

Pricing AI Augmented Services

Moving from Billable Hours to Outcome and Usage Economics

Authored by Chennakeshav Adya

Independent Researcher

September 2026

Abstract

Artificial intelligence is changing the economics of professional and technology-enabled services. Work that once required a visible number of professional hours can now combine human judgment, reusable intellectual property, software orchestration and variable model consumption. The resulting service may deliver faster decisions or higher throughput while using fewer labour hours. A fee based only on time can therefore weaken the connection between price and client value. Pure usage pricing creates a different problem: the units charged by an AI supplier rarely correspond to the outcome that a client buys.

This paper develops a pricing architecture for AI-augmented services. It separates three economic objects: the client value unit, the provider cost unit and the contractual billing unit. It then evaluates fixed fees, subscriptions, metered usage, transaction charges, gainsharing and hybrid structures against demand uncertainty, outcome measurability, cost volatility, assurance effort, liability and revenue quality. The framework includes a unit-cost curve, an outcome metric tree, a contract-risk matrix and a pricing governance process.

A hypothetical document-review service illustrates the method. The author assumes annual volume, labour effort, model consumption, exception rates, price levels and performance outcomes solely to demonstrate the calculations. These figures are analytical assumptions. They are neither observed company data nor forecasts. The analysis shows that an effective price does more than recover model tokens and professional time. It pays for reserved capacity, integration, evaluation, human review, security, accountability and the transfer of outcome risk. A durable commercial model uses units that the client can understand, the provider can measure and both parties can audit.

JEL Classification: D40, D46, L11, L21, M21, M41, O32, O33

Keywords: artificial intelligence, professional services, outcome pricing, usage pricing, unit economics, gross margin, service contracts, variable consideration, AI governance

1. Define the commercial decision

The central decision is how a provider should charge for a service when artificial intelligence changes delivery effort, cost variability and the distribution of risk. The answer affects sales, gross margin, cash conversion, revenue recognition, service design and client trust. It also determines whether productivity gains become provider profit, client savings, additional scope or a negotiated share of value.

Traditional professional services often use hours, days, team composition and reimbursable expenses as observable billing units. These units are imperfect proxies for value, yet they are easy to record and familiar to procurement. AI can reduce the time needed for research, drafting, classification, reconciliation or monitoring. It can also increase volume, create new assurance

work and introduce cloud consumption that moves independently of employee hours. A price structure that ignores those changes may reward inefficiency or expose the provider to unpriced demand.

Management should begin with a specific service and client decision. Useful questions include whether a due-diligence review should be priced per document, workstream, transaction or verified exception; whether a managed support service should charge per resolved case or reserve capacity; and whether an advisory engagement should retain a fixed fee while incorporating a measurable performance component. Each question has different evidence, attribution and risk.

The commercial case needs an explicit alternative. A client may perform the work internally, buy software, use a conventional outsourcer, retain a human-led adviser or defer the activity. The provider should identify the cost, speed, quality, control and switching implications of each realistic option. This gives the price discussion an economic reference without claiming that every benefit belongs to the provider.

The initial output is a pricing charter. It names the service boundary, buyer, user, value hypothesis, demand profile, cost drivers, delivery responsibilities, material risks, accounting review and renewal objective. It also states which elements remain uncertain and which evidence the pilot or first contract must generate.

2. Separate value, cost and billing units

Three different units sit beneath an AI-augmented service. The value unit describes what improves for the client, such as a case resolved, a day of cycle time removed, an error avoided, a qualified lead converted or a transaction completed. The cost unit describes what the provider consumes, such as professional time, model input and output, retrieval, storage, software licences, review effort and reserved infrastructure. The billing unit is the quantity written into the contract and invoice.

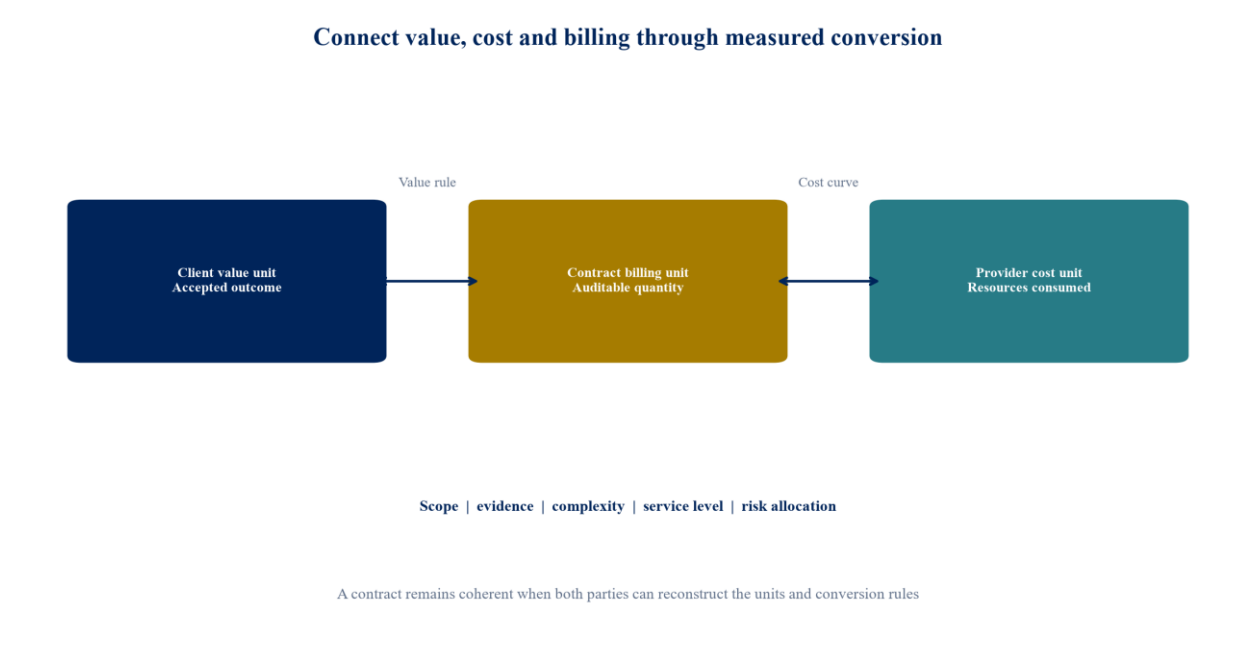
Weak pricing models treat the three units as if they were identical. A provider may pay for tokens while the client values an approved decision. A client may request an unlimited subscription while the provider bears highly variable inference and review cost. A gainshare may refer to savings that are influenced by demand, policy changes and client actions outside the provider's control. The commercial architecture must show how one unit converts into another.

The value unit should be observable, material and connected to the service. It can sit at several levels. Activity measures include documents classified or calls summarised. Output measures include reports accepted or cases closed. Operational outcomes include fewer days outstanding or higher first-time-right performance. Financial outcomes include verified cost avoided, revenue collected or loss reduced. Higher-level units can support stronger value alignment, but attribution and settlement become harder.

The cost unit needs the same discipline. Public cloud and model platforms distinguish input, output, cached input, batch processing, latency tiers, reserved throughput and related services [10][11][12][13][14]. The provider's cost also includes integration, evaluation, data preparation, human oversight, security, compliance, incident response, account management and intellectual property maintenance. A token-only cost model understates the service.

The billing unit should balance comprehensibility and control. A client can usually audit cases, transactions, sites, portfolios, users or service periods more easily than tokens. The provider can connect those units to internal cost through a measured conversion curve. Where neither party can rely on a single unit, a hybrid structure can charge for access and assurance, then add usage or outcome consideration within agreed limits.

Figure 1. The pricing architecture connects three different economic units



Client value, provider cost and contractual billing require explicit conversion rules and evidence.

Table 1. Unit hierarchy for AI augmented services

Layer	Example units	Primary owner	Main pricing question
Client value	Resolved case, accepted decision, cycle-time reduction, verified cash benefit	Client business owner	Which improvement is worth paying for and how will it be measured?
Service output	Completed review, approved report, validated exception, monitored asset	Joint service owners	What constitutes an accepted deliverable?
Billing	Subscription period, transaction, case, band of usage, verified outcome	Commercial and finance teams	Which quantity can both parties measure and audit?
Provider cost	Human review, model input and output, retrieval, storage, reserved capacity	Provider delivery and finance	How does cost change with volume, complexity and service level?
Risk capital	Warranty, remediation, indemnity, credit, performance fee at risk	Provider leadership and risk owners	What exposure is assumed and how is it funded or limited?

The preferred billing unit depends on the service, evidence and risk allocation.

3. Rebuild the fully loaded cost model

The cost model should follow the end-to-end service rather than the invoice from a foundation-model provider. Direct variable cost includes model inference, embeddings, retrieval, storage, external data, communications, workflow execution and human review. Semi-variable cost includes support, monitoring, evaluation and incident capacity that increases in steps. Fixed cost includes product development, integration, security, legal design, training, sales engineering and governance.

Model consumption should be measured by use case and stage. Input may contain instructions, context, retrieved records and conversation history. Output may include reasoning, generated text, tool calls or structured data. Caching can reduce repeated input cost. Batch processing can lower cost for delay-tolerant work. Reserved throughput can improve predictability for sustained loads while creating an under-utilisation risk. Public provider pages demonstrate that AI infrastructure already offers materially different charging bases and service tiers [10][11][13][14].

Human cost remains significant. Professionals define the problem, select data, design the workflow, review sensitive cases, communicate with the client and accept responsibility. The review rate may fall as the system improves, yet a higher-risk case mix can offset that benefit. The model should therefore estimate human minutes by case class and show the effect of false positives, false negatives, escalations and rework.

Assurance should have a visible cost pool. NIST's Generative AI Profile calls for governance, pre-deployment testing, content provenance and incident disclosure among its primary considerations [2]. Depending on the service, assurance can include evaluation sets, access controls, data protection, security testing, model-change review, audit records and client reporting. These activities protect the service promise and should be designed into price rather than treated as an exceptional overhead.

The unit-cost curve should show volume and complexity. Fixed cost is spread across more units as volume rises, while variable consumption grows. Capacity steps create discontinuities. Exception-heavy or long-context work can cost more even when case count is unchanged. A useful curve therefore reports cost per accepted unit by volume band, complexity band, service level and review intensity.

Cost governance requires versioning. A model change, supplier price change, new control, revised jurisdictional requirement or client data change can alter the curve. Finance should retain the assumptions, source date, observed usage and reconciliation to invoices. Sales should quote from an approved range with a defined validity period.

4. Identify the economic value pool

Value can arise from lower delivery cost, faster throughput, better quality, expanded capacity, improved decisions or reduced risk. The provider should identify the mechanism for each proposed benefit. A faster review may release professional time, accelerate a transaction, reduce a backlog or improve a customer response. Each path has different financial consequences.

Released hours do not automatically become cash. Finance should classify the conversion as reduced external spend, avoided hiring, redeployed capacity, increased revenue, faster working

capital, reduced expected loss or service improvement. The business owner should approve the baseline and conversion method. The provider can then decide whether the fee reflects the service output, a portion of verified benefit or both.

Evidence on generative AI productivity shows why baselines matter. A field study involving 5,179 customer-support agents reported an average productivity increase of 14 percent, with larger gains among novice and lower-skilled workers [5]. The result applies to that setting and intervention. OECD analysis also reports heterogeneous effects and warns that organisational change, skills and implementation affect realised gains [4][18]. A proposal should use client-specific evidence rather than transplanting a headline percentage.

Risk reduction needs an expected-loss formulation. A service may lower the probability of an error, improve detection or reduce severity through faster response. The value pool can be expressed as the change in probability multiplied by the financial consequence, adjusted for confidence and any control cost. Large but weakly evidenced risk claims should not support an immediate fee.

Client value can exceed provider cost without supporting unlimited gainshare. The client contributes data, process access, management attention, implementation resources and adoption. External conditions also influence results. A defensible value discussion allocates benefit according to contribution, risk and evidence. It leaves enough value with the client to sustain renewal.

The output is a value register. Each item records the metric, baseline period, source system, owner, counterfactual, calculation, confidence, attribution rule and settlement treatment. The register connects sales claims to later governance and reduces disputes when outcomes are reviewed.

5. Choose among six price archetypes

A fixed fee gives the client budget certainty and rewards the provider for efficient delivery. It works when scope, volume and complexity can be bounded. It can produce strong margins after automation, though it exposes the provider to demand, data-quality and exception risk. Scope definitions, assumptions and change control are therefore essential.

A subscription charges for access to a continuing capability, often with service, support and governance included. It suits recurring demand and ongoing model maintenance. The provider should define users, environments, data sources, service levels, included volume and fair-use boundaries. A subscription without usage visibility can conceal unprofitable consumption and weaken the renewal conversation.

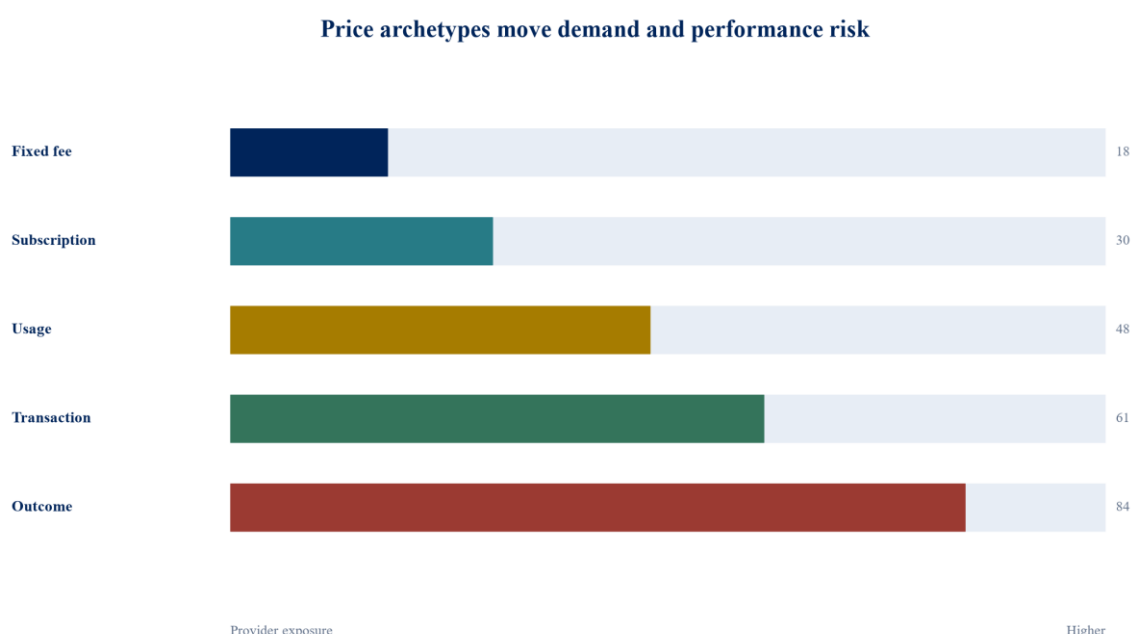
Metered usage charges for an auditable unit such as a case, transaction, page, minute, workflow or portfolio. It aligns price with activity and protects the provider when demand grows. It may discourage use if the client experiences every interaction as a marginal cost. Minimum commitments, volume bands and pooled allowances can create a better balance.

Transaction pricing charges for an economically meaningful event, such as a claim completed, invoice reconciled, loan reviewed or diligence item cleared. It moves closer to the client's operating outcome. The contract must define duplicates, reversals, partial completion, rejected cases and events handled by humans or other systems.

Outcome pricing links consideration to a verified operational or financial result. It can align incentives and support a premium when the provider accepts measurable performance risk. It also increases exposure to baseline disputes, attribution, data access and factors outside the service. Floors, caps, neutral measurement and agreed adjustments usually improve durability.

A hybrid structure combines elements. A fixed platform or assurance charge can fund readiness, reserved capacity and governance. Metered fees can cover variable consumption. An outcome component can reward verified performance. Hybrid pricing is often appropriate when AI cost, demand and value behave differently, provided the structure remains understandable.

Figure 2. Price models allocate demand and performance risk differently



Greater outcome exposure requires stronger measurement, attribution and contractual protection.

6. Design the hybrid price stack

The first layer is a mobilisation fee. It covers process discovery, integration, data preparation, controls, evaluation, training and launch. Milestones should correspond to accepted deliverables. Where substantial uncertainty remains, a paid discovery phase can produce the evidence needed to quote the operating service.

The second layer is a recurring capacity and assurance charge. It pays for service availability, account management, model and workflow maintenance, security, monitoring, audit evidence, reporting and a defined level of human oversight. This charge should reflect the standing capability the provider must maintain even when monthly usage is low.

The third layer is usage. The contract can include an allowance and charge for incremental bands. Bands reduce invoice noise and accommodate normal variability. The unit should be defined by a reliable event in the workflow rather than a provider's raw infrastructure meter, unless the client is explicitly procuring infrastructure consumption.

The fourth layer is service level. Faster turnaround, higher availability, dedicated throughput, additional jurisdictions or deeper human review can command a premium because they change cost and capacity. Credits should correspond to the service element that failed. They should also have clear measurement windows, exclusions and caps.

The fifth layer is outcome consideration. It can take the form of a bonus, gainshare, fee at risk or renewal adjustment. The formula should specify the baseline, measured population, attribution, data source, review period, exceptional events, floor, cap, settlement and dispute mechanism. A provider should avoid placing most economics on an outcome that the client alone controls.

The final layer addresses change. Model prices, legal obligations, data volumes and service scope can evolve. Indexation, price-review triggers and change control should state which changes are absorbed, which reopen commercial terms and which permit termination. This protects trust by making the adjustment mechanism visible before it is needed.

7. Construct an outcome metric tree

An outcome metric tree connects the service activity to an operational result and then to financial value. For a document-review service, the chain may run from documents ingested, to items accurately classified, to exceptions identified, to review cycle time, to transaction timing or professional capacity. Each link needs evidence and an owner.

The tree should include quality alongside speed. A model can increase throughput while creating rework, missed issues or inconsistent treatment. Useful measures include first-time-right rate, precision, recall, acceptance rate, severity-weighted errors, override rate and downstream correction. The selection depends on the decision and consequence.

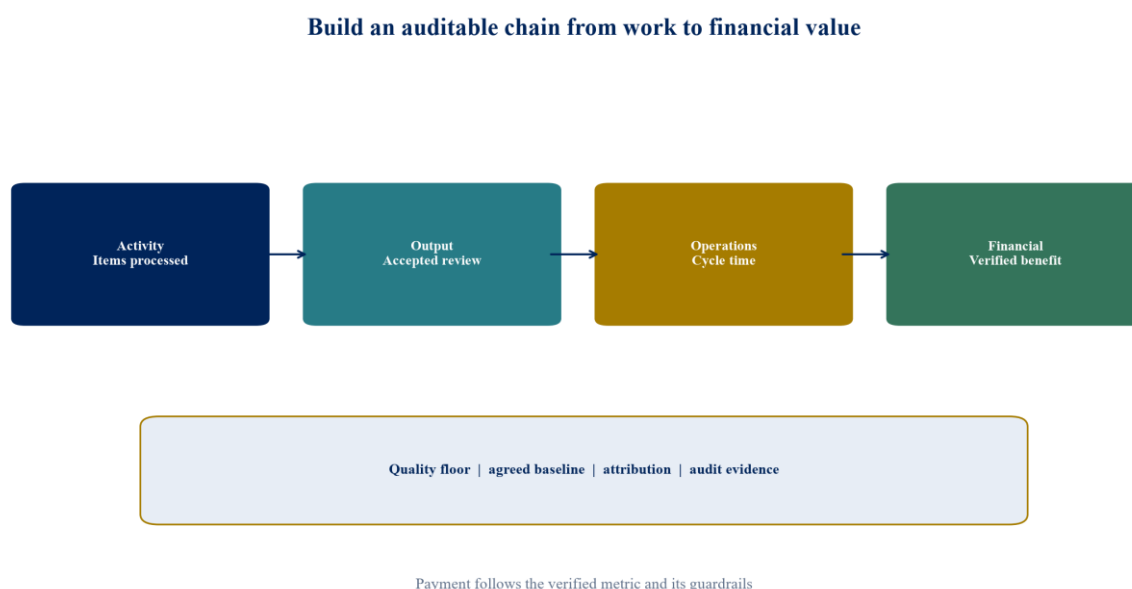
The baseline requires a defined period, population and process. A prior-year average may be distorted by mix or seasonality. A contemporaneous control group may be more reliable but harder to operate. A pre-agreed statistical or accounting method can adjust for volume, complexity and external shocks. The aim is a calculation that both parties can reproduce.

Attribution should follow control. If the provider controls classification and recommendation while the client controls approval and execution, the provider should be measured on the former unless the contract includes client dependencies. A joint outcome can still be priced, but the formula needs contribution rules and relief events.

Outcome measurement should resist gaming. A fee per resolved case can reward premature closure. A fee for cost reduction can reduce service quality. A revenue-share can favour short-term conversion over suitability. Guardrail metrics and audit sampling reduce this risk. Payment should depend on the primary outcome and the agreed quality floor.

The metric tree becomes part of service governance. Monthly reporting shows activity, output, outcome, quality and cost. Quarterly review tests whether the relationships still hold. A metric that no longer reflects value should be changed through the contract, not quietly reinterpreted.

Figure 3. Outcome pricing needs an auditable chain from activity to value



Guardrails prevent the primary outcome measure from rewarding lower quality or uncontrolled risk.

Table 2. Example outcome metric tree for an AI augmented review service

Level	Metric	Evidence	Pricing use
Activity	Items ingested and processed	Workflow event log	Capacity and usage reconciliation
Output	Reviews accepted without material rework	Client acceptance record	Billable completed unit
Quality	Severity-weighted error and override rate	Sample audit and correction log	Quality floor and service credit
Operations	Median cycle time by complexity band	Timestamped case data	Performance bonus or premium service level
Financial	Verified external spend avoided or capacity redeployed	Finance-approved baseline and ledger evidence	Capped gainshare
Risk	Expected loss reduced within defined cases	Approved probability and severity method	Deferred or contingent fee after validation

All measures require an agreed population, source and calculation method.

8. Protect gross margin through the cost curve

Gross margin should be modelled by client, service, volume band and scenario. Revenue is only one side of the calculation. Delivery cost changes with input size, output length, model routing, tool calls, retries, latency, review effort, exception mix and reserved capacity. Average cost can conceal an unprofitable tail.

The contract should identify measurable cost drivers. A document may be unusually long, multilingual, low quality, image-heavy or subject to additional confidentiality controls. A case may require several model calls and professional reviews. Complexity bands can price these differences while preserving a simple primary unit.

Unit cost often falls at first as fixed implementation and assurance cost is spread across more volume. It can rise when usage crosses a capacity threshold or when the next cases are more complex. Reserved throughput can lower volatility for predictable demand while creating a commitment. Pay-as-you-go capacity protects flexibility but may deliver a different latency or availability profile [11][13][14].

Management should set a contribution floor for standard work and a separate risk premium for uncertain work. Discounts should be tested against volume commitment, payment terms, data readiness, standardisation, reference rights and renewal value. A large stated volume without a minimum commitment cannot justify the same unit price as contracted volume.

Margin monitoring needs leading indicators. These include tokens or compute per accepted unit, review minutes, retries, exception rate, data defects, model mix, cloud discounts, service credits and support incidents. The finance system should reconcile operating telemetry to supplier invoices and client billing.

Price correction should follow agreed triggers. Examples include a sustained change in average context size, mandatory control work, a new data source, faster response requirement or supplier price movement. A collar can absorb ordinary variation and reopen terms only beyond a threshold. The mechanism should be symmetric where appropriate so clients share durable cost reductions.

9. Price assurance and accountability explicitly

AI-augmented services require assurance that conventional labour billing may hide inside senior review time. Evaluation, monitoring, access control, incident response and evidence retention are distinct service activities. Their cost persists even when model inference becomes cheaper.

NIST recommends continuous risk management across governance, mapping, measurement and management [1][2]. ISO/IEC 42001 specifies requirements for an AI management system [17]. These frameworks do not prescribe a price. They show why a provider must fund documented roles, objectives, risk assessment, monitoring and continual improvement when the service depends on AI.

The assurance charge should correspond to the actual service. A low-consequence internal drafting tool needs a different control package from a service that affects lending, employment, health, regulated advice or external commitments. The provider can define assurance tiers by consequence, data sensitivity, autonomy, audit requirement and recovery difficulty.

Client obligations also affect assurance. The client may need to provide lawful data, maintain access controls, approve configurations, appoint reviewers and report suspected incidents. The ICO's AI contract guidance recommends clear controller and processor roles, detailed instructions, documented decision boundaries, technical and organisational controls, subprocessor terms and in-life reviews [7]. These items belong in the operating model and the commercial scope.

Liability should be priced separately from routine delivery. A provider that accepts warranties, indemnities, service credits or remediation obligations bears expected and tail risk. The fee, cap, exclusions, insurance and operating controls should be considered together. A low price cannot compensate for an undefined or uncapped exposure.

Assurance evidence can support a premium when it is useful to the client. Evaluation reports, audit trails, control attestations, model-change notices and incident records reduce the client's verification burden. The provider should avoid selling generic trust language. It should specify the evidence delivered, its frequency, scope and limitations.

10. Allocate contract risk with precision

The contract should define the service before it defines the fee. The description should cover inputs, outputs, excluded decisions, human roles, systems, data, model or model family where relevant, service levels, jurisdictions and acceptance. Ambiguity in service scope quickly becomes ambiguity in price.

Usage definitions need event rules. The parties should agree whether retries, failed requests, duplicate documents, client cancellations, test traffic, human-only cases and reprocessing are chargeable. The provider should retain records that support the invoice while minimising unnecessary personal or confidential information.

Outcome terms need a measurement schedule. It should identify the baseline, source systems, calculation agent, client dependencies, relief events, audit rights, challenge window and settlement date. A concise worked example can expose disagreements before signature. The calculation should also address negative values, partial periods and termination.

Change control should cover service and technical change. A foundation model may be updated, deprecated or replaced. The provider may route work across models to manage cost or quality. The client needs to know which changes require notice, approval, re-evaluation or a price review. High-risk or regulated uses may impose additional duties under applicable law [6].

Data and intellectual-property terms should identify permitted use, training, retention, deletion, output ownership, third-party rights and client-provided materials. A pricing discount offered in exchange for broader data rights requires explicit consent and a clear valuation. Bundling the exchange into general terms can undermine trust.

Exit terms need portability and continuity. They should cover data export, records, model or workflow configuration where transferable, transition support, outstanding outcome measurement, prepaid capacity and retained evidence. The cost and time of exit affect the client's willingness to adopt and the provider's renewal quality.

Table 3. Contract risk matrix for AI augmented service pricing

Risk	Evidence before signature	Common allocation mechanism	Pricing response
Demand volume	Historical volume, forecast and minimum commitment	Included allowance, bands, minimum fee	Capacity charge and volume discount
Input complexity	Sample cases, language, size and exception mix	Complexity classes and change control	Differential unit price
Model and cloud cost	Usage profile, routing policy and supplier terms	Collar, index or review trigger	Cost pass-through or shared efficiency
Outcome attribution	Baseline, dependencies and source systems	Formula, relief events and audit rights	Capped gainshare or fee at risk
Quality failure	Evaluation method, severity and acceptance	Service credit, remediation and warranty	Assurance fee and risk premium

Risk	Evidence before signature	Common allocation mechanism	Pricing response
Data and regulatory change	Roles, jurisdiction and processing map	Compliance obligations and change order	Repricing for material control changes
Liability	Consequence, cap, exclusions and insurance	Indemnity, limitation and claims process	Explicit risk capital charge
Termination	Portability, records and transition plan	Notice, exit assistance and settlement	Mobilisation recovery and exit fee

Each allocation should match operational control and applicable law.

11. Address revenue quality and accounting

Commercial design and accounting should be reviewed together. IFRS 15 requires an entity to identify the contract, identify distinct performance obligations, determine the transaction price, allocate it to performance obligations and recognise revenue when each obligation is satisfied [3]. An AI-augmented service may combine implementation, access, support, usage and outcome consideration.

The contract should describe what the customer receives. Implementation may create an asset controlled by the client, configure the provider's platform or prepare the ongoing service. Access and managed service may be a combined promise or separate obligations depending on the facts. Finance should assess the arrangement rather than copying the billing schedule into the accounting conclusion.

Variable consideration requires attention. Usage fees, performance bonuses, service credits, refunds and gainshare can change the transaction price. IFRS 15 includes a constraint intended to limit revenue recognised when a significant reversal is possible [3]. The applicable accounting depends on the detailed terms and requires professional judgment.

Principal-versus-agent questions can also arise when the provider resells model, cloud or data services. The IFRS Interpretations Committee has emphasised that the analysis depends on whether the entity controls the specified good or service before transfer and on the contractual facts [19]. Gross billing does not by itself establish gross revenue presentation.

Revenue quality includes more than recognition. Management should monitor recurring committed revenue, usage volatility, concentration, implementation dependence, gross retention, expansion, service credits, collection, contract assets and outcome-fee uncertainty. A high headline contract value can have weak cash and margin characteristics.

The deal approval paper should contain the accounting view, cash schedule and downside. It should state which fees are non-refundable, when invoices arise, what evidence supports acceptance, which amounts vary and what could delay collection. This gives commercial leaders a complete view before concessions are granted.

12. Prevent deceptive or unsupported value claims

The commercial narrative should describe tested capability and measured outcomes. Claims that an AI service replaces a professional, guarantees savings or achieves a stated accuracy require evidence for the represented context. The US Federal Trade Commission has taken action against companies that made unsupported AI claims, including claims about substituting for professional services [20].

Sales material should identify the evaluation population, metric, date, model and limitations. A result achieved on curated internal cases should not be presented as a client-wide outcome. Where the result is a scenario, the document should say so in ordinary reader-facing language and explain the assumptions.

Outcome pricing can create pressure to overstate attribution. Governance should separate proposal approval, measurement and fee settlement. The client should have access to the agreed evidence, and the provider should retain adverse results. A transparent downside strengthens the credibility of the commercial case.

Dynamic or personalised pricing can also raise consumer and competition concerns in some contexts. The UK Competition and Markets Authority advises businesses to consider transparency, fairness and consumer-law obligations when using dynamic pricing [9]. Professional and business-to-business services have different facts, yet the principle remains useful: clients should understand the units, triggers and total-cost implications.

Marketing and contract language should use the same definitions. Terms such as autonomous, guaranteed, compliant, secure and human-reviewed can imply specific capabilities. The delivery team should confirm that the operating design supports them. Material changes should trigger a review of both the service and the representations.

The provider should maintain a claims register. It records each quantitative or comparative claim, supporting evidence, owner, approved context and expiry. The register protects proposals, webpages and renewal material from carrying obsolete performance assertions.

13. Manage supplier and model concentration

The unit economics of an AI service depend on a supply chain that can include cloud infrastructure, foundation models, data providers, orchestration tools, security services and business applications. Concentration affects cost, continuity, negotiating leverage and the ability to honour client commitments.

Competition authorities have examined partnerships and integration across the foundation-model value chain. The UK CMA's update identified risks to fair, open and effective competition and described the importance of access, diversity, choice, transparency and accountability [8]. The US FTC has also used its information-gathering authority to examine generative-AI investments and partnerships [21]. These inquiries do not determine an individual contract, but they highlight the need to understand dependencies.

A provider should map which supplier controls the model, hosting, identity, retrieval, data and workflow. It should record prices, volume commitments, service levels, data terms, change rights, termination, export and alternatives. A nominally multi-model service may still depend on one cloud, one embedding model or one proprietary connector.

Portability has a real cost. Changing a model can alter quality, latency, tokenisation, prompt design, tool behaviour, safety settings and evaluation thresholds. The provider should estimate migration and revalidation effort rather than assuming an instant switch. A tested fallback can support a lower concentration premium.

Client contracts should align with upstream terms. The provider should avoid promising service levels, retention, geographic processing, audit rights or liability that the supply chain cannot

support. Where there is a gap, the provider must control it operationally, price it or narrow the commitment.

Supplier cost reductions can create a sharing question. A transparent hybrid may pass through indexed consumption while retaining a fixed price for intellectual property and assurance. Another structure can fix the client price for a period and let the provider earn efficiency. The chosen model should be stated clearly.

14. Use a pricing council and deal desk

Pricing authority should sit across commercial, delivery, product, finance, risk and legal functions. Sales understands buyer context and competition. Delivery understands effort and exceptions. Product and engineering understand model and platform cost. Finance tests margin, cash and accounting. Risk and legal assess obligations and exposure.

The council should approve service archetypes, cost curves, target margins, discount corridors, outcome formulas, assurance tiers and contract clauses. The deal desk applies these standards to individual proposals. Exceptions require a named approver, quantified impact and expiry.

The approval pack should show client value, alternatives, scope, demand, unit-cost curve, price waterfall, margin scenarios, implementation funding, service levels, liability, data rights, accounting and cash. It should also show which assumptions will be tested after launch. This turns pricing into an operating decision rather than a sales-only negotiation.

Portfolio reporting should compare quoted and realised economics. Useful fields include contracted volume, actual units, revenue per unit, cost per accepted unit, review effort, gross margin, service credits, collection, renewal, outcome fees and exceptions. The comparison reveals where the price model or delivery design needs revision.

The council should protect simplicity. A contract with many units, thresholds and exclusions may price every risk while becoming difficult to sell and administer. The aim is a small number of economically meaningful components, supported by a richer internal model.

Renewal should be evidence-led. The provider can show changes in usage, quality, client outcomes, cost and assurance. Durable productivity gains may support a new fixed fee, lower unit rate with greater commitment, expanded scope or a revised gainshare. The renewal discussion should begin before the measurement period ends.

15. Apply the framework to a hypothetical review service

Consider a hypothetical service that reviews commercial documents, extracts obligations, identifies exceptions and prepares a structured review pack. The client currently processes 120,000 documents a year. The author assumes an average of 22 human minutes per document, an internal fully loaded labour cost of USD 72 per hour and annual external specialist spend of USD 900,000. These figures are analytical assumptions.

The proposed workflow uses document processing, retrieval, a generative model, rules and professional review. Management assumes that 55 percent of documents are standard, 30 percent require judgment and 15 percent contain material exceptions. Every material exception and a sample of standard work receive human review. No legal or commercial commitment is made by the system.

The cost model assumes annual fixed development and assurance cost of USD 480,000. Variable infrastructure, software and data cost averages USD 0.68 per document in the base case. Human review averages 6.5 minutes per document at USD 78 per hour. Support and monitoring add USD 0.24 per document. These values are assumptions and should not be treated as market prices.

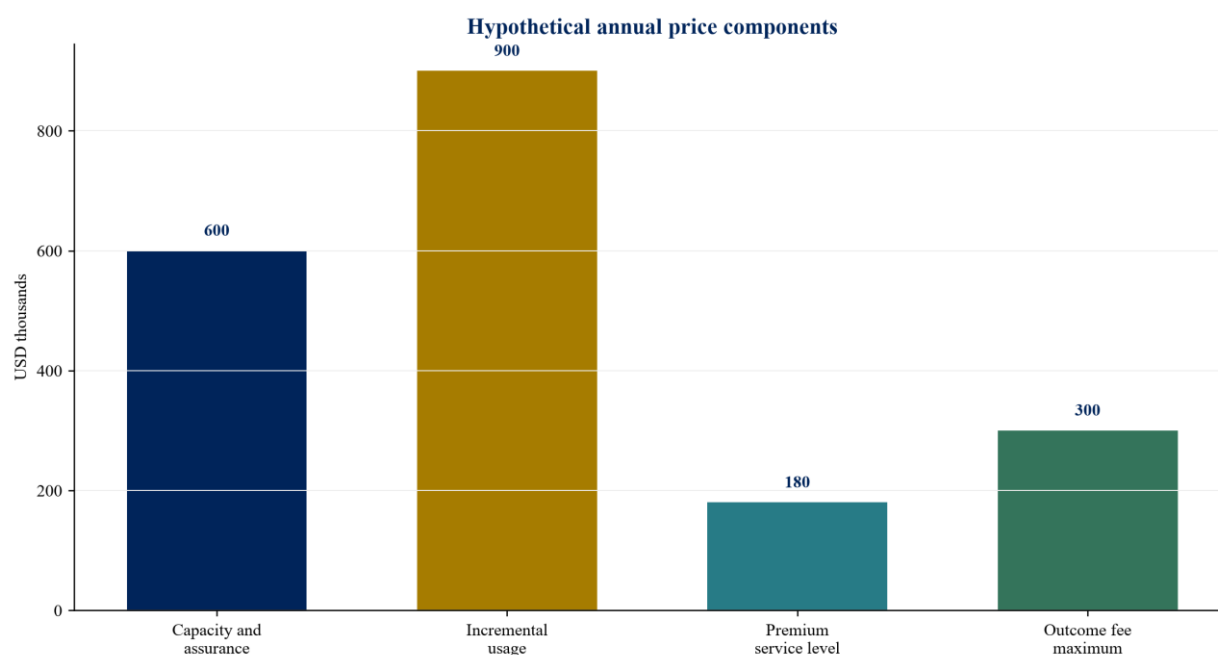
At 120,000 documents, the assumed direct annual cost is USD 1.584 million: USD 480,000 fixed, USD 81,600 infrastructure, USD 1.014 million review and USD 28,800 support. This calculation excludes sales, corporate overhead and risk capital. The direct cost per document is USD 13.20. At lower volume, fixed cost produces a higher unit cost; at higher volume, capacity and review steps may change the curve.

The pricing architecture charges a USD 600,000 annual capacity and assurance fee, includes 60,000 documents, and charges USD 15 per additional accepted document. A premium two-day service level costs USD 180,000. A capped outcome fee equals 12 percent of finance-verified external spend avoided above an agreed baseline, subject to quality floors and a USD 300,000 annual cap.

At the assumed base volume, recurring revenue before outcome consideration is USD 1.68 million. The direct cost estimate leaves USD 96,000 before sales and corporate overhead, an insufficient margin for the risk and investment. The provider can improve the structure through a higher base fee, higher included-volume discipline, lower review effort supported by evidence, a differentiated price for complex cases or a minimum outcome payment. This illustrates why a superficially attractive per-document price can fail when assurance and review are fully loaded.

The pilot should test consumption, context size, acceptance, review time, errors, exceptions, cycle time and client dependencies. The first commercial period can include a reopener after a representative volume. Any gainshare should settle after finance validates the baseline and quality guardrails. The client receives a transparent unit definition and audit report; the provider retains protection against unbounded complexity.

Figure 4. The hypothetical price waterfall separates capacity usage service level and outcome consideration



Illustrative author assumptions. Amounts are neither observed results nor forecasts.

Values illustrate the framework and do not represent observed client economics or a forecast.

Table 4. Commercial gate for the hypothetical service

Gate	Required evidence	Pricing consequence	Decision
Scope	Document types, languages, decisions and exclusions	Defines the chargeable unit and change process	Approve bounded service
Cost	Consumption, review time, support and capacity	Sets floor price and volume bands	Approve cost curve
Quality	Acceptance, severity-weighted error and audit sample	Sets quality floor, credits and review tier	Approve operating promise
Value	Cycle time, capacity and verified external spend	Sets outcome formula and cap	Approve contingent consideration
Risk	Data, regulation, liability and recovery	Sets assurance tier and risk premium	Approve contract allocation
Accounting and cash	Performance obligations, variable consideration and invoices	Sets billing and revenue treatment	Approve deal economics

The parties should validate each item with observed pilot evidence before committing to scale.

16. Negotiate from evidence

The provider should enter negotiation with a clear walk-away position. It needs the minimum price at realistic volume and complexity, acceptable liability, required data access, payment terms and change-control protections. These limits should reflect the downside scenario, not the most efficient pilot result.

Procurement may request rate-card comparability. The provider can show a transparent construction without reducing the service to hours. It can identify mobilisation, recurring

assurance, included capacity, incremental units, premium service levels and outcome consideration. A shadow day-rate equivalent can support comparison while the contract uses more suitable units.

Clients often ask for unlimited use. The provider can respond with a pooled allowance, fair-use limit, concurrency boundary or minimum commitment. The commercial goal is to give the client operational freedom while preserving capacity planning and margin. Usage visibility should remain available even when incremental invoices are not generated.

Discounts should buy something measurable. A longer term, prepaid amount, minimum volume, standard integration, reference permission, limited customisation or narrower service level can justify a concession. A discount given solely to reach signature becomes the new reference point at renewal.

Outcome pricing should be offered when the provider can influence the result and access the evidence. A client that cannot provide baseline data or timely operating actions may be better served by an output-based fee during the first period. The parties can add outcome consideration after measurement matures.

Negotiation records should preserve the assumptions behind the final price. Delivery and finance need to know which obligations were exchanged for concessions. The record supports implementation, margin review and renewal.

17. Implement a ninety day pricing programme

During the first 30 days, management should select one service, map its workflow and define the buyer's decision. The team should identify value, cost and candidate billing units; collect historical volume and effort; inventory model and cloud cost; classify risks; and review existing contract and accounting practices.

During days 31 to 60, the team should build the unit-cost curve, value register and outcome metric tree. It should test fixed, subscription, usage, transaction, outcome and hybrid scenarios. Finance should model margin, cash and revenue characteristics. Legal and risk should draft the contract matrix and assurance tiers.

During days 61 to 90, the provider should price a controlled pilot or apply the architecture to a renewal. The operating system should capture chargeable events, consumption, review effort, quality, service levels and outcome evidence. The invoice should be reconstructed from the same records used for service governance.

The day-90 review should compare quoted and observed economics. Management should decide whether to retain the unit, change volume bands, alter review rules, adjust assurance, narrow scope or stop the offer. Any improvement should be documented before broader sales use.

The provider can then create a service catalogue. Each offer has an approved scope, price stack, cost curve, evidence package, contract schedule, discount corridor and owner. Custom work can still be sold, but it starts from a governed commercial baseline.

The programme should train sales and delivery together. Sales needs to explain the unit and value without unsupported claims. Delivery needs to understand how operating choices affect margin and client outcomes. Finance needs timely telemetry rather than a retrospective cost allocation.

18. Recognise limitations and open questions

AI service economics change quickly. Public model prices, performance, context limits and service tiers can change during a contract. Published price lists may not reflect negotiated enterprise terms or the provider's complete architecture. Every cost model needs a source date and a contractual approach to material change.

Productivity evidence may not transfer across professions, tasks or organisations. A controlled study of customer support or writing tasks cannot establish the result for diligence, engineering, health, finance or regulated advice. The client service should generate its own baseline and adverse-case evidence.

Outcome attribution remains difficult when multiple interventions occur. Changes in demand, staff, policy, market conditions and client execution can influence the metric. A complex adjustment model can create the appearance of precision without producing agreement. The parties may prefer a narrower output measure until data quality improves.

Pricing can change behaviour. Per-unit charges may suppress useful adoption. Outcome fees may encourage metric gaming. Subscriptions may encourage excess use. Fixed fees may reduce incentives to expand quality. Governance should monitor behavioural effects and revise the structure when the service objective is distorted.

Legal, accounting, tax, competition and data-protection treatment depends on the facts and jurisdiction. This paper provides a commercial decision framework and cannot determine an individual contract. Organisations should obtain current professional advice before relying on a specific structure.

The hypothetical case uses assumed values. It demonstrates the calculation and the information required for a decision. It does not show achievable market price, cost, savings, margin or client demand.

19. Conclusion

AI-augmented services require a clearer connection between what the client values, what the provider consumes and what the contract bills. Hours remain useful for some discovery and bespoke judgment. Subscriptions, usage units, transactions and outcomes can support other parts of the service. The selected structure should follow the economics and evidence.

A durable price funds the complete operating capability. It covers integration, model and cloud consumption, human review, evaluation, security, governance, support, accountability and risk. It also leaves the client with a credible share of the benefit.

Hybrid structures can manage different forms of uncertainty. A mobilisation fee funds setup. A recurring charge supports capacity and assurance. Usage bands recover variable cost. Service-level premiums reflect operating commitments. A capped outcome component rewards verified value. Each component needs a clear definition and source of evidence.

The commercial system should continue after signature. Finance, delivery, product, sales, risk and legal should compare quoted and observed economics, review claims, govern exceptions and update the cost curve. Renewal then becomes a reasoned discussion of usage, quality, value, risk and price.

The board and executive team can test the architecture with a concise set of questions. What does the client buy? Which unit proves delivery? Which costs vary? Who controls the outcome? Which risks are accepted? How is the invoice reconstructed? How can either party exit? A price is ready for scale when those answers are measurable, contractual and financially coherent.

References

1. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework AI RMF 1.0. 2023. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>
2. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
3. 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/>
4. OECD. The Effects of Generative AI on Productivity Innovation and Entrepreneurship. OECD Artificial Intelligence Papers. 2025. <https://doi.org/10.1787/b21df222-en>
5. Brynjolfsson, E., Li, D. and Raymond, L. R. Generative AI at Work. Quarterly Journal of Economics 140(2), 2025. <https://www.nber.org/papers/w31161>
6. European Union. Regulation EU 2024/1689 laying down harmonised rules on artificial intelligence. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
7. Information Commissioner's Office. Artificial intelligence contracts and third parties. <https://ico.org.uk/for-organisations/advice-and-services/audits/data-protection-audit-framework/toolkits/artificial-intelligence/contracts-and-third-parties/>
8. UK Competition and Markets Authority. AI Foundation Models Update Paper and Technical Update Report. 2024. <https://www.gov.uk/government/publications/ai-foundation-models-update-paper>
9. UK Competition and Markets Authority. Dynamic Pricing Project Update. 2025. <https://www.gov.uk/government/publications/dynamic-pricing-project-update>
10. OpenAI. API pricing. Accessed September 2026. <https://openai.com/api/pricing/>
11. Amazon Web Services. Amazon Bedrock Pricing. Accessed September 2026. <https://aws.amazon.com/bedrock/pricing/>
12. Amazon Web Services. Understanding Amazon Bedrock Cost and Usage Report Data. <https://docs.aws.amazon.com/bedrock/latest/userguide/cost-mgmt-understanding-cur-data.html>
13. Microsoft Azure. Azure OpenAI Service Pricing. Accessed September 2026. <https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/>
14. Google Cloud. Generative AI on Vertex AI Pricing and Throughput Quota. Accessed September 2026. <https://cloud.google.com/vertex-ai/generative-ai/pricing>
15. Anthropic. Claude API Pricing. Accessed September 2026. <https://www.anthropic.com/pricing>
16. FinOps Foundation. FinOps Framework 2025 and Unit Economics Terminology. <https://www.finops.org/framework/>
17. International Organization for Standardization. ISO IEC 42001 2023 Information technology Artificial intelligence Management system. <https://www.iso.org/standard/81230.html>
18. OECD. Generative AI and the SME Workforce New Survey Evidence. 2025. <https://doi.org/10.1787/2d08b99d-en>
19. IFRS Interpretations Committee. Principal versus Agent Software Reseller IFRS 15. 2021. <https://www.ifrs.org/news-and-events/updates/ifric/2021/ifric-update-november-2021/>
20. US Federal Trade Commission. FTC Announces Crackdown on Deceptive AI Claims and Schemes. 2024. <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes>
21. US Federal Trade Commission. FTC Launches Inquiry into Generative AI Investments and Partnerships. 2024. <https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships>
22. UK Department for Science Innovation and Technology. AI Management Essentials Tool. Updated 2026. <https://www.gov.uk/government/consultations/ai-management-essentials-tool/ai-management-essentials-tool-accessible>
23. Information Commissioner's Office. Data protection by design and by default. Updated 2026. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/data-protection-by-design-and-by-default/>
24. UK Competition and Markets Authority. Joint Statement on Competition in Generative AI Foundation Models and AI Products. 2024. <https://www.gov.uk/government/publications/joint-statement-on-competition-in-generative-ai-foundation-models-and-ai-products>

About the Author

Authored by Chennakeshav Adya

Independent Researcher

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.