Guest editor
Eystein Thanisch

Contributing authors
Joe Allen
Paul Clermont
V. Kavida
Chirag Kundalia
Dan North
Michael Papadopoulos
Olivier Pilot

CUTTER
AN ARTHUR D. LITTLE COMMUNITY

# AMPLIFY

Vol. 38, No. 6, 2025

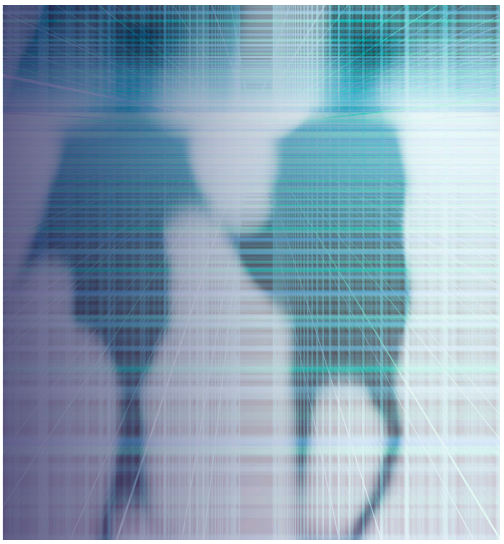
Anticipate, Innovate, Transform

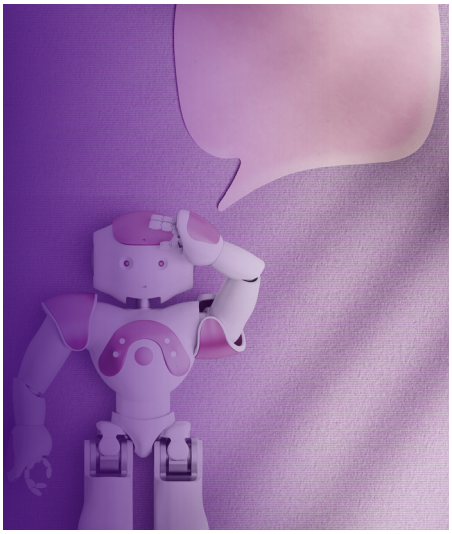
## Disciplining AI, Part II: Looping in Humans, Systems & Accountability

# CONTENT

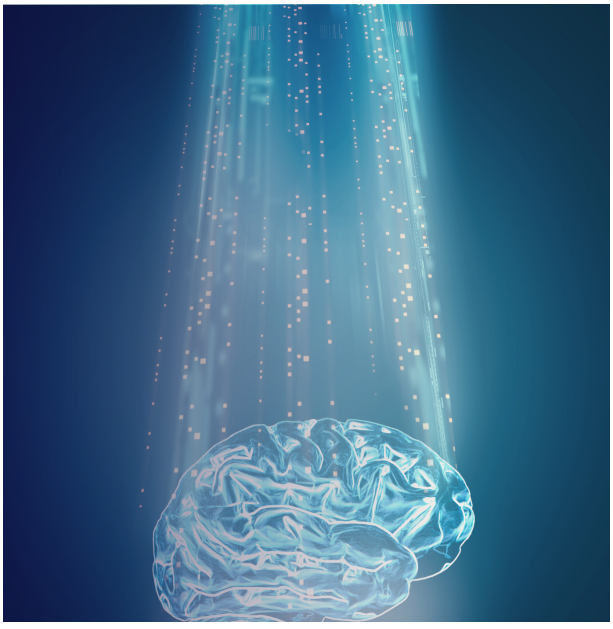












# DISCIPLINING AI, PART II: LOOPING IN HUMANS, SYSTEMS & ACCOUNTABILITY

BY EYSTEIN THANISCH, GUEST EDITOR

**Part I of this two-part *Amplify* series on disciplining AI considered criteria for AI success across industries and social domains.[1] Here in Part II, we closely examine how to evaluate against one's criteria when building and governing AI systems.**

In the short time since Part I was published, concerns have continued to be raised about AI's application in the real world. Consumer safety is one aspect.[2] In an especially troubling case, the parents of a teenager who died by suicide under advice from ChatGPT are suing OpenAI, with OpenAI already acknowledging flaws in its chatbot's guardrails.[3]

In the enterprise context, recently released preliminary findings from an MIT report (the latest in a string of surveys showing troubles with AI adoption) suggest 95% of organizations that invested in generative AI (GenAI) over the past three years have yet to see a positive ROI.[4]

Both GenAI model performance and enterprise adoption appear to be plateauing.[5] The former was exemplified for many by the launch in August of GPT-5, with intelligence gains that initially appeared modest compared to both the hype and previous major model releases. The past few months have been dubbed by some as AI's "cruel summer."[6]

Speculation that the AI bubble might be bursting has intensified, but the technology won't necessarily be leaving us here on our own (to paraphrase Bananarama's pop music hit "Cruel Summer"). Hype cycles are well understood, and many core technologies have been through one. AI could continue to improve as one technology among many, essential for some tasks, complementary for others, but a long way from transcending into artificial general intelligence or superintelligence.[7] Self-styled AI realists argue that the route to the latter will be through more composite approaches and not through reliance on large language models (LLMs), which will also require commercial adjustment.[8]

If GenAI is going to remain fallible for the foreseeable future and exist as a component in a complex system, then oversight of its outputs is vital. This is by no means just about policing the AI's bad behavior; it's about unlocking real value in situ. It is illustrative that, after initial disappointment, GPT-5 appears to be held in increasingly high regard (see Figure 1). Progress is not always apparent at first glance; truly productive applications sometimes require adaptation and localization.

This does not occur naturally. According to the previously mentioned MIT report, a key differentiator between successful and stalled GenAI pilots is the capacity of the pilot system to learn and improve in response to feedback, as well as to dynamically adapt to context.[9] Part I of this *Amplify* series showed us that oversight can be complex and that context can be varied.

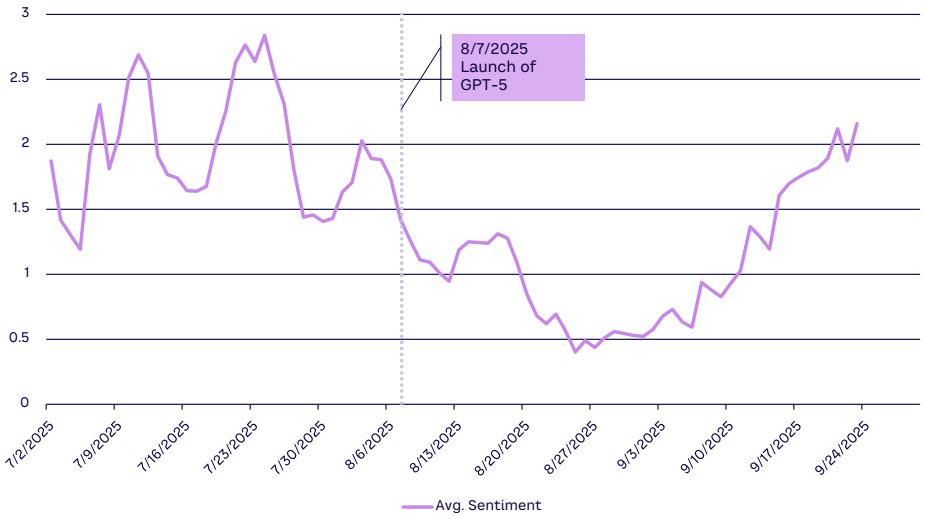



**Figure 1. GDELT (Global Database of Events, Language, and Tone) sentiment on GTP-5, seven-day rolling average (source: Arthur D. Little, GDELT)**

## IN THIS ISSUE

The contributors to this second issue explore what we're trying to do when we systematically evaluate AI systems and look at some of the philosophical-technical challenges and solutions involved.

Frequent *Amplify* contributor Paul Clermont opens the issue by reminding us of a crucial but under-recognized trait of true AI: learning and improvement in response to feedback independent of explicit human design. The implementation of human requirements is thus both highly feasible and dialogic. Clermont stresses the high level of responsibility borne by humans when interacting with AI. Critical thinking about inputs and outputs, and awareness of both objectives and social context, remain firmly human (and sometimes regulatory) responsibilities — no matter how widely terms like "AI accountability" have gained currency.

How can we ensure accountability for AI systems, especially as they scale and grow more complex? Looking specifically at LLMs, my Arthur D. Little (ADL) colleague Dan North examines the engineering discipline of AI evaluations, arguing that this is the locus of an LLM's translation from task-agnostic capabilities measured by benchmarks to tech that is setting specific and ready-to-deliver success. He emphasizes the ultimately human nature of this discipline. An organization must define the kind of outputs it is looking for through an inclusive process involving customers and stakeholders, alongside engineers.[10]

Such definition may be far from simple. For complex AI systems, the nature of the required output may differ markedly from step to step, and this might not be apparent "above the hood." AI can provide support by generating test data and evaluating LLM outputs against criteria, but without informed, judicious human involvement, there are no meaningful criteria.

My ADL colleagues Michael Papadopoulos, Olivier Pilot, and I agree that with model performance plateauing, the critical differentiator for AI in 2025 is how the technology is integrated into the specifics of an organization — and the context in which it will operate. Benchmarking, which is concerned with models out of context, is of diminishing applicability.

After reviewing some entertaining but troubling examples of theoretically capable AI systems disconnecting disastrously from data, business rules, and basic plausibility, we propose an evaluation framework more attuned to contemporary challenges to prevent similar outcomes. The key properties are oversight, explainability, and proximity (to data, process, and policy). Again, evaluations play a central role in orchestrating the components of the system, holistically understood.

Next, Joe Allen joins the critique of benchmarks as the key means of measuring AI systems, with a focus on agents. Given the open world in which they operate, agents have autonomy over both planning and execution, and Allen argues they have already outgrown even the criteria used to evaluate LLMs. As self-organizing systems,

agents should be assessed on their ability to follow their own plans consistently while adapting to unforeseen eventualities — an evaluation that cannot be fully scripted in advance.

Furthermore, agents are increasingly acting in the real world. Unlike in even mission-critical conventional systems, they do not behave deterministically. The stakes are thus higher in terms of risk, reward, and uncertainty. Drawing on direct experience, Allen details a suite of techniques that can be used to track and improve agent performance in terms of internal coherence, adherence to a context model, and more. The approach is iterative and continual, but that is what's required for such a dynamic technology.

So far, the focus of this *Amplify* issue has been on evaluating and incrementally improving an existing AI system. Looking further ahead, how can one foresee when the entire system will no longer be enough?

In our final article, Chirag Kundalia and V. Kavida propose using survival analysis to understand when, and under what circumstances, an AI system might need maintenance or replacement. This approach involves modeling the intrinsic and extrinsic factors that could render a system no longer fit for purpose. It also requires organizations to define the standard of output the system must deliver, monitor that performance, and scan the horizon for relevant externalities (e.g., superior technology). Although survival analysis cannot anticipate every eventuality, modeling the future of an AI system enhances budgeting, compliance reporting, and strategic planning.

## KEY THEMES

Organizations implementing AI systems must overtly define what "good" looks like. Guardrails and quality control are important — sometimes critically so — and AI evaluations are a principle means by which the organization can steer this nondeterministic technology and define its purpose for them. Confining it to an engineering subdiscipline would be a waste of a key opportunity. We might be tempted to let AI take the lead and tell us something we don't know. This might prove beneficial when using GenAI for research or "vibe coding," but for serious applications, one must be directive about purpose and context.

Having a framework in place to measure one's AI system against purpose and context leverages this technology's ability to learn and adapt without extensive rewiring. Data from evaluations can feed both into strategic planning and straight back into the AI system.

Explainability also emerges as key.[11] Evaluations are greatly enriched when one can go beyond playing trial-and-error with outcomes and understand why the system behaved the way it did. This is particularly important for agents, which may be following an evolving plan they have put together themselves. Also, if a range of stakeholders is going to be involved in steering the AI system, explainability is needed for that to be an open conversation among equals.

Identifying what an organization is looking for from a system is challenging, as is integrating that system with context and making it explainable in an effective way. Multiple contributors to this *Amplify* issue reference Goodhart's Law, under which well-intentioned application of metrics perverts a system's priorities.

The challenge intensifies the more general-purpose a system is, even as the various possible purposes in any given interaction multiply, increasing the complexity of the metrics required.

The peak challenge may be posed by consumer chatbots like ChatGPT or Claude. To illustrate, a recent OpenAI paper argued that much-maligned hallucinations in LLMs arise from post-training that discourages models from admitting they don't know and pushes them to guess instead.[12]

A human design decision is required as to whether what is sought is transparent caution, proactive risk-taking, or the kind of gradation of certainty that characterizes much human expression. Although not always reaching such levels of profundity, organizations should consider this kind of question when evaluating an AI system. Familiar disciplines like quality assurance, user experience research, software testing, and requirements analysis are all relevant.

Given increased responsibilities and autonomy of the systems, however, we find them remixed and applied to new questions and behaviors. The contributions presented in this issue add valuable insights into how this might be done rigorously and with cognizance of what is really needed.

## ACKNOWLEDGMENTS

Alongside the contributors and the ADL Cutter team, I would like to thank Lara Arnason, Natalie Demblon, Brian Lever, Greg Smith, and Oliver Turnbull for their thoughtful input.

## REFERENCES

[1] Thanisch, Eystein (ed.). "Disciplining AI, Part I: Evaluation Through Industry Lenses." *Amplify*, Vol. 38, No. 5, 2025.

[2] Nelson, Alondra. "An ELSI for AI: Learning from Genetics to Govern Algorithms." *Science*, Vol. 389, No. 6765, September 2025.

[3] Yousif, Nadine. "Parents of Teenager Who Took His Own Life Sue OpenAI." BBC, 27 August 2025.

[4] Challapally, Aditya, et al. "State of AI in Business 2025." MIT Project NANDA (Networked AI Agents in Decentralized Architecture), July 2025.

[5] Kahn, Jeremy. "Is the AI Bubble About to Burst? There's Mounting Data Suggesting It Might Be." *Fortune*, 9 September 2025.

[6] Nelson (see 2).

[7] Narayanan, Arvind, and Sayash Kapoor. "AI as Normal Technology." Knight First Amendment Institute at Columbia University, 15 April 2025.

[8] McMahon, Bryan. "What If There's No AGI?" *The American Prospect*, 4 September 2025.

[9] Challapally et al. (see 4).

[10] For more on how modeling intended AI outcomes brings stakeholders together, see: Farrington, Joseph. "A Simulation-First Approach to AI Development." *Amplify*, Vol. 38, No. 5, 2025.

[11] As shown in Part I, it is also a legal requirement in many contexts; see: Nance, Rosie, et al. "Explain Yourself: The Legal Requirements Governing Explainability." *Amplify*, Vol. 38, No. 5, 2025.

[12] Kalai, Adam Tauman, et al. "Why Language Models Hallucinate." arXiv preprint, 4 September 2025.

## *About the guest editor*

### EYSTEIN THANISCH

Eystein Thanisch is a Senior Technologist with Arthur D. Little (ADL) Catalyst. He enjoys ambitious projects that involve connecting heterogeneous datasets to yield insights into complex, real-world problems and believes in uniting depth of knowledge with technical excellence to build things of real value. Dr. Thanisch is also interested in techniques from natural language processing and beyond for extracting structured data from texts. Prior to joining ADL, he worked on Faclair na Gàidhlig, the historical dictionary of Scottish Gaelic, on a team tasked with building a tagged corpus of transcriptions from pre-modern manuscripts. He also was involved in IrishGen, a project on the use of knowledge graphs to represent medieval genealogical texts. Dr. Thanisch also worked as a freelance editor and analyst for a number of IGOs and academics. He earned a master of science degree in computer science from Birkbeck, University of London, and a PhD in Celtic studies from the University of Edinburgh, Scotland. He can be reached at experts@cutter.com.

# ACCOUNTABLE AI?

Author

Paul Clermont

**The term "artificial intelligence" goes back to 1956, but the general public heard little of it beyond occasional headlines when IBM's Deep Blue beat the world champion chess grand master in 1997 and Watson beat the all-time *Jeopardy* winner in 2010. Remarkable achievements, but like putting men on the Moon, unrelated to daily life.**

Behind the scenes, serious money was being invested by serious people, and in late 2022, ChatGPT startled the world. We could type in a natural language request and — within seconds — get sensible answers in readable, natural-sounding, grammatically correct language. Suddenly, AI was all over the news, and a broad swath of the public started using one or more of the products that rolled out.

Simultaneously, businesses and governments began exploring and implementing AI for everyday processes. Today, it's not difficult to envision all kinds of routine office work being done better/faster/*much* cheaper with AI. It's also easy to envision overdependence on AI to the point where no one understands how it works when something goes wrong, propagating problems that take months or years (if ever) to unravel.[1]

## WHAT IS AI REALLY?

As often happens when an idea emerges that promises great profit opportunities, there's a bandwagon effect. Executives claim to be implementing it, even if the application is trivial or does not embody the idea, strictly speaking. Indeed, some seem to be using the term "AI" to describe any smart application that is supposed to make the kind of intelligent decisions once the sole province of humans.

No matter how sophisticated it is, if every step is prescribed by the designer, it's **not AI** (it's the designer's natural intelligence — nothing artificial). If the application is designed to learn to improve its performance, it's AI.

**Generative AI (GenAI),** by far the most visible, is trained on text and images and can respond to natural language questions and directions, producing answers or appropriate images. Since OpenAI introduced ChatGPT in 2022, it has released improved versions and been joined by competing products. They require huge tables (large language models [LLMs]) to find relevant information and turn it into good-quality text, and they're designed for wide public use. A subset of GenAI is domain-specific GenAI. It's the same basic idea but with smaller, thoroughly vetted databases relevant to a specific knowledge domain (e.g., protein folding). These systems are designed by experts for experts, not the public.

**Real-time process control** has gone public in the form of robotaxis plying the streets of San Francisco, California, and other cities. They use AI to integrate continuous inputs from multiple cameras plus multiple radar, LiDAR (light detection and ranging), and acoustic sensors to establish situational awareness of their vicinity.

**Abnormality identification** involves distinguishing signal from noise in tasks like scanning X-rays or surveilling facilities. Both this and real-time process control involve on-the-job training by humans. A different but conceptually related application is the use of drones to identify the best targets for military action and course-correct in flight to hit them.

**Facial recognition,** while a natural for AI, is highly controversial because it performs less well with darker-skinned people, leading to false matches (and/or failures to match) for racial minorities. Darker-skinned people are often minoritized in training data. Increasing their representation may improve these products, but there is no guarantee.

**Combinatorial challenges** are the oldest AI application and exist in games like chess and *Go*, where success depends on identifying optimal moves, anticipating countermoves, and planning effective responses. Today, similar approaches are used to uncover vulnerabilities in our systems — or to proactively identify weaknesses in competitors' systems.

## AI'S ADVANTAGES ARE COMPELLING

Compared with humans:

- AI is faster by orders of magnitude.
- AI doesn't get bored, distracted, or tired (many auto accidents are due to momentary attention lapses[2]).
- AI is thorough; it doesn't cut corners that it hasn't learned are safe to cut.

Perhaps most importantly, AI can learn. In games, it learns from mistakes, avoiding moves that lead to dead ends. In robotaxi training, it improves through real-time feedback from a human driver. Over time, an AI system can build on its own history, adding value beyond what its original programmers designed in.

## GENAI IS NOT WITHOUT FRAILTIES

GenAI is a major accomplishment that descriptors like "stochastic parrot" tend to diminish.[3] That said, its shortcomings must be acknowledged.

First, GenAI lacks inherent common sense and has no intuitive ability to distinguish the implausible — or even the absurd — without targeted training. It may appear moral or politically correct if its training data leans that way (or in the opposite direction), but such behavior is not intrinsic. It can be tuned to favor positive, neutral, or skeptical feedback. For example, GPT-5 adopted a much less positive flavor than GPT4o, to the annoyance of some users, and GPT-4o's allegedly indiscriminate positivity may have contributed to the suicide of a teenage boy.[4,5]

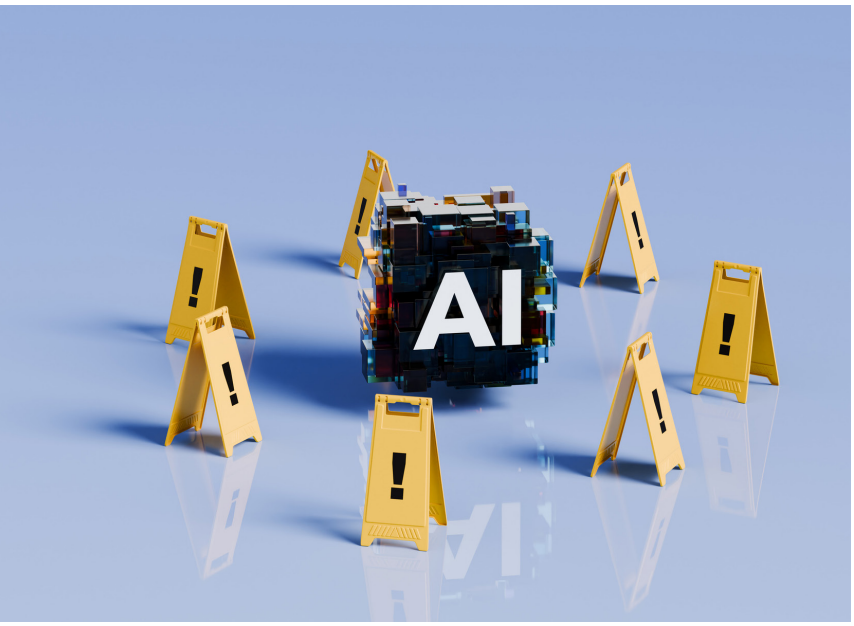
Second, GenAI has a propensity to make stuff up, which the industry euphemizes as "hallucinations." Unfortunately, it's good at this, to the dismay of lawyers who submitted briefs referring to nonexistent but plausible-sounding cases.[6] It's also good at making up plausible-sounding and correctly formatted citations in scholarly work.[7]

Recent attempts to build in some semblance of reasoning have, paradoxically, increased the incidence of hallucinations in some cases. Chain-of-thought reasoning that shows explicit steps reduces the incidence if the questioner has some notion of what the chain should look like and the initial premises are clear.[8] Some research has shown severe limitations in problem-solving capabilities.[9]

Third, GIGO (garbage in, garbage out) still applies. An AI application is no better than the information model it uses. There have been cases where innocent prompts elicited dangerous, outrageous, or obscene responses. This is called "going rogue" or adopting a "bad-boy persona," and a fair amount of attention has been paid to understanding how this happens (some bad stuff finds its way into the LLM) and developing techniques to get the model back on track.[10] There's also danger that future LLMs will incorporate too much unvetted output from earlier LLMs (e.g., scholarly or technical papers that have been challenged and retracted).[11] As with hallucinations, it's "caveat user" (or, let the user beware).

## 3 CHALLENGES

As organizations and societies race to adopt AI, it's easy to be swept up by bold promises and breathtaking demonstrations. But beneath the surface, critical weaknesses remain that demand clear-eyed attention. To separate genuine progress from misplaced optimism, we must confront three challenges.




### CHALLENGE 1: RETAIN SOME SKEPTICISM

One thing is crystal clear: if the stakes are high, trusting the early output of a GenAI request is not a good idea. One should pose the request more than once, using different wording and syntax and coming at it from different directions. Independent verification from other sources is ideal; if this isn't possible, there's no substitute for a good common-sense sniff test.

We must realize that the issues discussed above arise from the fact that a GenAI has no idea what we're talking about in our prompts. It's just a string of words that help it find useful data (if it's in the LLM) and formulate answers based on the probability distribution of what words follow others; hence the term "stochastic parrot."

That it works at all seems like a miracle, and it may be that the LLM approach simply cannot be refined to a sufficient level of trustworthiness for some tasks (e.g., agentic). That's not to minimize the achievements to date; it's to recognize that the technical approach has inherent limitations that, at the very least, call into question enthusiasts' forecasts of artificial general intelligence (AGI) as just a few steps down the current technical path.

I don't claim to understand the technology that could enable this huge leap, but I do know something about the IT industry's history of over-promising and underdelivering. If we take the term "general intelligence" literally (i.e., the intellectual capacity one would expect of a person considered generally intelligent), I suggest we not hold our breath for AGI. Any claims to having achieved it over the next few years should be met with skepticism. There will likely be large improvements over what we have today, but calling it AGI will almost certainly be based on a notion of general intelligence that's so diminished it would be hard to recognize as such.

### CHALLENGE 2: MEASURE AGAINST THE RIGHT STANDARDS

If AI is to be trusted, it must be measured against the right standards. Speed benchmarks are nearly meaningless — producing a hallucination faster is no achievement. What matters are quality and safety. Just as early automobiles were judged not by acceleration but by braking distance and reliability, today's AI must be assessed on trustworthiness.

That means asking questions such as: Does the system minimize hallucinations? Can it interpret prompts without being derailed by slight wording changes? Is it energy efficient, given the enormous power demands of large models? And, in specialized domains, is the training data sufficiently vetted to guarantee near-perfect accuracy?

Legal and ethical compliance is another key benchmark. At a minimum, AI systems must operate within the law. Ethics are harder to define, but violations can be equally damaging, especially in fields like medicine or law that have explicit professional codes.

For now, expecting AI to deliver measurable ROI is premature. *The New York Times* recently noted that billions invested in AI have yet to pay off in office productivity.[12] But this should not be surprising. Like earlier waves of computing, AI requires organizational adaptation before benefits emerge. That process takes time.

As AI evolves, evaluation methods must adapt. The temptation will be to seek cost savings quickly, often by cutting staff. But customer-facing applications show the risks of premature reliance. Who has not been infuriated by chatbots that fail to grasp a simple request? Quality of experience — whether for customers or employees — must come first.

### CHALLENGE 3: ADDRESS SOCIETAL ISSUES

Most inventions have provided good answers to "what can this do for us" questions, even if some ill effects took decades to materialize. AI is off to a different start: it raised "what can this technology do *to* us" questions right from the start.

Failure to address this can result in a backlash that ends in (1) throwing the baby out with the bathwater or (2) dystopian change. Problem areas include:

- Mass creation and instant global distribution of misinformation and disinformation could flood the zone with trash to the point where most people give up trying to find out what's really going on — creating an environment in which autocrats and oligarchs can easily wreak havoc. The rapidly improving quality of image and sound deepfakes increases the likelihood of this future.
- GenAIs could seem so human that some people forget they're just parrots that, if tuned to provide positive feedback, can encourage self-harm to the point of suicide.
- Material harmful to children and many adults could be mass-produced and quickly distributed.
- Invasions of privacy, harassment of individuals, and scams could be facilitated as databases are hacked (using AI to find security holes).
- The finding and exploiting of security holes in government and financial databases and physical process-control systems could lead to cyberwarfare.
- AI could be the first technology that creates massive persistent unemployment well beyond low-level jobs previously mechanized or automated away.

## CONCLUSION

For better and worse, AI is with us. The question mark in this article's title is deliberate. Obviously, the AI hardware and software can't be accountable, but those who use or distribute its products can be; hence, the importance of trustworthiness.

The current state of GenAI is far less trustworthy than it needs to be, and experts are increasingly questioning whether the technology can ever reach that level — casting doubt on the staggering scale of investments built on the assumption that it can and will.

Governments have a natural role in dealing with innovations like AI that have a potential downside for societies, but they're notoriously slow in addressing today's problems, never mind getting out in front of tomorrow's. Of course, that doesn't mean they shouldn't try. We must hope that the copious money Big Tech has available to spread around won't cause too many legislators and officials to look the other way.

We are living in increasingly interesting times.

## REFERENCES


1 "To err is human, but to really foul things up, you need a computer" — quote attributed to American biologist Paul Ehrlich.

2 "Transportation Institute Releases Findings on Driver Behavior and Crash Factors." Virginia Tech News, 20 April 2006.

3 Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery (ACM), March 2021.

4 Noreika, Alius. "OpenAI Restores GPT-4o After Encountering User Dissatisfaction With GPT-5." Technology.org, 11 August 2025.

[5] Yousif, Nadine. "Parents of Teenager Who Took His Own Life Sue OpenAI." BBC, 27 August 2025.

[6] Mangan, Dan. "Judge Sanctions Lawyers for Brief Written by AI with Fake Citations." CNBC, 22 June 2023.

[7] Gedeon, Joseph. "RFK Jr's 'Maha' Report Found to Contain Citations to Nonexistent Studies." *The Guardian*, 29 May 2025.

[8] Yao, Zijun, et al. "Are Reasoning Models More Prone to Hallucination?" arXiv preprint, 29 May 2025.

[9] Shojaee, Parchin, et al. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity." arXiv preprint, 18 July 2025.

[10] Hall, Peter. "OpenAI Can Rehabilitate a Model That Has Developed a 'Bad-Boy Persona.'" *MIT Technology Review*, 18 June 2025.

[11] Ananya. "AI Models Are Using Material from Retracted Scientific Papers." *MIT Technology Review*, 23 September 2025.

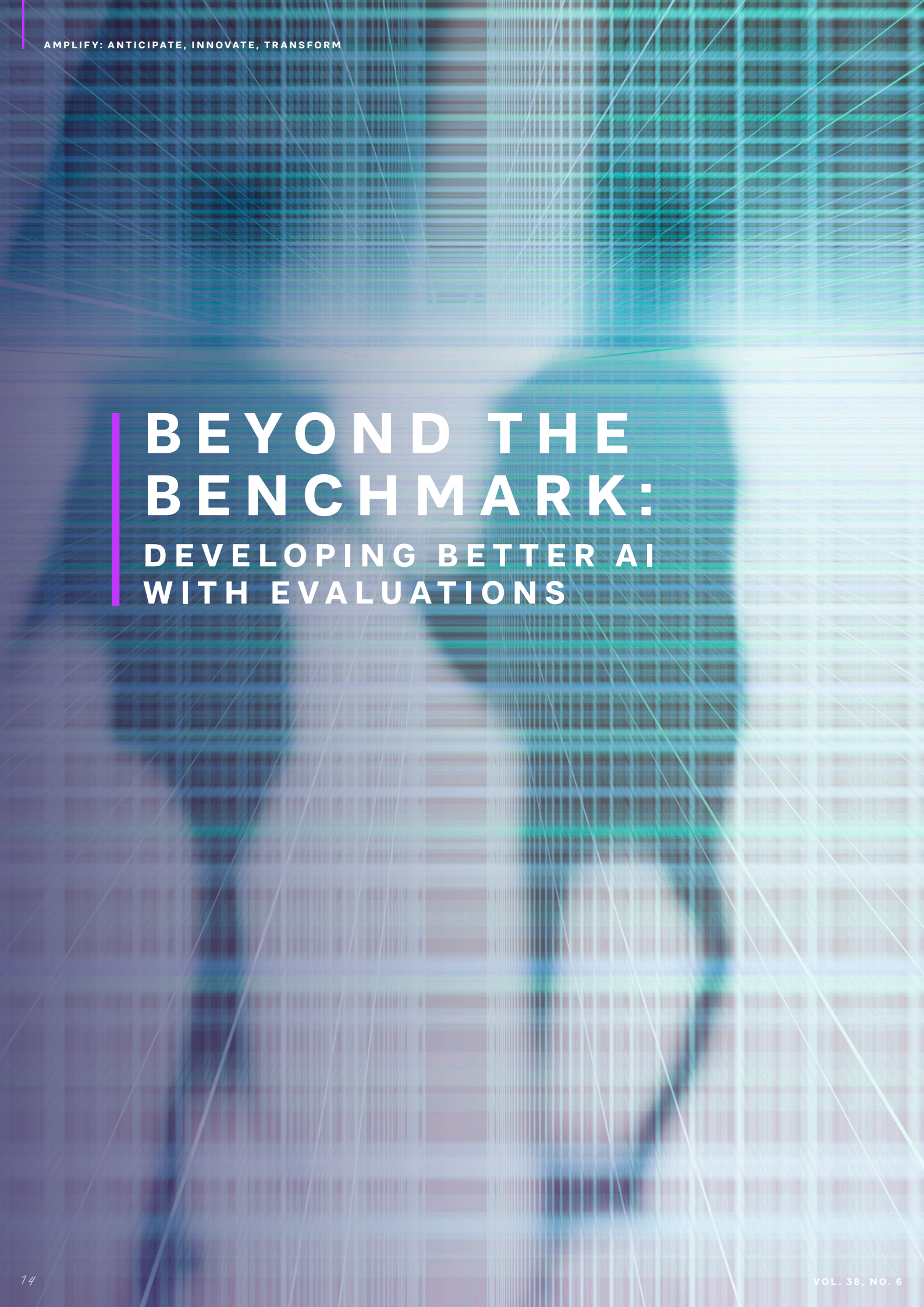
[12] Lohr, Steve. "Companies Are Pouring Billions into AI. It Has Yet to Pay Off." *The New York Times*, 13 August 2025.

# About the author

**Paul Clermont** has been a consultant in IT strategy, governance, and management for 40 years and is a founding member of Prometheus Endeavor, an informal group of veteran consultants in that field. His clients have been primarily in the financial and manufacturing industries, as well as the US government. Mr. Clermont takes a clear, practical view of how information technology can transform organizations and what it takes to direct both business people and technicians toward that end. His major practice areas include directing, managing, and organizing information technology; reengineering business processes to take full advantage of technology; and developing economic models and business plans.

Mr. Clermont is known for successfully communicating IT issues to general managers in a comprehensible, jargon-free way that frames decisions and describes their consequences in business terms. In his consulting engagements, he follows a pragmatic approach to the specific situation and players at hand and is not wedded to particular models, methodologies, or textbook solutions.

Before going into individual practice, Mr. Clermont was a Principal with Nolan, Norton & Co., a boutique consultancy that became part of KPMG. Before joining Nolan, Norton & Co., he directed IT strategy at a major Boston bank and launched its IT executive steering committee. Mr. Clermont has spoken and written about the challenges of getting significant and predictable value from IT investments and has taught executive MBA courses on the topic. His undergraduate and graduate education at MIT's Sloan School of Management was heavily oriented toward operations research. He can be reached at experts@cutter.com.



# BEYOND THE BENCHMARK:

## DEVELOPING BETTER AI WITH EVALUATIONS

Author

Dan North

**Evaluation criteria are a core part of deriving value from AI, unifying low-level code tests with high-level customer needs.[1] This article explains how to select evaluation criteria (a central yet underdiscussed step) and AI workflow design. These processes are likely to change quickly as the AI development ecosystem matures, but ultimately, human values remain central, so close collaboration between end user and developer is key.**

Unlike classical software, which is deterministic, large language models (LLMs) are stochastic — the same input could produce a range of possible outputs. That makes LLMs generative; it also makes them difficult to control.

Applying LLMs in a product requires controlling their outputs, so evals are required: evaluating the range of outputs you get from an input allows you to adjust the input to get the ones you want. Much of AI product development is now driven by evals — how you adapt the off-the-shelf technology to the specifics of your use case.

Crucially, data about these specifics is idiosyncratic to a given workflow and its end users. In AI, data quality is usually discussed in the context of LLM pretraining or fine-tuning, and evals are still often conflated with LLM benchmarking. Although these are important and could be included here, this article targets the newer sense of evals: assessment of LLM outputs in actual production workflows against criteria specific to those applications. (I exclude agent evals from this scope, since they're considerably different.[2])

Evals in this sense are different from the benchmarking used to train or fine-tune LLMs. Primarily, formal benchmarks are for generic, task-agnostic properties you would want the model to have in *any* setting; evals are for the task-specific properties you want the model to have in *your* setting.

Furthermore, most enterprise LLM workflows require evals, but fewer of them will require fine-tuning; fine-tuning can be more expensive and incurs sunk cost that slows your ability to switch models and adjust your workflow. So eval results will typically be used for in-context adjustments to LLM inference — that is, changing the information provided to an LLM at inference instead of pretraining, allowing businesses to iterate and respond to the market more quickly. That means the results will be applied to qualitative natural language inputs of the model, providing an ideal way to inject customers' success criteria directly into the generation request.

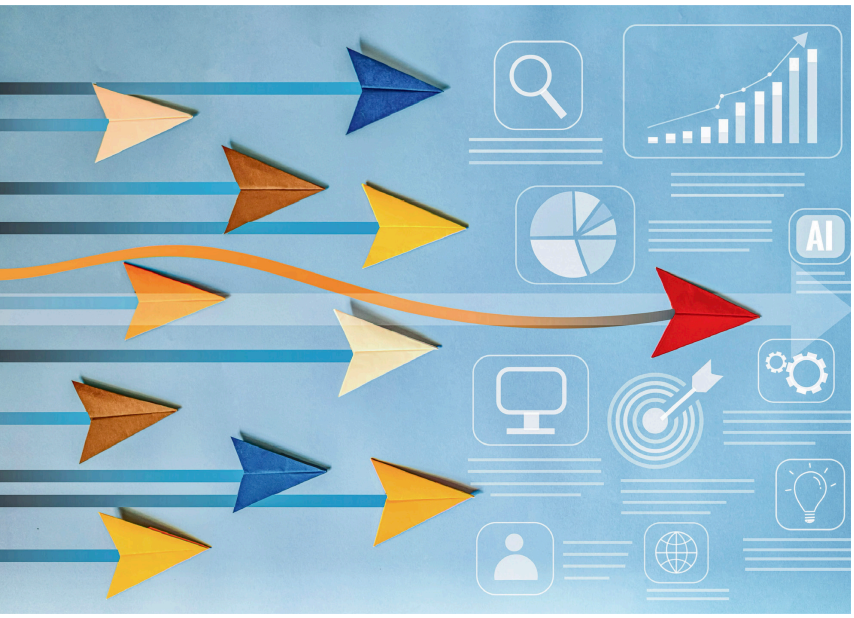
> **APPLYING LLMs IN A PRODUCT REQUIRES CONTROLLING THEIR OUTPUTS**

As any consultant or solutions architect knows, deriving a customer's success criteria is not simply reading off what they tell you. Much of our knowledge is tacit and enmeshed in a background of presuppositions. This difficulty is underscored when designing good evaluation criteria, because they require you to make explicit (in text) all the implicit "vibes" that divide success from failure. However, since all major LLM calls made in your application require evaluation, the criteria also resemble unit tests. So the assertions must incorporate preferences far beyond the scope of classical testing.

In short, authoring criteria requires a unification of the broad, customer-defined goals of a use case with the technical, narrow functionality of the LLM call being evaluated.

Good evaluations bring the tech to your use case. More broadly, AI promises a world where intelligence (sound, valid, and relevant inference from given data) is commoditized and cheap. If every business has access to graduate student–level intelligence for $.30 per million tokens, the level of differentiation is at the application, and, for reasons we've seen, user feedback is essential for developing AI-based applications.[3,4]

Unfortunately, the available guides and documentation give little guidance on how to use customer feedback to select and write the criteria you want to evaluate for.




## WHERE TO FIND THE CRITERIA

Manual annotation and authorship for evaluation criteria will always be required. However, it is best used on high-level outputs of the workflow to adduce the criteria for automated evaluations, which offer scale over low-level outputs. In addition, design principles of AI workflows inform exactly what these criteria should include.

A common theme in the extent documentation (such as from model providers like OpenAI or Anthropic) for implementing evaluation systems for AI applications is the importance of manual human analysis at some stage of the evaluation pipeline.[5]

The goal is to transfer human intuitions into the LLM, during either inference or pretraining, so that its range of output aligns with use cases and preferences. As mentioned, most uses of evaluation results involve context enrichment, so we must map the results into text. This involves performing error analysis to identify failure modes and adduce context refinements.

Performing this task manually for each evaluation is not scalable, so most solutions architects use an LLM as a judge (LaaJ) to do most of the annotation. This requires the judge LLM to be prompted and even evaluated itself. To do this, some AI experts suggest using a domain expert to hand-annotate a sample set of original LLM responses, explaining each success or failure, and another LLM to summarize these annotations.[6] Some recommend EvalGen, a sophisticated suite for iterating this process, and identify the now well-known criteria drift, indicating that end users refine their own self-reported success criteria on judging sequences of examples.[7]

Aptly named (from Greek krites, "judge"), criteria drift shows what psychologists and philosophers learned from Socrates: we don't really know our definition until challenged with examples.[8] Restricting answers to true and false, but requiring an explanation, forces us to say why the example does or doesn't meet the criteria.[9] This has the effect of making users' implicit assumptions explicit, and EvalGen capitalizes on LLMs' intended talent for detecting implicature.

Although an important tool in evaluation design, this method has limitations. The implicatures detected by LLMs in composing evaluation criteria are necessarily read from their text input; this limits the detection to presuppositions deducible from text. But implicature operates on many levels of information beyond text. In particular, much of the context that fills in the ellipses of our definitions is drawn from world knowledge, recent events, speaker relationships, conversation history, physical gesture, the surrounding environment, current time, and more.[10]

This means different information is communicated in a verbal conversation, especially in person, versus text. Humans evolved by encountering and interpreting each other in person, and our capacity for language use developed around this fact.[11] Fundamentally different types of information are available to and inferred by a developer in actual conversation with end users versus LLMs reading those users' text annotations.

There are also practical reasons for using insights gained in conversation from the customer when composing evaluation criteria. One engineering problem (which EvalGen proposes to address) is that LLM responses are often embedded deeply within a multistep workflow and thus not intelligible to a nontechnical user or anyone unfamiliar with the project code. One business problem is that, anecdotally, the hours of manual annotation required by automated processes are a significant demand from enterprise clients, most of whom do not have dedicated AI teams and are in completely unrelated industries.

It is also beneficial for the design process to include regular touchpoints and transparency between developer and user. In general, involving clients in the development process strengthens the developer relationship and increases their buy-in and likelihood to use the application.[12]

This approach is central to the forward-deployed engineer role popularized by Palantir Technologies. In its model, the engineer has direct, frequent meetings with customers, initially to discover the details of their use case, and subsequently to demonstrate the iterative work-in-progress application and receive their immediate feedback.

The key is close collaboration between the developer and the end user, not to give the customer more work, but to shape the development of the application from the user's direct comments. In my own experience, I found that conversations with the customer were crucial to capturing the implicit, unwritten heuristics at the core of their business.

The regression of LLM judges all-the-way-down can only terminate with humans in the loop, and eliciting customer feedback in live conversation is a useful way to do this, both for semantic and practical reasons. The manual step of developers synthesizing their insights gained from customer interactions and incorporating these into evaluation criteria is how extra-textual implicit information is passed to their workflow's AI.

## EVALUATING LLM WORKFLOWS

The human evaluation we have been discussing should be of the final output of the application: the presentation, report, and so on, to elicit the users' expectations of it. However, most value-adding AI apps do not produce this final output with a single LLM call.

Even relatively simple minimum viable products (MVPs) usually require a scaffold of multiple calls, each used for a specific and idiosyncratic purpose within the execution flow of the application and wrapped by various helper and parsing functions to integrate with that flow. This is because LLM outputs tend to be more successful the more predictable the desired I/O (input/output) pattern is (a corollary of the fact that artificial general intelligence is hard).

As we have seen, these are tedious to evaluate manually, so while LaaJ can and should be run on the final product, the way it adds value is by automating evaluation of these low-level LLM calls. Implementing LaaJ for these calls requires understanding how they're embedded in your workflow.

In fact, LLM workflows tend to decompose into individual calls that perform one of a few standard natural language processing (NLP) tasks that LLMs are good at. This is by design, since the transformer architecture used in current LLMs was originally created for machine translation, a core subfield of NLP.[13] These tasks are usually one of:

- **Classification.** What is this?
- **Extraction.** Where is this?
- **Summarization.** What does this mean?
- **Generation.** What's the most relevant response? (the bit that makes LLMs "intelligent")

In reality, the boundaries between these tasks are blurry. Nevertheless, when building an LLM workflow, it is helpful to think of them as building blocks to achieve the final output.

For example, earlier this year, I built an MVP for a veterinary hospital client who tracked patient data on 100% handwritten forms. The use case was to OCR (optical character recognition) the forms and map extracted text to invoice items for the client's CRM (customer relationship management) system, saving nurses the dozens of hours per week they spent doing this manually.

Because the inputs were in handwriting of varying legibility and included medical terms and abbreviations, the architecture required was more complex than simply passing an LLM the form and asking it what to invoice. The document image was first split into separate sections for semantically disparate parts of the form. Each part was then sent for an initial multimodal LLM pass A to extract the actual text (text recognition/extraction).

Because extraction alone achieved only 70% accuracy, even with lightly fine-tuned models, a further LLM call B mapped the results to a predefined list of common form items (classification). After combining the separate streams of extracted text, a final LLM call C mapped them to line items used in the client's invoicing system (classification). Although elaborate, this design was necessary to achieve the accuracy and reliability required by such a business-critical use case.

In this way, the LLM outputs used in a given workflow quickly become idiosyncratic to that particular workflow, which itself can be idiosyncratic to a given use case. So the criteria evaluating these outputs must assess the way they each contribute to the customer's success criteria, which constitute the ultimate purpose of the application.

At a first pass, the purpose here was to achieve complete, consistent, and accurate population of the client's invoicing CRM from handwritten forms. The intermediate purpose of, for example, LLM call B was to ensure each extracted element of text was mapped to standard items, so this information could be passed to the next call C. So the evaluation criterion for B is whether it completely, consistently, and accurately classifies raw text into standard items. Notice, however, that this criterion depends on the system's ultimate goal. If the final purpose of the system were something entirely different — for instance, to generate outputs that are funny or entertaining rather than accurate — then it would no longer make sense to judge call B by accuracy. Instead, we would evaluate it by how well it contributes to that new purpose (e.g., by how entertaining or funny the results are).

This is a purely theoretical criterion, with use-case-specific details waiting to be added. More concretely, different use cases can have very different success criteria and different types of data involved in their LLM calls. Currently, the core NLP tasks listed above provide a value proposition for enterprise in two broad types of use case: automation of administrative tasks (e.g., filling out forms, transferring data, sending routine messages or notifications, and performing time-consuming, boring tasks) and work-product creation (e.g., generation of leads, boilerplate documents and reports, due diligence and analysis, and presentation decks).

Success for the first type is mainly about achieving the correct outcome based on initial conditions. For the second type, it is more about meeting professional quality standards for the work product. The specifics of both change depending on the organization using them. The correct outcome for administrative tasks often depends on unwritten, heuristic rules developed by employees over time. In contrast, quality standards for work products often are contextual and vary between company, team, and individual.

A generic high-level criterion of "complete, consistent, and true" is perhaps sufficient for the task-agnostic process of LLM pretraining but not for the task-specific evaluations of LLM outputs as used in real-world applications.

Furthermore, as mentioned, there will be technical criteria reflecting errors specific to your implementation (e.g., generating too many words in a text response, including escape characters in a JSON output, or selecting the same item from a list when instructed not to). Pulling from a variety of sources, a list of automated eval criteria for call B could be something like:

1. The image of every mapping is a common item.
2. The pre-image of every mapping is in the raw text.
3. Every element of raw text that seems relevant is the pre-image of a mapping.
4. Each common item is uniquely the image of just one mapping (i.e., the map is bijective).
5. All common items returned are drawn from the list provided.
6. Any reference to copyright name a should be replaced by generic name b.
7. Raw text string s or similar should be interpreted as common item i instead of common item k.
8. If string s1 appears before s2, then map to common item j.
9. Only map to common item l when string s3 is present in the raw text.
10. No escape character appears in the response JSON.

Criteria 1-5 are basic definitions for the type of map required by this call in the workflow. Criteria 6-9 are heuristic criteria specific to how this administrative task is actually accomplished by the nurses (in my veterinary example), and criterion 10 is for response formatting. Because this is a classification task, it is framed as a simple map from a raw text string to a list of items. Of course, making the mapping more fuzzy or not exclusive to the provided list would require much richer heuristic criteria.

As we know from software product design in general, discovering the heuristics and type of mapping idiosyncratic to a particular use case and audience cannot be done from an armchair. The idiosyncrasies here are contingent on the facts of the specific environment the application is deployed in and thus not deducible a priori. Instead, the developer must collect specific feedback from end users — that is, people who will actually use the application.

It's critical that the evaluation for a given LLM output target how that particular call contributes to the broader purpose of the system it's in. This is neither reducible to generic benchmarks nor deducible from either the prompt or the use case alone. Instead, it is an application of the success criteria from the customer to that output's place in the workflow.

## THE FUTURE OF EVALUATIONS

This discussion assumes the current state of the field, in which evals and the tooling that makes them useful must be written from scratch for each application, with lots of humans in the loop. Of course, the state of this field doesn't stay in one place for long.

Two years ago, LLM capability was barely mature enough to provide enterprise value. Today, it is relied on for many of the tasks previously assigned to junior software developers. In another two years, even with linear improvement, we expect many of the processes described in the previous section to be automated.

Building a pipeline to generate test cases, run evals on them, and send the results to a self-refinement step currently requires significant manual integration with your app and how it implements an AI workflow. But what about these guidelines changes in a world where an AI coding agent built that entire workflow, so it knows how to build the integrated evals pipeline, too?

Several vendors offer evaluation tools that automate some of the work. As mentioned, OpenAI has an evaluations suite in its API and Web platform. Currently, it runs the evaluation and calculates scores for you, but this should expand over time to automatically run evals on calls made to its models in your application. Perhaps it could offer a functionality to group such calls into a workflow or project, providing the kind of architectural information discussed in the previous section. Or perhaps it could offer integrations with developers' self-refinement pipelines, expose its prompt optimizer in the API, or even host these pipelines itself. These functionalities make sense for closed-source model providers, since they already know exactly what you're sending to and receiving from their models.

There are also third-party platforms offering more advanced capabilities. Scale is the largest and best-funded, offering human-annotated data for LLM pretraining. It also brings human-in-the-loop evaluations to enterprise applications and offers some ability to use eval results for prompt refinement. However, the humans in question are anonymous and unrelated to the end users of your app.

Flow AI specializes in generating synthetic test data for agents, which is more complex, and offers an open source model fine-tuned for LaaJ. These third parties target developers or organizations that prefer open source, or at least the choice of model provider, but they run the risk of being made obsolete by major model providers that already have access to the critical data and inexpensive compute.

It is unclear how the market forces between closed versus open source (or weight) and single versus multiple model provider will play out in the AI space, especially if the cost of writing code trends toward zero. Nevertheless, developers will likely have far more tools and prebuilt infrastructure at their disposal in the near future for building evaluation suites and even entire AI applications and integrations, which themselves may look different in a world where AI communications protocols like MCP (Model Context Protocol) and A2A (Agent2Agent) are as important to our digital infrastructure as HTTP (Hypertext Transfer Protocol). What happens when the entire process of writing and running evals is as simple as calling an agent hosted in the cloud?

Advances in fundamental research may also change how models and applications are evaluated, potentially streamlining the collection and incorporation of user feedback. For example, achieving interpretability of deep learning models may allow us to more precisely target the weights implicated in desirable or undesirable behavior.

One approach to this is neuro-symbolic AI, which integrates neural nets with classical logic-based approaches to intelligence.[14] This is interesting because it could allow us to bridge stochastic and deterministic types of processing, getting the best of both worlds. Although there are many different branches, a common goal of the field is to logically structure model inference in a more direct way than scaffolding or simple prompting.

In a similar vein, alignment of deep learning models to human goals and preferences could benefit from formal control theory, which attempts to state these explicitly, such that model behavior necessarily adheres to them.[15] Much work remains to be done, but the hope is that, instead of an experimental and iterative approach to evaluations, we could give the model itself formal constraints similar to the assertions of traditional unit tests.

I believe the effect of a more automated world will be to increase the value of human judgments — and originality more generally. It may take a different form than the current human-in-the-loop architecture, but we are building these applications for our preferences, our values, and ourselves.

Even if much of the engineering becomes abstracted away, end users and their needs will remain. App development would still be behavior-driven, but the behavior would be much higher level and more accessible to the end user, reducing the friction involved in translating implicit, heuristic user needs into evaluation criteria and raising the value of direct touchpoints with developers. Perhaps in the not-too-distant future, end users won't even need developers.

## REFERENCES

[1] Low-level code tests are a fundamental level of software testing in which individual components or modules of an application are tested in isolation to ensure they function as intended.

[2] Thanisch, Eystein (ed.). "Disciplining AI, Part I: Evaluation Through Industry Lenses." *Amplify*, Vol. 38, No. 5, 2025.

[3] "Cost of Building and Deploying AI Models in Vertex AI." Google Cloud, accessed 2025.

[4] Seemann, Florian. "Defensibility in the Application Layer of Generative Artificial Intelligence." Medium, 12 April 2023.

[5] Shankar, Shreya, and Hamel Husain. "Application-Centric AI Evals for Engineers and Technical PMs." Course notes, May 2025.

[6] Husain, Hamel. "Creating a LLM-as-a-Judge That Drives Business Results." Blog post, 29 October 2024.

[7] Shankar, Shreya, et al. "Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences." arXiv preprint, 18 April 2024.

[8] Geach, P.T. "Platos 'Euthyphro': An Analysis and Commentary." *The Monist*, Vol. 50, No. 3, July 1966.

[9] Shankar and Husain (see 5).

[10] Hunter, Julie, Nicholas Asher, and Alex Lascarides. "A Formal Semantics for Situated Conversation." *Semantics and Pragmatics*, Vol. 11, No. 10, 2018.

[11] Pleyer, Michael, and Stefan Hartmann. *Cognitive Linguistics and Language Evolution*. Cambridge University Press, 2024.

[12] Joseph Farrington notes this in Part I of this *Amplify* series, stating: "Just as importantly, building the simulator prompted early engagement between stakeholders. Blood bank staff, clinicians, and data scientists worked together to define the decisions that mattered, the constraints that applied, and the metrics that should be used to determine success. This collaboration ensured that any model developed would be evaluated against practical criteria and shaped from the outset to fit the real-world context in which it would operate"; see: Farrington, Joseph. "A Simulation-First Approach to AI Development." *Amplify*, Vol. 38, No. 5, 2025.

[13] Vaswani, Ashish, et al. "Attention Is All You Need." arXiv preprint, 2 August 2023.

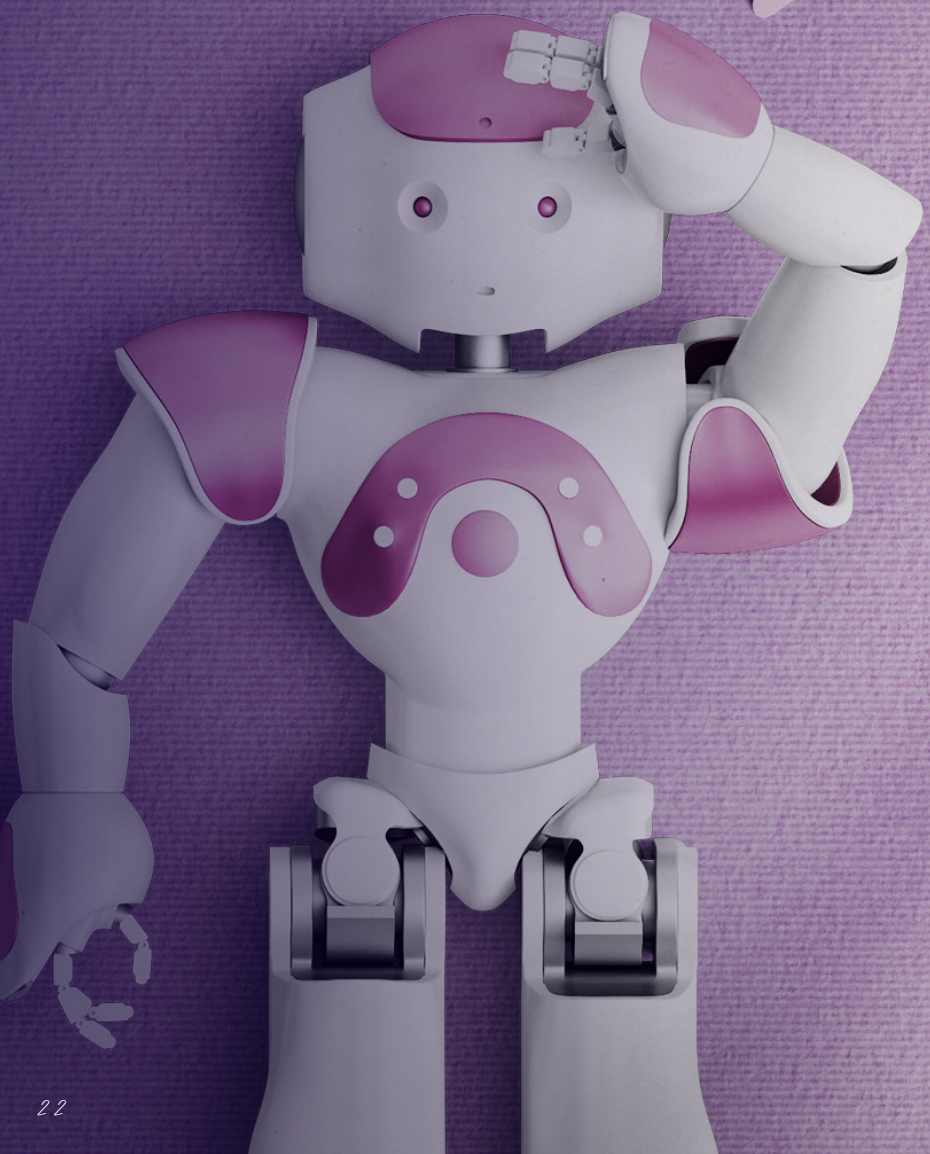
[14] Hitzler, Pascal, and Md Kamruzzaman Sarker. "Neuro-Symbolic Artificial Intelligence: The State of the Art." vaishakbelle.org, accessed 2025.

[15] Perrier, Elija. "Out of Control — Why Alignment Needs Formal Control Theory (and an Alignment Control Stack)." arXiv preprint, 21 June 2025.

## About the author

**Dan North** is Global IT Developer at Arthur D. Little, where he develops LLM-based applications for internal use. Prior to this role, he worked at an early-stage startup building AI agents for enterprise applications, and at AWS Bedrock, contributing to its LLM training pipeline. Mr. North has a deep passion for philosophy and strongly believes in its relevance and applicability to everyday life, including business. He is particularly interested in applications of AI that provide value for underprivileged and vulnerable communities. Mr. North earned a master of arts degree in linguistics from the University of Edinburgh, Scotland, where he specialized in formal semantics, pragmatics, and logic. He can be reached at North.Dan@adlittle.com.

# TACO BELL, 18,000 WATERS & WHY BENCHMARKS DON'T MATTER



Authors

Michael Papadopoulos, Olivier Pilot,
and Eystein Thanisch

**Starting in 2023, a number of fast-food chains deployed AI in an attempt to reduce mistakes and speed up service. Since then, many comical videos of customer interactions with AI at drive-throughs have gone viral. In one widely shared clip, a customer orders 18,000 water cups to "shut down" the system so he can speak to a human. In another, a frustrated driver grows increasingly angry as AI repeatedly asks him to add more drinks to his order. After millions of views, Taco Bell found itself rethinking its AI rollout in the face of ridicule.**[1]

It's hilarious — but also instructive. The AI didn't fail because it lacked linguistic skill. If anything, it was too good at turning speech into an order. The failure was that no one connected the AI's impressive fluency to the everyday constraints of the restaurant: no one orders 18,000 drinks at a drive-through.

This scenario captures the paradox of AI in 2025. Models like ChatGPT-5, Claude 4 Opus, and Gemini 2.5 are astonishingly capable (nearly indistinguishable on many benchmark tests). But in practice, they make mistakes humans would never make. The problem isn't performance. It's proximity and explainability — how well models align with the messy, human, real-world environments they're dropped into.

## THE PLATEAU NO ONE WANTS TO ADMIT

For years, AI progress felt unstoppable. GPT-3 blew people away with its fluency. GPT-4 showed surprising reasoning. Today, ChatGPT-5, Claude 4 Opus, and Gemini 2.5 systems are capable of sophisticated reasoning, multilingual fluency, and even multimodal interaction. But take a step back, and you'll notice the differences between them are slim (see Figure 1).

On MMLU (measuring Massive Multitask Language Understanding, a benchmark that tests AI models' ability to answer multiple choice questions across 57 academic/professional subjects), as of September 2025, the leading models hover around 90% in terms of accuracy.[2] On HumanEval, which tests coding, they differ by only a handful of points. Even open source challengers like DeepSeek function at similar levels.

The Stanford "AI Index Report 2025" notes that frontier models are converging on most static benchmarks and that incremental improvements are disproportionately costly while yielding marginal real-world value.[3]

**FOR YEARS, AI PROGRESS FELT UNSTOPPABLE**

Meanwhile, the leaderboards that once guided the industry are under fire. In early 2025, a tuned, nonpublic variant of Meta's Llama 4 Maverick was submitted to the popular Chatbot Arena, where models are ranked via human head-to-head voting. It scored far higher than the released model, sparking accusations of benchmark gaming and forcing Arena's operators to tighten policies.[4,5] Researchers have warned of the "leaderboard illusion" (also known as Goodhart's Law): once a measure becomes the target, it stops being a good measure.[6] Benchmarks were once a proxy for progress. Now they're often little more than marketing theater.

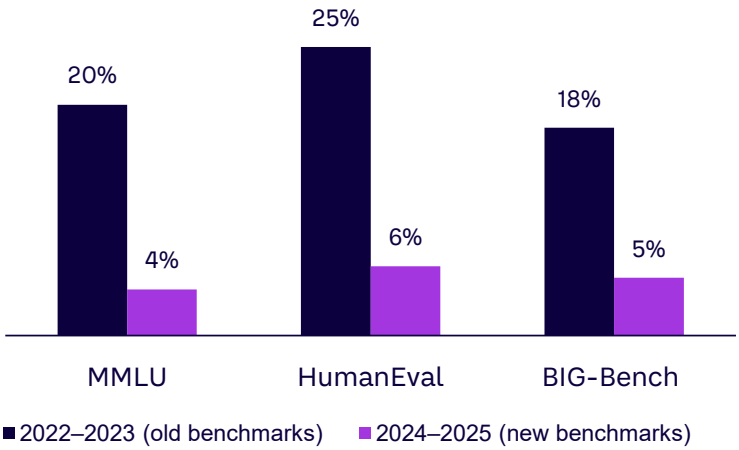



Figure 1. Score gaps between top models have narrowed, making traditional leaderboards less meaningful

## WHY THE TACO BELL STORY MATTERS

The 18,000 water cup fiasco makes for easy laughs. But it demonstrates something important: AI often fails not on capability, but on context.

Look at what went wrong:

- **Norms failure.** The system didn't realize that the ask was unlikely to be the result of a genuine need.
- **Data check failure.** AI didn't know that 18,000 was absurd given the store's inventory.
- **Workflow failure.** The order was accepted by the POS system without guardrails.
- **Governance failure.** No escalation was triggered before the blunder reached a paying customer.

And although funny in a fast-food setting, the same pattern can be devastating elsewhere. For example, in 2024, Air Canada's chatbot misinformed a customer about how to access a bereavement discount. A Canadian tribunal sided with the passenger, rejecting the argument that the chatbot was a "separate entity."[7]

It's certainly easy to imagine many such missteps in implementation. If an HR recruitment bot, for example, isn't integrated with the corporate calendar system, it could easily schedule overlapping interviews, leaving recruiters scrambling, candidates frustrated, and the company's brand at risk.[8]

These aren't failures of intelligence. They're failures of proximity — when AI isn't properly tethered to the data, processes, or rules of the organizations deploying it.

## CRACKS IN BENCHMARK THINKING

These examples illustrate why we can't rely on benchmarks to guide real-world adoption.[9] Traditional benchmarks are:

- **Static.** Sets like MMLU quickly saturate; new models simply memorize leaked questions.
- **Narrow.** They test abstract skills but not organizational integration.
- **Gameable.** Vendors can optimize leaderboard scores without real improvement.
- **Biased.** Many leaderboards now use large language models as judges, introducing circularity.

Efforts to fix this include MMLU-Pro (harder reasoning tasks), LiveBench (dynamic, contamination-resistant), SWE-bench-Live (real GitHub issues), and ARC-AGI-2 (AI hard-reasoning problems). These are valuable, but they still don't measure whether an AI can appropriately deal with someone ordering 18,000 waters at a drive-through.

## THE NEW BENCHMARK: PROXIMITY

To move beyond the plateau, organizations need a new lens: proximity. If the models themselves (GPT-5, Claude 4 Opus, Gemini 2.5) are functionally equivalent, their performance should not be evaluated outside the context of what they are tasked to achieve. The real competitive edge lies in how closely you can align them with your business.

Four layers matter most:

1. **Norms proximity.** Can the AI enforce guardrails to assess the appropriateness of requests it receives, as well as the responses it produces, against a backdrop of societal and business norms that a human would be expected to abide by, naturally or after training?
2. **Data proximity.** Can the AI access and use your proprietary data and knowledge? How quickly do new updates show up? How well is data protected?
3. **Workflow proximity.** Can the AI be embedded to reliably execute tasks that are part of wider workflows or collaborations in a truly beneficial way? Does it save time and reduce human error?
4. **Governance proximity.** Are you able to know when the AI hallucinates or breaks policy, and are mechanisms in place to escalate safely? Can you predict and manage cost?

## THE MISSING PIECE

There's one more piece: accountability through weighted explainability and oversight.

The phenomenon of algorithmic aversion describes how people often judge machine errors more harshly than human mistakes. The higher the stakes and the more autonomous the system, the sharper this reaction becomes. It's the driverless car problem all over again: even if the technology is statistically safer overall, a single failure can shatter public trust. At the core of this reaction lies a perceived accountability gap, fueled by the lack of transparency in algorithmic decision-making.

When a human errs, we can ask why, assign responsibility, and expect accountability. That's why most people accept the risk of a barista mishearing their coffee order but balk when an algorithm makes a comparable mistake (especially in contexts that feel more consequential).

Both weighted explainability (explainability that is tied to the gravity of the potential consequences of a decision) and weighted oversight (which decides when a human must be put back in the loop based on such potential consequences) are fundamental for trust and adoption.

Without them:

- **Employees won't trust** the AI's decisions.[10]
- **Regulators won't approve** opaque systems.
- **Customers won't forgive** errors they can't understand or feel empowered to challenge them.

That's why weighted explainability and oversight are the enablers of the fifth layer of proximity: accountability proximity.

## HOW DOES ACCOUNTABILITY PROXIMITY HELP?

To answer this, let's go back to some of our previous examples:

- In **Taco Bell's** case, norms themselves could have sufficed to correctly qualify the nature of the request. And if this had not been enough, correct integration with inventory data, or into a workflow with input data validation, would have caught it. In both cases, the system should have been able not only to reject the request but also explain why, giving the customer the ability to reframe it.
- In **Air Canada's** case, it is the proximity to the data that was the problem. What made things worse was the inability to explain why a key part of the company's discount terms and conditions had been ignored or misunderstood.

Each case highlights the same truth: these aren't failures of intelligence, but failures of proximity and accountability.

## AI IN THE POST-BENCHMARK ERA

The era of obsessing over raw model performance is over. The marginal differences between GPT-5, Claude 4 Opus, Gemini 2.5, and the best open source challengers matter less than whether or not a system delivers measurable, reliable outcomes.

CXOs now ask: Did AI reduce our ticket backlog? Did it improve customer experience without adding risk? Can it explain itself to regulators?

This is the rise of evaluation operations: continuous, production-aligned testing that measures AI against business KPIs rather than trivia sets. Offline testing leads to shadow deployments, then live A/B tests, then full production with ongoing monitoring.

The Stanford AI Index 2025 underscores this shift: enterprise adoption increasingly depends on integration efficiency, explainability, and outcome alignment — not leaderboard scores.

## THE PROXIMITY SCORECARD

Table 1 shows a practical framework for evaluating AI readiness. Each category is worth up to 20 points, for a total of 100.

This shifts the focus from "Can the model ace a benchmark?" to "Can it deliver measurable, reliable, and sustainable value in your actual workflows?"

## CONCLUSION: FROM BENCHMARKS TO BUSINESS OUTCOMES

The Taco Bell glitch is prophetic. It shows what happens when an AI's linguistic horsepower isn't tethered to real-world context and accountability. Benchmarks are plateauing, leaderboards are being gamed, and explainability has emerged as a make-or-break factor in adoption.

The organizations that win going forward won't be the ones chasing a percentage point on MMLU. They'll focus on proximity and explainability — embedding AI into their norms, data, workflows, and governance while ensuring people can trust its decisions.

Those that succeed will quietly unlock massive productivity gains; those that don't may find themselves the subject of the next viral story about an AI glitch. And no one wants to be remembered for that.

| CATEGORY | EXAMPLE METRICS |
|---|---|
| **Norms proximity (0-20)** | Invalid-input-rejection rate, false- acceptance rate, recovery-success rate, red-flag trigger rate |
| **Data proximity (0–20)** | Retrieval accuracy, knowledge coverage, update-to-response latency, privacy safeguards |
| **Workflow proximity (0–20)** | Task-success rate, time-to-task reduction, human effort saved, tool/API reliability |
| **Governance proximity (0–20)** | Hallucination rate, prompt-injection resistance, cost per successful task; robustness under load |
| **Accountability proximity (0–20)** | Explainability of outputs, auditability, user trust; ability to identify the need for human oversight; regulator acceptance |

Table 1. Proximity scorecard

## REFERENCES


1 Kaplan, Zach. "Taco Bell Re-Evaluating Plans for AI Drive-Through Experience." NewsNation, 28 August 2025.

2 "MMLU." Kaggle, accessed 2025.

3 "Artificial Intelligence Index Report 2025." Stanford Institute for Human-Centered AI (HAI), 2025.

4 Mann, Tobias. "Meta Accused of Llama 4 Bait-and-Switch to Juice AI Benchmark Rank." *The Register*, 8 April 2025.

5 "LMArena Tightens Rules After Llama-4 Incident." Digwatch, 9 April 2025.

6 Singh, Shivalika, et al. "The Leaderboard Illusion." arXiv preprint, 12 May 2025.

7 Proctor, Jason. "Air Canada Found Liable for Chatbot's Bad Advice on Plane Tickets." CBC, 15 February 2024.

8 For an example of an AI sales agent struggling with scheduling, see Joe Allen's article in this issue of *Amplify*: "Why Judgment, Not Accuracy, Will Decide the Future of Agentic AI."

9 Eriksson, Maria, et al. "Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation." arXiv preprint, 25 May 2025.

10 Byrum, Joseph. "AI's Impact on Expertise." *Amplify*, Vol. 38, No. 5, 2025.


# About the authors


**Michael Papadopoulos** is a Partner at Arthur D. Little (ADL) Catalyst. He is passionate about designing the right solutions using smart-stitching approaches, even when elegance and architectural purity are overshadowed by practicality. Mr. Papadopoulos leads the scaling of multidisciplinary organizations by focusing on continuous improvement, establishing quality standards, and following solid software engineering practices. He mentors team members, leaders, and managers along the way. Mr. Papadopoulos is a strong advocate of the DevOps culture and Agile principles and has demonstrated experience in solving problems in challenging global environments. Coming from a development background, he remains highly technical, with hands-on involvement in code review, design, architecture, and operations. Mr. Papadopoulos has 15 years' experience in technology and digital consulting and has worked in a variety of sectors, including telecom, gaming, energy, and media. He can be reached at experts@cutter.com.

**Olivier Pilot** is a Principal at ADL and Head of Product for ADL Catalyst. He focuses on the identification, prioritization, design, and delivery of digitally enabled solutions for complex challenges, with experience in enterprise and solution architecture, product strategy and management, and artificial intelligence. Mr. Pilot has more than 20 years of consulting experience across multiple sectors, helping clients transform operations, customer engagement, employee experience, and business models through technology. His recent work includes shaping AI-driven strategies and roadmaps, applying AI in product management and delivery, and leading large-scale solution implementations that enable innovation and new growth opportunities. Mr. Pilot earned a master of engineering degree in IT from École Centrale de Lyon, France. He can be reached at experts@cutter.com.

**Eystein Thanisch** is a Senior Technologist at ADL Catalyst. He enjoys ambitious projects that involve connecting heterogeneous datasets to yield insights into complex, real-world problems and believes in uniting depth of knowledge with technical excellence to build things of real value. Dr. Thanisch is also interested in techniques from natural language processing and beyond for extracting structured data from texts. Prior to joining ADL, he worked on Faclair na Gàidhlig, the historical dictionary of Scottish Gaelic, on a team tasked with building a tagged corpus of transcriptions from pre-modern manuscripts. He also was involved in IrishGen, a project on the use of knowledge graphs to represent medieval genealogical texts. Dr. Thanisch also worked as a freelance editor and analyst for a number of IGOs and academics. He earned a master of science degree in computer science from Birkbeck, University of London, and a PhD in Celtic studies from the University of Edinburgh, Scotland. He can be reached at experts@cutter.com.

# WHY JUDGMENT, NOT ACCURACY, WILL DECIDE THE FUTURE OF AGENTIC AI

*Author*

Joe Allen

**The research headlines are intoxicating. Stanford's "AI Index 2025" reports a 67-point leap on a coding challenge, which measures whether AI systems can fix real-world problems.[1] On LinkedIn, the narrative is tidy: scores soar, adoption soars; therefore, AI must be working. But spend 10 minutes in a board meeting, and the temperature drops. The CFO's first question is my own: "Can you show me proof it works in the wild?" Benchmarks inspire optimism; behavior earns trust.**

Why the disconnect? Leaderboards were engineered for single-shot competence (translate a paragraph, label an image, patch a snippet of code). Agentic systems (AIs that plan, click, and iterate) break that paradigm in two uncomfortable ways:

1. **Tool chaos.** Every deployment inherits its own APIs, data silos, log-in journeys, and user interface quirks. My agent behaves differently from yours, even if the underlying foundation model is identical.
2. **Temporal depth.** Success now hinges on dozens of micro-decisions: whom to email, which link to click, when to pause, whether to back off entirely.

Drop bleeding-edge systems into realistic flows, and cracks appear. State-of-the-art agents complete about 41% of tasks on the new REAL benchmark (a replica of everyday websites).[2] Domain-specific suites such as τ-Bench swing wildly: 69% success in retail but 46% in airline bookings. Static competence does not equal dynamic reliability. That gap defines the next five years of enterprise AI.

> **THOUSANDS OF LIVE DECISIONS COMPOUND INTO RISK**

## UNDERSTANDING BENCHMARK BLINDNESS

Traditional benchmarks reduce performance to a single score. In reality, thousands of live decisions compound into risk. Even as newer benchmarks log decision histories, they miss the brittle behaviors that only emerge when agents act in production.

Autonomous agents control calendars, inboxes, purchase-order APIs, and sometimes factory endpoints. Each additional actuator magnifies the consequence of every hidden edge case. Thus, benchmark blindness plays out at two levels:

1. **Temporal blind spot.** Benchmarks sample a moment; reality is a stream. A self-driving car that aces lane-keeping may miss an unpainted junction.
2. **Authority blind spot.** Benchmarks ignore reach. A language model that tops MMLU cannot wake a prospect at 6 am, but an outbound agent running on that same model can because it was wired to act.

## THE STATISTICAL REBELLION HAS BEGUN

Research labs and operations teams are discovering that single-shot metrics are not enough. Instead, we must watch agents in the wild, continuously, like epidemiologists tracing an outbreak.

In the lab, this translates into methodologists calling for live monitoring that captures rare edge-case failures and second-order knock-ons (rather than a one-off test run). In industry, this translates into European telecom pilots showing that continuous logging cuts outage time in half, resulting in boards expanding automation, not throttling it.[3,4]

The lesson for senior decision makers is this: transparency around error dynamics is nonnegotiable. Numbers still matter; they just need to be rooted in streaming reality, not sandbox perfection.

## BENCHMARKS WON'T SAVE US

A truly autonomous agent is no longer a scripted macro. It writes its own flight plan and tries to execute it in the open world. If it cannot show sound judgment (stay on-brief, handle tools cleanly, recover when reality shifts), it deserves no more latitude than a mail merge. "Trust the benchmark" is a comfort blanket; behavioral telemetry is the grown-up stance. Managers can track agent performance with four simple checks:

1. **Plan fidelity.** Did the agent follow its own plan or drift off course?
2. **Tool dexterity.** Did the agent use the software cleanly, without mis-clicks or redundant actions?
3. **Recovery ability.** How fast did the agent bounce back from errors like dead links or timeouts?
4. **Integrity.** Are logs and data trails strong enough to meet compliance standards?

Note the difference from classic benchmarking: the exam is no longer human-written. The agent drafts its own plan and is judged on how faithfully (and safely) it follows that self-authored blueprint. This step is what turns an LLM into an autonomous colleague and why static accuracy tests miss emerging risk.

This is the point where behavioral evaluation becomes essential. Behavioral signals reveal trouble days before revenue drifts and months before regulators arrive. They move the discussion from "How high did it score?" to "How safely does it fly when no one is watching?"

## CASE STUDY

Nova is our in-house, autonomous outbound sales agent. It's a small cluster of specialist LLMs wrapped in a planner, tool API layer, and reward engine. When I say "agent," I'm speaking about a live production stack, not a slide-deck prototype.

### THE "WRONG TITLE" DEMO

At 9:17, an alert popped up: "Agent has scheduled your demo." The guest? An intern in facilities "keen on AI." Any seasoned seller would have thanked her, disqualified politely, and moved on. But the agent sent a calendar invite to any positive reply. Task complete; objective failed.

### WHY THIS "TINY MISS" MATTERED

The facilities intern booking is a serious fault for two reasons:

1. **Economic reality.** Calendar time is billable. An hour spent demonstrating to someone with no budget displaces appointments worth five (sometimes six) figures. The agent's misfire cost far more than the penny-level compute it consumed.
2. **Cultural drift.** We are still hardening the agent. If the team shrugs off early slips, complacency will calcify just as the model moves into production.

Today's harmless mismatch is tomorrow's flash-crash once the system controls thousands of outbound threads. It's essential to treat every error like a defect, not a curiosity.

### WHAT WE CHANGED

We hardened the agent with four fixes:

1. **Quantify the miss.** We built in-house metrics to capture plan fidelity, tool dexterity, and recovery time. After hardening, the agent restated goals before acting, counted redundant clicks as errors, and added a "search then retry" routine for dead links. Those changes lifted scores from 72% to 92% on plan fidelity, 58% to 71% on tool dexterity, and cut recovery latency from 22 seconds to 8 seconds.
2. **Measure plan versus actions.** Agents must restate the goal, and telemetry shows when declared steps diverge from reality.
3. **Align consequences.** Errors now trigger the same escalation paths as humans, so small slips get flagged before they snowball.
4. **Keep logs for accountability.** Every action, prompt, and recovery is traced — meeting compliance standards and making blame visible.

Together, these shifts embed evidence into everyday culture: decisions rest on data, not gut feel.

## WHAT THE CASE TAUGHT US

Several important lessons emerged from this case:

1. **Benchmark early, behave often.** Leaderboards pick a base model; only behavior refines an agent.
2. **Instrument before you innovate.** If clicks aren't logged, failures stay silent.
3. **Punish the stumble, not the crash.** Micro-penalties on sloppy actions drive outsized gains.
4. **Chaos engineer the edge cases.** Synthetic obstacles today prevent embarrassments tomorrow.
5. **Fuse compliance with DevOps.** Evidence generated automatically is evidence you can hand an auditor before they ask.

### THE DATA-TRAINING-GUARDRAIL TRIANGLE

Beyond immediate fixes, the case showed three broader levers for keeping agents on track:

1. **Richer domain data.** Teach the model why a facilities intern can't authorize spend (e.g., ingest customer relationship management data on close-won versus close-lost by title).
2. **Tighter feedback loops.** Label every mis-booked meeting "nonqualified" and fine-tune nightly.
3. **Explicit guardrails.** Block invites below a title threshold but allow human override for exceptional cases.

These loops may seem mechanical, but they raise a deeper question: at what point does responsibility shift from the human trainer to the agent itself? The more decisions delegated, the more incremental drift becomes a slide from citizen control to code.

## BUILD YOUR OWN MEASUREMENT LOOP

To harden the agent, we borrowed from DevOps: log the plan up front, replay its clicks, diagnose tool errors, and ship evidence with every build. The idea is simple: problems must be visible, reproducible, and accountable. Once telemetry makes errors impossible to hide, fixing them becomes routine.

### ADOPT SHARED TESTS OF JUDGMENT

Technical unit tests prove the JSON is valid; they say nothing about whether the action makes business or ethical sense. We therefore add judgment tests: domain-specific assertions that check the quality of the agent's decision, not just its syntax (see Table 1).

Traditional ops checked "Did the code run?" Modern ops must also ask, "Would a reasonable human make this choice?" Building judgment into tests moves ethics and business logic upstream so issues can be caught in seconds, not in postmortems.

| EXAMPLE CHECK | WHY IT MATTERS | HOW TEST RUNS INSIDE CI/CD |
|---|---|---|
| **Job-title gate:** Was the meeting booked with someone who can sign? | Saves human hours; protects pipeline value | After the agent schedules, the test calls LinkedIn API (or cached org chart) and fails the build if title is not on the allowed list. |
| **Opt-out compliance:** Did outbound copy respect suppression lists? | GDPR/CAN-SPAM fines are expensive | CI job feeds recent opt-out CSVs into the agent's contact queue; if any appear in an email, the pipeline halts. |
| **Regional privacy flag:** Did workflow honor "no-tracking" regions? | Required under EU/UK data regimes | Integration test spins up a mock user in EEA, then inspects headers for prohibited tracking pixels. |

Table 1. Judgment tests

## REWARD SELF-EXPLANATION

"I prioritized this lead because her title includes VP procurement" lets debuggers see what the agent believed and gives auditors proof of nondiscriminatory logic. We typically capture the full chain of thought inside a secure store and retain only a hashed, one-sentence summary in production logs. That balance keeps debugging fast, satisfies compliance requests, and protects both customer data and intellectual property. (Industry norms vary, but regulators are already signaling that some level of human-readable justification will soon be mandatory.)

## TOOLING STARTER KIT

The tools mentioned in Table 2 are not endorsements; they are what we used to reach the results reported. All are either open source or have free tiers, and each solves a specific live-ops gap that would otherwise require weeks of custom code.

## STARTER KPI DASHBOARD

When you set performance targets for agents, there's always the risk they'll "game the numbers." This is Goodhart's Law in action: once a measure becomes the goal, the system learns to optimize the score rather than the behavior it was meant to capture.

The solution is to pair each KPI with an adversarial check, a test that makes cheating harder than simply performing well. Table 3 shows how four agent KPIs can be matched with such checks, ensuring the metrics remain meaningful. Once these are public inside the company, teams start competing to drive them up.

## STANDARDS, AUDITS & THE MORAL PERIMETER

Europe's AI Act names general-purpose models and demands detailed transparency reports for

| GAP TO CLOSE | TOOL WE PICKED | WHY WE CHOSE IT | VIABLE ALTERNATIVES |
|---|---|---|---|
| **Event capture & easy dashboards** | **OpenTelemetry** (open source) | Language-agnostic tracing spec; drops straight into Grafana or Honeycomb with one config file. | Jaeger, Zipkin |
| **Data-schema drift before fine-tuning** | **Great Expectations** (open source) | Declarative tests ("column X must be email") stop garbage before it hits retraining queue. | Soda-CL, Deequ |
| **Deterministic reruns of browser sessions** | **pytest-replay** (MIT license) | Re-executes recorded DOM events so QA can reproduce bug in seconds. | Playwright's trace mode, Cypress |
| **Prompt & trace diffing** | **LangSmith/ PromptLayer** (mixed open/paid) | Both visualize prompt chains, show token-level deltas, and let you tag "good" vs. "bad" traces for regression tests. | Traceloop (open), LlamaIndex observability module |

Table 2. Tooling starter kit

| KPI | TARGET BAND | LEADING ADVERSARIAL CHECK | RATIONALE |
|---|---|---|---|
| **Plan fidelity** | 95%-98% (not 100%) | Randomly inject hidden "must-skip" steps; flag if agent executes them | Ensures fidelity isn't faked by parroting oversized task lists |
| **Tool dexterity** | ≥80% clean clicks | Replay 1% of traces through Playwright; diff for hidden retries | Catches agents that "click-spam" until a selector works |
| **Recovery latency** | <10 s median | Chaos job revokes tokens; alerts if agent simply aborts task | Stops gaming via silent dropouts that keep latency at 0 |
| **Integrity flags** | 0 critical/ 30 d | Quarterly red team audit of full logs | Prevents quiet deletion of problematic traces to keep count clean |

Table 3. Pairing KPIs with adversarial checks

systems with systemic risk. A voluntary Code of Practice already sketches the format: logs, energy use, red team results. Treat these documents as free product requirements: expose plan fidelity, dexterity, recovery, and intent logs by design.

Financial statements rely on auditors; flight recorders keep aircraft honest. Agentic AI needs a neutral clearinghouse. Imagine an Agent Audit Office that:

- Runs shadow tasks on production agents
- Verifies carbon, privacy, and safety disclosures
- Publishes dashboards across plan, dexterity, recovery, and integrity

Table 4 is a "should exist/already exists" cheat sheet — so you can tell which pieces are live today and which ones this article proposes.

Actuarial models love repeatable signals, and telemetry turns black-box AI into something insurers can price. Within three years, cyber policies will likely demand "continuous behavior logging" alongside firewalls and multifactor authentication. Failure to comply leads to a rise in premiums.

| BUILDING BLOCK | STATUS | WHAT IT IS/WHY IT MATTERS |
|---|---|---|
| **OpenTelemetry Standard** | Proposed extension | Builds on existing OpenTelemetry spec but adds 4 agent-specific fields: plan, action, result, context. Gives every vendor same JSON trace so auditors and insurers can compare like-for-like. |
| **Federated red team exchange** | Emerging (pilot consortia in finance & health) | Think "threat-intel feed but for AI failures." When 1 bank sees a prompt that bypasses guardrails, every subscriber gets the signature the same day. |
| **Performance-bonded SLAs** | Already used in cloud carbon offsets; proposed for AI | Vendor refunds credits if plan fidelity or recovery dips below contract thresholds; this aligns incentives without waiting for regulation. |
| **Independent arbitration panel** | Concept (mirrors TAG) | A cross-disciplinary body (engineers, ethicists, insurers) reviews disputed failures. Rulings feed back into OTS schema, so lessons become standards, not folklore. |

Table 4. Proposed building blocks for a trusted agentic AI ecosystem

Handing routine decisions to software is seductive; nobody romanticizes manual data entry. But history shows us that unseen systems gain power by invitation, not conquest. If we cannot measure an agent's judgment today, we may struggle to revoke its privileges tomorrow.

## APPLE EXPOSES 3 BLIND SPOTS

Benchmark top scores hide the brittle behaviors that sink autonomous systems in production. Apple's June 2025 paper, "The Illusion of Thinking," hands us laboratory proof.[5] By engineering a family of logic puzzles with difficulty levels that can be nudged upward in single, measurable increments, Apple exposed three blind spots that puncture benchmark comfort:

1. **Complexity cliffs.** Accuracy cruises near 100% then collapses to about 0% with a one-step jump in difficulty. A four-step approval chain may work flawlessly; add a fifth, and the agent fails. Leaderboards rarely push models this far, so the cliff remains hidden.
2. **Effort paradox.** Token-level "thinking" rises with difficulty, then peaks and collapses alongside accuracy. Long reasoning traces don't guarantee success; they can mask fruitless search. Benchmarks miss this because they only score the final answer.
3. **Three performance zones.** Vanilla LLMs win easy tasks, and reasoning-tuned models win mid-level ones, but no model survives past the cliff. A single headline score can't tell you which zone your workflow sits in.

## HOW APPLE'S STUDY VALIDATES OUR THESIS

First, Apple's experiment shows that single-score tests hide collapse points (accuracy looks fine until it suddenly fails). Second, its traces match exactly the signals we recommend (plan fidelity, recovery). Third, it shows that stress curves beat single scores: difficulty ramps reveal where systems snap, not just how they score.

The fact is, execution matters more than description. It's not enough for a model to outline an algorithm — it must actually run the steps. Apple exposes that execution gap, the same one that decides whether agents can operate safely in live processes. Our telemetry closes that gap by tracking what the agent does, not what it says it could do.

Apple confirms that you cannot grade an autonomous system by the cleverness of the algorithm it could write; you must grade it by the safety and fidelity of the actions it actually performs. That requires continuous behavioral logging, not prettier exam scores.

## OPERATIONALIZING APPLE'S INSIGHTS

Four steps can help operationalize Apple's insights:

### STEP 1: BUILD YOUR OWN PUZZLE LADDER

- **Sales use case?** Add one extra approver, one extra currency, and one extra compliance clause each test cycle.
- **Support chatbot?** Increase simultaneous tickets or escalate sentiment hostility.
- **Chart accuracy, recovery time, and explanatory length at every rung;** the first discontinuity is your cliff.

### STEP 2: TRACK EFFORT VERSUS OUTCOME

- **Log reasoning token count** (or tool-call depth) alongside success. A sudden dip in effort at higher complexity is your "effort paradox" alarm.

### STEP 3: PENALIZE VACUOUS VERBOSITY

- **Longer traces sometimes mask confusion,** according to Apple.
- **Score explanations for coherence** (no repeated phrases, no self-contradictions) rather than length.
- **Dock points when verbosity rises** without a matching lift in correctness.

### STEP 4: PROMOTE COLLABORATION TRIGGERS

When an agent drifts off course, you need clear triggers for handoff. For example:

- **Recovery spikes** (tool calls that stall too long).
- **Effort dips** (the agent is "thinking less" while tasks get harder).
- **Plan drift** (skipping too many declared steps).

When any of these occur, the task should pass to a backup — a fact-checker (knowledge graph), a compliance gate, or a human reviewer.

## WHY THIS CHANGES THE ACCOUNTABILITY CONVERSATION

Regulators and insurers want proof that an autonomous system knows when it is out of its depth. Apple's methodology offers exactly that.

Cliff plots could become mandatory disclosures, showing where accuracy collapses and how vendors detect failure. Effort-paradox curves reveal whether a model's "reasoning" mode is genuine or just rambling until it guesses. Behavioral early-warning metrics meet the EU AI Act's demand for continuous monitoring far better than a once-a-year benchmark certificate.

We recommend starting small. Measure the length of reasoning tokens in your agent today; add one extra twist to a live workflow tomorrow. At the end of the week, graph effort versus accuracy. If you see the cliff, you've found the evidence needed for guardrails and safe handoffs. If the curve stays flat, keep ramping (the cliff is still there, just not reached).

## THE ONE-WEEK TEST

The wrong title demo was a nuisance, not a scandal, but it shattered the illusion that high benchmarks equal safe autonomy. One missed meeting costs an hour; scale that indifference and civilization faces two darker forks:

1. **AI winter.** Minor errors accumulate (flaky bookings, garbled invoices) until trust evaporates and budgets freeze.
2. **Loss of agency.** Quiet success slides authority from citizen to code. Humans stop checking, and by the time anyone notices, reversal is too costly.

Both risks share unaudited, untraceable behavior. Either we choke adoption out of fear, or we embrace it until it's too late to fix.

The one-week test shows how small tests expose the hidden cliffs. Behavioral evidence, not benchmarks, will decide whether we face another AI winter or a future where code governs quietly in the background.

## REFERENCES


[1] Maslej, Nestor, et al. "The AI Index 2025 Annual Report." Stanford University, Human-Centered AI (HAI) Institute, April 2025.

[2] Garg, Divyansh, et al. "REAL Benchmark Benchmarking Autonomous Agents on Deterministic Simulations of Real Websites." arXiv preprint, 17 April 2025.

[3] Said, Sherif, and Mario Guido. "How Vodafone Is Using Gen AI to Enhance Network Life Cycle." Google Cloud blog, 22 November 2024.

[4] SIGNL4 is an incident-alert platform used by several EU carriers; see: "Modern Incident Management." Derdack SIGNL4, accessed 2025.

[5] Shojaee, Parhin, et al. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity." Apple Machine Learning Research, June 2025.

| TERM | DEFINITION |
|---|---|
| **Agent** | An AI that acts (reading, planning, clicking) rather than merely answering |
| **API** | Application programming interface |
| **Benchmark** | Offline test set used to compare model competence |
| **CAN-SPAM** | Controlling the Assault of Non-Solicited Pornography and Marketing Act (US) |
| **CI/CD** | Continuous integration/continuous delivery |
| **CSS selector** | Snippet of webpage code that tells browser “press this button” |
| **DOM** | Document Object Model |
| **EEA** | Enterprise Edition Agreement |
| **GDPR** | General Data Protection Regulation (EU) |
| **JSON** | JavaScript Object Notation |
| **KPI** | Key performance indicator |
| **LLM** | Large language model |
| **MMLU** | Massive multitask language understanding |
| **Model card** | One-page document summarizing model’s data sources, limits, and risks; fast becoming mandatory for compliance |
| **OTS** | OpenTelemetry Standard |
| **QA** | Quality assurance |
| **REAL** | Benchmark built from cloned production websites to test web-navigation agents (frontier agents ≈ 41% success) |
| **SLA** | Service-level agreement |
| **TAG** | Trustworthy Accountability Group |

Table 5. Glossary

# About the author

**Joe Allen** is founder and CEO of Launched, the company behind Nova, the world's first autonomous AI salesperson. That vision grew out of his early career in sales, where he experienced firsthand the inefficiency of relying on sales rep teams for prospecting, follow-ups, and cold calls — work that often led to burnout and inconsistent results. This experience inspired his mission to replace repetitive human effort with a system capable of truly selling. In 2021, Mr. Allen founded Launched and led the development of Nova, an AI platform designed to manage the entire sales cycle — from prospecting and qualification to demonstrations, negotiations, and closing. Drawing on both technical expertise and commercial acumen, he has positioned Nova to redefine how sales organizations operate. Mr. Allen envisions a future where revenue growth is driven not by large human teams managing tools, but by autonomous AI systems at the frontline of sales. Known for his persistence and commercially focused approach, Mr. Allen combines a willingness to challenge established models with a commitment to building more efficient and adaptive paths to growth. His work reflects a broader ambition: to set a new standard for business growth that is smarter, more sustainable, and less constrained by traditional limitations. Mr. Allen can be reached at jallen@launchedtech.io.

# AI ASSET SURVIVAL IN THE AGE OF EXPONENTIAL TECH

## Authors

Chirag Kundalia and V. Kavida

**Contemporary discourse around AI is often infused with narratives that celebrate its boundless promise, its disruptive capacity, and its role in shaping competitive advantage. From financial services and healthcare to transportation and supply chains, AI has already demonstrated its capacity to reconfigure industries. However, this enthusiasm obscures a sobering truth that rarely receives adequate attention: AI systems are inherently dynamic; they are not static tools to be deployed once and trusted indefinitely.**

Consider the parallels. When an airline invests in a new fleet, it models the lifespan of each aircraft, plans for maintenance cycles, and budgets for eventual replacement. When a hospital installs advanced diagnostic equipment, it tracks performance drift, schedules recalibrations, and anticipates obsolescence. When companies deploy AI models (e.g., fraud-detection algorithms, recommendation engines, customer service bots), these lifecycle dynamics are often ignored. This is particularly true for AI systems deployed and managed internally by the organization itself rather than as cloud-managed services.

Many AI capabilities are now accessed through managed cloud platforms (with providers handling model maintenance and updates), but this article focuses on enterprise-deployed, narrow AI systems that are typically managed in-house. These systems often support domain-specific use cases (e.g., fraud detection, diagnostics) and require the deploying organization to actively monitor performance, retrain models, and plan for obsolescence.

Recent research shows that AI models exhibit temporal degradation: performance decays over time as data patterns shift, environments change, or model assumptions erode.[1,2] A high-performing model in January, for example, may be a silent liability by December. In some industries like finance or healthcare, this can cause lost business value, regulatory noncompliance, and reputational harm.

AI investments are growing larger and more embedded, but AI governance frameworks remain immature, particularly when it comes to tracking long-term asset health.[3,4]

Moreover, external forces accelerate AI aging. Regulatory landscapes shift (e.g., new EU AI Act requirements), and competitors launch disruptive models. Foundational shifts in AI paradigms, such as the rapid emergence of large language models, can cause once-leading systems to become outdated almost overnight. In this volatile context, executives must treat AI systems as living, dynamic assets requiring ongoing stewardship.

> **AI INVESTMENTS ARE GROWING LARGER & MORE EMBEDDED, BUT AI GOVERNANCE FRAMEWORKS REMAIN IMMATURE**

Failure to do so carries tangible business risk:

- **Financial risk** — sunk costs in underperforming models
- **Operational risk** — degraded customer experience, errors in critical processes
- **Compliance risk** — outdated models violating new regulations
- **Strategic risk** — losing market edge due to slow refresh cycles

## USING SURVIVAL ANALYSIS TO MANAGE THE AI LIFECYCLE

The longevity of AI systems has emerged as an under-explored yet strategically vital dimension of AI governance. Business investments in AI continue to accelerate, but mechanisms to anticipate and manage the degradation of these assets remain fragmented and inconsistent. This article proposes a conceptual framework that uses the well-established methodology of survival analysis to improve AI asset management.

Originally developed for biomedical and reliability engineering, survival analysis estimates the probability that a given entity will continue to function over time. In healthcare, it predicts patient survival; in engineering, it models the time until component failure.

Recent research proved its usefulness in domains involving time-to-event outcomes, including complex systems such as lithium-ion battery degradation and predictive maintenance in industrial fleets.[5] We believe survival analysis can be successfully applied to AI systems — conceptualizing them as dynamic, degradable assets operating within evolving environments.

We model each AI asset (e.g., a deployed machine learning model) as a function S(t), representing the probability that the model continues to deliver acceptable performance at time t. The corresponding curve captures the evolving risk of obsolescence or failure as the system ages (see Figure 1).

Importantly, performance degradation in AI is typically multifactorial, driven by a combination of internal and external variables. Our framework incorporates both categories.

### 1. ENDOGENOUS FACTORS (INTERNAL)

- **Architectural adaptability** — the flexibility of a model's architecture to accommodate new data or retraining cycles without catastrophic forgetting
- **Algorithmic robustness** — the resilience of model performance under data drift (changes in the input data distribution over time), concept drift (changes in the relationship between inputs and outputs), and shifting feature distributions (e.g., variable ranges evolving due to environmental or user behavior changes)[6,7]
- **Data ecosystem vitality** — the freshness, quality, and stability of input data streams, which directly affect model relevance and accuracy[8]

### 2. EXOGENOUS FACTORS (EXTERNAL)

- **Regulatory flux** — changes in legal or compliance requirements that render models noncompliant or require recertification[9]

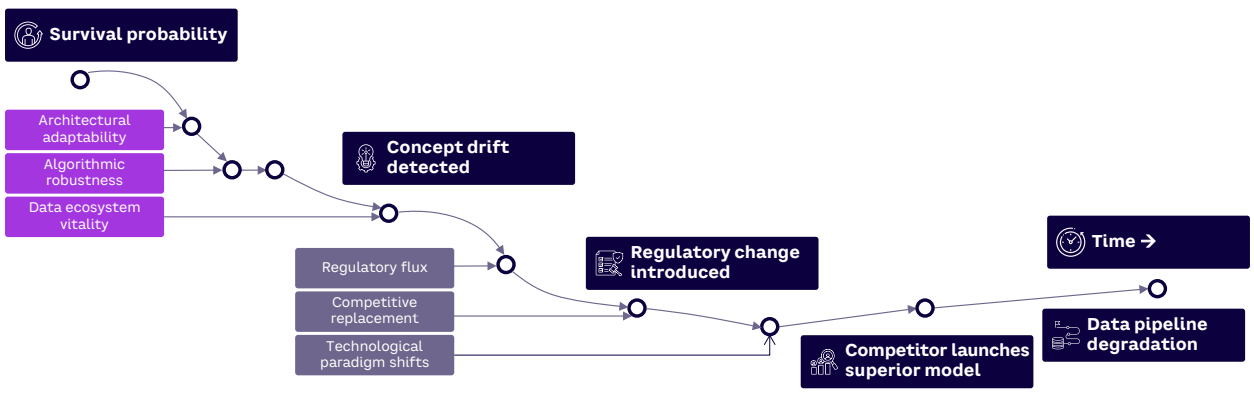



Figure 1. Conceptual AI survival model

- **Competitive displacement** — the emergence of superior models or market standards that diminish the relative value of an existing system
- **Technological paradigm shifts** — foundational shifts in the AI field, such as the transition from narrow models to foundation ones, which can accelerate obsolescence

It is important to distinguish between two forms of degradation. Intrinsic degradation occurs when a model's performance declines due to internal factors, such as data drift or architectural limitations. Relative degradation arises when a model becomes obsolete not because it performs worse than before but because superior alternatives emerge (raising the standard for what is considered acceptable or competitive). Both forms are critical to understanding AI asset survival and are integrated into our framework.

Survival analysis provides several advantages over conventional monitoring methods. First, it enables forecasting of AI asset longevity rather than reactive detection of failure. Second, it supports early-warning detection (identifying indicators that an AI system is entering a high-risk phase of degradation). Third, it offers a foundation for adaptive governance frameworks in which survival-informed insights guide maintenance, retraining, and retirement decisions.

Recent progress in deep learning–driven survival modeling has broadened our ability to forecast the longevity of AI systems. Recent techniques, including model architectures like dynamic survival networks, now accommodate time-varying covariates and complex, nonlinear hazard functions, offering a far more nuanced and realistic depiction of AI lifecycle risks.

Just as crucial are strides in model interpretability. Tools like median-SHAP have introduced much-needed transparency into survival models, helping practitioners and decision makers alike identify the most influential factors shaping an AI system's projected lifespan. This interpretability is a necessity for responsible AI governance and informed risk management.

By reframing AI lifecycle management through the lens of survival analysis, organizations gain a structured, quantitative approach to understanding and mitigating AI asset risk.

## APPLYING SURVIVAL ANALYSIS TO AI ASSETS

The first step in applying survival analysis to an AI asset is defining an appropriate "failure" or "degradation" event for the system in question. Unlike mechanical systems, in which failure may be binary (operational versus nonoperational), AI assets tend to degrade along a continuum of performance. Organizations must therefore establish performance thresholds below which a model is no longer fit for purpose. For example, a financial fraud-detection model might set a minimum acceptable precision-recall value; a healthcare diagnostic model may establish clinical-accuracy benchmarks.

Once the target metric and threshold are defined, survival curves can be estimated using historical model performance data, operational monitoring logs, and contextual factors. For narrow or highly specialized models, however, relevant historical data may be limited. In such cases, organizations often rely on internal benchmarking against similar past deployments or generate synthetic datasets to simulate expected performance trends. Recent methods such as Cox-Time, DeepHit, and dynamic survival networks enable modeling of static and time-varying covariates, capturing how both internal attributes and external shocks influence survival probabilities.[10,11]

Early-warning signs can be derived from survival curves by identifying inflection points where the estimated probability of maintaining acceptable performance declines rapidly. Such signals can inform proactive governance interventions.

## INTEGRATING SURVIVAL ANALYSIS INTO AI GOVERNANCE

A major strength of survival modeling is its compatibility with emerging AI governance frameworks. For example, researchers have proposed an hourglass governance model that embeds accountability throughout the AI lifecycle.[12] Although the hourglass model emphasizes continuity and lifecycle control, it is not always possible to recover a degrading model. In some cases (particularly when architectural constraints or external paradigm shifts are too great), full model replacement or redevelopment may be required. Survival curves provide quantitative input for this process, supporting lifecycle budgeting, model monitoring, and retirement planning.

In practical terms, survival-informed governance can:

- **Trigger adaptive maintenance cycles** — like retraining before critical degradation occurs
- **Support risk-adjusted ROI calculations** — factoring in expected asset lifespan
- **Inform compliance strategies** — ensuring AI systems remain aligned with evolving regulatory standards
- **Prioritize model-refresh investments** — focusing resources on high-risk assets

Moreover, survival-informed metrics can be incorporated into service-level agreements (SLAs) and internal audit processes, creating a more robust accountability structure.[13]

## LIMITS OF MODELING: TRANSPARENCY & REALISM

It is important to acknowledge that survival models are not crystal balls. Indeed, the modeling of AI asset survival cannot fully account for all externalities or emergent factors. Technological paradigm shifts (e.g., the rapid rise of generative AI) and abrupt regulatory changes are likely to outpace any predictive model.

To address this, survival analysis should be used as a strategic tool for directional insight, not as a deterministic forecast. Note that recent advances in explainability enable business leaders to interrogate survival models — understanding which factors are driving risk and where uncertainties lie.[14]

Organizations should also combine survival modeling with robust scenario planning, using model outputs as one input among many in strategic decision-making.

## LESSONS FROM AI ASSET OBSOLESCENCE

The rapid pace of innovation and the volatility of technological ecosystems have already produced high-profile examples of model obsolescence. The following examples illustrate a key pattern: the models in question did not degrade internally; they lost relevance relative to newer, more powerful alternatives. This form of relative degradation is especially common in fast-moving domains. Later, we briefly discuss intrinsic degradation, in which models decay due to internal factors like data drift or architectural limits.

### OPENAI

Few companies have moved faster than OpenAI. In 2020, OpenAI released GPT-3, a language model that quickly became a benchmark for natural language processing (NLP) performance. Organizations rushed to build applications on top of GPT-3, embedding it into chatbots, content generation tools, and customer service platforms.

Within three years, GPT-4 and GPT-4o were released, each demonstrating significant performance gains, broader capabilities, and improved safety protocols. The improvements rendered many GPT-3-based applications obsolete (no longer competitive in markets increasingly expecting the sophistication of newer models). Companies that had deeply integrated GPT-3 faced the costly decision of whether to migrate, retrain, or rebuild their AI infrastructure.

The degradation was not due to internal technical failure but to the external benchmark shifting upward. From a survival modeling perspective, this reflects a competitive displacement risk: the survival curve steepens not because the asset degrades intrinsically, but because the environment around it changes.

Organizations employing survival analysis could have anticipated this risk by monitoring indicators such as:

- Frequency and scale of major model updates from key AI vendors
- Industry benchmarking reports signaling performance-expectation shifts
- Acceleration in model capability metrics (e.g., parameters, context windows, multimodal abilities)

Proactive monitoring would have supported earlier budget allocation for migration and reduced business disruption.

## THE RISE OF FOUNDATION MODELS

In the mid-2010s, VGGNet was widely regarded as a top-performing deep convolutional neural network for image-recognition tasks. Organizations, particularly those in healthcare imaging and autonomous driving, invested heavily in VGGNet-based architectures.

However, the advent of ResNet (residual neural network) and later transformer-based vision models (e.g., vision transformers) dramatically shifted the performance frontier. ResNet introduced innovative skip connections that solved degradation problems in deep networks, achieving higher accuracy with more manageable training dynamics. Vision transformers brought further gains by capturing long-range dependencies more effectively.

The result was a structural obsolescence of VGGNet models, despite them remaining operational. Organizations that failed to adapt faced competitive disadvantages, with models delivering inferior performance, longer inference times, and/or higher error rates relative to newer architectures. This effect was especially pronounced in highly regulated sectors like healthcare and autonomous driving, where evolving safety, accuracy, and explainability standards pushed organizations to adopt models aligned with the latest technical and compliance benchmarks.

Survival analysis frameworks could have flagged early-warning signs such as:

- Emerging research momentum around alternative architectures (publication trends)
- Shifts in benchmarking leaderboards (e.g., ImageNet)
- Performance degradation relative to industry best practices

Incorporating these signals into survival curves would have allowed for more strategic refresh planning, enabling a smoother transition to next-generation architectures.

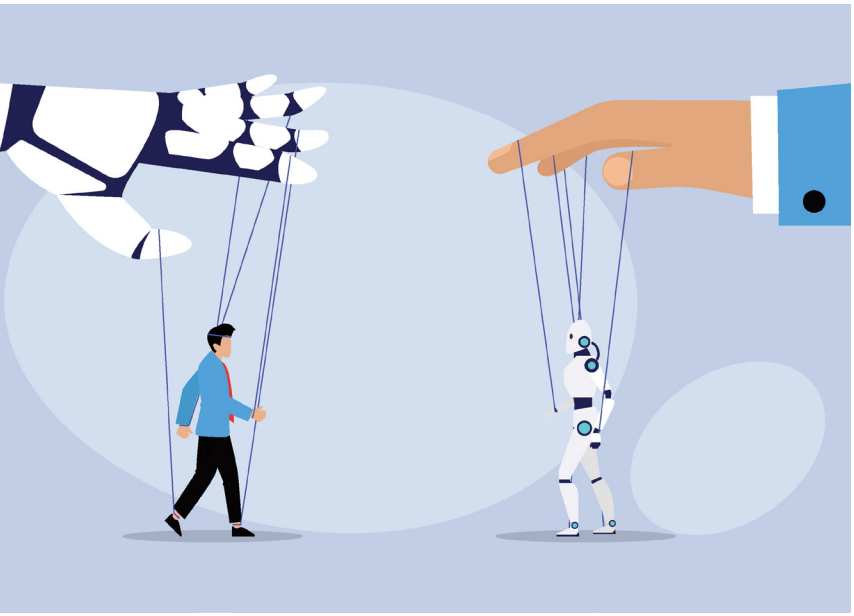
The OpenAI and VGGNet cases highlight relative obsolescence, but intrinsic degradation is also a significant concern. For example, a predictive maintenance model in manufacturing may deteriorate over time due to cumulative data drift, aging sensor hardware, or repeated deployment without proper recalibration. In such cases, performance may decline even in the absence of superior external models, underscoring the need for robust internal monitoring.

In all cases, the question of AI system longevity extends well beyond technical performance. It is deeply connected to external ecosystem shifts, evolving competitive pressures, and the relentless pace of technological change. In the absence of structured monitoring and predictive frameworks, organizations are left vulnerable to sudden system failures, costly reengineering efforts, erosion of market position, and/or reputational setbacks.

Conversely, firms that integrate survival analysis into their AI governance protocols gain a forward-looking advantage. By identifying early signals of obsolescence and planning transitions accordingly, they safeguard their existing AI investments while strengthening their capacity to adapt. In doing so, they are better positioned to harness the next wave of technological innovation with agility and confidence.

## IMPLICATIONS FOR AI GOVERNANCE & BUSINESS STRATEGY

Integrating survival analysis into AI lifecycle management is a strategic shift that extends beyond technical oversight into the domains of corporate governance, enterprise risk, and long-term business planning. It marks a departure from the prevailing set-it-and-forget-it mentality, introducing a more anticipatory and responsible approach to AI asset stewardship. This section shows how survival modeling can serve as a foundation for strengthening governance, enhancing strategic foresight, and embedding accountability across the AI value chain.



### SURVIVAL ANALYSIS AS A GOVERNANCE TOOL

A growing body of research about AI governance highlights the need for continuous, lifecycle-aware oversight, and survival analysis provides this.[15,16] Survival curves offer an integrative lens that facilitates shared understanding across data science, compliance, enterprise risk, and corporate strategy.

Areas where survival modeling can be embedded include:

- **Lifecycle budgeting.** Survival forecasts support more accurate planning for retraining and model updates, helping CFOs align financial decisions with AI system needs.
- **ROI adjustment.** Unlike static ROI models, survival-informed analysis accounts for performance decay, enabling more realistic and risk-aware investment decisions.
- **Regulatory compliance.** With regulators demanding ongoing validation (e.g., EU AI Act), survival analysis offers a structured way to demonstrate long-term model reliability.

In essence, survival curves function as a "health record" for AI assets — capturing performance over time and surfacing unexpected deviations that can trigger reexamination of assumptions, deployment context, or model design. When embedded into governance and strategic-planning processes, this approach transforms AI from a static capital expense into an actively managed enterprise asset.

### FORECASTING RISK & PRIORITIZING INVESTMENTS

Survival-informed governance can also create strategic advantage. In fast-moving industries, the ability to anticipate when AI assets will lose competitiveness (and plan transitions accordingly) can determine market leadership.

For example, firms that monitored survival indicators in NLP models during the rapid evolution from GPT-3 to GPT-4 were better positioned to manage customer expectations, allocate resources for migration, and avoid reputational risks from outdated offerings. Similarly, proactive lifecycle management in computer vision has enabled leading healthcare firms to maintain clinical accuracy levels by adopting newer models that redefined performance standards as imaging modalities and model architectures emerged.

By aggregating survival curves across AI assets, organizations can:

- Prioritize high-risk models for early intervention
- Identify robust models suitable for long-term deployment
- Balance investment between innovation (new models) and sustainability (maintaining existing assets)

Such portfolio insights support adaptive, resilient AI strategies that are aligned with both business goals and ethical governance imperatives.

### CREATING ACCOUNTABLE AI

Survival-informed governance also strengthens accountability by making AI system risks more transparent to both internal and external stakeholders. Studies have shown that transparency about AI performance over time is essential to maintaining trust with users, regulators, and society.[17,18]

Survival curves, particularly when enhanced with explainable AI techniques, can make degradation risks visible not only to technical teams but also to business leaders and external stakeholders. This fosters more informed dialogue about acceptable risk thresholds, update cadences, and end-of-life planning for AI systems.

In this way, survival analysis is not merely a technical enhancement. It is a governance innovation that helps organizations operationalize the principles of accountability, transparency, and continuous oversight that are central to the emerging global consensus on responsible AI.

## LIMITATIONS & FUTURE DIRECTIONS

Survival analysis offers a powerful framework for estimating the lifespan of AI assets, but it is important to recognize its inherent limitations. No statistical method, regardless of its complexity, can entirely account for the nuanced, evolving nature of real-world AI environments. Openly acknowledging these constraints preserves analytical integrity and encourages responsible integration of such tools into governance practices.

### MODELING BOUNDARIES & UNPREDICTABLE EXTERNALITIES

One challenge lies in the modeling of externalities. Technological paradigm shifts, abrupt regulatory interventions, and sudden market disruptions may introduce discontinuities that survival models, trained on historical data, cannot fully anticipate. For instance, few survival curves could have predicted the sudden leap in NLP capability delivered by large-scale foundation models in 2023/2024 or the accelerated emergence of multimodal AI capabilities.

Moreover, survival models depend on well-defined degradation signals. In domains where such signals are difficult to quantify (or where performance metrics are subjective), modeling reliability may be constrained. Recent studies highlight key limitations in survival modeling, including difficulty calibrating time-to-event predictions, managing censored observations, and accounting for nonlinear degradation patterns under real-world constraints.[19,20]

Survival analysis frameworks are sensitive to feature selection and context. Factors driving AI asset survival may vary dramatically across industries, regulatory regimes, and organizational maturity levels. Models must therefore be tailored with domain-specific expertise and regularly validated against evolving benchmarks.

### FUTURE RESEARCH DIRECTIONS

Recognizing these limitations opens several promising avenues for future development:

- **Dynamic survival modeling.** The integration of time-varying covariates and longitudinal monitoring offers a pathway to more responsive survival models that are better suited to environments characterized by rapid change.
- **Explainability and transparency.** Continued advances in interpretable survival modeling will be critical to ensuring that survival analysis can support accountable governance, providing not only risk forecasts but also insight into why certain risks are emerging.[21]

- **Governance integration.** Many governance frameworks still lack mechanisms for tracking long-term asset degradation.[22] Future work should explore how survival-informed monitoring can be more deeply embedded into governance processes, SLAs, and regulatory reporting.
- **Ethical alignment.** This is a complex topic requiring careful, dedicated treatment that cannot be included here due to space constraints. As a first step, survival-informed governance can contribute to ethical AI leadership by enhancing transparency and accountability (i.e., providing stakeholders with clearer insights into the risks and limitations of AI systems over time). A fuller exploration of the societal impacts of AI longevity remains an important area for future research.

Survival analysis should be viewed not as a panacea, but as a valuable addition to the AI governance tool kit. Its adoption must be accompanied by critical reflection, transparent communication, and adaptive learning. By doing so, organizations can responsibly harness survival analysis's strengths while remaining mindful of its boundaries.

## CONCLUSION

In an era when AI increasingly underpins critical business functions, the question of AI asset longevity is urgent. Unfortunately, most organizations lack a structured, predictive approach to managing AI lifecycle risks. Without such tools, they risk financial loss, operational disruption, compliance failure, and diminished competitiveness.

Survival analysis offers a practical and powerful way to address this gap. By modeling AI systems as dynamic assets subject to degradation and external shocks, survival analysis enables business leaders to:

- Forecast performance risks over time
- Identify early signs of obsolescence
- Align governance and budgeting with realistic lifecycle expectations
- Embed accountability and transparency into AI stewardship

Crucially, this approach supports a shift from reactive to proactive AI governance, a transition that is essential as regulatory expectations grow and technological change accelerates.

However, survival modeling is not a definitive predictor. It must be applied with transparency, domain sensitivity, and a recognition of its boundaries. Used appropriately, it complements other governance tools, enhancing both strategic planning and responsible leadership.

Ultimately, understanding AI survival is an accountability imperative. Organizations that embrace this mindset will protect their AI investments and foster trustworthy AI ecosystems that are better aligned with the expectations of customers, regulators, and society.

## REFERENCES


1 Vela, Daniel, et al. "Temporal Quality Degradation in AI Models." *Scientific Reports*, Vol. 12, July 2022.

2 Chen, Pin-Yu, and Payel Das. "AI Maintenance: A Robustness Perspective." *Computer*, Vol. 56, No. 2, February 2023.

3 Mäntymäki, Matti, et al. "Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance." arXiv preprint, 31 January 2023.

4 Batool, Amna, Didar Zowghi, and Muneera Bano. "AI Governance: A Systematic Literature Review." *AI and Ethics*, Vol. 5, January 2025.

5 Xue, Jingyuan, et al. "Survival Analysis with Machine Learning for Predicting Li-ion Battery Remaining Useful Life." arXiv preprint, 6 May 2025.

6 Bayram, Firas, Bestoun S. Ahmed, and Andreas Kassler. "From Concept Drift to Model Degradation: An Overview on Performance-Aware Drift Detectors." *Knowledge-Based Systems*, Vol. 245, June 2022.

7 Vela et al. (see 1).

8 Chen and Das (see 2).

[9] Mäntymäki et al. (see 3).

[10] Xue et al. (see 5).

[11] Mesinovic, Munib, Peter Watkinson, and Tingting Zhu. "DySurv: Dynamic Deep Learning Model for Survival Analysis with Conditional Variational Inference." *Journal of the American Medical Informatics Association (JAMIA)*, 21 November 2024.

[12] Mäntymäki et al. (see 3).

[13] Batool et al. (see 4).

[14] Mesinovic et al. (see 11).

[15] Mäntymäki et al. (see 3).

[16] Batool et al. (see 4).

[17] Mesinovic et al. (see 11).

[18] Chen and Das (see 2).

[19] Wiegrebe, Simon, et al. "Deep Learning for Survival Analysis: A Review." *Artificial Intelligence Review*, Vol. 57, No. 65, February 2024.

[20] Mesinovic et al. (see 11).

[21] Mesinovic et al. (see 11).

[22] Batool et al. (see 4).

# About the authors

**Chirag Kundalia** is Assistant Professor at Inspiria Knowledge Campus, India. His research examines technological resilience and quantitative modeling through the intersection of innovation economics, AI governance, and data-driven decision-making. With a focus on the implications of technology in both advanced and emerging economies, Mr. Kundalia's contributions aim to foster scholarly understanding and inform policy interventions. He can be reached at chiragkundalia@pondiuni.ac.in.

**V. Kavida** is Associate Professor in the Department of Commerce, Pondicherry University, India. She is an accomplished academic with extensive experience in economics, finance, and technological policy analysis. Dr. Kavida's research spans innovation economics, intellectual property rights, and intangible assets. She is passionate about exploring how emerging technologies influence economic resilience and societal progress. Dr. Kavida has contributed to numerous interdisciplinary studies and enjoys mentoring young scholars. Her dedication to ethical research and collaborative learning underscores her commitment to creating a positive impact through knowledge dissemination. Dr. Kavida can be reached at kavida4@yahoo.com.

# AMPLIFY

Anticipate, Innovate, Transform

**Cutter is Arthur D. Little's Open Consulting community, bringing expert academics and business leaders together to advance thinking in key areas of business and technology.**

**Arthur D. Little has been pushing the boundaries of innovation since 1886, linking people, technology and strategy to help our clients overcome today's most pressing challenges, while seizing tomorrow's most promising opportunities.**

**Our people are present in the most important business centers around the world, combining strong practical industry experience with excellent knowledge of key trends, technologies and market dynamics. We are proud to work alongside most of the Fortune 1000 companies and other leading firms and public sector organizations, supporting them to accelerate performance, innovate through convergence and digital and make a positive impact on the world.**

**It's what we believe makes *The Difference*.**

Founding Editor: Ed Yourdon
Managing Editor: Christine Generali
Publishing Manager: Linda Mallon Dias
Copyeditor: Tara K. Meads



CUTTER

AN ARTHUR D. LITTLE COMMUNITY

**For more content, visit www.cutter.com**