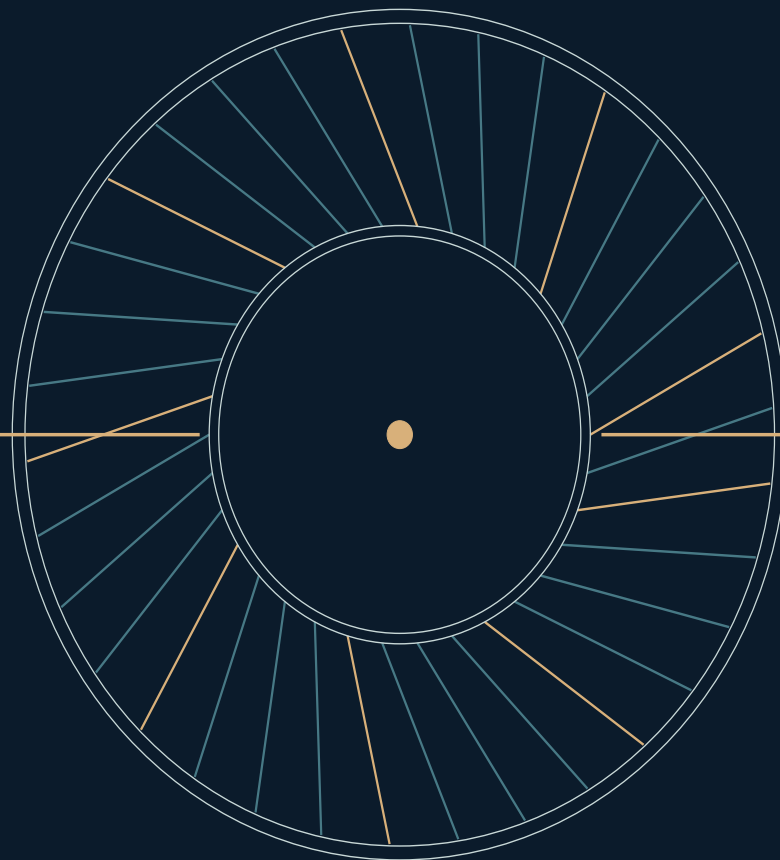


ENGINEERING / OCTOBER 2026

Evaluate the decision before buying the capability

A CEO's guide to measuring whether an AI system deserves operational authority



THE DECISION IN BRIEF

Evaluate the decision the business proposes to delegate. The test earns its value when it changes operating scope, authority or investment - not when it produces a higher model score.

- 01

Write the decision contract before the evaluation: the population served, accepted outcome, cost of failure and authority the system may exercise.
- 02

Check whether benchmark evidence transfers across task, population, scoring and operating conditions. Keep excluded cases and prohibited actions outside reassuring averages.
- 03

Use proportional evidence to choose a release mode. Compare complete alternatives, calibrate automated judges and set a practical restriction or stop rule when the evidence no longer supports the delegated decision.

THE MAYDEN VIEW

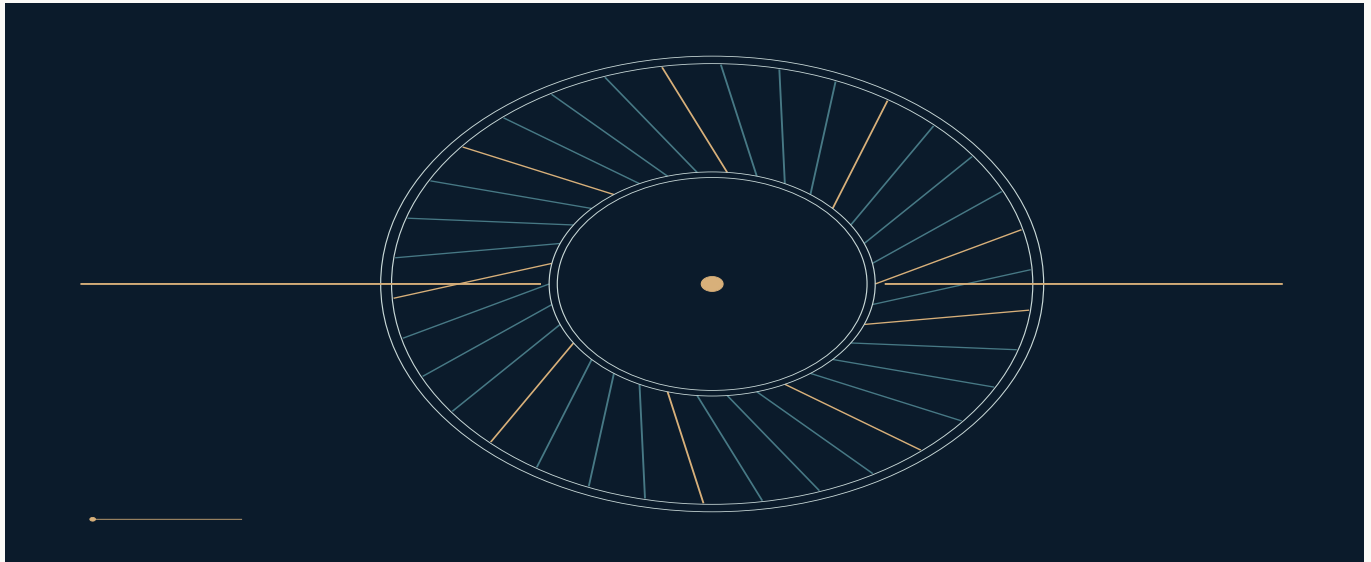
An evaluation is valuable when it changes what the business is prepared to delegate.

INSIDE THIS PERSPECTIVE

The argument at a glance	3
01 / Define the delegated decision	4
02 / Test consequential failures	5-6
Exhibit 01 / A supplier-risk release memo limits what the evidence can authorise	7
03 / Compare useful alternatives	8
Exhibit 02 / A benchmark result crosses four gaps before it supports a decision	9
Exhibit 03 / Three studies show why one evaluation score is insufficient	10
04 / Make the release choice explicit	11
The leadership agenda	12
Method and references	13

ONE DECISION, FOUR CONNECTED QUESTIONS

Evidence opens the gate



EVALUATION APERTURE / CONCEPTUAL ARTWORK

01 Define the delegated decision

Write the decision before testing the model

02 Test consequential failures

Keep failure costs outside the average score

03 Compare useful alternatives

Compare complete outcomes at complete cost

04 Make the release choice explicit

Release supplier prioritisation without granting approval

An evaluation is valuable when it changes what the business is prepared to delegate.

CHAPTER 01

01

Define the delegated decision

Write the decision before testing the model

The same model can summarise a claim or recommend refusing it. The second use carries a different customer consequence, authority and recovery burden. A high benchmark score cannot settle that choice. Before selecting a capability, define the decision the organisation is prepared to delegate and the evidence that would change its mind.

HELM's evaluation across scenarios and multiple dimensions established a useful principle: performance has several dimensions that can conflict. Its historical results are not a ranking of today's systems. The management implication is to request an evaluation profile with explicit trade-offs before selecting an operating role. ^[1] A fast, concise answer that omits a contractual exception may pass a preference test while failing the business decision.

Name the current decision maker, required evidence, escalation conditions and cost of delay. Specify what AI may produce and what it may change. Disagreement about acceptable decisions is a policy or operating issue for leadership to resolve. Discovering it before procurement is a useful evaluation result.

Test the four gaps between benchmark and deployment

mayden proposes a decision-transfer test with four questions: Does the test represent the intended task? Does it represent the intended population? Does its scoring rule reflect business loss? Does the operating environment preserve the conditions under which the result was obtained? This is an original analytical framework, not a validated predictive model. It makes the assumptions behind an automation decision visible and open to challenge.

A 2025 systematic review of 445 language-model benchmarks by 29 reviewers identifies recurring construct-validity weaknesses. The study examines what benchmarks substantiate about their intended concepts; it does not establish that every benchmark is useless. ^[2] For an executive, the practical concern is the distance between the label on a score and the decision that label is used to justify. A test called reasoning may also depend on memorized knowledge and presentation format.

Use a transfer memo alongside each result. Record which deployment cases are represented, which are excluded and who reviewed the correspondence. A contact-center test drawn from short English questions will leave unresolved the performance of a bilingual dispute containing attachments and contradictory records. A missing segment should remain an unknown in the release decision. Converting unknowns into average confidence creates apparent precision while quietly expanding the scope of the evidence.

CHAPTER 02

02

Test consequential failures

Keep failure costs outside the average score

A practical evaluation portfolio starts with ordinary cases drawn from the intended service. Add edge cases whose cost is material even when their frequency is low. Include cases with missing evidence, conflicting evidence and instructions that exceed permitted authority. Preserve an untouched holdout for comparing releases. Keep the people constructing adversarial cases separate from the team tuning the system wherever staffing permits. This separation protects the information value of a test without requiring a large research function.

Model Cards proposes reporting intended uses and performance under relevant conditions. ^[3] mayden's extension is to attach that documentation to the complete deployed service: retrieval sources, system instructions, model version, tool permissions and review workflow. Model-level documentation cannot establish the behavior of a system that surrounds the model with additional evidence and capabilities. A small change in retrieval can alter the decisions presented to a customer without changing the model identifier.

A 2026 study evaluated 14 agentic models on two benchmarks and separated reliability into consistency, robustness, predictability and safety. Higher capability coincided with only modest reliability gains in that setting. ^[8]

mayden's implication is to require separate evidence for repeated execution, changed inputs, recognisable failure and bounded consequences. The study does not supply a universal release threshold: it covers limited benchmarks and scaffolds, and some safety judgements are model-based.

Score errors by consequence and keep prohibited actions visible as individual events. A weighted average may help compare routine quality, but it should never make an unauthorized disclosure disappear behind excellent spelling. Define hard boundaries separately from service-quality targets. Where an absolute claim is impossible, specify the evidence required to proceed, the residual uncertainty and the containment available. The release record should explain why the selected operating mode is acceptable, not merely announce that a target score was exceeded.

CHAPTER 02

02

Test consequential failures

Calibrate the judge against consequential errors

Human review is expensive and inconsistent. Automated judges can make testing more frequent and expand the number of cases examined. Research on MT-Bench and Chatbot Arena found substantial agreement with human preferences in its setting, while also documenting position, verbosity and self-enhancement biases. Preference agreement in those experiments does not demonstrate correctness for financial, clinical or legal decisions. ^[4] Treat a model judge as an instrument whose behavior needs calibration.

Create a human-adjudicated set with explicit reasons for each rating. Ask domain specialists to resolve disagreements and distinguish ambiguous policy from model failure. Compare the judge with this set by failure class. Reverse

answer order, vary response length and include plausible but unsupported answers. A judge that rewards fluent prose can otherwise improve the reported score while the underlying service becomes less trustworthy. Keep some difficult cases hidden from both the generator and judge-development process.

Use deterministic checks wherever the outcome permits them. A submitted total can be reconciled; a required field can be checked; a payment destination can be compared with an approved record. Human judgment remains necessary for many open-ended tasks, but the organization should avoid asking a language model to settle a question that its system of record can answer directly. This division also lowers review cost. Spend scarce specialist attention on meaning, uncertainty and consequence, where it adds the most information.

ILLUSTRATIVE DESIGN · NO MEASURED RESULTS

A supplier-risk release memo limits what the evidence can authorise

Illustrative decision record

Memo field	Evidence required	Conditional release decision
Population and exclusions	Authorised supplier files; missing certificates, conflicting dates, ambiguous entities	Exclude unrepresented languages or case classes from initial scope
Comparator	Same cases through existing specialist process	Compare complete disposition, not draft speed
Error and review burden	Material omissions, unnecessary referrals and total reviewer effort	Narrow class or revise routing if added signals overwhelm review
Disagreement	Specialists retain unresolved policy/evidence cases	Resolve or keep human-led; do not manufacture a gold label
Permitted authority	Rank cases and prepare supported drafts	No approval, rejection or supplier-record mutation
Uncertain estimate	Limited cases or outcomes that have not matured	Restrict scope or obtain more evidence
Unacceptable action	Any action outside the permission boundary	Stop the action path and investigate

THE MANAGEMENT IMPLICATION

Evidence supporting prioritisation does not authorise supplier rejection or approval.

Fictional manufacturer's proposed memo; no measured results or approved deployment. Case counts and confidence requirements depend on consequence; no universal threshold. Source: mayden analysis. Proposed framework or illustrative scenario.

CHAPTER 03

03

Compare useful alternatives

Compare complete outcomes at complete cost

An evaluation budget should include retries, tool calls, retrieval, review and remediation. AI Agents That Matter argues for evaluating cost alongside accuracy and highlights holdout and reproducibility problems in agent benchmarks.

^[5] Its research does not supply a universal business-case multiplier. It does support a discipline: compare complete ways of accomplishing the work under the same constraints, including a simpler workflow or a human-only baseline where appropriate.

Repeat trials on representative cases and record the distribution of outcomes. An average score hides variability; a single attractive demonstration reveals almost nothing about consistency. Track time to an accepted result, total resource cost and the amount of human intervention. Define accepted before running the experiment. Otherwise teams can unconsciously redefine success around the system's strongest behavior. The denominator should include abandoned runs and recoveries, not only the tasks that reach a polished final answer.

Research on long-horizon execution shows why short-task tests can mislead. In the studied settings, execution behavior changed with task length and reasoning approaches mitigated some accumulated-error effects. ^[6] The implication is conditional, not a forecast for an individual enterprise agent. Test the longest operational paths, the recovery path after a mistaken step and the interaction between repeated retries and business state. A retry that improves answer quality may duplicate a real transaction unless the surrounding system prevents it.

Demand enough evidence to decide, not certainty

Excessive testing can become a defense against making a decision. Emerging services lack complete ground truth, and an exhaustive evaluation would be both expensive and obsolete before completion. Small organizations cannot reproduce every major benchmark or employ a permanent panel of specialists. This objection is strong. A policy demanding certainty can preserve a costly human process whose own errors have never been measured. The right comparator is the actual alternative, including its limitations.

mayden's proposed response is proportional evidence. For a reversible drafting assistant, use a modest case set, clear user review and outcome sampling. For consequential external actions, require more extensive testing, enforced authority boundaries and an exercised recovery plan. NIST's Generative AI Profile is voluntary guidance across the lifecycle; it supports structured risk management but is not a certification of any deployment. ^[7] The operating scope should expand as relevant evidence accumulates.

Give each unresolved evaluation question an owner, a method and a restriction where its answer could change release. Time-box work unlikely to alter the operating mode. Fund targeted research where resolving uncertainty changes a material commitment. A dashboard with no consequence for scope, release or investment is administration, not decision support.

MAYDEN FRAMEWORK

A benchmark result crosses four gaps before it supports a decision

Conceptual framework



mayden original decision-transfer framework; no empirical weights or thresholds. Sources: [1], [2], [3]

RESEARCH COMPARISON

Three studies show why one evaluation score is insufficient

Primary-study comparison

HELM; 2023; 30 models	Observed evidence Seven metrics across 16 core scenarios, where feasible	Transfer limit Historical suite; no enterprise outcome estimate
MT-Bench/Arena; 2023; chat preferences	Observed evidence Strong judges exceeded 80% agreement with human preferences	Transfer limit Agreement concerns preference, not universal factual correctness
Construct-validity review; 2025; 445 benchmarks	Observed evidence 29 reviewers identify recurring validity weaknesses	Transfer limit Review diagnoses benchmark claims; it does not invalidate every test
THE MANAGEMENT IMPLICATION Coverage, measurement validity and judging quality are separate sources of uncertainty.		

Abstract-level findings. Studies use different populations and methods; results must not be pooled. Values retain published precision. Sources: [1], [2], [4]

CHAPTER 04

04

Make the release choice explicit

Release supplier prioritisation without granting approval

Consider a fictional manufacturer evaluating an assistant that reads supplier evidence and recommends which files need specialist review. The proposed service can rank cases and prepare a draft explanation; it cannot approve suppliers. Management defines success as identifying relevant issues with enough support for a reviewer to act. The evaluation must also measure the burden of unnecessary referrals. A service that flags every supplier would miss no issue while delivering little operational value.

The team constructs cases from appropriately authorized historical material and adds explicitly synthetic cases covering missing certificates, conflicting dates and ambiguous entity names. Specialists establish the expected treatment and retain disagreements as a separate category. The existing manual process and the assistant are compared on the same cases. Reviewers record time to a defensible disposition, material omissions and the evidence they had to retrieve themselves. No performance figures are assumed here; the example specifies what an experiment should measure.

Exhibit 01 is the proposed release memo, with unresolved evidence left visible. If additional weak signals increase review effort, inspect their consequence and case class before choosing a narrower release, evidence-gathering only or more review capacity. An uncertain estimate calls for further evidence or restricted scope; it is different from an observed action outside authority, which stops that action path. Retain specialist disagreement rather than treating an ambiguous policy as model error. The initial authority remains prioritisation and supported drafting only: neither supplier rejection nor approval is

permitted. No outcome in this illustrative memo is claimed as measured.

Approving low-risk suppliers would require a new decision contract. Evidence supporting prioritisation under human review cannot establish autonomous approval performance. Test the approval errors, authority controls and delayed information that the new role introduces. Keep the successful assistive service in place while that separate case is assessed.

Let the evidence change the operating decision

Put the operating choice on a short release record: scope, current alternative, evidence of improvement, consequential failure classes and the controls that contain them. Show segment results before an average. Name both the owner of customer outcomes and the person who can restrict the service. Keep reproducible technical detail behind that decision.

Set a stop or restriction rule before launch: an unauthorised action, a material error cluster, unexplained quality change or a review queue that cannot protect the service. Exercise the fallback. Sample real outcomes against the release assumptions as users, sources and suppliers change. The approval is a dated judgement about an operating mode.

Fund tests that can alter the decision. Remove tests that do not distinguish useful alternatives; add them when incidents expose a missing failure class. Preserve earlier results under their original conditions. A better model may justify wider authority, a narrower scope or no change at all. The evaluation must establish which conclusion the business can support.

FROM INSIGHT TO ACTION

Decisions to take forward

BUSINESS OWNER

Authorise only the tested decision

Evidence to require: Population, exclusions and current comparator

EVALUATION LEAD

Narrow scope or obtain missing evidence

Evidence to require: Errors, disagreement and reviewer burden

RELEASE OWNER

Stop prohibited actions; restrict uncertain claims

Evidence to require: Permission violations and unresolved uncertainty

mayden connects senior judgement, research and engineering to turn strategic choices into working capabilities.

EVIDENCE YOU CAN INSPECT

Research notes

Method and boundaries

Structured narrative synthesis of eight primary research or official guidance sources, accessed 10 October 2026; original mayden decision framework. No pooled effect estimate.

Selected work on multidimensional evaluation, construct validity, documentation, judges, agent economics, long-horizon behavior and lifecycle risk. Older studies support enduring methods, not current rankings.

No proprietary experiments or client outcomes. Heterogeneous tasks, model generations and scoring methods prevent quantitative meta-analysis. Framework requires local validation.

Primary sources and author research

[1] Holistic Evaluation of Language Models

Liang et al.; TMLR / 2023-10-01. Abstract, methodology overview. Historical multi-model benchmark; its 30 models and 16 core scenarios do not represent current systems or enterprise outcomes. Seven metrics were evaluated where feasible, not for every scenario-model combination.

[Read the source](#)

[2] Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Bean et al.; NeurIPS / 2025-11-03. Abstract. A systematic review of benchmark validity, not an experiment proving that all benchmark results are unusable or predicting a deployment failure rate.

[Read the source](#)

[3] Model Cards for Model Reporting

Mitchell et al.; FAT*2019 / 2019-01-14. Abstract. A documentation proposal; completing a model card does not validate the complete deployed service.

[Read the source](#)

[4] Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

Zheng et al.; NeurIPS / 2023-12-24. Abstract. Agreement concerns human preferences in the studied settings and model generation; it does not establish factual, legal or clinical correctness.

[Read the source](#)

[5] AI Agents That Matter

Kapoor et al.; arXiv / 2024-07-01. Abstract. Research preprint on evaluation design and agent costs; it supplies no universal enterprise return-on-investment multiplier.

[Read the source](#)

[6] The Illusion of Diminishing Returns: Measuring Long Horizon Execution in LLMs

Sinha et al.; ICLR 2026 / 2026-03-13. Abstract, version3. Controlled long-horizon tasks and specified models; findings cannot be extrapolated directly to a business agent or treated as a current model ranking.

[Read the source](#)

[7] Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

NIST / 2024-07-26. Publication abstract. Voluntary lifecycle guidance, not a certification, legal opinion or assurance that a particular deployment is safe.

[Read the source](#)

[8] Towards a Science of AI Agent Reliability

Rabanser et al.; arXiv / 18 February 2026, v1; rendered text marked preprint as of 24 August 2026. Abstract; Sections 3-4; Section 6 limitations. Two benchmarks, scaffold dependence, subjective metric choices, LLM safety judges; operational rather than adversarial robustness. Abstract and conclusion differ on sub-metric count; no count is asserted here.

[Read the source](#)