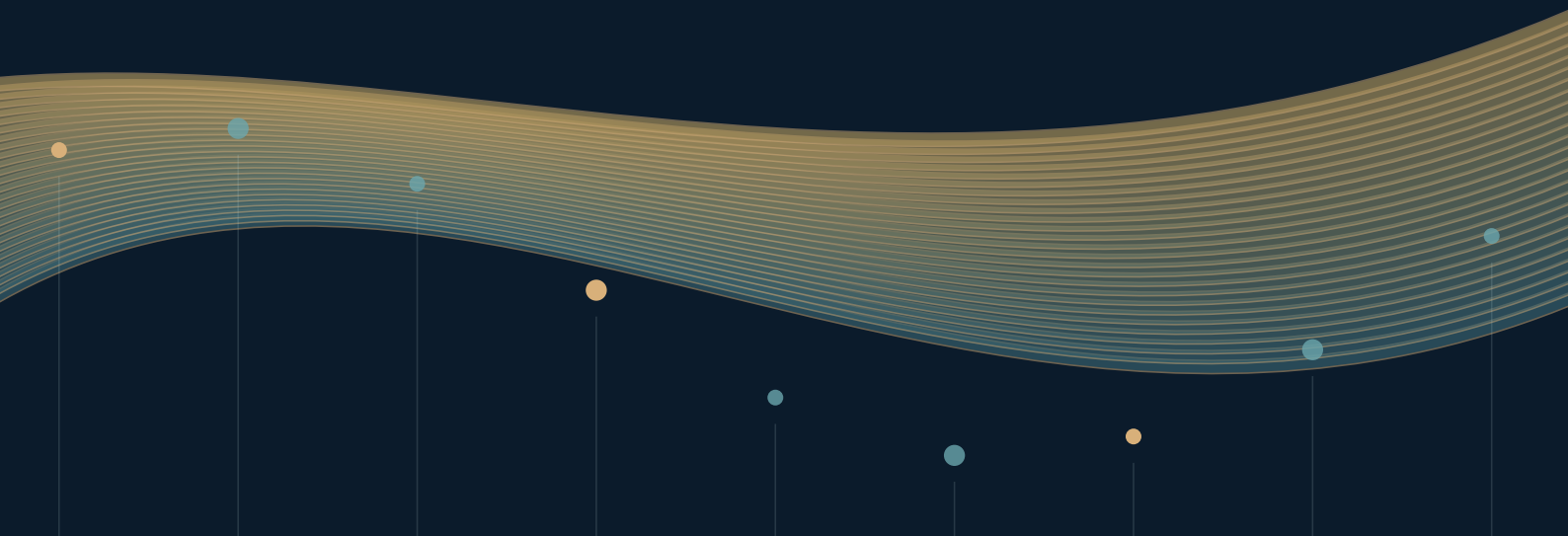




CROSS-SECTOR RESEARCH

The operating evidence atlas

Eight sectors. Five markets. A closer examination of how AI enters recurring work.



ORIGINAL MAYDEN GRAPHIC / EVIDENCE IN MOTION

01 / THE EXECUTIVE VIEW

Start with the operating evidence.

Eight sectors. Five markets. A closer examination of how AI enters recurring work.

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

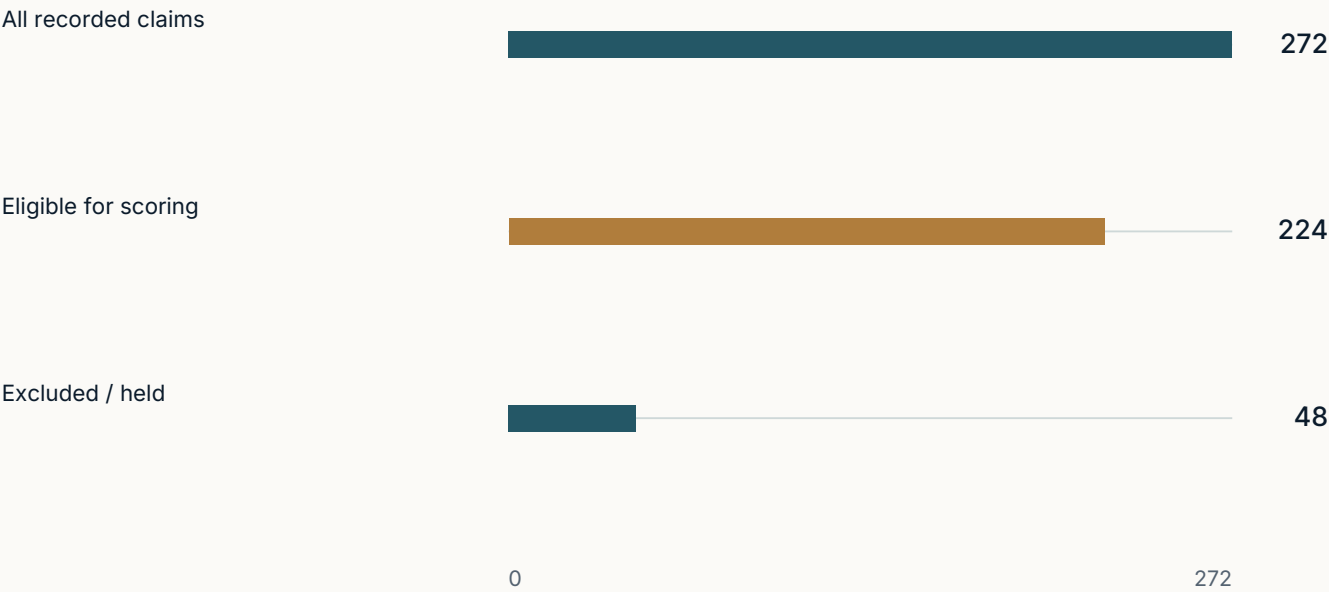
Cells show counts within each named cohort; / n is the cohort denominator.

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

Follow the decision. Measure the operating change.

Mayden research perspective for the eight-sector evidence edition · October 2026

An enterprise becomes meaningfully different when AI changes how work moves, who makes decisions and how the organisation learns from results. Buying access to a model can begin that change. A convincing demonstration can make it imaginable. Neither tells a board whether the business now operates differently.

Mayden proposes a practical definition: an AI-native enterprise repeatedly combines AI, people, data and decision authority inside its operating processes, with measurable results and a working capacity to correct failure. This is a research proposition to test,

not a validated threshold that companies have already crossed.

The immediate executive question is concrete: **Which important decision works differently today, and what evidence would persuade someone outside the project team?** Answering it connects technology investment to business operations. It also exposes the difference between a useful isolated application and an organisational capability that survives a change of model, supplier or leadership.

This edition approaches that question across banking, telecommunications, energy and utilities, logistics and aviation, manufacturing and materials, retail and consumer businesses, healthcare, and real estate and construction. Its public evidence is a starting point for inquiry. It cannot observe every internal practice, and it does not estimate regional adoption or certify the most capable enterprises.

Exhibit 03 - Five questions between an AI demonstration and an enterprise result

Evidence layer	Executive question	Evidence that moves the discussion forward
Task capability	Can the system do the defined job?	Representative tests, quality criteria and failure cases
Operating use	Does it handle real work?	Production dates, eligible workload and actual usage
Decision authority	What may it decide or change?	Permissions, accountable owner and escalation rules
Business consequence	Did the relevant outcome improve?	Baseline, comparison, full costs and observation period
Organisational learning	Can the enterprise detect and correct deterioration?	Monitoring, incidents, changed procedures and retesting

Mayden analytical framework. These are different evidence questions, not five measured maturity bands.

A board can use this sequence without waiting for a perfect enterprise score. Choose a workflow, ask for the evidence at each layer, and note the first unanswered question. That point often identifies the next useful investment: better data, a clearer owner, broader rollout, or a more credible evaluation.

1. A capable model does not guarantee a capable workflow

The central unit of investigation should be a recurring business process with a clear beginning and end. "Our staff use AI" leaves too much unspecified. "The system proposes a replenishment order; a planner checks exceptions; an approved order updates inventory" describes work that can be inspected.

Research illustrates why this distinction matters. In a randomised study of 50 US physicians, chatbot access did not significantly improve the primary diagnostic-reasoning score on six clinical vignettes. The study concerns a bounded test setting, not patient outcomes, but it shows why performance of the combined workflow deserves its own evaluation. [Goh and colleagues, JAMA Network Open](#).

Task benchmarks answer another useful question. GDPval evaluates professional work products

across selected occupations, while explicitly limiting claims about tacit context and interactive work. It offers ideas for realistic test materials; it does not measure whether a named enterprise has integrated AI successfully. [GDPval, version 1](#).

For an executive, the operating test has four parts. Identify the eligible work, observe the human and system together, include the exceptions, and examine the final accepted output. An impressive draft that requires extensive repair is different from an accepted result. A fast decision that creates downstream rework may move cost rather than remove it.

The same logic applies to governance. A policy is useful evidence of intent. An operating control has a trigger, a responsible person and a consequence. The relevant question becomes: **Show one occasion when this control changed what the system or an employee did.**

2. Eight sectors need different proof

A common measurement backbone makes comparisons understandable. Sector-specific operating questions make them useful. A bank and an industrial operator may both need reliable data, accountable owners and trained staff; their material decisions and failure costs differ.

Exhibit 04 - Where to begin the operating interview

Sector	Start with this workflow	Ask for this denominator	Examine this exception
Banking	Fraud alert investigation or service resolution	Eligible alerts or cases, by type	Legitimate customer incorrectly blocked
Telecommunications	Network fault detection and resolution	Covered sites, incidents or service lines	Escalation during a severe outage
Energy and utilities	Equipment maintenance or process optimisation	Eligible assets and operating hours	Recommendation outside safe operating limits
Logistics and aviation	Routing, disruption response or cargo handling	Eligible journeys, shipments or disruption cases	Late data or an infeasible alternative
Manufacturing and materials	Defect inspection or production scheduling	Products, lines and inspected units	Unseen defect or changed material
Retail and consumer	Forecasting, replenishment or returns	Eligible products, stores or transactions	Promotion, stockout or disputed refund
Healthcare	Documentation, scheduling or clinical support	Encounters and sites within the approved scope	Missing information or disputed recommendation
Real estate and construction	Design review, site monitoring or asset operations	Projects, buildings or inspected work packages	Conflicting design versions or unsafe site condition

Illustrative evaluation agenda developed by Mayden; no row asserts a measured sector result or recommends autonomous clinical or safety decisions.

Begin with one process that matters economically or operationally, then ask whether the practice travels. A deployment at one flagship site may be valuable without being enterprise-wide. A parent group may announce a programme that reaches only one subsidiary. A technology service sold to customers may demonstrate commercial capability without showing the supplier's own internal transformation.

Those distinctions should appear beside company findings. Otherwise, readers can mistake a broad corporate name for a broad deployment.

Industry surveys also require careful scope. RICS separates pilot activity from regular process use in its construction survey; its respondents are professionals, not a census of enterprises. Our local

evidence collection needs similarly precise units. [RICS construction research](#).

3. Measure the value that survives the whole workflow

Time saved is an intermediate result. It can improve response speed, increase output, reduce strain or create room for better work. Translating it into financial value requires another observation: what happened to the released capacity?

In its Copilot evaluation, the UK Department for Work and Pensions reports user benefits alongside verification overhead. The evaluation relies on self-reported outcomes, nonrandom licence allocation and no pre-trial baseline. Its

methodological candour is a useful model for separating a promising experience from a causal estimate. [DWP evaluation](#).

A Danish study linking adoption surveys to administrative records likewise distinguishes

changes in work from recorded labour-market outcomes. Its March 2026 revision finds substantial task reorganisation alongside earnings and hours estimates near zero over the observed period. Different outcome layers can move at different speeds. [Humlum and Vestergaard](#).

Exhibit 05 - A value account with separate, auditable lines

Line in the account	What to record
Gross task effect	Time, quality, throughput or loss change on the defined task
Additional work	Review, correction, escalation, integration and training effort
Operating cost	Model use, software, infrastructure, support and monitoring
Capacity use	More output, shorter queues, different work or reduced paid hours
Business outcome	Margin, service, reliability or another chosen outcome
Attribution limits	Demand changes, staffing changes and other concurrent initiatives

Accounting structure, not a forecast. Do not add unlike units or convert every saved minute into cash.

The chief financial officer and process owner should agree on the last two lines before a pilot starts. Choose one primary outcome and a small set of guardrails. Record the baseline and comparison design. Preserve unsuccessful runs. Where randomisation is practical, use it; where it is not, make the identification limits visible.

A programme can be worth continuing before it generates a proven financial return. The investment decision becomes stronger when the organisation can say what it has learned, what remains uncertain and which next observation will resolve the uncertainty.

4. Agentic systems make authority a measurement problem

An agent that drafts a recommendation and an agent that changes a customer record may use the same underlying model. Their operating implications differ. Executives need a map of permitted actions, dependencies and recovery routes.

Provider telemetry reinforces the need to examine deployment context. Anthropic's original agent-use study distinguishes individual API tool calls from complete sessions and acknowledges that observed traffic can include evaluations. Its findings are evidence about one provider's activity, not a count of autonomous enterprises. [Anthropic research](#).

Interactive benchmarks offer another lesson. Tau-bench tests repeated task completion and

notes that a correct final database state can still conceal a policy violation. Success therefore needs more than one indicator. [Tau-bench, version 1](#).

Exhibit 06 - Put scrutiny at the point of consequence

Action	Evidence to request
Read or retrieve	Allowed sources, access boundaries and stale-data handling
Recommend	Acceptance criteria, reviewer competence and rejection records
Change a record	Authorisation, validation and an attributable change log
Commit resources	Limits, approvals, reconciliation and exception ownership
Affect a person or physical process	Domain-specific controls, intervention route and tested recovery

Illustrative action categories; greater autonomy is not automatically a better operating design.

Ask the process owner to walk through one successful case and one failure. Where did the system stop? Who noticed? Could the affected state be restored? Did the team repair only the immediate result, or change the process that produced it?

Evaluation itself needs inspection. The SABER authors identify annotation problems in an established agent benchmark. Their results table also includes cases where the proposed safeguard lowers performance, despite broader positive wording in the narrative. The lesson for Mayden is to retain exceptions and check the underlying comparison. [SABER, version 1, sections 5-8](#).

5. Workforce evidence should describe changed work

Training participation is an input. It does not establish whether someone can judge a difficult output, handle an exception or redesign a process. A workforce assessment should distinguish access, practice, competence and responsibility.

A useful interview asks an operator to describe the last task completed with AI, show how the output was checked and explain what happens when the answer is uncertain. A manager should be able to describe how roles or workload changed. A learning lead should be able to show how the training responds to observed mistakes.

The OECD's original employer and worker surveys demonstrate the value of sector-specific wording and more than one organisational perspective. They also predate mass generative-AI adoption, so their reported attitudes are historical evidence rather than current regional norms. [OECD workplace study](#).

For Mayden's future GCC fieldwork, bilingual interviews should test whether terms such as deployment, approval and active use mean the same thing to respondents. The English label alone is not a stable measure. Interviewers should offer equivalent examples, allow "I do not know", and record whether the respondent speaks for a team, site, business unit or group.

The report back to a company should identify a next action its managers can own: an exception-handling exercise, a better usage denominator, a revised approval route or a repeatable evaluation. That makes participation valuable before the broader benchmark has statistical norms.

6. Read the public index as an evidence map

Mayden's public edition records the strength of located documentation across nine dimensions. The highest accepted anchor in each dimension

contributes to a fixed-denominator score. Unknown evidence contributes no documented points, while remaining explicitly unknown. The result is a comparison of documented evidence in the research corpus.

It is not a percentage of enterprise transformation. A score twice as large does not mean twice the capability. A sophisticated private company may disclose little. A highly communicative group may provide detailed examples from a narrow operating area. Both possibilities matter when reading a ranking.

Exhibit 07 - Three kinds of coverage that must stay separate

Coverage	What it answers	What it cannot establish
Search coverage	Which routes, languages and document types were checked?	That all relevant information was found
Evidence coverage	In how many dimensions was qualifying documentation located?	That the practices cover the whole organisation
Operating coverage	What share of the eligible business uses the practice?	Its quality or causal business impact

Mayden measurement distinction. A source can support one coverage statement without supporting the others.

Readers should be able to open a company result and follow the score to its contributing claim, source and review decision. If a source is a supplier case study, that origin should remain visible. Several pages repeating one announcement do not become several independent confirmations.

Sensitivity views help readers see dependence on choices: remove commitment-only evidence, cap detailed anchors, or vary one dimension's weight. These are alternative assumptions applied to the same corpus. They are not confidence intervals, and stability across them cannot validate the underlying construct.

7. The next research step is a company-level learning loop

The public edition should lead into voluntary, evidence-backed company research. The two products should retain separate identities: public documentation provides an external view; participating organisations can supply operational evidence unavailable to an outside reader.

Start the interview with one material workflow and follow it deeply. Then broaden to the organisational arrangements supporting it: sponsorship, data access, skills, decision rights, evaluation and improvement. Ask for examples before asking respondents to rate themselves.

A useful company report has three parts. First, show the evidence supplied and its scope. Second,

identify the important unanswered questions. Third, propose a small set of operating experiments with owners, measures and review dates. Keep sensitive

operational information out of the public comparison unless the company has authorised its use.

Exhibit 08 - The executive challenge for the next ninety days

Period	Decision to make	Concrete output
First month	Which workflow is important and measurable?	Scope, owner, baseline and eligible-work denominator
Second month	What should change, and what could fail?	Bounded trial, authority map and recorded exceptions
Third month	What evidence justifies continuation or scale?	Outcome account, deployment decision and next evaluation

Proposed operating agenda, not a claim about completed Mayden fieldwork.

A credible AI-native benchmark will be built through repeated observation, challenge and correction. Its value to an executive is the quality of the questions it makes unavoidable: what changed, where it changed, who remains accountable, and how the business knows the change worked.

experiments, surveys, provider telemetry and simulations are labelled at their point of use. They do not supply GCC adoption estimates or validate Mayden's index. Company comparisons belong to the accompanying frozen release and retain their documented scope, eligibility and source limitations.

Research note

This report is an original Mayden interpretation informed by a scoped, AI-assisted evidence review. It combines distinct study designs to develop questions, not to pool effect sizes. International

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 09 / 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 10 / 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 & 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 / Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial

JAMA Network Open / 2024-10-28 / peer-reviewed-randomized-trial

Locator: Stratified randomised, blinded-rating trial

[Open source](#)

S02 / GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks

OpenAI author preprint / 2025-10-05 / provider-original-benchmark-preprint

Locator: Expert-authored task benchmark with pairwise deliverable evaluation

[Open source](#)

S03 / RICS artificial intelligence in construction report 2025

Royal Institution of Chartered Surveyors / 2025 / professional-body-original-survey

Locator: Professional-body online sentiment surveys

[Open source](#)

S04 / An Evaluation of DWP's Microsoft 365 Copilot Trial

UK Department for Work and Pensions / 2026-01-29 / official-original-evaluation

Locator: Nonrandom post-implementation surveys, adjusted regressions and interviews

[Open source](#)

S05 / Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI

NBER working paper 33777 / 2026-03 / empirical-working-paper-observational

Locator: Linked adoption surveys and administrative records; observational difference-in-differences

[Open source](#)

S06 / Measuring AI agent autonomy in practice

Anthropic original research / 2026-02-18 / provider-original-observational

Locator: Provider telemetry with model-assisted classification

[Open source](#)

S07 / tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains

Author preprint / Sierra research / 2024-06-17 / original-benchmark-preprint

Locator: Synthetic interactive agent benchmark with repeated trials

[Open source](#)

S08 / SABER: Small Actions, Big Errors - Safeguarding Mutating Steps in LLM Agents

Amazon AGI Foundations author preprint / 2025-11-26 / original-benchmark-methods-preprint

Locator: Agent benchmark re-audit and safeguard experiments

[Open source](#)

S09 / The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers

OECD Social, Employment and Migration Working Papers / 2023-03-22 / official-original-survey

Locator: Separate employer telephone and worker online-panel surveys

[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](#)