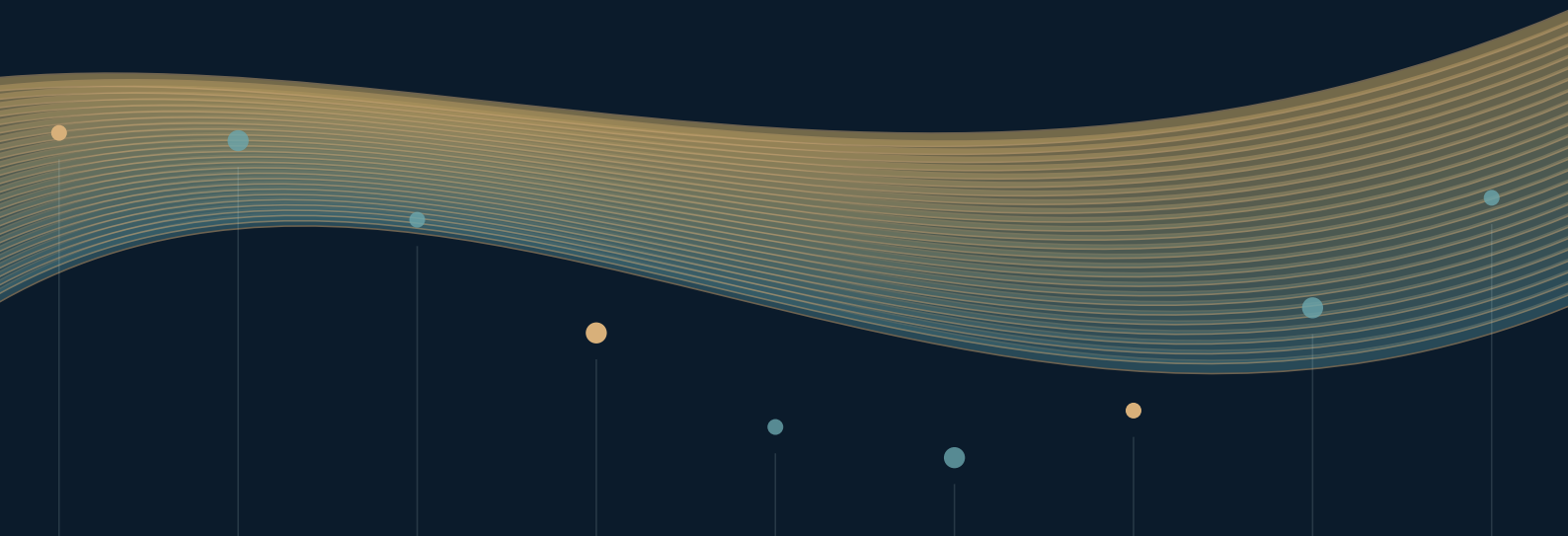


CROSS-SECTOR RESEARCH

# How much should you trust an AI index?

Evidence quality, disclosure bias and what changes when scoring assumptions change.



ORIGINAL MAYDEN GRAPHIC / EVIDENCE IN MOTION

## 01 / THE EXECUTIVE VIEW

# Start with the operating evidence.

Evidence quality, disclosure bias and what changes when scoring assumptions change.

Two frozen comparisons: the historical chapter uses 20 records from PE-2026.10-v1; the current-corpus chapter uses 80 records from PE-2026.10-v3. The accompanying charts and metrics use PE-2026.10-v4 and its 80 selected entity records. The editorial revision retains the v3 coding. These are different research snapshots, not a time series of company performance.

---

**80**

selected entity records

**272**

recorded claims

**224**

eligible claims

**21**

unranked records

## 01 / WHAT THE RECORD SHOWS

59 of 80 selected entity records contain qualifying AI operating evidence in this review. 21 remain unranked because no qualifying claim was located; this does not establish absent capability.

## 02 / WHAT THE RECORD SHOWS

The corpus contains 272 recorded claims, of which 224 meet the declared eligibility rules. Specific commitments, deployed practices and operating detail are coded separately.

## 03 / WHAT THE RECORD SHOWS

Decision Lineage appears in 4 of these 80 profiles, compared with 44 for AI Core. These are counts within the selected research corpus, not enterprise adoption rates.

Public-evidence review. Selected entity scopes. No company participation, population inference or independent assurance implied.

Read the argument, inspect the exhibits, then use the leadership questions to identify the next evidence your organisation should produce. All report bodies are in English; Arabic executive summaries are available online.

## EXHIBIT 01 / PUBLIC EVIDENCE

# Where evidence is visible across the eight sector cohorts

Number of selected entity records with a qualifying claim in each theme; cohort size shown per row

|                            | Core | People | Partners | Interface | Lineage | Testing | Authority | Knowledge | Control |      |
|----------------------------|------|--------|----------|-----------|---------|---------|-----------|-----------|---------|------|
| Aviation & Logistics       | 7    | 3      | 5        | 3         | 1       | 2       | 0         | 2         | 2       | / 10 |
| Banking                    | 7    | 8      | 4        | 7         | 2       | 3       | 2         | 4         | 7       | / 10 |
| Energy & Utilities         | 7    | 6      | 5        | 3         | 0       | 3       | 1         | 3         | 1       | / 10 |
| Healthcare                 | 6    | 3      | 3        | 4         | 0       | 3       | 0         | 1         | 4       | / 10 |
| Manufacturing & Materials  | 5    | 2      | 2        | 1         | 0       | 3       | 0         | 2         | 2       | / 10 |
| Real Estate & Construction | 4    | 1      | 4        | 3         | 0       | 3       | 1         | 1         | 0       | / 10 |
| Retail & Consumer          | 3    | 1      | 3        | 5         | 0       | 0       | 1         | 1         | 2       | / 10 |
| Telecommunications         | 5    | 5      | 3        | 4         | 1       | 2       | 3         | 3         | 3       | / 10 |

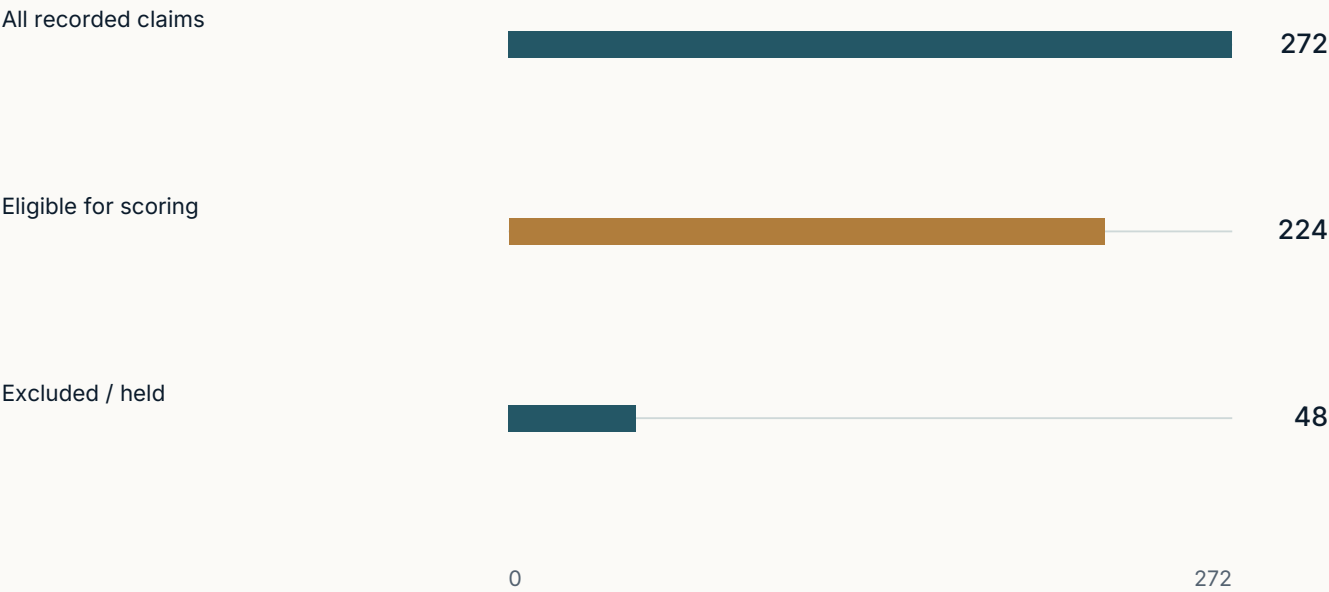
Cells show counts within each named cohort; / n is the cohort denominator.

Source: frozen Mayden public-evidence corpus. Counts describe located evidence, not comparable sector maturity or prevalence. Cohorts are purposeful, heterogeneous and sometimes overlap.

EXHIBIT 02 / PUBLIC EVIDENCE

# The evidence record includes what did not qualify

Recorded claims pass explicit access, scope, deployment and review gates



Source: release claim ledger. These counts do not measure research recall, factual accuracy or company capability. All exclusions remain inspectable in the public dataset.

## 02 / ANALYSIS

Research development note · 10 October 2026

Mayden's ambition is to establish a durable benchmark for how enterprises turn AI into useful, accountable work. The current public corpus contains eighty selected records across eight sectors and five headquarters markets: Saudi Arabia, UAE, Qatar, Kuwait and Bahrain. The frozen first edition (PE-2026.10-v1), with twenty banking and telecom records, remains a historical worked example below. It measures located public evidence. The next stage must establish what the evidence can tell us about an organisation, where it remains incomplete and how a company could demonstrate progress.

A useful benchmark gives executives better questions before it gives them a position. Which workflows have changed? Who uses the systems? What happens when they fail? What decisions may they execute? What evidence shows that the change helped? Public announcements can suggest answers. They rarely supply the complete operating record.

## Start with the object being measured

Three objects need separate treatment: an enterprise's actual operating capability, what its people report about that capability, and what outsiders can locate in public documents. They are related but not interchangeable. A quiet company can have strong capabilities. An articulate company can describe a narrow deployment in impressive detail.

The public index is a designed composite of disclosed evidence. It is not yet a validated psychological or organisational scale. Measurement research distinguishes indicators that define a construct from indicators that reflect an underlying trait; the appropriate tests depend on that choice. High internal consistency would not, by itself, establish that the nine themes measure AI-native enterprise capability. [Bollen and Lennox](#)

The programme therefore separates two linked instruments. The public edition maps documented practices and their evidential limits. A confidential company study investigates operating capability using interviews, artefacts, multiple respondents and, where appropriate, usage records. Neither instrument should silently stand in for the other.

## Keep the number interpretable

The current public calculation takes the highest eligible evidence anchor in each of nine themes and divides the documented points by 27. Expressing the result on a 0-100 scale makes the arithmetic readable. It does not turn it into a percentage of AI nativeness.

Fixed anchors have advantages: another company's new announcement does not mechanically rescale everybody else, and repeated announcements within a theme do not add points. But a maximum has limitations. One detailed practice can fill a theme, even when enterprise-wide coverage is unknown. The arithmetic also treats the difference between a commitment and an implementation as equal to the difference between an implementation and a more detailed implementation account. That is a design assumption.

**Historical worked example: PE-2026.10-v1, twenty records.** A frozen-data sensitivity exercise illustrates the consequence. Across the twenty frozen v1 records, removing commitment-only points while retaining the denominator changes five sector positions. Capping the higher implementation anchor at two points, with the fixed denominator, changes nine. Eighteen checks that individually halve or double one theme's weight change between one and eight positions each. These are deterministic alternative assumptions, including changes in ties. They are not confidence intervals or estimates of the probability that a company belongs in a position.

**Current-corpus comparison: frozen PE-2026.10-v3, eighty records.** Repeating the same deterministic checks changes 16 sector positions

when level-1 points are omitted and 20 when detailed anchors are capped at two. The eighteen individual weight changes alter between 1 and 23 positions each. This calculation includes changed ties and unranked states; it compares each scenario with its own frozen baseline. These counts describe method dependence in the expanded corpus, not company improvement since v1. The fourth-edition editorial revision retains the company coding and baseline scores.

The practical implication is to foreground the nine-theme profile, raw points and evidence coverage, with positions as a secondary view. Methodology research also shows that nominal weights and statistical importance need not coincide when indicators vary and correlate. Equal weights are a transparent starting choice, not a discovery that each theme contributes equally to performance. [Becker and colleagues](#)

## Separate three kinds of coverage

An executive should be able to distinguish research coverage, evidence coverage and operational coverage.

Research coverage describes work completed against a declared search plan. It includes source types checked, languages searched, reporting periods considered and documents retrieved. Evidence coverage describes how many themes have qualifying claims. Operational coverage describes the proportion of relevant people, workflows, transactions or decisions affected by a practice.

Only the last speaks directly to penetration inside the firm. A report discussing all nine themes does not prove enterprise-wide use. Conversely, an inaccessible annual report is a research limitation, not evidence that an operating practice is absent.

The public research-input ledger publishes recorded search routes and access gaps. A future standardised search matrix should add planned-versus-completed routes alongside each company profile. The matrix would make failed retrievals, unresolved scope questions and

under-searched themes visible. It would report completion against the plan without calling that an estimate of the percentage of all relevant evidence found. PRISMA-S offers useful reporting disciplines for documenting searches; adopting those disciplines does not make this corporate review an exhaustive systematic review. [PRISMA-S](#)

## Trace every conclusion to its support

A credible claim needs more than a link. Readers should see the relevant page or passage, the corporate entity it concerns, the operating period, whether the practice is a plan or in use, and what remains unestablished. A group announcement should not automatically become evidence for every subsidiary. A vendor's customer story remains a commercially connected account. A mirrored annual report is another route to the same document, not another independent observation.

The evidence ledger records these distinctions and links each transformation: fetched source, extracted passage, atomic claim, review decision and published contribution. W3C's provenance model provides a useful vocabulary for entities, activities, agents and derivations. Traceability helps assessment; it does not certify truth. [W3C PROV-DM](#)

The public interface exposes contributing claims beside theme results. That view should include qualifying, contradictory and excluded material, with the reason for each decision. "Company-reported" and "independently corroborated" should be explicit labels, never an unexplained confidence percentage.

## Measure work, people and boundaries together

Recent empirical research supports studying AI at the level of tasks, roles and operating conditions. It does not establish one universal enterprise productivity multiplier.

A 2026 observational preprint links ChatGPT Enterprise account and usage records with worker

roles and company information. Its author abstract describes differences in adoption speed, breadth and purpose. The selected customers of one product cannot represent all enterprise AI, and observed usage does not demonstrate causal business impact. The study nevertheless motivates a useful distinction between available accounts, active use, tasks performed and roles reached.

[Chatterji and colleagues, version 1](#)

Another preprint describes randomized field experiments across seven online-retail workflows, with effects varying by application, including some with no effect. Its author abstract supports a workflow-specific evaluation question, not a transferable enterprise-wide return estimate. The inspected version concerns one platform and sales outcomes. [Fang and colleagues, version 6](#)

For Mayden, the corresponding interview should ask for a concrete example, a relevant denominator and an observable boundary. “We rolled out AI” becomes: which process, to whom, over what period, how frequently, with which review or fallback, compared with what baseline? The answers should remain separate from unrelated marketing claims or future forecasts.

## Test the questionnaire before making it authoritative

A company instrument must be understandable to the people answering it. That requires more than expert approval of attractive terminology. An executive, an operations manager and a technology owner may interpret “autonomy”, “deployment” or “governance” differently.

Content-validity guidance in health measurement distinguishes relevance, comprehensiveness and comprehensibility, including direct testing with intended respondents. These principles can inform organisational questionnaire development, while the clinical thresholds and grading rules require adaptation. [COSMIN content-validity manual](#)

The next company pilot should begin with English and Arabic cognitive interviews in the UAE and Saudi Arabia, across more than one sector and respondent role. Participants should explain how

they interpret each question and how they chose an answer. Unknown, no and not applicable need distinct response options. A limited feasibility pilot can reveal misunderstood questions, onerous evidence requests and missing topics. It cannot establish population norms.

Subsequent validation should examine respondent agreement, stability over an appropriate interval, relationships with independently measured operating practices and differences across language, sector and company scope. The required sample depends on those analyses and the desired precision. There is no defensible universal pilot size that makes the instrument scientific.

## Preserve comparability as AI changes

A durable index needs stable definitions and an honest account of revisions. The U.S. Census Bureau's 2026 business-AI analysis explains that its survey question broadened in November 2025 from production-related use to use in any business function. Such changes demonstrate why a trend can reflect question design as well as business behaviour. U.S. survey estimates are not GCC adoption norms. [U.S. Census Bureau](#)

The programme needs a stable core and version emerging-practice modules separately. Each update should distinguish newly located evidence, corrected coding, changed organisational scope, revised methodology and documented operational change. Historical releases should remain reproducible. A better search next year must not be described as proof that the company improved.

This development note is a scoped methodological review with selected empirical examples. It is not a systematic literature review, new meta-analysis or validation of Mayden's instrument. Reporting guidance such as PRISMA provides useful transparency standards; using the name is not a substitute for executing and documenting the corresponding research design. [PRISMA 2020](#)

The near-term publication can be valuable on these terms: a transparent account of what selected companies disclose, where evidence is strong

enough to inspect, and which important questions remain unanswered. A stronger enterprise benchmark will follow from the quality of those questions, the evidence collected against them and the discipline with which Mayden separates observation from inference.

**Review transparency.** Extraction and source challenge in this edition were AI-assisted. Separate AI readers do not constitute human expert sign-off or independent assurance. Company claims, research findings and Mayden interpretations retain their distinct scope.

EXHIBIT 03 / INSPECT THE ASSUMPTION

# Operating detail deserves a separate test

Eligible claim records, n = 224; disclosure anchors are not maturity stages



Source: reviewed coding. Detail is not independent assurance, and multiple claim records may describe a shared underlying practice.

EXHIBIT 04 / INSPECT THE ASSUMPTION

# Build a benchmark from evidence to a decision

Mayden research architecture; conceptual framework, not an observed causal model

|    |           |                                                    |
|----|-----------|----------------------------------------------------|
| 01 | Define    | Entity, workflow, population and period.           |
| 02 | Observe   | Sources, tasks, usage and operating boundaries.    |
| 03 | Challenge | Alternative explanations, controls and exclusions. |
| 04 | Calculate | Declared rules, sensitivity and reproducibility.   |
| 05 | Decide    | Specific action, evidence gap and next test.       |

Source: Mayden methodological synthesis. Confidential company research remains a separate instrument requiring permission, respondent testing and validation.

## 04 / THE LEADERSHIP AGENDA

# Ask for evidence that changes the decision.

## 01 / Choose one recurring workflow.

Evidence to request: Its users, decisions, frequency, business importance and operating boundary.

## 02 / Make authority and evidence inspectable.

Evidence to request: Input provenance, evaluations, interventions, accountable owner and recovery path.

## 03 / Track value with a credible comparator.

Evidence to request: Quality, cost, risk and adoption denominators over a stated period.

## Method and boundaries

This is an AI-assisted, bounded public-source review using a purposeful corpus. Nine themes retain their highest eligible anchor: 1 for a specific commitment or pilot, 2 for an implemented practice, 3 for implemented practice with operating detail. U means unknown. Points equal 100 times the sum of anchors divided by 27. Ordinal steps and equal weights are design assumptions, not validated importance estimates.

Reported outcomes, actual company capability and disclosure depth are different objects. Source-linked corporate or vendor statements are not independently assured operations. Deterministic sensitivity tests are not confidence intervals. Sector scopes, reporting periods and access differ; no national rankings, population prevalence or causal inference are made. The confidential company instrument remains a separate development track.

Dated sources run from 2019-05-14 to 2026-10-01. 79 of 224 eligible claim records have undated sources. Publication dates do not confirm continuing operation or a common observation period; undated does not mean stale.

[Inspect the full method](#) / [Reproduce the evidence](#)

## 05 / SOURCES &amp; PROVENANCE

## Follow the argument back to its source.

Sources below include eligible and excluded research leads. Inclusion in this list does not make a source eligible or independently verified. Exact claim decisions, source locators and limitations are retained in the online evidence ledger. Research cutoff: 10 October 2026.

### S01 / Conventional Wisdom on Measurement: A Structural Equation Perspective

Psychological Bulletin / 1991-09 /  
peer-reviewed-methodological

Locator: Conceptual measurement-model analysis

[Open source](#)

### S02 / Weights and importance in composite indicators: Closing the gap

Ecological Indicators / European Commission JRC / 2017-09 /  
peer-reviewed-methodological

Locator: Methodological analysis with three case studies

[Open source](#)

### S03 / PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews

Systematic Reviews / 2021-01-26 / reporting-guideline

Locator: Reporting-guideline development

[Open source](#)

### S04 / PROV-DM: The PROV Data Model

W3C / 2013-04-30 / technical-standard

Locator: Normative data-model specification

[Open source](#)

### S05 / How Organizations Use AI: Evidence from ChatGPT

arXiv author preprint / 2026-08-12 /  
empirical-preprint-observational

Locator: Observational linked enterprise-account and usage analysis

[Open source](#)

### S06 / Generative AI and Sales Productivity: Field Experiments in Online Retail

arXiv author preprint / 2026-06-29 /  
empirical-preprint-experimental

Locator: Series of randomized field experiments

[Open source](#)

### S07 / COSMIN methodology for assessing the content validity of PROMs: User manual

COSMIN / 2018-02 / methodological-guideline

Locator: Methodological manual

[Open source](#)

### S08 / Large Firms With at Least 20 Employees Biggest AI Users

U.S. Census Bureau / 2026-05-26 / original-survey-analysis

Locator: Original statistical-agency survey analysis

[Open source](#)

### S09 / The PRISMA 2020 statement: an updated guideline for reporting systematic reviews

The BMJ / 2021-03-29 / reporting-guideline

Locator: Reporting-guideline development

[Open source](#)

## About Mayden

Mayden connects senior judgement, research and engineering to help organisations move from strategic choices to working capabilities. This research programme examines how AI changes recurring work, decisions and accountability.

[Explore the Index](#) / [Submit a correction or discuss the evidence](#)