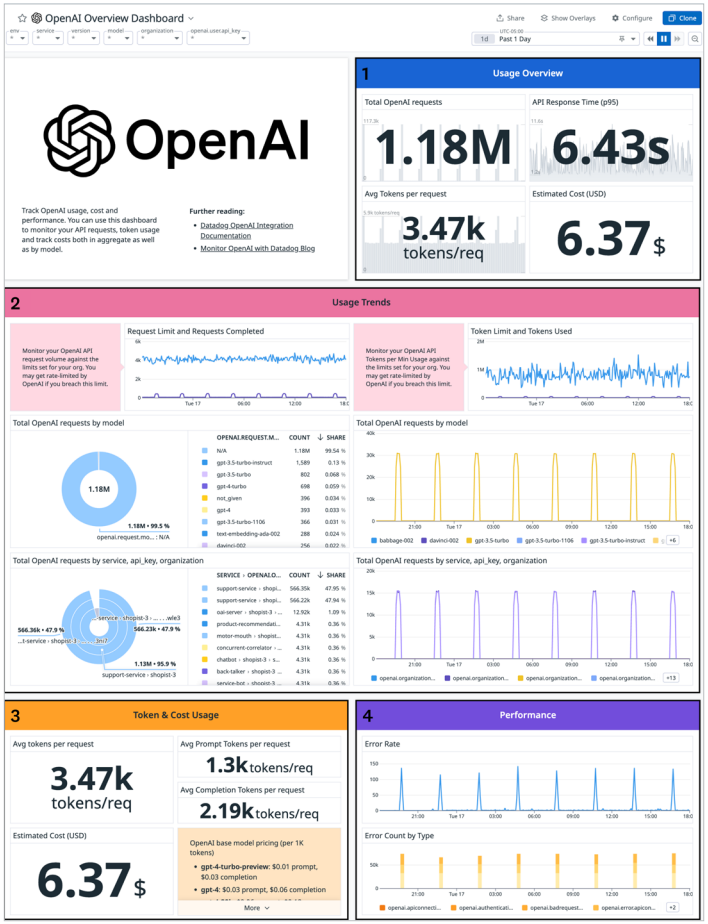
Cheatsheet: OpenAI

This cheatsheet provides comprehensive cost estimation, prompt and completion sampling, error tracking, and performance metrics for OpenAI account-level requests, as well as those made through Python, Node.js, and PHP libraries. It leverages Datadog metrics, Application Performance Monitoring (APM), and logs for detailed monitoring and analysis.



Why Datadog?

Datadog is a leading SaaS-based observability and security platform that brings together telemetry from across your tech environment – including infrastructure metrics, application traces, and logs – together in a single platform. Our monitoring capabilities include customizable alerting and monitoring reports and visualization tools like out-of-the-box dashboards, making it easy and fast to investigate and resolve issues. **With 800+ vendor-backed integrations, including OpenAI, you can gain visibility into any part of your tech stack.**

1. Usage Overview

NAME ON DASHBOARD	DATADOG METRIC NAME	METRIC DESCRIPTION
Total OpenAI requests	openai.request.duration	Request duration distribution. Shown as nanosecond
Avg Tokens per request	openai.tokens.total	Total number of tokens used in a request to OpenAI. Shown as token
	openai.tokens.prompt	Number of tokens used in the prompt of a request to OpenAI. Shown as token
Estimated Cost (USD)	openai.tokens.completion	Number of tokens used in the completion of a response from OpenAI. Shown as token
	openai.tokens.total	Total number of tokens used in a request to OpenAI. Shown as token

2. Usage Trends

NAME ON DASHBOARD	DATADOG METRIC NAME	METRIC DESCRIPTION
Request Limit and Requests Completed - Monitor your OpenAI API request volume against the limits set for your org. You may get rate-limited by OpenAI if you breach this limit.	openai.ratelimit.requests	Number of requests in the rate limit. *Shown as request*
Token Limit and Tokens Used - Monitor your OpenAI API Tokens per Min Usage against the limits set for your org. You may get rate-limited by OpenAI if you breach this limit.	openai.ratelimit.tokens	Number of tokens in the rate limit. *Shown as request*
Total OpenAI requests by model	openai.request.duration	Request duration distribution. *Shown as nanosecond*
Total OpenAI requests by service, api_key, organization		

3. Token & Cost Usage

NAME ON DASHBOARD	DATADOG METRIC NAME	METRIC DESCRIPTION
Avg tokens per request	openai.tokens.total	Total number of tokens used in a request to OpenAI. Shown as token
Avg Prompt Tokens per request	openai.tokens.prompt	Number of tokens used in the prompt of a request to OpenAI.
Avg Completion Tokens per request	openai.tokens.completion	Number of tokens used in the completion of a response from OpenAI.
	openai.tokens.prompt	Number of tokens used in the prompt of a request to OpenAI. Shown as token
Estimated Cost (USD)	openai.tokens.completion	Number of tokens used in the completion of a response from OpenAI. Shown as token
	openai.tokens.total	Total number of tokens used in a request to OpenAI. Shown as token
Prompt Token Usage	openai.tokens.prompt	Number of tokens used in the prompt of a request to OpenAI.
Total Tokens usage by model	openai.tokens.total	Total number of tokens used in a request to OpenAI.
Tokens usage by Prompt & Completion	openai.tokens.prompt	Number of tokens used in the prompt of a request to OpenAI.
	openai.tokens.completion	Number of tokens used in the completion of a response from OpenAI.

4. Performance

NAME ON DASHBOARD	DATADOG METRIC NAME	METRIC DESCRIPTION
Error Rate	openai.request.error	Number of errors. *Shown as error*
	openai.request.duration	Request duration distribution. *Shown as nanosecond*
Error Count by Type	openai.request.error	Number of errors. *Shown as error*
API response time by model (p95)		Request duration distribution. *Shown as nanosecond*
API response time by service (p95)	openai.request.duration	Request duration distribution. *Shown as nanosecond*
API response time by organization (p95)		Request duration distribution. *Shown as nanosecond*
Response Time to Prompt Token Ratio	openai.request.duration	Request duration distribution. *Shown as nanosecond*
	openai.tokens.prompt	Number of tokens used in the prompt of a request to OpenAI. Shown as token