ISSN: 2473-4901
v12 n6724

How AI Generates Value and How Firms Show It: A Morphological Analysis of ROI Evidence in S&P 500 Earnings Calls

Tan Gürpinar
Tan.Gurpinar@qu.edu

Johannes January
Johannes.January@qu.edu

Guido Lang
Guido.Lang@qu.edu

Luis Sa-Couto
Luis.SaCouto@qu.edu

Quinnipiac University
Hamden, CT, USA

Abstract

Voluntary AI disclosure in corporate earnings calls varies widely in whether claims come with measurable supporting evidence. The credibility structure of these disclosures has been studied only as a binary distinction between concrete and speculative mentions, leaving the dimensions along which AI claims vary uncharacterized. This study examines how AI is described as generating value in corporate disclosures and how firms substantiate those claims through quantified return-on-investment (ROI) evidence. We use large language model-based extraction of 12,898 AI use cases from 3,512 S&P 500 earnings calls across 479 companies from Q1 2020 through Q1 2026. Each use case is classified along a six-dimensional morphological taxonomy capturing both what is being claimed and how credibly it is substantiated. Logistic regression identifies the disclosure characteristics that independently predict measurability. Only 36.5% of AI use cases come with quantified ROI evidence; the remaining 63.5% describe AI deployments without supporting numbers. Revenue-growth and cost-reduction claims are about twice as likely to be quantified as product-capability claims, and the most credible region of the corpus combines in-production status with revenue or cost-reduction framing. Strikingly, AI claims about a firm's own products carry higher ROI evidence rates than claims about its internal operations — reversing the intuition that firms can most easily substantiate AI in their own workflows. Additionally, an exploratory mechanism analysis shows how AI is described as generating value within each value driver, with cost reduction operating predominantly through labor substitution and revenue growth through personalization.

Keywords: AI disclosure, earnings calls, ROI evidence, voluntary disclosure, morphological taxonomy, AI value mechanisms

How AI Generates Value and How Firms Show It: A Morphological Analysis of ROI Evidence in S&P 500 Earnings Calls

Tan Gürpinar, Johannes January, Guido Lang and Luis SaCouto

1. INTRODUCTION

Quarterly earnings calls have become a primary channel through which corporate executives signal artificial intelligence (AI) activity to investor audiences, and the volume of AI-related disclosure has grown sharply since the launch of generative AI tools in late 2022. As AI discussion has become near-ubiquitous among large publicly traded companies, attention has shifted from whether firms mention AI to the quality of what they disclose. Some AI claims arrive with specific percentages, dollar figures, or headcount numbers attached, whereas others describe deployments in general terms without supporting evidence. The distinction matters because, in a setting where executives have wide latitude, quantified evidence functions as a credibility signal that distinguishes substantive operational claims from narrative positioning.

Cao et al. (2024) establish that this distinction has market consequences, with AI disclosures that have clear implementation plans generating significant valuation benefits while speculative mentions do not. Recent work has extended this concern by quantifying the gap between firms' AI talk and AI walk, as Li (2025) shows that AI talk in earnings calls often does not predict subsequent AI workforce investment, while Barrios et al. (2025) find that AI disclosures are more informative when aligned with AI employment and that firms with disclosure-employment mismatches experience weaker operational and market outcomes. These contributions establish that disclosure quality is economically meaningful and that variation across firms can be detected, but they treat AI disclosure as a single firm-level signal compared against external proxies for substance, leaving open the question of what makes a specific AI claim concrete and how credibility varies across types of disclosure within a single firm. The dimensions along which individual AI claims vary, including what is being claimed, when the outcome is realized, who benefits, how the claim is quantified, and where the evidence comes from, have not been systematically characterized at the use-case level.

This study addresses that gap by extracting structured records of AI use cases from S&P 500 earnings calls and coding each one for whether the disclosure includes quantified ROI evidence and, if so, what form the evidence takes. We then develop a morphological taxonomy following Nickerson et al. (2013) that characterizes AI ROI claims along six dimensions grouped into two meta-dimensions (MD) that distinguish what is being claimed from how credibly it is substantiated. The taxonomy supports systematic comparison across use cases and identifies the credibility positions that voluntary AI disclosure does and does not occupy. The present study examines ROI evidence as a distinct dimension of disclosure quality across all AI use cases in the corpus.

Our central research question asks: *How is ROI evidence distributed across AI use case disclosures in S&P 500 earnings calls, and what makes some ROI claims more credible than others?*

We address this question using 12,898 AI use cases drawn from 3,512 earnings calls across 479 S&P 500 companies spanning Q1 2020 through Q1 2026. The study contributes to three streams of research. Methodologically, it operationalizes AI disclosure credibility at the use-case level, extending Cao et al.'s (2024) binary distinction into a six-dimensional morphological taxonomy that supports systematic characterization of how credibility varies across disclosures. Empirically, it provides a longitudinal mapping of where voluntary AI disclosure is substantiated and where it remains speculative, documenting that ROI evidence concentrates in in-production deployments with revenue or cost-reduction framing and is less common in product-capability claims and rare in aspirational claims. Theoretically, it characterizes voluntary AI disclosure as exhibiting credibility stratification in which firms can substantiate certain claim types but not others, with implications for how investors, analysts, and regulators interpret AI narratives in voluntary corporate communications.

2. RELATED LITERATURE

2.1 Voluntary Disclosure and Credibility Signaling

Voluntary disclosure theory provides the foundation for explaining why some AI claims are accompanied by quantified evidence while others are not. Healy and Palepu (2001) identify capital market access, control contests, compensation, litigation and proprietary costs, and agency considerations as the main forces shaping disclosure choices. In their framework, managers disclose information when the expected benefits exceed the expected costs, with decisions depending on both information content and disclosure credibility. Verrecchia (1983, 2001) formalizes this tradeoff: managers withhold information when disclosure costs outweigh benefits, creating gaps between what firms know and communicate.

Spence (1973) provides the canonical account of signaling under information asymmetry. When quality cannot be directly observed, firms rely on signals that are more costly for low-quality than high-quality types to produce. Quantified evidence serves this role in voluntary disclosure. Firms with operational AI deployments can attach percentage improvements, dollar values, or headcount effects to their claims, whereas firms making speculative claims face greater costs from doing so because such claims are more easily subject to later verification. The asymmetric cost of producing quantified evidence therefore makes it a credibility signal.

Applied to AI disclosure, this framework predicts that ROI evidence will be concentrated in deployments with measurable outcomes and in value drivers that lend themselves to direct quantification. Cost-reduction and revenue-growth initiatives are easier to express in percentages or dollar terms than broad capability narratives, implying systematic differences in quantification. Voluntary AI disclosure thus constitutes a structured field in which credibility varies systematically with the characteristics of the underlying deployment rather than being uniformly credible or uniformly speculative.

2.2 Earnings Calls and AI Disclosure

Earnings conference calls have long served as a key channel for voluntary disclosure beyond mandatory reporting (Bushee et al., 2003; Matsumoto et al., 2011). Their combination of prepared remarks and analyst questioning creates a structured but heterogeneous text record that has increasingly been analyzed using natural language processing. Hassan et al. (2019) show that text-based measures of earnings-call content enable valid cross-firm comparisons and predict subsequent firm behavior, establishing transcript content as structured analyzable data. Hollander et al. (2010) show that strategic silence itself conveys information, framing the absence of ROI evidence as a disclosure choice rather than a neutral omission.

Recent studies have applied these methods to AI disclosure. Cao et al. (2024) distinguish between AI disclosures with clear implementation plans and speculative claims, showing that valuation benefits accrue only to the former. While their binary classification establishes the economic importance of disclosure quality, it does not characterize variation in credibility within each category. Jia et al. (2026) find that discussions of generative AI in earnings calls increased sharply after ChatGPT's release, became more action-oriented, and varied systematically across firms and industries. Soto (2025) develops an AI Research Index from semantic similarity between earnings calls and AI publications, showing that substantive AI discussion predicts subsequent capital investment beyond mere mention frequency.

Related studies further connect earnings-call AI language to investment outcomes. Chen et al. (2026) link AI discussion to investment efficiency, while Kalyani and Li (2026) show that recent U.S. AI investment growth is concentrated among large public firms with positive AI sentiment. Together, these findings support treating earnings calls as a useful setting for examining how firms communicate AI's expected economic value.

2.3 AI Washing and Disclosure Quality

The phenomenon of AI washing, defined as claims that exceed the substance of underlying deployments, has received increasing attention but remains difficult to operationalize empirically. Li (2025) conceptualizes AI washing as a gap between AI talk in earnings calls and AI walk measured through workforce expertise, finding that discussion often fails to predict subsequent AI investment. Barrios et

al. (2025) find that AI disclosures are more informative when aligned with AI employment, while mismatches predict weaker outcomes. Blades et al. (2026) show that analysts distinguish between substantive and speculative AI discussion. Together, these studies demonstrate that disclosure quality matters but evaluate credibility primarily at the firm level rather than at the level of individual AI use-case claims.

Existing measures of AI-related activity rely on occupational exposure scores (Eisfeldt et al., 2023) or résumé-based proxies for AI investment (Babina et al., 2024). These approaches capture AI exposure and talent investment but not the deployments firms describe or the evidence accompanying them. The present study addresses this gap by operationalizing AI disclosure credibility at the use-case level and examining how it varies across deployment types, value drivers, and disclosure contexts.

3. METHODOLOGY

The analysis covers S&P 500 quarterly earnings calls from Q1 2020 through Q1 2026, drawn from a larger data collection on AI mentions across these calls. Transcripts for S&P 500 member firms were sourced from financial data providers and processed in two stages. A first stage scanned each transcript with a rule-based keyword search covering terms such as artificial intelligence, machine learning, large language model, generative AI, foundation model, neural network, and deep learning, together with compound patterns including natural language processing, computer vision, and predictive analytics. A second stage passed each retained call to a large language model (Claude Sonnet, claude-sonnet-4-6) via the Anthropic Messages Batches API with a structured extraction prompt that returned records of the AI use cases disclosed. This approach follows related work applying generative AI to scalable text extraction, where it reproduces manual coding with comparable accuracy at substantially greater scale (Sa-Couto et al., 2025; Sa-Couto et al., 2026). The present study analyzes these use cases and codes each for ROI evidence and its position within the morphological taxonomy developed in Section 3.2.

3.1 Use Case Corpus and ROI Evidence Extraction

The extraction applied a three-part qualification test requiring an identifiable AI or machine learning capability, a specific task or process, and an identifiable beneficiary. The corpus includes all AI use cases regardless of subject classification: focal-internal (where the reporting company deploys AI in its own operations), focal-product (where the reporting company sells or embeds AI in a product), and customer-deployment (where the reporting company references a customer's AI deployment). This broader scope is appropriate because ROI evidence can be compared across subjects and that variation is analytically informative.

Two levels of ROI evidence extraction were applied. A binary indicator flags whether the disclosure includes measurable evidence attributable to AI—a specific percentage, dollar figure, headcount or full-time-equivalent number, time-savings claim, multi-metric statement, or explicit reference to performance measurement without a disclosed numeric value. The evidence is then classified by metric type as percentage, dollar value, headcount/FTE, time savings, multiple metrics, or qualitative reference. The binary indicator captures the credibility-signaling dimension of interest; the metric type classification adds depth for the morphological taxonomy in Section 3.2.

To assess extraction quality, two authors independently hand-coded a stratified random sample of 100 use cases drawn proportionally across sectors and value driver categories. Coders evaluated whether the ROI evidence binary indicator was correct and whether the metric type matched the quoted evidence. Cohen's kappa between human coders and the LLM output was 0.81 for the binary indicator and 0.74 for metric type, both in the substantial agreement range (Landis & Koch, 1977). Inter-coder agreement between the two human raters was 0.85, indicating extraction quality comparable to inter-human agreement on similar coding tasks. The stratified design supports this check: agreement rates were consistent across tech vs. non-tech firms, across financial vs. non-financial value drivers, and across pre- and post-2023 disclosures, providing evidence against systematic bias in specific subgroups.

3.2 Morphological Taxonomy Development

We developed a morphological taxonomy of AI ROI claims following the iterative method of Nickerson et al. (2013), which combines conceptual derivation from prior literature with empirical refinement based

on observed instances. The taxonomy organizes use case characteristics along two MD, namely what is being claimed and how credibly it is substantiated. Each MD contains three dimensions, and each dimension contains five mutually exclusive characteristics. The taxonomy was developed through six iterations, beginning with initial conceptual versions derived from voluntary disclosure theory and the descriptive structure of the corpus, followed by an empirically refined version based on coding 500 use cases, and concluding with a final version after coding the full 12,898 use cases. The ending conditions specified by Nickerson et al. were met at the third iteration.

3.3 Analytical Approach

Descriptive analysis tabulates ROI evidence rates across the dimensions of the morphological taxonomy and across additional disclosure characteristics including use case subject and year. A logistic regression specifies the probability of ROI evidence as a function of deployment maturity, value driver type, use case subject, and year, with standard errors clustered at the firm level to account for within-firm correlation across multiple calls. The regression identifies which disclosure characteristics independently predict measurability after controlling for the others, distinguishing direct effects from compositional ones.

4. FINDINGS

4.1 A Morphological Taxonomy of AI ROI Claims

The first finding of the analysis is a structural one, since applying the development procedure described in Section 3.2 to the full corpus produced a six-dimensional morphological taxonomy of AI ROI claims, presented in Table 1.

Table 1. Taxonomy of AI ROI Claims in Earnings Call Disclosures.

Meta-Dimension	Dimension	Characteristics					N
Content	**Value driver type**	**Cost reduction** 2,841 (22.0%)	**Revenue growth** 2,449 (19.0%)	**Risk reduction** 604 (4.7%)	**Product capability** 6,413 (49.7%)	**Other strategic value** 591 (4.6%)	n=12,898
	Beneficiary scope	**Workforce** 1,815 (14.1%)	**Operations** 5,401 (41.9%)	**Customers** 4,993 (38.7%)	**Partners/ecosystem** 403 (3.1%)	**Other external** 286 (2.2%)	n=12,898
	Time horizon	**Already realized** 638 (4.9%)	**In progress** 546 (4.2%)	**Near-term expected** 699 (5.4%)	**Multi-year aspirational** 347 (2.7%)	**Indefinite** 10,668 (82.7%)	n=12,898
Form	**Deployment maturity**	**Scaling** 472 (3.7%)	**In production** 9,999 (77.5%)	**Piloting** 1,210 (9.4%)	**Planned** 655 (5.1%)	**Aspirational** 515 (4.0%)	n=12,898
	Quantification format	**Monetized value** 906 (19.2%)	**Percentage** 1,442 (30.6%)	**Operational metrics** 194 (4.1%)	**Multiple metrics** 540 (11.5%)	**Qualitative only** 1,625 (34.5%)	n=4,707
	Evidence source	**External validation** 323 (2.5%)	**Internal measurement** 238 (1.8%)	**Customer testimony** 117 (0.9%)	**Anecdotal** 303 (2.3%)	**No source named** 11,917 (92.4%)	n=12,898

The taxonomy is anchored by the credibility of AI ROI claims in voluntary corporate disclosure, understood as the extent to which claims are supported by observable evidence about both the substance of the deployment and the form of substantiation. Six dimensions derive from this meta-characteristic and are organized into two MD. The content MD captures what an AI use case claims to deliver through three dimensions covering value driver type, beneficiary scope, and time horizon. The form MD captures how the claim is substantiated through three dimensions comprising deployment maturity, quantification format, and evidence source.

The two MD reflect the distinction central to credibility signaling in voluntary disclosure, namely that substantive content and the supporting credibility infrastructure are analytically separate and can combine in different ways. A claim about cost reduction (content), for example, may be supported by an in-production deployment and internally measured performance metrics (form), or it may remain an

aspirational statement without an identified source. Thus, identical content can occupy different credibility positions depending on the accompanying form characteristics.

Table 1 displays the taxonomy as a heatmap in which each characteristic cell is shaded according to its empirical frequency across the 12,898 use cases in the corpus. Five frequency bands are applied across all dimensions, namely dominant (≥40% of cases in the dimension), major (20-40%), moderate (10-20%), minor (2-10%), and rare (<2%). The visual asymmetry of the resulting heatmap is itself informative, with four of the six dimensions showing a single characteristic at dominant frequency, indicating that voluntary AI disclosure concentrates heavily on particular positions in the credibility space rather than distributing evenly.

4.2 Overall ROI Evidence Distribution

Across the 12,898 AI use cases in the corpus, 4,707 (36.5%) include quantified ROI evidence. The remaining 8,191 (63.5%) describe AI deployments without quantified outcomes. The subsequent subsections show how this 36.5% rate varies substantially when the corpus is examined along the form and content MD of the taxonomy, with some positions occupied almost exclusively by ROI-backed claims and others by claims that are not substantiated.

4.3 Findings Along the Form Meta-Dimension: Deployment Maturity

Within the form MD, deployment maturity is the single strongest descriptive predictor of ROI evidence. Table 2 shows that in-production deployments are eight times more likely to carry ROI evidence than aspirational ones, with rates of 41.9% versus 5.4%. Piloting deployments carry a 17.4% ROI rate and planned deployments a 12.5% rate. Scaling deployments, an intermediate category between piloting and full production, carry a 46.0% ROI rate, the highest of any deployment stage and slightly above in-production status itself. The pattern reflects the simple operational fact that quantified outcomes can be reported only once a system is running and generating measurable effects, but it also reveals a credibility structure in which 93.2% of all ROI-backed claims describe in-production or scaling deployments, concentrating measurable AI evidence in mature systems rather than nascent ones.

Table 2. ROI Evidence by Deployment Maturity.

Deployment maturity	ROI-backed	Non-ROI-backed	ROI rate	Share of ROI-backed
Scaling	217	255	46.0%	4.6%
In production	4,386	6,085	41.7%	88.6%
Piloting	210	1,000	17.4%	4.5%
Planned	82	573	12.5%	1.7%
Aspirational	28	487	5.4%	0.6%
Unclear	1	46	2.1%	0.0%

4.4 Findings Along the Content Meta-Dimension: Value Driver Type

Within the content MD, value driver type is also an important descriptive predictor of ROI evidence. Table 3 shows that revenue-growth claims carry the highest ROI evidence rate at 56.1%, followed by cost-reduction claims at 45.8%. Product-capability claims, which dominate the corpus at 49.7% of all use cases, carry a much lower ROI evidence rate at 26.8%. Risk-reduction and other strategic value claims also fall below 30%. The pattern reflects the relative ease of quantifying discrete operational gains such as a percentage improvement in conversion, a dollar figure of cost savings, or a number of hours saved, compared to the difficulty of attaching specific evidence to broader product capability narratives.

Table 3. ROI Evidence by Value Driver Type.

Value driver	ROI-backed	Non-ROI-backed	ROI rate	Share of ROI-backed
Revenue growth	1,375	1,074	56.1%	29.2%
Cost reduction	1,301	1,540	45.8%	27.6%
Product capability	1,718	4,695	26.8%	36.5%
Competitive moat	153	424	26.5%	3.3%
Risk reduction	160	444	26.5%	3.4%

The product-capability finding has implications for how AI disclosure is interpreted overall. Product capability accounts for 36.5% of all ROI-backed claims but 57.3% of non-ROI-backed claims, meaning the dominant narrative framing in the corpus is the framing least likely to be substantiated. This pattern is consistent with the credibility-signaling interpretation, in that when firms describe AI as enhancing what their products can do, the breadth of the claim makes quantified evidence difficult to produce, while when firms describe AI as reducing a specific cost or driving specific revenue, the discrete operational structure supports direct measurement.

To examine how AI is described as generating the claimed value within each value driver, we conducted an exploratory heuristic classification of all 4,707 ROI-backed use cases by operational mechanism and corporate examples, identifying the pathways through which AI deployments produce value. Appendix A reports the full distribution: cost reduction is dominated by labor substitution and process automation (33.2%), revenue growth by personalization and matching (20.9%), risk reduction by direct loss avoidance (26.9%), and "other strategic value" by decision quality and prediction (20.3%), while product capability distributes more evenly across multiple mechanisms — indicating that voluntary AI disclosure varies systematically not only in what value is claimed but in how that value is described as being generated.

4.5 Interactions Between Form and Content: The Deployment-Value Configuration

The form and content MD of the taxonomy do not operate independently. Where they intersect, particular configurations of substantiation and content produce credibility profiles that neither dimension predicts on its own. Table 4 reports ROI evidence rates for the combinations of deployment maturity and value driver with the largest cell counts. The configuration of in-production deployment combined with revenue-growth framing reaches a 63.0% ROI evidence rate, the highest in any deployment-by-value pair and substantially above either dimension's marginal rate. In-production cost-reduction follows at 50.3%, whereas the same value drivers attached to piloting or planned deployments fall well below these levels, with piloting cost-reduction at 26.1% and planned revenue-growth at 25.6%.

The pattern shows that ROI evidence is a function of the joint configuration of form and content, not of any single dimension on its own, and identifies the in-production-plus-direct-financial-outcome configuration as the high-credibility quadrant of the taxonomic space. In contrast, the in-production-plus-product-capability configuration, which is the single largest cell in the corpus at 5,008 use cases, carries an ROI evidence rate of only 31.3%, indicating that even mature deployments fail to generate measurable evidence when the value driver is framed in broad capability terms rather than discrete operational outcomes.

Table 4. ROI Evidence by Deployment-Value Combination (Top Configurations).

Deployment maturity	Value driver	Use cases	ROI-backed	ROI rate
In production	Revenue growth	2,074	1,306	63.0%
In production	Cost reduction	2,404	1,209	50.3%
In production	Competitive moat	438	145	33.1%
In production	Product capability	5,008	1,570	31.3%
Piloting	Cost reduction	234	61	26.1%
Planned	Revenue growth	90	23	25.6%

Taken together, the findings support an AI Credibility Inference Guide:

- High-credibility disclosures: combine in-production or scaling deployment, revenue-growth or cost-reduction framing, and monetized or percentage quantification — a profile associated with ROI evidence rates of 50–63%.
- Low-credibility disclosures: combine aspirational deployment, product-capability framing, and qualitative-only evidence with no source named — a profile associated with ROI evidence rates below 10%.

4.6 Use Case Subject as a Credibility-Stratifying Characteristic

Use case subject, whether the AI deployment described is the focal firm's own operation, a product the focal firm sells, or a customer's adoption that the focal firm references, is not a dimension of the morphological taxonomy itself but operates as an additional credibility-stratifying characteristic that cuts across both MD. The same content claim, attached to the same form characteristics, can occupy a different credibility position depending on whose deployment the disclosure describes.

Focal-product use cases carry the highest ROI evidence rate at 38.4%, followed by focal-internal use cases at 33.5% and customer-deployment use cases at 32.8%. This ordering reverses the direct-measurement intuition that internal deployments would necessarily be the easiest to substantiate. One plausible interpretation is that product disclosures often come with named customer outcomes, standardized commercial metrics, or sales-facing evidence that firms can report when positioning AI-enabled products. Focal-internal and customer-deployment disclosures sit lower and closer together, indicating that subject classification contributes to the credibility profile but does not create as large a descriptive gap as deployment maturity or value-driver type.

4.7 Temporal Evolution of Taxonomic Positions

The distribution of AI use cases across taxonomic positions has shifted substantively over the 2020 through 2026 period. The share of all use cases including ROI evidence has risen from 32.3% in 2020 to 42.6% in Q1 2026, a 10.3-percentage-point increase over the sample window. A brief dip to 30.3% in 2023 coincides with the wave of post-ChatGPT generative AI announcements that introduced many speculative early-stage claims into the corpus. By 2024 the rate had recovered to 37.2%, rising further to 40.1% in 2025 and 42.6% in Q1 2026.

The temporal evolution suggests that the average voluntary AI disclosure has been migrating toward more credibility-bearing taxonomic positions over time, with more disclosures in the in-production cell of the deployment dimension, more disclosures with revenue or cost framing on the content side, and consequently more disclosures with quantified ROI evidence.

4.8 Cross-Cutting Patterns by Use Case Category

Beyond the taxonomic dimensions, ROI evidence rates also vary across use case categories. AI Hardware and infrastructure use cases carry the highest rate at approximately 64%, reflecting capital-intensive disclosures and standardized financial metrics. Code Generation follows at approximately 50%, with Customer Service and Cloud AI Services in the 38–40% range. Scientific R&D carries a lower rate

at 28.5%, while Foundation Model R&D is slightly higher at 31.4%, reflecting the long-horizon and exploratory nature of these activities where operational outcomes are often not yet measurable.

4.9 Predictors of ROI Evidence: Regression Results

A logistic regression tests which taxonomy dimensions independently predict ROI evidence. The dependent variable is the binary ROI indicator, and predictors include deployment maturity, value driver type, use case subject, and year as a temporal control. Standard errors are clustered at the firm level, and the model is estimated on the full corpus of 12,898 use cases.

Value driver type emerges as the strongest predictor of ROI evidence relative to the product-capability baseline. Revenue-growth framing increases the odds of ROI evidence by a factor of 3.82 and cost-reduction framing by 3.71. Competitive-moat framing is also positive but smaller (OR = 1.37, $p = .018$), whereas risk reduction does not differ significantly from product capability. Deployment maturity also matters, in that aspirational claims have much lower odds than in-production claims (OR = 0.07, $p < .001$), as do planned claims (OR = 0.20, $p < .001$) and piloting claims (OR = 0.32, $p < .001$).

Use case subject contributes independently. Focal-internal use cases have lower odds than focal-product use cases (OR = 0.44, $p < .001$), as do customer-deployment use cases (OR = 0.56, $p < .001$), consistent with voluntary disclosure cost-benefit logic: internal deployments may carry proprietary disclosure costs that suppress quantification, while product-related claims are supported by commercial metrics firms have stronger incentives to report. Year shows a modest upward effect (OR = 1.06, $p = .002$). The McFadden pseudo-R^2 of 0.112 indicates ROI evidence is not randomly distributed but most strongly predicted by value-driver framing and deployment maturity, with use case subject adding an independent credibility-stratifying effect.

5. DISCUSSION

The analysis of 12,898 AI use cases reveals that voluntary AI disclosure exhibits structured credibility stratification rather than a uniform mix of substantive and speculative claims, a structure that emerges across three particularly noteworthy findings.

First, deployment maturity and value driver type jointly determine where claims are substantiated. This pattern is consistent with voluntary disclosure theory, in that firms can attach specific measurements to realized operational gains but struggle to do so for broad capability claims or forward-looking aspirations. The asymmetric cost of producing quantified evidence generates the observed variation in credibility.

Second, product capability dominates the non-substantiated portion of the corpus. Consequently, the most common framing of AI value is also the least likely to be substantiated. This pattern offers a structural interpretation of AI washing, suggesting that rather than implying fabrication, the dominant framing of AI value is inherently difficult to substantiate.

Third, credibility varies systematically by use case subject. Focal-product disclosures exhibit the highest ROI evidence rates, while focal-internal and customer-deployment claims cluster at lower levels. Product-oriented disclosures may benefit from named customer outcomes and standardized commercial metrics, whereas internal transformation narratives often lack quantified performance measures. Subject classification therefore contributes to credibility in ways more nuanced than a simple measurement-access explanation would predict.

The findings extend the voluntary disclosure literature by showing that AI disclosure possesses a multidimensional credibility structure rather than a simple concrete-versus-speculative divide. While Cao et al. (2024) demonstrate the valuation relevance of this distinction, the present study shows that credibility also varies systematically with deployment maturity, value driver type, beneficiary scope, quantification format, and evidence source. The morphological taxonomy provides a framework for locating individual disclosures within this credibility space and comparing them across firms, sectors, and periods. For investors and analysts, observable disclosure characteristics signal claim credibility: in-production deployments framed around revenue growth or cost reduction with quantified outcomes

occupy the high-credibility region of the taxonomy, while aspirational product-capability claims without metrics occupy the low-credibility region. This offers a more granular basis for reading AI narratives than aggregate mention measures or binary classifications. For regulators, disclosure requirements emphasizing deployment status, value drivers, and quantified outcomes would address the dimensions where voluntary AI disclosure exhibits the largest credibility gaps, strengthening informational value without prescribing content.

The findings are subject to several limitations, the first being that because the corpus consists of S&P 500 firms, the results may not generalize to smaller or private companies. Although hand validation of 100 use cases indicates substantial agreement between LLM and human coding, larger multi-coder validation would better quantify classification error. In addition, the positive bias of investor-facing communications limits the observation of unfavorable AI outcomes. Finally, the ROI evidence indicator captures the presence of measurable evidence, not its accuracy, leaving claims potentially subject to selective measurement or framing.

6. CONCLUSION

Voluntary AI disclosure in S&P 500 earnings calls from Q1 2020 through Q1 2026 exhibits a structured credibility profile in which 36.5% of AI use cases include ROI evidence and 63.5% do not. The central finding is a six-dimensional taxonomy that organizes AI ROI claims along two meta-dimensions: content (how AI is described as generating value) and form (how firms show it through deployment maturity, quantification, and evidence) Within this framework, substantiated claims concentrate in in-production revenue-growth and cost-reduction deployments, whereas non-substantiated claims are concentrated in product-capability framing and aspirational use cases. The overall ROI evidence rate rose from 32.3% in 2020 to 42.6% in Q1 2026 as deployments matured.

The study makes three contributions, the first methodological, operationalizing AI disclosure credibility at the use-case level using LLM-based extraction validated against human coding. As a research artifact, it develops a six-dimensional morphological taxonomy and heatmap. Empirically and theoretically, it demonstrates that AI disclosure exhibits credibility stratification consistent with signaling under information asymmetry, with quantified evidence concentrating where outcomes are directly measurable and remaining scarce for broad capability narratives and forward-looking aspirations. An exploratory mechanism analysis further shows how AI is described as generating value within each value driver, with cost reduction operating predominantly through labor substitution and revenue growth through personalization.

Practically, the taxonomy enables investors and analysts to infer credibility from observable disclosure characteristics, with value driver framing and deployment maturity as the strongest predictors, and to distinguish the operational mechanism through which AI is described as generating the claimed value For regulators, the findings identify where disclosure standardization could strengthen credibility signals without prescribing content. Future research linking ROI-backed AI claims to firm performance could test whether these credibility signals translate into the productivity and valuation effects predicted by voluntary disclosure theory.

7. REFERENCES

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics, 151*, Article 103745. https://doi.org/10.1016/j.jfineco.2023.103745

Barrios, J. M., Campbell, J. L., Johnson, R. G., & Liu, C. (2025). *Signals or smoke? The determinants and informativeness of corporate artificial intelligence (AI) disclosures*. SSRN. https://doi.org/10.2139/ssrn.5133107

Blades, R., Bilinski, P., & Kraft, A. (2026). *Analysts' response to "AI washing" in earnings conference calls*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6359598

Bushee, B. J., Matsumoto, D. A., & Miller, G. S. (2003). Open versus closed conference calls: The determinants and effects of broadening access to disclosure. *Journal of Accounting and Economics, 34*(1–3), 149–180. https://doi.org/10.1016/S0165-4101(02)00073-3

Cao, Y., Liu, M., Qiu, J., & Zhao, R. (2024). *Information in disclosing emerging technologies: Evidence from AI disclosure*. SSRN. https://doi.org/10.2139/ssrn.4987085

Chen, S.-S., Kim, J., & Peng, S.-C. (2026). The real effects of AI: Evidence from corporate investment efficiency. *Journal of Empirical Finance, 88*, Article 101730. https://doi.org/10.1016/j.jempfin.2026.101730

Eisfeldt, A. L., Schubert, G., & Zhang, M. B. (2023). *Generative AI and firm values* (NBER Working Paper No. 31222). National Bureau of Economic Research. https://doi.org/10.3386/w31222

Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics, 134*(4), 2135–2202. https://doi.org/10.1093/qje/qjz021

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics, 31*(1–3), 405–440. https://doi.org/10.1016/S0165-4101(01)00018-0

Hollander, S., Pronk, M., & Roelofsen, E. (2010). Does silence speak? An empirical analysis of disclosure choices during conference calls. *Journal of Accounting Research, 48*(3), 531–563. https://doi.org/10.1111/j.1475-679X.2010.00365.x

Jia, N., Li, N., Ma, G., & Xu, D. (2026). Corporate responses to generative AI: Early evidence from conference calls. *Review of Accounting Studies, 31*(2), 649–703. https://doi.org/10.1007/s11142-025-09916-1

Kalyani, A., & Li, H. (2026, May 18). Is optimism for artificial intelligence boosting investment? *FRBSF Economic Letter, 2026*(13), 1–6. https://www.frbsf.org/research-and-insights/publications/economic-letter/2026/05/is-optimism-for-artificial-intelligence-boosting-investment/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Li, B. (2025). *AI washing* [Working paper]. University of Florida.

Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review, 86*(4), 1383–1414. https://doi.org/10.2308/accr-10034

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*(3), 336–359. https://doi.org/10.1057/ejis.2012.26

SaCouto, L., Gürpinar, T., Jia, S., & Quadri, S. (2026). Preparing the XR workforce: Curricular patterns of extended reality in higher education. Journal of Information Systems Education. Forthcoming.

Sa-Couto, L., Gürpinar, T., Jia, S. J., Quadri, S., & McKenna, A. (2025). Automating curriculum intelligence: Scalable extraction and categorization of university courses using GenAI. *Proceedings of the ISCAP Conference, 11*(6346), 1–13. https://iscap.us/proceedings/

Soto, P. E. (2025). Research in commotion: Measuring AI research and development through conference call transcripts. *Finance and Economics Discussion Series* No. 2025-011. Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2025.011

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics, 87*(3), 355–374. https://doi.org/10.2307/1882010

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics, 5*, 179–194. https://doi.org/10.1016/0165-4101(83)90011-3

Verrecchia, R. E. (2001). Essays on disclosure. *Journal of Accounting and Economics, 32*(1–3), 97–180. https://doi.org/10.1016/S0165-4101(01)00025-8

ISSN: 2473-4901
v12 n6724

APPENDIX A

How AI Generates Value: Operational Mechanisms by Value Driver and Corporate Examples

Mechanism	n	%	Description	Example
Cost reduction (N=1,301)				
Cost substitution (labor/automation)	432	33.2%	AI automates work previously done by employees.	"HSBC uses prebuilt agents with Dynamics 365 to manage customer inquiries across products, markets, and regulatory requirements, reducing resolution time by over 30%" (MSFT 2026 Q3).
Decision quality/prediction	157	12.1%	AI improves resource allocation, reducing waste.	"Our tech stack and workflows drive prices down. In GLP-1s, we have had a 200 bps improvement in share across the category—from winning new DTC customers and continuing to serve payer customers" (CVS 2026 Q1).
Cycle time reduction	121	9.3%	AI shortens processing time per task, lowering unit cost.	"We are getting about a 30% benefit with our programmers from a coding and testing perspective. That has allowed us to get to more stuff" (AXP 2026 Q1).
Personalization/ matching	59	4.5%	Better targeting reduces wasted spend.	"We have built an AI-assisted recommendation system for exception item processing... banks report that AI is reducing the time to close exceptions each day by 70% to 80%" (JKHY 2026 Q3).
Quality/accuracy improvement	37	2.8%	Fewer errors reduce rework and warranty costs.	"Already as a result of this initial rollout, customer satisfaction is up 16%, and call center flows are managing end to end with 90% accuracy, resulting in 30% faster service times and 15% fewer network failures" (ORCL 2026 Q2).
Risk/loss avoidance	20	1.5%	AI prevents losses that would otherwise show as cost.	"We launched an AI chat tool to handle customer inquiries related to the Drug Supply Chain Security Act. By enabling natural language answers to complex DSCSA data questions, we prevented 75% of inquiries from being escalated" (MCK 2026 Q3).
Throughput/capacity expansion	17	1.3%	Higher volume at same cost lowers per-unit cost.	"We handle more than 250 million service interactions a year with over half resolved through self-service. More than 30% of those are powered by AI and that number keeps increasing" (EXPE 2026 Q1).
Other/unclear	458	35.2%	-	-
Revenue growth (N=1,375)				
Personalization/ matching	287	20.9%	AI matches offers to customer preferences, lifting conversion.	"In Q2, our personalized product recommendation carousels drove over $470,000,000 of ecommerce sales, and our newly modernized product display pages are driving incremental sales" (COST 2026 Q2).
Decision quality/prediction	123	8.9%	AI predicts demand or customer behavior, enabling targeted growth.	"We introduced a new Gemini Enterprise agent platform that empowers users to build, orchestrate, govern and optimize agents... In Q1, Gemini Enterprise paid monthly active users grew 40% quarter-over-quarter" (GOOGL 2026 Q1).
Cost substitution (labor/automation)	88	6.4%	AI automates sales/service tasks, freeing capacity to capture more revenue.	"Our agentic products in LinkedIn Talent Solutions, which help hirers automate time-consuming tasks like sourcing, screening, and drafting messages, have already surpassed a $450 million annualized revenue run rate" (MSFT 2026 Q3).
Cycle time reduction	60	4.4%	Faster service reduces abandonment and shortens sales cycles.	"With the launch of Uber Autonomous Solutions, we're building the technical and operational infrastructure to help our partners commercialize faster... AV Mobility trips grew more than 10x" (UBER 2026 Q1).
Quality/accuracy improvement	20	1.5%	Better recommendations improve satisfaction and lifetime value.	"Microsoft 365 Copilot's accuracy powered by WorkIQ is unmatched, delivering faster and more accurate work-grounded results than competition. We have seen our biggest quarter-over-quarter improvement in response quality to date" (MSFT 2026 Q2).
Risk/loss avoidance	19	1.4%	Reduced fraud or churn translates to retained revenue.	"USDC and Base are now powering the majority of onchain stablecoin transactions for AI agents. When agents pay with crypto onchain, they use USDC 99% of the time and over 90% of those transactions are happening on the Base chain in Q1" (COIN 2026 Q1).
Throughput/capacity expansion	17	1.2%	Higher transaction volume directly grows revenue.	"We are accelerating the deployment of Gemini across our entire ads infrastructure to help businesses reach more customers in more places than ever before... Smart bidding now uses Gemini to match user intent to an advertiser's product and services more accurately and further drive performance" (GOOGL 2026 Q1).
Other/unclear	761	55.3%	-	-
Risk reduction (N=160)				
Risk/loss avoidance	43	26.9%	AI detects fraud, breaches, before they translate to losses.	"Apple Pay eliminated more than $1 billion in fraud for our partners last year, and we have made it available in more markets than ever before" (AAPL 2026 Q1).
Decision quality/prediction	19	11.9%	Predictive models surface emerging risks for proactive mitigation.	"When we worked with Palantir on the ontology, we were able to get a look at the portfolio within a week about every upcoming month as to what the submission activity is going to be and what we like for pricing and what we thought we needed to restructure" (AIG 2026 Q1).
Cycle time reduction	18	11.2%	Faster threat detection narrows the window of exposure.	"In these markets, the feature has reduced interaction with bad actors by 20% to 30% with minimal impact on revenue. Originally developed by Tinder, FaceCheck showcases portfolio-wide innovation, enabling Hinge to quickly iterate" (MTCH 2026 Q1).
Cost substitution (labor/automation)	14	8.8%	Automated controls replace human risk-monitoring effort.	"The number of Security Copilot customers increased 2x year over year. Our data security triage agents alone handled over 2 million unique alerts this quarter. We are helping customers secure their AI deployments as well. Thirty-five billion Copilot interactions have been audited by Purview" (MSFT 2026 Q3).

ISSN: 2473-4901
v12 n6724

Quality/accuracy improvement	11	6.9%	Higher detection accuracy reduces both false positives and missed risks.	"Mythos and SPUD and even other current-generation models with AIP are capable of finding novel vulnerabilities in complex cyber kill chains. They have discovered thousands of zero days across major operating systems and browsers" (PLTR 2026 Q1).
Personalization/ matching	10	6.2%	Risk-scoring tailored to individual exposure profiles.	"The latest version of Claude can elevate the performance of an entire claims team, surfacing patterns across submissions that are easy to miss... Examples of what Claude routinely flags include time line inconsistencies, geolocation mismatches, linguistic fingerprints, prior claim patterns, document tampering signals, coverage gaps" (AIG 2026 Q1).
Throughput/capacity expansion	2	1.2%	Larger volume of risk events monitored simultaneously.	"These adept authorizations and Visa Risk Manager utilize artificial intelligence and machine learning capabilities, which helped reduce fraud by $26 billion screening 30% more transactions in 2021 than in 2020" (V 2021 Q4).
Other/unclear	43	26.9%	-	-
Product capability (N=1,718)				
Cost substitution (labor/automation)	292	17.0%	AI embedded in the product automates tasks for end-users.	"In our healthcare business, we now have 274 customers live in production on our clinical AI agent, and that number continues to rise daily. Also in healthcare, our brand new AI-based ambulatory EHR is generally available and has received US regulatory approval" (ORCL 2026 Q2).
Decision quality/prediction	220	12.8%	AI in the product produces better recommendations or insights for users.	"More than 1/3 of our CapIQ Pro users engage with the AI features we've launched, including ChatIQ and Document Intelligence. We also saw tremendous growth in the usage of S&P Global data in the quarter" (SPGI 2026 Q1).
Cycle time reduction	200	11.6%	Product runs faster or processes user inputs in real time.	"In Q4, we added several new technologies and advancements, including DLSS 4.5, which uses AI to bring game visuals to a new level, and 35% faster LLM inference across leading AI PC frameworks" (NVDA 2026 Q4).
Personalization/ matching	161	9.4%	Product tailors output to individual user context.	"The Meta AI business assistant has now been fully rolled out to all eligible advertisers on supported Meta buying services, providing personalized recommendations to advertisers, resolving account issues, and surfacing campaign insights to help optimize results" (META 2026 Q1).
Quality/accuracy improvement	59	3.4%	Product produces more accurate results than non-AI alternative.	"Acrobat AI Assistant provides users conversational experiences that help them comprehend information faster and more accurately with an individual PDF or across documents in the PDF Space" (ADBE 2026 Q1).
Risk/loss avoidance	32	1.9%	Product reduces customer's downstream risks.	"Our new Visa Large Transaction Model is beginning to act as the foundational model for a variety of AI-powered fraud and risk services at Visa. Early results have shown that it can power up to a 5x increase in fraud value capture" (V 2026 Q2).
Throughput/capacity expansion	17	1.0%	Product handles higher volume than prior generation.	"We are also seeing more customers use Copilot Studio to Microsoft 365 Copilot and build their own agents. This year, customers created 3 million agents using SharePoint and Copilot Studio" (MSFT 2025 Q4).
Other/unclear	737	42.9%	-	-
Other strategic value (N=153)				
Decision quality/prediction	31	20.3%	AI enables better strategic inference in underwriting or planning.	"By some metrics, the AI5 chip will be 40 times better than the AI4 chip. Not 40%, 40 times. Because we have a detailed understanding of the entire software and hardware stack, we're designing the hardware to address all of the pain points in software" (TSLA 2025 Q3).
Cycle time reduction	15	9.8%	Faster strategic responses to market or operational change.	"Using Oracle Fusion, we continue to close and file our financial results faster than any other company in the S&P 500, providing us with a significant strategic advantage as well as an opportunity to help our Fusion customers do the same" (ORCL 2026 Q3).
Personalization/ matching	11	7.2%	Strategic positioning tailored to specific market segments.	"We already have 115 major AI projects. But my guess is in five years, it'll be 1,000 AI projects. So we're going as fast as we can to do a great job for customers" (JPM 2021 Q1).
Cost substitution (labor/automation)	8	5.2%	Automation of strategic-analytic functions previously done by analysts.	"We will also continue to grow and enhance the capabilities of Super Cruise, the industry's best L2 driver assistance technology. We will also develop L3 and even more advanced automated autonomous technologies" (GM 2025 Q1).
Quality/accuracy improvement	6	3.9%	Higher-fidelity strategic analysis or competitive intelligence.	"Our engaged acres stepped up again this quarter to 500 million engaged acres — over 10% increase from a year ago and about a 25% increase on highly engaged. So nearly 1/3 of those engaged acres are highly engaged" (DE 2026 Q1).
Throughput/capacity expansion	2	1.3%	Scaling strategic analytics across more units or markets.	"Gemini 3.1 Pro continues to push the frontier in reasoning, multimodal understanding and cost... First-party models now process more than 16 billion tokens per minute via direct API used by our customers, up from 10 billion last quarter" (GOOGL 2026 Q1).
Risk/loss avoidance	1	0.7%	Strategic risk avoidance through AI-informed positioning.	"Last 5 years, we've invested $7 billion into Safety and Security Solutions and that makes a competitive advantage for us" (MA 2024 Q1).
Other/unclear	79	51.6%	-	-