pronix.ai

BENCHMARK REPORT · 2026

Enterprise LLM Cost & TCO Benchmarks *2026.*

Per-request, per-user and per-workflow LLM cost bands from 200+ enterprise programs. Cascade-routing savings, cache-hit economics, GPU vs API tradeoffs and a working 3-year TCO model you can adapt to your own portfolio.

From strategy and pilots to production outcomes.

CFO | CIO | Head of Platform / Cloud | VP FinOps

200+
Programs benchmarked

42%
Median savings from cascade routing

3 yr
TCO model included

Published Q1 2026 · 9 min read

pronix.ai/resources/enterprise-llm-cost-tco-benchmarks-2026

LLM economics changed twice in 2026. Prices fell, and the cost profile shifted from inference to everything around it. This benchmark unpacks the real cost of enterprise LLM workloads across foundation model, retrieval, evaluation, human review and run — with the routing patterns that reliably take 20% to 40% out of the bill.

200+
Programs benchmarked

42%
Median savings from cascade routing

3 yr
TCO model included

What this report answers

- Per-request cost bands by use-case pattern — chat, RAG, agent loop, code, voice
- Cascade routing savings — GPT-class → mid-tier → small OSS — with hit-rate data
- Cache and retrieval economics — where hits actually land
- GPU-hosted vs API — break-even bands by concurrency
- 3-year TCO model — platform, model, run, guardrails, observability, talent

01 The real cost anatomy

Inference is now the smallest surprise in most enterprise AI invoices. Retrieval, evaluation, telephony (in voice workloads) and human review each frequently exceed foundation-model spend.

Programs that report only token cost are budgeting against a fraction of their real TCO.

- Foundation model tokens: 25–35% of workload TCO in mature programs
- Retrieval infrastructure and refresh: 15–25%
- Evaluation, observability and human review: 15–20%
- Telephony (voice workloads only): often 25–40%
- Platform run and FinOps: 8–12%

02 Model routing is the highest-ROI FinOps move

Cheap-first cascade routing — attempt the cheapest capable model, escalate on a confidence check — reliably moves 60% to 75% of traffic to the cheap tier with no measurable quality regression.

The gain is durable because provider prices move quarterly. A router that can swap providers captures price arbitrage without a code change.

03 License rightsizing across CCaaS, CRM and AI providers

Named-versus-concurrent is the most consistent savings lever in CCaaS. Enterprises with strong shift patterns typically recover 15% to 25% by moving to concurrent licensing.

Align renewal cycles across your CCaaS, CRM and AI providers so you can negotiate as a portfolio, not a series.

04 Telephony is usually the biggest under-optimized line item

In modernized voice bot estates, telephony minutes routinely exceed LLM inference by four to seven times. Regional dial plans and carrier arbitrage move the number materially with no CX impact.

05 The board dashboard that ends the arguments

Six metrics is enough: cost per resolution, containment rate, AI gross margin, model mix, seat utilization and forecast variance. More than six becomes noise.

The winning move is agreeing on the six with FinOps and product before anyone builds a dashboard.

THE BOTTOM LINE

AI unit economics are becoming the new cloud unit economics. The programs that build the discipline early will fund every subsequent workload out of savings.

Questions we get asked

Is the TCO model provided as a spreadsheet?
Yes — the appendix links to an editable model you can adapt to your own platform, model mix and run costs.

How much can cascade routing actually save?
Median savings are 42% at equivalent quality when routing GPT-class → mid-tier → small OSS with a calibrated confidence threshold. Top-quartile programs reach 61% savings.

When does GPU hosting beat API pricing?
For sustained concurrency above ~35 requests/sec on 7–13B class models, self-hosted GPU typically breaks even within 4–7 months. Below that, API pricing wins on TCO.

Benchmark your own program. Model the business case with our ROI calculators, or book a working session with the practitioners who ship these programs.

pronix.ai/book
pronix.ai/roi