TokenLean

Same answers. Smaller bill.

by CBeyond
SEE BEYOND THE HUMAN EYE

CBEYOND TECHNOLOGIES

YouTube

TOKENLEAN · THE LLM TOKEN OPTIMISATION LAYER

Spend compounds

Agents re-send context on every step, so token spend grows non-linearly.

Gateways don't shrink

They route traffic and manage keys. Cutting each request isn't their job.

Savings need proof

A CFO trusts the provider's invoice, not a vendor's dashboard.

Cut your LLM bill with a one-line change

A drop-in proxy between your app and any LLM. It shrinks every request and proves the saving on the provider's own bill.

OSS: https://github.com/sumitdevgupto/TokenLean Commercial: https://www.cbeyond.cloud/tokenlean

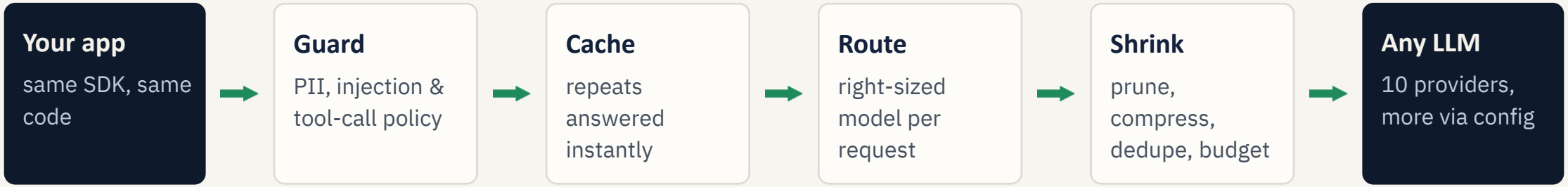
HOW IT WORKS

Two values change. Nothing else.

```
# Before
client = OpenAI(api_key="sk-...")

# After — two values change
client = OpenAI(
    api_key=PROXY_API_KEY,
    base_url=PROXY_ENDPOINT + "/v1",
)
```

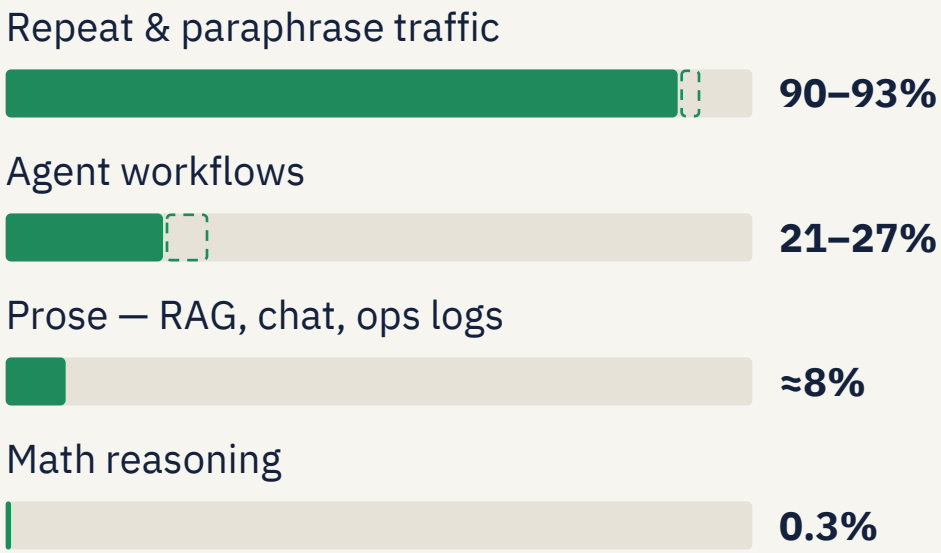
- **Native OpenAI, Anthropic** (incl. Claude Code) and **Gemini** APIs; tool calls round-trip intact
- **10 first-class providers**; any OpenAI-compatible API by config alone
- **Config hot-reloads** in ~60 seconds, no redeploy
- **Configurable optimization pipeline** via ON/OFF toggle-switch with custom providers/models for each step (OSS:via JSON edit, Commercial: via UI)
- **Multitenant** by default



PROOF

Measured on the provider's own bill.

Token savings by workload — provider-billed A/B



gpt-4o-mini, temperature 0, 2026-09-10. Every item sent twice, direct and through TokenLean, on public datasets (HotpotQA, MT-Bench, SWE-bench Lite, HumanEval, GSM8K, BFCL). Solid bar = low end of the observed range; outline = high end.

What does the saving

- **Semantic cache**: repeats and paraphrases answered with no model call
- **Structured pruning**: trims bulky tool and JSON payloads
- **Tool-catalogue pruning**: an agent sees only the tools it needs
- **Model routing**: easy requests go to cheaper models (a cost lever)

Quality is checked, not assumed: both arms are graded on the same facts, and any dropped fact is reported next to the saving.

PRODUCT

Live in minutes. Every saving visible.

1 Get a proxy key

Provider keys stay server-side, or bring your own.

2 Point base_url at us

Same SDK, same code: OpenAI, Anthropic or Gemini.

3 Run verify.sh

A true A/B on your own proxy, provider-billed, $1 cap.

Enterprise customer portal

Usage	Observability	Trust & Safety
Routing	Tool Policy	Agents
Models & Keys	Invoices	Docs assistant

Built for enterprise buyers

- **Savings disclosed** on every response, even cache hits
- **Per-tenant controls** for every technique; 10 dashboards
- **Trust & safety that can't be bypassed**: PII redaction, injection guardrails, RAG re-scan, tool-call policy

BUSINESS MODEL

Give away the magic. Sell the operations.

OSS (FREE) · APACHE-2.0

Self-host, $0

- Every optimisation, all trust & safety
- Docker or your own cloud
- The adoption engine
- Config via JSON Edits
- Multitenant

ENTERPRISE · MANAGED SAAS

Operated for you, SLA-backed

- OSS

+

- Consoles: tenants, routing, tool policy, agents
- Learning loop, invoicing, webhooks
- GDPR, always-updated jailbreak defences, BYOK/KMS
- UI for all config tasks, Dashboards and more

Billed per request, never per token

We earn nothing by inflating your tokens, so incentives stay aligned.

Open source is distribution

Developers adopt free; teams pay for operations and compliance.

Provider-neutral

10 providers, works with the OpenAI, Anthropic and Gemini SDKs. No lock-in for the customer.

TokenLean

Same result. Fewer tokens.

BECOME A FOUNDING PARTNER

tokenlean.cbeyond.cloud

① Visit the site → ② Contact → ③ Apply for early access

First cohort deliberately small · founding-partner pricing

Open-source code: github.com/sumitdevgupto/TokenLean

by CBeyond
SEE BEYOND THE HUMAN EYE

https://youtu.be/22on_HzyuuQ