AIMONGER
WHITE PAPER

AI Discovery

Stop Sending Every Prompt to the Flagship Model

Finance is starting to see the AI bill before it sees the surplus. Sending every prompt to a flagship model buys certainty of spend without certainty of value. Routing by task complexity is not a research trick. It is financial control for cognitive work, especially for mid-market firms that cannot absorb Silicon Valley-style usage waste.

David Saliba . 2026-07-20 . aimonger.com/whitepapers/model-routing-cost-governance/

The problem in one sentence

AI spend is becoming visible in OpEx before EBIT moves, and flagship-default routing makes that gap worse.

Finance leaders are seeing the first consolidated AI invoices: Copilot seats, API lines, retrieval services, agent platforms. Finance asks a fair question: “What did a successful case cost last month?” Engineering often cannot answer because every prompt hit the same premium model and nothing was tagged by workflow.

McKinsey’s State of AI 2025 survey finds 88 per cent of organisations regularly use AI in at least one function, but only 39 per cent report any enterprise-level EBIT impact, and most of those report less than 5 per cent of EBIT. About 6 per cent are high performers attributing more than 5 per cent of EBIT to AI. Routing will not create EBIT by itself, but unrouted flagship spend makes the absorption gap harder to close.

Model routing, plainly: choosing which model handles each request based on task type, risk, and quality need, not sending everything to the most expensive endpoint. **In practice:** invoice field extraction runs on a small model; ambiguous contract disputes escalate to a strong model only when the cheap path fails evaluation.

FinOps, in short: the operating practice of measuring, allocating, and optimising technology spend with engineering and finance at the same table. **Worked case:** weekly review of top workflows by spend, with tags that map tokens to “AP exceptions” not “misc AI.”

The default that burns money

Teams often hard-code a flagship model because it felt safe in a demo. In production, that default:

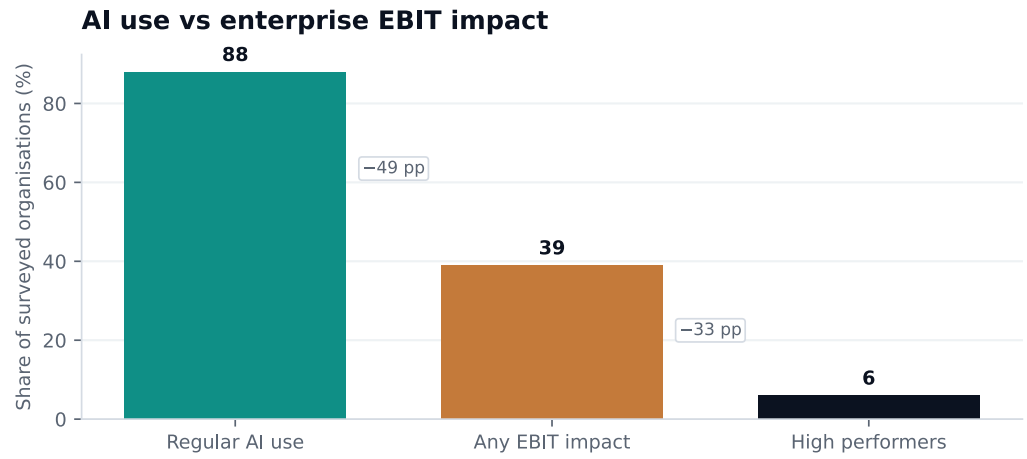
- **Inflates unit economics.** Every routine classification and extraction task pays frontier pricing even when a cheaper tier passes the same proposition tests on holdout. The premium is paid for capability the task never exercises, so the cost-per-case line carries frontier overhead on work a small model would handle. Over a month of high-volume extraction, the overspend is the difference between a controlled OpEx line and a surprise invoice the CFO escalates.

- **Creates latency that breaks workflows.** Premium endpoints add queue time and token volume that push interactive and batch SLAs past what operators can tolerate in daily work. A flagship model's median latency can be two to four times a small model's, and at P95 the gap widens further. Workflows that need sub-second response start missing SLAs, and operators route around the official path into shadow tools that are faster but ungoverned.
- **Hides simple task shapes.** Many production jobs are classification, extraction, or short drafting, work that never needed frontier reasoning but never got tagged by workflow. Because everything looks like "one prompt, one model," the team cannot see which slice of spend is routine and which is genuinely hard. The absence of tags is the absence of the cost narrative finance needs to defend the line.
- **Politicises FinOps conversations.** When finance asks for cost per successful case and engineering has no tier tags, the debate becomes trust-based instead of evidence-based. Engineering says "we need the best model for quality" and finance says "the bill is too high," and neither side has the data to settle it. Tier tags turn that argument into a table that both sides can read.

In a CIO office, the waste is usually not dramatic at first. It is a few teams using premium endpoints for tagging, short summaries, and routine extractions because nobody wrote the routing policy. By the time finance notices, the pattern has become the default.

What you see globally right now is a strange cost story: hyperscalers are spending at infrastructure scale, frontier model prices keep changing, and enterprise teams are being told that more AI usage is progress. The better question is not "how many tokens did we consume?" It is "what did a successful case cost?"

Eurostat's 2025 figure, 20.0 per cent of EU enterprises with 10+ employees using AI, shows many mid-market firms are still early. Early is the right time to install routing discipline, before spend hardens into habit.



McKinsey State of AI 2025 (survey). AIMonger redraw.

Figure 1. Regular AI use versus enterprise EBIT impact (McKinsey State of AI 2025). Plain read: most firms use AI somewhere; few can point to profit. **On the P&L:** if your tier mix is 100% flagship and your EBIT line is flat, routing is the first control to install, not the last. Source: McKinsey State of AI 2025 (survey). AIMonger redraw.

A simple routing policy

Tier	Example tasks	Model band (illustrative, July 2026)	Gate
T0	Formatting, tagging, obvious extraction	Small / cheap (e.g. Gemini Flash-class, DeepSeek-V4-Flash)	Spot checks
T1	Summaries, routine drafts, FAQ discovery	Mid (e.g. Claude Sonnet-class, Qwen mid-tier)	Automated eval sample
T2	Ambiguous analysis, multi-doc synthesis	Strong open or mid-closed (e.g. DeepSeek-V4-Pro, GLM-5.2)	Proposition tests
T3	High-stakes advice, novel reasoning	Closed flagship (e.g. GPT-5.x, Claude Opus-class, Gemini 3.x Pro)	Human gate

Plain read: SKUs churn; the tier logic does not. Re-pin model IDs in the routing policy each quarter, not in the board thesis.

Open-weight T1/T2 bands matter for FinOps and for alpha. Routing routine and strong synthesis work to self-hosted DeepSeek-V4 or GLM-5.2 keeps high-volume OpEx off closed flagship meters and keeps prompts on your infra. Reserve T3 closed US or EU-managed endpoints for novel high-stakes cases with human gates. Chinese-hosted cheap APIs are not a routing win for confidential classes; they are a different data path with a different jurisdiction. Origin policy belongs beside the tier table.

Escalate up when:

- **Refusal or low-confidence rules fire.** The cheaper tier correctly abstains or scores below threshold, which is the intended signal to spend more on a stronger model for that case only. Refusal is a feature of a well-tuned cheap model, not a failure; it tells the router where the cheap tier's competence ends. Escalating on refusal means premium spend lands on the cases that genuinely need it, rather than on every case by default.
- **Evaluation traps fail for the proposition family.** Automated checks on the frozen holdout show the lower tier cannot meet accuracy or format requirements for that workflow slice. The trap is a fixed test the tier must pass before it owns production traffic for that family. When it fails, the router promotes that slice to a higher tier and the cheap tier keeps the rest, so spend rises with evidence rather than with prompt length.
- **Operator or user requests review.** A human explicitly escalates for ambiguous, high-stakes, or novel cases rather than the router silently upgrading every prompt by default. Human escalation is the safety valve the operator controls, and it should be a one-click action with a logged reason. The alternative, silent auto-upgrade, is how a router becomes a flagship-default in disguise.

Never escalate because "the vendor logo looks better."

Proposition test means: a fixed question with a known correct answer used to check whether a model tier can handle a task family. **On Monday:** fifty paraphrases of "extract VAT number from this invoice layout"; if T0 passes, do not burn T3 tokens on every file.

Worked example: accounts payable extraction (synthetic)

A mid-market manufacturer processes 80,000 supplier invoices monthly across PDF, scan, and EDI formats.

Wrong default: every document through a flagship model because "accuracy matters." Blended cost tracks premium pricing; finance sees AI OpEx rise with invoice volume linearly.

Right routing:

Stage	Tier	Task	Gate
Intake	T0	Format detect, OCR cleanup	Spot check 1%
Core	T1	Line-item and VAT extraction	Automated proposition tests on 200-invoice daily sample
Exception	T2	Ambiguous tax or multi-currency	Escalate only when confidence or eval fails
Dispute	T3	Novel contractual interpretation	Human gate; never auto

Metrics after 90 days: cost per successfully posted invoice, escalation rate, exception rework rate.
Finance reading: report "EUR 0.04 per posted invoice" not "4.2M tokens."

Worked example: customer support drafting (synthetic)

A regulated services firm drafts email responses from policy and case history.

Wrong default: flagship model for every draft because tone matters. Latency breaks SLA; cost scales with ticket volume.

Right routing: T1 drafts from retrieved policy snippets; T2 only when retrieval confidence is low or the case tag is "complaint/legal"; T3 only when operator clicks "escalate analysis." Refusal rules unchanged across tiers.

Operating case: customers still get human-sent email; routing saves margin on routine acknowledgements without touching high-risk cases.

Metrics that belong in the CFO pack

1. **Cost per successful case by workflow.** Finance sees euros or dollars per accepted outcome tagged to AP, support, or discovery, not an undifferentiated token line on the cloud bill. The tag is what makes the number a business metric rather than a platform metric, and it is the figure the CFO can compare across workflows to decide where to invest or cut. Without it, the AI line on the OpEx is a black box the board cannot interrogate.

2. **Escalation rate to T2 and T3.** Track what share of jobs require expensive tiers; a rate above twenty-five per cent usually means the default tier is mis-set, not that routing failed. The rate is the diagnostic that tells engineering where the default tier is too low and tells finance where premium spend is concentrating. A rising rate on a stable workflow is the signal to retest the default tier on the holdout.

3. **Quality delta when escalating.** Measure whether higher spend actually buys accuracy on the holdout set, or only buys latency and vendor logo comfort without measurable gain. The delta is the return on the premium spend, and if it is near zero the escalation rule is spending money for nothing. This is the metric that stops the team from defaulting to the strongest model "just in case."

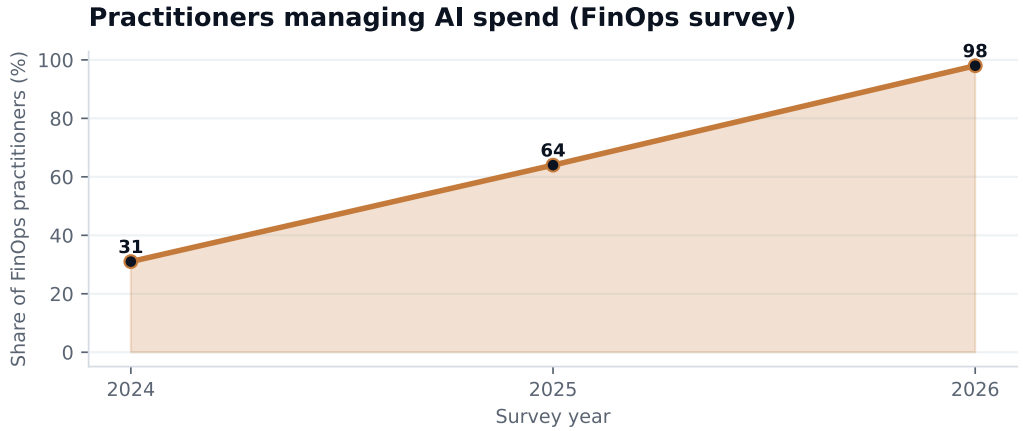
4. **Ceiling breaches per month.** Count how often workflows hit token, cost, or run caps, because repeated breaches signal missing budgets or runaway agent loops before invoice shock. A breach is a leading indicator that a workflow is mis-budgeted or that an agent loop is stuck, and either condition costs money if it runs to month-end. The count turns a surprise invoice into a weekly operations signal.

5. **Shadow spend outside the router.** Monitor consumer tools and untagged API keys that bypass routing policy, since governed savings mean nothing if shadow traffic doubles net spend. Shadow spend is the leakage that can erase a well-designed routing programme, and it usually grows

where the official path is too slow or too restrictive. The indicator tells finance and security where the official router is being routed around.

Gartner's warning that more than 40 per cent of agentic projects will be canceled by end-2027 for cost and unclear value applies to unrouted agent loops as much as to crews.

the discipline is also competitive. A company that routes well can afford broader coverage in discovery, support, or document operations without asking the board to accept an open-ended utility bill.



FinOps Foundation State of FinOps 2026 (1,192 respondents). AIMonger redraw.

Figure 2. Share of FinOps practitioners managing AI spend (FinOps Foundation State of FinOps 2026). In short: where FinOps exists, AI cost is already in the remit, and routing belongs in that rhythm. **On the P&L:** if 98% of FinOps peers track AI spend and you cannot tag by workflow, you are behind peer practice, not ahead of it. Source: FinOps Foundation State of FinOps 2026 (1,192 respondents). AIMonger redraw.

Implementation without a science project

Start with rules, not a learned router:

1. **Tag workflows with default tiers.** Assign each production workflow a starting tier based on proposition tests, not on whichever model felt best in the kickoff demo. The tag is the first routing decision and it is reversible, so the team can start conservative and move down as evidence accumulates. Tagging is also what makes spend visible by workflow, which is the prerequisite for every later optimisation.

2. **Wire escalation to evaluation failures.** Promote to a higher tier only when refusal rules or holdout traps fail, so spend rises with evidence rather than with prompt length alone. Escalation on evidence means the premium spend is justified by a measured gap, not by a hunch. It also keeps the default tier honest, because the team can see exactly which cases forced the upgrade.

3. **Log tier, cost, and outcome.** Every request records which tier ran, what it cost, and whether the case succeeded, giving FinOps the attribution layer routing is meant to provide. The log is the system of record for cost-per-successful-case, and it is the data the weekly review runs on. Without it, routing is a policy on paper and FinOps is guessing at the impact.

4. **Review weekly for mis-tiers.** Engineering and finance scan workflows where escalation or quality anomalies cluster, retagging defaults before bad habits harden into architecture. The weekly review is the feedback loop that keeps the tier map honest as workloads and models drift. A tier set at launch and never revisited is a stale policy that no longer matches production reality.

5. **Automate complexity classifiers last.** Learned routers earn their place only after static tiers and gates prove where cheap models suffice, not as day-one science project cover. A learned router adds operational complexity, a training pipeline, and a new failure mode, and it only pays

for itself when the static policy has already shown where the savings are. Build the evidence first, then decide whether the learned layer is worth its overhead.

FinOps operating rhythm

Weekly (engineering + finance partner):

- **Top workflows by spend.** Review the five costliest tagged workflows first, because routing discipline starts where the invoice already hurts rather than on experimental sandboxes. The top five are where the budget is actually concentrated, and they are the workflows where a tier change moves real money. Starting on the sandboxes optimises spend that does not matter and delays the conversation that does.
- **Escalation rate anomalies.** Flag workflows whose escalation rate jumped week-on-week, which often precedes a mis-tiered default or a broken confidence gate. A sudden jump is rarely a real change in task difficulty; it is usually a model update, a prompt change, or a gate that stopped firing. Catching the anomaly weekly prevents a quiet quality drift from becoming a month-end surprise.
- **Quality regressions after tier changes.** Compare holdout scores before and after any down-tier, rolling back automatically when accuracy drops beyond the pre-registered threshold. The comparison is the safety net that makes down-tiering safe, and the automatic rollback is what makes it fast. Without them, the team will hesitate to down-tier and the savings will never materialise.
- **Shadow spend indicators.** Track new API keys, blocked consumer tools, and untagged cloud lines that bypass the router, since shadow traffic can erase governed savings entirely. The indicators are the early warning that the official router is being circumvented, usually because it is too slow or too restrictive for a real workflow. Each new untagged key is a governance gap and a budget leak.

Monthly (steering):

- **Cost per successful case versus target.** Report whether each workflow beat its unit-cost target after routing changes, tying AI OpEx to outcomes finance can defend in the board pack. The target is the number the workflow signed up to, and the report is the accountability moment. Workflows that miss their target explain why and propose a correction, rather than carrying the miss silently into the next quarter.
- **Ceiling breaches and root causes.** Explain every cap hit, runaway agent loop, missing budget, or misconfigured tier, and assign an owner to prevent repeat breaches next month. A breach without an owner repeats; a breach with an owner and a root cause gets fixed. The monthly review is where the pattern across breaches becomes visible, which a weekly view can miss.
- **Tier policy changes requiring approval.** Treat default-tier moves like production config changes that need steering sign-off when they affect customer-facing or regulated workflows. Sign-off is the gate that prevents a cost optimisation from silently degrading a regulated workflow. It also forces the change to carry a written rationale, which is the record the auditor will ask for.

Quality gates before cheap routing

Never route down without a safety net:

1. **Cheap tier must pass proposition tests.** The lower tier clears the frozen evaluation pack for that proposition family on holdout before any production traffic shifts down. The test is the licence to move traffic, and without it the down-tier is a gamble with quality. Passing on holdout is not a guarantee of production quality, but failing is a guarantee the down-tier will break the workflow.
2. **Refusal rules must still fire.** Down-tiering never disables abstention or low-confidence gates, because cheaper models that never refuse can look accurate until the first consequential error. A cheap model that refuses on unknown input is safe; a cheap model that guesses confidently on

everything is a liability. Refusal is the behaviour that makes a cheaper tier trustworthy on the long tail.

3. **Spot human review after down-tier.** Sample human review runs for two weeks after any tier reduction, catching regressions automated checks miss on long-tail layouts or languages. Automated eval covers the holdout, but production traffic includes layouts and dialects the holdout does not. Human review is the backstop for the cases the eval set never anticipated.

4. **Instant rollback to prior policy.** Keep one-command rollback to the previous tier map so a quality dip reverts routing before operators lose trust or finance sees rework costs spike. The rollback is the confidence the team needs to try a down-tier at all. Without it, down-tiering feels irreversible and the team will leave spend on the table rather than risk a change.

Worked numbers (illustrative arithmetic, not a benchmark)

If a workflow runs 100,000 jobs/month:

- **Eighty per cent at T0 or T1 (1x unit cost).** Routine tagging, extraction, and short drafts run on cheap tiers when proposition tests pass, forming the bulk of volume at baseline unit economics. This is the slice where routing earns its keep, because the volume is high and the task is simple. Moving this slice off the flagship tier is the single largest saving in the policy.

- **Fifteen per cent at T2 (8x unit cost).** Ambiguous cases escalate to a strong model only when confidence gates or holdout traps fail on the cheaper path, keeping premium spend on the long tail. The escalation is evidence-driven, so the premium spend tracks the cases that genuinely need it. Fifteen per cent is illustrative; the real rate is the diagnostic that tells you whether the default tier is well-set.

- **Five per cent at T3 (20x unit cost).** Novel disputes and high-stakes interpretation stay on the flagship tier with human gates, never as the default for the entire monthly job volume. This is the slice where frontier reasoning is actually required, and the human gate is the control that keeps it honest. Five per cent of volume at twenty-times unit cost is affordable; one hundred per cent is not.

Blended cost ≈ 0.8×1 + 0.15×8 + 0.05×20 = 0.8 + 1.2 + 1.0 = 3.0x versus 20x all-flagship.

The saving only matters if T0/T1 quality holds. That is why evaluation is a finance control.

Procurement implications

Vendors that force all traffic through one premium endpoint without routing hooks create structural overspend. Ask for:

- **Per-request model selection.** The contract must allow your router to choose tier per call, not lock every workflow to the vendor's most expensive default SKU. Without this clause, the vendor sets your tier policy and you pay their margin on every request. Per-request selection is the contractual foundation that makes routing legal under the vendor relationship.

- **Usage exports by workflow tag.** Finance needs CSV or API exports keyed to your workflow IDs, so cost per successful case is computable without manual spreadsheet archaeology. The export is the data pipeline that turns the bill into a business metric. If the vendor cannot tag by your workflow IDs, finance will rebuild the attribution in spreadsheets every month, which is fragile and late.

- **Hard budget controls.** Per-workflow and per-day caps must stop runaway agent loops at the gateway, not only warn after the month's invoice is issued. A warning after invoicing is a report on damage already done; a gateway stop prevents the damage. The distinction is whether the cap protects the budget or merely documents the breach.

- **Latency SLOs per tier.** Each tier needs stated P95 latency bounds so down-tiering for cost does not silently break SLAs operators depend on in production. The SLO is the contract that makes

down-tiering safe, because it gives the team a measurable guarantee that cheaper does not mean slower beyond a bound. Without it, a cost saving can become an SLA breach and a customer complaint.

Appendix: routing policy template

Complete one policy row per workflow. Leave the Entry column blank until owners fill it.

Field	Entry
Workflow	
Default tier	
Escalate when	
Never auto-run when	
Monthly ceiling	
Owner	
Review cadence	

Plain read: empty cells are intentional. Untagged traffic with no policy row is an organisational failure mode, not a tooling gap.

Attribution design

Tag every production request with workflow ID, tier, model, retrieval cost, and success/failure. Without tags, FinOps can see a bill and cannot see a business case. FinOps Foundation evidence that almost all surveyed practitioners now manage AI spend makes untagged traffic an organisational failure mode, not a tooling gap.

Quality-preserving savings playbook

1. **Freeze evaluation for the workflow.** Lock the holdout proposition set before changing tiers so quality comparisons remain honest across the down-tier experiment. Freezing the set means the before and after scores are comparable, and the team cannot quietly swap in easier questions after a poor result. The frozen set is the honesty layer of every saving claim.
2. **Down-tier only passing families.** Shift traffic down only for proposition families that already pass on the cheaper model in eval, not for every task in the workflow indiscriminately. Indiscriminate down-tiering saves money on the easy slice and loses quality on the hard slice, and the net can be negative. The proposition family is the unit of routing decision, not the workflow.
3. **Keep refusal and temporal tests mandatory.** Cheap tiers must still abstain on unknown layouts and respect document dates, because silent guessing on T0 is how routing saves money and loses audits. A cheap model that ignores dates will cite a superseded policy confidently, and the audit will ask why. Refusal and temporal checks are the controls that keep the cheap tier safe on the long tail.
4. **Sample human review for two weeks after change.** Operators or QA review a fixed sample post-change, catching edge-case regressions automated evals miss on real supplier PDFs or dialects. The sample is the backstop for production diversity the holdout does not cover. Two weeks is long enough to see a real distribution of cases, not just the first day's traffic.
5. **Roll back automatically on regression.** Pre-register the quality threshold that triggers instant return to the prior tier policy without waiting for the next steering meeting. The threshold is the number the team committed to before the change, and the rollback is the action that honours it. Automatic rollback is what makes down-tiering reversible and therefore safe to attempt.

Escalation tier, put simply: a defined step up to a more capable (and expensive) model when cheaper tiers fail quality or confidence rules. **Operating case:** escalation rate above 25% on a workflow means your default tier is wrong, not that you need a fancier router.

Shadow spend and the router bypass

Shadow AI spend (plain English): model usage outside the governed router, consumer tools, untagged API keys, vendor sandboxes charged to department cards. **On the P&L:** routing saves 40% on governed traffic while shadow spend doubles; net bill still rises.

Require the same workflow tags on any approved path. Finance and security share a monthly shadow indicator: new API keys, consumer-tool blocks, untagged cloud lines.

Agent loops amplify routing mistakes

When agents call models in loops, tier mistakes compound. A T3-default agent that re-plans five times per task can burn more than a single long flagship call. Agent workflows need per-run and per-day ceilings **before** autonomy expands.

Gartner forecasts that more than 40 per cent of agentic AI projects will be canceled by end-2027, citing escalating cost among drivers. Routing policy is agent policy.

Source	Finding	Implication
FinOps Foundation State of FinOps 2026	98% of 1,192 surveyed practitioners manage AI spend (up from 31% two years earlier); AI cost management is a top priority and top desired skillset	Routing and attribution are finance controls, not optional engineering elegance
Linux Foundation FinOps press release (Feb 2026)	Same survey represents more than $83 billion in annual cloud spend; many organisations are asked to self-fund AI via optimisation savings	Cost-per-successful-case becomes a capital allocation language
Gartner (2023)	Projected mainstream GenAI API/application use by 2026	Unrouted flagship-default traffic scales with adoption
Gartner (2025)	Forecast of heavy agentic project cancellation driven partly by escalating cost	Agent loops amplify the need for tiered routing
Stanford AI Index 2026	High organisational adoption with agent use still early	Install routing before agent volume hardens bad defaults

Methodology note

FinOps respondents are practitioners already managing technology spend; the 98% figure is not a census of all enterprises. It shows that where FinOps exists, AI cost has entered the remit. Routing policies should still be validated against quality gates for each workflow.

Counter-position: routing adds latency and complexity

Poor routing does. Good routing starts with static workflow tiers and escalation rules, not a machine-learned router on day one. Measure quality delta when escalating. If cheap tiers fail proposition tests, keep those families on higher tiers until evaluation says otherwise.

CFO operating standard

Require monthly reporting of:

1. **Cost per successful case by workflow.** Each major workflow reports unit economics tagged to business outcome, not aggregate token volume that finance cannot map to AP or support. The tag

is what turns a cloud line into a business line, and it is the figure the CFO compares across workflows to allocate the next budget. Without it, the AI spend is a number with no narrative.

2. **Tier mix and escalation rate.** Show what share of jobs ran at T0 through T3 and how often escalation fired, exposing mis-tiered defaults before they become cultural habit. The mix is the picture of where spend actually landed, and the escalation rate is the diagnostic of whether the defaults are well-set. Together they let finance see the policy in action, not just on paper.

3. **Ceiling breaches.** List every cap hit with root cause and owner, because repeated breaches on the same workflow signal missing stop rules or runaway agent behaviour. A breach is a budget control doing its job, but a repeated breach on the same workflow is a design fault the owner must fix. The list turns a monthly invoice surprise into an assigned action list.

4. **Quality outcomes after down-tier changes.** Report holdout accuracy and human correction rate after any tier reduction, proving savings did not purchase silent quality debt. The report is the evidence that the saving is real, and it is the defence against the charge that routing traded quality for cost. Without it, every saving is suspect and every regression is unattributable.

5. **Shadow AI spend outside the router.** Include consumer-tool and untagged API indicators monthly, since governed routing savings disappear when shadow traffic grows unchecked. The indicator is the leakage rate, and a rising rate is the signal that the official router is losing ground to convenience tools. Finance and security share this number because it is both a budget risk and a data risk.

Research addendum: what the routing literature actually supports

Public research supports routing as a **measured control**, not a guaranteed saving.

Study	Result	Caveat for boards
RouteLLM (2024)	Learned routers reduced API cost by more than 2x on evaluated public benchmarks without measured quality loss in the study setting	Strong/weak model pairs, benchmark-dependent quality, price snapshots
FrugalGPT (2023/ TMLR)	Cascades matched top-model quality with **up to** 98% lower API cost on selected tasks	Maximum across selected datasets, not a typical enterprise multiplier; cascades raise tail latency
LLMRouterBench (Jan 2026)	>400k query-model instances across 21 datasets and 33 models; many sophisticated/ commercial routers did not reliably beat simple baselines	Benchmark study, not a randomized enterprise deployment

Operating rule: a router earns its complexity only if it beats a one-model baseline after including router overhead, retries, evaluation labour, and p95/p99 latency. More models in the pool do not automatically improve outcomes; curation matters more than catalogue size (LLMRouterBench).

FinOps Foundation guidance also implies a sequence: meter and allocate before sophisticated optimisation. Token price alone is not the unit of economics; report **cost per accepted outcome** with quality, escalation rate, and human correction.

Gartner's January 2026 public forecast put worldwide AI spending near **$2.5 trillion in 2026**, with AI infrastructure the largest category. Falling unit prices can coexist with rising aggregate spend as usage, agent loops, and context lengths expand.

Ninety-day routing installation (CFO + CIO)

Days 1-30: Tag top five workflows by spend; assign default tiers; start logging cost per outcome. **Days 31-60:** Add proposition tests and escalation triggers; publish weekly tier mis-classification review. **Days 61-90:** Down-tier families that pass eval; report blended cost vs all-flagship counterfactual; freeze policy or roll back.

Investment committee approves scale only when cost per successful case trends down without quality regression on the promotion holdout.

Stop sending every prompt to the flagship model.

Routing is how leadership converts commodity intelligence into controlled unit economics, a prerequisite for absorption that shows up in capacity or EBIT rather than in a surprise invoice.

References

1. FinOps Foundation, "State of FinOps 2026 Report." https://www.finops.org/insights/state-of-finops/
2. Linux Foundation / FinOps Foundation press release on State of FinOps 2026 (1,192 respondents; 98% manage AI spend). https://www.linuxfoundation.org/press/state-of-finops-survey-ai-value-and-skills-top-priorities-as-finops-matures-across-technology-value-98-manage-ai-90-saas-64-licensing-48-data-center-1
3. Gartner, "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026," press release, 11 October 2023. https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026
4. Gartner, "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," press release, 25 June 2025. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027
5. Stanford Institute for Human-Centered Artificial Intelligence, "The 2026 AI Index Report," Economy chapter. https://hai.stanford.edu/ai-index/2026-ai-index-report/economy
6. McKinsey & Company / QuantumBlack, "The State of AI: Global Survey 2025." https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
7. Eurostat, "20% of EU enterprises use AI technologies," 11 December 2025 (20.0% in 2025; 13.5% in 2024). https://ec.europa.eu/eurostat/en/web/products-eurostat-news/w/ddn-20251211-2
8. Ong et al., "RouteLLM: Learning to Route LLMs with Preference Data," 2024. https://arxiv.org/html/2406.18665
9. Chen, Zaharia and Zou, "FrugalGPT," 2023 (TMLR). https://arxiv.org/abs/2305.05176
10. Li et al., "LLMRouterBench," January 2026 preprint. https://arxiv.org/abs/2601.07206
11. Gartner, "Gartner Says Worldwide AI Spending Will Total $2.5 Trillion in 2026," press release, 15 January 2026. https://www.gartner.com/en/newsroom/press-releases/2026-1-15-gartner-says-worldwide-ai-spending-will-total-2-point-5-trillion-dollars-in-2026
12. FinOps Foundation, "FinOps for AI Overview." https://www.finops.org/wg/finops-for-ai-overview/
13. CNBC, "Chinese AI models gain ground with U.S. companies as costs surge" (7 July 2026). https://www.cnbc.com/2026/07/07/chinese-ai-models-costs-us-openai-anthropic.html
14. DeepSeek, "DeepSeek V4 Preview Release" (24 April 2026). https://api-docs.deepseek.com/news/news260424/
15. Z.ai / Zhipu, "GLM-5.2: Built for Long-Horizon Tasks" (16 June 2026). https://z.ai/blog/glm-5.2

Published by AIMonger . https://aimonger.com . Malta . Europe