@demotoprod · BUDGET

Agent budget checklist

A frontier model went on sale on 13 October 2026 at two dollars per million input tokens and ten per million output, for seven weeks. From 1 December the same tokens cost four and twenty. Output is five times input and includes the reasoning nobody reads. A pilot priced on the launch rate meets a December invoice at twice the price. Five checks.

1 Budget at the list price

The discount has an end date; the budget line does not. A cost model built on the launch rate is wrong by a factor of two from the first of December.

- Put the list price in the cost model today, with the discount's end date on the ticket that owns the line.
- Set the spend alert at the list price now, so the alarm is tested while the bill is still small.

```
price_per_million: {input: 4.00, output: 20.00, cached: 0.20}   # list price
alert_usd_per_day: 120   # list-price projection, from today
```

2 An output ceiling per route

Agents are output machines, and output costs five times input. A request cap written when the ceiling was sixty-four thousand tokens does nothing against a million-token task.

- Size max output tokens to the job, per route, and log every request that hits the ceiling.
- A million-token task is a batch job with its own approval, at half the standard rate.
- Count ceiling hits every day; a rising count is a prompt problem before it is a budget problem.

```
routes: {support-answer: {max_output_tokens: 4000}, code-migration:
{max_output_tokens: 64000, tier: batch}}
```

3 Cache hit rate on the dashboard

Cached input is ninety-five percent off: at the list price, twenty cents per million against four dollars. The order of the prompt decides the bill more than the model does.

- Stable content first: system prompt, tool schemas, reference documents; the user's turn last.
- Put the cache hit rate per route next to latency and error rate; a falling rate pages someone before the invoice does.
- Treat a prompt change that breaks the cache as a cost regression in review.

```
cache_hit_rate{route="support-answer"} < 0.8 for 30m   # page: prefix changed
```

4 Thinking tokens are output

The output rate includes the model's reasoning; the published price table says so in as many words. The answer on screen is a fraction of what was billed.

- Log output tokens per task, split into visible answer and reasoning, and watch the ratio per route.
- Set the reasoning effort per route; the cheapest setting that passes the eval is the default.

5 Rehearse 1 December

A price step with a known date is the easiest incident to prepare for. The launch rate runs to 30 November 2026 and the list rate applies from 1 December; both dates were published on launch day.

- Run one day of traffic through the cost model at the list price, compare it with the alert thresholds, and name who approves the line after the step.

Source: a frontier model vendor's published API price list of 13 October 2026 ($2 / $10 per million to 30 November; $4 / $20 from 1 December; cached input 95% off). Illustrative thresholds; set yours from

Nitish Gautam · indica-tech.com · hello@indica-tech.com

INDICA TECH · @DEMOTOPROD

Nitish Gautam

Founder-led AI engineering. Demo to production, under NDA, in regulated industries, on fixed scope.

Where this list comes from

Ten-plus years taking products from zero to one across seven sectors, including defence, fintech, medical and telecom, with AI shipped into regulated and air-gapped environments. The @demotoprod channel takes one production AI failure apart each day. These checklists are the same work written down: every item is a dated failure or a documented default, with the control that stops it and the primary source.

If this list found a gap

1. **Your demo works and production is stuck.**
   Production Readiness Audit. Fixed price, two weeks. A five-dimension scorecard, cost and failure analysis, and a 90-day execution roadmap.

2. **You need senior AI judgement without a full-time hire.**
   Fractional AI Lead. Monthly, 30-day notice either side. Weekly 1:1 with the founder or CTO, architecture and code review, technical hiring support, investor-ready due diligence.

3. **An AI feature has to work, by a date.**
   AI Build Sprint. Fixed price, four to eight weeks. Scope signed off up front, weekly demos, CI/CD and observability included, 30-day production warranty.

4. **A regulator, an auditor or a customer's security team is asking.**
   AI Governance & Compliance. Fixed price, four weeks. ISO 27001 / SOC 2-aligned controls, a model risk framework, human-in-the-loop and audit logging, regulator-ready documentation.

BOOK A CALL
indica-tech.com/#contact

EMAIL
hello@indica-tech.com

WEBSITE
indica-tech.com

ALL CHECKLISTS
indica-tech.com/checklists/

YOUTUBE
youtube.com/@demotoprod

LINKEDIN
linkedin.com/in/startup-founderx

INSTAGRAM
instagram.com/demotoprod

TIKTOK
tiktok.com/@demotoprod

Free to share. The live list, with every source link: indica-tech.com/checklists/#agent-budget