AIMONGER
WHITE PAPER

AI Discovery

Your Enterprise Knowledge System Is Lying to You

Most enterprise knowledge systems can produce a dashboard that looks green. Fewer can survive an independent question set the team did not rehearse. Where institutional memory sits across markets, languages, legacy systems, and senior staff, the board question is not simply search coverage. It is whether the firm can trust what the system asserts, and prove why.

David Saliba . 2026-07-13 . aimonger.com/whitepapers/enterprise-knowledge-systems-trust/

The lie that looks like progress

Your enterprise knowledge system can look excellent on an internal scorecard and still be unsafe for decisions.

That is not a vendor scandal. It is a measurement problem. A CIO may have policies in English, contracts governed by another jurisdiction, customer records in a legacy system, and local operating knowledge in people's heads. Teams build regression fixtures from questions they already know. The system learns the contours of those fixtures, through prompt iteration, retrieval tuning, and quiet overfitting to the evaluation set. The dashboard stays green. Leadership concludes the system "works." Then a real user asks a question that was never in the fixture pack, or asks about a rule that changed last quarter, and the system answers with confident fluency and the wrong proposition.

What you see globally right now is a rush to make every document repository conversational. US and Asian platform vendors package "ask your data" into suites, while internal teams run demos over curated folders. AI discovery, finding the right policy, clause, prior case, or operating fact at the moment of work, is becoming load-bearing infrastructure. McKinsey's State of AI 2025 survey finds AI use is now mainstream across organisations, while enterprise-level EBIT impact remains concentrated among a small high-performer cohort. Knowledge systems sit in that gap: widely piloted, unevenly trusted, rarely measured with the discipline finance applies to other operational controls.

This paper is a board and CIO briefing on how knowledge systems lie, and how to stop mistaking evaluation theatre for readiness.

> **AI discovery, plainly:** finding authoritative organisational knowledge at the moment of work, policies, contracts, cases, procedures, with answers grounded in sources the firm actually trusts. **In practice:** a Luxembourg fund administrator asks whether a client fee schedule changed after the 2024 amendment; the system must cite the signed version, not the draft still indexed from SharePoint.

> **Evaluation theatre, in short:** dashboards that look green because the team tested questions it already knows, not because the system survives independent scrutiny. **Worked case:** leadership hears "92% accuracy" while a thirteenth paraphrase on retention periods fails in production; the score measured rehearsal, not readiness.

What "AI discovery" means for the enterprise

> **Retrieval-augmented generation (RAG) means:** the model answers from retrieved passages in your corpus instead of guessing from memory alone, search plus synthesis. **On Monday:** a Spanish retail group indexes store procedures and returns answers with chunk IDs; if retrieval fails, the system refuses instead of inventing a policy.

AI discovery is not a science project. It is institutional memory with a modern interface:

- "Have we approved this clause pattern before?" A lawyer used to email three colleagues and wait; the system returns the answer with the signed contract attached, in seconds.
- "What is the current retention rule for this data class?" The compliance team gets the rule from the governing policy with a version stamp, not from the last person who happened to remember it.
- "Which prior incident is closest to this event?" An operations lead finds the analogous case with the postmortem linked, instead of retracing the investigation from scratch.
- "Summarise our position on this customer obligation from the signed agreement and the latest amendment." A relationship manager walks into the renewal meeting with the consolidated position, not a folder of PDFs and a hope.

These are absorption problems as much as model problems. The model is the engine. The corpus, permissions, retrieval path, evaluation regime, and escalation rules are the vehicle. If the vehicle is weak, a stronger engine only produces wrong answers faster.

Gartner projected that by 2026 more than 80 per cent of enterprises would use GenAI APIs or GenAI-enabled applications in production. That includes a wave of "ask your documents" products. Access is no longer the scarce resource. Trustworthy measurement is.

the first useful question is rarely "which model won the benchmark?" It is "which source is authoritative for this market, this customer, this date, and this role?"

Failure mode 1: a perfect regression score is not perfect accuracy

> **Holdout set, put simply:** questions the build team must not tune against, used only for a promotion go/no-go after changes freeze. **Operating case:** promotion validation runs once per release; if engineers peek at failures and rewrite prompts until it passes, the holdout became a second development set.

> **Proposition coverage (plain English):** testing whether each material claim your business depends on is verified, not whether fifty paraphrases of the same claim look impressive. **Operating case:** twelve evaluation questions that all test one retention rule are one proposition family, not twelve independent proofs of readiness.

Distinguish three kinds of exposure that boards and CIOs should not confuse:

Exposure type	What it measures	What it cannot prove alone
Exposed regression fixtures	Known cases still pass after a change	Behaviour on unseen questions
Independent validation	Performance on held-out questions	Long-term drift after promotion
Live shadow evaluation	Behaviour on real traffic without user harm	Perfect labels (labels are expensive)

Exposed fixtures are good for change control. They catch "we broke yesterday's working path." They cannot establish that the system is accurate enough for a new decision class.

A practical rule: never promote a knowledge system to a consequential workflow on regression score alone.

Failure mode 2: fifty questions may test twenty propositions

Evaluation sets often contain near-duplicates. Fifty questions that restate the same twenty propositions create a false sense of coverage. A system can score highly by mastering a small idea cluster.

Board-usable fix:

1. **Group by proposition.** Group questions by proposition (the claim that must be true), not by wording. Propositions are what the business actually depends on, and grouping by surface wording hides the fact that one weak rule can hide behind twelve passing paraphrases.
2. **Report by family.** Report accuracy by proposition family, not only overall percentage. A family-level report forces the team to admit that a critical claim family is red even when the blended score looks green.
3. **Independent paraphrase.** Require at least one independent paraphrase per critical proposition in the holdout set. Without an independent paraphrase, the team has no evidence the system understood the claim rather than memorised its wording.

If the scoreboard cannot show proposition coverage, it is a vanity metric.

Failure mode 3: retrieval is only one layer

Retrieval quality matters. It is not the whole system. A useful measurement map for AI discovery has multiple layers. Boards do not need to operate each layer; they need to know that someone owns them.

Nine-layer measurement map (executive view)

1. **Corpus coverage.** Confirm the relevant material is actually in the system before tuning retrieval, because a perfect ranker cannot find documents that were never ingested or permissioned. Coverage gaps are the silent failure: the system answers with confidence from the wrong source because the right source was never loaded.
2. **Freshness.** Mark or remove superseded documents so "latest PDF wins" does not silently answer historical questions with obsolete rules. Without freshness marking, a 2026 policy overwrites a 2022 question and the error is invisible until audit.
3. **Permissions.** Ensure retrieval respects who is allowed to see what, because a fluent answer that leaks restricted content is worse than a refusal the operator understands. Permission failures do not show up in accuracy scores; they show up in incident reports.
4. **Retrieval.** Verify the right passages are found for the question, since downstream grounding cannot fix a miss at the search layer. A grounding check on the wrong passage produces a cited wrong answer that looks well sourced.

5. **Grounding.** Check that the answer sticks to retrieved evidence rather than fluent prose near a real document, which is how cited answers still mislead boards. Grounding is the difference between "the document supports this" and "the document is nearby."

> **Grounding, briefly:** whether each material claim in an answer is supported by the retrieved evidence, not merely written in fluent prose near a real document. **On Monday:** the system cites the travel policy PDF but omits the exception in the next paragraph; grounding checks catch "cited but unsupported" claims.

6. **Claim validity.** Verify each material claim is true given the evidence, not merely well written beside a real paragraph that does not support the assertion. A claim can be fluent, cited, and still unsupported, which is the most dangerous failure mode for a board.
7. **Temporal validity.** Confirm the answer is correct for the time the question implies, because default "latest" retrieval fails when someone asks what the rule was in 2022. Temporal failure hides behind a current, authentic PDF that was never the right one for the question.
8. **Refusal behaviour.** Require the system to abstain when evidence is weak, so operators prefer a visible refusal over a confident guess on a consequential decision. A system that always answers teaches operators to trust it on the cases where it should not.
9. **Operational safety.** Enforce rate limits, audit logs, secret redaction, and escalation paths so discovery systems fail safely under load rather than leaking or overspending quietly. Safety failures turn a useful assistant into an incident and a surprise invoice at once.

A system can win on layer 4 and fail on layers 6 to 8. That is how "cited" answers still mislead.

Failure mode 4: the citation can be correct and the answer still wrong

> **Provenance:** the reconstructable trail showing which document version, passage, and retrieval path produced an answer, not a decorative link. **Board reading:** an Italian professional firm can show regulators exactly which handbook version supported an HR answer; without version IDs, a real citation can still assert an obsolete rule.

Three common patterns:

1. **Right document family, wrong proposition.** The system cites a policy PDF but asserts a clause from an older version or a different product line, which looks grounded until an auditor compares version IDs. The citation is real; the proposition is not, and only version discipline catches the gap.
2. **Current passage for a historical question.** The system answers "what was the rule in 2022?" with 2026 text because retrieval ranks latest by default, hiding temporal failure behind fluent prose. The answer reads well and is wrong for the date the user asked about.
3. **Selective quotation.** The system retrieves a real paragraph and omits the exception in the next paragraph, which produces a citation theatre that passes a casual read and fails in production. The omission is where the consequential exception lived.

Minimum provenance record for consequential answers:

- **Document ID and version.** Every consequential answer records which document version was retrieved, so regulators and internal audit can reconstruct whether the rule in force was actually the one cited. A citation without a version is a guess about which edition the system read.
- **Passage locator.** Store page, section, or chunk ID for each supporting passage, because "from the handbook" is not provenance when three handbook versions coexist in the index. A locator lets an auditor open the exact paragraph in minutes.

- **Retrieval timestamp.** Log when retrieval ran, so temporal disputes can be resolved without guessing whether the corpus changed between question and answer. A timestamp protects the firm against "the answer was right when we asked, but the archive moved" disputes.
- **Model and prompt version.** Record model and prompt version on each answer, so score movements after a release can be tied to a specific configuration rather than debated as mystery drift. Without this, every regression becomes an argument about what changed.
- **Human review flag.** State whether a human reviewed before action, so consequential workflows can prove propose-and-wait discipline instead of implying autopilot where none existed. The flag is what separates a reviewed answer from an unsupervised one in the audit trail.

If you cannot reconstruct why an answer was given, you do not have a knowledge system. You have a chatbot with a filing cabinet.

Failure mode 5: uncalibrated judges

> **LLM-as-judge, in one line:** using one model to score another model's answers, useful when calibrated, dangerous when it always agrees. **Architecture case:** a trap suite includes "right citation, wrong claim" pairs; if the judge passes them, automated scores are theatre until recalibrated.

Many teams use a model to grade another model's answers. That can work. It can also manufacture confidence.

Before trusting an automated judge:

- **Verifier trap suite.** Build a small verifier suite with deliberately wrong answers the judge must fail, so agreement on easy cases cannot masquerade as calibration on hard ones. A judge that never fails has not been tested; it has been flattered.
- **Citation-claim traps.** Include "right citation, wrong claim" pairs in the trap suite, because judges that reward fluency near a real document will green-light dangerous answers in production. These pairs catch the specific failure mode that causes the most board-visible incidents.
- **Temporal traps.** Include temporal traps where the correct answer depends on document date, since default retrieval bias toward "latest" is a common false pass. A judge that ignores date will pass answers that are right today and wrong for the question asked.
- **Human gold agreement.** Measure agreement with a human gold set on a rotating sample, and recalibrate when agreement drops after model or prompt changes. Agreement that drifts after a release is the signal that the judge needs retraining before its scores mean anything again.

If the judge cannot fail on purpose, it cannot be trusted when it passes.

Failure mode 6: small evals produce large confidence theatre

A pilot with forty questions can show a dramatic percentage swing from a handful of flips. That does not mean the system is production-ready.

Practical pilot protocol:

- **Report counts and intervals.** Report Wilson confidence intervals or at least absolute counts ("37/40"), not only percentages. A percentage on forty questions can move ten points from a single flip, which is noise dressed as a trend.
- **Separate development from promotion.** Separate development set performance from promotion validation performance. Reporting development scores as if they were promotion scores is how a team claims readiness while the holdout stays untouched.

- **Surface unreliable families.** Do not let "any success" hide unreliable behaviour on the critical proposition families (legal, clinical, safety, finance). A green average on easy questions can hide a red family on the questions that actually carry risk.

Eurostat's 2025 figure, 20.0 per cent of EU enterprises with 10+ employees using AI, reminds boards that many peers are earlier in the journey. That is not a reason to lower the measurement bar. It is a reason early movers can set a trust standard competitors will have to match.

The global narrative is adoption speed. The opportunity is different: make the knowledge system trustworthy enough that lawyers, operators, finance, and client teams can use the same memory without inventing parallel shadow processes.

Failure mode 7: a holdout stops being a holdout when you learn from it

Govern evaluation sets as a portfolio:

Set	Purpose	Rule
Development set	Iterate prompts and retrieval	Team may inspect freely
Promotion validation set	Go/no-go for production	Inspect only under change control after scoring
Rotating shadow holdout	Catch drift after launch	Periodically refreshed; not used for tuning
Regression scoreboard	Protect known good paths	Update when workflows change, with review

The moment a team tunes against the promotion set until it passes, it is no longer a holdout. It is a second development set wearing a badge.

Failure mode 8: freeze the right layer or the experiment proves nothing

When you change three things at once, corpus, retrieval, prompt, a score movement is uninterpretable.

Use explicit experiment modes:

- **Frozen-trace replay.** Replay the same retrieved passages with a new generator to test generation changes in isolation, so you know whether a score move came from wording or from search. This mode answers the question "did the new prompt help, or did the retrieval change underneath it?"
- **Frozen-plan, live retrieval.** Keep the query plan shape fixed while swapping the index to test retrieval and indexing changes without confounding the experiment with prompt edits. It isolates the search layer so a retrieval upgrade can be judged on its own merits.
- **Full live.** Run full live configuration only after component tests pass, because changing corpus, retrieval, and prompt together makes a metric swing uninterpretable for the board. Full live is for confirmation, not for diagnosis.

Boards should ask one question in steering meetings: "What exactly was frozen when this metric moved?"

Failure mode 9: security and ops gates are not optional extras

Knowledge systems fail operationally as often as they fail semantically:

- **Secrets and personal data in logs.** Block secrets and personal data from appearing in prompts or logs, because discovery traffic often contains identifiers that become an incident when retained carelessly. A log that captures a national insurance number turns a useful trace into a reportable breach.
- **Cross-tenant leakage.** Prevent cross-tenant leakage in shared indexes, since a retrieval miss that returns another customer's clause is a trust failure no accuracy percentage can undo. The customer

does not care that the system scored 95 per cent; they care that their counterparty saw their contract.

- **Unbounded tool exfiltration.** Stop unbounded tool calls that exfiltrate corpus content to arbitrary destinations, which turns a helpful assistant into an uncontrolled export path. One misconfigured tool grant can move the entire corpus to an endpoint the board never approved.
- **Retrieval rate limits.** Apply rate limits on expensive multi-step retrieval, so a popular demo day cannot silently exhaust the monthly FinOps envelope on nested searches. Without a hard stop, a single busy afternoon produces a surprise invoice.
- **Audit trail for outcomes.** Maintain an audit trail when an answer influences a customer or regulator outcome, so post-incident review can reconstruct prompts, sources, and human overrides. An audit trail is what turns "we think the system did X" into "we can prove what the system did."

Gartner has warned that more than 40 per cent of agentic AI projects will be canceled by the end of 2027 due to escalating costs, unclear business value, or inadequate risk controls. Knowledge assistants that grow into agentic workflows inherit those cancellation risks unless cost and risk controls are designed in early.

A board scorecard for trustworthy AI discovery

Use this as a quarterly control, not a one-off audit.

Control	Green looks like	Red looks like
Ownership	Named business + technical owners	"The AI team" with no business owner
Proposition coverage	Critical claims mapped and tested	One blended accuracy percentage
Holdout integrity	Promotion set untouched during tuning	Score climbed after endless peeking
Provenance	Versioned sources on consequential answers	Answers without reconstructable evidence
Temporal rules	Superseded docs marked; time-scoped questions tested	Latest PDF always wins
Refusal policy	Written abstain rules	Always answers
Human gates	High-risk actions require named review	Autopilot on contractual or regulated steps
Ops security	Logs, permissions, redaction tested	Demo credentials in production
Cost envelope	Token/retrieval budget with alerts	Surprise invoice after a successful demo
Training	Operators trained on failure modes	Launch email only

Staged discovery pilot (ninety days)

Days 1-15: Scope and corpus truth

- **Single workflow scope.** Pick one workflow such as policy Q&A, contract discovery, or case precedent rather than all three, so the pilot can produce a promotion decision inside ninety days. Three workflows in parallel means three incomplete measurement regimes at day ninety.
- **Source-of-truth inventory.** Inventory source-of-truth systems and remove or mark obvious duplicates and superseded PDFs, because polluted corpora produce confident wrong answers that pass casual demos. A corpus audit is the unglamorous work that makes every later metric trustworthy.
- **Critical propositions.** Write ten critical propositions the system must get right, grouped by claim family so the scoreboard cannot hide one weak rule behind twelve paraphrases. Ten propositions is enough to find a real failure without building a scoreboard nobody maintains.

Days 16-45: Measurement before features

- **Separated evaluation sets.** Build development, promotion, and regression sets under change control, so holdout integrity is a process rule rather than a hope on slide seventeen. Change control means a set is a versioned artefact, not a shared spreadsheet anyone can edit.
- **Provenance instrumentation.** Instrument provenance on every answer before feature sprawl, because retrofitting version IDs after launch is how citation theatre becomes permanent. Provenance built in from day one is cheap; added after an incident is expensive and partial.
- **Judge calibration.** Calibrate the judge on a trap suite including wrong-claim pairs, so automated green scores correlate with human risk judgement rather than model politeness. An uncalibrated judge agrees with everything, which feels like success and proves nothing.
- **Frozen baseline.** Establish baseline with a frozen configuration and recorded versions, giving the board a reproducible starting point before iterative tuning begins. A frozen baseline is the anchor that makes "we improved" a verifiable claim.

Days 46-75: Improve one layer at a time

- **Single-layer changes.** Change retrieval or generation per experiment, not both, so score movements can be attributed to a layer the steering group actually understands. Two changes at once produce a number with no causal story.
- **Frozen-layer logging.** Record frozen-layer mode for each run in the experiment log, because "accuracy improved" without freeze metadata is not evidence for investment committee. The log is what lets a future audit reconstruct why a release was promoted.
- **Shadow human review.** Add human review on a sample of live shadow queries when traffic exists, catching temporal and permission failures regression sets miss. Live traffic exposes failure modes that rehearsed fixtures cannot, because real users ask unexpected things.

Days 76-90: Promotion decision

- **Untouched promotion score.** Score the untouched promotion set once changes freeze, treating any peek-and-patch during the window as contamination of the holdout. A promotion score earned after peeking is a development score in a promotion costume.
- **Red scorecard review.** Review red scorecard items with business and risk owners before go-live, documenting accepted residual risk in writing rather than assuming green averages. Accepted residual risk, signed in writing, is how the business owner takes ownership of what the system cannot do.
- **Graduate or close.** Graduate to limited production with human gates, or close the pilot with a written reason, preferring a reversible no-go over an irreversible "platform" commitment. A closed pilot with a reason is discipline; a platform commitment without proof is a debt.

This is reversible by design. Prefer reversible pilots over irreversible "platform" commitments.

What this means for leadership teams

Firms face practical operating complexity before they face any formal compliance checklist: multilingual documents, cross-border customer obligations, sector rules in finance, health, gaming, and professional services, and long-lived archives that outlast the people who understood them. A Nordic bank may hold customer policies in three languages across two regulators; a Malta gaming operator may inherit vendor PDFs from acquisitions with conflicting effective dates. Even when a knowledge assistant is not classified as high-risk, answers that influence customer outcomes, employee decisions, or filings still need an audit story.

AIMonger's view is practical: treat AI discovery as operating capacity. Workshops and team training should teach operators how the system fails, duplicate confidence, temporal errors, citation theatre, not only how to click ask.

Case patterns (synthetic, for board literacy)

These cases are teaching patterns, not claims about a named client.

Case 1: The duplicate confidence trick

An evaluation set contains twelve questions that all test the same retention-period proposition with light paraphrase. The system scores 11/12. Leadership hears “92 per cent.” Proposition coverage is actually one family. A thirteenth paraphrase, worded differently and held out, fails.

Board ask: Show accuracy by proposition family, not only overall percentage.

Case 2: Right citation family, wrong proposition

The system cites the employee handbook PDF. The quote is real. The answer asserts the probation rule from an older handbook version still stored in the index. Provenance without version discipline is theatre.

Board ask: Are superseded documents marked or removed? Do answers carry version IDs?

Case 3: Current passage for a historical question

A user asks what the travel policy was in 2023. The system answers with the 2026 policy because retrieval ranks “latest” by default. Fluency hides temporal failure.

Board ask: Do we test time-scoped questions? Is temporal validity a measured layer?

Case 4: One hundred per cent of the rows that survived

A team drops hard questions from the scoreboard after failures, then reports a perfect score on the remainder. This is selection bias, not quality.

Board ask: Who is allowed to edit the promotion set, and is there an audit log of removals?

Security and operations gates (expanded)

Before production traffic:

1. **Identity.** Attribute every query to a user or service principal, so incident review can answer who asked what rather than treating discovery traffic as anonymous noise. Anonymous traffic makes every incident unresolvable, because there is no way to ask the user what they meant.
2. **Authorisation.** Filter retrieval by the caller’s entitlements, because a correct answer drawn from a document the caller may not see is a confidentiality breach dressed as helpfulness. The system being right about the content does not make it right to show that content to this person.
3. **Egress control.** Block tools from sending corpus content to arbitrary destinations, which closes the path from “helpful search” to unapproved export or shadow integration. An egress allowlist is the difference between a discovery tool and an export tool.
4. **PII and secret scanning.** Scan inputs, outputs, and logs for personal data and secrets, so operators get warned before sensitive content lands in a shared trace or vendor subprocessors. Scanning at the boundary is cheaper than a breach notification after the fact.
5. **Retention policy.** Keep logs long enough for incident review but not forever by accident, balancing audit need against the risk of turning discovery into a permanent sensitive archive. Forever-by-accident retention turns a useful audit trail into a standing liability.
6. **Rate and cost limits.** Enforce hard rate and cost stops rather than email alerts alone, because alerts after budget exhaustion are finance controls in name only. A hard stop is what prevents a demo day from becoming a budget incident.

7. **Incident runbook.** Document who disables the system and who notifies stakeholders, so the first production failure does not invent roles and communication paths under pressure. A runbook written before launch is read; one written during an incident is ignored.
8. **Vendor subprocessors.** Review data residency and training-use terms for subprocessors, because the corpus you trusted to internal search may still flow to a vendor training policy you never approved. Training-use terms are the clause that decides whether your data improves someone else's model.

Gartner's agentic cancellation forecast (cost, unclear value, weak risk controls) applies to knowledge assistants that quietly grow tool use and autonomy. Design the gates while the system is still "just search."

Operator curriculum (ten questions)

Train operators, not only builders, to answer:

1. What sources is this system allowed to use?
2. How do I see provenance for an answer?
3. What do I do if provenance is missing?
4. When must the system refuse?
5. How do I escalate a wrong answer into the evaluation set?
6. Which actions require a human gate?
7. What is the difference between regression green and holdout green?
8. How do I spot a temporal answer error?
9. Who owns corpus freshness for my domain?
10. What is the cost envelope for a heavy discovery session?

If operators cannot answer these, training is incomplete. AIMonger treats workshops as control-plane work: capacity to run the system under real conditions.

Procurement questions for knowledge vendors

1. Can we export our evaluation sets and run them outside your UI?
2. Do answers expose document version and passage locators?
3. How do you prevent tuning against our holdout?
4. What is the permission model for retrieval?
5. What happens to our data for model training?
6. Show a failed answer with full trace, not only successes.
7. What is the unit economics at our expected query volume?
8. How do you handle superseded documents?
9. What isolation exists between tenants?
10. What is the rollback path if a release regresses quality?

Vendors that cannot answer without marketing slides are not ready for consequential workflows.

Linking knowledge trust to absorption capacity

A knowledge system without absorption is a demo portal. Absorption without knowledge trust is faster mistakes.

Leadership should sequence:

1. Pick the workflow (absorption paper).
2. Install measurement and provenance (this paper).
3. Graduate only when both the workflow metric and the trust scorecard are green.

McKinsey's gap between widespread AI use and scarce EBIT impact often lives in this sequencing failure: tools arrive before measurement, and measurement never becomes a management habit.

Appendix: promotion checklist (print and sign)

- ☐ **Business owner named.** A business owner is named with authority to accept residual workflow risk before consequential discovery traffic is promoted. Without a named owner, no one carries the decision when the system fails.
- ☐ **Technical owner named.** A technical owner is named to certify versioned evidence, releases, and rollback paths for each production configuration. The technical owner is who the board calls when a release regresses.
- ☐ **Proposition map reviewed.** The proposition map is reviewed so critical claims are tested by family rather than hidden inside one blended accuracy percentage. A proposition map is what turns one green number into a usable picture of risk.
- ☐ **Evaluation sets separated.** Development, promotion, and regression sets are separated under change control, so tuning cannot quietly consume the holdout meant for go/no-go decisions. Separation under change control is the process rule that keeps a holdout honest.
- ☐ **Promotion set integrity.** The promotion set stayed untouched during the final tuning window, proving the reported score was not manufactured by peek-and-patch iteration. Integrity here is the difference between evidence and rehearsal.
- ☐ **Provenance on consequential answers.** Provenance fields appear on consequential answers, including document version and passage locator, not optional metadata for demos only. Provenance on consequential answers is what makes the system auditable rather than merely conversational.
- ☐ **Temporal tests included.** Temporal tests are included in the validation pack, so "latest wins" retrieval cannot pass promotion while failing historical questions. Temporal tests catch the failure that looks like a correct citation and is not.
- ☐ **Refusal policy written.** A written refusal policy defines when the system must abstain rather than guess, and operators know how to interpret a refusal in workflow context. A refusal policy protects the firm from the system that is too helpful to be safe.
- ☐ **Judge trap suite passed.** The automated judge passed a trap suite including "right citation, wrong claim" pairs, so green automated scores mean something other than agreeable theatre. A trap suite is how the board knows the judge can fail, not just agree.
- ☐ **Permission tests passed.** Permission tests passed for each role that will query production, proving retrieval cannot leak documents the caller is not entitled to see. Permission tests are the control that prevents a useful answer from becoming a breach.

- ☐ **Cost envelope with hard stop.** A cost envelope is configured with a hard stop at expected query volume, so a successful demo cannot become a surprise invoice in month two. A hard stop converts a budget hope into a budget fact.
- ☐ **Incident runbook exists.** An incident runbook exists with named disable authority and stakeholder notification paths before go-live traffic is accepted. A runbook that exists before launch is read; one improvised during an incident is not.
- ☐ **Operator curriculum completed.** The go-live operator cohort completed curriculum on provenance, refusal, and escalation, not only a launch email with a login link. Trained operators are what keep the workflow running after the project team leaves.
- ☐ **Human gates staffed.** Human gates are configured and staffed for high-risk actions, so propose-and-wait is operational behaviour rather than a slide footnote. An unstaffed gate is a gate that will be bypassed the first busy week.
- ☐ **Rollback version identified.** A rollback version is identified before promotion, so a regression can revert without debating which release was last known good under incident pressure. A rollback version turns a regression from a crisis into a procedure.

Unsigned checklists do not graduate systems. They decorate decks.

Appendix: metrics dictionary

Metric	Formula (plain language)	Abuse to avoid
Overall accuracy	Correct / total on a named set	Mixing sets
Proposition accuracy	Correct per proposition family	Hiding weak families in the average
Grounding rate	Answers with adequate evidence / total	Counting decorative citations
Refusal precision	Correct refusals / all refusals	Refusing everything to look safe
Refusal recall	Correct refusals / questions that should refuse	Answering everything to look helpful
Provenance completeness	Answers with version+locator / consequential answers	Optional provenance
Cost per successful answer	Fully loaded cost / successful answers	Ignoring human review time

Publish the dictionary beside the dashboard. Otherwise every green number is negotiable.

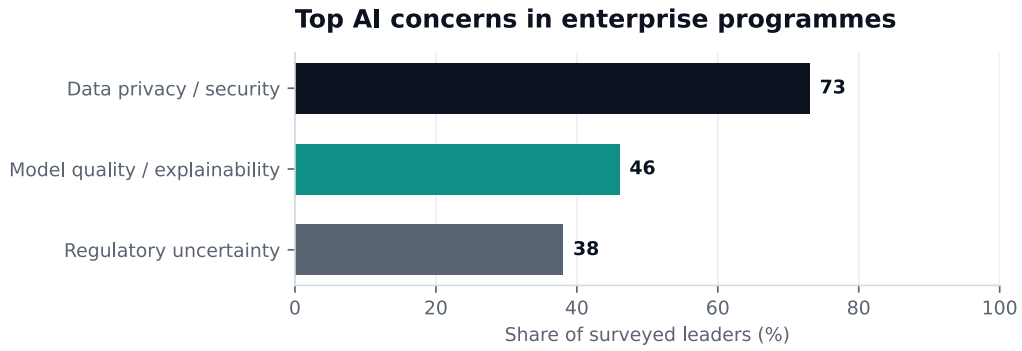
Deepening AI discovery as operating memory

Organisations already pay for institutional memory, in senior time, in repeated research, in inconsistent answers across teams. AI discovery is a bet that memory can be partially mechanised without surrendering accountability.

The bet fails when leaders treat the system as an oracle. The bet works when leaders treat it as indexed memory with brakes: retrieval, provenance, refusal, and human gates. That is closer to how good libraries and good counsel already work, accelerated, not replaced.

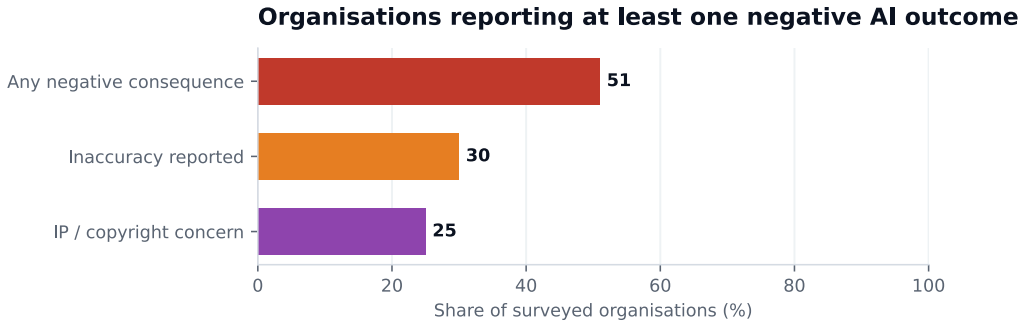
For boards, the strategic question is whether operating memory becomes a managed asset with the same seriousness as financial controls. If yes, fund trust instrumentation early. If no, keep GenAI in low-stakes drafting and do not pretend otherwise.

Evidence base: trust must be engineered before deployment



Deloitte State of AI in the Enterprise 2026 (survey). AIMonger redraw.

Figure 1. Top reported AI concerns among enterprise leaders. Plain read: privacy and quality fears outrank "we need a bigger model." **On the P&L:** budget for logging, permissions, and evaluation before another retrieval upgrade. Source: Deloitte State of AI in the Enterprise 2026 (survey). AIMonger redraw.



McKinsey State of AI 2025 (survey). AIMonger redraw.

Figure 2. Share of organisations reporting negative AI consequences. In short: more than half of surveyed firms already have a failure story; inaccuracy is prominent. **On Monday:** a French insurer treats discovery promotion like a credit control, green only with holdouts, not demos. Source: McKinsey State of AI 2025 (survey). AIMonger redraw.

NIST's Generative AI Profile (NIST AI 600-1, July 2024) organises suggested actions around content provenance, pre-deployment testing, supplier assessment, and incident disclosure. It recommends inventorying GenAI systems with model versions, access modes, data provenance, known issues, human-oversight roles, and sensitive-data considerations. This is a stronger operating baseline than a single answer-accuracy score.

The profile is voluntary US guidance, not an EU compliance certificate. Its value is that it makes the control surface explicit. For deployments under EU rules, pair that operating inventory with applicable EU and sector obligations, plus the organisation's own risk appetite.

Evidence category	Corporate implication
NIST AI 600-1	Inventory, provenance, pre-deployment testing, supplier assessment, and incident processes should exist before consequential use
EU AI Act	Documentation, transparency, and oversight obligations depend on role and risk classification; requirements are use-specific, not triggered by the marketing label "RAG"
Deloitte 2026 survey	Data privacy/security was the top reported AI concern (73%); model quality, consistency, and explainability concerned 46% of respondents
McKinsey 2025 survey	51% reported at least one negative AI consequence, with inaccuracy among the prominent issues reported in the study

Methodology standard for an evaluation report

Every published internal score should state:

- **Evaluation-set provenance.** State evaluation-set origin, size, date, and proposition-family distribution, so readers know whether the score represents broad coverage or one rehearsed cluster. Provenance of the set is the first defence against a score that looks green and tests nothing.
- **Exposure and tuning history.** Disclose whether the set was exposed to builders or used for tuning, because a holdout that was peeked during development is no longer independent evidence. Disclosure turns a hidden compromise into a known limitation.
- **Metric bundle.** Report exact-match, semantic, grounding, refusal, and temporal metrics as applicable, so one green percentage cannot hide failure on the layer that matters for the workflow. A bundle prevents the team from reporting the best-looking number.
- **Small-sample honesty.** Publish confidence intervals or at least absolute counts for small samples, because "92%" on forty questions can flip from a handful of changes. Honesty here protects the board from acting on noise.
- **Human adjudication protocol.** Document human adjudication protocol and inter-rater agreement where relevant, so disputed labels do not become silent score inflation. Inter-rater agreement is how the board knows the labels themselves are trustworthy.
- **Version stamp.** Record model, prompt, retrieval, corpus, and judge versions on every published score, making regressions reproducible instead of argued from memory. A version stamp is what makes a score a piece of evidence rather than an anecdote.
- **Excluded rows.** List excluded rows and reasons, because teams that drop hard failures and report the remainder are reporting selection bias, not quality. Excluded rows are where the hard, honest part of the evaluation often hides.

Without these fields, the score is not reproducible evidence.

Counter-position: formal evaluation can slow useful adoption

It can. The answer is risk-tiered evaluation, not no evaluation. A low-stakes internal drafting assistant should not face the same gate as a system that influences contractual interpretation. The minimum control scales with consequence. The economic objective is not zero error; it is known error, controlled exposure, and a credible escalation path.

Decision rights

The business owner accepts residual workflow risk. The technical owner certifies versioned evidence. Risk/legal determines applicable obligations. Operators own override feedback. No committee may collectively own the final decision: one named executive signs promotion for consequential discovery systems.

Research addendum: separate the scores that boards confuse

Peer-reviewed and evaluation literature converge on a simple rule: do not collapse retrieval, grounding, answer correctness, citation support, temporal validity, and judge reliability into one green percentage.

- **Claim-level citation evaluation.** A citation can be present and still fail to support the claim, which is why claim-level citation evaluation must separate support from fluency near a real document. The discipline exists because cited-but-unsupported answers are the failure mode that most easily fools a board.

- **Calibrated LLM judges.** LLM-as-judge agreement with humans varies by task and position bias is systematic, so calibrate judges with traps, order swaps, versioned prompts, and uncertainty intervals before trusting automation. Calibration is what separates a judge from an agreeable mirror.
- **Temporal QA discipline.** Temporal QA research distinguishes document time, event time, supersession, and recency bias, because "latest" retrieval fails historical questions even when the cited PDF is authentic. The discipline exists because time is the dimension that default retrieval gets silently wrong.

NIST AI 600-1 already points inventories at provenance, versions, access modes, and human oversight. Use that inventory language in promotion packets because it is useful operating evidence, not because a board needs another compliance appendix.

Closing position

A green dashboard is not a trustworthy knowledge system.

As GenAI access becomes common, the organisations that win on AI discovery will be those that measure proposition truth, protect holdouts, require provenance, and refuse to promote evaluation theatre into production decisions. The model will keep improving. The board question remains whether your measurement regime improves with it, across the real documents and operating constraints your firm actually has.

References

1. NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, July 2024. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.600-1.pdf
2. Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj
3. Deloitte AI Institute, "The State of AI in the Enterprise: The Untapped Edge" (2026 edition); survey of 3,235 leaders, Aug-Sep 2025. https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html
4. McKinsey & Company / QuantumBlack, "The State of AI: Global Survey 2025." https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
5. Gartner, "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026," press release, 11 October 2023. https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026
6. Gartner, "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," press release, 25 June 2025. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027
7. Reuters, "Over 40% of agentic AI projects will be scrapped by 2027, Gartner says," 25 June 2025. https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/
8. Eurostat, "20% of EU enterprises use AI technologies," 11 December 2025 (20.0% in 2025; 13.5% in 2024). https://ec.europa.eu/eurostat/en/web/products-eurostat-news/w/ddn-20251211-2

9. Gao et al., “Enabling Large Language Models to Generate Text with Citations” (ALCE citation evaluation), arXiv:2305.14627. https://arxiv.org/abs/2305.14627

10. NIST, “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” NIST AI 600-1 (July 2024). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.600-1.pdf

Published by AIMonger . https://aimonger.com . Malta . Europe