

How to Think About AI Risk: A business owner's guide to a cost you can already measure, price, and manage

An MLCommons perspective on treating AI risk as a managed cost rather than a catastrophe.

Abstract

The public conversation about artificial intelligence risk is largely conducted in the language of catastrophe: rogue systems, existential threats, and science-fiction imagery. For the people who actually bear the risk on the balance sheet, that framing is unhelpful because it uses the wrong unit of analysis. For a business owner or a Director of AI Governance, AI risk is simply another component of the risk registry to be managed. The argument here proceeds in five steps:

1. AI risk behaves like the risks your organization already manages, and its costs are already on your books;
2. The reward for managing those risks puts AI on the trajectory of the airline industry. Developing trust through reliability standards can turn an expensive novelty for early adopters into a common, safe way to travel. Or in the case of AI, a pillar of the global economy;
3. The era of grading your own homework is ending for the companies that build AI and the companies that deploy it;
4. Independent, consensus-governed measurement — MLCommons' AILuminate® — is homework nobody gets to design for themselves, which is precisely the kind of signal the emerging chain of assurers, insurers, and regulators will be looking for; and
5. The MLCommons approach is anchored on the principle of testing what you actually deploy, not just what was sold. With sufficient investment, AI risk can be identified, quantified, controlled, and reported, just as organizations have done for credit risk, currency exposure, and workplace safety for decades.

1. AI Behaves Like Other Risks You Currently Manage

If you ask executives who are not steeped in the topic of AI governance to describe "AI risk," the answer typically arrives in one of two registers. The first is apocalyptic and focuses on existential threats like systems autonomously producing catastrophe. The second is an apathetic shrug, because the apocalypse is not on this quarter's agenda. Both registers share a common defect: they treat AI risk as a category distinct from the ordinary business practices of managing risk exposure. This pushes what could be a standard business process decision toward either paralysis or dismissal. This outcome is not unique to the business side; many engineering teams find themselves in the same position of paralysis or dismissal.

We argue there is a more useful construction of AI risk: what strategic business objective are we trying to achieve, can AI help us do that, what is the risk of the given AI implementation, what could it cost us to understand and mitigate it, and what controls provide that mitigation? Phrased

that way, AI risk stops being speculation about the distant future and becomes present-day accounting. An AI system that produces defamatory statements, leaks proprietary or regulated data, gives dangerous or brand-unsafe advice, or generates content that violates the law creates liabilities. A coding agent that ships code that leaves the organization vulnerable to a cyber attack creates liabilities. Plaintiffs, regulators, customers, and costs follow. With this framing, we aren't trying to predict the AGI singularity; we're managing a business process. The benefits of that process are real - faster cycles, lower costs, new capabilities - and the risks are real too. Accounting means weighing both sides of the ledger, not just listing debits.

Every mature organization already runs a risk-management loop in some form. It identifies its exposures, estimates the expected loss from those exposures, applies controls in proportion to the risk's probability and impact, and reports the residual risk to someone accountable for carrying it. Insurers price that residual, auditors test it, and boards sign off on it. **AI doesn't need to break this loop; it just needs new measurement tools to quantify risk.** AILuminate is a step in the right direction for quantifying these risks - we are aiming to inform, not replace, these existing risk management loops. Integrating measurement like AILuminate and automated red teaming, for example, is about understanding the risk levels you can accept to keep high-value, AI-driven use cases unblocked.

The standards bodies and regulators have already reached a similar conclusion. In the US, the National Institute of Standards and Technology's AI Risk Management Framework organizes the entire discipline around four familiar functions: govern, map, measure, and manage (*NIST, 2023*). The European Union's AI Act takes a similar proportional stance, categorizing AI uses into risk tiers and attaching obligations that scale with potential harm (*European Union, 2024*). These positions acknowledge the premise that you do not eliminate risk; you manage it. For AI, standardized measurement is a critical part of that risk management process.

Available evidence suggests the cost of doing nothing is neither small nor hypothetical. IBM's 2025 Cost of a Data Breach study put the global average breach at USD 4.44 million and the United States average at a record USD 10.22 million (*IBM, 2025*). The same study found AI adoption outpacing AI oversight: 63% of breached organizations either had no AI governance policy or were still developing one; among organizations that suffered an AI-related security incident, 97% lacked functioning AI access controls; and organizations with high levels of unsanctioned "shadow AI" saw breaches cost roughly USD 670,000 more on average than those with little or none (*IBM, 2025*).

Regulatory exposure compounds the picture. The EU AI Act began phase-in in August 2024. It applies in stages: its prohibitions since February 2025, its general-purpose model obligations since August 2025, and most remaining provisions from August 2, 2026 — though the "digital omnibus" package adopted by the European Parliament and Council in June 2026 moves the compliance deadlines for high-risk systems out to December 2027 and August 2028 (*Gibson Dunn, 2026*). The penalty architecture is unchanged: the most serious violations carry fines of up to €35 million or 7% of total worldwide annual turnover, whichever is higher, with lower tiers at €15 million or 3% and €7.5 million or 1% (*European Union, 2024*). Fines expressed as a

percentage of global turnover are designed to be too painful to ignore. The EU AI Act is not the only regulatory regime to be mindful of in this space, of course. Existing laws already account for AI, even if they don't use the same terminology. For example, the GDPR's Article 22 and the CCPA demand strict controls around automated decisions. Product liability, defamation, and privacy regimes already cover many of the risks where AI will be, or already is, implicated.

Meanwhile, the risk-management infrastructure has not kept pace with adoption. In McKinsey's most recent State of AI survey, 88% of organizations report regularly using AI in at least one business function (*McKinsey, 2025*). However, when the same firm measured responsible-AI maturity, the average organization scored just 2.3 on a 0-to-4 scale in 2026, up only marginally from 2.0 the year before (*McKinsey, 2026*). As of early 2024, only 18% had an enterprise-wide council with the authority to govern AI decisions (*McKinsey, 2024*). The gap between adoption and governance is a liability being carried on the books today.

We aren't your risk team, or the owners of risk within your organization. However, we do see that AI-based risk accounting requires new tools to measure and monitor risk for those entities. These tools don't imply a brand-new philosophy to manage AI risk. We believe these tools must test what you deploy - they need to be available in your tenant, accessible by your production AI systems for both pre-launch testing and continuous monitoring. Managing risk is an activity your organization is likely already performing competently elsewhere. For AI, ***insist on the same standard of evidence and acumen you would demand anywhere else in your business operations.***

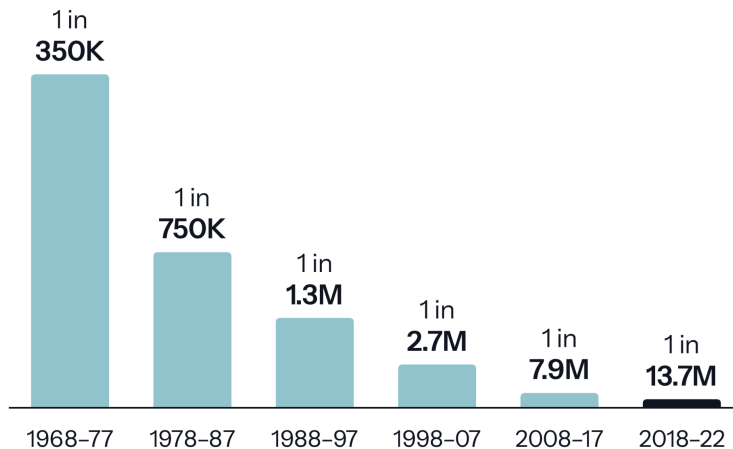
2. Managing Those Risks Allows AI to Become the Airline Industry

Why go to this trouble? Because the payoff for managing a new risk well is more than avoiding losses. The entire commercial aviation industry was built on new risk practices.

In 1959, U.S. commercial aviation suffered roughly 40 fatal accidents per million departures; within about a decade, that figure had fallen below 2, and it has continued to decline ever since (*Allianz Commercial, 2015*). MIT's Arnold Barnett, the dean of aviation risk statistics, calculates that the worldwide risk of dying on a commercial flight fell from 1 per 350,000 boardings in 1968–1977 to 1 per 13.7 million boardings in 2018–2022. The risk has halved roughly every decade for half a century, a compounding improvement Barnett likens to an aviation Moore's Law (*Barnett & Reig Torra, 2024*). By 2025, the industry recorded roughly one accident of any kind per 760,000 flights, and its own trade association could state flatly that "flying is the safest form of long-distance travel" (*IATA, 2026*). Aviation did not become safe because aircraft stopped suffering from mechanical or other failures; it became safe because it built the machinery of managed risk and never turned it off.

Flying became ~40× safer by treating risk as something to measure.

RISK OF DEATH PER PASSENGER BOARDING, WORLDWIDE



The risk of dying on a flight has halved roughly every decade.

Source: Barnett & Reig Torra (2024), Journal of Air Transport Management 119.

The modern aviation safety record, along with the sector's commercial success, followed investment in risk management. Framing "Risk" as innovation and progress produced a virtuous cycle. First, they measured and priced risk. The price then enabled insurance. Insurance brought audits and a safety culture of trust, supporting several billion passengers a year who treat flying as unremarkable. That trust is manufactured, and it is what turned a daredevil pursuit into infrastructure for the global economy. Similarly, the purpose of AI risk management is not to slow AI down; it is to make AI trustworthy enough (measurably, provably, insurably) to accelerate adoption for your highest value use cases. Trust at scale requires more than self-attestation; it is built from independent measurement, shared data, and accountable follow-through. If the AI industry builds that machinery, the industry can put itself on an airline industry-style trajectory.

3. The Era of Grading Your Own Homework Is Ending for Developers and Deployers

Accepting the premise of section 2, managing AI risk requires trustworthy measurement. This is where AI risk programs can quietly fail, even those with measurement in place. Much of the measurement currently on offer comes from the party with the most to gain from a flattering result. A model developer's own home-grown safety scores, produced on an undisclosed test set under conditions it controls, are not evidence that a regulator, an auditor, or an insurer is likely to accept. Therefore, even if they are 100% accurate and meticulously executed, those

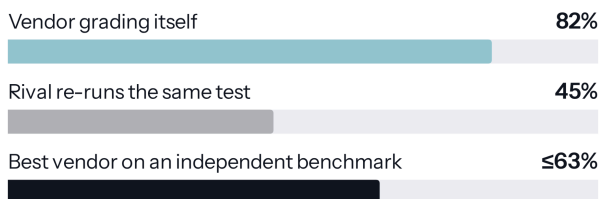
measurements are primarily marketing and sales assets.

Pharmaceutical companies do not adjudicate their own clinical trials. Public companies do not audit their own financial statements. Even the laboratories that issue calibration certificates are themselves accredited by someone else. Scholars of AI governance have mapped this same requirement onto AI, arguing that a functioning ecosystem of accredited, third-party oversight is a structural precondition for trustworthy AI claims (*Raji et al., 2022*). **We are not arguing that AI vendors are dishonest.** We are pointing out that self-graded evidence carries a structural, self-flattering bias that no amount of good faith fully removes. Internal audit plays a real role, and good companies do it, but internal audit answers to management and the board, not to the customers and regulators who rely on those claims. You cannot reliably grade your own homework — the incentives simply do not allow for it.

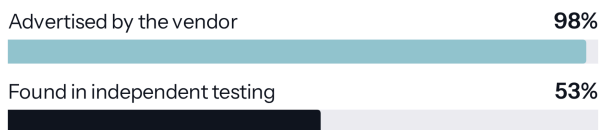
The public record furnishes many examples. In the market for AI code review, one vendor's self-run benchmark reported an 82% bug-catch rate; a rival re-ran the evaluation on the same repositories and scored it at 45%; and when a disinterested research lab built its own benchmark for the category, no vendor found more than 63% of the known issues (*DeepSource, 2026; Martian, 2026*). Scores measuring the same property differing so substantially draw scrutiny for the entire measurement enterprise..

Same product. Different grader. Different score.

AI CODE REVIEW · SHARE OF KNOWN BUGS CAUGHT



AI-DETECTION TOOL · ACCURACY (FTC CASE)



Sources: DeepSource (2026); Martian (2026); U.S. Federal Trade Commission (2025).

Regulators have taken notice. In 2025, the U.S. Federal Trade Commission ordered the AI-detection vendor Workado to substantiate its accuracy claims, after independent testing described in the FTC's complaint found its advertised 98% accuracy to be 53% on general-purpose content (*Federal Trade Commission, 2025*).

These distortions are fairly well understood now. We know to be wary of benchmark washing, the selective use of convenient results to imply a general safety or readiness that the evidence does not support. A second, more safety-specific distortion is safetywashing, capability gains

dressed up as safety progress, because many "safety" benchmarks correlate so strongly with raw capability that improving the model improves the score without actually improving safety (*Ren et al., 2024*). For a decision-maker, the practical lesson is to view safety claims through a lens of evidence quality, not just model quality.

As AI capabilities continue to improve and evolve, the demand for independent evidence is arriving on both sides of the vendor relationship. It is tempting for a deploying enterprise to treat measurement as the model developers' problem. It is not, for two reasons. First, the moment your team configures a model — a system prompt, a retrieval pipeline, fine-tuning, tool access, an agent scaffold — you have created a new system whose safety behavior nobody's report card describes, including the vendor's. Second, the obligations are landing on deployers directly: the EU AI Act attaches duties, and its €15 million / 3% penalty tier, to the organizations that deploy high-risk systems, not just to the developers that build them (*European Union, 2024, Art. 26*). Note also that the IBM breach statistics cited earlier are deployer statistics — shadow AI, missing access controls, and absent governance policies are failures of deployment, not of model training (*IBM, 2025*). When the assurance chain reaches you, your own internal evaluations, however diligent, will be your homework. The question every auditor asks applies on both sides of the vendor relationship: who measured this, and could anyone else reproduce it?

This is the standard AILuminate was built to meet. Independence is not a single thing: the methodology must be public and open to scrutiny; its development must accept open contributions under documents and institutionally adhered governance; governance and funding must be distributed across parties so no single entity controls the evaluation; and the benchmark needs ongoing stewardship, because methodology nobody maintains rapidly saturates or drifts out of date.

MLCommons checks these boxes by construction: a membership organization of over 125 participants spanning big tech, startups, academia, and nonprofits, producing industry grade benchmarks - not evaluations, audits, or red-teaming - through an open, consensus-based process that any participating company can inspect but none can dictate. We are not the only form independent evaluation will take, and we would be suspicious of anyone who claimed to be. But, for the question this paper addresses - can a buyer rely on a safety score when thinking about risk - a benchmark built this way is the strongest evidence on the market today.

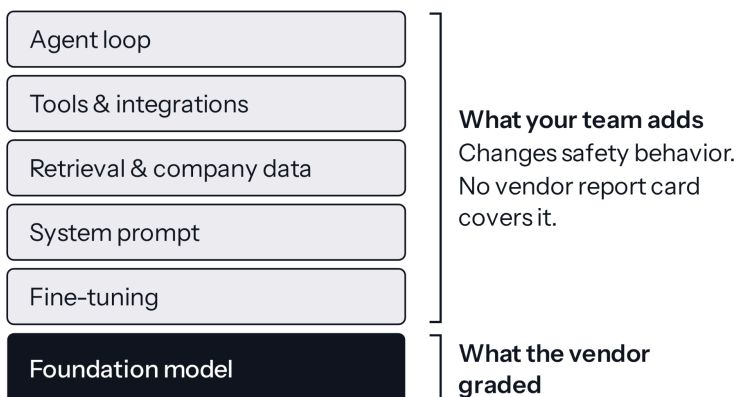
4. AILuminate Is Independent System Reliability Homework

This is the gap we intend to close with AILuminate®. MLCommons launched AILuminate in December 2024 as a family of independent, consensus-governed benchmarks that grade how reliably an AI system avoids producing harmful responses (*MLCommons, 2024; Ghosh et al., 2025*). Crucially, no single model developer produces this measurement. The MLCommons AI Risk & Reliability program's working groups designed it, with broad participation from AI companies, academia, and civil society. To resist gaming and contamination, MLCommons relies on private test sets, independent evaluators, and a continuous benchmark stewardship

program. No vendor wrote its own exam, no vendor can see the questions, and no vendor holds the grading pen.

ALLuminate tests an AI system against twelve hazard categories: physical hazards such as violent crime and indiscriminate weapons; non-physical hazards such as privacy, defamation, intellectual property, and hate; and contextual hazards such as specialized financial, health, and legal advice (*Ghosh et al., 2025*). The result is a single, legible grade on a five-tier scale from Poor to Excellent, reported overall and per hazard. The current v1.2 safety benchmark grades systems under normal conditions and now runs in English, French - with a Chinese and Hindi version in development with in-country partners (*MLCommons, 2025a, 2025b*). A companion v1.1 jailbreak benchmark, launched in June 2026, measures how a system's performance against the hazard categories holds up under deliberate adversarial pressure: in the Jailbreak v1.0 benchmark, eight open-weight systems were evaluated with 264 seed prompts spanning eleven hazard categories, and the unsafe-response rate rose from 11.08% under baseline conditions to 18.65% under jailbreak conditions - an average Resilience Gap of 7.57% (*MLCommons, 2026*).

**The vendor graded the model.
You deploy the system.**



35 of 39

systems earned lower safety grades under jailbreak attack than in normal use.

Test the system you deploy, not just the model you bought.

Source: MLCommons ALLuminate jailbreak v0.5 pilot (2025).

ALLuminate grades serve at two different points, and it matters which system is being graded. When you are buying, a vendor's published grade on the model as they shipped it is the comparable baseline: it is evidence about their system, and it is what makes one procurement choice defensible against another. But the grade that belongs in your risk registry is the one earned by the system you actually run — your configuration, your tenant, your tools. "Is this

model safe?" is unanswerable. "Did the system we deploy earn a Good or better on the hazards relevant to our use, on an independent benchmark?" is a question you can put in a risk registry or a contract. And answering it is your job, not the vendor's, which is why testing your own deployment, not just reading the vendor's report card, is the subject of the next section.

Now consider who else will be asking for that number, because the assurance chain that prices every other business risk has begun assembling around AI, and its direction of travel is easy to read. In late 2025, major insurers including AIG, Great American, and W. R. Berkley asked U.S. regulators for permission to exclude AI-related liabilities from corporate policies, with one underwriter telling the *Financial Times* that large language model outputs are "too much of a black box" to price (*Financial Times*, 2025). **This is not insurers saying AI risk is unmanageable; it is insurers saying unmeasured risk is uninsurable.** Where measurement exists, insured products can follow.¹

MLCommons is building for exactly that chain. In early 2026, it launched the AILuminate Global Assurance Program, which extends our benchmarks into benchmarking-as-a-service: private, non-saturated test sets; practice runs followed by official verified testing; and a Risk Label designed to slot directly into compliance and procurement workflows, developed with participants including PWC, KPMG, Google, Microsoft, and Qualcomm (*MLCommons*, 2026). AILuminate is not a complete risk program, and it does not claim to be. It is the industry-standard baseline instrument for turning AI reliability from a vibe into a metric. It is homework somebody else designed, graded, and reported.

5. Test What You Deploy with AILuminate®

The final discipline is the one most programs skip. A grade on a foundation model is evidence about the model as its developer shipped it. It is not evidence about your system. The moment your team adds a system prompt, a retrieval pipeline, fine-tuning, tools, or an agent loop, safety behavior changes, and the jailbreak results described above show how quickly grades can move when conditions do. **The unit of AI risk management is the deployed system, not the purchased model.** Procurement benefits from independent grading of the models you buy; governance should retest the configuration you actually ship on the same instrument to keep numbers comparable end to end. Closing that gap - testing the system you actually deploy, on an independent, continuous basis - is where the market is now heading, and it is precisely the evidence that deployer-side obligations, from EU AI Act Article 26 to your own insurer's questionnaire, will ask you to produce

¹ Munich Re has underwritten AI performance since 2018 and markets that coverage explicitly as a way for AI providers to earn customer trust (*Munich Re*, n.d.); Lloyd's underwriters began writing affirmative AI liability cover in 2025 (*Armilla AI*, 2025); and reinsurance brokers now argue publicly that without better, independent model evaluation, insurers are left "pricing uncertainty rather than risk" — an outcome that "almost always leads to higher premiums, narrower coverage or both" (*Gallagher Re*, 2026). Demand is not the constraint: in a Geneva Association survey of 600 corporate insurance decision-makers, more than 90% wanted insurance coverage for generative AI risks, and two-thirds would pay premiums at least 10% higher for it (*Geneva Association*, 2025). Governments are building the same muscle on the audit side — the UK counts a £1.01 billion domestic AI assurance market, which it projects could exceed £6.5 billion by 2035 (*DSIT*, 2024). This is the aviation sequence running again, a century later: the underwriters are asking for the airworthiness certificate before the regulators finish writing their rules.

OR

MLCommons' assurance programs were built to take independent measurement beyond public leaderboards, and closing the deployer gap - testing the system you actually deploy, on the same instrument - is the direction that work is heading. This is precisely the evidence that deployer-side obligations, from EU AI Act Article 26 to your own insurer's questionnaire, will ask you to produce.

A grade from Alluminate is a beginning, not an end, for AI risk management. A benchmark establishes a comparable baseline. It is not a replacement for your own custom benchmarking, red teaming, continuous monitoring, etc. It helps average away the rare, severe failures that approaches like red teaming are designed to find. Similarly, continuous monitoring helps detect failure modes that can emerge in agentic systems over time. Microsoft's AI Red Team maintains a comprehensive taxonomy of these failure modes (*Microsoft AI Red Team, 2026*). We encourage AI governance teams to consider all the tools at their disposal when developing a robust, in-depth risk-management strategy, and we are excited to support enterprises with our Alluminate family of benchmarks as they continue to deploy those strategies.

Conclusion

The most important step a leader can take regarding AI risk is to treat it like any other risk. Risk-based costs can be identified, priced, proportionally managed, and reported to accountable and interested parties. The exposure is already on your books — the choice is whether to measure it. Independent reliability measurement is needed to engage with the realities of what a board, a regulator, or an insurer will accept; it is how aviation earned a century of trust, one independent measurement at a time. This is the problem we aim to solve with Alluminate, and we look forward to going on that journey together.

References

Allianz Commercial. (2015). *How aviation safety has improved.* Allianz Global Corporate & Specialty.
<https://commercial.allianz.com/news-and-insights/expert-risk-articles/how-aviation-safety-has-improved.html>

Armilla AI. (2025, April 30). *Armilla launches affirmative AI liability insurance with Lloyd's underwriter, Chaucer* [Press release]. PR Newswire.
<https://www.prnewswire.com/news-releases/armilla-launches-affirmative-ai-liability-insurance-with-lloyds-underwriter-chaucer-302442586.html>

illa-launches-affirmative-ai-liability-insurance-with-lloyds-underwriter-chaucer-302442586.html)

Barnett, A., & Reig Torra, J. (2024). Airline safety: Still getting better? *Journal of Air Transport Management, 119*, 102641.
https://doi.org/10.1016/j.jairtraman.2024.102641

DeepSource. (2026, February 26). *Every AI code review vendor benchmarks itself, and wins.* DeepSource.
https://deepsources.com/blog/ai-code-review-benchmarks

Department for Science, Innovation and Technology. (2024, November 6). *Assuring a responsible future for AI.* UK Government.
https://www.gov.uk/government/publications/assuring-a-responsible-future-for-ai

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union.
https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Federal Trade Commission. (2025, April 28). *FTC order requires Workado to back up artificial intelligence detection claims* [Press release].
https://www.ftc.gov/news-events/news/press-releases/2025/04/ftc-order-requires-workado-back-artificial-intelligence-detection-claims

Financial Times. (2025, November 23). *Insurers retreat from AI cover as risk of multibillion-dollar claims mounts.*
https://www.ft.com/content/abfe9741-f438-4ed6-a673-075ec177dc62

Gallagher Re. (2026, June). *Anthropic's fourth way: Why restricted AI models are a challenge for insurers.*
https://www.ajg.com/gallagherre/-/media/files/gallagher/gallagherre/news-and-insights/2026/june/restricted-ai-models-insurance-risk-analysis-2026-gallagher.pdf

Geneva Association. (2025, October 14). *Gen AI risks for businesses: Exploring the role for insurance.*
https://www.genevaassociation.org/publication/digital-ai-transformation/gen-ai-risks-businesses-exploring-role-insurance-0

on/gen-ai-risks-businesses-exploring-role-insurance-0)

Ghosh, S., Frase, H., Williams, A., Luger, S., Röttger, P., Barez, F., McGregor, S., Fricklas, K., Kumar, M., Feuillade-Montixi, Q., Bollacker, K., Friedrich, F., Tsang, R., Vidgen, B., . . .

Vanschoren, J. (2025). *AILuminate: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons* (arXiv:2503.05731). arXiv.

<https://arxiv.org/abs/2503.05731>

Gibson Dunn. (2026, May 27). *EU AI Act omnibus agreement: Postponed high-risk deadlines and other key changes* [Client alert].

<https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/>

IATA. (2026, March 9). *IATA releases 2025 safety report* [Press release]. International Air Transport Association.

<https://www.iata.org/en/pressroom/2026-releases/2026-03-09-01/>

IBM. (2025, July 30). *Cost of a data breach report 2025.* IBM Corporation (research conducted by Ponemon Institute).

<https://www.ibm.com/reports/data-breach>

Martian. (2026, February 26). *Code review bench: Towards billion dollar benchmarks.*

<https://withmartian.com/post/code-review-bench-v0>

McKinsey & Company. (2024, May 30). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value.**

<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024>

McKinsey & Company. (2025, November 5). *The state of AI in 2025: Agents, innovation, and transformation.*

<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>

McKinsey & Company. (2026, March 25). *State of AI trust in 2026: Shifting to the agentic era.*

<https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era>

Microsoft AI Red Team. (2026, June 4). *Taxonomy of failure modes in agentic AI systems* (Version 2.0) [Whitepaper]. Microsoft.

<https://www.microsoft.com/en-us/security/blog/2026/06/04/updating-taxonomy-failure-modes-agentic-ai-systems-year-red-teaming-taught-us/>

MLCommons. (2024, December 4). *MLCommons launches AILuminate, first-of-its-kind benchmark to measure the safety of large language models.*

<https://mlcommons.org/2024/12/mlcommons-ailuminate-v1-0-release/>

MLCommons. (2025a, February 10). *MLCommons releases AILuminate LLM v1.1, adding French language capabilities to industry-leading AI safety benchmark.*

<https://mlcommons.org/2025/02/ailuminate-v1-1-fr/>

MLCommons. (2025b, May 29). *MLCommons announces expansion of industry-leading AILuminate benchmark.*

<https://mlcommons.org/2025/05/nasscom/>

MLCommons. (2025c, October 15). *MLCommons unveils new jailbreak benchmark, quantifying AI's "resilience gap" to adversarial attacks.*

<https://mlcommons.org/2025/10/ailuminate-jailbreak-v05/>

MLCommons. (2026, February 19). *A new standard for AI risk: How the AILuminate Global Assurance Program is reshaping reliability.*

<https://mlcommons.org/2026/02/ailuminate-global-assurance/>

Munich Re. (n.d.). *aiSure: Insurance for artificial intelligence.* Retrieved July 23, 2026, from <https://www.munichre.com/en/solutions/for-industry-clients/insure-ai.html>

NASA ASRS. (n.d.). *ASRS: The case for confidential incident reporting systems* (Publication 60). NASA Aviation Safety Reporting System.

[https://asrs.arc.nasa.gov/docs/rs/60_Case_for_Confidential_Incident_Reporting.pdf](https://asrs.arc.nasa.gov/docs/rs/60_Case_for_Confidential_Incident_Reporting.pdf)

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce.

<https://doi.org/10.6028/NIST.AI.100-1>

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. E. (2022). Outsider oversight: Designing a third-party audit ecosystem for AI governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571.

<https://doi.org/10.1145/3514094.3534181>

Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., ... Hendrycks, D. (2024). *Safetywashing: Do AI safety benchmarks actually measure safety progress?* (arXiv:2407.21792). arXiv. <https://arxiv.org/abs/2407.21792>

Underwriters Laboratories. (2019, February 15). *Taking flight: The early years of aviation.** UL. <https://www.ul.com/news/taking-flight-early-years-aviation>