

Model Routing and the AI Bill Nobody Was Watching

Quality Held. 92% Less Spend.

From an uncontrolled frontier default to measured routing in eight weeks

Dr. Tim Nedyalkov | Chief AI Officer & Cybersecurity Executive
tim@cyberkeynote.com | July 2026



The AI worked. The economics did not.

1,000+

employees across the GCC, Europe and Africa

500%

month-over-month inference growth

\$250K

monthly spend before routing

Every task followed the same premium path

Every request

Two-line emails and hundred-page contracts

One premium model

Trusted by default, regardless of task

One invisible bill

No task-level owner or unit economics

Finance was no longer asking where AI could create value. It was asking whether the program should pause.



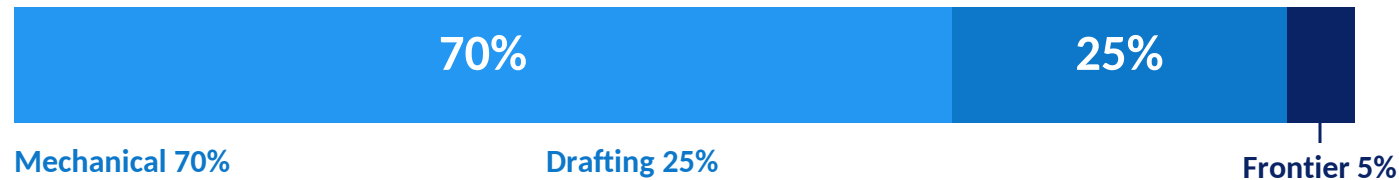
Buy only the intelligence each task requires.

The routing decision is a control sequence, not a model preference.

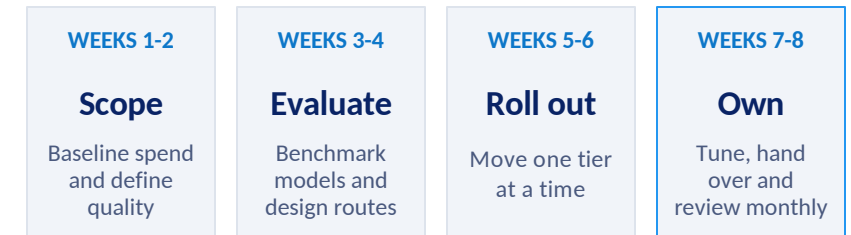
We deployed NVIDIA Nemotron 3 Ultra and Mistral open-weight models through no-train enterprise hosts.^[4]



ROUTED WORKLOAD MIX



THE EIGHT-WEEK SEQUENCE



Quality is structural: failed validation retries one tier higher.

[4] The Open Weight Models that Matter - OpenRouter, 2026.

The quality line held while the cost curve broke.

92%

LOWER MONTHLY SPEND

\$250,000 → under \$20,000

Measured at equal task volume. Cost per unit of work fell ~92%; absolute spend fell despite 15 new use cases onboarded in the same period.

30-40%

faster routine-task P95

15

new use cases

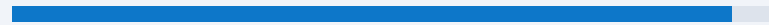
THE QUALITY-HELD PROOF

QUALITY HARNESS

Evaluation harness: 245 real production tasks per tier, blind side-by-side review against the frontier output, pass threshold = acceptance within 2 percentage points of the frontier acceptance rate. Re-run monthly.

COST / DOCUMENT

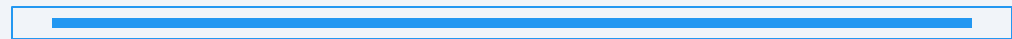
~\$0.30



~\$0.02

QUALITY

WITHIN 2% OF FRONTIER BASELINE



Cost dropped while the measured acceptance band stayed flat.

"The budget was never the problem. The absence of task-level routing was. The same budget now funds three times the portfolio."



Model prices fell 280-fold. Enterprise behavior barely moved.

THE MARKET REPRICED INTELLIGENCE

\$20

per 1M tokens
Nov 2022



\$0.07

per 1M tokens
Oct 2024

280×

price decline.^[2]

ENTERPRISE DEMAND MOVED SLOWER

~30%

of real production workloads
use open models.^[1]

100 trillion

production tokens studied

3-6 months

open-weight capability gap.^[4]

18+ months

gap has not widened.^[4]

DeepSeek V4

80.6%

on SWE-bench Verified.^[4]

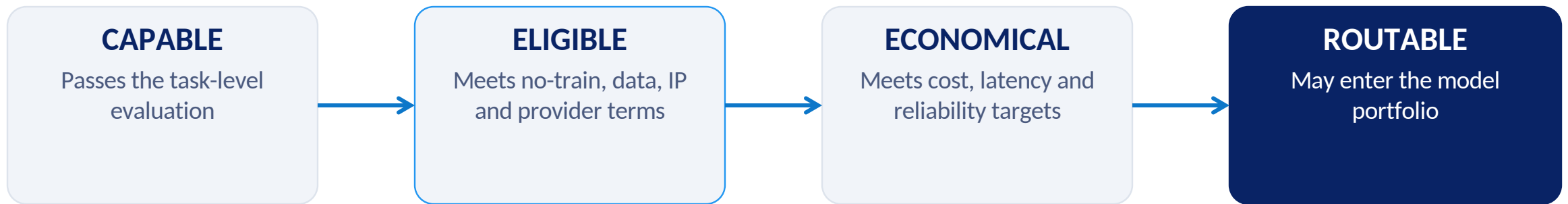
ACE (Agentic Context Engineering) showed a smaller open-source model matching the top-ranked production agent while sharply cutting adaptation latency and rollout cost.^[3]

[1] State of AI - OpenRouter & a16z, 2025. [2] AI Index Report - Stanford HAI, 2025.

[3] Agentic Context Engineering - Stanford / SambaNova, 2025. [4] The Open Weight Models that Matter - OpenRouter, 2026.

A capable model is not automatically routable.

THE ROUTABILITY TEST



CAPABLE → ELIGIBLE → ECONOMICAL → ROUTABLE

The cheapest capable model can still be unusable for a regulated firm.

Inference FinOps must become a monthly operating rhythm.

Prices move monthly. The routing table, evaluation evidence and ownership model must move with them.



The durable advantage is not access to one model. It is the discipline to change routes safely.

Not everything routed: legal reasoning, client-facing strategy work, and ambiguous multi-document analysis stayed on the frontier tier by design. Routing discipline includes knowing what never leaves it.

Routing was one decision. The mandate covers them all.

A fractional Chief AI Officer, embedded in your leadership team: governance, decision rights, spend, evidence, and the board report, owned end to end.

| In your leadership team in one week or less

| First results by day 30, first board report delivered

| Four artifacts: decision rights, controls, evidence, board reporting

| Outcomes like this case: quality held, 92% less spend

CHIEF AI OFFICER AS A SERVICE



Dr. Tim Nedyalkov

Chief AI Officer & Cybersecurity Executive

Book a free 30-minute consultation

✉ tim@cyberkeynote.com

🌐 cyberkeynote.com/caio

Dubai based • travels for selected engagements