BRIEFING NOTE · ENTERPRISE AI ECONOMICS

The Cost Economics of Generative AI in the Enterprise

Five structural trends reshaping how enterprises budget, build, and buy AI capabilities — with data visualizations for each.

By Angel Armendariz · March 2026

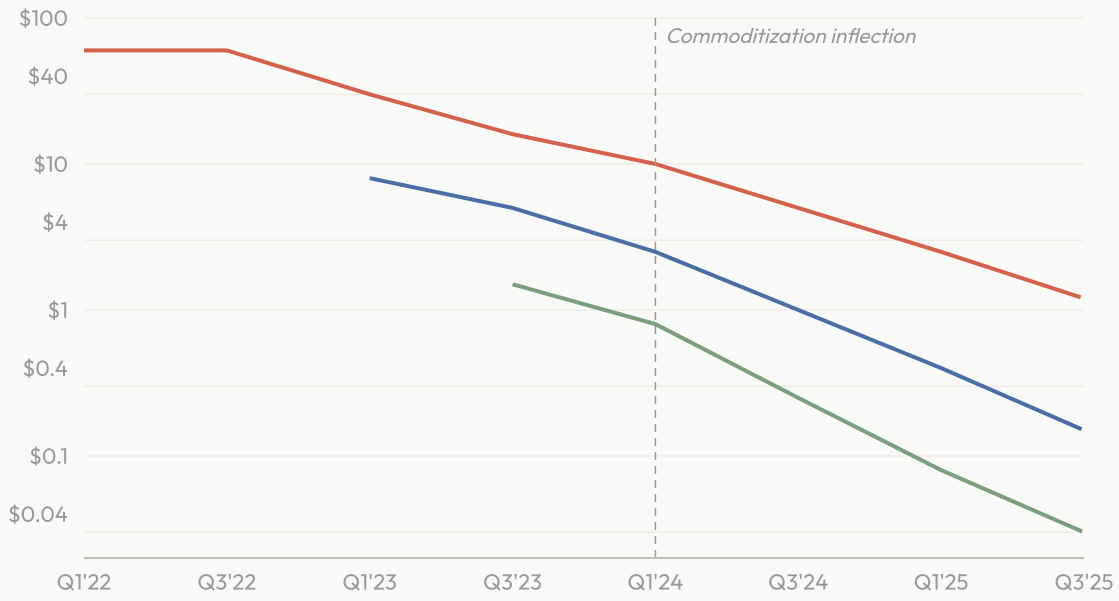
TREND 01 · Inference Deflation

> Token costs fell 50× in three years — not gradually, but in lurches, each triggered by a model release that made the previous pricing regime look absurd.

Enterprises that locked infrastructure contracts in 2022 are paying ten times the current market rate for equivalent capability. The implication is structural: inference cost is no longer a variable to optimize around. It is a floor, and the floor is dropping. The correct response is not to find cheaper compute — it is to rebuild the financial model from the current floor upward, and to treat any contract signed before 2024 as a liability under review.

Inference Cost Has Collapsed 50× in Three Years

Cost per million tokens (USD, log scale) — frontier, mid-tier, and small models, 2022–2025



Source: OpenAI, Anthropic, Google, Mistral published pricing; analyst estimates. Frontier model costs fell from ~$60/M tokens in 2022 to under $2 today — a pace that structurally alters enterprise unit economics and makes 2022-era infrastructure contracts a liability.

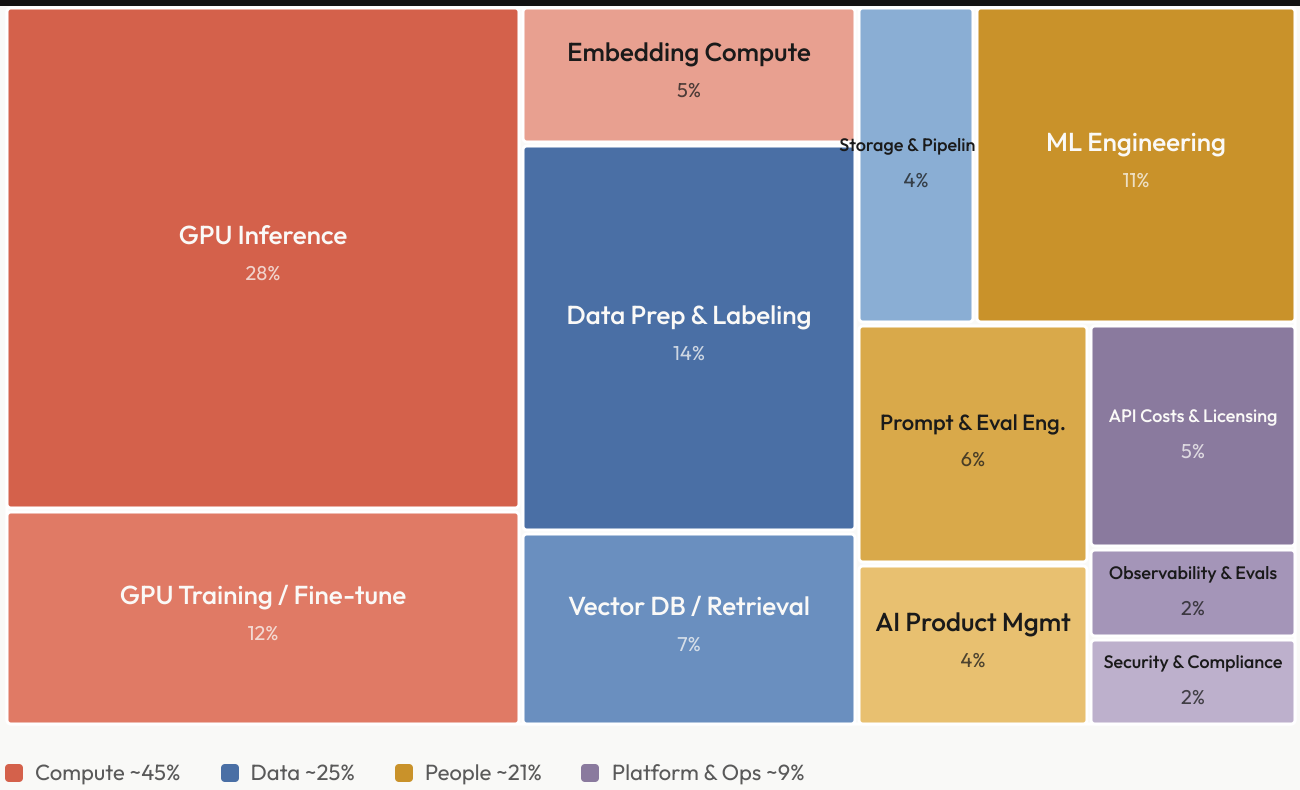
TREND 02 · TCO Composition Shift

As inference deflates, a different cost center rises in relative weight: the people who decide what to build, how to prompt it, and whether it is working.

Engineering judgment — knowing which model to route to, how to structure evaluation, when fine-tuning outperforms prompt engineering — now commands a growing share of first-year TCO. The enterprise that treats GenAI as an infrastructure problem will underspend on the people costs that determine whether the infrastructure produces anything. Compute is becoming a commodity. Judgment is not.

Compute Dominates Year-One TCO — But People Costs Are the Growing Constraint

Total cost of ownership composition for a mid-scale enterprise GenAI deployment, year one



Source: AWS, Azure, GCP published rates; industry practitioner surveys 2024–2025. Compute (~45%) leads first-year TCO, but as inference deflates, engineering judgment — which model to call, how to evaluate it, when to fine-tune — grows as a cost share and becomes the primary optimization lever.

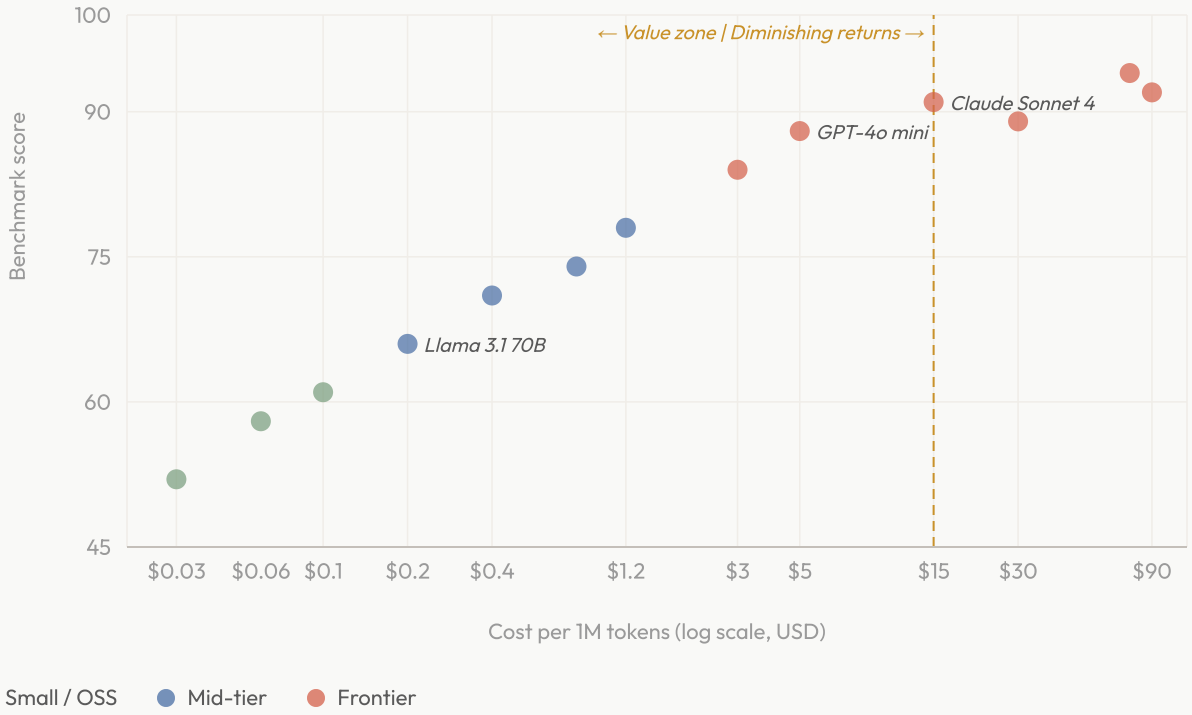
TREND 03 · Capability-Cost Frontier

At $15 per million tokens, the performance curve bends. Capabilities stop compounding. Every dollar above that threshold produces less in return.

Most enterprise GenAI workloads — document processing, data extraction, internal search, code generation — sit firmly in the zone where mid-tier models are indistinguishable from frontier ones in practice. Overspending on frontier capability is the most common and least examined procurement error in enterprise AI today. The discipline required is not technical — it is the willingness to route tasks to the cheapest model that clears the accuracy bar, rather than defaulting to the one the team knows.

The Capability-Cost Frontier Bends at $15 — Most Enterprises Are Past It

Composite benchmark score vs. cost per million tokens (log scale) — 13 production models, 2025



Source: MMLU, HumanEval, MT-Bench composite; provider published pricing. Performance gains above $15/M tokens are marginal for most enterprise workloads. The mid-tier cluster captures ~85% of frontier capability at 5–20× lower cost — the economic case for task-matched model routing is direct and quantifiable.

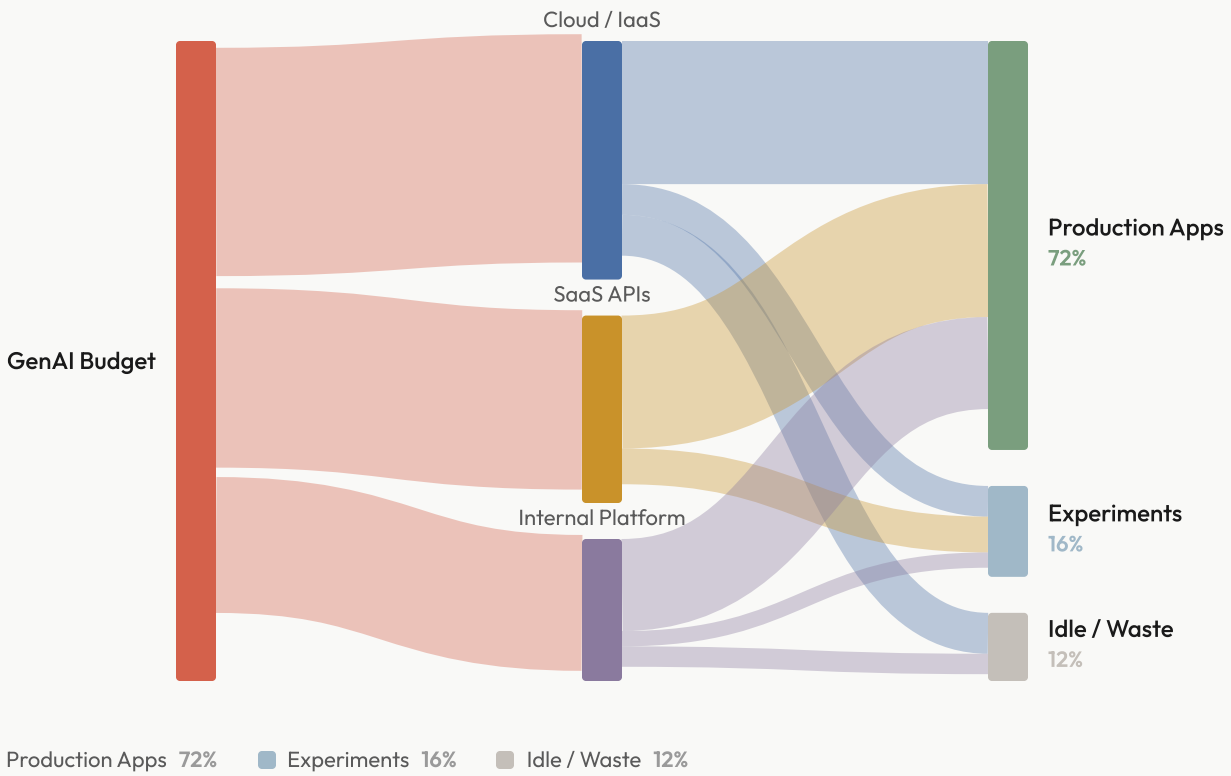
TREND 04 · Enterprise Spend Routing

Between twelve and eighteen cents of every enterprise GenAI dollar ends as idle compute or abandoned experiment.

This is not waste from ignorance — it is waste from the absence of routing discipline. Workloads default to the models the team knows, not the models the task warrants. Experimentation accumulates without a threshold for when experiments become production commitments. The recovery mechanism is unglamorous: model routing policies, spend governance, and the discipline to decommission what is not working. Cloud/IaaS carries the highest waste rate, largely from idle GPU reservations held for capacity that never materialized at scale.

Where GenAI Budget Routes — and Where It Disappears

Normalized spend flow for a representative enterprise GenAI program — allocation → channel → outcome



Source: Gartner, Forrester enterprise AI spend surveys 2024–2025; practitioner interviews. Cloud/IaaS carries 42% of budget but delivers the highest waste rate due to idle GPU reservations. An estimated 12% of total enterprise GenAI budget ends as unrecoverable spend — the primary recovery mechanism is routing discipline, not cost cuts.

TREND 05 · Use Case Viability Matrix

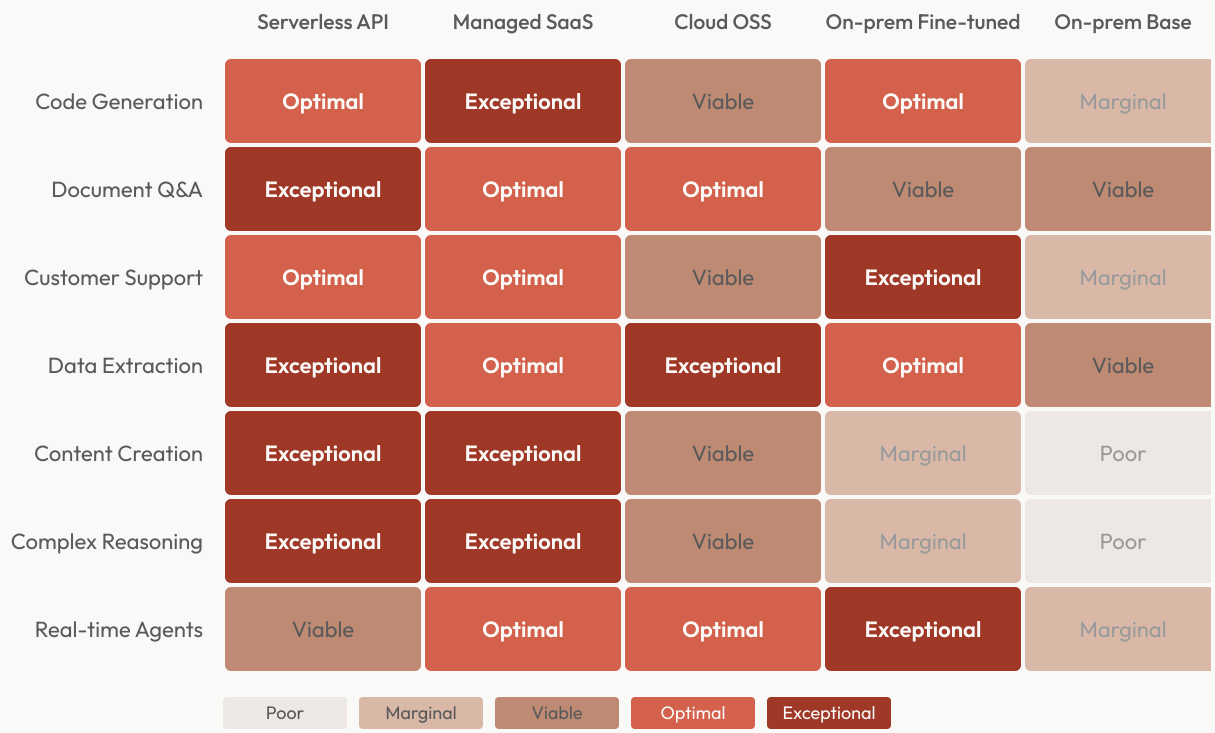
There is no universally optimal deployment architecture. The cost-optimization decision begins with the use case — not the model, not the vendor.

Fine-tuning a 70B parameter model on-premises for customer support achieves better economic viability than routing those queries through a frontier API — but the same on-premises deployment is economically indefensible for content creation or complex reasoning, where frontier models have no viable substitute today. The enterprise that selects infrastructure before use case is working the problem backwards.

Sovereignty requirements, latency thresholds, and accuracy floors each constrain the deployment decision in different directions, and no single architecture satisfies all three simultaneously.

Use Case Viability Varies by Deployment — No Architecture Wins Everything

Economic viability score (1–5) by use case × deployment model — cost, latency, sovereignty, and capability weighted



Source: Caerus Alpha practitioner analysis; AWS, Azure cost benchmarks 2025. Content creation and complex reasoning have no viable on-premises economics today — frontier API is the only defensible path. Customer support and real-time agents favor on-prem fine-tuned despite 3–5× higher infrastructure cost, where data sovereignty governs the decision.