

When Reasoning Gets Expensive: SaaS Gross Margins and the Valuation Reset

A Transaction Framework for Pricing Test-Time Compute, Quality, Customer Value and Durable Contribution

Authored by Chennakeshav Adya

Independent Researcher

September 2026

Abstract

Reasoning models can spend additional inference effort on difficult tasks. OpenAI reported that performance in its o1 research improved with more test-time compute. Current provider documentation also exposes reasoning controls, thinking levels or budgets, and pricing rules that can charge generated reasoning or thinking tokens as output. These capabilities allow a software product to buy more computation for a customer task when management expects a better answer to create enough value.

This changes a central assumption behind conventional software-as-a-service economics. A seat can generate requests with materially different production cost. A short classification, a multi-document investigation and an agentic workflow may share the same user interface while consuming different model, retrieval, tool, evaluation and human-review resources. The cost is partly endogenous because product policy decides which model to invoke, how much reasoning effort to allow, how many candidates or retries to generate, which tools to call and when to escalate.

This paper develops a reasoning-contribution framework for boards, investors, lenders and operators. The framework links task value, compute budget, quality uplift, customer acceptance, retention, expansion, complete direct cost and cash contribution. It separates ordinary automation from premium reasoning, designs seat, credit, overage and outcome packages, and sets out telemetry, contracting, accounting, governance, diligence and valuation requirements.

A hypothetical enterprise software case demonstrates the method. Every volume, price, token, cost, quality, retention, expansion and valuation input in that case is a management assumption created solely to show the calculation. The case is not a forecast, market quotation or representation of any named company. Six figures show the reasoning-value chain, task portfolio, budget router, gross-margin bridge, cohort scorecard and 180-day roadmap. Seven tables provide the task taxonomy, unit ledger, pricing architecture, hypothetical operating case, sensitivity matrix, diligence checklist and valuation bridge.

The resulting valuation reset is analytical rather than automatically negative. Additional reasoning can support higher value, retention and price. It can also create uncontrolled cost, slow service and weak forecastability. Transaction value depends on evidence that the product allocates reasoning economically and retains a defensible share of the customer value created.

AI, privacy, intellectual property, consumer, sector, cyber-security, export, tax, accounting, valuation, financing and investment decisions require current advice from qualified professionals in the relevant jurisdictions. Provider models, prices, terms, availability and regulation can change. This paper provides general information for professional audiences and does not provide legal, regulatory, tax, accounting, valuation, credit, technical or investment advice.

JEL Classification: G24, L11, L21, L86, M13, O32

Keywords: reasoning models, test-time compute, SaaS, gross margin, pricing, retention, valuation, AI applications, contribution margin

1. Treat reasoning as a production decision

Traditional SaaS analysis often begins with recurring revenue, gross retention, net revenue retention, gross margin, sales efficiency and cash generation. The model assumes that software can serve an additional user at a small marginal cost. Cloud hosting, support and third-party services remain relevant, yet the economic story usually rests on substantial operating leverage.

Reasoning software adds another production decision. The product can allocate more inference effort to a task before returning an answer. A harder legal review, engineering diagnosis, financial analysis or research request may receive a larger reasoning budget, more tool calls, more candidate paths or additional verification. A routine extraction or classification can receive a smaller budget. Management therefore controls part of the direct cost at the moment of service.

The decision should begin with customer value. A product should not spend ten times more because a model can think longer. It should spend more when the expected improvement in accepted outcome, avoided loss, time saved, revenue created or risk reduced exceeds the complete incremental cost and remains inside the product's safety and service envelope.

OpenAI's o1 research described performance improvements with additional test-time compute.[1] Google documents thinking controls and states that thinking tokens can be included in output pricing.[2] Anthropic documents extended thinking and explains the treatment of thinking tokens within its service.[3] These provider mechanisms differ, and the applicable documentation and contract at the transaction date remain authoritative. Their common commercial implication is that reasoning intensity can be an explicit product variable.

Figure 1. The reasoning-value chain



Evidence gap at any link converts the claimed benefit into a scenario assumption

Author framework. Each link requires observable evidence before higher compute is credited with higher enterprise value.

2. Define the task before measuring the token

A defensible unit model begins with a task taxonomy. “AI request” is too broad. The same user can ask for a two-line rewrite, a contract comparison, a multi-source market study or an agentic sequence that changes a business system. Each task has a different value, risk, latency tolerance, evidence requirement and cost envelope.

The taxonomy should be stable enough for finance and commercial teams to use. Product events should record the task class, customer, package, model route, reasoning setting, tool calls, attempts, evaluation result, latency, direct review and final acceptance. The taxonomy can evolve through controlled versioning. Silent changes weaken cohort comparisons and transaction diligence.

Table 1. Reasoning-task taxonomy

Task class	Customer promise	Typical consequence	Reasoning policy	Commercial treatment
Routine assist	draft, classify, summarise or extract	low and reversible	bounded model and low effort	included within seat allowance
Professional analysis	compare evidence and explain a recommendation	material decision support	higher effort, retrieval and cited sources	reasoning credits or premium tier
Controlled decision support	analyse a consequential regulated or financial question	high consequence with human accountability	approved route, verification and mandatory review	premium workflow or outcome fee
Agentic execution	plan and perform permitted multi-step actions	operational side effects	tool limits, confirmation, rollback and audit trail	metered run plus service commitment
Exceptional investigation	ambiguous, cross-domain or high-volume problem	high value and uncertain effort	explicit budget, checkpoint and stop rule	scoped project or separately approved overage

The taxonomy should be versioned and reconciled to product, finance and commercial systems.

Task classification should happen before routing. A customer label alone is insufficient. One enterprise account can contain routine and consequential work. Contracting should define which tasks are included, which require premium credits and which remain outside the automated service.

3. Measure incremental quality, not theatrical effort

More reasoning can improve some tasks. It can also add latency, verbosity and cost without changing customer acceptance. The product therefore needs a quality-response curve by task class.

The curve compares approved routes and effort settings on a versioned evaluation set. It should measure task success, factual support, instruction compliance, tool correctness, safety, latency and review effort. Customer acceptance and subsequent workflow completion provide essential production evidence. A laboratory score without customer behaviour is an incomplete commercial measure.

The relevant metric is incremental accepted value per incremental complete cost. Suppose a medium reasoning setting improves an approved professional-analysis task from 76 per cent to 86 per cent acceptance, while a high setting improves it from 86 per cent to 88 per cent. The medium increment may be economic. The high increment may serve only high-value or high-consequence work. These percentages are illustrative and require company-specific evidence.

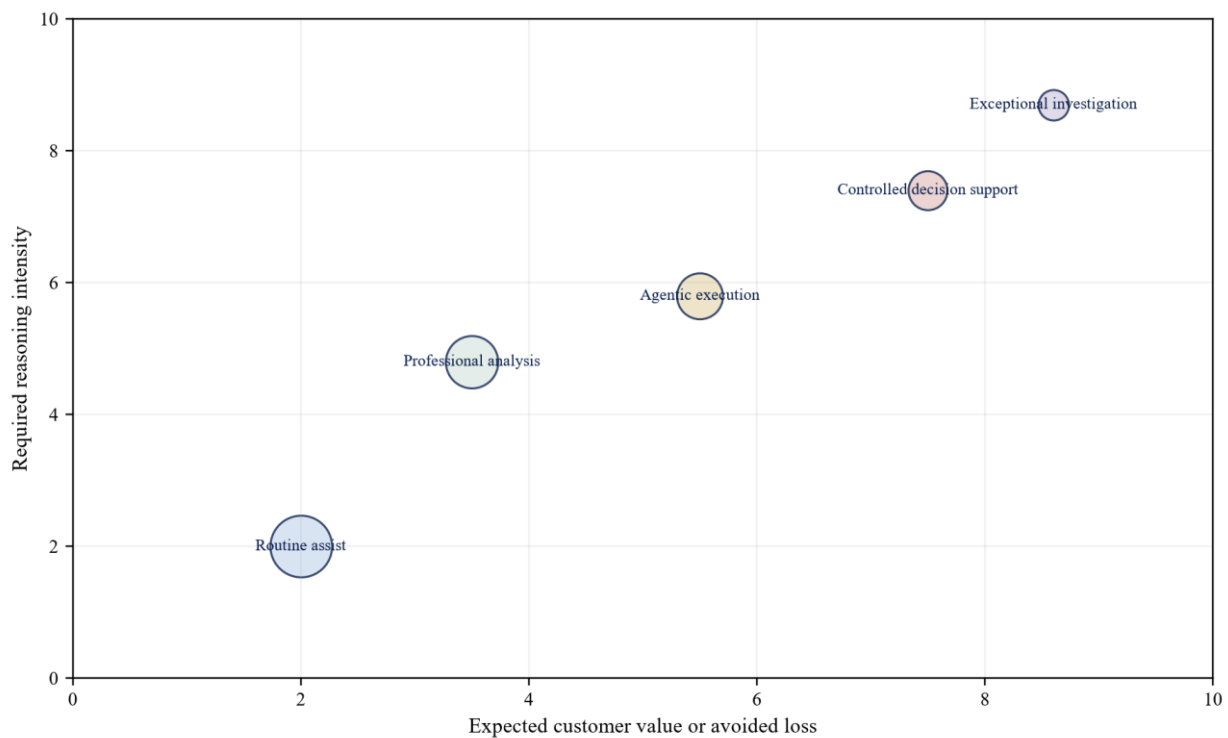
Evaluation design should protect against contamination and selection bias. Historic accepted answers can inform a benchmark, subject to rights and privacy controls. New cases should be held out. Reviewers need defined rubrics and agreement testing. Production samples should be stratified by task, customer and difficulty. Version changes require regression testing before economic assumptions are rolled forward.

4. Build a task-value router

A task-value router combines technical policy with commercial policy. It asks how much the task is worth, what can go wrong, which quality threshold applies and how much complete cost the package can support. The router then selects an approved model, reasoning level, retrieval method, tool set and verification path.

The router can use deterministic rules for high-consequence categories and learned predictions for bounded low-risk tasks. It should retain clear override, monitoring and rollback. Routing purely to the cheapest model can damage quality. Routing every task to the most capable model can destroy contribution margin. The objective is the lowest complete cost that meets the approved outcome standard for that task and customer promise.

Figure 2. Task portfolio by value and required reasoning intensity



Author framework. Bubble size represents hypothetical task volume; positions are illustrative.

The router should record rejected routes and fallbacks. A preferred route may fail because of availability, rate limits, regional constraints or an evaluation regression. Fallback economics belong in the base model when they occur regularly. Rare disaster recovery can remain a separately weighted scenario.

5. Construct the complete reasoning ledger

Provider output or thinking tokens are only one cost layer. The reasoning ledger should include foundation-model input and output, cached input, retrieval, reranking, search, code execution, external APIs, storage, data transfer, orchestration, evaluation, safety, observability, failed attempts and direct human review. Capacity commitments and minimum spends should be allocated according to a documented policy.

Table 2. Complete reasoning ledger

Ledger field	Measurement source	Economic purpose
Customer and contract	CRM, order form and billing system	connects task cost to realised price and package

Ledger field	Measurement source	Economic purpose
Task class and risk tier	product event and policy register	determines route and control requirement
Model and reasoning setting	orchestration trace	measures chosen production intensity
Input, output and thinking units	provider usage export	reconciles technical consumption to invoice
Retrieval and tool calls	trace and vendor invoice	captures non-model variable cost
Attempts, retries and fallbacks	workflow trace	identifies failure and reliability cost
Evaluation and control cost	evaluation log and service bill	prices the approved quality and safety envelope
Direct review minutes	workflow and time record	captures required human delivery work
Accepted outcome	product event and customer action	supplies the economic denominator
Collected or recognised revenue	finance ledger	supplies the contribution numerator

Inclusion and allocation depend on the service and documented accounting policy.

Finance should reconcile a sample from customer contract through task event, provider meter, invoice, evaluation and collection. Allocation choices should be documented and consistent. If evaluation or human review is required to deliver the contracted service, it should be visible in product contribution even when statutory presentation uses another line.

6. Separate reasoning contribution from reported gross margin

Reported gross margin is governed by the company's accounting policy and applicable standards. Transaction analysis can add a management measure called reasoning contribution margin. The measure deducts the complete direct cost of delivering accepted outcomes from the relevant revenue. It should be reconciled to reported financial statements and presented consistently.

The measure needs cohort views. Aggregate improvement can hide deterioration in recently acquired customers, discounted enterprise contracts or heavy users. Cohorts can be grouped by start quarter, package, task mix, geography, model route or sales channel. A durable business should show that mature cohorts produce stable or improving accepted value and cash contribution after required controls.

Reasoning contribution should also separate fixed and variable commitments. A reserved capacity contract can lower effective cost at high utilisation and increase cost at low utilisation. The model should show consumed capacity, unused capacity, overflow, renewal rights and termination exposure. Credits and temporary incentives should be excluded from the mature run-rate case unless contractually durable.

7. Design a pricing architecture that makes value legible

Pure seat pricing remains useful when customers value predictable procurement and usage variance is modest. Reasoning-intensive products need an internal usage ledger even if the customer sees a simple subscription. Without that ledger, growth in deep tasks can turn a successful product into a margin surprise.

A hybrid architecture can preserve simplicity. The subscription grants access, workflow features and an included reasoning allowance. Premium tasks consume published or contract-specific credits. Overage can be purchased or paused. High-consequence workflows can carry separate pricing because they include verification, audit evidence and human accountability. Outcome pricing can apply where the result is objectively measurable and legal, regulatory and incentive questions are addressed.

Table 3. Hybrid pricing architecture

Package component	Customer value	Cost protection	Evidence required
Base seat	access, collaboration and routine assistance	included task and credit allowance	active use and routine cost distribution
Reasoning credits	transparent access to deeper analysis	maps premium effort to economic unit	credit consumption and accepted outcome
Overage	flexibility beyond included allowance	prevents unlimited variable cost	notice, limit, price and approval trail
Premium workflow	controlled consequential analysis	funds verification, audit and review	service definition and control record
Outcome fee	payment linked to defined result	aligns price with measurable value	objective measure, attribution and contract
Enterprise commitment	predictable capacity and service	supports planning and minimum revenue	committed volume, true-up and renewal terms

Commercial design should align customer value, contract clarity and delivery economics.

Credits should correspond to a stable commercial entitlement rather than expose raw provider tokens. A provider can change tokenisation, model architecture or price. The product should preserve the right to update the conversion schedule with notice and customer protection. Contract language requires qualified legal review.

8. Model usage elasticity before promising margin expansion

A lower price per generated unit does not translate mechanically into higher margin. Customers can ask more questions. Product teams can add longer context, more agents, more retries and richer output. A cheaper model can make previously uneconomic workflows viable. These effects can increase total consumption even as the unit rate falls.

The forecast should separate four elasticities. Task-frequency elasticity measures how many more tasks users initiate when the service improves or becomes less constrained. Reasoning-intensity elasticity measures the change in compute per task. Quality elasticity measures the change in accepted outcomes from additional effort. Price elasticity measures the customer's response to package price and included allowance.

Management can test elasticity with controlled package experiments. The test should avoid consequential workflows where experimentation could create unacceptable risk. It should measure activation, accepted outcomes, retention signals, expansion, cost and support. A favourable result has more customer value and more retained contribution, rather than merely more tokens.

The experiment design should distinguish an allowance effect from a capability effect. A larger included allowance can increase use even when the underlying product has not improved. A stronger reasoning route can improve acceptance without increasing task frequency. A price reduction can stimulate use while reducing realised revenue. Randomised or phased tests can help management separate these effects where customer contracts, sample size and risk permit.

Finance should translate experiment results into a monthly contribution bridge. The bridge begins with observed customer price and task mix, applies the measured change in frequency and intensity, includes all route and control costs, and then applies the observed customer response. Management should report confidence intervals or sample limitations where relevant. A short test with self-selected power users should not be extrapolated automatically to the entire customer base.

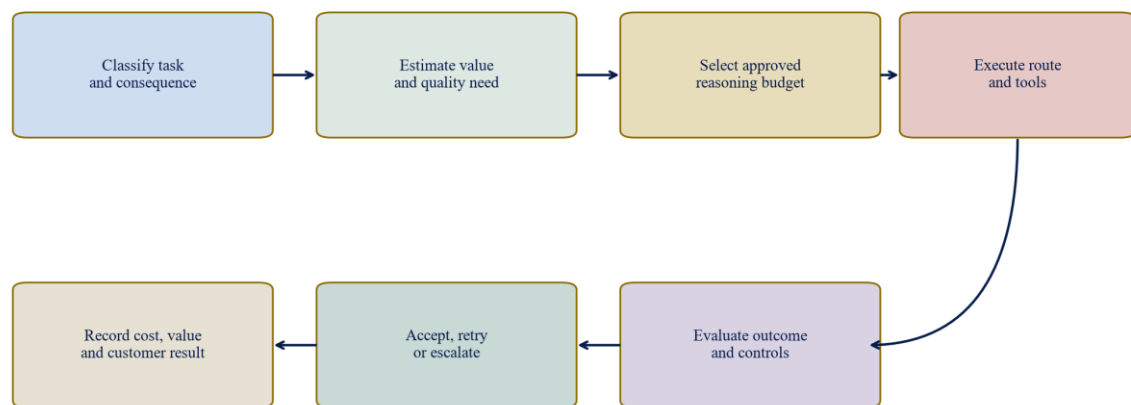
9. Connect quality to retention and expansion

Valuation requires a commercial bridge from model performance to customer behaviour. An evaluation uplift creates enterprise value only when it influences adoption, renewal, expansion, pricing or risk. The bridge can be measured through task-level acceptance, repeated use, workflow completion, saved time, avoided error and economic outcome.

Retention analysis should distinguish product dependency from promotional usage. A customer may experiment heavily during onboarding and then disengage. Another may use fewer tasks because the product resolves them efficiently. Simple request count can misclassify both customers. Cohort reporting should combine active qualified users, completed workflows, accepted outcomes, customer value indicators, realised price and complete cost.

Net revenue retention can be decomposed into seat expansion, package upgrade, reasoning-credit consumption, contractual price, contraction and churn. This decomposition reveals whether AI value is broadening the account or merely increasing variable consumption within a fixed contract. Sales compensation should reward durable contribution and collection, subject to appropriate governance, rather than uneconomic usage growth.

Figure 3. Reasoning-budget router and control points



Budget and fallback policy are versioned, monitored and reversible

Author framework. High-consequence tasks retain mandatory controls and human accountability.

10. Price latency, reliability and evidence

Deep reasoning can increase latency. Some workflows tolerate minutes or hours; others require interactive response. Premium service therefore includes more than answer quality. It can include queue priority, completion time, audit evidence, data location, failover and support.

The product should measure end-to-end latency by percentile, including retrieval, tools, retries, evaluation and review. Average latency can hide a damaging tail. Contracts should align service commitments with the actual route, capacity and fallback architecture. Service credits, penalties and customer remedies belong in the unit model.

Reliability can alter both cost and retention. A failed long-running task wastes model and tool spend and can require human recovery. Checkpointing, idempotent tools, bounded retries, fallbacks and clear stop conditions can improve completion economics. These mechanisms create engineering and infrastructure cost, so their benefit should be measured through avoided failures and protected customer value.

Evidence is a product feature in professional work. Citations, traceability, version history, approval records and reproducible calculations can justify a premium when they reduce customer review burden or support governance. The claims should be validated with observed willingness to pay and customer behaviour.

11. Govern provider and model concentration

A reasoning product can depend on one provider's capability, price, region, throughput or safety controls. A multi-provider claim is meaningful only when an alternate route meets the required quality, latency, data and tool standards. A dormant integration that fails current evaluations does not create operating resilience.

The board should receive a provider concentration register. It should show spend, tasks, customers, regions, data categories, minimum commitments, service performance, termination rights, depreciation exposure and tested alternatives. Switching cost includes prompt and workflow redesign, evaluation, regression, customer validation, migration and commercial disruption.

Open models or self-hosting can improve control for selected workloads. They also introduce infrastructure, optimisation, security, model-management and talent requirements. The make-or-buy decision should compare complete cost at representative load and the value of control. It should include the cost of maintaining quality through model and data change.

12. Protect the product contract

Commercial terms should define the service rather than promise an abstract level of intelligence. Relevant terms can cover task scope, usage allowance, credits, overage, service level, permitted data, retention, output responsibility, human review, audit evidence, model change, third-party providers and termination.

Price-adjustment language matters when provider rates, product architecture or regulatory requirements change. A customer needs predictability. The supplier needs protection against unlimited reasoning or new control cost. Minimum commitments, fair-use rules, budget alerts, admin controls and renewal adjustments can allocate the risk transparently.

Material consequential use requires explicit responsibility. The product can support analysis while the authorised person retains the decision. Contract language, user experience, workflow controls and operating practice should agree. A disclaimer that conflicts with product design or sales claims may provide weak protection.

Renewal mechanics deserve specific attention. If the supplier can change credit conversion, model route or fair-use thresholds, the customer may seek notice, termination or price protection. If the customer can roll unused credits indefinitely, the supplier may accumulate a delivery obligation whose future cost is uncertain. Contract design should align commercial flexibility with accounting, customer trust and operational capacity.

Enterprise procurement can also require committed throughput, dedicated environments or regional processing. These features can turn a nominally variable service into a capacity business. The order form should identify which capacity is reserved, how it is measured, what happens during provider interruption and who bears unused or overflow cost. The economic model should use the same definitions.

13. Align revenue and cost recognition

Pricing innovation does not remove accounting requirements. IFRS 15 requires an entity to identify the contract, performance obligations, transaction price, allocation and revenue recognition pattern under the standard.[44] A subscription, reasoning credits, implementation, premium support and outcome payment may create different accounting questions depending on the contract.

Unused credits, expiry, refunds, rollover and material rights require careful analysis. Outcome fees can contain variable consideration. Enterprise commitments may include service levels or other obligations. Management should obtain current professional advice and maintain an accounting memorandum that agrees with product configuration and invoicing.

Cost classification also needs consistency. Provider compute that directly delivers the service normally belongs in the economic contribution calculation. Product development, research and general platform work require a documented policy. Direct human review should be visible even when the statutory line differs. Investors need a reconciliation from the management measure to audited or reviewable accounts.

14. Treat governance cost as part of the product

The NIST AI Risk Management Framework and its Generative AI Profile organise governance around govern, map, measure and manage.[11][12] ISO/IEC 42001 specifies requirements for an AI management system.[50] Applicable legal and sector obligations vary by product and jurisdiction. The economic model should fund the required controls rather than assume compliance appears without cost.

Control intensity should follow the task. Routine drafting can use proportionate evaluation and monitoring. Consequential financial, legal, employment, health, safety or infrastructure analysis may require stronger data controls, validation, human review, logging and incident response. The product should identify restricted and prohibited uses and enforce them through system design.

Governance can create commercial value when it helps customers adopt the service. An enterprise buyer may pay for evidence, control, security and accountability. The company should track whether these features shorten sales cycles, support renewals or justify premium pricing. Unsupported claims remain scenarios.

15. Hypothetical enterprise-software case

Consider a hypothetical enterprise analysis platform serving finance, legal and operations teams. The platform sells annual seats with an included reasoning allowance and premium credits. All numbers in this section are management assumptions for illustration. They do not describe a named company or market forecast.

The base month has 1,000 seats at an average realised subscription price of AED 1,400 per seat per month, producing AED 1.4 million of recognised subscription revenue. Customers also buy AED 180,000 of premium reasoning credits. Total monthly revenue is AED 1.58 million. The product processes 90,000 routine tasks, 18,000 professional analyses and 2,000 controlled decision-support tasks.

Complete direct cost includes models, retrieval, tools, infrastructure, evaluation, safety, observability, customer-specific environments and required review. The base month incurs AED 632,000 of direct cost and produces AED 948,000 of reasoning contribution, or 60.0 per cent.

Table 4. Hypothetical enterprise-software operating case

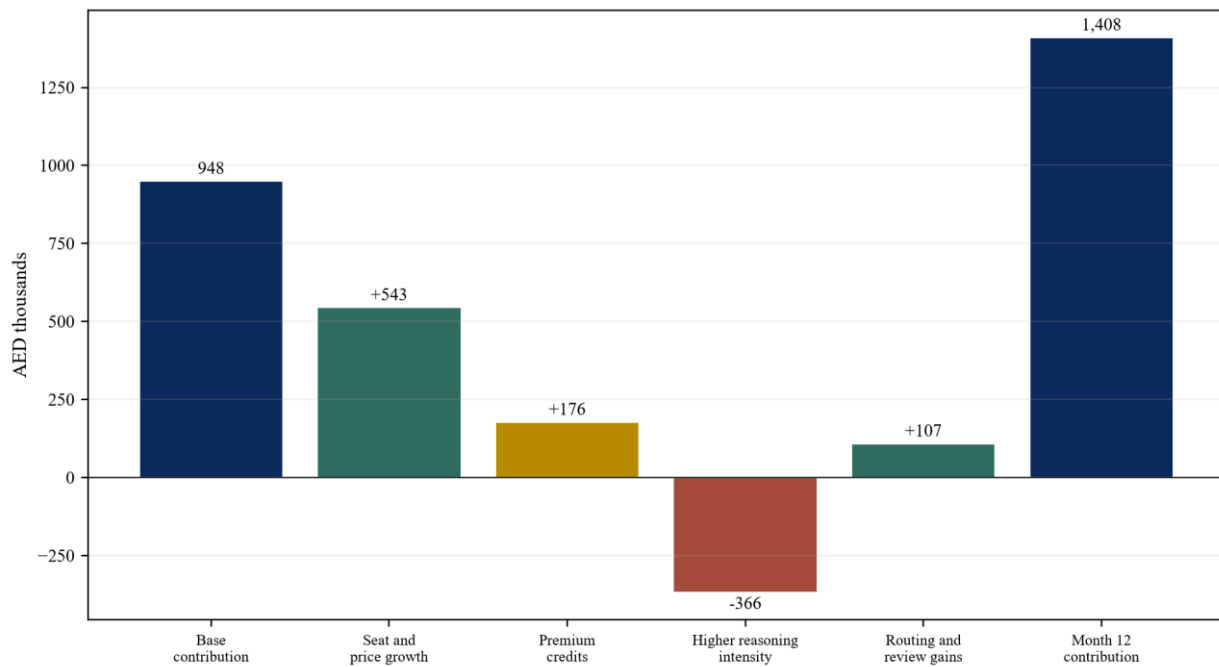
Metric	Base month assumption	Month twelve case	Evidence gate
Seats	1,000	1,340	active contracted seats and collection
Subscription revenue	AED 1,400,000	AED 1,943,000	invoice, recognised revenue and realised discount
Premium-credit revenue	AED 180,000	AED 356,000	paid credits and accepted premium tasks
Total accepted tasks	91,400	139,500	product event and approved evaluation

Metric	Base month assumption	Month twelve case	Evidence gate
Model and reasoning cost	AED 244,000	AED 391,000	provider usage and invoice reconciliation
Retrieval, tools and infrastructure	AED 138,000	AED 205,000	trace, cloud and vendor invoice
Evaluation, controls and direct review	AED 250,000	AED 295,000	test, control and workflow records
Complete direct cost	AED 632,000	AED 891,000	reconciled cost ledger
Reasoning contribution	AED 948,000	AED 1,408,000	total revenue less complete direct cost
Reasoning contribution margin	60.0%	61.2%	consistent accounting and allocation policy

Every value is an illustrative management assumption for method demonstration.

The month-twelve case assumes more seats, a higher share of premium tasks, improved routing and less review per accepted task. It also assumes a small increase in average subscription price and higher credit purchases. Each assumption requires operating evidence. The margin gain is modest because more difficult work consumes more reasoning and control.

Figure 4. Hypothetical reasoning-contribution bridge



All values are illustrative management assumptions in AED thousands.

The bridge should be rebuilt with actual cohort and task data. The transaction case should not credit routing savings that have not passed production tests. It should not treat temporary provider credits as mature cost. It should include the cash required to build evaluation, telemetry and commercial packaging.

16. Stress the task mix, not only the provider price

The central downside is an adverse shift in task mix. Customers may discover that the product is most valuable for difficult work. Premium usage can grow faster than routine usage, while the contract remains seat-based and unlimited. Average provider cost per generated unit can fall at the same time that complete cost per seat rises.

The sensitivity model should vary reasoning intensity, acceptance, task frequency, realised price, direct review, provider rate and capacity utilisation independently. Correlations also matter. A quality problem can increase retries and human review while reducing retention. A capacity shortage can increase latency, trigger premium service charges and cause customer credits.

Table 5. Hypothetical reasoning-economics sensitivities

Scenario	Realised price	Reasoning intensity	Accepted outcome rate	Complete direct cost	Contribution margin	Interpretation
Base assumption	index 100	index 100	83%	AED 632k	60.0%	illustrative current case
Premium mix with credit uptake	index 108	index 132	87%	AED 755k	62.1%	higher cost supported by price and acceptance
Unlimited-seat leakage	index 100	index 145	84%	AED 814k	48.5%	premium reasoning consumed inside fixed price
Quality regression	index 98	index 118	70%	AED 792k	48.9%	retries, review and concession pressure
Provider saving absorbed by use	index 96	index 138	85%	AED 704k	54.1%	lower unit rate does not reach margin fully
Controlled router	index 101	index 92	84%	AED 558k	65.0%	requires verified quality and customer response
Combined downside	index 90	index 150	68%	AED 910k	39.0%	liquidity and package redesign required

Every value is an illustrative management assumption and requires company-specific evidence.

These figures are illustrative assumptions. A buyer or lender should replace them with actual contract, telemetry, invoice, evaluation and collection data. The downside should flow into cash runway, debt service, covenant headroom and funding need.

17. Rebuild valuation around durable contribution

Revenue multiples can provide market context, subject to comparability. They do not solve the reasoning-economics question. A buyer can normalise revenue and gross margin, construct a cash-flow case and use transaction structures to allocate uncertainty.

The normalisation should remove temporary credits, capitalised or excluded costs that belong in the economic delivery model, unusual service work and unsupported routing savings. It should identify the mature package mix, provider commitments, review burden and control cost. Cohort contribution should be reconciled to financial statements.

A discounted cash-flow model can forecast seats, realised price, task frequency, reasoning intensity, accepted outcome rate, complete direct cost, retention, expansion, operating expense, working capital, capital expenditure and tax. The terminal case should reflect a mature task mix and sustainable control environment. A contribution multiple can be applied to verified cash contribution with an explicit reinvestment requirement.

Table 6. Reasoning-intensive SaaS valuation bridge

Valuation item	Base observation	Transaction adjustment	Required evidence
Recurring revenue	contracted and recognised subscriptions	remove non-recurring or weakly collectible amounts	contracts, billing, recognition and collections
Premium revenue	credits, overage and controlled workflows	retain only demonstrated repeat use and willingness to pay	usage, renewal and customer evidence
Direct cost	reported cost of revenue	add complete reasoning, tool, control and direct-review cost	invoice-to-task ledger and policy
Cohort retention	gross and net revenue retention	separate price, seats, credits, contraction and churn	cohort bridge and contract history
Routing benefit	current production performance	credit verified savings net of quality and implementation	controlled test and production result
Provider concentration	spend and capability dependency	price switching, disruption and commitment exposure	contracts, alternatives and tested fallback
Control maturity	current governance and evidence	fund required remediation and continuing operation	control register, incidents and roadmap
Terminal margin	forecast mature economics	use stable task mix, price and complete cost	supported operating case and sensitivities

Adjustments should be supported by reproducible evidence and reconciled to financial statements.

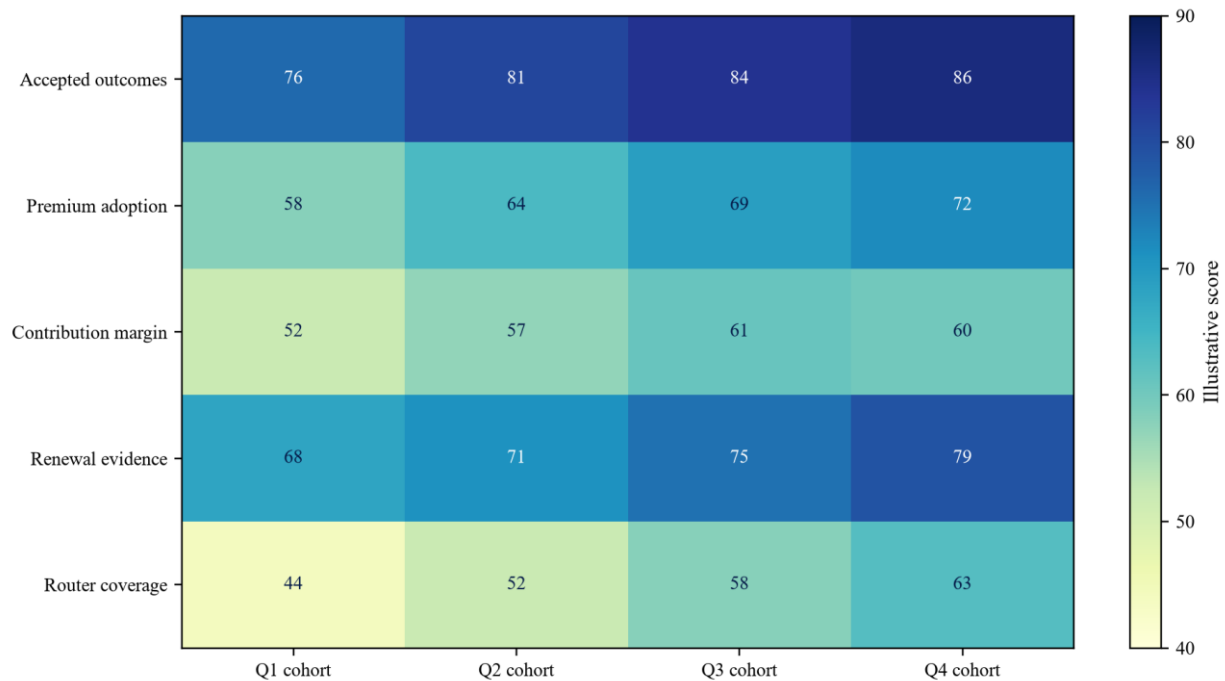
Valuation can rise when reasoning creates demonstrable customer value, supports premium price, improves retention and produces durable contribution. It can fall when unlimited commitments, weak telemetry, provider dependence or required review make margin unpredictable. The analysis should locate the evidence behind each adjustment.

Comparable-company analysis should screen for revenue model and delivery intensity. A conventional workflow SaaS business, a generative assistant, an agentic platform and a technology-enabled service can report similar subscription labels while carrying different direct cost and labour exposure. Public disclosure may not provide the required task-level detail. The analyst should state the limitation and use wider sensitivity ranges where comparability is weak.

Purchase price allocation and impairment analysis can also be affected by the durability of technology, customer relationships and forecast cash flows. IFRS 3, IFRS 13, IAS 36 and IAS 38 provide relevant accounting frameworks.[40][41][42][43] Their application depends on facts and professional judgement. Technical obsolescence, model substitution, customer portability and provider concentration can influence useful life and risk assumptions.

The capital structure should reflect downside cash behaviour. A business with annual prepaid contracts can receive cash before consuming reasoning capacity, but heavy later use can create a delivery burden. A business with consumption billing may match revenue and cost more closely while carrying demand volatility. Debt sizing should use collections, committed cost, renewal concentration and downside contribution rather than headline annual recurring revenue alone.

Figure 5. Cohort reasoning scorecard



Illustrative heatmap. Higher scores are favourable except intensity variance, which is evaluated against package design.

18. Use deal terms to allocate reasoning uncertainty

Transaction structure can bridge evidence gaps. An acquirer can tie contingent consideration to customer renewal, premium adoption or verified contribution, subject to precise definitions. Growth capital can be staged against telemetry coverage, package migration, margin and retention milestones. Debt can use covenants and liquidity cases that reflect the downside task mix.

Metric definitions must be durable. “Gross margin” can become disputed when the parties disagree about model, evaluation, support or review cost. Transaction documents can define eligible revenue, direct reasoning cost, accepted outcome, task class, allocation policy, measurement period, permitted changes, data source, audit right and dispute process. Qualified legal and accounting advice is essential.

Representations and diligence should cover provider contracts, data rights, evaluation integrity, material incidents, customer commitments, pricing controls and accounting. The buyer should understand whether historical economics depend on founder review, promotional credits, an undisclosed human service or a provider feature that can change.

An escrow or retention mechanism can fund known remediation. A price adjustment can recognise an uneconomic unlimited contract portfolio. A seller can benefit from an earn-out when a reasoning product proves premium adoption and contribution after closing. The structure should avoid incentives to suppress required control or customer support.

19. Build an evidence-led diligence room

The diligence room should allow a reviewer to reproduce economic claims. Static presentations are insufficient. A selected sample should connect contract, task, route, provider meter, tool calls, evaluation, review, invoice, customer charge and collection.

Table 7. Transaction diligence checklist

Workstream	Core evidence	Reperformance test
Customer and pricing	contracts, packages, allowances, credits, invoices and collections	rebuild realised price by cohort and task mix
Product telemetry	task, route, reasoning setting, tools, retries, latency and acceptance	trace selected tasks from initiation to accepted outcome
Provider economics	contracts, rate cards, commitments, usage, credits and invoices	reconcile provider bill to task and customer allocation
Quality and controls	evaluation sets, rubrics, results, incidents and approvals	rerun material task classes on current and fallback routes
Human delivery	review, exceptions, implementation and support records	allocate direct minutes and cost to workflows and customers
Finance	revenue policy, cost policy, ledger, budgets and cash forecast	reconcile reasoning contribution to financial statements
Resilience and governance	data flow, regions, failover, security, privacy and AI governance	test fallback, change approval and incident evidence

The reviewer should select representative samples and preserve data protection and confidentiality.

Data protection and confidentiality apply to the room. Access should be scoped, logged and time-limited. Customer and personal data can be minimised or anonymised where appropriate. Evidence requests should align with contractual and legal permissions.

The reviewer should select samples rather than accept only management-curated examples. A balanced sample can include high and low margin customers, new and mature cohorts, routine and premium tasks, successful and failed workflows, normal and fallback routes, and periods before and after a model change. The objective is to test whether the ledger behaves consistently across the portfolio.

Customer calls can validate the economic story. The discussion should distinguish satisfaction with the interface from dependence on the completed workflow. Relevant questions include the alternative process, time saved, error reduction, approval burden, budget owner, renewal decision, willingness to pay and response to usage limits. Statements should be documented as customer evidence and should not replace operating data.

Cyber-security and privacy diligence should follow the actual data path. The team should understand input collection, retrieval stores, provider processing, tool permissions, logs, retention, model improvement terms, regional transfer and deletion. A high-margin route that violates a customer commitment is not an acceptable route. Remediation cost and commercial impact belong in the transaction model.

20. Execute a 180-day margin programme

The first 30 days establish definitions and facts. Management approves the task taxonomy, accepted-outcome standard, cost policy and provider register. Finance reconciles selected customers. Product and engineering instrument the missing events. Commercial teams identify unlimited or unclear packages.

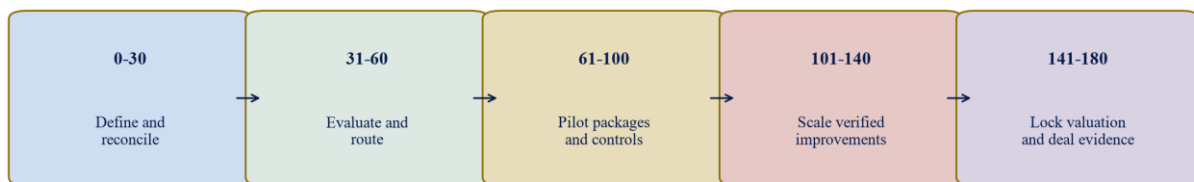
Days 31 to 60 establish the router and evaluation baseline. The company compares approved models and reasoning settings by task. It defines fallback, stop and escalation rules. Finance publishes cohort contribution. Legal and commercial teams prepare revised package terms.

Days 61 to 100 pilot packaging and routing. Selected customers receive clear allowances, premium credits or workflow tiers. Experiments measure acceptance, adoption, retention signals, support and contribution. High-consequence tasks retain approved controls.

Days 101 to 140 expand verified improvements. The company migrates eligible customers, negotiates provider commitments only against evidenced demand and removes uneconomic routes. The board reviews provider concentration, control investment and forecast variance.

Days 141 to 180 lock the transaction evidence. Management produces the cohort scorecard, valuation bridge, downside liquidity case and data-room sample. Remaining unsupported benefits stay outside the base case.

Figure 6. 180-day reasoning-economics roadmap



Gate: contract + telemetry + evaluation + invoice + collection

Author framework. Each phase closes with an evidence gate.

21. Convert the thesis into board decisions

The board should approve three numbers for each material task class: the quality threshold, the compute budget and the commercial allowance. These numbers connect product, finance and customer promise. Management can change them through an approved process supported by evaluation and operating evidence.

The monthly board pack should show accepted outcomes, realised price, reasoning intensity, complete cost, contribution, retention signals and incidents by cohort. Variance should be separated into customer price, task frequency, route mix, reasoning setting, provider rate, retries, tools, review and customer mix. Each material variance needs an owner and action.

Capital allocation should follow marginal evidence. Engineering effort can target task classes where better routing or workflow design improves contribution without lowering acceptance. Commercial effort can migrate packages where premium use is trapped inside a fixed seat price. Procurement can negotiate commitments after eligible demand and fallback exposure are understood.

The company should maintain a model-version register and package register. The first records approved tasks, reasoning settings, evaluation results, expected cost, provider, fallback and retirement. The second records included allowance, credits, overage, renewal, price-review terms and economic assumptions. Linking them identifies contracts that become uneconomic after a model or usage change.

22. Apply a transaction decision matrix

An investor can classify the opportunity across four evidence states. Proven value with proven contribution supports the base case. Proven value with weak contribution requires pricing, routing or

contract remediation. Weak value with low cost still presents product-market risk. Weak value with high cost requires a fundamental product and capital-allocation review.

Each state leads to a different action. A strong operating case can support ordinary valuation methods and growth investment. A remediable margin case can support a phased plan, deferred consideration or milestone funding. A value-evidence gap requires customer research and controlled adoption tests. A combined value and cost problem should remain outside the investment base case until evidence changes.

The final investment committee paper should state which benefits are observed, which are contracted, which are tested and which remain management assumptions. It should show the source, period, owner and sensitivity of every material claim. This discipline allows a transaction team to move quickly while preserving the boundary between current evidence and prospective value creation.

Conclusion

Reasoning models make inference effort a controllable input to a customer service. This capability can improve difficult professional work. It also weakens the assumption that every additional SaaS interaction has a negligible and predictable marginal cost.

The appropriate economic unit is the accepted customer outcome. Management should link task value, compute budget, quality uplift, acceptance, retention, expansion, complete direct cost and cash contribution. The product should allocate reasoning through an approved task-value router and price premium work through allowances, credits, overages, workflow tiers or measurable outcomes.

Investors should evaluate cohort contribution after the complete cost of models, retrieval, tools, infrastructure, evaluation, controls and required human review. Revenue multiples require normalisation for task mix, provider dependence, unlimited commitments and unsupported margin improvements. Cash-flow valuation should use explicit operating drivers and downside cases.

The valuation reset rewards evidence. A company that proves premium customer value, disciplined compute allocation, strong retention and durable contribution can earn a stronger case. A company that cannot connect reasoning spend to accepted value leaves the benefit as an unverified management scenario. The transaction model should preserve that distinction.

That evidence should be refreshed as models, provider terms, customer behaviour and product controls change. A current ledger and repeatable decision process create more reliable management information than a one-time diligence calculation.

Frequently asked questions

Q: What is test-time compute in a SaaS product? A: It is the inference effort allocated while the product processes a customer task. Depending on the provider and architecture, it can include reasoning or thinking tokens, multiple candidates, retries, tools, retrieval and verification. The applicable provider documentation defines the technical meter.

Q: Why can reasoning models reduce SaaS gross margin? A: Users and workflows can consume materially different inference effort within the same seat price. Deeper tasks can also require more tools, evaluation, infrastructure and review. Margin weakens when the contract does not price or limit that complete cost.

Q: When is additional reasoning cost economic? A: It is economic when the expected improvement in accepted outcome, customer value, retention, expansion or avoided risk exceeds the complete incremental cost within the approved service and control envelope.

Q: Should customers be billed for raw thinking tokens? A: A product can use stable commercial credits or task tiers instead of exposing a provider-specific meter. The conversion should be transparent enough for customer governance and adaptable to model or price changes under the contract.

Q: How should an investor test the claimed margin? A: The investor can reproduce selected customers and tasks from contract and price through route, reasoning setting, provider usage, tools, evaluation, direct review, invoice and cash collection. Cohort results should reconcile to reported accounts.

Q: Does a higher reasoning budget always improve quality? A: No. The response varies by task, model, evaluation and implementation. The company should measure the quality-response curve and stop allocating additional effort when incremental accepted value no longer supports incremental cost.

Q: How should reasoning economics affect valuation? A: Valuation should use durable cohort contribution, retention and cash flow after complete direct cost. Unsupported routing savings, promotional credits and unlimited-package assumptions should remain outside the base case or receive explicit risk treatment.

Q: Which Matchpoint practice is relevant to this work? A: This research connects to Matchpoint Partners' strategy and execution advisory work, including pricing, unit economics, operating-model design, transaction diligence, valuation, capital planning and post-investment execution.

References

- [1] OpenAI, Learning to Reason with LLMs, 12 September 2024, <https://openai.com/index/learning-to-reason-with-llms/>
- [2] Google AI for Developers, Gemini Thinking, <https://ai.google.dev/gemini-api/docs/thinking>
- [3] Anthropic, Extended Thinking, <https://platform.claude.com/docs/en/build-with-claude/extended-thinking>
- [4] OpenAI, API Pricing, <https://openai.com/api/pricing/>
- [5] OpenAI, Reasoning Models Guide, <https://platform.openai.com/docs/guides/reasoning>
- [6] Google AI for Developers, Gemini Developer API Pricing, <https://ai.google.dev/gemini-api/docs/pricing>
- [7] Anthropic, Claude Pricing, <https://platform.claude.com/docs/en/about-claude/pricing>
- [8] Amazon Web Services, Amazon Bedrock Pricing, <https://aws.amazon.com/bedrock/pricing/>
- [9] Microsoft, Provisioned Throughput Billing and Cost Management, <https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/provisioned-throughput-billing>
- [10] MLCommons, MLPerf Inference: Datacenter, <https://mlcommons.org/benchmarks/inference-datacenter/>
- [11] National Institute of Standards and Technology, AI Risk Management Framework, <https://www.nist.gov/itl/ai-risk-management-framework>
- [12] National Institute of Standards and Technology, Generative Artificial Intelligence Profile, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
- [13] Stanford Institute for Human-Centered Artificial Intelligence, AI Index Report 2026, https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf
- [14] Stanford Institute for Human-Centered Artificial Intelligence, AI Index Report 2025, https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf
- [15] OpenAI, Prompt Caching Guide, <https://platform.openai.com/docs/guides/prompt-caching>
- [16] OpenAI, Batch API Guide, <https://platform.openai.com/docs/guides/batch>
- [17] Anthropic, Message Batches, <https://platform.claude.com/docs/en/build-with-claude/batch-processing>
- [18] Anthropic, Prompt Caching, <https://platform.claude.com/docs/en/build-with-claude/prompt-caching>
- [19] Google Cloud, Vertex AI Generative AI Pricing, <https://cloud.google.com/vertex-ai/generative-ai/pricing>
- [20] Google Cloud, Vertex AI Context Caching, <https://cloud.google.com/blog/products/ai-machine-learning/vertex-ai-context-caching>
- [21] 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>
- [22] Microsoft, Plan and Manage Costs for Microsoft Foundry, <https://learn.microsoft.com/en-us/azure/foundry/concepts/manage-costs>
- [23] GitHub, Copilot Plans, <https://github.com/features/copilot/plans>
- [24] GitHub Docs, Copilot Billing, <https://docs.github.com/en/copilot/reference/copilot-billing>
- [25] Adobe, Creative Cloud Pricing, <https://www.adobe.com/creativecloud/pricing.html>
- [26] Salesforce, Agentforce Pricing, <https://www.salesforce.com/agentforce/pricing/>
- [27] Salesforce, Annual Reports, <https://investor.salesforce.com/financials/default.aspx>
- [28] ServiceNow, Annual Reports, <https://www.servicenow.com/company/investor-relations/financial-results.html>

- [29] Palantir Technologies, Annual Reports, <https://investors.palantir.com/financials/annual-reports>
- [30] European Commission, Guidelines on Obligations for General-Purpose AI Providers, <https://digital-strategy.ec.europa.eu/en/faqs/guidelines-obligations-general-purpose-ai-providers>
- [31] European Commission, Transparency Obligations under Article 50 of the AI Act, <https://digital-strategy.ec.europa.eu/en/faqs/transparency-obligations-under-article-50-ai-act>
- [32] Organisation for Economic Co-operation and Development, Measuring the Environmental Impacts of AI Compute and Applications, <https://wp.oecd.ai/app/uploads/2025/05/7babbf571-en.pdf>
- [33] International Energy Agency, Energy and AI Executive Summary, <https://www.iea.org/reports/energy-and-ai/executive-summary>
- [34] International Energy Agency, Energy Demand from AI, <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>
- [35] Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing, <https://cloudsecurityalliance.org/research/guidance>
- [36] OWASP Foundation, Top 10 for Large Language Model Applications, <https://owasp.org/www-project-top-10-for-large-language-model-applications/>
- [37] MITRE, ATLAS, <https://atlas.mitre.org/>
- [38] United States Securities and Exchange Commission, Commission Guidance on Management's Discussion and Analysis, <https://www.sec.gov/rules/interp/33-8350.htm>
- [39] IFRS Foundation, Conceptual Framework for Financial Reporting, <https://www.ifrs.org/issued-standards/list-of-standards/conceptual-framework/>
- [40] IFRS Foundation, IFRS 13 Fair Value Measurement, <https://www.ifrs.org/issued-standards/list-of-standards/ifrs-13-fair-value-measurement/>
- [41] IFRS Foundation, IFRS 3 Business Combinations, <https://www.ifrs.org/issued-standards/list-of-standards/ifrs-3-business-combinations/>
- [42] IFRS Foundation, IAS 36 Impairment of Assets, <https://www.ifrs.org/issued-standards/list-of-standards/ias-36-impairment-of-assets/>
- [43] IFRS Foundation, IAS 38 Intangible Assets, <https://www.ifrs.org/issued-standards/list-of-standards/ias-38-intangible-assets/>
- [44] 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/>
- [45] Financial Accounting Standards Board, Accounting Standards Codification Topic 606, <https://asc.fasb.org/topic&trid=2129886>
- [46] International Valuation Standards Council, International Valuation Standards, <https://www.ivsc.org/standards/>
- [47] Association of International Certified Professional Accountants, SaaS Metrics and Financial Reporting Resources, <https://www.aicpa-cima.com/resources>
- [48] National Institute of Standards and Technology, Secure Software Development Framework, <https://csrc.nist.gov/Projects/ssdf>
- [49] International Organization for Standardization, ISO/IEC 27001 Information Security Management Systems, <https://www.iso.org/isoiec-27001-information-security.html>
- [50] International Organization for Standardization, ISO/IEC 42001 AI Management Systems, <https://www.iso.org/standard/81230.html>

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.