How 498 S&P 500 companies disclose AI value

Methods note. Luke Wilcox, Roiva. November 2026.

This note describes a study of how the largest US public companies describe the financial return on their own AI deployments, in the two places they write to investors: the annual report (10-K) and the quarterly earnings release (the exhibits attached to an 8-K).

Every claim below rests on a document filed with the SEC and retrievable from EDGAR by accession number. Nothing rests on a survey, an interview, or a vendor dataset. A reader who disagrees with any single judgment can pull the filing, read the passage, and re-code it.

Disclosure. I run Roiva, which sells software for measuring the return on AI initiatives. This study is about a problem my company sells into. That is a reason to check the work rather than take it on trust, which is why the full per-company dataset is available on request.

How this was done, and by what. The filings were collected from EDGAR programmatically and the AI-term passages extracted by pattern match. The per-company coding was done by a large language model (Claude) working to the written rubric, under my direction, with me adjudicating disagreements and ruling on every amendment. I independently blind-coded 30 companies myself across two separate blind draws, and the reliability figures in §7 are agreement between that human coding and the model's. The later recall passes used multiple independent model instances, each blind to the existing codes.

Two things follow, and I would rather state them than have them found. First, a study about how companies describe AI value was itself conducted with AI, which is a reason to check it against the filings rather than take it on trust. Second, the human second-coding covers 30 of the 498 companies on the 10-K and 18 on earnings releases. That is the study's thinnest point and it is listed as such in the limitations.

---

1. The question

One question, asked of each company: **does it put a number on a financial benefit from its own use of AI, and state that benefit as achieved rather than expected?**

Two distinctions do most of the work and are easy to lose:

- **Own use, not product sales.** A company that sells AI software reports AI product revenue. That is a return on its customers' spending, not on its own. The two are coded on separate axes and never combined.
- **Achieved, not forecast.** "We saved $40m" and "we expect to save $40m by 2028" are different disclosures. Only the first counts as achieved.

2. Population and documents

Universe. S&P 500 constituents, membership fixed at 30 September 2026. Starting from 503 constituent rows: one dropped for having no resolvable SEC identifier, three

share-class pairs merged into their parents, and one company excluded as a spin-off with no succession filing. **Final population: 498 companies.**

The S&P 500 rather than the Fortune 500, because the Fortune list is ranked by revenue and includes private companies and foreign issuers that file no 10-K — the denominator would contain companies that cannot appear in the numerator.

Annual reports. Each company's most recent 10-K filed on or before 30 September 2026. Fiscal years differ, so each period end is recorded. Items 1 (Business), 1A (Risk Factors) and 7 (MD&A), plus the financial-statement notes. Where Items 1A and 7 are incorporated by reference from an exhibit, the exhibit is followed. Amendments (10-K/A) are excluded.

Earnings releases. For the same 498 companies and the same fiscal year, every 8-K reporting results, identified by SEC Item 2.02, and **every EX-99 exhibit attached** — not just the press release. Investor presentations and supplemental financial data arrive as later exhibits, and that is where quantified program detail often sits.

Earnings-call transcripts are **not** in the study. Commercial transcript providers are licensed and the free aggregators restrict reuse, so the only transcripts included are the handful of companies that file their own prepared remarks as an exhibit. **Nothing here supports a claim about what executives say on calls.** The finding is about the written record.

3. Finding the AI content

Case-sensitive `AI` as a standalone word — an unbounded search hits "said", "chair", "maintain" — plus, case-insensitively: *artificial intelligence, machine learning, generative, large language model, LLM, copilot, agentic, neural network, foundation model, deep learning, natural language processing*.

Three terms were collected but held separately, because they predate the current wave and would inflate every figure: *automation, algorithm, predictive analytics*. This decision was made before coding and it costs the study a case (see §8).

Yield. 10,414 AI passages across the 498 annual reports. 3,962 AI passages across 3,157 earnings releases at 286 companies.

Known gaps in the term list, disclosed rather than quietly fixed. Four spellings the search does not match: bare `ML` (15 occurrences), `GenAI` as one word (6), `A.I.` with periods (2), `LLMs` plural (1). All 24 were read individually; none contains a quantified benefit, so no count changes. They are listed because retro-fitting a term list after seeing the data is exactly the move an amendment log exists to prevent.

4. How claims were coded

A five-rung ladder, applied independently to the internal axis (the company's own AI use) and the product axis (AI-containing products sold to customers). Each company takes its highest rung reached.

Tier	Definition
0	No AI mention anywhere in the document
1	Mentions AI, no benefit claimed
2	Benefit claimed qualitatively, stated as current or achieved
3	Quantified benefit, bundled with non-AI levers or with no stated basis
4	Quantified benefit attributed to AI, with a stated basis or period

Rules fixed before coding: non-GAAP figures count and are flagged; cost avoidance counts as quantified; a benefit attributed to a broader program "including AI" is tier 3, never tier 4. That bundling line, rather than the "stated basis or period" wording in the tier-4 definition, is what separates tier 3 from tier 4. The model second coder, reading that wording strictly, would have moved Amazon, Airbnb and Revvity to tier 3, because none states a measurement period or baseline. The tier 3-or-4 total and the empty dollar cell in §6 are unaffected either way.

One rule was added after the pilot, as amendment A2 (extended to every tier as A2a during the full run): an intention ("seeking to", "expect to", "aims to") is not a claim of achievement at any rung, so a forecast cannot reach tier 3 or 4 however precisely it is quantified.

Fourteen amendments were made to the rubric, each recorded with what it moved. Seven were prompted by a defect someone found rather than by an ambiguity that arose in normal coding — four of those mine, three from the model auditing its own outputs. Some reversed a published figure, including the most consequential correction in the study: I noticed that a draft's thesis and its population paragraph could not both be true, which took the annual-report count of achieved quantified AI benefits from one to zero. The log is available; rubric changes made after seeing data are a standard way to launder a result, so each is listed with its effect rather than folded in silently.

5. The attribution standard

This is the crux, and the place to push hardest.

Co-location is not attribution. A figure counts as attributed to AI only when one of two things is true: the thing the figure is attributed to **names AI in its own definition**, or the text joins the two with a **causal predicate** ("driven by", "enabled by", "reflecting stronger benefits from").

Proximity fails. A filing that reports 3% of annual cost savings from a named program, and discusses AI elsewhere in the document, has not attributed those savings to AI — not if the program's own definition lists five functional areas and none of them is AI.

That example is not hypothetical. It is Quest Diagnostics, and it was the single tier-3 case in the annual reports.

Quest's 10-K states that its Invigorate program "has consistently delivered 3% of annual cost savings and productivity improvements." The filing discusses AI. But Invigorate's own definition in that filing names patient services, logistics, laboratory operations, revenue services, information technology and procurement — and no AI. No phrase anywhere in the document joins AI to the 3%. And both the MD&A and the audited financial-statement note render it "**aims to deliver**" rather than has delivered.

Quest held that tier-3 code through a reliability pass, where a model second coder flagged the "aims to deliver" wording and was overruled. Applying this standard consistently moved Quest to tier 2 and took the annual-report count of achieved quantified AI benefits to **zero**.

The same test removes the study's best near-miss. Hasbro's 10-K states that, through 2025, it delivered "almost $800 million of gross cost savings" for a program called Operational Excellence; an AI term sits 1,266 characters away, in a *growth* strategy paragraph about AI-enabled entertainment rather than in the program's definition. Co-located, not attributed.

The standard cuts against the headline as often as for it, which is the only reason it is worth anything.

6. Results

Internal axis — the company's own AI use, n = 498 in each venue:

Tier	Annual report	Earnings release
0 — no AI mention	15 (3.0%)	212 (42.6%)
1 — mentions AI, no benefit	396 (79.5%)	210 (42.2%)
2 — qualitative benefit	87 (17.5%)	65 (13.1%)
3 — quantified, bundled	**0**	**6** (1.2%)
4 — quantified, AI alone	**0**	**5** (1.0%)

The finding, stated four ways:

	Annual report	Earnings release
Achieved and quantified (tier 3 or 4)	**0** (0.0%)	**11** (2.2%)
Quantified but forecast or target	**1** (0.2%)	**12** (2.4%)
Any quantified AI value figure	**1** (0.2%)	**20** (4.0%)
Achieved dollar figure attributed to AI alone	**0**	**0**

Three things to note about that table:

1. **Zero of 498 annual reports** state an achieved, quantified financial benefit from the company's own AI use. Exactly one attaches a quantified figure of any kind, and it is a forecast (Sysco, ~$100m of expected fiscal 2027 efficiency improvements).

2. **Eleven companies' earnings releases do.** So the annual report is not structurally incapable of carrying such a figure — the same companies in the same fiscal year produce them in the other venue.
3. **The bottom row never moved.** Across both venues and all 498 companies, no company attaches an *achieved dollar figure* to AI alone. The five companies that attribute an achieved figure to AI with nothing else bundled in — Airbnb, Amazon, Bank of America, Broadridge, Revvity — every one of them reports a **percentage**, not a dollar amount.

Costs are a different story. 38 companies quantify what AI *costs* them in at least one venue — capital expenditure, AI-attributed R&D, acquisition prices, restructuring charges — against the one that quantifies a return in an annual report. 26 in annual reports, 19 in releases, 7 in both. This count was populated by reading after an automated screen failed, so it is a **floor, not a ceiling**.

Venue coverage runs the other way, which matters. AI terms appear in 483 of 498 annual reports but only 286 of 498 release sets. 200 companies discuss AI in the annual report and in no release; 3 do the reverse. Releases discuss AI *less* overall. The quantified claims are concentrated in a small number of companies and should not be described as general practice.

7. What was done to try to break the finding

A null result is only worth as much as the effort spent trying to falsify it. Four independent passes, after the population was coded:

Re-read every long passage untruncated. A display defect in the coding materials had cut passages at 300 characters (900 in some reliability materials). On the annual-report side, all 2,275 passages over 300 characters, across 421 companies, were re-read in full by six independent model instances, each blind to the existing codes and to each other. **Zero achieved internal AI benefits found.** On the earnings-release side, where long slide-deck passages made the cut costly, a targeted audit of 107 passages and then a complete re-read of all 1,066 long passages at 212 companies changed several codes, including Bank of America, Broadridge and U.S. Bancorp. The release counts in §6 are counts after that re-read.

Document-wide proximity screen. Across 491 of the 498 annual reports, searched for any achieved savings figure within 1,200 characters of any AI term, ignoring the coding entirely. Two companies surfaced; both fail the attribution standard in §5. With the AI requirement removed, 44 of the 491 state an achieved savings figure; in the four whose named program description includes an AI term, the figure sits 1,281 / 8,469 / 27,538 / 71,644 characters from the nearest AI word.

Benefit-first vocabulary screen. Inverted the logic: find quantified benefits first, then ask whether the mechanism is AI described in other words — *chatbot, service agent, smart routing, computer vision, recommendation engine, digital assistant, RPA*, brand names like *Erica* and *Einstein*, and about fifty more terms across two rounds. No falsifying case in either round.

That screen also produced the most useful failure in the study. Its first version returned a reassuring five hits. Calibrated against nine known AI value claims, it

caught **two of nine** — it split text on full stops, and slide-deck text has no sentences. The rebuilt version catches eight of ten. **A screen that is silent is worthless until it has been calibrated against known positives**, and this one nearly produced a false null.

Image-only exhibits. 88 release exhibits carried under 3,000 characters of extracted text, flagged because investor slide decks are where this study's achieved release figures mostly live. Most are legitimately short documents — non-GAAP reconciliation tables, dividend notices. **Twelve, at six companies, contained nothing but the EDGAR header row**: slides rendered as JPEGs inside an HTML wrapper, with no machine-readable text at all.

One of those twelve changed a finding. A 129-slide investor deck, downloaded and read as images rather than ruled on from OCR, carried as a slide headline "Customized precision displays enabled by AI and automation / 50% reduction in time from concept to commercialization" — moving Hershey from tier 2 to tier 3. The same deck's largest figures, $241m of achieved productivity savings and a $400m transformation program, were inspected and rejected: they name procurement, automation and "technology", with no AI term.

Eleven remain unreadable. Ten of the eleven have a readable sibling exhibit covering the same filing. The one that does not — Linde's Q3 2024 press release — belongs to a company with zero AI-term hits across its six readable exhibits.

Reliability, and what kind of reliability it is. I coded 30 of the 498 companies myself, blind, in two separate draws, and compared my codes against the model's. These are **human-versus-model agreement figures on 30 companies. They are not human inter-coder reliability at scale** and should not be described as such.

The first draw (12 companies, 10-K venue only) gave Cohen's kappa of **0.707** on tier — above the pre-registered 0.6 threshold — and 0.510 on claim type, which did not clear it. Two of the four disagreements behind the lower figure traced to rules that existed only in the model's working protocol rather than in the written rubric — not to inconsistency on my side — and each became a written amendment.

The second draw (18 companies, both venues, every amendment then in force) gave four judgments per company:

Axis	Agreement	Kappa
10-K internal	14 / 18 (78%)	0.507
Release internal	12 / 18 (67%)	0.376
10-K product	11 / 18 (61%)	0.311
Release product	18 / 18 (100%)	1.000
Internal axis, both venues	**26 / 36 (72%)**	**0.553**

That second pass does not clear the 0.6 threshold on the internal axis, and I would rather say so than round it. What it does show is where the agreement sits:

Tier 3/4 boundary — the achieved-and-quantified line every finding in this study rests on — 72 of 72 judgments, 100%. None of the seventeen disagreements

is at that boundary; all are at 0/1 or 1/2. So agreement is complete where the findings live and moderate where they do not.

The disagreements are also one-directional within each group, which is the signature of an unwritten rule rather than coder noise. Six are at the 0/1 boundary in the release venue, and their causes differ. Two (Apple and AES) turn on whether an AI mention unrelated to the company's own operations counts as "mentions AI", a gap now closed by amendment A12, which changed no code because it describes what was already done. The other four (Devon, PPL, Truist and United Rentals) arose because the AI term sat beyond a 900-character display cut in the materials I was given, so my 0 reflects the materials rather than a judgment. On the 14 release cells where the AI content was visible, agreement is 12 of 14 (86%, kappa 0.682); the table above is the unadjusted result, and both are given. Of the rest, four are on the tier 1/2 boundary in the 10-K: two of them on a purpose-clause rule that I read more strictly than the model and that has since been withdrawn as amendment A14, because it contradicted the rubric's own requirement that a benefit be stated as achieved, and two where I under-coded a plain assertion of effect. Seven are on the 10-K product axis, whose kappa of 0.311 is now measured rather than suspected. One of those — Apple's product tier — looks miscoded by the model and is flagged and deliberately left unchanged rather than quietly fixed.

This is the third consecutive reliability exercise to find that the rubric, rather than the coding, is this study's weak component.

A separate pass handed blocks of companies to four independent model instances, each blind to the existing codes and instructed not to open the study's notes. That measures **rubric reproducibility** — whether the rules are written well enough that an independent reader applying them to the same passages reaches the same codes. It is a real test, and it found three errors in the original coding, two of which changed the published results. It is not a second opinion from a second person.

Separately, 68 companies in which no passage had passed the benefit screen were read in full: **68 of 68 confirmed**. A random 30-company out-of-sample audit (model-read, not human) agreed **30 of 30 on the tier 3/4 boundary** and 29 of 30 on tier overall.

8. The closest cases that failed, named

These belong in any fair account of the study. An editor who finds them independently should find them here first.

Hasbro is the best near-miss. Its 10-K states "almost $800 million of gross cost savings" through 2025 for a program called Operational Excellence. An AI term appears 1,266 characters away — in a *growth* strategy paragraph about AI-enabled entertainment, not in the program's definition. Excluded on §5.

PPL is the strongest disclosable objection and the one I would raise if I were arguing against this study. Its earnings release reports $170m of achieved O&M savings against a 2021 baseline, attributed to "smart grid technology, automation and data science". That is a quantified, achieved, internal technology return. It is excluded **only** because *automation* was pre-registered as a pre-wave term before any data was

seen, and *data science* sits on the same side of that line. The release's only AI term is in its forward-looking boilerplate. The figure is not in PPL's 10-K, so the annual-report zero holds either way. Include those terms and the study has a release case; the term list was fixed in advance precisely so that call could not be made after the fact.

Quest Diagnostics held the only tier 3 in the annual reports until a re-read. Its 3% figure attaches to a named program whose definition lists patient services, logistics, laboratory operations, revenue services, IT and procurement — and no AI. The audited note and MD&A both render it "aims to deliver". It is now tier 2, and the annual-report count is zero.

9. The control: do these filings quantify anything?

Yes, which is what makes the AI silence interesting rather than trivial.

A screen for quantified cost savings tied to a named program, with AI not required, returns **more than forty of the 498** annual reports. **None of them names AI in the same sentence.** So the document type plainly carries quantified program returns; AI just is not one of them.

The comparison is not symmetrical, and the asymmetry runs against this study's own standard. AI figures were held to the achieved test — a forecast cannot reach tier 3 or 4, however precisely it is quantified. The control screen was not: it counted targets and realized amounts together. Of the named controls, Merck's ~$1.7bn is "expected to … result in", Stanley Black & Decker's $2.0bn is "designed to achieve", and Bristol Myers Squibb's, Thermo Fisher's, Aon's and Constellation Brands' are all dated to a future year. Only a minority are stated as achieved — Pfizer's ~$500m of R&D savings achieved in 2025, PPG's ~$75m of restructuring savings realized in 2025, and PNC having "exceeded our 2025 continuous improvement program savings goal of $350 million" (a goal exceeded, note, rather than a realized figure disclosed).

One control that looks achieved should be read carefully: Xcel Energy's lean operating program "has generated $1.5 billion of cumulative savings **for our customers**" since 2020. That is ratepayer savings at a regulated utility, disclosed in a customer-affordability section — a real achieved program figure, but not a parallel for a company reducing its own cost base, which is what an internal AI efficiency program would produce. Earlier drafts of this note quoted that sentence short of "for our customers"; the full quote is given here because the truncation changed its meaning.

So "more than forty companies quantify non-AI program savings" should not be read against "zero quantify an achieved AI benefit" as though the two counts were built the same way. **What survives the asymmetry is the part the argument rests on**: some of those programs do state achieved figures, so the document type demonstrably carries one — and none of the forty-plus attributes its figure to AI.

Honesty about the count itself: it comes from two screens with different definitions. The first returns 41 of 498. The second, with a stricter test of "achieved", ran over 491 annual reports and returns 44. Neither finds a figure attributed to AI. I would quote "more than forty" rather than a precise figure, and the definition behind each

is available.

One argument was tested and withdrawn. I expected to show that the same filer quantifies non-AI returns while withholding AI ones. It does not hold: the three testable companies among the four that attribute a restructuring to AI also gave no savings figure for their *earlier, non-AI* programs. They disclose charges and not returns as house style. That claim is not supported and is not made.

What this is and is not. This is a disclosure finding, not a measurement finding. Three explanations were open when only the annual reports had been read: the value is not measured; it is measured and not disclosed; or disclosure convention routes quantified returns to restructuring events and most AI spend is not one.

The earnings-release data narrows that to two. The convention explanation predicts that AI value goes unquantified because AI spending is not the kind of event that triggers a savings disclosure. HP contradicts it directly: HP routed its AI program *through* a restructuring — the exact vehicle that normally carries a savings figure — and the figure appears in the press release while the audited restructuring note omits it. Three other companies did the same thing and also produced no figure in the annual report. The convention cannot explain a number that exists in one document and not the other for the same program in the same quarter.

So two explanations survive and this data cannot separate them: the value is not measured, or it is measured and will not survive the standard a filed annual report imposes. There is direct evidence for the second — Blackstone and Apollo both warn in their risk factors of being "accused of making inaccurate or misleading statements" about AI benefits, and Gen Digital names "AI washing" allegations outright.

The caveat on that argument, stated so it is not mistaken for a clean kill: three of those four companies produce savings figures for nothing at all, AI or otherwise, so their silence on AI is partly house style rather than evidence about AI specifically. The convention explanation is weakened rather than eliminated.

10. Limitations

- **The tier-2 count is an upper bound.** A late correction (amendments A13 and A14) found annual-report tier-2 codes resting on risk-factor language — a company describing what it does as the *premise* of a risk disclosure is not claiming a benefit, and "we use AI to improve X" does not assert that X improved. All 21 companies whose AI content is wholly or almost wholly in risk factors were re-read; 16 fell to tier 1, taking tier 2 from 103 to 87. The 82 with substantive non-risk content were not re-read, because the deciding passage is not stored per company, so the true figure is 87 or lower. **No published finding changed**: this is entirely at the 1/2 boundary, tier 3 and tier 4 remain zero, and the releases are untouched.
- **The coding is model-generated and human-adjudicated.** Human blind double-coding covers 30 of 498 companies on the 10-K (6%) and 18 on earnings releases (3.6%), across two draws, and the second draw did not clear the pre-registered reliability threshold on the internal axis (kappa 0.553). Agreement at the tier 3/4 boundary was complete, which is the relevant number for every published finding, but the lower boundaries are moderate at best. A

separate 30-company audit measures consistency out of sample, not human agreement across the population.
- **One fiscal year, two document types.** No time series.
- **No earnings calls.** Transcripts are out of scope (§2).
- **The AI-cost count is a floor** (§6).
- **The product axis is provisional.** It has not had the scrutiny the internal axis has, and a spot check found rows whose stated basis does not match the rule they are coded under. One case is flagged and deliberately left unresolved (Dell). **No published claim in this study rests on the product axis.**
- **Eleven exhibits could not be read at all.** Ten are covered by a readable sibling exhibit in the same filing; the eleventh (Linde, Q3 2024) is not, though that company carries no AI term in any of its six readable exhibits.
- **The pre-wave term exclusion costs the study PPL** (§8).

11. What is available

On request, in full:

- Per-company codings for all 498 companies in both venues, with the evidence note behind each judgment
- The filing list with SEC accession numbers, filing dates and period ends
- The extracted passages with character offsets, every one of which slices back to the stored filing text
- The rubric and all fourteen amendments with their effects
- The blind double-coding, the kappas and the adjudication log
- The audit draw and its random seed
- The screen definitions, including the two that failed and the one that nearly produced a false null

Extraction is deterministic: a second run produced byte-identical output. Every published count is regenerated from the coded data by a script rather than maintained by hand, after an audit found that hand-maintained figures had drifted.