

Cloud + AI bill: 10 checks in 30 minutes

Ten questions to ask about your cloud and AI bill. Each one takes a few minutes and points to a known source of spend.

Open your billing console next to this list. Answer each question with yes, no or don't know. A "don't know" is already a finding.

01

OPENAI · ANTHROPIC

Do AI calls that can wait go through the Batch API?

At OpenAI and at Anthropic, it costs half price.

☐

02

ANTHROPIC

Is the long context your agents resend at every step cached?

At Anthropic, the cached part is read back at 10% of the input price at most.

☐

03

AI MODELS

Do you know which model serves which feature?

A large model on a sorting task costs several times the price of a small one.

☐

04

AWS

Does traffic to S3 go through a NAT Gateway?

It bills every GB; an S3 gateway endpoint is free.

☐

05

AWS

Do your CloudWatch log groups have a retention period?

By default, they are kept with no limit.

☐

06

AWS

How many public IPv4 addresses are you paying for?

AWS has charged for them by the hour since February 2024, even when in use.

07

AWS

Are there any gp2 EBS volumes left?

gp3 costs less per GB.

08

EVERY CLOUD

Do your dev environments run at night and on weekends?

Turning them off outside 8 am–8 pm on weekdays means 108 hours out of 168.

09

DATADOG

On Datadog, does a custom metric carry a tag like a customer ID?

Each combination of tags is billed.

10

GOOGLE CLOUD

On Google Cloud, do your logs go beyond the free 50 GiB per project per month?

Checked against each provider's official pricing pages on October 20, 2026. Prices change: confirm on your own bill.

See these costs in one place

Fivemetrics brings your cloud and AI bills together and shows what changed.

[Create a free account](#)