Local LLMs aren't cheaper. They're a different cost curve.

The economics only work past a specific volume threshold. Get the conditions wrong and it's significant capital misallocation.

SWIPE →

Four signals that make the local case credible

01 Sustained high token volume, verified not estimated

02 Regulatory or data residency mandates

03 Open-weight models clear your accuracy bar

04 In-house MLOps capability to run it

$120–160K

for a production-grade inference cluster

Two nodes, H100-class GPUs, redundant configuration — before you've served a single production request.

Local LLM vs. API: When Does the Economics Justify the Switch?, 2026

$40K/yr

operational overhead, on top of hardware capex

Running, monitoring, and maintaining inference infrastructure doesn't stop after procurement.

Local LLM vs. API: When Does the Economics Justify the Switch?, 2026

70B

parameter open-weight models — Llama 3.3, Qwen 2.5, Mistral Large — approach frontier performance, but not on complex reasoning, structured output under constraint, or multi-step tool use.

Local LLM vs. API: When Does the Economics Justify the Switch?, 2026

"

The local LLM is not cheaper in the short term. It carries a lower marginal cost at scale, with higher fixed cost and operational complexity.

Terence Kok
Enterprise AI Strategist

When each one actually wins

LOCAL WINS	API WINS
High, sustained token volume	Variable or lower volume
Data residency mandates	Frontier reasoning required
In-house MLOps maturity	No dedicated MLOps team

Tiered, not binary.

Route by complexity, sensitivity, and volume. A lightweight classifier or rule-based dispatcher sends each task to the right tier — established practice in production AI platforms, with minimal latency overhead when implemented correctly.

Benchmarked, or estimated?

Confirm token volume against actual workload data, not headcount estimates. Benchmark open-weight models against real operational tasks, not general leaderboards.

Full decision checklist and sources:
terencekok.com/blog/local-llm-vs-api-when-does-economics-actually-justify-switch

It's an infrastructure decision. Not an AI strategy.

Justified at scale, under regulatory constraint, or where agentic economics make API dependency unviable. Below that threshold, managed APIs win on capability per dollar.

Terence Kok · Enterprise AI Strategist

SAVE · COMMENT · FOLLOW