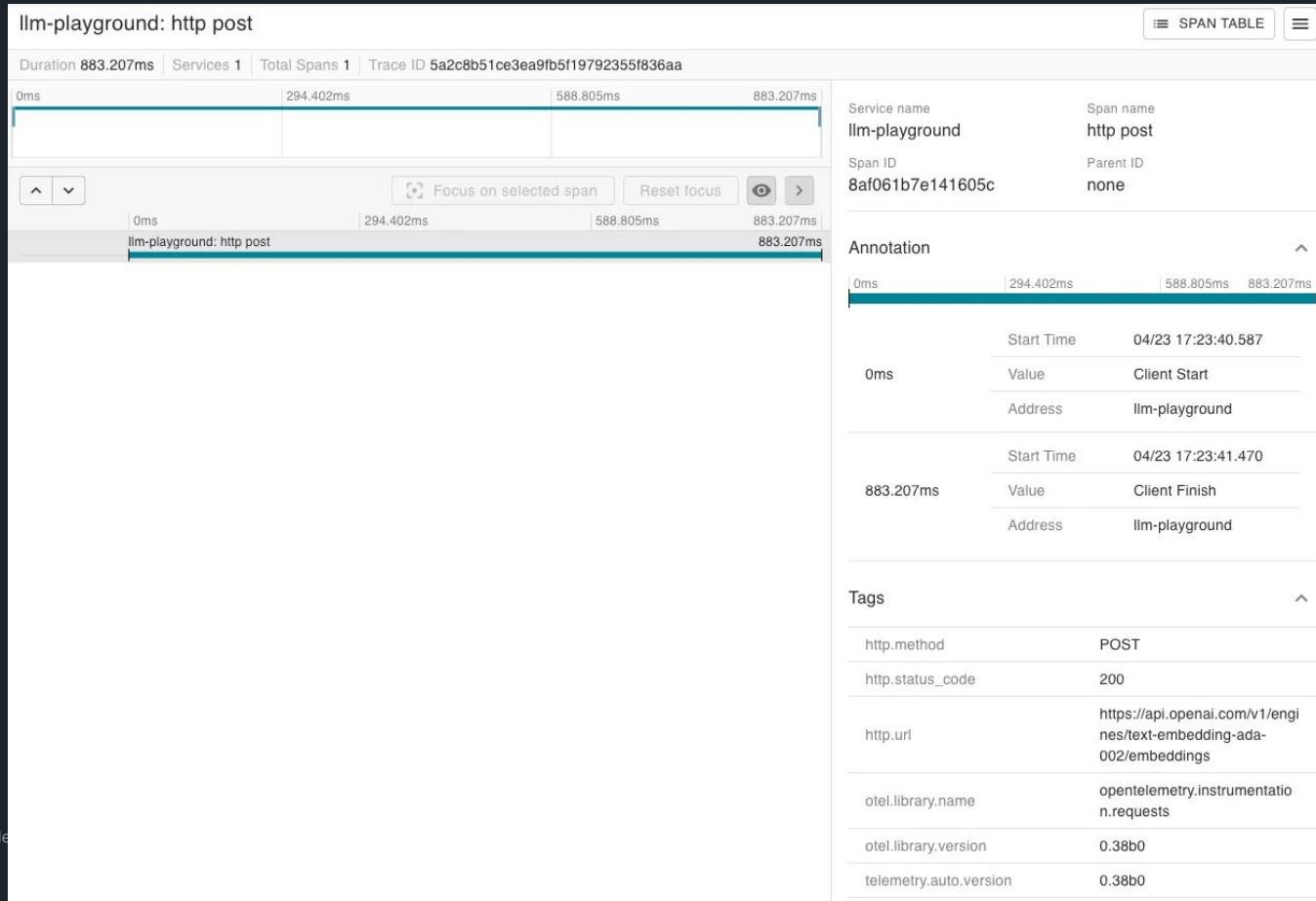
# Automatic Instrumentation



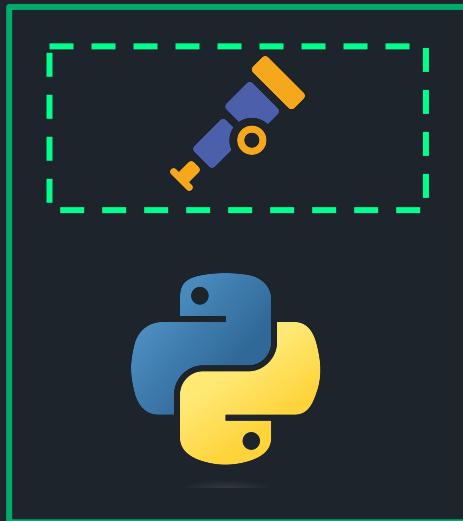
bytecode



```
$ poetry add opentelemetry-instrumentation-requests  
$ poetry add opentelemetry-exporter-otlp  
$ poetry add opentelemetry-distro  
$ poetry run opentelemetry-instrument  
--traces_exporter console,otlp \ --metrics_exporter  
console \ --service_name llm-playground \  
--exporter_otlp_endpoint 0.0.0.0:4317 \ python  
main.py
```



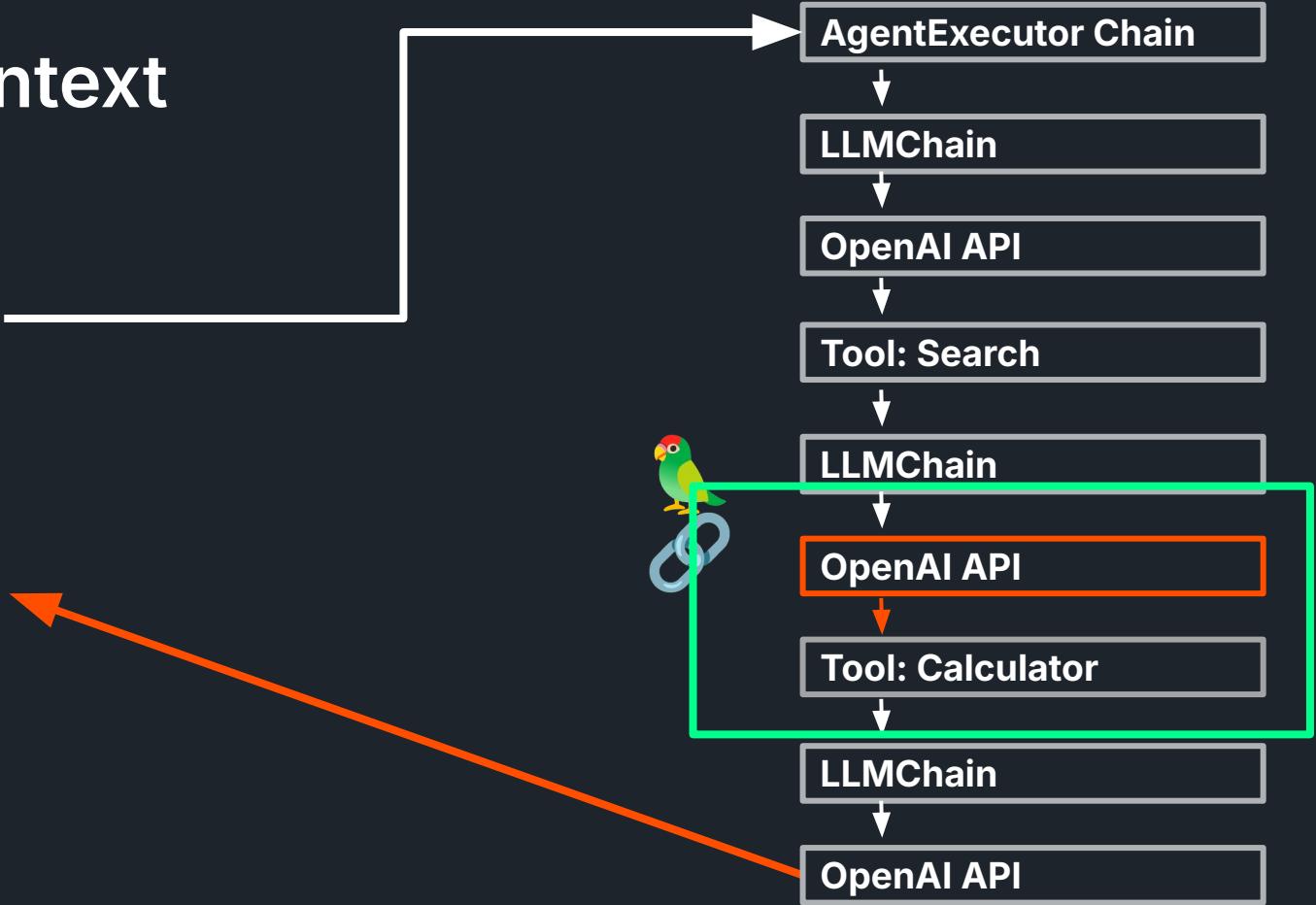
# Manual Instrumentation



```
from opentelemetry import trace  
from opentelemetry.sdk.trace import TracerProvider  
from opentelemetry.sdk.trace.export import BatchSpanProcessor  
from opentelemetry.exporter.otlp.proto.grpc.trace_exporter  
import OTLPSpanExporter  
from opentelemetry.trace.status import Status, StatusCode  
  
trace.set_tracer_provider(TracerProvider())  
tracer = trace.get_tracer("my.tracer.name")  
  
def aFunction():  
    with tracer.start_as_current_span("/aFunction") as span:
```

[opentelemetry.io/docs/instrumentation/python/manual/](https://opentelemetry.io/docs/instrumentation/python/manual/)

# Adding Context

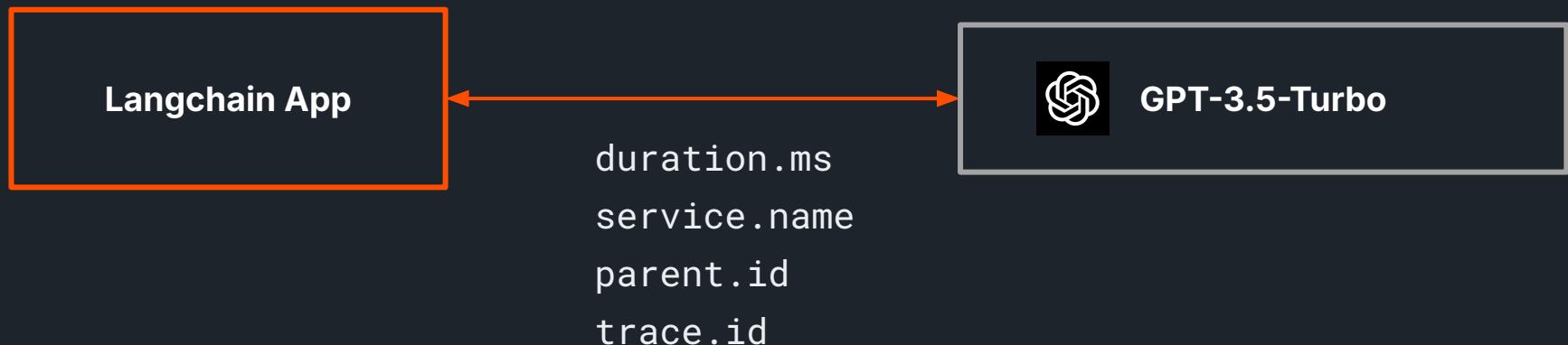


# Adding Context

The prompts used

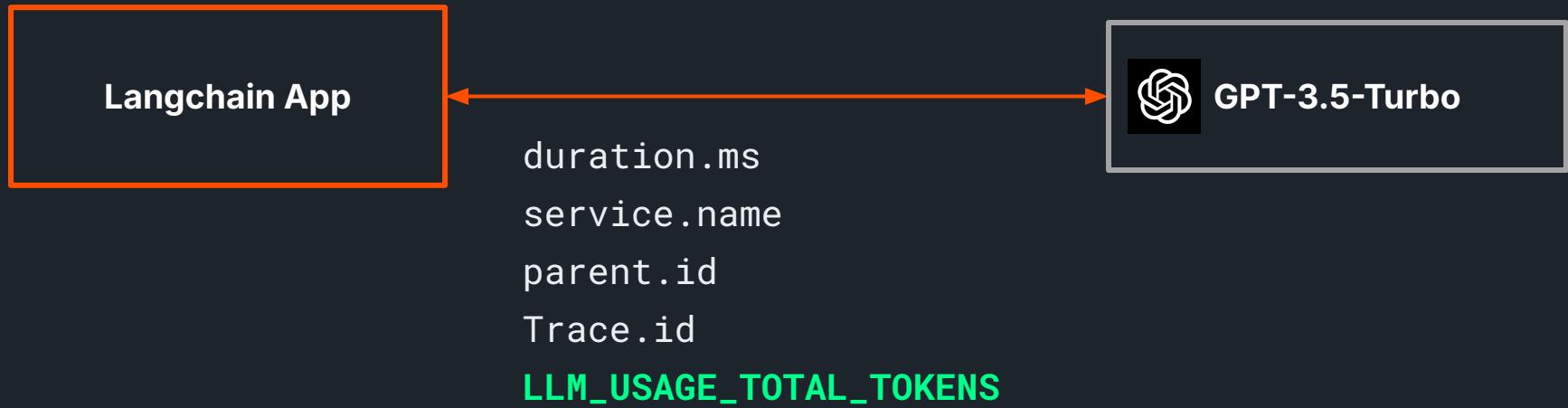
# of tokens used?

LLM temperature



# Adding Context

```
span.set_attribute(LLM_USAGE_TOTAL_TOKENS, usage.get("total_tokens"))
```





# OpenLLMetry

Wrapper around OpenTelemetry to automatically instrument parts of the AI Stack.

GitHub - traceloop/openllmety

github.com/traceloop/openllmety

Product Solutions Open Source Pricing

Search or jump to... Sign in Sign up

traceloop / openllmety Public

Code Issues 4 Pull requests 30 Actions Projects Security Insights

main 58 branches 23 tags Go to file Code About

github-actions[bot] bump: version 0.1.9 → 0.1.10 ✓ 13dd345 yesterday 134 commits

.github fix(cli-cd): release workflow fetches the outdated commit on release ... 4 days ago

img docs: first readme draft 2 months ago

packages bump: version 0.1.9 → 0.1.10 yesterday

scripts fix(cli-cd): finalize release flow (#133) 5 days ago

.cz.toml bump: version 0.1.9 → 0.1.10 yesterday

.gitignore chore: batch upgrade to otel 0.41b0 (#116) last week

CHANGELOG.md bump: version 0.1.9 → 0.1.10 yesterday

CODE\_OF\_CONDUCT.md docs: first readme draft 2 months ago

CONTRIBUTING.md chore(docs): renamed docs Python SDK → OpenLLMetry 2 weeks ago

Readme Apache-2.0 license Code of conduct Security policy Activity

python monitoring metrics ml datascience observability model-monitoring opentelemetry open-telemetry opentelemetry-python llmops

# I have the data... Now what?



# Tracking End User Latency

Plot	Trace name	Start time ↓	Trace duration	Ent...	Spans	Errors
●	langchain_llm.task streamlit_agent/mrkl_demo.py	16 hours ago Nov 1, 2023 04:..	6 s	1	12	0
●	langchain_llm.task streamlit_agent/mrkl_demo.py	16 hours ago Nov 1, 2023 04:..	7 s	1	12	0
●	langchain_llm.task streamlit_agent/mrkl_demo.py	16 hours ago Nov 1, 2023 04:..	7 s	1	12	0
●	langchain_llm.task streamlit_agent/mrkl_demo.py	16 hours ago Nov 1, 2023 04:..	4036 ms	1	6	0
●	langchain_llm.task streamlit_agent/mrkl_demo.py	16 hours ago Nov 1, 2023 03:..	10 s	1	21	0
●	langchain_llm.task streamlit_agent/mrkl_demo.py	16 hours ago Nov 1, 2023 03:..	6 s	1	12	0
●	langchain_llm.task streamlit_agent/mrkl_demo.py	20 hours ago Oct 31, 2023 23:..	6 s	1	10	0
●	langchain_llm.task streamlit_agent/mrkl_demo.py	20 hours ago Oct 31, 2023 23:..	4512 ms	1	8	2

# Tracking Errors

The screenshot shows the New Relic APM interface for monitoring the `langchain.agent` application. The main view displays a timeline of spans across different service components: `streamlit_agent`, `langchain.agent`, `langchain.tool.calculator`, and `calculator_tool.task`. A red bar highlights a specific error event in the `langchain.agent` component, which occurred at 5:52:00pm on November 3 at 5:54pm. The trace ID is 7522a68c71700e0a01c47fbecb7ad3a3, and there are 8 spans in total.

The error details panel, highlighted with an orange border, provides the following information:

- otel.status\_code**: ERROR
- otel.status\_description**: ValueError: unknown format from LLM: 7 \* 6 / 3 `` ...numexpr.evaluate("7 \* 6 / 3")...
- SpanEvent exceptions**: There was 1 span event exception

Average duration: 4000 ms



# OpenLLMetry

TRACE

Collapse Stats Filter Show All ▾

AgentExecutor 9.61s 1,915 zero-shot-react-description

LLMChain 2.00s

ChatOpenAI gpt-3.5-turbo 1.84s

Search 2.46s

LLMChain 1.58s

ChatOpenAI gpt-3.5-turbo 1.52s

Search 2.46s

LLMChain 0.97s

ChatOpenAI gpt-3.5-turbo 0.90s

## AgentExecutor

Run Feedback Metadata

Input

```
1 {  
2   "input": "What was the high temperature in Bangalore, India this month in Celsius?"  
3 }
```

JSON

Rendered Output

The highest temperature in Bangalore, India this month was 37.2 degrees Celsius.

Run ID Trace ID

START TIME  
08/09/2024, 09:48:15 PM

END TIME  
08/09/2024, 09:48:25 PM

TIME TO FIRST TOKEN  
N/A

STATUS  
Success

TOTAL TOKENS  
1,915 tokens / \$0.0010795

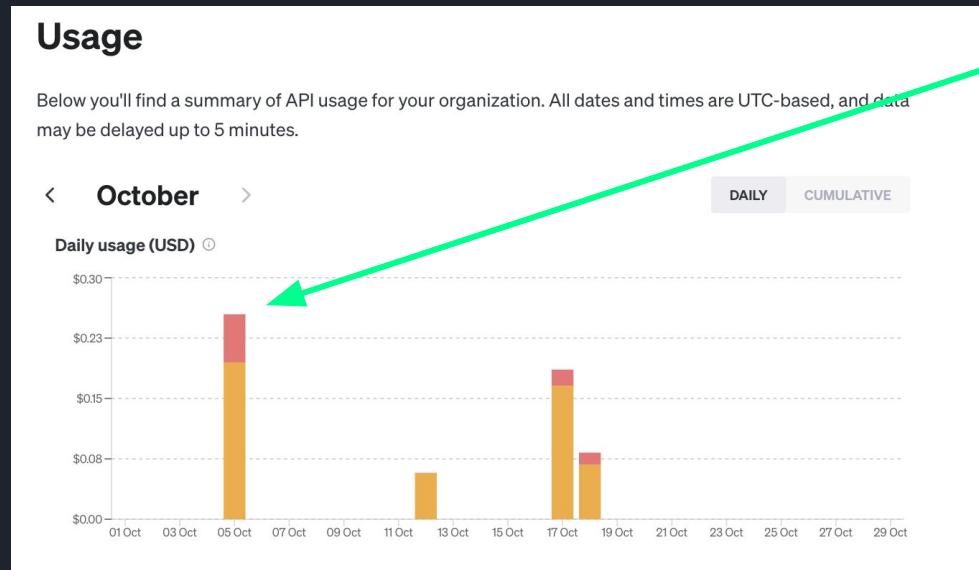
LATENCY  
9.61s

TYPE  
Chain

TAGS  
zero-shot-react-description

# Tracking Cost

# OpenAI Dashboard



What is causing  
the spike?



# Tracking Token Usage

Red is my favorite color.

Tokenizing algorithm

[7738, 318, 616, 4004, 3124, 13]

platform.openai.com/tokenizer

OpenAI Platform Relaunch to update

platform.openai.com/tokenizer

Overview Documentation API reference Examples Account >

token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

It's important to note that the exact tokenization process varies between models. Newer models like GPT-3.5 and GPT-4 use a different tokenizer than our legacy GPT-3 and Codex models, and will produce different tokens for the same input text.

GPT-3.5 & GPT-4 GPT-3 (Legacy)

Many words map to one token, but some don't: `indivisible`.

Unicode characters like emojis may be split into many tokens containing the underlying bytes: `😊`

Sequences of characters commonly found next to each other may be grouped together: `1234567890`

[Clear](#) [Show example](#)

Tokens  
57

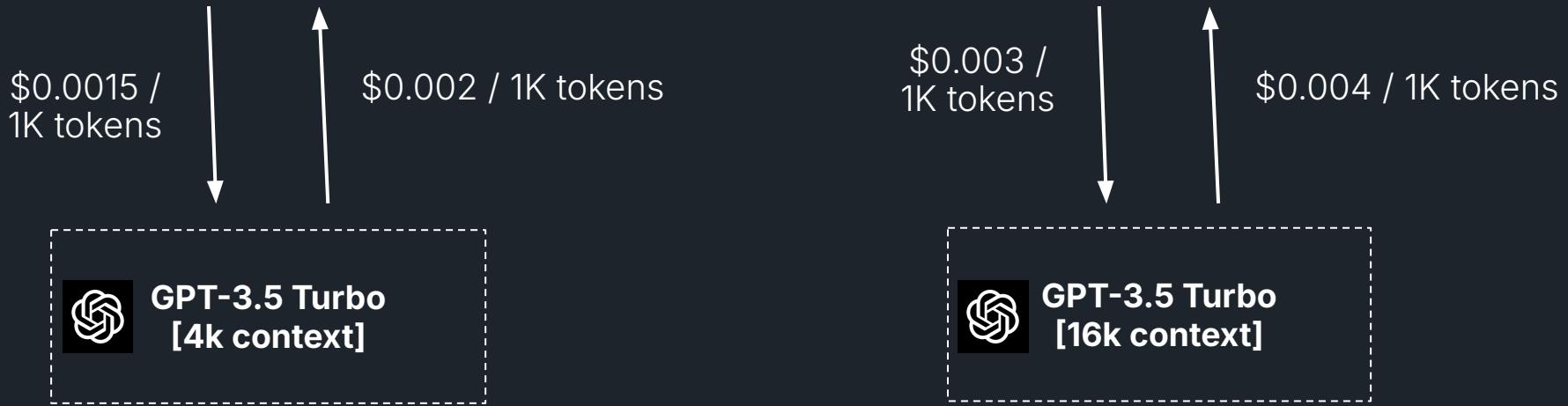
Characters  
252

Many words map to one token, but some don't: `indivisible`.

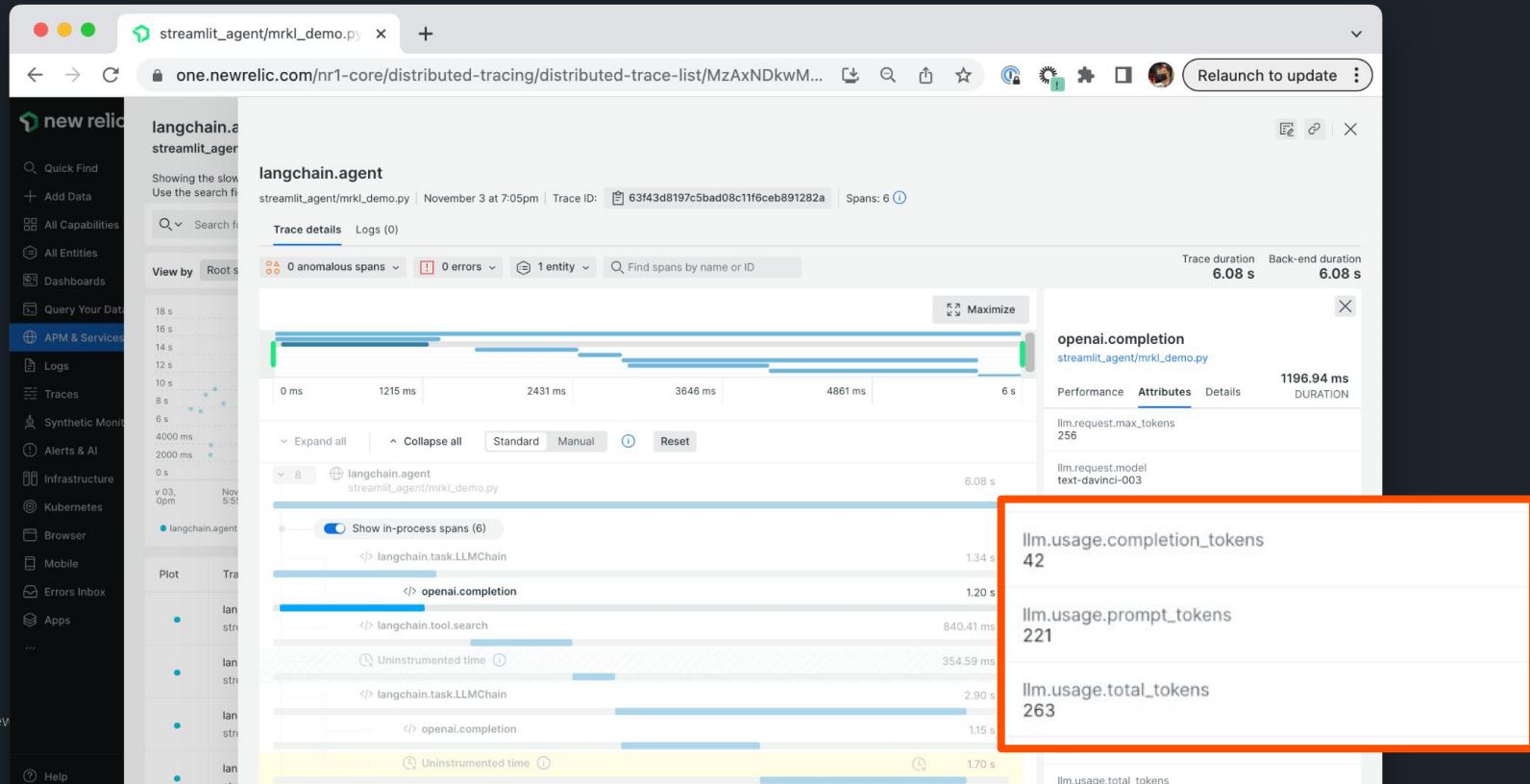
Unicode characters like emojis may be split into many tokens containing the underlying bytes: `😊`

Sequences of characters commonly found next to each other may be grouped

# Tracking Token Usage



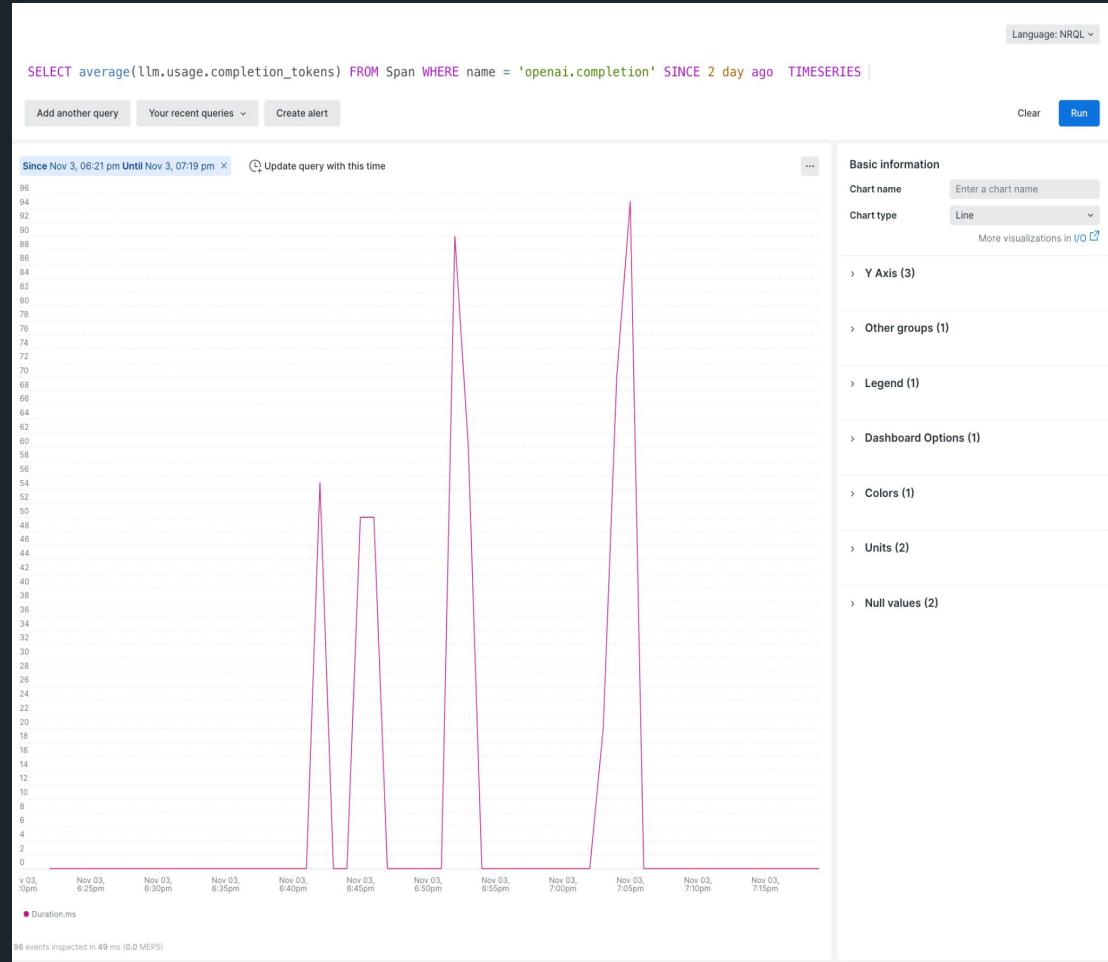
# Tracking Token Usage



# Investigating Token Usage

## Query for clues

- High standard deviation
  - Is there a rise in token usage after a specific code change?
  - What are the inputs/outputs of the largest token usage trace







# Experimentation

**langchain.agent.streamlit\_agent**

**langchain.agent**

Showing the slowest spans. Use the search field to find specific spans.

Trace ID: 63f43d8197c5bad08c11f6ceb891282a | Spans: 6

Trace duration: 6.08 s | Back-end duration: 6.08 s

Logs (0)

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

View by: Root spans | Expand all | Collapse all | Standard | Manual | Reset

Show in-process spans (6)

entity.name: streamlit\_agent/mrkl\_demo.py

entity.type: SERVICE

entity.guid: MzAxNDKwMXxFWFR8U0VSVkIDRxw4NDE3NTY5OTYz0TA3OTIxODg4

entity.name: streamlit\_agent/mrkl\_demo.py

entity.type: SERVICE

entity.guid: MzAxNDKwMXxFWFR8U0VSVkIDRxw4NDE3NTY5OTYz0TA3OTIxODg4

guid: 0593e5b1ca7cf29a

id: 0593e5b1ca7cf29a

instrumentation.provider: opentelemetry

duration.ms: 1149.209781

entity.name: streamlit\_agent/mrkl\_demo.py

entity.type: SERVICE

entity.guid: MzAxNDKwMXxFWFR8U0VSVkIDRxw4NDE3NTY5OTYz0TA3OTIxODg4

guid: 0593e5b1ca7cf29a

id: 0593e5b1ca7cf29a

instrumentation.provider: opentelemetry

llm.completions.0.content  
I now know the answer Final Answer:

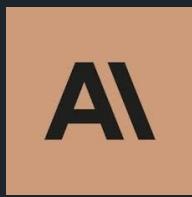
## Average token usage



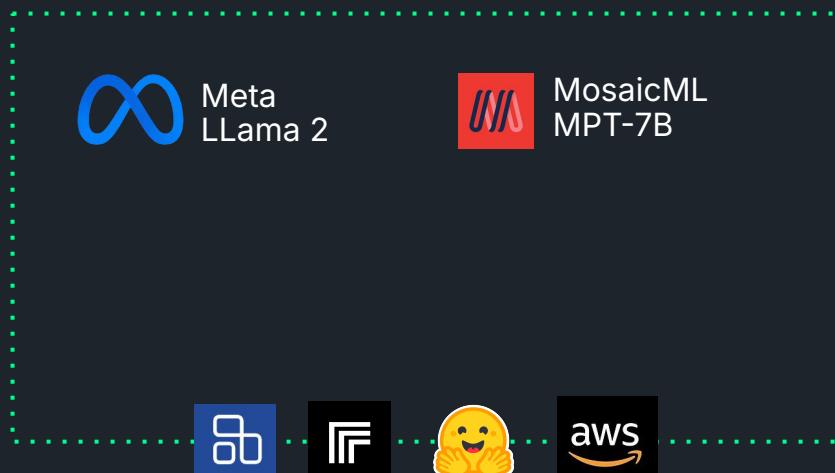
~~~~~ 152 tokens/query

~~~~~ 200 tokens/query

# A/B Testing LLM Performance



Commercial closed-source LLM  
offerings



Hosted open-source LLM  
offerings

# A/B Testing LLM Performance



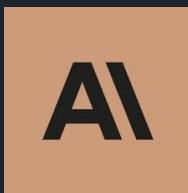
AgentExecutor Chain

LLMChain

33%

33%

33%



```
from langchain.llms import OpenAI  
from langchain.llms import Anthropic  
from langchain.llms import Replicate  
llm = Replicate(model=[insertmodelhere])
```

Token usage (cost)  
% of users satisfied  
Prompt/Response Quality

# \$whoami\_



**Zameer Fouzan**

Lead Developer Relations Engineer -  
APJ

- Full Stack Developer 😎
- Open Telemetry Advocate at NR
- AI RAG Picker

