

AI Consumption Pricing Broke The Capacity Model

Runtime behavior generates the bill now. Nobody built a governance process to review it.

Boards Governed Infrastructure With One Question

CAPACITY QUESTION

**“How much capacity are we buying?”
buying?”**

Stable. Auditable.

A board could approve it and hold someone accountable to it
— for twenty years.

AI CONSUMPTION PRICING

No capacity commitment to approve.

The bill is generated after the fact — by what a system did,
system did, not what was provisioned in advance.

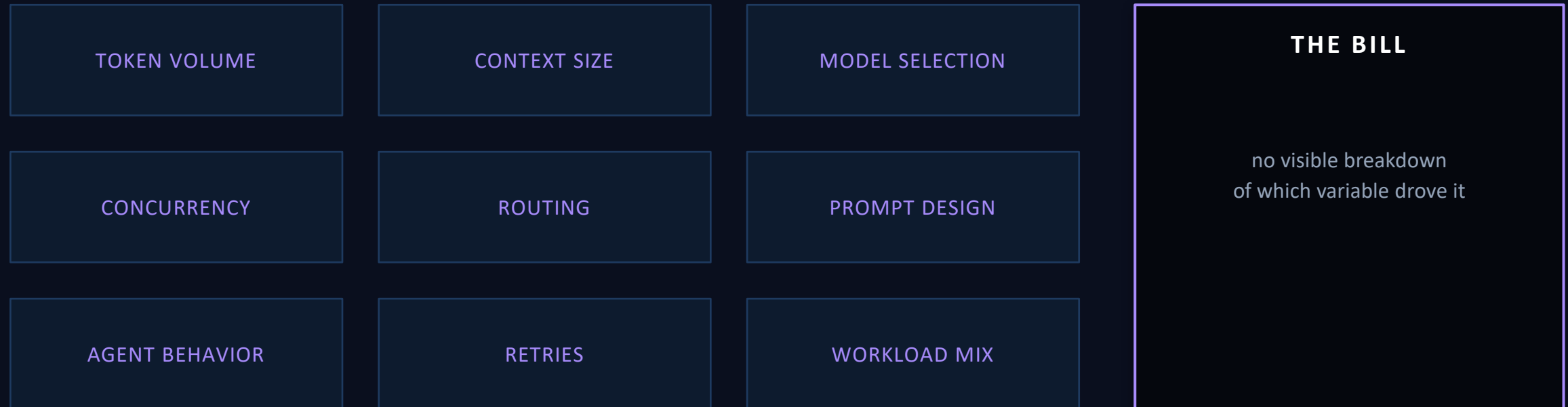
A Third Of Leaders Can't Explain Their AI Costs

33%

of business leaders cite limited understanding of usage costs as a key deployment challenge for AI agents.

KPMG Global AI Pulse, Q2 2026 — 2,145+ C-suite and senior executives across 20 countries

Nine Runtime Variables — Zero Capacity Questions



What Boards Assume vs. What's Actually True

EXECUTIVE ASSUMPTION	AI REALITY
More users means more cost	Model behavior matters more
Same workload, predictable cost	Runtime choices change cost fast
Budget controls spending	Runtime behavior controls spending
Capacity planning limits exposure	Consumption scales unexpectedly

42% Have Only Partial Visibility Into AI Spend

42%

KPMG Global AI Pulse, Q2 2026

- Visibility shows what happened. It doesn't show why.
- A dashboard can show a cost spike — not that a retry loop with no ceiling caused it.
- Routing shifts, context growth, agent retries: runtime decisions, not reporting events.

The Governance Question Changed

OLD GOVERNANCE QUESTION

“How much infrastructure are we willing to buy?”

NEW GOVERNANCE QUESTION

“What workload behaviors are we willing to permit?”

Visibility Correlates With 5x Higher ROI

STRONG VISIBILITY



5x

WEAK VISIBILITY



Organizations with strong AI spend visibility were roughly five times more likely to report achieving ROI on their AI investment.

KPMG Global AI Pulse, Q2 2026

THE TAKEAWAY

Runtime Behavior Controls Spending Now.

Governance that still asks how much infrastructure was bought is trying to govern AI economics economics from a layer that no longer generates the cost.

rack2cloud.com | Follow The Architect for more