Routing the Margin: What AI Optimisation Platforms Are Worth as Arbitrage Narrows

A Transaction Framework for Defensible Routing, Evaluation, Control and Retained Contribution

Authored by Chennakeshav Adya

Independent Researcher

September 2026

Abstract

AI optimisation platforms promise to route workloads across models, providers and deployment modes while balancing quality, latency, policy and cost. Official AWS and Microsoft documentation describes managed intelligent routing. Provider platforms also offer batching, caching, regional inference, evaluation and security controls. These developments change the strategic boundary for independent optimisation platforms and require investors, buyers and boards to distinguish temporary price arbitrage from durable operating capability.[1][2][3][4]

This paper develops a transaction framework for testing whether an optimisation platform owns defensible workload evidence, evaluation, orchestration, policy, distribution and customer value after model prices and managed routing converge. It defines the economic unit as an accepted business outcome after complete platform cost. It then connects workload classification, routing performance, fallback, caching, batching, provider concentration, switching, rights, security and customer ownership to gross margin, cash flow and valuation.

The framework treats routing as a control problem with economic consequences. A router must decide among eligible models under quality, latency, privacy, availability and policy constraints. The cheapest eligible model can improve contribution only when its accepted-outcome rate, retries, escalation and customer service remain within threshold. A platform that cannot reproduce these results across representative workloads has an unstable margin claim.

A hypothetical enterprise case demonstrates the method. Every task volume, quality score, price, cost, probability and valuation input is a management assumption created solely to show the calculation. It is not a forecast, quotation or representation of a named company. Six figures and seven tables connect workload classification, route scoring, lifecycle cost, switching, acquisition value and a 180-day implementation programme.

AI, privacy, intellectual property, competition, cyber-security, export, tax, accounting, valuation and investment decisions require current specialist advice from qualified professionals. Models, provider terms, prices and regulation change. This paper provides general information for professional audiences and does not provide legal, regulatory, tax, accounting, technical or investment advice.

JEL Classification: G24, L11, L21, L86, O32

Keywords: AI routing, inference optimisation, model selection, caching, batching, gross margin, switching, valuation, M&A

1. Define the routing transaction thesis

The investment case should identify the scarce asset that survives provider-price convergence. Candidate assets include proprietary workload evidence, a reproducible evaluation corpus, orchestration and policy

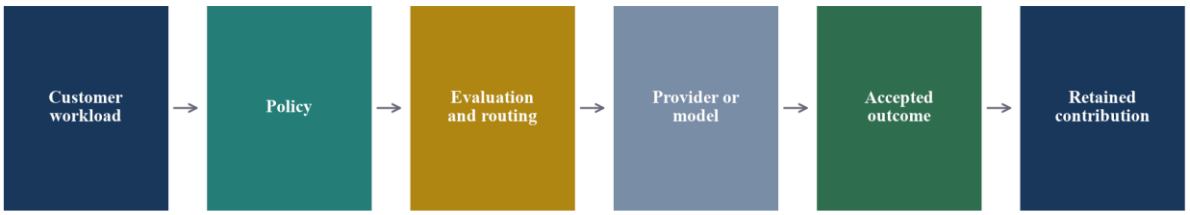
controls, distribution, customer contracts, operating data and switching capability. Temporary access to a cheaper model is a trading advantage whose value decays when providers change prices, rivals obtain similar access or a cloud platform bundles comparable routing.

The evidence chain begins with an eligible task, proceeds through policy, routing, model response and acceptance, and ends with realised customer value, complete cost and cash. Model access alone does not establish durable value. The platform must control enough of this chain to sustain outcomes and economics as providers and models change.

The counterfactual must use the same workload, quality threshold, service level and time horizon. A routed case that excludes evaluation and support cannot be compared with direct deployment that includes them. A platform case that credits list-price savings cannot be compared with a customer baseline using negotiated rates. Finance should normalise scope, timing, tax, working capital, infrastructure and exit obligations.

The thesis needs an expiry test. Management should specify the effect of provider-price convergence, standard cloud routing, customer direct discounts and changes in the best-performing model. A defensible platform retains value through evidence, workflow integration, governance, distribution or switching capability. A fragile platform loses contribution when the spread between buy price and customer price narrows.

Figure 1. Optimisation platform value chain



The chain identifies evidence and control points; company-specific diligence determines ownership and durability.

2. Classify workloads before routing them

Workload classification should precede vendor selection. Relevant dimensions include consequence of error, task variability, required context, latency, volume, data sensitivity, explainability, tool access, language, modality, offline need and human oversight. The same enterprise can have thousands of low-consequence classification events and a small number of consequential legal or credit decisions.

Classification should include the cost of abstention and escalation. A model that refuses uncertain cases can be economically attractive when a human channel exists and the residual volume is manageable. The same refusal pattern can destroy a fully automated service proposition. Evaluation should therefore measure coverage, accepted accuracy and the cost of unresolved cases together.

Task stability matters. A stable schema and narrow domain can reward investment in a smaller model, deterministic validation and specialised data. A rapidly changing knowledge task may favour retrieval and an external model whose capability evolves. A workflow with changing tools can place more value on orchestration than on the underlying model. The board should invest at the layer where differentiation can persist.

Table 1. Workload classes and routing hypotheses

Workload	Primary requirement	Plausible route	Evidence gate
Repetitive classification	consistency and unit cost	small specialised model	held-out accuracy and drift
Enterprise retrieval	faithfulness and permission control	small or frontier model with retrieval	grounded answer and access test
Complex synthesis	broad capability and judgement support	frontier API with review	expert acceptance
High-volume extraction	throughput and stable schema	small model or rules plus model	field-level precision
Sensitive local task	privacy and latency	private or on-device model	data-path verification
Rare frontier task	broad knowledge and reasoning	managed frontier API	outcome quality and escalation

Routes are starting hypotheses and require evaluation against company evidence.

3. Map the optimisation stack

The stack begins with request intake and identity. It applies customer, data, geography and policy eligibility before comparing candidate models. Downstream components can add retrieval, tools, response validation, fallback, human escalation, caching, logging and billing. Investors should map each component to its owner, supplier and customer dependency.

Routing can operate through static rules, learned selection, cascades or ensembles. A static rule sends a known workload class to one model. A learned router estimates which eligible model is likely to meet a threshold. A cascade tries an economical model before escalating. An ensemble uses several outputs. Each design has different latency, failure and cost characteristics.

Managed routing changes the make-or-buy boundary. AWS documents intelligent prompt routing within model families, and Microsoft documents a router that selects from an eligible model set.[1][2] An independent platform therefore needs evidence of value beyond a standard provider feature. Cross-provider routing, customer-specific policy, model-neutral evaluation, application integration and commercial distribution are potential sources of differentiation.

The stack map should expose hidden concentration. A platform may advertise many models while relying on one cloud, one provider for most accepted outcomes, one embedding model or one observability service. Revenue diversity can coexist with technical concentration. Diligence should trace actual request, cost and control paths.

4. Establish the direct-deployment baseline

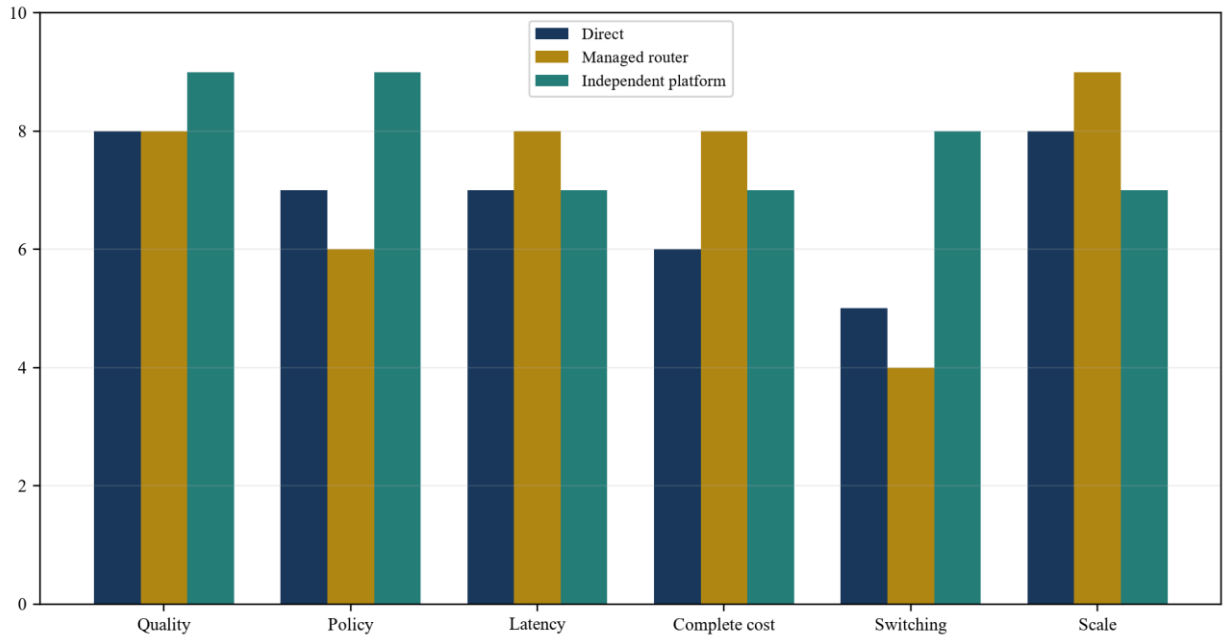
Every optimisation claim requires a direct-deployment baseline. The baseline should reflect the customer's feasible use of providers, including negotiated pricing, batch discounts, caching, cloud commitments, engineering, observability and support. Public list price can overstate a platform's saving when a customer can obtain lower direct economics.

The comparison should use the same workload, quality threshold, latency and policy envelope. The lowest nominal price is irrelevant when the selected model fails acceptance. The baseline is the lowest-cost direct route that meets the stated service requirements. Evidence ranges should reflect variation by cohort, prompt length, fallback and period.

Provider pricing is dynamic and multidimensional. OpenAI, AWS, Microsoft, Google and Anthropic publish structures that vary by model, input, output, caching, batch, region, capacity and service

form.[4][5][6][7][8] The transaction model should preserve the date, currency, unit and contractual source for every price. Future price assumptions belong in scenarios.

Figure 2. Illustrative route-performance scorecard



Scores are hypothetical management assumptions for method demonstration.

5. Evaluate on the enterprise task distribution

Public benchmarks help screen candidates. Investment decisions require a representative enterprise evaluation set with easy, difficult, frequent, rare and adverse cases. Amazon Bedrock documents automatic, human and model-judge evaluation and permits custom prompt datasets.[10] The enterprise should preserve ground truth, rubric, reviewer qualifications, model settings and version.

Quality should be expressed as accepted outcomes, not aesthetic preference. Measures can include correctness, completeness, faithfulness, instruction following, refusal, harmfulness and task-specific business criteria.[11] Evaluation must remain separate from training or prompt optimisation to reduce leakage.

The acceptance threshold should reflect consequence. A marketing draft can tolerate correction before publication. A payment, clinical, legal or safety workflow requires qualified review and controls appropriate to the actual decision. Aggregate scores can hide severe failure in a small but important cohort. The report should therefore show distribution, worst-case and subgroup results.

Model judges can improve scale, while they require calibration against human experts. The organisation should measure agreement, investigate systematic disagreement and retain human review for consequential cases. Evaluation data can also become a sensitive asset because it reveals customer workflows, internal rules and weaknesses. Access and retention should be controlled.

6. Reconstruct complete platform cost

Token price is one component. Complete cost includes prompts and outputs, embeddings, retrieval, tools, retries, evaluation, moderation, human review, orchestration, infrastructure, observability, security, support, compliance, model updates and failed outcomes. It also includes reserved commitments, unused capacity and exit work.

Cost should be measured per accepted outcome and by cohort. A cheaper model can require more retries, longer prompts or greater human review. A capable model can reduce tool calls or raise resolution. Route economics therefore combine price, consumption, quality and operating consequences. The ledger should reconcile provider records and infrastructure usage to customer or process events.

Fixed and variable cost affect scale differently. Internal deployment can create capacity steps and underutilisation. APIs can offer elasticity and expose the enterprise to price or currency changes. Reserved capacity can reduce unit price while increasing commitment risk. The scenario model should identify the volume at which preferred routes change and test demand volatility around that point.

Table 2. Complete platform cost ledger

Cost layer	Driver	Verification	Common distortion
Model input and output	tokens or units	provider invoice and logs	list price used instead of contracted rate
Caching and batch	eligible traffic	cache and queue records	gross saving without misses or delay
Retrieval and tools	calls and storage	service records	excluded from inference cost
Retry and fallback	failed or low-confidence calls	trace-level logs	duplicated calls hidden in average
Evaluation and observability	tasks and traces	platform records	treated as discretionary overhead
Human review and support	cases and service time	workflow and payroll evidence	assigned outside cost of service
Commitments and capacity	reserved spend	contracts and invoices	unused commitment omitted

The ledger requires company-specific measurement and consistent accounting policy.

7. Test routing quality and cost together

A router should be evaluated against direct eligible alternatives on the same task set. The test records accepted outcome, latency, provider cost, complete cost, fallback, escalation and policy compliance. The purpose is to measure whether routing improves the efficient frontier for the workload.

The policy can minimise expected cost subject to acceptance, latency and risk constraints. These variables are uncertain and often correlated. An economical model may need longer prompts or more retries. A fast model may trigger more human correction. A capable model can reduce tool calls. End-to-end traces are required to measure the sequence.

Fallback deserves separate treatment. A cheap-first cascade can appear attractive when only the first call price is reported. Its complete economics include the failed call, subsequent call, added latency and any customer loss. Fallback creates value when the first model resolves enough eligible cases and the escalation threshold is calibrated.

Routing performance can drift after model, prompt, policy or workload changes. The platform should maintain versioned evaluations, canary tests, incident thresholds and rollback. A router that learns from production data needs controls over feedback quality and customer rights. Repeatable adaptation supports durable value.

8. Price caching and batching correctly

Caching can reduce repeated-input cost and latency when requests share eligible prefixes or responses. Its value depends on hit rate, retained duration, invalidation, privacy, provider terms and workload stability.

A claimed saving should reconcile to eligible traffic and actual provider bills. The analysis should also test whether the customer receives the saving through lower pricing.

Batch services can lower provider cost for workloads that tolerate delayed completion. OpenAI publishes batch pricing and prompt-caching features; other providers publish their own terms.[4][25][26] A platform creates value when it identifies delay-tolerant work, manages queues and preserves service commitments while retaining an agreed share of the saving.

Caching and batching can affect working capital. The platform may prepay capacity or consume minimum commitments before collecting customers. Queue failures create reprocessing and service-credit exposure. Finance should model cash timing, utilisation and incident cost alongside the nominal unit discount.

Table 3. Optimisation and switching-cost register

Optimisation or dependency	Evidence	Economic effect	Mitigation
Prompt caching	eligible traffic and hit logs	lower cost and latency	test invalidation and misses
Batch processing	queue and service records	discount with delayed output	classify delay-tolerant work
Provider-specific API	feature inventory	engineering and delay	abstraction and fallback
Reserved capacity	commitment schedule	stranded spend	staged commitment
Data-zone availability	regional deployment map	service restriction	eligible alternatives
Customer promise	contract review	consent or liability	route-neutral drafting

Treatment should reflect contracts, architecture and tested migration evidence.

9. Treat routing as a strategic asset

A router can place simple tasks on small models and complex tasks on frontier capability. Static rules provide transparency. Learned routers can respond to prompt characteristics. Microsoft describes a lightweight routing model that selects among eligible models and reports the selected model.[2] Routing value depends on quality, latency, cost and policy evidence after routing overhead and fallback.

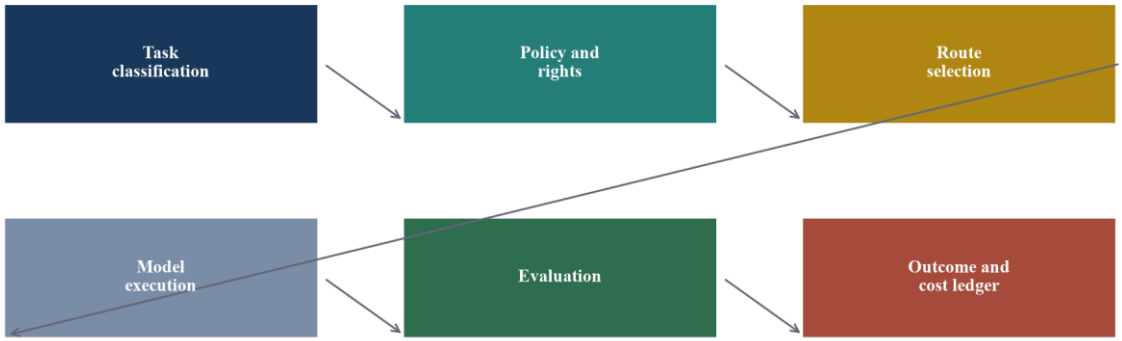
The router must respect eligibility before optimisation. Geography, data class, customer contract, model approval, safety and capacity can narrow the candidate set. Within that set, the route can optimise expected accepted value net of cost and latency. Policy violations should fail safely and create an auditable event.

Routing also creates model-selection data. The enterprise can learn which workloads need frontier capability and which perform well on smaller models. This evidence informs negotiation, product design and future build or acquisition. The router therefore becomes a capital-allocation measurement layer when its decisions and outcomes reconcile.

The platform should disclose route share and outcome concentration. A claim of multi-model optimisation has limited substance when one provider handles nearly all economically important work. Route share should be measured by requests, provider cost, accepted outcomes, customer revenue and retained contribution. Each view answers a different diligence question and can reveal concentration hidden by request counts.

Feedback quality is another asset. Customer corrections, human escalations and downstream process results can improve routing when they are lawfully collected, accurately attributed and protected from manipulation. Weak feedback can reinforce poor choices. The transaction team should examine provenance, consent, retention, reviewer quality and the effect of feedback on production decisions.

Figure 3. Model-routing control loop



The loop is a proposed operating architecture.

10. Secure model data output and evaluation rights

Rights should cover weights, adaptations, training and evaluation data, prompts, retrieval content, outputs, telemetry, derivative work, audit and transfer. Open availability does not mean unrestricted commercial use. Google describes Gemma as a starting point that users adapt and deploy under applicable legal and responsible-use obligations.[5]

An acquisition should trace employee, contractor and third-party contributions. A partnership should state use of customer data, retention, training, confidentiality, infringement allocation and change of control. Unclear rights reduce portability and transaction value.

The diligence schedule should list every material component and its licence, owner, restriction and termination consequence. Open-source notices, model licences, data permissions and software dependencies must match the planned commercial use. A licence that works for internal evaluation can fail for redistribution or a customer-facing product.

Output rights do not resolve factual reliability or third-party claims. The operating model still needs review, attribution and use controls appropriate to the workflow. Contractual indemnity has economic value only when scope, exclusions, process and counterparty capacity are understood.

11. Distinguish control from commodity plumbing

The operating model needs product, model, data, evaluation, security, privacy, infrastructure, finance and procurement ownership. Small models can reduce dependency while increasing responsibility for serving and updates. Frontier partnerships can reduce infrastructure burden while requiring vendor, route and usage governance.

The board should see accepted outcomes, route mix, complete cost, incidents, provider concentration, version changes and migration readiness. Technical metrics without contract and cash reconciliation are insufficient.

Release governance should bind a model version to its evaluation, prompt, tools, retrieval data, policies and infrastructure. A change in one component can alter quality and cost. Canary tests, rollback and incident authority reduce the risk of invisible production drift. Material changes should trigger customer or regulatory processes where required.

Talent strategy differs by route. Build requires durable engineering and research capacity. Buy requires retention and integration of scarce teams. Partner requires strong product, architecture, evaluation and

commercial management even when model operations are outsourced. The board should avoid assuming that procurement removes the need for internal competence.

12. Govern safety, privacy and cyber-security

The NIST AI Risk Management Framework and Generative AI Profile provide structured risk-management resources.[6][7] Controls should follow the actual workload, consequence and jurisdiction. Relevant issues include data leakage, prompt injection, insecure tools, harmful output, model extraction, supply-chain compromise and unreliable automation.

Build, buy and partner allocate responsibility differently. The enterprise retains accountability for its product and decisions. Contracts and shared-responsibility statements should match the technical architecture and incident process.

Privacy review should map the complete data path, including prompts, retrieval, tools, logs, human review and support. Retention and training settings should be verified technically and contractually. Data minimisation and purpose limitation can reduce both risk and cost. Sensitive workflows may require private or local deployment, subject to actual regulatory advice.

Security testing should include the orchestration layer and connected tools. A safe model can still execute an unsafe action through excessive permissions. Identity, least privilege, transaction limits, approval and logging should sit around consequential tools. Acquired products should be tested against the buyer's threat model before integration expands access.

13. Underwrite provider concentration and counterparty exposure

Provider concentration can affect price, capacity, geographic availability, model continuity and negotiating leverage. Diversification has costs because models differ in behaviour, interfaces and controls. The decision should compare the expected loss from dependency with the cost of maintaining alternatives.

Counterparty review includes financial capacity, service history, security, subcontractors, data processing, intellectual property, sanctions, export controls and termination support. A small-model partner can also create concentration when it supplies unique weights or expertise.

Concentration should be measured at several layers. Multiple applications can appear diversified while using the same model, cloud region, accelerator supplier or identity service. The enterprise should map common failure and commercial dependencies. A portfolio limit can apply to critical workloads, spend or revenue exposed to one dependency.

Financial exposure includes committed spend, prepayments, credits and transition cost. Operational exposure includes throughput, latency and service recovery. Strategic exposure includes roadmap influence and access to new capability. The board can accept concentration where benefits are strong, provided liquidity, service and migration plans match the downside.

14. Test customer ownership and distribution

Customer ownership is a central diligence question. The platform may control the commercial contract while a model provider, cloud or application partner controls the technical relationship. Investors should identify who can change price, terminate service, contact the customer and use workload data. Revenue attributed to the platform can have limited durability when another party owns the critical relationship.

Competition analysis can consider model access, data, ecosystems, foreclosure and innovation under current merger rules.[8][9][10] Transaction counsel should assess the actual jurisdictions and structure.

Distribution can provide defensibility when the platform reaches customers through embedded workflows, trusted channels or specialist implementation. Its value should be tested through acquisition cost,

activation, retention, cross-sell and contribution by channel. A referral channel with weak renewal evidence should receive a different value from an embedded product relationship.

Technical diligence should reproduce claimed quality and cost. Commercial diligence should test willingness to remain after price, provider or control changes. Financial diligence should normalise cloud credits, capitalised development, founder compensation and shared infrastructure. Legal diligence should test model, data and code rights and change-of-control terms.

Talent concentration can be material when a few individuals hold undocumented knowledge. Retention arrangements need knowledge transfer, succession and institutional documentation. Employment incentives should align with integration outcomes and sustainable operating evidence rather than a short demonstration milestone.

15. Construct a hypothetical operating case

Consider a hypothetical enterprise processing 12 million monthly tasks across extraction, retrieval, customer support and complex analysis. Every value is a management assumption for method demonstration. It is not a forecast or market quotation.

Table 4. Hypothetical workload and routing case

Workload	Monthly tasks	Accepted outcome	Selected route	Complete cost per accepted outcome
Extraction	6.0m	96%	economical model with validation	AED 0.05
Retrieval	3.5m	91%	routed specialist portfolio	AED 0.18
Customer support	2.2m	88%	cascade with fallback	AED 0.42
Complex analysis	0.3m	81%	frontier model plus review	AED 6.80

All values are illustrative management assumptions.

The portfolio chooses routes by workload and retains evaluation, routing, data and outcome evidence as common assets. The assumed annual customer revenue is AED 31 million. Direct provider and delivery cost is AED 13 million before routing optimisation, producing AED 18 million of contribution before sales, administration, tax and capital investment.

The hypothetical platform reduces direct model spend by AED 4 million through workload selection, caching and batching. It adds AED 1.8 million of orchestration, evaluation and observability, AED 0.9 million of additional fallback cost and AED 0.7 million of customer support and service credits. Retained annual contribution improvement is therefore AED 0.6 million before any customer repricing or incremental sales. These assumptions demonstrate the calculation and do not represent market prices.

Management expects routing and evaluation to reduce frontier use for extraction and retrieval while preserving frontier access for complex analysis. The case assigns no value to unsupported quality improvement. It measures accepted outcomes by workload and applies complete cost. Benefits outside the evidence period remain in a separate scenario.

The portfolio also retains a migration budget. The enterprise tests an alternative frontier route and a second small-model stack on selected tasks. The cost reduces current contribution, while providing evidence about downside continuity and negotiating leverage. The investment committee can decide whether the resilience benefit supports that recurring cost.

16. Build the retained-contribution bridge

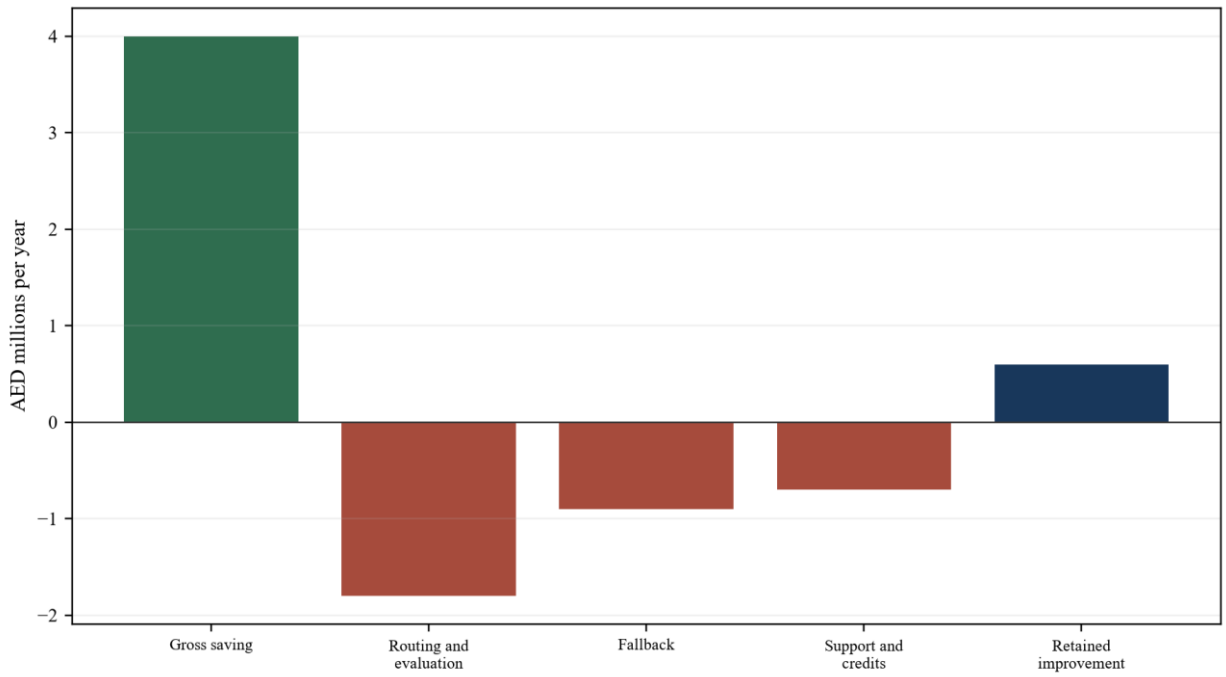
The bridge begins with customer revenue and subtracts direct model cost, routing and orchestration, retrieval and tools, evaluation, observability, human review, support, service credits and reserved-capacity leakage. It then isolates the change created by optimisation. Customer repricing, churn and incremental adoption belong in separate lines because they can absorb or amplify the technical saving.

Cash timing matters. Provider prepayments and capacity commitments can precede customer collections. Credits can arrive after the reporting period. The discounted-cash-flow model should reflect implementation gates, demand ramp, tax and working capital. Monthly cohort evidence is stronger than an annual average when traffic and model mix change quickly.

Economic ownership should match accounting presentation. Gross or net revenue, capitalised development, cloud commitments, leases and acquired intangibles can affect reported metrics without changing cash in the same period. The investment committee should receive both reported-accounting and cash views under approved policy.

The break-even volume is not static. Provider prices, hardware, model efficiency and labour cost change. A build route can require periodic retraining and infrastructure refresh. A partner route can benefit from competition and new models. The board should use scenarios and decision triggers instead of one permanent crossover point.

Figure 4. Hypothetical retained-contribution bridge



Values are illustrative management assumptions in AED millions.

17. Stress price quality fallback and churn

Stress should combine provider price increases, volume growth, model retirement, quality regression, capacity restriction, security events and migration delay. Correlated events matter because a provider change can increase cost while forcing engineering and reducing quality.

The downside should flow through customer behaviour and cash. Quality regression can increase review, complaints, refunds and churn. Capacity restriction can reduce service and collections. Migration can consume engineering that would otherwise deliver product growth. The model should fund these effects and test liquidity.

Recovery time is a decision variable. A portable architecture can still require months of evaluation, remediation and customer approval. The enterprise should model degraded service and dual-running during transition. Contractual notice periods, customer obligations and regulatory approvals affect the actual timeline.

Reverse stress testing asks which combination breaks the investment case or covenant. The board then identifies early-warning measures and actions. Trigger examples include accepted-outcome decline, concentration, cost per outcome, fallback, incidents, provider notices and migration-test failure.

Table 5. Hypothetical routing-platform sensitivities

Scenario	Accepted outcomes	Annual complete cost	Migration spend	Decision implication
Base portfolio	index 100	AED 18m	AED 2m	proceed with controls
API price increase	index 100	AED 23m	AED 2m	re-route eligible tasks
Small-model regression	index 91	AED 20m	AED 3m	retrain or partner
Frontier retirement	index 94	AED 22m	AED 6m	invoke transition plan
Combined downside	index 82	AED 29m	AED 8m	capital and terms protection

Every value is an illustrative management assumption.

18. Value the platform and its option set

Income valuation should use cash from accepted outcomes after complete cost, investment and tax. Market references require adjustment for revenue quality, rights, concentration, growth and operating maturity. Replacement cost can inform model, data, tooling and team value while missing distribution and customer relationships.

Option value can arise from rights to deploy, adapt or switch, access new workloads, or avoid dependency. The probability, investment and time required to exercise the option should be explicit. IFRS 3, IFRS 13, IAS 36 and IAS 38 govern relevant accounting and fair-value questions.[11][12][13][14]

Valuation should distinguish model capability from the operating system around it. Evaluation data, workflow integration, customer contracts, distribution and governance can remain valuable when a model changes. A model without rights, production reliability or customer access can have limited enterprise value despite strong benchmark performance.

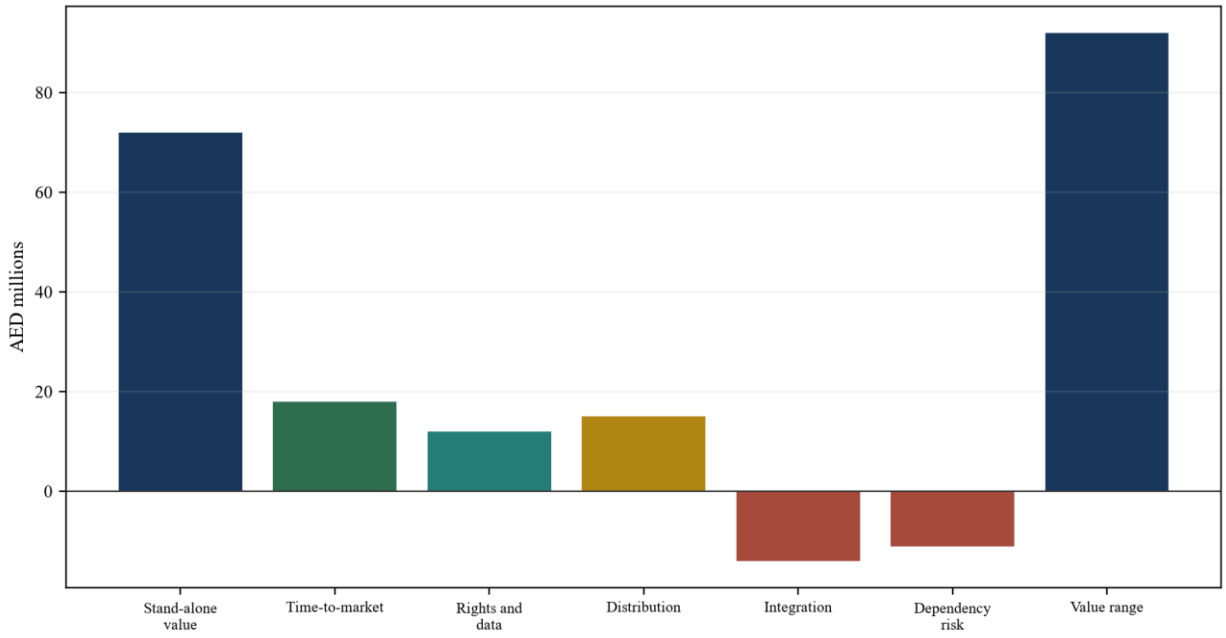
The valuation model should separate recurring platform revenue, provider pass-through and implementation services. Each stream has different margin, working-capital and retention characteristics. Reported growth driven by rising provider usage can increase revenue while leaving retained contribution unchanged. Cohort schedules should therefore reconcile revenue, accepted outcomes and contribution over time.

Scenario weights need governance. A base case can use observed renewal, route share, complete cost and price terms. Upside can reflect evidenced cross-sell or improved routing performance after a funded programme. Downside should include price convergence, direct customer contracting, model retirement, provider concentration and migration delay. Probabilities are management assumptions and should remain visible.

Buyer synergies require an owner and implementation cost. Distribution synergy requires eligible customers, sales motion, price and retention. Infrastructure synergy requires compatible architecture and capacity. Data synergy requires lawful rights and technical usefulness. Paying a seller for benefits created by the buyer's existing assets overstates stand-alone value.

Contingent consideration can protect against uncertain adoption, quality or contribution. Metrics should be auditable and resistant to manipulation. Revenue can reward discounting or uneconomic usage. Accepted contribution and retention can align more closely with durable value, subject to actual legal and accounting advice.

Figure 5. Illustrative routing-platform value bridge



Values are hypothetical management assumptions in AED millions.

19. Build a reproducible transaction diligence room

Diligence should permit a reviewer to select a workload cohort, reproduce evaluation, identify model and version, trace routing and fallback, reconcile usage and direct cost, and link accepted outcomes to customer or operating value. Evidence covers code, weights, licences, data provenance, evaluations, incidents, contracts, invoices, telemetry and financial policy.

Sampling should cover important customers, languages, geographies, models, versions, workloads and failure states. Management demonstrations are starting points. Reproduction by the buyer or an independent reviewer provides stronger evidence. Exceptions should remain visible rather than being removed from averages.

The data room should maintain version history. Evaluation results can become obsolete after model, prompt, tool or retrieval changes. Financial schedules should identify the same production versions and period. This alignment supports valuation, representations and post-close measurement.

Red flags include undocumented training data, non-transferable licences, unexplained benchmark differences, dependence on promotional credits, incomplete usage reconciliation, missing incident history, customers without AI terms and provider commitments that exceed contracted demand. Each red flag needs a price, term, remediation or decline treatment.

Table 6. Minimum routing-platform diligence checklist

Workstream	Minimum evidence	Decision output
Workloads	taxonomy, volume and consequence	route scope
Quality	held-out set and acceptance	capability envelope
Economics	contract-to-cash and cost ledger	contribution case

Workstream	Minimum evidence	Decision output
Rights	model, data, code and output rights	control and portability
Operations	routing, monitoring and incidents	resilience plan
Counterparty	service, finance and concentration	dependency treatment
Organisation	roles, retention and documentation	operating model
Valuation	counterfactual, synergies and downside	price range

The checklist requires transaction-specific adaptation.

20. Allocate risk through contracts and deal terms

Partner contracts can address price, capacity, model notice, eligible regions, service, security, data, audit, transition and termination. Acquisition agreements can address rights, privacy, security, customers and financial information through warranties, covenants, indemnities, holdbacks and contingent consideration as appropriate.

Earn-outs should use auditable accepted outcomes, contribution or retention. Raw token volume can reward uneconomic use. Integration and migration funding should sit beside purchase price and provider commitments.

Pricing terms should address input, output, caching, tools, fine-tuning, hosting, support, taxes and currency. Volume tiers should be tested against downside demand. Most-favoured or benchmark provisions require enforceable definitions. Capacity and service credits should reflect the consequence of failure without creating false comfort.

Model-change provisions should define notice, documentation, evaluation access and continuation. The enterprise may require a period of parallel availability. Data and logs should remain exportable. Assistance fees and transition capacity should be agreed before termination risk rises.

Acquisition closing conditions can address critical rights, customer consent, key staff and remediation. Escrow or holdback can support specific exposures. Warranty and indemnity insurance changes recourse and does not replace evidence. Transaction advisers should tailor terms to the actual structure.

21. Execute a 180-day value-protection programme

The first thirty days establish workload, route, rights, evaluation and cost baselines. Days 31 to 90 test candidate routes, negotiate terms, remediate material controls and implement observability. Days 91 to 180 scale only workloads that pass quality, control and contribution gates.

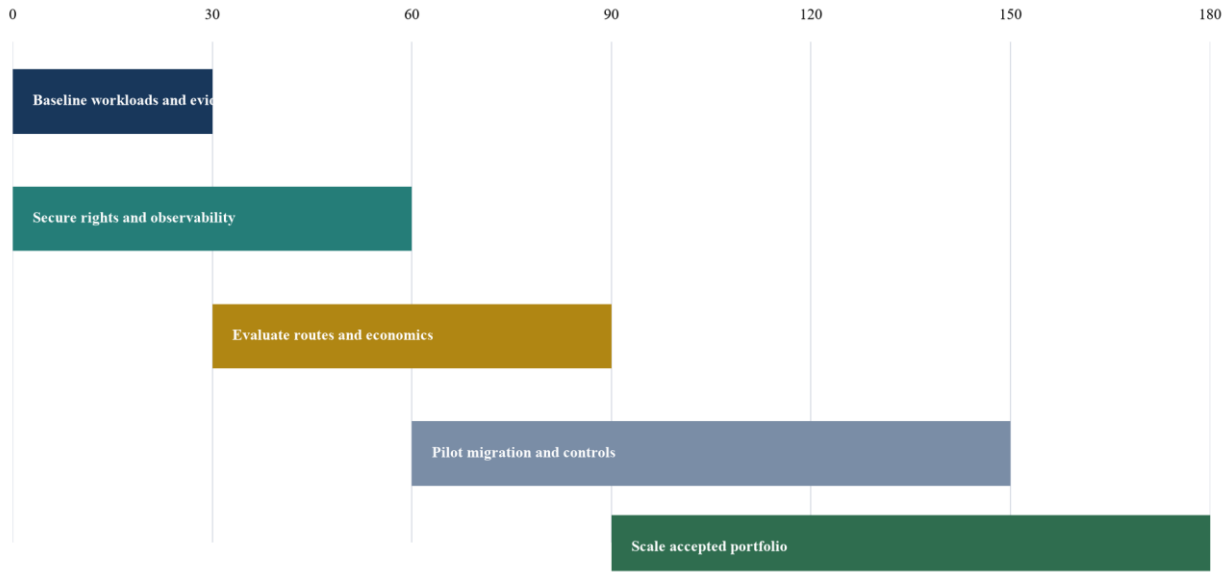
The programme needs one executive owner and a cross-functional control room. Weekly review covers delivery, quality, cost, incidents, rights and dependencies. Monthly investment review reconciles actual cash and accepted outcomes to the case. Decisions and exceptions are recorded with owners and expiry dates.

The first phase should protect continuity. The enterprise inventories production models, versions, prompts, tools, data, contracts and customer commitments. It freezes unsupported expansion and resolves critical access or security issues. Key staff and counterparties receive clear operating authority.

The second phase creates comparable evidence. Teams run the shared evaluation set, implement route telemetry, reconcile complete cost and test migration. Commercial and product teams align claims and pricing with the verified capability. Procurement negotiates options using observed demand and alternatives.

The third phase scales selectively. High-confidence workloads move to the economically preferred route. Complex or consequential cases retain appropriate review and fallback. The board receives realised contribution, investment, concentration and next capital decision.

Figure 6. 180-day routing-platform programme



The sequence is a general framework.

22. Make the board decision explicit

The memorandum should state the workload, selected route, counterfactual, evidence, complete economics, rights, dependencies, capital, downside and triggers. It should distinguish decisions that can be reversed cheaply from commitments that create structural dependency.

Approval can be staged by evidence. A limited partner pilot can validate demand. A small-model build can proceed through data, evaluation, production and scale gates. An acquisition can use exclusivity, confirmatory diligence, signing and integration gates. Each release of capital should correspond to a reduction in a material uncertainty.

The recommendation should state excluded value. Unsupported customer demand, untested migration, buyer-owned synergies and unapproved future workloads should remain outside the base case. Their potential can be shown separately with requirements to move into the case.

The board should receive a clear decline condition. Examples include rights that cannot transfer, quality that fails consequential cohorts, economics dependent on permanent credits, unacceptable data use, unmanageable concentration, or integration that the organisation cannot execute. Declining an attractive demonstration can protect capital when the evidence chain does not reach cash.

Table 7. Board decision matrix

Condition	Build	Buy	Partner
Stable high-volume bounded task	strong candidate	selective	benchmark alternative
Scarce strategic asset	possible	strong candidate	interim route
Rapidly changing frontier capability	selective	cautious	strong candidate
Strict deployment control	strong candidate	strong with rights	conditional

Condition	Build	Buy	Partner
Limited internal capability	staged	integration dependent	strong with governance
High switching requirement	architecture-led	rights dependent	multi-provider design

Thresholds require company-specific approval.

Conclusion

An AI optimisation platform earns durable value when it improves accepted outcomes, complete cost or control after provider prices converge. Routing access and temporary price spreads can support early growth. Transaction value depends on the evidence, workflow position, rights, distribution and switching capability that remain when those spreads narrow.

Durable value comes from accepted outcomes, complete economics, clear rights, operating capability and tested switching. Enterprises that own evaluation, routing, data and outcome evidence can change models while preserving the product and investment thesis.

Managed routing from model and cloud providers raises the standard for independent platforms. Cross-provider evidence, model-neutral evaluation, customer-specific policy, embedded distribution and tested portability can support differentiation. A platform should demonstrate these assets through reproducible diligence and customer retention.

The operating case should reconcile request-level routing to accepted outcomes, customer revenue, provider invoices, platform delivery costs and cash. Caching, batching and lower-cost models can improve the bridge. Retries, fallbacks, human review, support and service credits can consume the apparent saving. Cohort evidence reveals which workloads create contribution.

A durable routing strategy preserves the ability to learn. Every production route should produce comparable evidence about task, version, quality, latency, cost, escalation and customer response. That evidence supports provider negotiation, policy refinement, product design and decisions to stop workloads that do not create accepted value.

Finance and technology teams need a common unit. Cost per token supports invoice control. Accepted outcome per customer or process supports capital allocation. The reconciliation between them explains why a more expensive model can be economic for one workload and wasteful for another. It also prevents reported automation volume from being treated as value before quality, price and cash are demonstrated.

Rights and switching determine how much of the operating evidence remains with the enterprise. Evaluation sets, customer relationships, workflow design, routing logic and outcome history should be governed as strategic assets. Their ownership can preserve bargaining power even when model access is purchased. An acquisition that fails to secure these assets may transfer employees and code while leaving the economic control elsewhere.

The route also shapes financing. Predictable contracted contribution can support debt capacity, subject to lender diligence and downside coverage. Experimental workloads, concentrated providers and material migration exposure can require equity or staged capital. Acquisition debt should use cash after integration, minimum provider commitments, security remediation and transition reserves. Financing structure should reflect the same evidence used for the strategic decision.

Risk governance should remain proportionate to consequence. A low-value drafting aid can use broad experimentation within privacy and security policy. A model that triggers payments, commitments or safety actions needs stronger validation, approval, limits and monitoring. Applying one control standard to every workload either raises cost unnecessarily or leaves consequential activity under-controlled.

Routing economics should remain visible after deployment. Monthly management reporting can track route share, accepted outcomes, complete cost, provider notices, migration tests, concentration and retained contribution. Quarterly board review can compare realised economics with the approved case. An unexpected price change, model retirement, capability shift, incident or regulatory development can reopen the investment decision.

This approach converts a volatile provider market into a sequence of governed investment decisions. The board receives a reproducible method for comparing routes, protecting strategic assets, funding execution and changing course when evidence changes.

The method also creates a common language for transactions and operations. Sellers can present reproducible workload evidence. Buyers can compare the target with build and partnership alternatives. Lenders can test cash resilience and dependency. Management can convert the approved thesis into measurable execution. Clear definitions support accountability across each stage.

Frequently asked questions

Q: What makes an AI routing platform defensible? A: Defensibility can come from proprietary workload evidence, reproducible evaluation, policy and orchestration, embedded distribution, customer contracts and tested switching capability. Each claimed asset requires transaction evidence.

Q: Why is token-price arbitrage fragile? A: Providers can change prices, customers can negotiate directly and cloud platforms can bundle routing. Durable value requires contribution after complete delivery cost and plausible price convergence.

Q: What is the right economic denominator? A: Cost per accepted outcome connects usage with quality, retries, fallback, escalation and customer value. Token cost remains useful for invoice control and requires reconciliation to the outcome measure.

Q: How should caching and batching be valued? A: Measure eligible volume, actual hit or batch rate, provider-bill saving, service effects, operating cost and the share retained after customer repricing. Future rates should remain explicit scenarios.

Q: How should routing quality be compared? A: Use representative held-out tasks, defined acceptance rubrics, expert review where needed and versioned settings. Compare routed performance with direct eligible alternatives on quality, latency, complete cost and policy compliance.

Q: What is the most important switching control? A: A reproducible workload and evaluation layer supported by model-neutral interfaces, portable data, observable routing, tested alternatives and contractual transition rights.

Q: How does routing affect valuation? A: It changes contribution, working capital, rights, concentration, churn, operating risk and option value. Valuation should use complete economics, downside scenarios and separately evidenced buyer synergies.

Q: Which Matchpoint practice is relevant to this work? A: This research connects to Matchpoint Partners' strategy and execution and M&A advisory work, including AI economics, technology diligence, valuation, transaction structuring and integration.

References

[1] Amazon Web Services, Intelligent Prompt Routing in Amazon Bedrock, https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-routing.html

[2] Microsoft, How Model Router Works in Microsoft Foundry, https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/model-router-how-it-works

[3] Microsoft, Evaluate Model Router for Your Use Case, https://learn.microsoft.com/en-us/azure/foundry/openai/how-to/model-router-evaluate

[4] OpenAI, API Pricing, https://openai.com/api/pricing/

[5] Google Cloud, Model Armor Overview, https://cloud.google.com/security/products/model-armor

[6] National Institute of Standards and Technology, AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework

[7] National Institute of Standards and Technology, Generative AI Profile, https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

[8] United States Department of Justice and Federal Trade Commission, Merger Guidelines, https://www.justice.gov/atr/merger-guidelines

[9] European Commission, Merger Control, https://competition-policy.ec.europa.eu/mergers/merger-control_en

[10] United Kingdom Competition and Markets Authority, Merger Assessment Guidelines, https://www.gov.uk/government/publications/merger-assessment-guidelines

[11] IFRS Foundation, IFRS 3 Business Combinations, https://www.ifrs.org/issued-standards/list-of-standards/ifrs-3-business-combinations/

[12] IFRS Foundation, IFRS 13 Fair Value Measurement, https://www.ifrs.org/issued-standards/list-of-standards/ifrs-13-fair-value-measurement/

[13] IFRS Foundation, IAS 36 Impairment of Assets, https://www.ifrs.org/issued-standards/list-of-standards/ias-36-impairment-of-assets/

[14] IFRS Foundation, IAS 38 Intangible Assets, https://www.ifrs.org/issued-standards/list-of-standards/ias-38-intangible-assets/

[15] Google, Get Started with Gemma Models, https://ai.google.dev/gemma/docs/get_started

[16] Google, Deploy Gemma with Google Cloud, https://ai.google.dev/gemma/docs/integrations/google-cloud

[17] Amazon Web Services, Bedrock or SageMaker Decision Guide, https://docs.aws.amazon.com/pdfs/decision-guides/latest/bedrock-or-sagemaker/bedrock-or-sagemaker.pdf

[18] Microsoft, Foundry Models Sold Directly by Azure, https://learn.microsoft.com/en-us/azure/ai-foundry/foundry-models/concepts/models

[19] Microsoft, Model Catalog and Collections, https://learn.microsoft.com/en-us/azure/ai-foundry/how-to/model-catalog-overview

[20] Google Cloud, Vertex AI Model Garden, https://cloud.google.com/model-garden

[21] Amazon Web Services, Amazon Bedrock Pricing, https://aws.amazon.com/bedrock/pricing/

[22] Microsoft, Azure AI Foundry Pricing, https://azure.microsoft.com/en-us/pricing/details/ai-foundry/

[23] Google Cloud, Vertex AI Pricing, https://cloud.google.com/vertex-ai/generative-ai/pricing

[24] OpenAI, API Pricing, https://openai.com/api/pricing/

[25] Anthropic, Pricing, https://www.anthropic.com/pricing

[26] National Institute of Standards and Technology, Cybersecurity Framework 2.0, https://www.nist.gov/cyberframework

[27] National Institute of Standards and Technology, Secure Software Development Framework, https://csrc.nist.gov/Projects/ssdf

[28] National Institute of Standards and Technology, Privacy Framework, https://www.nist.gov/privacy-framework

[29] European Commission, AI Act, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[30] European Union, General Data Protection Regulation, https://eur-lex.europa.eu/eli/reg/2016/679/oj

[31] OWASP Foundation, Top 10 for Large Language Model Applications, https://owasp.org/www-project-top-10-for-large-language-model-applications/

[32] MITRE, ATLAS, https://atlas.mitre.org/

[33] International Organization for Standardization, ISO/IEC 42001, https://www.iso.org/standard/81230.html

[34] International Organization for Standardization, ISO/IEC 27001, https://www.iso.org/isoiec-27001-information-security.html

[35] MLCommons, AI Safety Working Group, https://mlcommons.org/working-groups/ai-safety/

[36] MLCommons, Inference Benchmarks, https://mlcommons.org/benchmarks/inference/

[37] United States Federal Trade Commission, Premerger Notification Program, https://www.ftc.gov/enforcement/premerger-notification-program

[38] IFRS Foundation, IFRS 15 Revenue from Contracts with Customers, https://www.ifrs.org/issued-standards/list-of-standards/ifrs-15-revenue-from-contracts-with-customers/

[39] International Valuation Standards Council, International Valuation Standards, https://www.ivsc.org/standards/

[40] Financial Accounting Standards Board, Accounting Standards Codification, https://asc.fasb.org/

[41] Organisation for Economic Co-operation and Development, AI Principles, https://oecd.ai/en/ai-principles

[42] Organisation for Economic Co-operation and Development, Due Diligence Guidance for Responsible Business Conduct, https://www.oecd.org/investment/due-diligence-guidance-for-responsible-business-conduct.htm

[43] United States Securities and Exchange Commission, Cybersecurity Disclosure Guidance, https://www.sec.gov/newsroom/press-releases/2023-139

[44] Cloud Security Alliance, Security Guidance, https://cloudsecurityalliance.org/research/guidance

[45] Linux Foundation, SPDX, https://spdx.dev/

[46] Open Source Initiative, Open Source AI Definition, https://opensource.org/ai/open-source-ai-definition

[47] Google, Responsible Generative AI Toolkit, https://ai.google.dev/responsible

[48] Amazon Web Services, Responsible AI, https://aws.amazon.com/ai/responsible-ai/

[49] Microsoft, Responsible AI Standard, https://www.microsoft.com/en-us/ai/principles-and-approach

[50] International Energy Agency, Energy and AI, https://www.iea.org/reports/energy-and-ai

About the Author

Chennakeshav (CK) is a corporate finance and investment banking executive with 25+ years of global experience in deal origination, structuring and execution across M&A, growth capital and corporate strategy. He has led value-creation mandates for founders, corporates and funds — bridging the boardroom view to hands-on execution and close.

His career spans Morgan Stanley, HSBC, Lloyds Banking Group, EWEC, ADQ portfolio companies and Emirates Growth Fund, across TMT, real estate, fintech, deeptech, cleantech, infrastructure and energy. He has partnered with C-suite leaders, private equity and venture funds, sovereign wealth funds and family offices to finance complex fund raises and scale-up ventures, and has led M&A due diligence, post-merger integration and business-transformation initiatives to create value.

At Matchpoint Partners he is Managing Partner, leading the firm's corporate finance, M&A and capital-raising practice. He holds an MBA from London Business School, an engineering degree from VTU and a Master of Laws (LLM, in progress) from UCL London.

An active start-up mentor, CK mentors at Techstars, DIFC FinTech Hive, Startup Grind, Founder Institute and IN5, serves as Entrepreneur Mentor in Residence (EMiR) at London Business School, and judges the Entrepreneurship World Cup.

https://www.linkedin.com/in/ckadya/

https://www.matchpoint-partners.com/team/ck-adya.html

This paper is part of a continuing series on the structure of private and alternative markets. The views expressed are the author's own. The paper is for information only, describes market structure in general terms, and does not constitute investment, legal, tax or regulatory advice or a recommendation in respect of any security, vehicle or counterparty.