@demotoprod · BUDGET · also LEVERS

Token optimization checklist

On one provider's caching page, cached reads cost 0.1 times the base input price, with per-model exceptions; another provider's guide says reused tokens are discounted up to 95%. Both need the whole prefix to match. Five checks, each sized for an afternoon, from prompt order to routing.

1 Put the fixed part first, the question last

One guide says cache reuse requires the entire rendered prefix to match. Another says a change at one level (tools, then system, then messages) invalidates that level and every later one. A timestamp near the top leaves nothing below it to reuse.

- Order the prompt: tools, system prompt, shared documents, history, then the new question.
- Move timestamps and per-request IDs to the end; run the snippet on two consecutive prompts.

```
import os   # prompt_1, prompt_2: two consecutive rendered prompts
print(len(os.path.commonprefix([prompt_1, prompt_2])), "shared characters")
```

2 Read the cache counters on every response

On one page, a prompt below the model's minimum cacheable length, 512 to 4,096 tokens by model, is processed without caching and no error is returned. Zero cache reads and writes in the usage fields show it was not cached.

- Log cached, written and uncached input tokens per call; alert when the cached share drops.
- Count your fixed prefix in tokens and compare it with the model's minimum.

```
u = response.usage   # the other provider: usage.input_tokens_details.cached_tokens
r, w, n = u.cache_read_input_tokens, u.cache_creation_input_tokens, u.input_tokens
print("cached share:", r / (r + w + n))
```

3 Check that reads repay each cache write

Both pages price the default cache write at 1.25 times base input (the second for its newest models), a standard read at 0.1 times, and the first's 1-hour write at 2 times. One read repays a default write; two repay a 1-hour write.

- List each prefix's call gaps; a gap past the lifetime (5 minutes, first page) means a new write.
- Keep caching only where reads per write reach 1 (default) or 2 (1-hour), counted as in check 2.

4 Trim old turns in batches, not every turn

One guide says that even with chained responses all previous input tokens are billed as input tokens. Another says clearing old tool results invalidates the cached prefix, so each clear should free enough tokens to be worth it.

- Clear or summarise old turns only past a token threshold, and set a minimum freed per clear.
- Compare cached share and cost per conversation before and after, from check 2's counters.

5 Send easy requests to a smaller model after testing it

Two papers test sending each query to a cheaper model when it can answer. One reports matching the best single model with up to 98% cost reduction; the other, costs cut by over 2 times in certain cases without lower response quality.

- Hand-label a sample of real requests as easy or hard and run both models on the whole sample.
- Route to the smaller model only where it passes your checks; send the rest to the larger one.

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.

Paste into your project

ONE TOKEN POLICY COVERING ALL FIVE CHECKS

```
# token-policy.yaml: names in capitals are yours to set
prompt_order: [tools, system, documents, history, question]
move_to_end: [timestamp, request_id]
cache:
  min_prefix_tokens: MODEL_MINIMUM   # 512 to 4,096 by model, first page
  min_reads_per_write: {default: 1, one_hour: 2}
  alert_below_cached_share: CACHED_SHARE_FLOOR  # set from your own traffic
trim:
  start_above_tokens: TRIM_THRESHOLD
  min_freed_tokens: MIN_FREED_PER_CLEAR
routing:
  smaller_model: {only_for: easy, after: side_by_side_test}
```

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: 90-day minimum, then 30 days' 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/#token-optimization