AI Benchmark Diligence: Reproducing Model and Product Claims before Signing

A Transaction Framework for Testing Leakage, Comparability and Production Performance

Authored by Chennakeshav Adya, Independent Researcher

September 2026

Abstract

Artificial-intelligence businesses are frequently valued through claims about model accuracy, latency, cost, safety, adoption and product superiority. A headline benchmark can conceal material differences in datasets, task definitions, prompts, hardware, software versions, sample exclusions and statistical treatment. Public test sets may have entered training data. A result obtained in a research environment may deteriorate under production traffic, customer data, security controls and unit-cost constraints. These differences can affect forecast revenue, gross margin, capital requirements, contractual performance and the durability of an acquisition thesis.

This paper develops an evidence-led benchmark-diligence framework for boards, investment committees, corporate-development teams, sponsors, lenders and technical advisers. It reconstructs the benchmark lineage, freezes the claimed configuration, tests contamination and selection effects, reproduces the result under controlled conditions, and bridges laboratory performance to production economics. Five original figures and five decision tables present the evidence graph, reproducibility protocol, leakage tests, production-gap bridge and transaction adjustment matrix.

The worked example concerns a hypothetical acquisition and uses analytical assumptions. All amounts, scores, percentages, probabilities, multiples and scenarios are illustrative assumptions. Reproduction does not prove general performance beyond the tested population, and a failed reproduction does not by itself establish misconduct. The framework is designed to improve evidence quality and decision control. It does not provide legal, regulatory, tax, accounting, cybersecurity or investment advice and does not recommend a transaction, valuation or financing structure.

JEL Classification: G34, G32, C12, C18, O32

Keywords: artificial intelligence, mergers and acquisitions, technical due diligence, benchmarks, reproducibility, data contamination, valuation, model evaluation, production performance, transaction pricing

1. Convert every benchmark claim into a testable proposition

A benchmark claim should identify the system, task, population, intervention, comparator, metric and decision relevance. The statement that a model is more accurate than a competitor is incomplete until the buyer knows which model checkpoint, prompt policy, tool configuration, dataset, exclusions, scoring rule and confidence interval produced the result. The statement that an application is faster or cheaper is incomplete until concurrency, hardware, caching, retries, human review and failure handling are included.

The diligence register should preserve the exact wording presented to customers, investors and the transaction committee. Each claim should link to source code, weights or API version, evaluation data, execution logs, hardware, software dependencies, metric implementation, analyst notebooks and approvals. Evidence should retain timestamps and hashes where practicable. A

presentation number that cannot be traced to a reproducible run belongs in an unresolved state.

Claims should also carry a transaction use. A result may support product differentiation, customer retention, pricing power, gross margin, regulatory readiness or a contingent payment. The required test strength depends on that use. A marketing comparison can require correction without changing value. A claimed capability that supports most of the forecast may require an independent blind evaluation and a contractual remedy before signing.

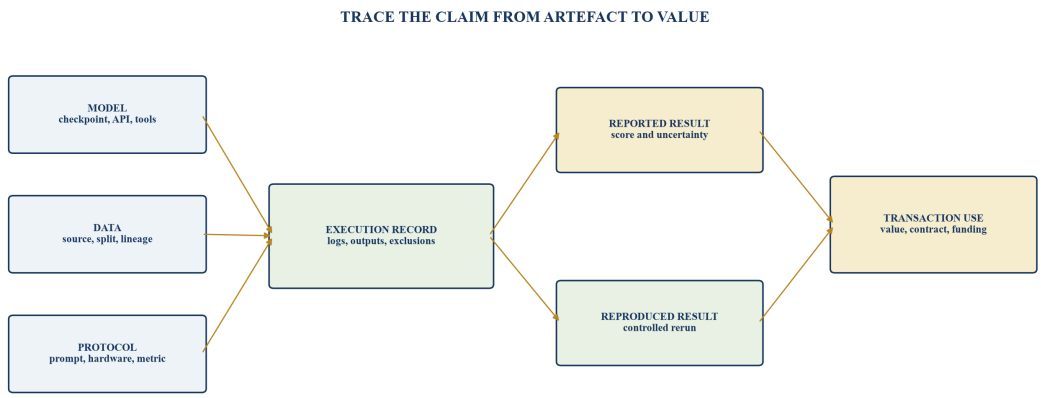
2. Build a benchmark evidence graph

The evidence graph connects the reported score to the artefacts and decisions that produced it. Nodes represent the model, dataset, task, prompt, tool, hardware, software version, metric, exclusion, human rater, run and published claim. Edges record dependencies and transformations. The graph should allow a reviewer to move from a board-slide number back to raw outputs and forward to the commercial assumption it supports.

Every node needs an evidence state: preserved and verified, preserved but unverified, management representation, third-party record, unavailable or superseded. A hash establishes file identity, while it does not establish that the file is complete, appropriate or free from contamination. Reviewers should therefore record both provenance and fitness for purpose.

The graph exposes silent substitutions. A public claim may cite one benchmark while the evaluation notebook uses a modified split. The product may call a newer model than the one described in diligence. A latency chart may omit failed requests. Machine-assisted reconciliation can flag mismatched identifiers, dates, sample counts and versions for expert review. The resulting graph becomes the control surface for reproduction, exceptions and transaction treatment.

Figure 1. Benchmark claim evidence graph



A transaction-relevant claim should be traceable from source artefacts through evaluation to the commercial assumption.

Table 1. Minimum benchmark-claim record

Field	Required evidence	Core test	Decision owner
system	model, checkpoint, API, tools and configuration	can the tested system be reconstructed?	technical lead
data	source, licence, population, split and hashes	is the evaluation population appropriate and uncontaminated?	data lead
protocol	prompts, sampling, hardware, software and retries	were conditions fixed and comparable?	evaluation lead
metric	implementation, aggregation and uncertainty	does the score measure the claimed outcome?	methodology lead
result	raw outputs, failures, exclusions and logs	can the published number be recalculated?	independent reviewer
commercial use	forecast, contract, margin or value linkage	how does the claim enter the transaction case?	deal lead
authority	preparer, reviewer, date and approval	who accepted the claim and its limitations?	diligence chair

The record separates identity, method, result and transaction use.

3. Define the decision before choosing the test

Evaluation design should begin with the transaction decision. A buyer deciding whether a product meets a contractual accuracy threshold needs a representative customer population, the contractual metric and agreed failure treatment. A buyer underwriting gross margin needs throughput, hardware, inference cost, human review and support burden. A board assessing technical differentiation needs relevant comparators and evidence that the advantage persists outside a curated demonstration.

One benchmark rarely answers all these questions. Capability tests measure performance on defined tasks. Robustness tests vary inputs and operating conditions. Safety tests probe harmful or policy-violating behaviour. Economic tests connect quality to latency, compute, labour and service levels. Product tests incorporate the interface, retrieval, tools, workflow and human escalation. The diligence plan should state which question each test answers and which question it cannot answer.

The team should agree materiality before seeing reproduced results. Materiality can relate to revenue concentration, contractual penalties, forecast margin, regulated use, customer harm, capital needs or valuation. A preregistered decision rule reduces the temptation to redefine success after an inconvenient result. It also helps the seller provide the right artefacts and protects the buyer from spending time on scores with little transaction consequence.

4. Freeze the claimed system and evaluation environment

Reproduction requires an identified object. The diligence team should freeze the model checkpoint or API version, system prompt, retrieval corpus, tool definitions, safety layers, sampling parameters, software packages, hardware profile and evaluation code. Containers, dependency locks, hashes and run manifests can support the freeze. When a third-party API prevents full preservation, the team should record provider, endpoint, model identifier, date, region and observed response metadata.

The frozen system should reflect the claim being diligenced. A seller may demonstrate a research checkpoint while customers use an optimised product stack. Both may deserve testing, but the results should remain distinct. Fine-tuning, retrieval, post-processing and human review can create value that does not reside in the base model. The buyer should avoid attributing product performance to proprietary model capability when the advantage comes from a replaceable external component.

Access controls matter. The reproduction environment can contain confidential code, customer data and model weights. The clean team should log users, data transfers, executions and exports. Reproduction evidence should survive signing while respecting licences, privacy, trade secrets and security. Counsel and security specialists should define permitted access, retention and destruction.

5. Reconstruct dataset lineage and intended population

Dataset lineage should describe where examples originated, how they were collected, filtered, labelled, deduplicated and split, and which versions entered training, tuning and evaluation. The buyer needs item-level identifiers or defensible substitutes to test overlap. Aggregate statements about proprietary data provide limited assurance when the valuation depends on data advantage.

The evaluation population should match the commercial claim. A medical workflow, credit decision, industrial inspection or legal drafting tool may face case mix, language, geography, prevalence and user behaviour that differ from a public benchmark. Performance can change when rare but costly cases receive greater weight. The diligence team should compare benchmark distributions with actual or contracted use and document the coverage gap.

Licensing and consent require separate review. A dataset can be statistically useful while creating contractual, privacy, intellectual-property or regulatory exposure. The benchmark record should link provenance, permitted use, retention and transfer restrictions. Qualified advisers should evaluate applicable obligations. Technical reproduction cannot cure a right that the target does not possess.

6. Test direct train-test overlap

Direct overlap occurs when evaluation items, answers or close duplicates appear in training or tuning material. It can inflate apparent generalisation because the model has encountered the test. Detection can combine exact hashes, canonicalised text matching, n-gram overlap, semantic similarity, image fingerprints and metadata. The method should account for formatting changes, translations, paraphrases and derived examples.

A match needs interpretation. Widely published knowledge can appear legitimately across sources, while a distinctive benchmark question and answer pair creates a stronger concern. The team should record match strength, direction, timing and whether the item could have influenced training. Where the seller cannot expose training data, alternatives include attestations, data-generation records, membership-inference research, canary items and new sequestered tests.

The appropriate response depends on prevalence and claim sensitivity. The reviewer can recalculate performance after removing suspected overlaps, compare affected and unaffected strata, and test a fresh equivalent set. The result should show a range rather than compress uncertainty into a binary contaminated or clean label. A material fall in an outcome supporting price requires escalation.

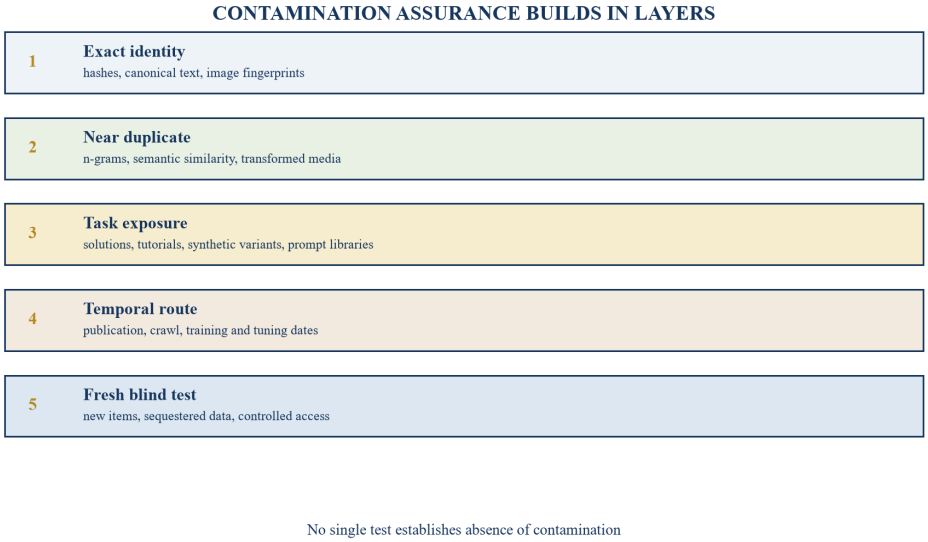
7. Detect semantic contamination and benchmark familiarity

Contamination can extend beyond exact copies. Tutorials, solution repositories, benchmark discussions, synthetic variants and model-generated training data may encode the task or answer structure. Models can also learn benchmark-specific formatting and strategies without memorising individual items. Semantic searches, repository histories, web-crawl dates and exposure analysis can help identify plausible routes.

The diligence team should compare performance on legacy public items with newly authored, sequestered and structurally equivalent items. A large gap may indicate exposure, distribution shift, test-construction differences or several effects together. It is evidence requiring investigation rather than proof of a particular cause. Blind administration and strict access controls strengthen the inference.

Benchmark familiarity also affects agentic systems. Tools, prompts or routing logic may contain benchmark names, answer patterns or task-specific shortcuts. The system freeze should therefore include orchestrator code and prompt libraries. A base model that performs modestly can appear strong when a hand-built harness is tailored to the test. That harness may still be valuable, but its transfer to customer workflows must be demonstrated.

Figure 2. Layered leakage and contamination tests



Stronger evidence combines artefact matching, temporal analysis and fresh sequestered evaluation.

Table 2. Leakage test matrix

Leakage route	Test	Stronger evidence	Limitation
exact item overlap	hashes and canonicalised matching	dated corpus records and item-level matches	misses transformations
near duplicates	lexical and semantic similarity	reviewed match clusters	similarity does not prove training use
answer exposure	solution and repository search	dated access or ingestion record	public knowledge can be legitimate
prompt tailoring	inspect prompts and harness code	frozen generic protocol comparison	product optimisation can be real value
temporal contamination	compare publication and training dates	immutable logs and snapshots	dates can be incomplete
broad familiarity	legacy versus fresh blind items	sequestered equivalent task set	new sets may differ in difficulty

Results should be interpreted with timing, access and benchmark design.

8. Test benchmark selection and omitted alternatives

Selection bias can arise before any run. A target may report benchmarks where it performs well, choose favourable subtasks or omit respected alternatives. The diligence team should reconstruct the candidate benchmark universe from product claims, customer requirements, regulatory expectations, academic practice and competitor disclosures. Management should explain inclusion and exclusion criteria.

Relevance matters more than volume. Running dozens of public leaderboards can create a false sense of coverage while leaving the buyer's commercial use untested. The team should map each selected benchmark to the intended population, error cost and decision. It should also examine results that were generated internally and not published, because abandoned or unfavourable experiments can reveal selection effects.

A benchmark portfolio can include public comparability tests, private customer-like tests, stress tests, economic tests and safety tests. The weights should reflect transaction exposure and remain fixed before results are known. If a seller chooses a different portfolio, the buyer should preserve both views and bridge the difference.

9. Reproduce metric calculation from raw outputs

The reported score should be recalculated from raw predictions, labels and metric code. Reviewers should test denominators, averaging, class weights, ties, abstentions, timeouts, invalid outputs and duplicate records. Macro and micro averages can tell different stories. Accuracy can obscure class imbalance. A composite score can hide a material failure in one dimension.

Uncertainty belongs beside the point estimate. Sample size, dependence, repeated runs, annotator variation and stochastic generation affect precision. NIST's work on statistical models for automated benchmark evaluations emphasises the need to define the measurement target and assumptions supporting conclusions. The diligence report should state confidence intervals or other suitable uncertainty measures and avoid false precision.

Metrics should connect to consequences. A one-point improvement may have little customer value, while a small deterioration in a high-cost error class can matter greatly. Thresholds should be justified by contracts, workflow economics, risk appetite and user needs. A leaderboard rank alone rarely establishes economic significance.

10. Recreate prompts, tools and human intervention

Generative and agentic systems can be highly sensitive to prompts, tool access, retrieval, sampling and evaluator instructions. The claimed protocol should preserve system and user prompts, few-shot examples, tool schemas, retry logic, temperature, seeds where supported, context limits and stopping conditions. Undocumented manual edits or cherry-picked completions should be treated as interventions.

Human assistance must be visible. Staff may select inputs, repair outputs, choose among candidates, label borderline answers or override failures. Human work can be an intentional part of a valuable service. Its time, skill, cost, consistency and scalability should enter the benchmark and the commercial model. A product described as automated should not receive automated economics when its demonstrated quality depends on unrecorded expert labour.

The reproduction plan should run the disclosed configuration first. Sensitivities can then vary prompts and controls to test robustness. A method that only succeeds under one finely tuned instruction may be appropriate for a stable workflow, while it offers weaker evidence for general capability.

11. Control hardware, concurrency and system optimisation

Latency, throughput and cost claims depend on hardware, precision, batch size, sequence length, caching, network, region, concurrency and utilisation. A target can produce an impressive laboratory result on scarce accelerators at low load while the forecast assumes commodity infrastructure and sustained traffic. The diligence environment should capture hardware and system telemetry alongside quality.

Warm and cold starts, cache hits, retries and failed requests require separate reporting. Percentile latency is generally more informative than an average for service commitments. Throughput should include backpressure and queueing. Cost per successful task should incorporate inference, retrieval, tools, storage, observability, human review and allocated infrastructure rather than token cost alone.

Optimisation can be proprietary and durable. Quantisation, compilation, routing and scheduling may create genuine advantage. The buyer should test whether the optimisation transfers with the acquired assets, depends on a third-party licence or requires specialists who may not remain. The benchmark claim and transaction deliverability should therefore be connected.

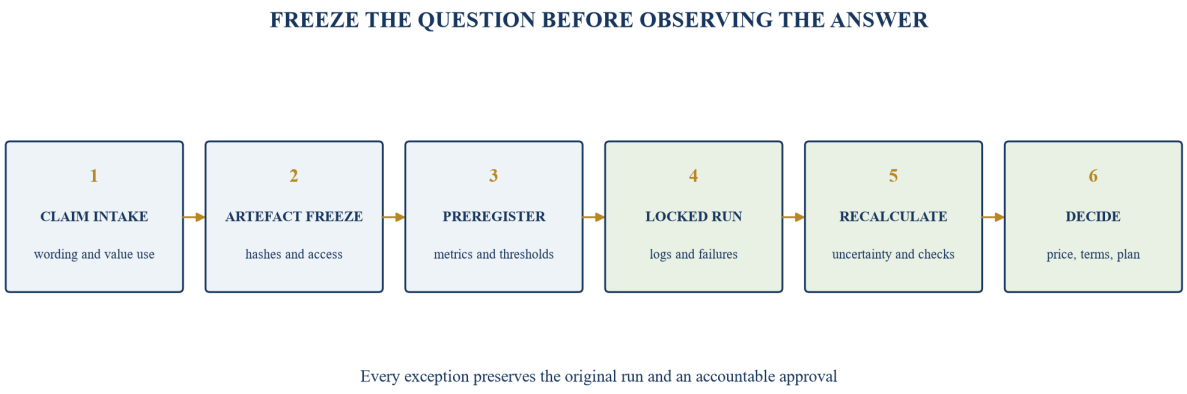
12. Run a preregistered reproduction protocol

The protocol should be signed off before the independent team sees final results. It states the claimed system, test population, sample, metrics, statistical method, environment, permitted interventions, exception rules and decision thresholds. It identifies who can access labels, who executes runs and who approves deviations. Preregistration reduces hindsight changes and supports a defensible audit trail.

Execution can proceed through intake, integrity checks, dry run, locked run, independent recalculation, sensitivity analysis and exception review. The dry run validates plumbing on non-scored items. The locked run uses controlled artefacts and produces immutable logs. Any rerun should retain the original output and a reason.

The team should distinguish reproducibility from replication. Reproduction can mean obtaining the result with the seller's artefacts and method. Replication can test the underlying claim with independently constructed data or methods. A strong diligence programme may need both, particularly when the public benchmark is exposed or the commercial population differs.

Figure 3. Controlled benchmark-reproduction protocol



Decision rules and permitted changes are fixed before the scored run.

Table 3. Reproduction protocol gates

Gate	Required output	Failure response	Authority
claim intake	exact claim and commercial use	remove ambiguity before testing	deal lead
artefact freeze	manifest, versions, hashes and access log	record missing or mutable artefacts	technical lead
preregistration	sample, metric, uncertainty and threshold	defer scored execution	methodology lead
locked run	complete raw outputs, telemetry and failures	quarantine run and investigate	evaluation lead
recalculation	independent score and sensitivity	reconcile code or data difference	independent reviewer
decision	treatment of value, contract and integration	escalate material gap	diligence chair

Each gate produces evidence for the next transaction decision.

13. Use blind and sequestered evaluation where exposure matters

Blind testing restricts the evaluated party's access to labels, expected outputs or final items. Sequestered testing keeps data in a controlled environment and can prevent inclusion in future training. NIST's AI Technology Evaluation programme uses blind data and a sequestered environment to reduce contamination risk and improve comparability. Transaction teams can adapt the principle to material claims.

The seller can provide a container or controlled endpoint while an independent team holds the test set. Inputs should reflect the agreed population, and the process should capture outputs, refusals, errors, timing and resource use. Security controls should prevent data exfiltration through logs, tools or model updates. The protocol should also prevent evaluator leakage through informal feedback during execution.

Fresh data does not automatically create a fair test. Item quality, difficulty and labelling require review. Where possible, a calibration subset can be jointly examined before the scored set is locked. Disputes should follow a predefined adjudication process that preserves blinded decisions and avoids selective relabelling.

14. Test robustness across meaningful operating conditions

A single score describes one point in a larger operating space. Robustness tests vary language, geography, user expertise, input quality, document length, class prevalence, adversarial behaviour, time pressure, tool failure and system load. The variations should reflect customer and regulatory exposure rather than arbitrary perturbations.

The team should distinguish graceful degradation from cliff effects. A model may decline modestly as documents become noisier, then fail abruptly beyond a context or retrieval threshold. A workflow may remain accurate while latency and cost breach service levels. These transition points can drive capacity planning, contract terms and the integration roadmap.

Robustness claims need uncertainty and coverage. Testing every condition is impossible. The report should state the sampled region, untested combinations and monitoring needed after close.

The buyer can protect value through staged deployment, holdbacks, support commitments or contingent consideration when material conditions remain untested.

15. Compare against credible alternatives

Differentiation requires a comparator that the buyer or customer could actually use. This may include a current product, open model, commercial API, specialist vendor, human workflow or simpler statistical method. The comparator should receive reasonable optimisation and the same data, metric, hardware boundary and service requirement. A deliberately weak baseline overstates advantage.

The test should also separate model value from system value. A target may create superior outcomes through proprietary data, workflow integration, feedback loops, distribution or expert operations even when the underlying model is replaceable. Conversely, a leading model score can create limited business value if the product lacks data rights, customer adoption or unit economics.

Comparator availability can change rapidly. The transaction case should include refresh triggers between signing and closing and after close. A material improvement in an accessible alternative may alter forecast pricing, capital requirements or the value of proprietary technology. The agreement and approval process should define how such changes are assessed.

16. Reconcile statistical significance with economic significance

A statistically distinguishable result may be economically immaterial, and an economically serious risk can remain statistically uncertain in a small sample. The diligence team should define the smallest effect that changes customer value, cost, safety, contract performance or valuation. Sample design should aim to resolve that effect where practicable.

Decision analysis can combine effect estimates, uncertainty and consequence. High-volume low-cost errors may be managed through operations, while rare severe failures may require targeted stress tests and controls. The team should avoid converting all dimensions into one opaque score. Quality, safety, latency, cost and human effort should remain visible before any weighting.

Where evidence remains weak, the transaction treatment can reflect uncertainty directly. Price ranges, earn-outs, warranties, covenants, staged investment and closing conditions can allocate risk more transparently than an unsupported probability adjustment. The board should see how the selected mechanism responds to the observed evidence gap.

17. Bridge laboratory quality to production performance

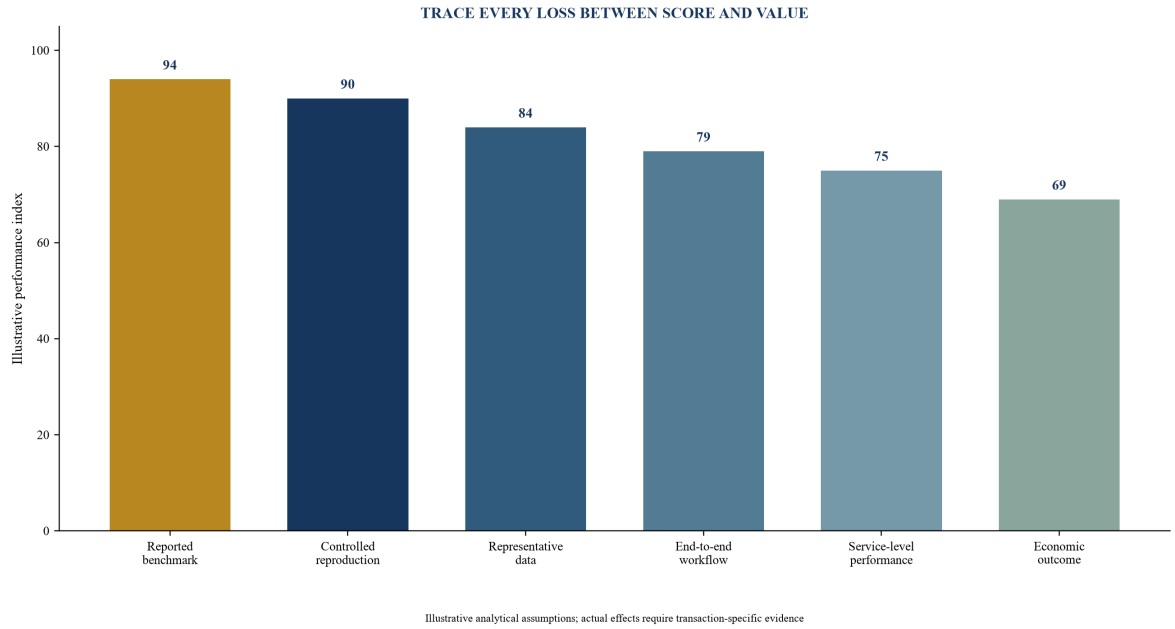
Production introduces traffic mix, imperfect inputs, user adaptation, retrieval drift, tool outages, security controls, monitoring, human escalation and changing models. The diligence team should create a bridge from benchmark quality to end-to-end task success. Each step should identify the evidence, loss, cost and owner.

Shadow testing can replay or mirror representative production cases without allowing the candidate system to control the real outcome. A limited pilot can test live workflow and user behaviour under agreed safeguards. Historical back-tests can add scale, while they may miss

changing distributions and feedback effects. The strongest evidence usually combines methods.

The bridge should reconcile model-level metrics with customer and financial measures: successful tasks, cycle time, rework, escalation, retention, price, service credits, compute and labour. A model can improve accuracy while reducing gross margin through expensive inference or review. A lower headline score can create more value when it is reliable, fast and economical in the target workflow.

Figure 4. Production-gap bridge



Reported benchmark quality is reduced by distribution, workflow and operating effects before it reaches customer and cash outcomes.

Table 4. Production-gap evidence bridge

Layer	Evidence	Typical gap	Transaction consequence
reported benchmark	publication, deck and run record	selection and protocol ambiguity	claim confidence
controlled reproduction	frozen artefacts and locked run	calculation or environment difference	technical value
representative data	customer-like blind set	population and prevalence shift	product fit
workflow	tools, retrieval, humans and failure handling	integration and rework	delivery capacity
service level	latency, uptime, security and support	operational resilience	contract performance
economics	compute, labour, price and retention	margin and cash conversion	valuation and financing

Each layer converts technical performance into an operating and financial consequence.

18. Rebuild inference and support economics

Unit economics should be measured per successful customer outcome. The cost stack can include model calls, accelerator time, retrieval, storage, network, third-party tools, observability, security, retries, human review, support and allocated platform cost. Contractual minimums and reserved capacity can create cost even when utilisation is low.

Quality and cost interact. A smaller model can reduce inference expense but increase errors and review. A larger model can improve first-pass quality while raising latency and vendor concentration. Routing can create value when it sends each task to the least costly system that meets the required threshold. The benchmark should therefore report a quality-cost frontier rather than one score at one configuration.

Forecasts should use observed distributions and capacity constraints. Volume discounts, hardware prices and model efficiencies may improve, while customer mix, context length and service expectations can raise cost. Scenarios should preserve both directions. Unverified future efficiency should remain a management estimate until supported by contracts or demonstrated engineering.

19. Examine benchmark governance and incentives

The buyer should understand who designed, executed, reviewed and approved each evaluation. Researchers, founders, sales teams and advisers can face incentives to present favourable results. Independence is strengthened through separated roles, preregistered rules, raw-output retention and review by people who did not create the claim.

Changes after an unfavourable run deserve particular attention. Some changes correct genuine errors; others narrow the population, replace a metric or exclude failures. The log should preserve the original result, the reason, the approver and the effect. Repeated undocumented reruns or missing raw outputs weaken reliance even when a later score is plausible.

The board should receive material limitations in plain language. A technical appendix cannot compensate for a headline that implies broader performance than the evidence supports. Governance should also define who can make public, customer or transaction representations and how those statements remain aligned with validated results.

20. Assess third-party benchmark and certification evidence

Independent evaluations, audits and certifications can strengthen evidence when their scope, methods, competence and independence fit the claim. The diligence team should obtain the actual report, tested version, period, exclusions and findings. A logo or certificate may cover information-security management while saying little about model quality.

The team should examine access to raw evidence, sampling, reproducibility, conflicts, reliance language and whether the evaluator tested the product or accepted management representations. Standards can support consistent controls but do not establish that a commercial forecast is achievable. Assurance should be mapped to the transaction claim rather than treated as a general badge.

Where the target relies on a benchmark organiser, the buyer should review rules, submission limits, test secrecy, hardware disclosure and audit mechanisms. MLCommons and other

benchmark bodies publish rules for specific programmes. The applicable version and category matter. The diligence report should avoid combining scores generated under different rules.

21. Evaluate regulatory and contractual relevance

AI evaluation obligations vary by role, jurisdiction and use. The European Commission's guidance on general-purpose AI explains documentation and evaluation expectations under the AI Act, including additional duties for models with systemic risk. UK government guidance describes AI assurance as a set of mechanisms supporting evaluation and communication of risks. Sector regulators and customer contracts can impose additional requirements.

The transaction team should map provider, deployer, importer and distributor roles; intended uses; high-risk or safety relevance; and geographic reach. Technical results should link to required documentation, monitoring, incident response and human oversight. Qualified counsel should interpret applicable law and enforcement dates.

Contracts may define acceptance tests, service levels, audit rights, model-change controls, data restrictions, warranties and remedies. The buyer should compare diligence tests with these commitments. A benchmark that excludes the exact cases covered by a customer warranty can create hidden exposure even when the average score is strong.

22. Test security and adversarial performance

An AI product can perform well on ordinary cases and remain vulnerable to prompt injection, data poisoning, model extraction, tool abuse, insecure outputs or denial of service. Security testing should reflect architecture and threat model. The UK National Cyber Security Centre's secure AI development guidance and NIST's AI risk-management resources provide useful control anchors.

The benchmark environment should preserve system boundaries. A sandboxed model-quality test does not establish the security of retrieval, plugins, agents, credentials or deployment infrastructure. Red-team and adversarial results should state scope, access, success criteria, mitigations and residual risk. A demonstration of one attack does not measure all exposure.

Transaction consequences can include remediation cost, delayed deployment, customer notices, insurance constraints, contractual breach and regulatory engagement. The buyer should connect material findings to the closing plan, warranties, escrow, investment budget and operational authority. Security evidence should remain protected and shared on a need-to-know basis.

23. Inspect evaluation code as transaction-critical software

Evaluation code can change outcomes through parsing, answer matching, timeout treatment, randomisation, data loading and aggregation. It should receive code review, tests and version control comparable to other transaction-critical calculations. The reviewer should run known positive and negative cases through the scorer and confirm that failures are represented correctly.

Dependencies can introduce drift. A library update may change tokenisation, numerical behaviour or default parameters. The environment manifest should include package versions and platform details. Seeds improve traceability where systems support determinism, while repeated runs may still be needed for stochastic outputs.

Generated evaluation code can accelerate test creation but should be inspected and validated. Machine assistance can detect duplicated items, inconsistent labels and anomalous score patterns. It cannot establish that a metric represents customer value. Methodological ownership remains with qualified reviewers.

24. Reconcile customer evidence with benchmark evidence

Customer usage, renewals, complaints and realised outcomes provide a different evidence layer. The team should compare benchmark strata with customer cohorts, workflows and support records. Strong customer retention may validate product value despite a modest public score. A leading benchmark may coexist with limited adoption or heavy services work.

References and case studies require verification. The buyer should inspect contracts, invoices, usage logs and agreed outcome measures subject to permissions. Customer interviews should distinguish product capability from founder involvement, custom engineering and price concessions. Concentrated use by a few sophisticated customers may not transfer to a broader market.

The bridge should identify which benchmark metric predicts renewal, expansion, price or cost. Correlation in a small historical sample is limited evidence. The forecast should preserve uncertainty and monitor the relationship after close. Where customer outcome data contradicts the benchmark, the team should investigate measurement, population and workflow before choosing one account.

25. Identify benchmark drift between signing and closing

AI systems, competitors and public benchmarks can change during a transaction. The target may release a new model, switch a third-party provider or alter the product stack. A public test set can become more exposed. The agreement should define permitted ordinary-course changes, notification, evidence preservation and retesting triggers.

The buyer should preserve the signed configuration and monitor the production configuration. A better score from a new version does not remove transition, licence, cost or reliability risk. A lower score may reflect a safety control or cheaper operating point that improves total value. Changes should be assessed across quality, cost, security, compliance and customer impact.

Closing conditions can focus on specific deliverables, access or remediation rather than a broad promise that technology remains satisfactory. Materiality and remedies require legal drafting. The diligence framework supplies the evidence and thresholds for that work.

26. Translate the reproducibility gap into valuation

The reproducibility gap is the difference between the claimed economic case and the case supported by controlled evidence. It can arise from lower quality, higher cost, slower deployment, narrower addressable use, additional remediation or weaker durability. The valuation bridge should identify each mechanism rather than apply an arbitrary haircut.

Revenue can change through conversion, retention, price and eligible customer population. Margin can change through inference, review, support and infrastructure. Capital needs can change through remediation, data acquisition and deployment. Timing affects present value and

financing. Scenario ranges should avoid double counting a single evidence gap across several lines.

The buyer can preserve strategic option value separately from underwritten value. A promising capability may justify continued investment without supporting the same purchase price as a reproduced production result. The board should see both categories and the milestones that can convert option value into underwritten value.

27. Allocate risk through transaction terms

Terms can respond to evidence that cannot be resolved before signing. Contingent consideration can link payment to independently measured customer, performance or margin outcomes. Warranties can address the accuracy of disclosed benchmark records, artefact completeness or absence of undisclosed manual intervention. Covenants can preserve systems and require access for testing.

Metrics for contractual use need precision. The agreement should identify the system, population, test administrator, environment, calculation, uncertainty, exceptions, change control and dispute process. A public leaderboard rank is often too mutable for an earn-out. Operational outcomes such as paid usage, renewal or gross profit may be more durable, though they introduce other dependencies.

Remedies should be proportionate and enforceable. Counsel, tax and accounting advisers should assess treatment. The technical team should provide facts and reproducible methods without drafting beyond its expertise.

28. Protect the financing case

Lenders and credit committees may rely on forecast revenue, margin, cash flow and capital expenditure influenced by AI claims. The financing model should use the evidence-supported case and test delayed or failed reproduction, customer loss, remediation, vendor price changes and additional compute. Debt capacity should remain distinguishable from strategic upside.

Information undertakings can require reporting on material model changes, service failures, regulatory events and customer concentration. Security packages and covenants rarely compensate for a product thesis that has not been evidenced. The buyer should assess liquidity through the period needed to validate and remediate the system.

A strong benchmark can support confidence but does not create cash by itself. The credit case should reconcile test results to contracted revenue, collection, gross margin, working capital and investment. This protects both the transaction and the acquired business from an operating plan that assumes immediate laboratory-to-market conversion.

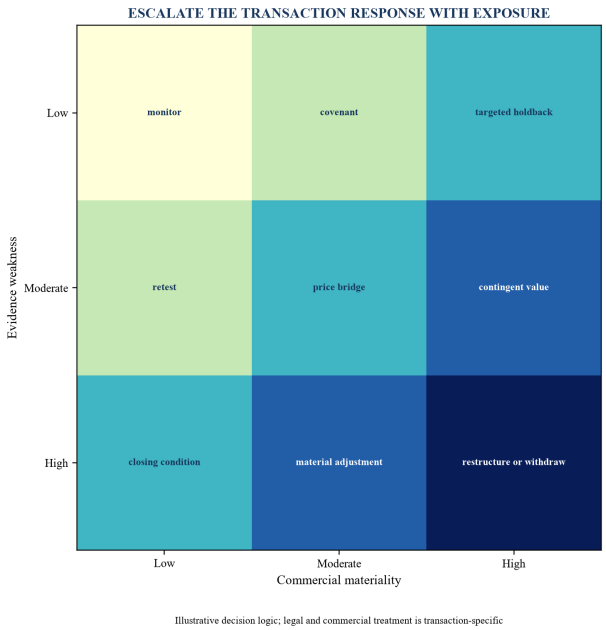
29. Worked example: a hypothetical AI workflow acquisition

Assume a hypothetical buyer evaluates an AI document-workflow company. Management reports a score of 92 on a public task set, median latency of two seconds and gross margin of 78 per cent. The forecast assumes enterprise expansion based on quality leadership. These figures are illustrative assumptions and do not describe a company or transaction.

The evidence graph identifies an undisclosed prompt library, a modified test split and exclusion of timed-out cases. Controlled reproduction yields an illustrative score of 87. A fresh blind set representative of customer documents yields 81, with weaker performance on long and multilingual inputs. Including retries, human review and allocated inference reduces illustrative gross margin to 66 per cent at forecast volume.

The buyer does not treat the result as a single discount. It rebuilds eligible revenue, retention, support capacity, compute and remediation. Under an illustrative scenario, evidence-supported standalone value is 18 per cent below the initial case. The parties could respond through price, contingent consideration linked to paid production outcomes, a remediation covenant and retained technical support. Actual treatment would depend on evidence, negotiation and professional advice.

Figure 5. Illustrative price adjustment matrix



Transaction treatment strengthens as commercial materiality and evidence weakness increase.

Table 5. Transaction response to benchmark evidence

Finding	Value mechanism	Possible diligence response	Possible transaction response
unreproducible score	weaker differentiation or demand	locked rerun and fresh blind set	price bridge or contingent value
contamination	overstated generalisation	remove overlaps and create new set	warranty, covenant or holdback
production degradation	lower adoption or contract performance	shadow trial and cohort tests	milestone payment or remediation
higher cost	lower gross margin and cash	quality-cost frontier and load test	valuation adjustment or funding reserve
missing artefacts	uncertain capability and transfer	evidence recovery and access condition	closing condition or risk allocation
fragile third-party dependency	weaker durability and control	licence, substitution and transition test	consent, covenant or support obligation

The response should address the mechanism through which the gap affects value.

30. Transfer evidence into the integration plan

The benchmark register should survive close. Integration owners need the frozen system, unresolved gaps, remediation actions, customer-like test suites, cost model, monitoring thresholds and authority for model changes. Repeating the diligence test after changes creates continuity between valuation and operations.

The first hundred days can prioritise material evidence gaps and dependencies. Work may include securing data rights, hardening evaluation infrastructure, reducing human review, negotiating provider terms, improving long-context performance or establishing customer outcome measurement. Each action should link to the transaction case and budget.

The buyer should avoid turning a diligence benchmark into a permanent product target without review. Markets, systems and customer needs change. The durable asset is a governed evaluation process that can update tests while preserving lineage and decision history.

31. Monitor post-close drift without rewriting the baseline

Post-close reporting should compare actual performance with the signed evidence case. The team can update forecasts while retaining the original benchmark, assumptions and transaction treatment. This prevents favourable rebasing from obscuring whether the investment thesis was realised.

Monitoring can include data-distribution shift, quality by cohort, safety events, latency, cost, human review, customer outcomes and model or provider changes. Thresholds should trigger investigation and accountable decisions. Statistical alarms need operational interpretation; small fluctuations can be noise, while stable averages can hide cohort deterioration.

Observed differences should improve future diligence. The buyer can record which pre-signing tests predicted production, which gaps mattered financially and which controls reduced uncertainty. Sensitive information should remain governed and anonymised where appropriate.

32. Adapt the method to regulated and high-consequence sectors

Financial services, health, energy, infrastructure and public-sector use can require sector-specific populations, error costs, explainability, human oversight, resilience and documentation. A general benchmark should be supplemented with tests reflecting applicable decisions and users. Qualified specialists should define regulatory and professional requirements.

In credit or insurance, class prevalence, fairness, adverse outcomes and model change can matter. In health, clinical relevance, site variation and human factors may dominate. In industrial systems, sensor drift, environmental conditions and safe fallback can matter more than a public accuracy score. In legal or compliance workflows, citation validity, confidentiality and review burden can determine value.

The common framework remains: identify the claim, preserve artefacts, define the population, reproduce the result, test production conditions and connect evidence to transaction treatment. Sector adaptation changes the tests and consequences, not the need for traceability.

33. Protect confidentiality, privacy and intellectual property

Benchmark diligence can expose model weights, code, customer data, trade secrets and security information. The clean-team protocol should define permitted users, environments, transfers, outputs, retention and destruction. Synthetic or de-identified data can reduce exposure, while it may alter the population and should be validated.

The target should demonstrate rights to datasets, labels, benchmark content and evaluation tools. Open licences can contain attribution, redistribution or use conditions. Customer contracts may restrict secondary testing or transfer on change of control. Counsel should assess the specific rights and remedies.

The buyer should also protect its own test design and strategic thresholds. Disclosure of a sequestered set can destroy future value. Controlled execution, audit logs and compartmentalised access help preserve integrity without preventing appropriate seller review and dispute resolution.

34. Govern machine assistance in the diligence process

Machine assistance can inventory artefacts, compare versions, detect possible overlaps, cluster errors, generate test variations, execute harnesses and prepare sensitivity analysis. Each output should retain source links, code versions and reviewer decisions. Automation is most useful where it improves coverage and traceability.

The diligence system itself can introduce error. Generated test items may be ambiguous, synthetic labels may be wrong, semantic matching may overstate contamination and automated scoring may misread valid outputs. Validation sets, human review and exception logs are required. Material decisions should not depend on an opaque aggregate produced by an untested tool.

Access and confidentiality controls apply to assistance models. Sending target code or customer records to an external service can breach obligations. The team should use approved environments and record providers, retention settings and data flows. Security and legal reviewers should approve sensitive uses.

35. Build the board decision pack around evidence gaps

The board pack should show the claimed result, reproduced result, production bridge, uncertainty, commercial linkage and unresolved limitations. It should identify how much forecast revenue, margin and value depend on each material claim. A concise heat map can link evidence weakness and financial exposure to the proposed decision.

Questions should include whether the tested system transfers, whether data represents contracted use, whether contamination has been bounded, whether alternatives were tested, whether unit economics include all work, and whether the financing case survives downside. Answers should link to preserved artefacts rather than unsupported assurances.

Approval should state which claims support price, which remain optional, which terms allocate risk, what must occur before close and who owns post-close validation. New information should trigger reapproval when it crosses agreed thresholds.

36. Operate a repeatable benchmark-diligence protocol

A repeatable protocol has six stages: claim intake, evidence graph, artefact freeze, preregistered testing, commercial bridge and transaction decision. Each stage has an accountable owner, evidence cut-off, exception process and output. The transaction team can scale depth according to materiality while preserving the same control structure.

Process measures can include claims traced to raw evidence, artefacts missing at intake, leakage findings, reproduced-score differences, production-gap drivers, exceptions after lock, elapsed time and value assumptions changed. These measures indicate diligence quality and workload. They do not prove that a test caused a transaction outcome.

Adoption can begin with the few claims that drive most value. A buyer can freeze the claimed system, reproduce one material test, run a representative blind set and rebuild unit economics before investing in a broad platform. Institutional capability grows through preserved protocols, test suites, reviewers and post-close feedback. The result is a faster route from technical assertion to an accountable price, contract and integration decision.

The protocol should also define a stopping rule. Additional testing has value when it can change price, risk allocation, financing, closing readiness or the integration budget. Testing can stop when the material decision is supported within the agreed uncertainty, or when remaining uncertainty is more efficiently allocated through terms and post-close controls. The diligence chair should document that judgement, the evidence still missing and the person accepting the residual exposure. This prevents open-ended technical work from delaying a transaction without improving the decision.

Repeatability also depends on maintaining a controlled benchmark library. Test sets should carry owners, intended uses, access classifications, exposure history, refresh dates and retirement rules. Public items can remain useful for comparability, while fresh and customer-like sets provide stronger evidence for transfer. The library should record which systems have seen each item so that future teams do not mistake an exposed test for a blind one. This operating discipline protects the value of evaluation assets across transactions.

Conclusion

AI benchmark diligence should treat a reported score as the beginning of an evidence process. A decision-ready claim connects an identified system, appropriate population, controlled protocol, recalculable metric and uncertainty to a commercial assumption. Leakage, selection, execution conditions and production economics must remain visible.

Controlled reproduction can establish whether the reported result survives the disclosed method. Fresh blind testing and production-like evaluation can establish stronger evidence for generalisation and customer value. Neither removes uncertainty beyond the sampled conditions. The board should see the remaining gap and the mechanism through which it affects revenue, margin, capital, timing and risk.

A benchmark evidence graph, preregistered protocol and production bridge give transaction teams a repeatable method. They also allow the parties to allocate unresolved risk through price, contingent value, warranties, conditions, funding and integration actions. Machine assistance can improve coverage and speed within this governed process.

The objective is an acquisition case that can be reconstructed from claim to cash. That discipline supports more credible valuation, financing and post-close accountability in transactions where AI performance is a material part of value.

References

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

[2] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[3] National Institute of Standards and Technology. AI RMF Playbook. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

[4] National Institute of Standards and Technology. AI Technology Evaluation overview. https://pages.nist.gov/ai-technology-evaluation/

[5] National Institute of Standards and Technology. Towards Best Practices for Automated Benchmark Evaluations. https://www.nist.gov/news-events/news/2026/01/towards-best-practices-automated-benchmark-evaluations

[6] National Institute of Standards and Technology. Expanding the AI Evaluation Toolbox with Statistical Models, NIST AI 800-3. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-3.pdf

[7] National Institute of Standards and Technology. Assessing Risks and Impacts of AI, ARIA Program Companion Document. https://ai-challenges.nist.gov/aria/docs/ARIA_Program_Companion_Document_Dec20.pdf

[8] UK Government. Introduction to AI assurance. https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance

[9] UK Government. Portfolio of AI assurance techniques. https://www.gov.uk/guidance/portfolio-of-ai-assurance-techniques

[10] UK Government. Responsible AI Toolkit. https://www.gov.uk/government/collections/responsible-ai-toolkit

[11] UK National Cyber Security Centre. Guidelines for secure AI system development. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

[12] European Commission. General-purpose AI obligations under the AI Act. https://digital-strategy.ec.europa.eu/en/factpages/general-purpose-ai-obligations-under-ai-act

[13] European Commission. Guidelines on obligations for General-Purpose AI providers. https://digital-strategy.ec.europa.eu/en/faqs/guidelines-obligations-general-purpose-ai-providers

[14] European Commission. Guidelines on the scope of obligations for providers of general-purpose AI models. https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act

[15] European Commission. General-Purpose AI Code of Practice. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[16] European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[17] Central Bank of the UAE. Guidance Note on Consumer Protection and Responsible Adoption and Use of Artificial Intelligence. https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence

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

[19] International Organization for Standardization. ISO/IEC 42001 Artificial intelligence management system. https://www.iso.org/standard/81230.html

[20] International Organization for Standardization. ISO/IEC 23894 Guidance on risk management for artificial intelligence. https://www.iso.org/standard/77304.html

[21] MLCommons. MLPerf policies and results. https://mlcommons.org/benchmarks/

[22] Competition and Markets Authority. Merger assessment guidelines. https://www.gov.uk/government/publications/merger-assessment-guidelines

[23] US Department of Justice and Federal Trade Commission. 2023 Merger Guidelines. https://www.justice.gov/atr/2023-merger-guidelines

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

[25] Public Company Accounting Oversight Board. AS 2501: Auditing Accounting Estimates, Including Fair Value Measurements. https://pcaobus.org/oversight/standards/auditing-standards/details/AS2501

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.