Signal vs. Substance: We Mistake AI Disclosure for Value

Henry Dalby

8 July, 2026

MBA Candidate - IMD Business School MBA

Priced on Disclosure

Enterprises are spending $30-$40 billion on AI, yet 95% of pilot projects fail to realize measurable business impact.[1] When it comes to AI Integration, the problem is that we price firms on what they disclose, not on the value they create. The same information asymmetry that misleads one firm benchmarking against a competitor is, at scale, mispricing the AI economy—the market is mistaking disclosure for real value.

Everyone is asking the same question: *what are my competitors doing?* In fact, 51% of executives say strategy is the biggest driver of AI ROI.[2] Yet only 22% of operations leaders report a fully developed AI strategy.[2] Given that the market signals are misleading, the competitive advantage comes from reading your own position rather than your competitor's.

The solution to external information asymmetry is internal capability.

Currently, many organizations are making AI investment decisions without structured intelligence on how competitors are deploying. I built *benchmarkai* to test the information asymmetry gap between strategy and value creation.

I am an MBA Candidate at IMD Business School, with 6 years of experience scaling technical operations in the sports industry. Systems thinking is at the core of how I approach strategy: a habit built across industries and sharpened further when I flew to IMD Singapore for a one-month immersion program focused on AI transformation. Not only is AI literacy becoming critical for businesses to maintain a competitive advantage, but that alone is not enough: as AI transforms workflows, firms must balance competitive intelligence with strategic integration to realize business impact.

Mapping AI Maturity

The firms with the loudest signals do not necessarily demonstrate the most substantial AI integration. Once the tool worked, I turned it on the market: a 2025 MIT report found that 90% of surveyed employees use personal AI tools.[1] While individual productivity is being measurably impacted—activity that has disproportionate impact in small organizations like startups—large corporate P&Ls are not reflecting this value creation.[1]

The tool was built to identify signals, but it began to reveal a pattern. Across the reports I ran, firms clustered into four groups assessed on the pillars of AI Integration Depth, Workforce Transformation, and Quantified Business Impact.

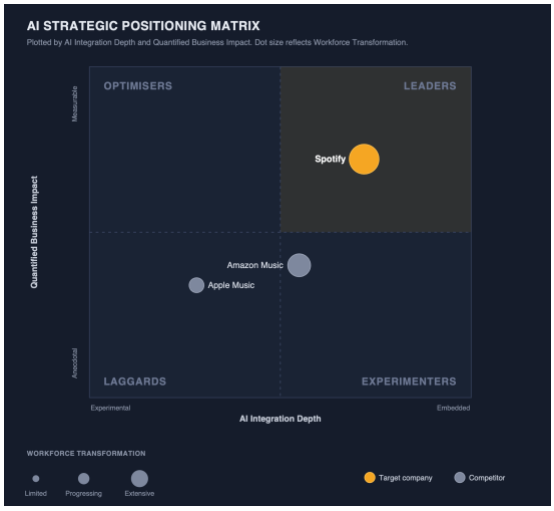
Each of the four groups provides unique insights but invite deeper scrutiny**.** Broad activity limits **experimenters** from maximizing business impact, while **optimizers** see high impact through specific processes but lack the stamina to integrate mature AI systems cross-organization. **Leaders** practice strong iterative processes alongside strategic depth to realize expansive value creation. Public signaling that indicates a **laggard** must be considered for two root causes: one, that the firm truly lacks AI integration or two, that the firm signals little quantified impact. This group highlights the core tension between signal and substance.

I hypothesize that most firms stop at experimentation or optimization because they lack the structure to strategize for the next stage.

Organizations that fully integrate AI beyond the pilot stage are 4x more likely to report revenue growth (58% versus 15%).[2] Most firms stop at experimentation and fail to scale AI integration into strategic depth of implementation: nearly two-thirds of firms have not begun scaling AI across the enterprise.[6] *The GenAI Divide: State of AI in Business 2025* report states that projects delivering value can be identified by three key characteristics: deep integration into specific processes, continuous learning capability and evaluation based on business outcomes.[1] This framework maps strongly onto the AI Strategic Positioning Matrix and is the structure that many firms lack. The *2026 AI at Work Report* further supports these findings: 16% of businesses reported negative returns on AI investment and 73% of businesses with measurable impact reported that ROI fell short of their expectations.[7]

The Apple Music Paradox

Building *benchmarkai* taught me to distrust confident output and running it enabled strategic insight. *benchmarkai* highlights this information asymmetry in the example report on Spotify, Apple Music, and Amazon Music.



Apple Music lands in Laggards despite significant AI spending – demonstrating the gap between disclosure and substance.

Apple Music's parent company, Apple Inc., announced significant AI spending [5], yet the music service division scores low confidence and has no disclosed AI features comparable to Spotify's AI DJ. Amazon Music's Maestro is still in beta with no quantified impact reported. A company could benchmark itself against Apple Music and conclude it is behind on AI, when the likelier explanation is Apple Music simply has not disclosed business impact. This reveals the distinction between signal volume and intelligence quality.

The score gap is a disclosure gap. Where Spotify shows shipped features and "Disclosed margin gains", Apple Music shows "Undisclosed" and "Not reported". Loud signal is not the same as substantial signal.

benchmarkai is correctly separating three different signal-substance relationships. Signal strength measures how well-covered a company is by public evidence found for the report. Spotify pairs strong signal with disclosed substance, an example of the report rewarding strong evidence. Apple Music generates plenty of general AI signals but discloses nothing that matters on these dimensions. Amazon Music is the inverse—quieter, but with concrete, if not fully substantiated, signals. The point is not to score Apple Music but to stop a strategist from concluding that Apple is behind. The user makes the final judgement call, not the AI-generated report.

Mispricing the AI Economy

The boundary of *benchmarkai* is the strategic point: the market is pricing AI maturity off signals that systematically mislead.

AI native startups are scaling at record speeds: Anthropic and OpenAI have both filed confidentially to go public at target valuations near $1 trillion.[8] Yet while this capital is flowing, 95% of firms are seeing no measurable P&L impact from AI.[1] With huge investment

by private firms driving market expansion yet a lack of firm-level value creation, the market is pricing signal disclosure not value. The information asymmetry that misleads one firm benchmarking against competitors is, at scale, mispricing the entire AI economy. *benchmarkai* scores public disclosure but its built-in limitation—that disclosure is not substance—is a firm level example of market-level failure. This is the boundary where the tool's limitation becomes the point.

An exception to the rule is China's DeepSeek, using cost-efficient training and open-weight distribution strategy to build the model at a fraction of the cost of the US equivalent.[4] This strategy drove genuine frontier parity and adoption followed on price and efficiency; the top Chinese model is competing for performance with US frontier models with a gap of just 2.7%.[4] Separating signal from noise is what will drive real value creation, for a single firm and for the market pricing them.

Concept

I built *benchmarkai* to provide quick reports with high-level insights regarding AI integration and impact. I structured the output to augment human judgement, not replace it. Visual aids, a clearly defined methodology, and accessible references establish a clear foundation for interpretation and transparency. Each report scores firms on explicit behavioral data, cites every source, and flags where evidence is weak.

benchmarkai evaluates larger, digitally-visible corporations against three criteria:

1. **AI Integration Depth**
2. **Workforce Transformation**
3. **Quantified Business Impact**

These pillars are determined based on *The GenAI Divide: State of AI in Business*; a 2025 report out of MIT on how firms can deliver value, not just noise.[1] Workforce Transformation and AI Integration Depth are interlocked capabilities that drive Quantified Business Impact. The criteria are not independent factors but progress checks at different points in an organization. Together, they create a multidimensional health check of AI maturity. In the AI Strategic Positioning Matrix, the off-diagonal positions are where capability and impact decouple. The final report is directionless on its own: human judgement turns output into actionable intelligence.

Design

benchmarkai reports are built on three layers of intelligence: benchmarking, live signals, and human judgement. The first layer uses reports from McKinsey and Co., IMD Business School,

Stanford University, and OECD to frame the competitors and direct the live signaling. This is the only continuous layer across all reports. The second layer pulls public data from the web for live signaling on firm-specific activity. The third layer employs the user as an active catalyst for synthesizing reported insights into actionable intelligence. A two-minute run time, repeatable methodology, and organized framework make it possible to check the pulse quickly on any business grappling with AI impact.

Build

AI tools are defined by their limitations. ***benchmarkai*** **reports public signals, not substance. Human judgement unlocks strategic intelligence.**

The first key to *benchmarkai* is transparent methodology.

In a time when noise can be mistaken for signals, high-integrity data is key to building user trust. Reports like the EY *Points of Attack*—retracted after 60% of references were alleged to be hallucinated—are calling into question the integrity of AI content.[3] While AI is a powerful tool for synthesizing data, clear framing and academically standardized references are essential for trustworthy reporting.

In general, a one-off report for a competitor firm is useless—a repeatable pulse-check of activity becomes an actionable strategy. 80 notable AI models were produced in the US and China in 2025 yet that record output has not translated into organizational impact: only 39% of firms report any EBIT contribution from using AI.[4][6] Strategists need replicable intelligence that includes the most recent signals without losing sight of the foundational context. *benchmarkai* explores the relationship between timely insights and trustworthy information.

benchmarkai's built-in limitations enable trustworthy reporting.

Qualitative visualizations and confidence level assessments deliberately invite criticism. While some references may be strongly supported by multiple primary sources, low level insights are also included when the live signal search derives insights from unverified sources. The blended reporting is paired with critical analysis to invite human judgement as the key mechanism for securing conclusive insights. Every report is an executive level summary of the loudest public signals. Every conclusion is the critical analysis of the user. This critical layer intentionally highlights the necessity of collaboration between AI reporting and human judgement.

See the full methodology: https://benchmarkai.live/methodology

Key Learnings

Programs like *Claude* make coding more accessible than ever because they allow developers to communicate using natural language and the LLM translates intent into output. And *Claude* can write code quickly. However, the coding output risks misalignment with user intent because at the heart of AI coding is the fact that LLMs are probabilistic in all outputs, not deterministic. My design choices when building *benchmarkai* were structured to align trust and build accountability between human and AI. I interrogated every build choice with this core question:

How can AI build effective tools when relying on probabilistic outputs?

It took five hours to draft *benchmarkai*; it took a weekend to build a structurally-sound architecture. Prompting the output was fast but refining the methodology was tedious—although *Claude* built an MVP in minutes, the underlying architecture was buggy and inconsistent at executing tasks. A core challenge when coding with *Claude* is that AI is path dependent: after the first output, subsequent code is layered over existing architecture. Prompting *Claude* to re-evaluate the foundational layers of the tool was a battle against this bias.

Three takeaways came out of the build process. Each learning began as a lesson about coding with AI but became an insight about AI strategy:

1. **AI can build quickly, but not always efficiently.**
   When a user prompts AI in natural language, the computer infers output, and the primary risk is misinterpretation. Accuracy and efficiency shift together: the less precise the output, the more work it takes to trust. The strategic trap that most firms fall into is mistaking the speed of AI activity for the value of AI impact.

2. **AI cannot infer user intent.**
   A prompt defines the parameters; a human defines the question. AI tools are only as robust as the critical thinking behind them. This is why *benchmarkai* makes a human the final analytical layer. AI can surface the loudest signals but only a human judgement can separate disclosure from reality. An organization that lets the model do the interpreting inherits its blind spots: the information asymmetry remains invisible.

3. **AI tools are never finished.**
   An AI tool is bound by the model, the signals, and the moment it was built in—all of which keep moving. Updating, testing, and iterating remain necessary decisions in the lifetime of an AI build. The strategic parallel is the sharpest lesson of all: strategic intelligence is not a document you produce but a capability you sustain. Firms that treat AI adoption as a project to finish stall at experimentation; the leaders pull ahead by updating strategic thinking, constantly.

Asking the Right Question

benchmarkai works because the signals are loud; but signaling is not a proxy for business impact. As firms update workflows at an unprecedented rate, just 5% of AI enterprise pilots are capturing measurable value.[1] With 70% of executives prepared to scale back AI integration if returns are not met, the window for course correction is narrowing.[7] Organizations that succeed at holding tension between experimentation and strategic depth of implementation will be in the winning group.

When it comes to AI integration, every business is asking the same question: what - are my competitors doing?

This is the wrong question. Instead, we need to ask:

What stage of integration have we reached, what does the next stage require, and how do we mobilize the organization?

That is the question *benchmarkai* was built to answer: not where your competitors stand, but where you do. The next move depends on where you stall. An Experimenter—broad activity, no measurable return—should integrate one deployable use case with a P&L metric. An Optimizer—real impact with narrow experimentation—should take its working integration to another business function. The quadrant is the diagnosis; human judgement creates value.

Run a report: https://benchmarkai.live/

Notes

[1] Arjun Challapally, Caleb Pease, Ramesh Raskar, and Praneeth Chari, *The GenAI Divide: State of AI in Business 2025* (MIT NANDA, July 2025).

[2] Grant Thornton, *2026 AI Impact Survey Report* (Grant Thornton Advisors LLC, April 2026).

[3] Stephen Foley, *EY retracts study after researchers discover AI hallucinations* (Financial Times, 2026).

[4] Nils Maslej et al., *The 2026 AI Index Report* (Stanford University Human-Centered AI, April 2026).

[5] MacKenzie Sigalos and Jennifer Elias, *Apple's R&D Spending Climbs to 10% of Revenue on AI Investments* (CNBC, May 6, 2026).

[6] Alex Singla, Alexander Sukharevsky, Bryce Hall, Lareina Yee, and Michael Chui, *The State of AI in 2025: Agents, Innovation, and Transformation* (QuantumBlack, AI by McKinsey & Company, November 5, 2025).

[7] Globalization Partners, Wakefield Research, *AI at Work: The 2026 Reality Check* (Globalization Partners, May 2026).

[8] Tanner Stening, *Why Anthropic,* OpenAI and SpaceX are all racing to go public now (Northeastern Global News, June 8, 2026).

Claude Sonnet 4.6 (Anthropic, 2025) was used throughout the drafting process as an editorial partner: providing structural feedback, evaluating argument cohesion, conducting source validation, and suggesting line-level revisions. All analytical claims, strategic interpretations, and conclusions are the author's own. benchmarkai was built using Claude Code. AI-assisted editing does not imply AI authorship.