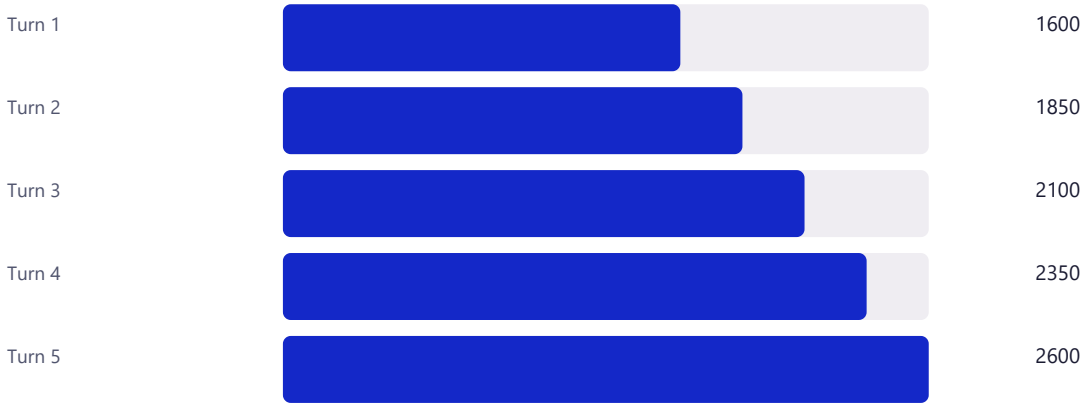
CHAPTER 16 / DEPLOYMENT

Where the bill comes from.

See how history, caching and capacity change an illustrative bill.

See how history, caching and capacity change an illustrative bill.



Handpicked example values · inspect exact values below

Illustrative rates and capacity. The example excludes staffing, retries, reasoning tokens and differences in model quality.

Hosted API

A provider runs inference; investigate contract, region, capacity and data handling.

Reserved capacity

Plan capacity and utilization rather than assuming a fixed allocation is cheaper.

Self-hosting

You operate serving and security; token billing disappears, but compute and staff costs do not.

↗ **A provider runs inference; investigate contract, region, capacity and data handling.**

PUT IT TO WORK

Compare five turns with and without a cache, then see when a hypothetical server at $4 per hour crosses the per-token bill.

CLOSE THE PAGE. EXPLAIN IT.

Why can writing to a cache cost more than leaving it off for one turn?