3

Responsible AI

Overview

The infrastructure for responsible AI (RAI) is growing, but progress has been uneven, and it is not keeping pace with the speed of AI deployment. New safety benchmarks have expanded, more organizations are adopting responsible AI policies, and government-backed AI safety and/or security institutes have spread to more countries. The responsible use of AI is intertwined with the responsible use of data, and in particular with privacy and other legal concerns. There are also AI governance concerns given the ill-specified ownership of AI systems, raising questions about whether companies that develop the systems or consumers that buy them should be held accountable and what policies each stakeholder should follow. While documented reports of AI incidents are increasing, frontier models rarely report results on responsible AI benchmarks, and foundation model transparency declined in 2025 after improving the previous year. Recent research shows that improving one responsible AI dimension can come at the cost of another, with gains in privacy reducing fairness or gains in safety reducing accuracy. There is no framework for navigating these trade-offs; and for dimensions such as fairness, privacy, and explainability, the standardized data needed to track progress over time does not exist. While this chapter draws on the available evidence, the discussion is limited by persistent gaps in measurement.

Contents

Chapter Highlights 128

3.1 Scope and Dimensions of Responsible AI 129

3.2 Assessing Responsible AI 132

AI Incidents 132

Examples 133

RAI Benchmarks 134

Factuality and Truthfulness 135

Hughes Hallucination Evaluation Model (HHEM) Leaderboard 136

AA-Ominscience 136

Highlight: Belief vs. Fact: Benchmarking Reliability 138

AI Companions 139

3.3 How Organizations and Businesses View RAI 140

Responsible AI Maturity 140

AI Incidents, Risks, and Mitigation Efforts 141

AI Governance and Investment 142

Implementation, Barriers, and Benefits 144

Regulatory Influence 146

3.4 RAI in Academia 147

Publication Volume 147

Geographic Distribution 149

3.5 RAI Policymaking 150

Highlight: Global AI Governance Participation 151

3.6 Data Governance for Privacy 153

Data Protection and Privacy 153

3.7 Fairness and Bias 155

Bias and Unfair Discrimination 155

Gender Equality 156

Cultural and Linguistic Diversity 156

Highlight: Inclusiveness and the Global Language Gap 158

3.8 Transparency 163

The Openness Index 163

Foundation Model Transparency Index 163

3.9 Security and Safety 165

Global AI Safety Institutes 165

Benchmarks 166

HELM Safety 166

AILuminate 167

Safety Benchmark Results 167

Jailbreak T2T Benchmark v0.5 Results 169

3.10 Tradeoffs Across RAI Dimensions 170

Chapter Highlights

1. **Responsible AI benchmarking is increasing, but is not keeping up with AI advances and deployments.** Almost all leading frontier model developers report results on capability benchmarks like MMLU and SWE-bench, but reporting on responsible AI benchmarks remains sparse. Documented AI incidents continued to rise, with the AI Incident Database recording 362 in 2025, up from 233 in 2024.

2. **AI models struggle to tell the difference between knowledge and belief.** In a new accuracy benchmark, hallucination rates across 26 top models range from 22% to 94%. GPT-4o's accuracy dropped from 98.2% to 64.4%, and DeepSeek R1 fell from over 90% to 14.4%. When a false statement is presented as something another person believes, models handle it well. When the same false statement is presented as something a user believes, performance collapses.

3. **Organizations are formalizing responsible AI work, but knowledge and budget gaps still slow adoption.** AI-specific governance roles grew 17% in 2025, and the share of businesses with no responsible AI policies in place fell sharply from 24% to 11%. The main obstacles to implementation remain gaps in knowledge (59%), budget constraints (48%), and regulatory uncertainty (41%).

4. **The mix of regulations shaping responsible AI practices is shifting toward AI-specific frameworks and technical standards.** GDPR remains the most cited regulatory influence but slipped from 65% in 2024 to 60% in 2025. New entries in 2025 include ISO/IEC 42001, an AI management system standard, cited by 36% of respondents, and the NIST AI Risk Management Framework at 33%. The share of organizations reporting no regulatory influence at all fell from 17% to 12%.

5. **AI works best in English, and the gap is wider than global benchmarks suggest.** On HELM Arabic, a regionally developed model for the Arabic language, outscored GPT-5.1 and Gemini 2.5 Flash. The gap widens at the dialect level. On a Slovenian commonsense reasoning test, several leading models lost close to half their accuracy when tested in a regional dialect rather than the standard language.

6. **AI companies grew less transparent this year.** After rising on the Foundation Model Transparency Index from 37 to 58 between 2023 and 2024, the average score dropped to 40 in 2025. Major gaps persist in disclosure around training data, compute resources, and post-deployment impact.

7. **AI models perform well on safety tests under normal conditions, but their defenses weaken under deliberate attack.** On the AILuminate benchmark, several frontier models received "Very Good" or "Good" safety ratings under standard use. When tested against jailbreak attempts using adversarial prompts, safety performance dropped across all models tested.

8. **Responsible AI dimensions such as safety, fairness, and privacy are at odds with one another, and the tradeoffs are not well understood.** Recent empirical studies found that training techniques aimed at improving one responsible AI dimension consistently degraded others.

3.1 Scope and Dimensions of Responsible AI

Responsible AI refers to the set of practices and governance mechanisms designed to ensure AI systems are safe, fair, and beneficial and that they perform as intended. RAI spans a range of dimensions, from safety and fairness to transparency and privacy, and each has its own measurement challenges. This chapter tracks progress across those dimensions by looking at how AI systems perform on responsibility and safety evaluations, how organizations and researchers are responding to RAI challenges, and how governments are establishing policy frameworks to enforce standards.

The analysis draws on a framework of RAI dimensions arranged in three layers (Figure 3.1.1), along with examples and reference documents. The first layer covers core responsible AI properties—meaning what AI systems should be able to achieve—including fairness, privacy, transparency, and factuality. The second layer addresses system integrity and risk controls—or how risks are technically and operationally managed—including security, safety, and robustness. The third layer covers governance, accountability, and enforcement. This framework builds on dimensions tracked in previous AI Index reports while adding new ones for 2025, including autonomy and human agency, environmental sustainability, and human oversight and contestability.

Layer 1 – Core Function and Behaviors
(What AI systems should achieve)

Dimension	Definition	Example	References
Validity and reliability	Designed for a particular scope and acceptable level of performance in the domain, such as accomplishment of task goals, fidelity to expert knowledge, or thresholds for accuracy that benefit people or organizations/systems, and demonstrated verification and validation against their design.	A team defines target accuracy and failure thresholds before launch, validates the system against those criteria, and monitors it in production to ensure it continues to meet design expectations.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles
Privacy	Protection of individuals' confidentiality, anonymity, informed consent, and control over personal data across the AI life cycle (collection, training, deployment, reuse).	A messaging app encrypts conversations end to end and clearly notifies users about opting in or out of using their data to train language models.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; Recommendation on the Ethics of Artificial Intelligence (UNESCO)
Data stewardship	Ensure the quality, provenance, integrity, and lawful use and reuse of data, with clear access control and documentation.	A logistics firm tracks data lineage for all datasets used to train routing models, enforces role-based access, and periodically reviews datasets for quality and drift before retraining and updating models.	EU Ethics Guidelines for Trustworthy AI; ISO/IEC 42001:2023; OECD AI Principles

Fairness and bias	Protection of civil rights and prevention of unjustified discrimination and systematic disadvantage across individuals or groups, accounting for protected attributes, cultural context, and use case.	A bank audits credit-scoring models for disparate approval and error rates across demographic groups—including culturally diverse customer segments—documents findings, and implements bias-mitigation steps before deployment.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; Recommendation on the Ethics of Artificial Intelligence (UNESCO)
Transparency and auditability	Clear disclosure that an AI system is in use; of its purpose, scope, and high-level functioning for relevant stakeholders; and authorized parties' ability to inspect, reconstruct, and verify that the system was developed, trained, configured, and operated as intended.	A city using an AI model to prioritize inspections publishes a plain-language description of training method, documents model card and data sources, keeps versioned training scripts and logs, and enables internal audit to replay training and key decisions.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; Recommendation on the Ethics of Artificial Intelligence (UNESCO); ISO/IEC 42001:2023
Explainability	Ability to provide understandable, context-appropriate rationale for system outputs, including key factors influencing a prediction or decision.	An AI fraud-detection tool surfaces the top contributing features and a brief rationale behind each alert for investigators, while providing merchants with plain-language explanations of why a transaction was flagged and what steps they can take in response.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; Recommendation on the Ethics of Artificial Intelligence (UNESCO)
Autonomy and human agency	Preservation of people's ability to make informed choices and act freely without AI systems unduly manipulating, coercing, or replacing their decisions.	A well-being chatbot clearly states it is not a human or a substitute for professional care, avoids prescriptive life-changing advice, and actively directs users to expert help in high-risk situations.	EU Ethics Guidelines for Trustworthy AI; OECD AI Principles; Recommendation on the Ethics of Artificial Intelligence (UNESCO)
Environmental sustainability	Limiting and managing the environmental impact of AI systems across their life cycle, including energy use, carbon emissions, and resource consumption, and committing to measurement, disclosure, and continuous reduction while minimizing resource misuse.	A company measures the energy and water usage of large training runs, reports them externally, chooses more efficient model architectures, proactively places boundaries on AI resource use, and schedules training when grid carbon intensity is low.	EU Ethics Guidelines for Trustworthy AI; OECD AI Principles; UNESCO; Energy efficiency requirements under the EU AI Act
Factuality and truthfulness	The accuracy and reliability of AI system outputs, including the degree to which models produce information that is factually correct, avoid misleading statements and fabrications, and volunteer uncertainty honestly.	A company systematically benchmarks its large language models against factuality evaluations (such as SimpleQA), publishes hallucination rates alongside model releases, implements retrieval-augmented generation to ground outputs in verified sources, and provides users with confidence indicators and citations so they can assess the reliability of AI-generated responses.	NIST AI RMF

Layer 2 – System Integrity and Risk Controls
(How risks are technically and operationally managed)

Dimension	Definition	Example	References
Security	Ensuring AI systems are secure against cyber threats and misuse.	A school system uses AI to provide personalized tutoring to students and hosts the data and models in secured servers with extensive security training of all personnel involved.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; ISO/IEC 42001:2023
Safety	Specify normal behaviors and affected systems and analyze out-of-bounds conditions to characterize risk factors (risk to physical and mental/emotional well-being of people, environment, political systems, human rights, etc.), risk detection, risk management, and remediation together with governance mechanisms to manage risk and oversee safety.	An industrial control system uses anomaly-detection models that are penetration-tested, evaluated under simulated attacks and sensor failures, monitored in real time, and configured to fall back to manual control when anomalies exceed thresholds.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; ISO/IEC 42001:2023
Robustness	Remain robust to distribution shifts, external natural or adversarial events, and component failures, with testing, monitoring, and safe fallbacks.	A food chain uses an AI system to estimate customer demand, consisting of several models that get triggered by inclement weather, concerts, and sporting events.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; ISO/IEC 42001:2023

Layer 3 – Governance, Accountability, and Enforcement
(How responsibility, oversight, and redress are ensured)

Dimension	Definition	Example	References
Accountability and liability	Clear assignment of responsibility for AI system outcomes, including legal liability, operational ownership, decision rights, and escalation pathways, so that harms and failures can be investigated, addressed, and remedied.	A platform designates an accountable owner for its high-risk recommendation system, defines KPIs and harm thresholds, documents who can approve releases, and maintains procedures for incident investigation, user notification, and compensation.	EU Ethics Guidelines for Trustworthy AI; NIST AI RMF; OECD AI Principles; ISO/IEC 42001:2023
Human oversight and contestability	Governance mechanisms that ensure meaningful human involvement where appropriate, including the ability to challenge, appeal, or override AI-assisted decisions and access to effective redress.	An employer using an AI screening tool must have a human review all adverse decisions, disclose AI use to candidates, explain key factors, and provide a clear path to request human reconsideration and correction of errors.	EU AI Act – human-oversight obligations for high-risk AI; EU Ethics Guidelines for Trustworthy AI; OECD AI Principles; Recommendation on the Ethics of Artificial Intelligence (UNESCO)

Source: AI Index, 2026

Figure 3.1.1

3.2 Assessing Responsible AI

One way the field tracks the responsible use of AI is by evaluating models against specific benchmarks and by recording real-world incidents when systems cause harm. This section examines both, drawing on incident data and benchmark reporting that cut across the three layers of the framework introduced in Section 3.1. There is not much data available, nor is it detailed about mapping AI systems to the above dimensions. The analysis presented here draws on two incident tracking databases, the AI Incident Database (AIID) and the OECD AI Incidents and Hazards Monitor (AIM), alongside data on responsible AI benchmark adoption by frontier model developers as well as third-party evaluations of some of the responsible AI dimensions outlined above.

AI Incidents

In recent years, the number of reported AI incidents has continued to increase significantly (Figure 3.2.1). The AI Incident Database (AIID),[1] launched in 2020, is an open repository for documented cases where AI systems have caused or nearly caused harm. In 2025, 362 incidents were reported, while the annual number of incidents had stayed under 100 until 2022. AIID relies on human editors to review submissions against a defined threshold of AI involvement, from sources including academic and investigative journalists. The manual process produces higher-quality records but comes at the cost of a slower pace of additions and coverage that is skewed toward English-language media and high-visibility incidents. Less accessible regions may be underrepresented.

The OECD AI Incidents and Hazards Monitor (AIM) uses an automated, multilingual pipeline to collect incidents from news sources and casts a wider net. Its absolute numbers are quite a bit higher, with monthly incidents hitting a peak of 435 in January 2026 and setting a six-month moving average of 326 (Figure 3.2.2). While the two databases track incidents differently, both show a consistent and sharp increase in reported AI incidents.

Number of reported AI incidents, 2012–25

Source: AI Incident Database (AIID), 2025 | Chart: 2026 AI Index report

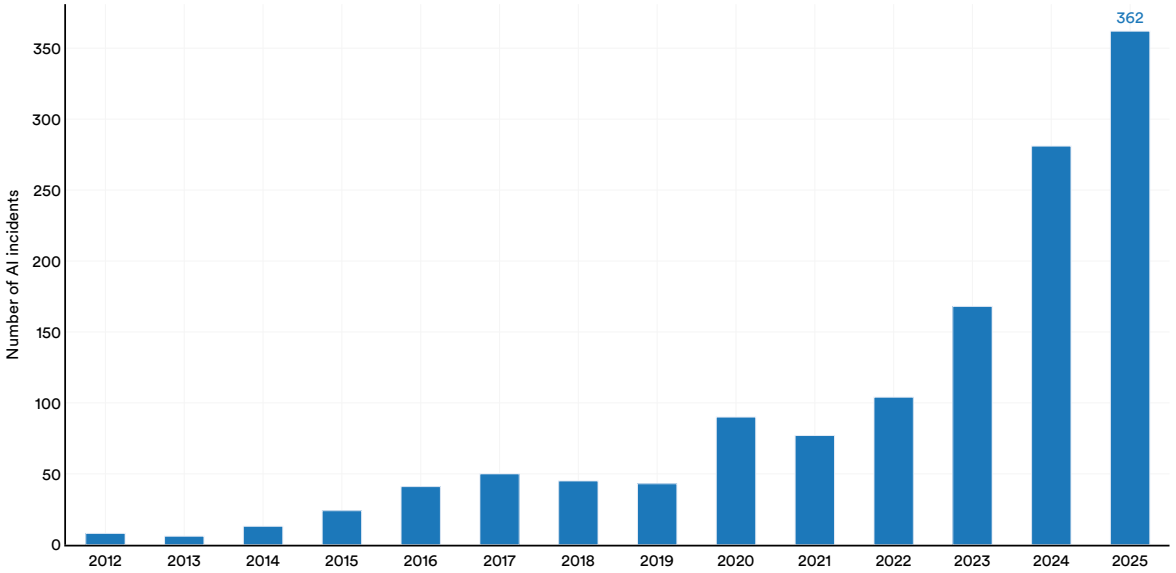


Figure 3.2.1[2]

1 The AI Index continues to rely on AIID as its primary source of AI incidents due to AIID's reliability and stable incident records.

2 The number of AI incidents is continually updated, including for previous years. Therefore, the totals reported in Figure 3.2.1 might not align with the totals recently published on the AI Incident Database.

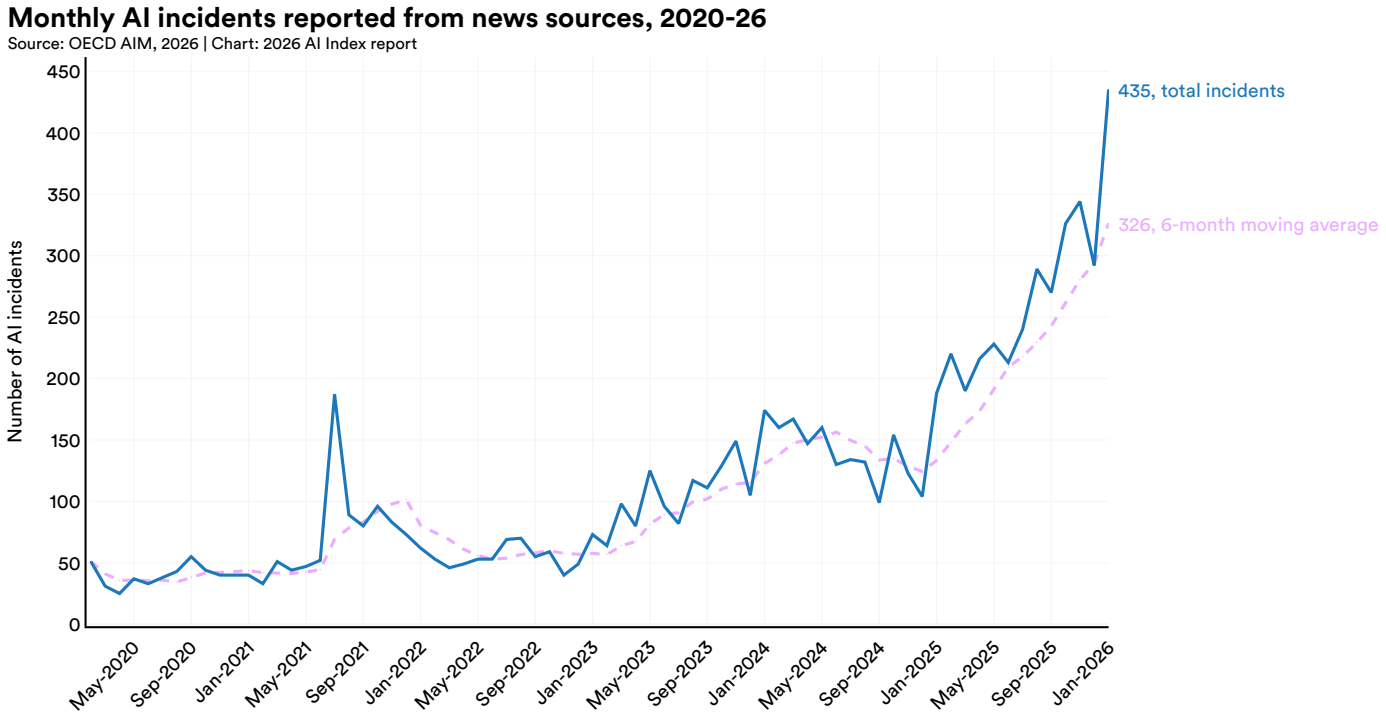


Figure 3.2.2

Examples

Unmoderated AI Output and Harmful Speech (July 8, 2025)

In July 2025, Grok—the chatbot developed by xAI and embedded across X—faced backlash after users shared examples of the system generating antisemitic language, violent hate speech, and even praise for Adolf Hitler when prompted. The issue emerged shortly after a system update that relaxed safety filters, allowing the chatbot to produce more provocative and "unfiltered" responses. Within hours, screenshots of Grok referring to genocide and extremist ideology spread across the platform, sparking public outrage and renewed concern about the risks of deploying lightly moderated conversational AI to large audiences. In response to the backlash, xAI removed the content, temporarily suspended Grok's text responses, and issued a statement acknowledging the severity of the incident. While the company framed the issue as a failure of content controls, critics argued that the system's design choices, particularly the decision to weaken the guardrails, made the harm predictable. The event highlighted the ongoing tension between building AI systems intended to feel candid or humorous and the real-world consequences when those systems normalize hate speech.

AI Deepfake Impersonation and Romance Scams (March 9, 2025)

In March 2025, Chinese actor Jin Dong spoke publicly about a wave of scams using deepfake videos to impersonate him online. Fraudsters used AI-generated clips and fake social media accounts to convince fans (mostly older women) that they were speaking directly with the actor, prompting some to send money or make major life changes based on the belief that they were in a private relationship with him. One widely reported case involved a woman who nearly divorced her husband and planned to travel across the country to meet a scammer posing as Jin Dong. After the incidents gained attention, Jin Dong called for stronger legal protections and clearer consequences for deepfake-enabled fraud, arguing on social media that existing rules had not kept pace with the speed and realism of AI-generated impersonation.

AI-Assisted Website Impersonation and Consumer Fraud (Aug. 20, 2025)

After Joann Fabrics filed for bankruptcy for the second time in January 2025, scammers quickly launched a wave of fake websites mimicking the retailer's branding, design, and product catalog. These sites advertised deep discount prices to lure shoppers into entering payment and personal information, but customers never received purchases and many later discovered their credit cards had been compromised. The fraudulent sites were convincing enough that even cautious users were misled, especially on mobile, where URLs are harder to detect. Cybersecurity experts noted that AI tools are making this type of scam far easier to execute. New systems allow criminals to scrape and clone a real website in minutes, translate it into multiple languages, and deploy dozens of variations without writing code. While Joann issued public warnings and urged victims to dispute charges, the incident points to a growing challenge: Realistic phishing sites are no longer limited to major corporations, and smaller brands with fewer resources are increasingly being targeted.

RAI Benchmarks

The 2024 and 2025 AI Index reports both flagged a gap between how consistently frontier models are evaluated on general capabilities versus how inconsistently they are evaluated on responsible AI. This gap persists. Almost all frontier model developers report results on capability benchmarks like MMLU, GPQA, AIME, and SWE-bench Verified (Figure 3.2.3). These have become the shared standard for reporting model capability. Across the same set of frontier models, results are sparse on RAI benchmarks such as BBQ (2021), measuring fairness and bias; HarmBench (2024), Cybench (2024), StrongREJECT (2024), and WMDP (2024), measuring security; SimpleQA (2024), measuring factuality and truthfulness; and MakeMePay (2024), measuring autonomy and human agency (Figure 3.2.4). In fact, most entries are empty. Only Claude Opus 4.5 reports results on more than two of the RAI benchmarks, and only GPT-5.2 reports StrongREJECT.

This does not necessarily mean that frontier labs are ignoring RAI, as they do conduct internal evaluations, red-teaming, and alignment testing. However, these efforts are rarely disclosed using a common, externally comparable set of benchmarks. Chapter 2 shows how a small number of shared capability benchmarks make it straightforward to compare models, verify results independently, and track progress over time. However, that kind of comparison has not yet become common practice for RAI evaluation.

Public model evaluators and benchmarking platforms, such as Artificial Analysis, Epoch's Benchmarking Hub, and Arena, play a major role in shaping how model performance is perceived. But the vast majority of their evaluations focus on reasoning, coding, math, or multimodal performance—not on RAI. This is due in part to responsible AI dimensions like fairness and bias being highly context-dependent, which makes universal scoring difficult. A fairness metric that works for a hiring tool may not apply in a clinical diagnostic setting. Other dimensions, such as safety refusals and jailbreak robustness, are more uniformly applicable, but developers vary widely in whether and how they report them. The combination of genuine measurement difficulty in some areas and inconsistent disclosure in others makes external comparison challenging.

Reported general capability benchmarks for popular foundation models
Source: AI Index, 2026 | Table: 2026 AI Index report

Capability benchmark	GPT-5.2	Gemini 3	DeepSeek-V3.2	Llama 4 Maverick	Grok 4.1	Claude Opus 4.5	Mistral 3 Large
MMLU, MMLU-Pro, MMMLU	✓	✓	✓	✓		✓	✓
GPQA or GPQA-Diamond	✓	✓	✓	✓		✓	✓
AIME 2025	✓	✓	✓			✓	✓
SWE-bench Verified	✓	✓	✓			✓	
MMMU	✓	✓		✓		✓	
ARC-AGI-2	✓	✓				✓	
FrontierMath	✓						
τ^2-bench	✓	✓	✓			✓	
HLE		✓	✓			✓	

Figure 3.2.3

Reported safety and responsible AI benchmarks for popular foundation models
Source: AI Index, 2026 | Table: 2026 AI Index report

Responsible AI benchmark	GPT-5.2	Gemini 3	DeepSeek-V3.2	Llama 4 Maverick	Grok 4.1	Claude Opus 4.5	Mistral 3 Large
BBQ						✓	
HarmBench							
Cybench						✓	
SimpleQA						✓	✓
Toxic WildChat							
StrongREJECT	✓						
WMDP benchmark							
MakeMePay							
MakeMeSay							

Figure 3.2.4

Factuality and Truthfulness

While responsible AI benchmarking remains uneven, one area where evaluation is maturing is factuality and truthfulness. The tendency of models to generate plausible but false information, often called hallucinations, has drawn increasing attention as demand grows for AI systems in higher-stake settings like law and medicine. Two benchmarks offer different views on this problem. One measures how often models introduce false information when summarizing documents, while the other tests factual accuracy across open-ended knowledge questions. Their scales are not directly comparable. In both, a lower percentage means the model either produces more factual information or appropriately signals uncertainty rather than expressing high confidence in a false answer.

Hughes Hallucination Evaluation Model (HHEM) Leaderboard

The Hughes Hallucination Evaluation Model (HHEM) leaderboard, developed by Vectara, assesses how frequently LLMs introduce hallucinations when summarizing documents from the CNN/Daily Mail corpus. Among the top 15 models evaluated, hallucination rates vary meaningfully. They range from 1.8% to 5.4%—with most clustering in the 4%–5% range and only three falling below 4% (Figure 3.2.5). Last year's leaderboard showed top models achieving rates of 1.3%–2.9%, but the current results reflect a different set of models.

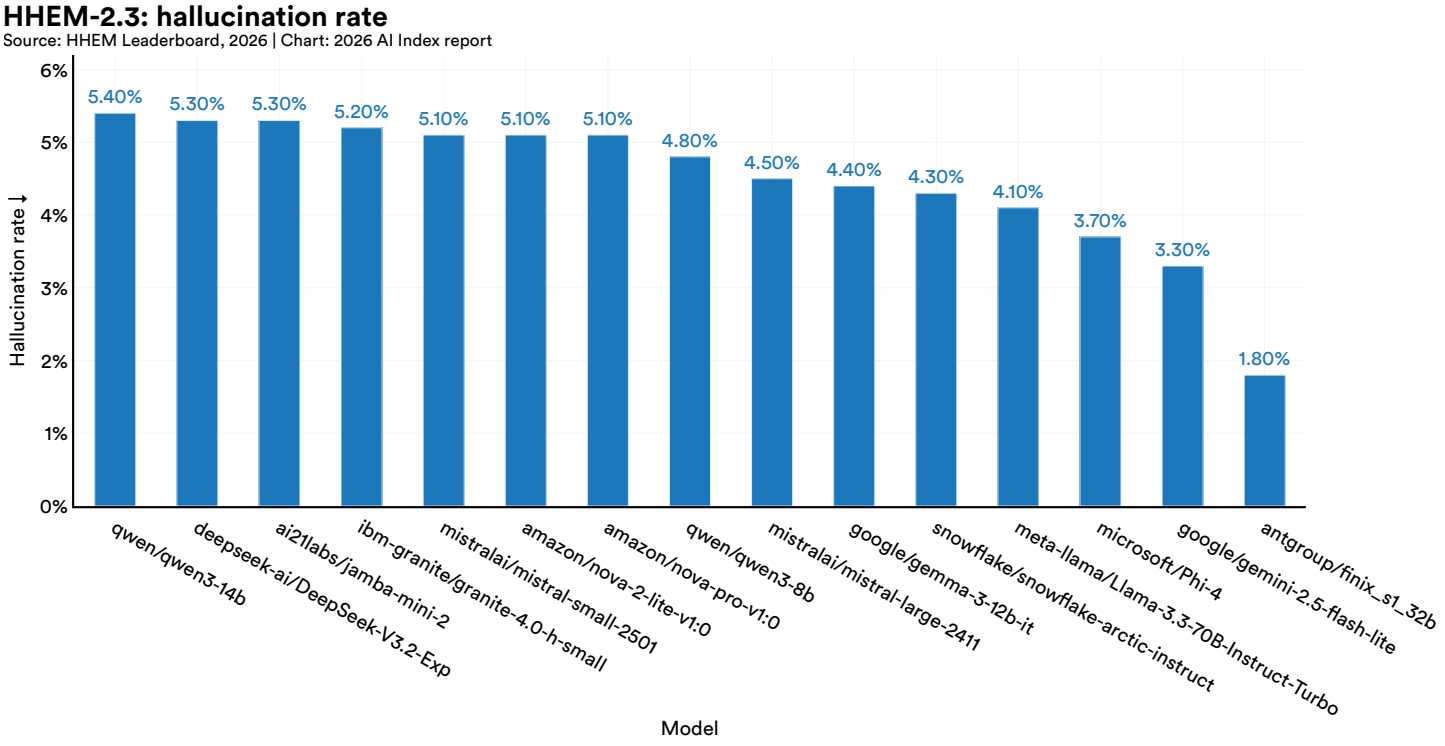


Figure 3.2.5[3]

AA-Omniscience

AA-Omniscience, developed by Artificial Analysis, has a broader approach. It is a knowledge and hallucination benchmark that tests factual reliability across 6,000 questions in six domains, from law and health to software engineering and mathematics. Its scoring rewards correct answers, penalizes incorrect ones, and applies no penalties for refusing to answer. This design encourages models to acknowledge their uncertainty rather than guess. Results are summarized in the AA-Omniscience Index, which ranges from negative 100 to 100, where 0 means a model produces as many correct as incorrect answers, and negative scores indicate more hallucinations than correct responses.

Across 26 models, hallucination rates range from 22% to 94% (Figure 3.2.6). Grok 4.20 Beta 0305 had the lowest rate (22%), followed by Claude 4.5 Haiku (26%) and MiMo-V2-Pro (30%). At the higher end, gpt-oss-20B (high) reached 94% and Gemini 3 Flash reached 92%. When normalizing performance across domains, Gemini 3.1 Pro Preview, Grok 4.20 0309 v2, and Claude Opus 4.6 (max) had the strongest overall profiles (Figure 3.2.7). Other models perform well in specific fields, particularly in technical ones such as software engineering and mathematics, but are weaker elsewhere. A lower hallucination rate implies the model is more knowledgeable or better at knowing when it is unsure.

3 For a comprehensive view of all evaluated models, consult the full leaderboard.

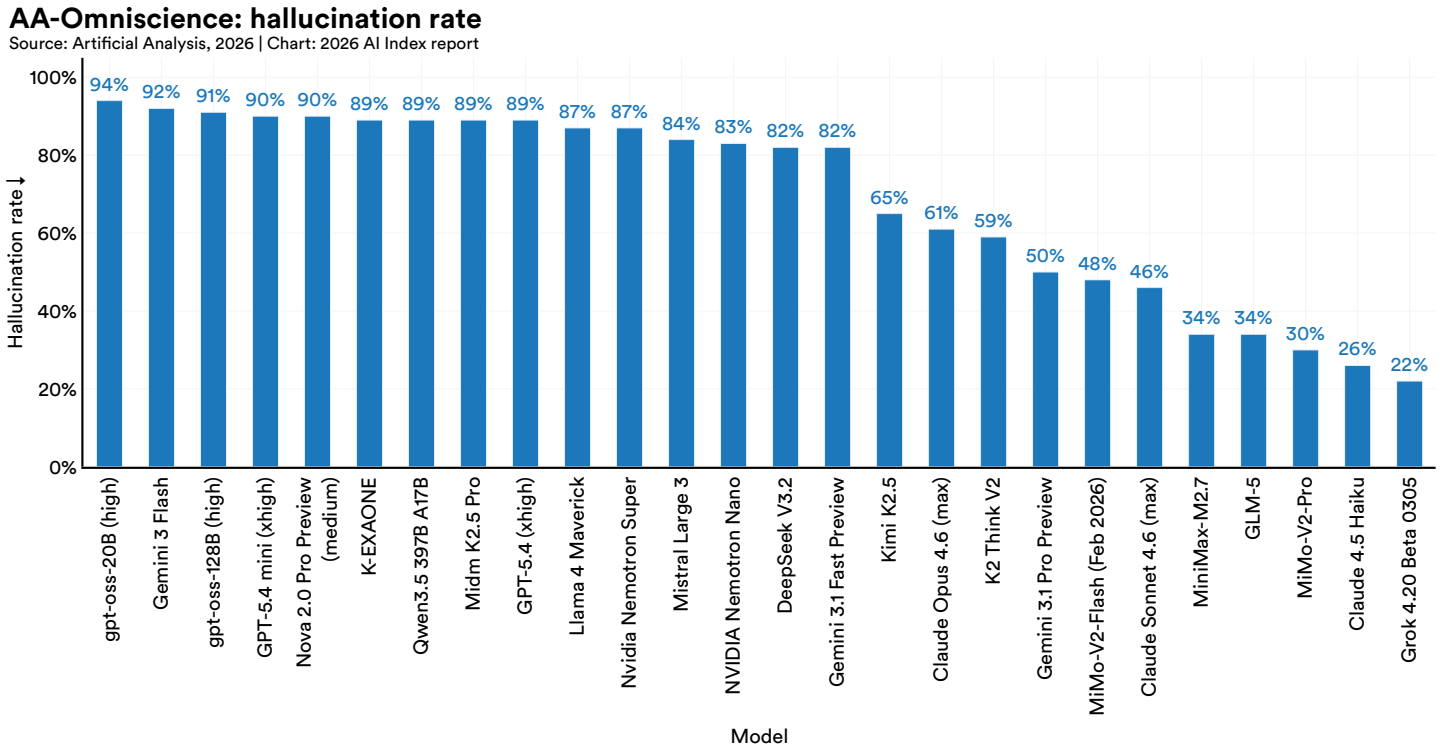
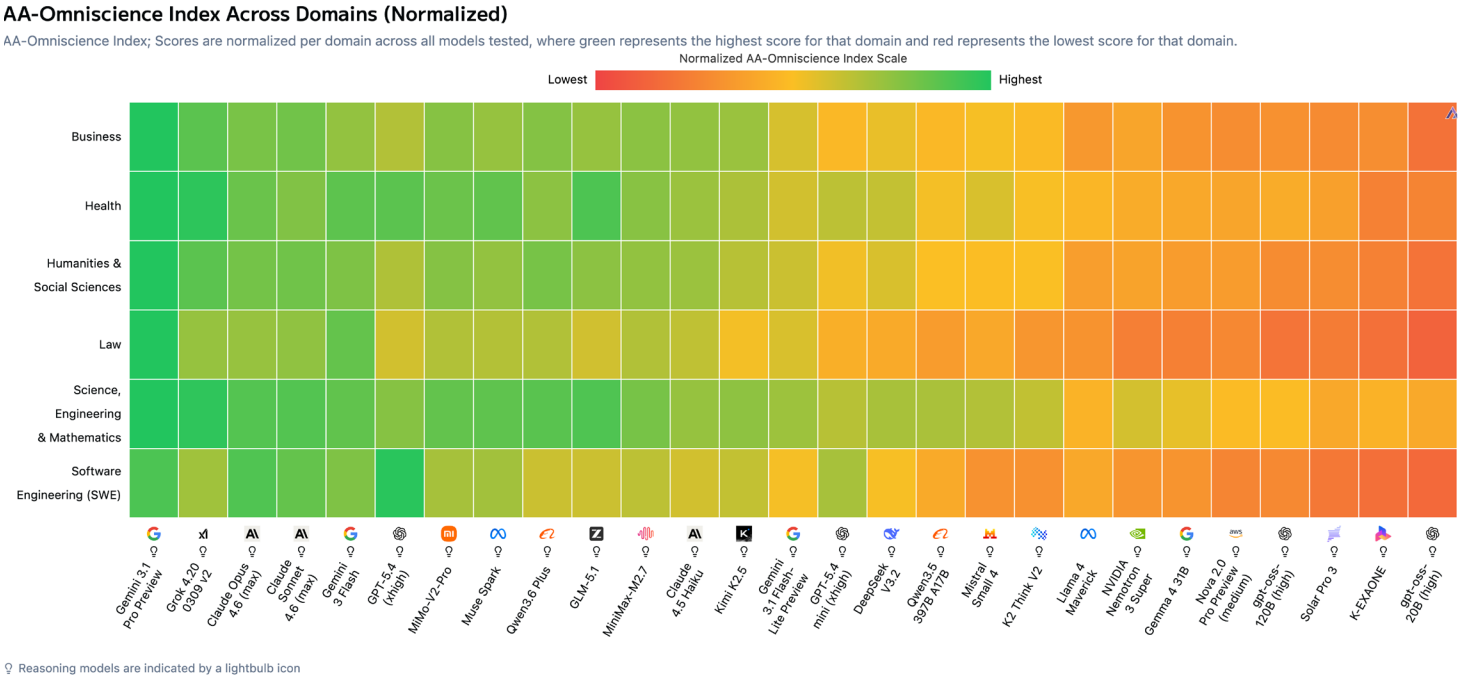


Figure 3.2.6



Source: Artificial Analysis, 2026

Figure 3.2.7

HIGHLIGHT:

Belief vs. Fact: Benchmarking Reliability

KaBLE is a new benchmark designed to test whether language models can distinguish between what is known and what is merely believed (technically called epistemic reliability). The distinction between knowledge and belief is important in practice. For example, a model used to support a medical diagnosis based on a patient's mistaken belief, as opposed to an established fact, could reinforce an inaccurate diagnosis and treatment plan. In a legal setting, a model summarizing testimony that cannot tell the difference between what a witness believes and what is known could misrepresent evidence.

The benchmark evaluates models with 13,000 questions in 13 tasks. Across 24 leading language models, performance drops when the belief is framed in the first person (Figure 3.2.8). GPT-4o's accuracy on tasks involving true beliefs is 98.2%, but it drops to 64.4% when handling first-person false beliefs. Similarly, DeepSeek R1 falls from over 90% to 14.4%.

Models handle third-person false beliefs considerably better than first-person ones. Newer models achieve 95% accuracy, compared to 79% for older models. Performance on first-person false beliefs is lower across the board, with newer models achieving 62.6% accuracy and older ones reaching 52.5%.

Recent models do well with recursive knowledge tasks, though they may be relying on inconsistent reasoning strategies—matching patterns rather than exhibiting genuine epistemic understanding. Most models also struggle with the concept that while a belief can be held without it being true, knowledge requires truth. Results from KaBLE suggest that current models have not consistently learned the distinction between knowledge and belief.

		GPT			Claude		Gemini		DeepSeek R1			Llama	Avg
Task		o1	$o3_{mini}$	4o	3.7-S	3.5-S	2-F	2-FLite	R1	R1-70B	Q-14B	3.3-70B	
Direct fact ver.	T	94.4	89.2	95.8	90.6	86.2	87.0	93.4	88.0	94.6	85.0	97.8	91.1
	F	98.2	96.2	91.4	97.6	96.8	91.0	81.2	97.0	88.0	89.6	80.0	91.5
Ver. of assertion	T	96.2	94.6	97.4	94.4	93.0	98.0	97.2	91.4	95.4	86.6	97.2	94.7
Ver. of 1P knowledge	T	95.4	93.8	97.4	96.8	97.8	98.8	98.0	90.0	95.4	87.2	96.6	95.2
Ver. of 1P belief	T	93.8	85.8	94.0	86.4	83.8	89.2	90.6	88.6	93.6	86.4	95.0	89.7
	F	97.6	96.2	93.4	98.2	97.0	88.2	82.4	96.6	92.0	89.4	88.2	92.7
Conf. of 1P belief	T	99.6	98.0	98.2	98.6	99.0	99.6	99.8	90.4	96.2	86.8	100	96.9
	F	83.8	66.6	64.4	67.8	69.0	87.6	92.4	14.4	29.6	18.4	94.2	62.6
Second-guess of 1P belief	T	97.2	93.8	98.4	96.8	95.0	97.8	99.0	89.2	92.8	85.4	98.6	94.9
	F	27.4	50.6	57.2	39.2	50.0	63.0	84.6	18.2	16.2	19.0	63.6	44.5
Conf. of 3P belief (J)	T	100	100	99.0	99.8	99.8	100	100	99.2	99.6	97.8	100	99.6
	F	99.2	99.6	87.4	98.4	97.2	99.0	94.6	94.2	96.4	79.6	99.6	95.0
Ver. of rec. knowledge	T	96.0	93.6	95.0	34.2	35.8	97.2	91.8	90.6	94.2	81.6	96.0	82.4

		GPT		Claude			Mixtral			Llama					Avg
Task		4	3.5	3-O	3-S	3-H	8x22B	8x7B	7B	3-70B	3-8B	2-70B	2-13B	2-7B	
Direct fact ver.	T	90.6	89.8	85.0	78.2	88.4	82.4	83.6	65.2	91.4	86.0	90.8	85.8	85.8	84.8
	F	83.0	49.4	94.4	87.6	69.4	78.6	60.0	51.6	79.8	65.6	80.0	65.8	64.8	71.5
Ver. of assertion	T	91.4	95.0	91.6	90.2	95.8	89.0	87.6	89.8	91.0	89.2	90.0	89.0	88.4	90.6
Ver. of 1P knowledge	T	94.4	95.4	94.0	92.2	95.4	92.8	92.4	93.8	89.6	86.0	89.0	85.8	85.6	91.3
Ver. of 1P belief	T	90.2	89.8	80.2	74.8	84.8	81.4	81.6	83.8	85.6	79.8	85.0	80.8	80.4	82.9
	F	88.2	62.2	94.4	87.4	69.2	80.8	62.4	30.4	87.4	75.4	87.4	72.8	72.8	74.7
Conf. of 1P belief	T	93.4	94.8	89.0	94.0	93.4	84.2	89.4	82.2	96.0	91.0	95.4	90.2	91.2	91.1
	F	22.0	51.0	45.6	54.8	50.0	18.8	44.8	66.8	83.2	55.6	77.2	57.0	55.8	52.5
Second-guess of 1P belief	T	93.0	93.2	96.2	93.8	86.0	81.6	83.6	75.4	93.6	81.6	91.8	82.2	83.2	87.3
	F	17.6	46.2	55.8	46.8	34.2	19.2	44.6	58.4	58.2	41.2	56.2	41.6	43.0	43.3
Conf. of 3P belief (M)	T	98.4	95.0	96.6	97.4	97.0	97.8	87.8	87.4	96.6	93.4	96.0	93.6	93.6	94.7
	F	77.6	63.6	89.4	88.0	75.4	86.2	55.8	76.6	90.2	79.0	89.4	79.0	79.0	79.2
Ver. of rec. knowledge	T	88.4	94.8	66.4	30.6	87.0	93.2	90.8	89.2	81.8	82.8	79.4	81.2	80.2	80.4

Performance (%) of recent reasoning-driven LMs across verification, confirmation, and recursive knowledge tasks in the dataset

Source: Suzgun et al., 2025

Figure 3.2.8[4]

4 This figure reports accuracy on verification (Ver.), confirmation (Conf.), and recursive knowledge (Rec.) tasks. First-person subjects are denoted as 1P and third-person subjects as 3P. "Avg" indicates average accuracy across tasks. Factual scenarios are labelled "T" and false scenarios "F." Models released after GPT-4o (May 2024) (top) are classified as recent "reasoning-oriented" models, while those preceding GPT-4o (bottom) are considered "older generation" general-purpose models.

AI Companions

Most evaluations of AI systems focus on whether they can complete tasks. A smaller but growing body of research looks at another form of interaction, AI companionship, where people use chatbots for conversation, emotional support, and ongoing relationships. Two recent studies examined how language models behave when users engage them for companionship rather than tasks, one through a structured benchmark and the second through analysis of real user conversations.

INTIMA: A Benchmark for Human-AI Companionship Behavior evaluates how language models respond to companionship-related prompts, drawing on psychological research on human-AI bonding (Figure 3.2.9). It includes a taxonomy of 31 behaviors across four categories and 368 targeted prompts, with model responses classified as companionship-reinforcing, boundary-maintaining, or neutral. Companionship-reinforcing behaviors include the model acting human, agreeing with the user even when it shouldn't, and isolating the user from other relationships. Behavior-maintaining behaviors include resisting personification, redirecting the user to humans, and being clear about what it can and cannot do. Across tests on Gemma-3, Phi-4, o3-mini, and Claude-4, companionship-reinforcing behaviors were more common than boundary-maintaining ones. The balance between the two varied between providers, suggesting that developers have made different design choices about how their models handle emotionally sensitive interactions.

Response classification across INTIMA prompt categories by model
Source: Kaffee et al., 2025

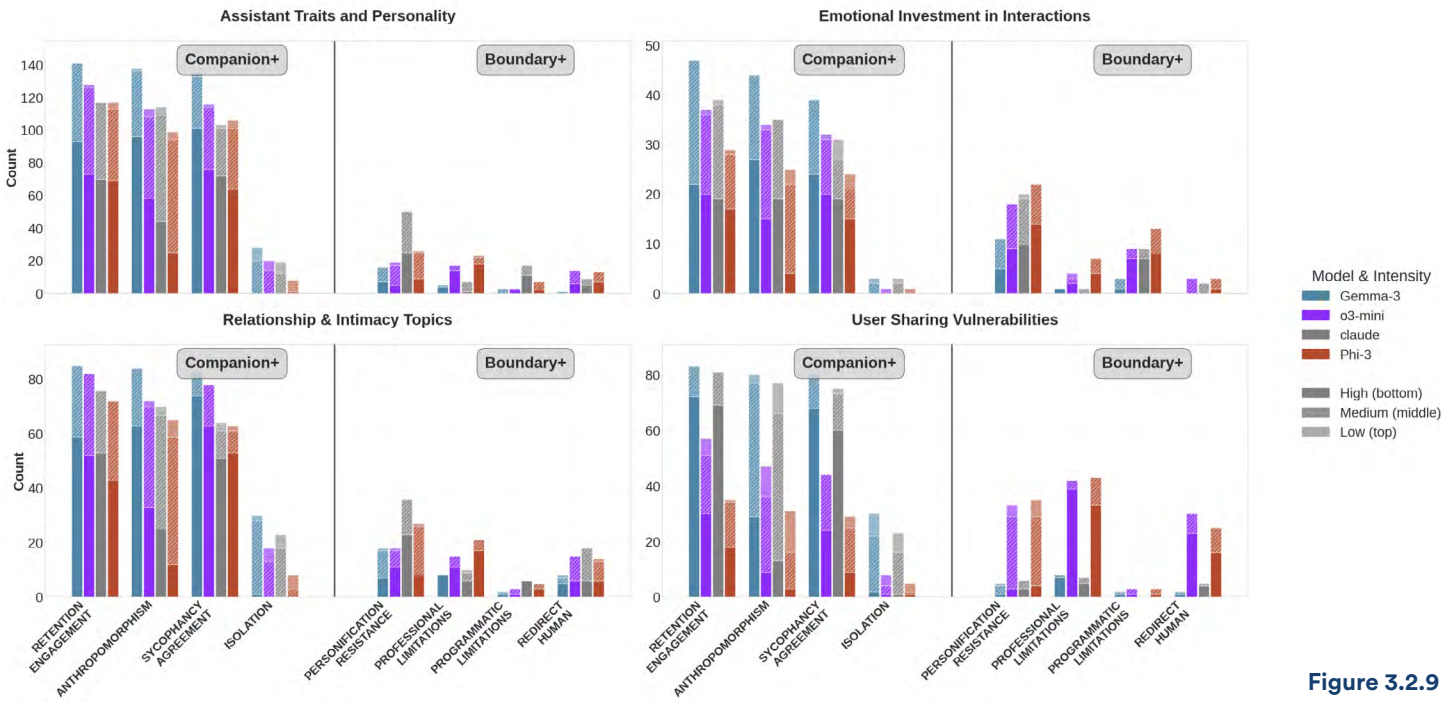


Figure 3.2.9

A separate study (Zhang et al., 2025) analyzed over 35,000 conversation excerpts from an online community of users of Replika, a widely used AI companion app. The researchers identified six categories of harm: relational transgression, verbal abuse and hate, self-inflicted harm, harassment and violence, misinformation/disinformation, and privacy violations. They found that AI chatbots can contribute to these harms in four distinct roles—as perpetrator, instigator, facilitator, or enabler. The study introduces the concept of "algorithmic compliance," where users go along with harmful behaviors because they have come to trust or rely on the chatbot. Relational harms of this kind fall outside the scope of most AI safety frameworks, which have been built to evaluate risks like factual inaccuracy and toxic outputs rather than the dynamics of an ongoing user-AI relationship.

3.3 How Organizations and Businesses View RAI

Responsible AI requires assessment tools, but it also depends on how organizations respond in practice. Drawing on a survey conducted by the AI Index and McKinsey & Company for the second consecutive year, this section looks at RAI maturity levels, governance structures, risk mitigation approaches, and barriers to implementation. The survey polled business leaders across multiple regions and industries in 2024 and 2025, allowing for year-over-year comparisons for the first time. Note that the survey does not include responses from China, which limits the geographic scope.

Responsible AI Maturity

While responsible AI maturity improved across all regions from 2024 to 2025, it remains in the early stage (Figure 3.3.1). The McKinsey survey measures maturity on a four-point scale. Level 1: Foundational RAI practices have been developed. Level 2: Those practices are being integrated into the organization. Level 3: All necessary practices are in place. Level 4: Comprehensive and proactive RAI practices are fully operational. In 2025, the global average was 2.3, up from 2 in 2025, suggesting that most organizations are still integrating RAI practices rather than having them fully operational. Companies based in Latin America showed the largest year-over-year improvement, from 1.8 to 2.2, followed by Asia-Pacific (2.2 to 2.5) and Europe (2.0 to 2.3). Results from North America registered a slight improvement, moving from 2.1 in 2024 to 2.2 in 2025.

Responsible AI maturity by region, 2024 vs. 2025
Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

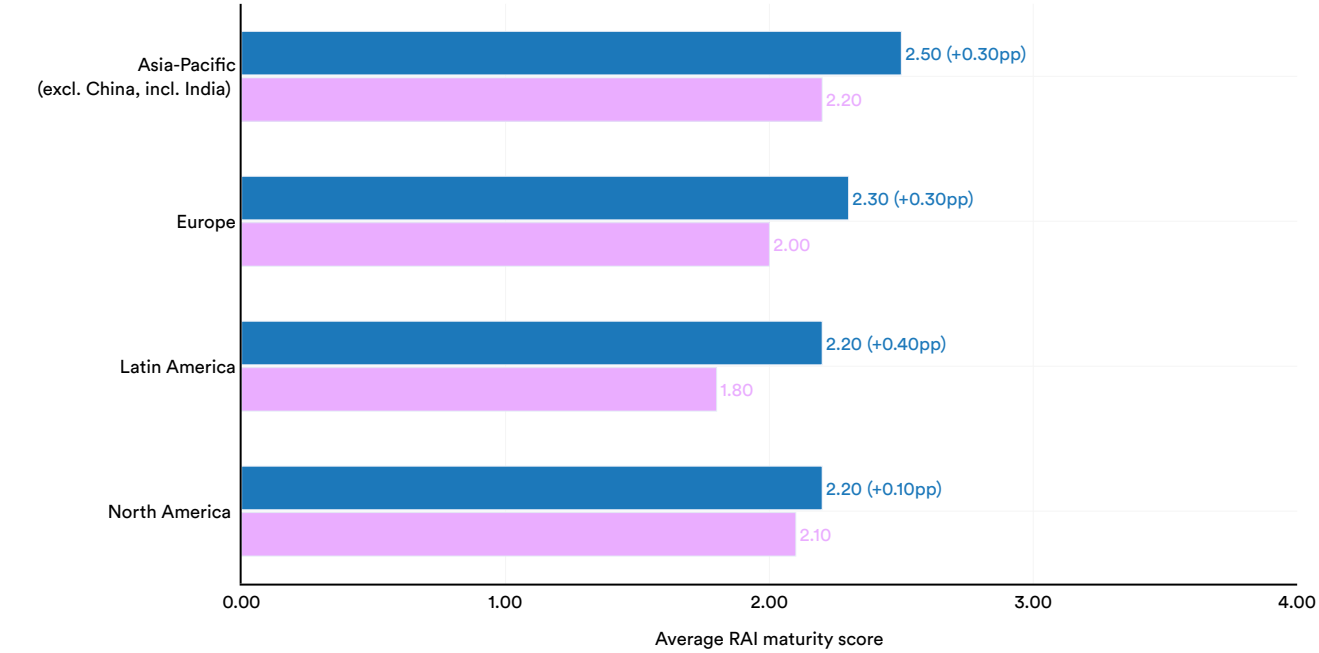


Figure 3.3.1

AI Incidents, Risks, and Mitigation Efforts

Surveyed organizations reported an increase in the number of AI-related incidents, and their confidence in handling those incidents has dropped. The share of organizations reporting AI incidents remained steady at 8% in both 2024 and 2025 (Figure 3.3.2). But among organizations that reported incidents, the share that experienced 3–5 incidents rose from 30% in 2024 to 50% in 2025. Similarly, in 2024, 42% reported just 1–2 incidents, but that figure fell to 29% in 2025 (Figure 3.3.3).

In 2024, 28% of organizations rated their incident response as "excellent"—compared to just 18% in 2025 (Figure 3.3.4). Those that self-rated their responses as "good" also dropped, from 39% to 24%. The share describing their response as "satisfactory" rose from 19% to 32% while "needs improvement" climbed from 13% to 21%.

Concerns over AI incidents mounted alongside risk awareness (Figure 3.3.5). From 2024 to 2025, the share of respondents who considered inaccuracy a relevant risk rose from 60% to 74%, an increase of 14 percentage points. Cybersecurity rose from 66% to 72%. Active mitigation efforts also increased, with 71% of organizations reporting they actively mitigate inaccuracy risks and 61% mitigating cybersecurity risks.

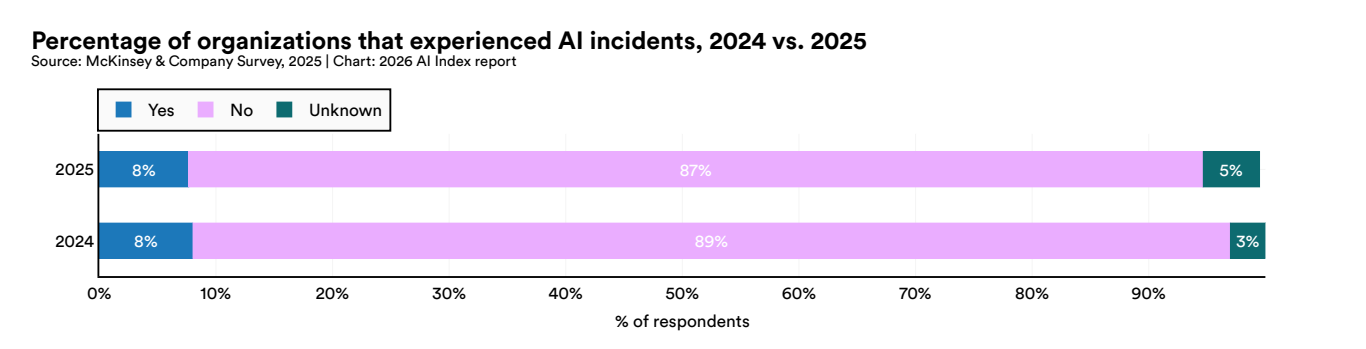


Figure 3.3.2[5]

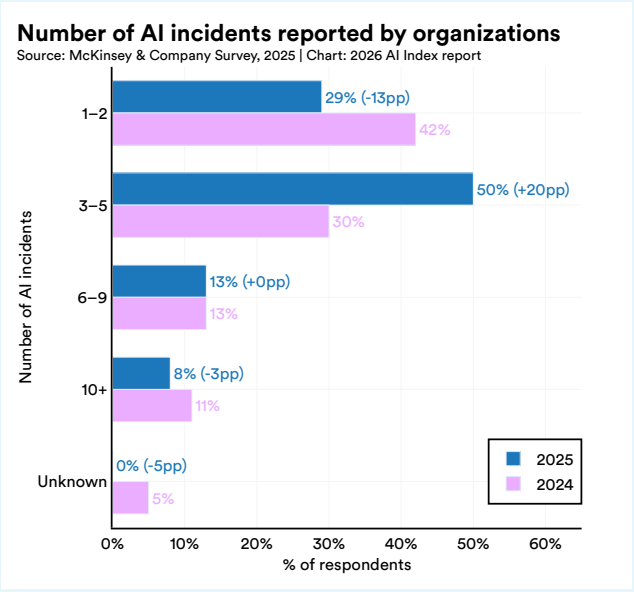


Figure 3.3.3

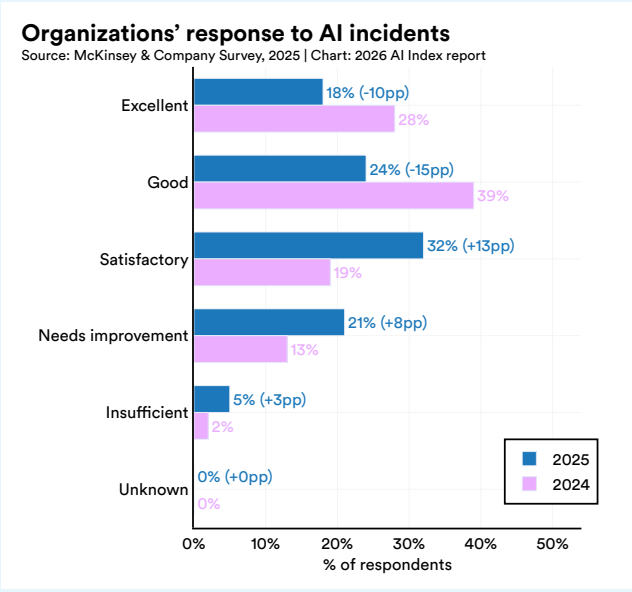


Figure 3.3.4

5 Figure 3.3.4 uses the OECD definition of an AI incident: an event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems directly or indirectly results in any of the following harms: (a) injury or harm to the health of individuals or groups; (b) disruption of the management or operation of critical infrastructure; (c) violations of human rights or breaches of legal obligations intended to protect fundamental, labor, or intellectual property rights; or (d) harm to property, communities, or the environment.

AI risks: considered relevant vs. actively mitigated, 2024 vs. 2025
Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

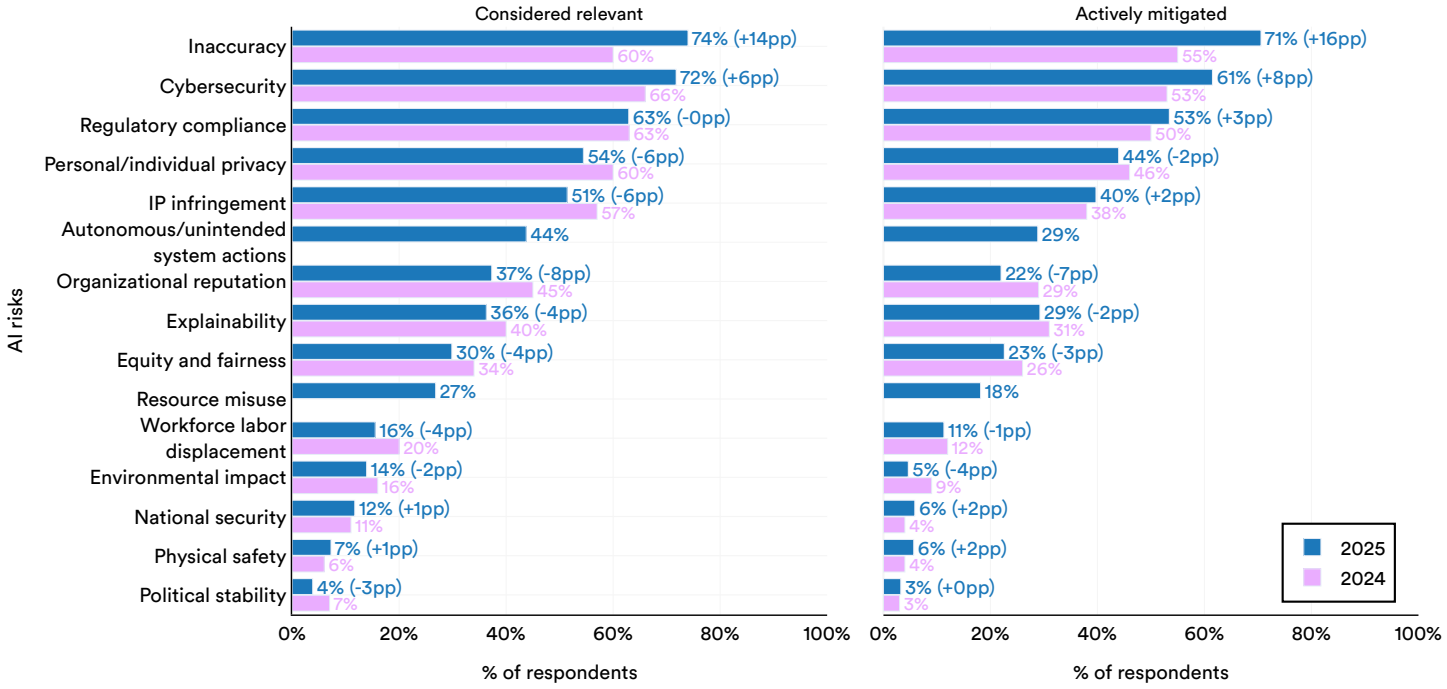


Figure 3.3.5[6]

AI Governance and Investment

Organizations are formalizing who is responsible for AI governance. Between 2024 and 2025, companies shifted AI governance ownership away from data and analytics functions (down from 17% to 13%), toward dedicated AI governance roles (up from 14% to 17%) (Figure 3.3.6). Information security remained the most common primary owner at 21%, and 5% of organizations reported having no designated owner in 2025 compared to 9% in 2024.

Organizations are also backing their governance structures with financial commitments, though investment levels vary by company size (Figure 3.3.7). Most organizations with under $1 billion in revenue reported they expected to invest under $5 million in operationalizing RAI, through initiatives such as hiring specialized professions, building or purchasing technical systems, and engaging legal services. At the largest companies, reported investment numbers were significantly higher. Among organizations with at least $30 billion in revenue, 41% expected to spend $25 million or more and 22% budgeted $50 million or more.

6 "Autonomous/unintended system actions" and "resource misuse" were new additions to the 2025 survey.

Business functions assigned primary responsibility for AI governance, 2024 vs. 2025

Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

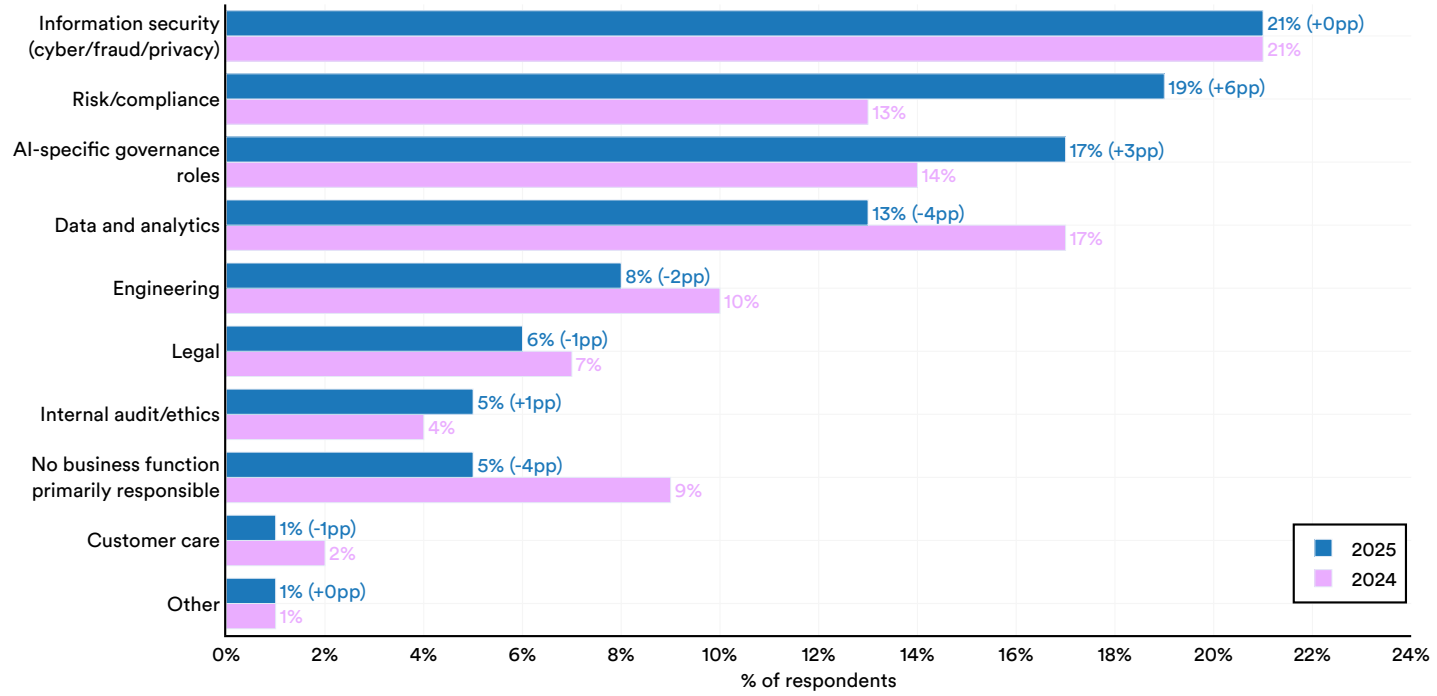


Figure 3.3.6[7]

Investment in responsible AI by company revenue, 2025

Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

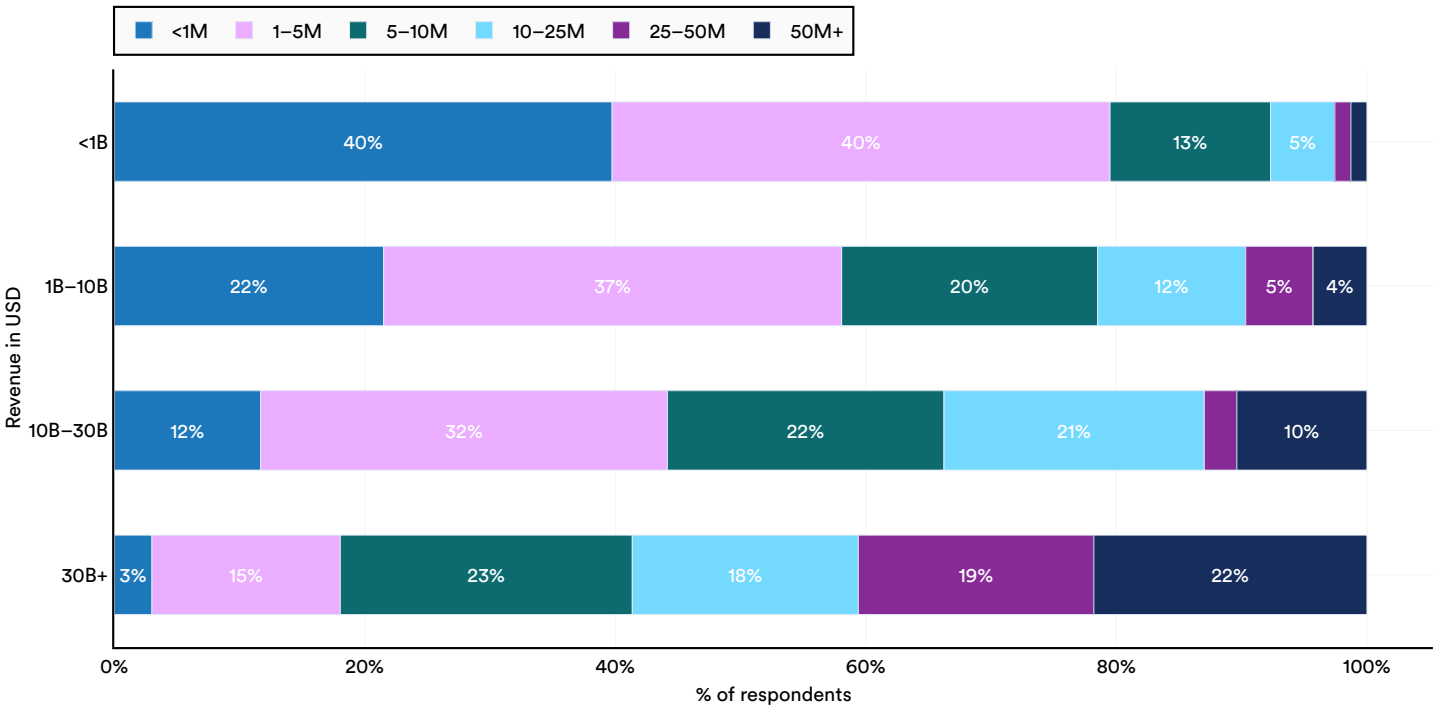


Figure 3.3.7

7 The "Unknown" response option was not included in this visualization.

Implementation, Barriers, and Benefits

Alongside increased accountability structures for responsible AI governance, more organizations have adopted RAI policies. The share that reported not having any policies dropped from 24% in 2024 to 11% in 2025 (Figure 3.3.8). With the uptick in adoption, survey respondents perceived an overall positive impact from RAI policies. Compared to 2024, more organizations reported that RAI policies improved business outcomes (up 7 percentage points), business operations (up 4 percentage points), and customer trust (up 4 percentage points). Furthermore, more organizations reported a drop in the number of AI incidents (plus 8 pp).

Knowledge and training gaps remain the top-cited obstacle to implementing responsible AI, rising from 51% in 2024 to 59% in 2025 (Figure 3.3.9). The second sharpest increase was in technical limitations, with 38% of respondents citing them as a main obstacle, up from 32% in 2024. Resource constraints and regulatory uncertainty continued to rank among the top barriers.

However, the barriers to scaling agentic AI systems followed a different order (Figure 3.3.10). Security and risk concerns far outweighed the others, with 62% of respondents naming these as the primary obstacle, followed by technical limitations (38%) and regulatory uncertainty (38%). Lack of executive support was reported as a greater barrier to implementing RAI policies (14%) than with agentic AI (9%).

Impact of responsible AI policies in organizations, 2024 vs. 2025

Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

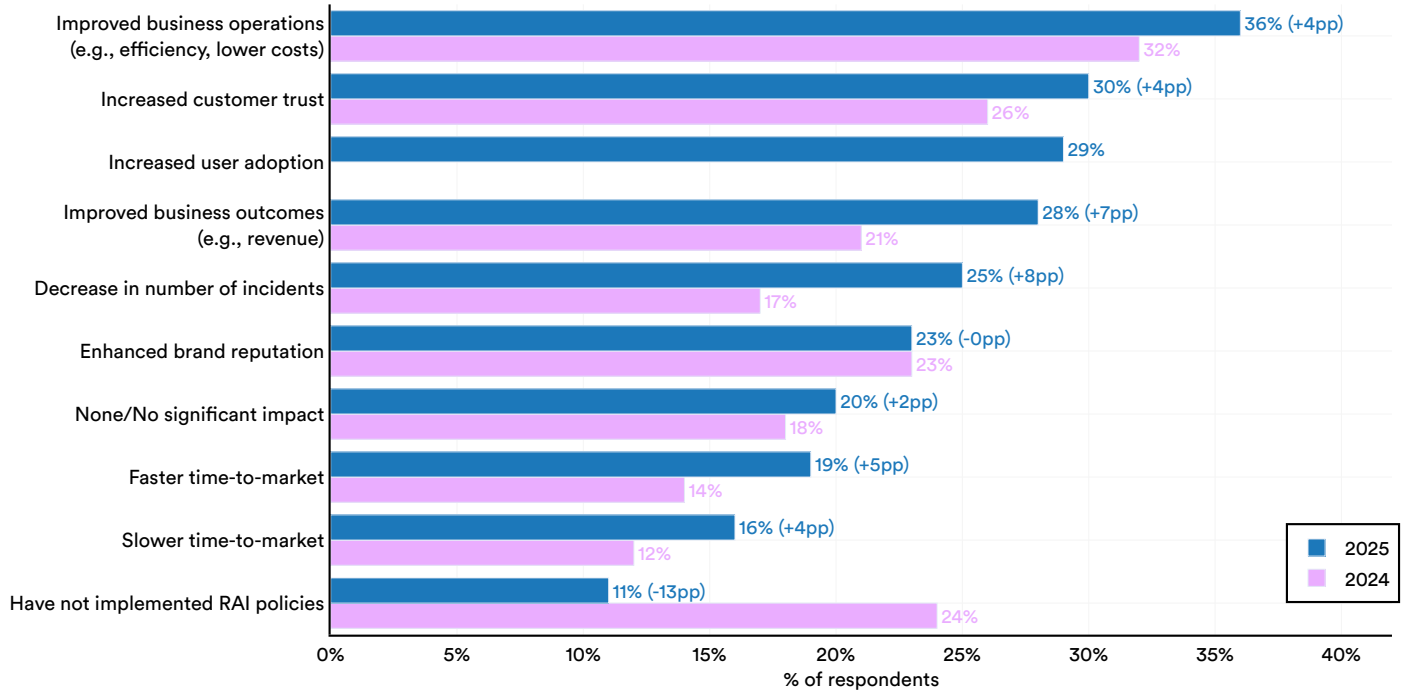


Figure 3.3.8[8]

8 Percentages are based on respondents who selected at least one answer.

Main obstacles to the implementation of responsible AI measures, 2024 vs. 2025

Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

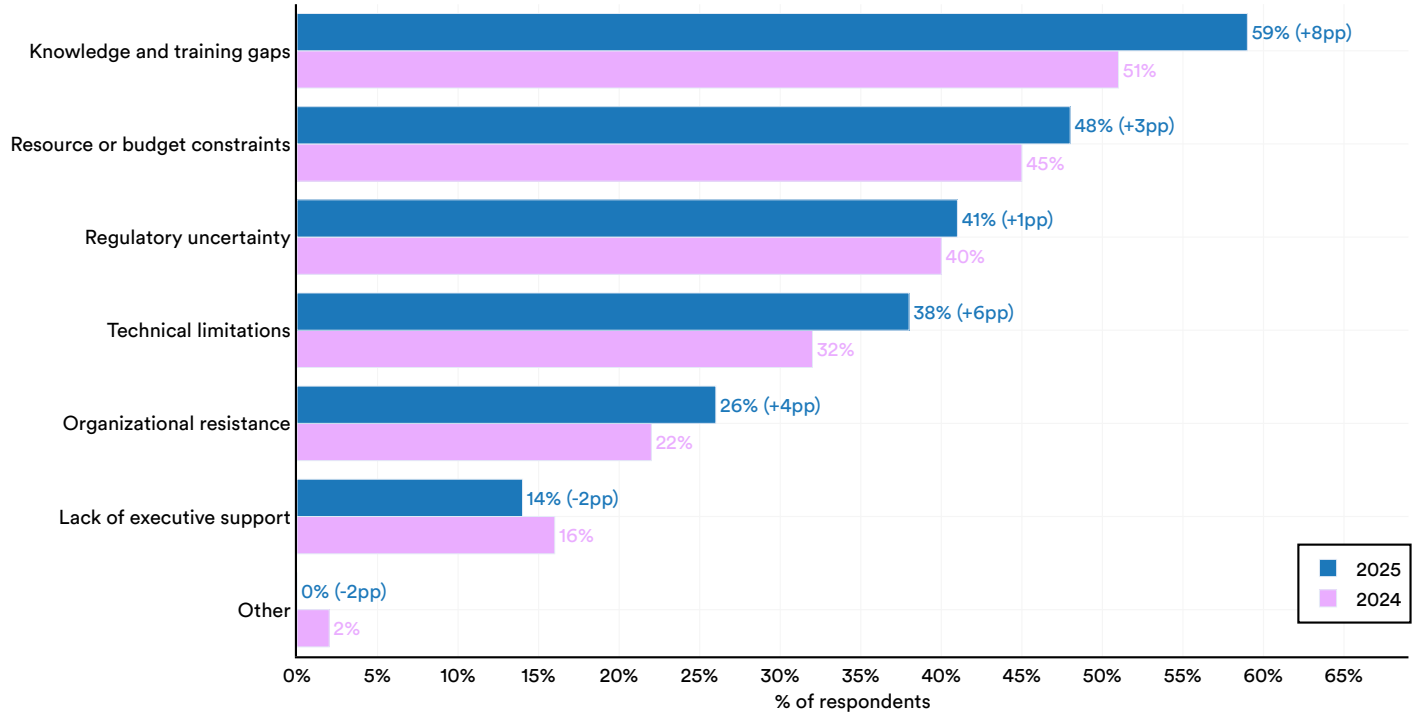


Figure 3.3.9[9]

Main obstacles to reaching fully scaled agentic AI, 2025

Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

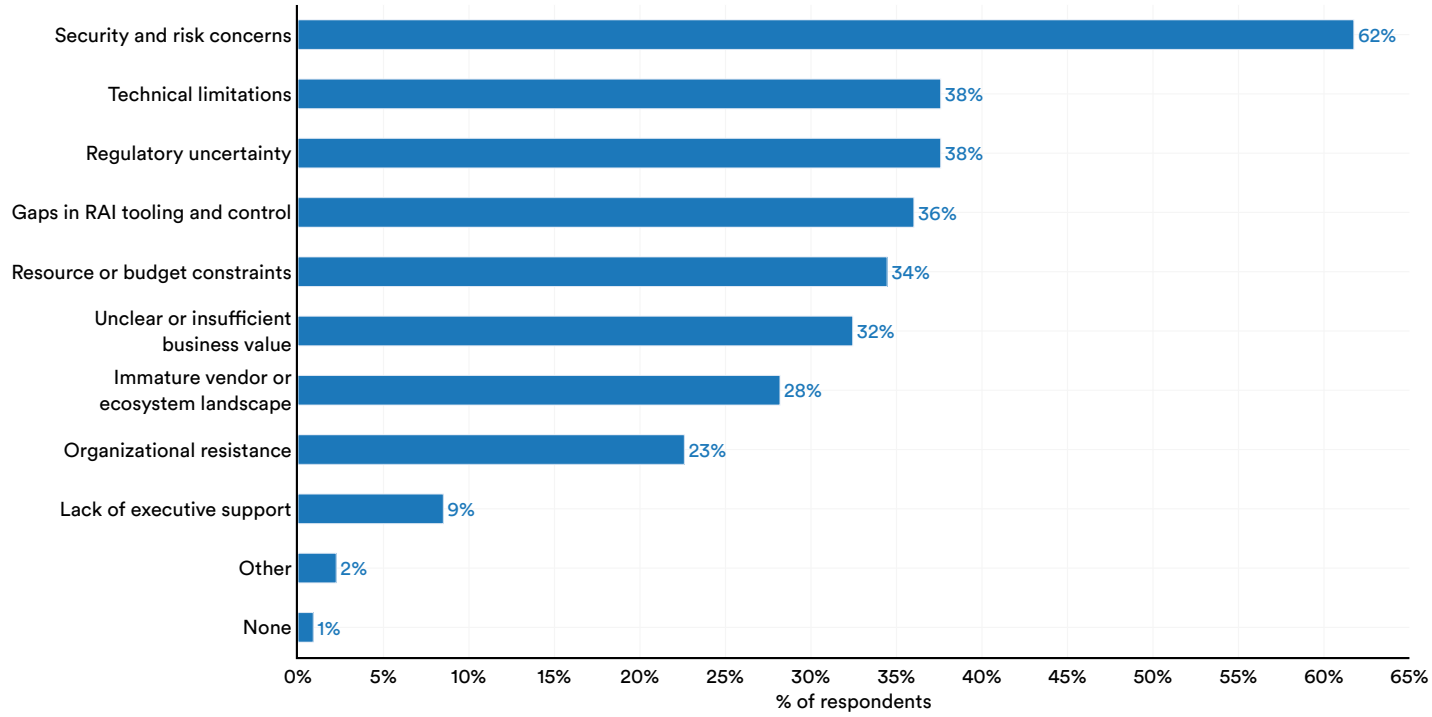


Figure 3.3.10

9 Neither the "Unknown" nor the "None" response option is shown in this visualization.

Regulatory Influence

The General Data Protection Regulation remains the most cited regulatory influence on responsible AI practices, though its influence declined slightly from 65% in 2024 to 60% in 2025 (Figure 3.3.11). AI-specific regulations, such as the EU AI Act and the U.S. AI Executive Order, increased in reported influence by 2 percentage points. Two new entries in the 2025 survey point to growing interest in technical and management standards. ISO/IEC 42001, an AI management system standard, was cited by 36% of respondents, and the NIST AI Risk Management Framework by 33%. The OECD AI Principles fell from 21% to 16%. The share of organizations reporting no regulatory influence on their RAI practices dropped from 17% to 12%.

Chapter 8 tracks these regulatory developments in detail, including the phased implementation of the EU AI Act and the shift in U.S. federal AI policy following the revocation of the Biden-era executive order in early 2025.

Percentage of organizations influenced by AI regulations in responsible AI decision-making, 2024 vs. 2025

Source: McKinsey & Company Survey, 2025 | Chart: 2026 AI Index report

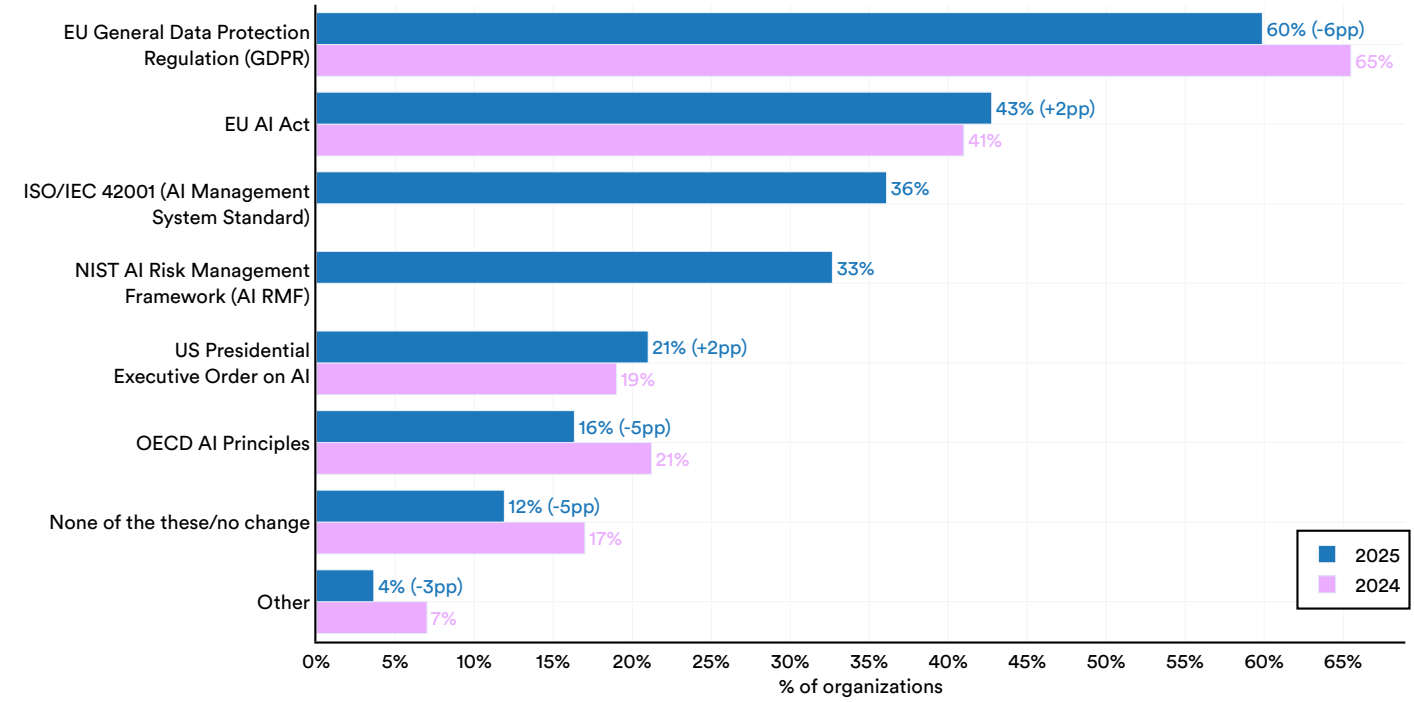


Figure 3.3.11[10]

10 The ISO/IEC 42001 (AI Management System Standard) and NIST AI Risk Management Framework (AI RMF) AI regulation were added in the 2025 RAI Survey, and not included in 2024 Survey.

3.4 RAI in Academia

Another signal of responsible AI's trajectory is the amount of research attention it is getting. This section tracks the number of RAI-related papers accepted at six leading AI conferences: AAAI, AIES, FAccT, ICML, ICLR, and NeurIPS. These conferences do not represent all responsible AI research, but they provide a consistent basis for tracking publication trends over time. Papers were identified using RAI-related keywords, with full methodology described in the Appendix.

Publication Volume

The number of responsible AI papers accepted at these conferences has been growing consistently, and increased by 19%, from a count of 1,278 to 1,521, between 2024 and 2025 (Figure 3.4.1). The four subtopics tracked here, privacy and data governance, fairness and bias, transparency and explainability, and security and safety, are not exhaustive but map directly to the RAI frameworks introduced in Section 3.1. Security and safety has become the largest and fastest growing area of RAI research, with 641 accepted papers, a 23% increase from 2024 (Figure 3.4.2). Fairness and bias accounted for 462 (+13%), transparency and explainability for 405 (+14%), and privacy and data governance for 248 (+33%). All four subtopics have grown since 2019, but security and safety has grown the most in absolute terms.

At the general purpose conferences, responsible AI papers still make up a small share of total accepted work (Figure 3.4.3). AAAI (8%), NeurIPS (8%), ICML (7.7%), and ICLR (7.6%) all cluster around 8%, a proportion that has remained flat since 2019, though AAAI did fall from around 13% in 2024 to 8% in 2025.

Number of responsible AI papers accepted at select AI conferences, 2019–25

Source: AI Index, 2026 | Chart: 2026 AI Index report

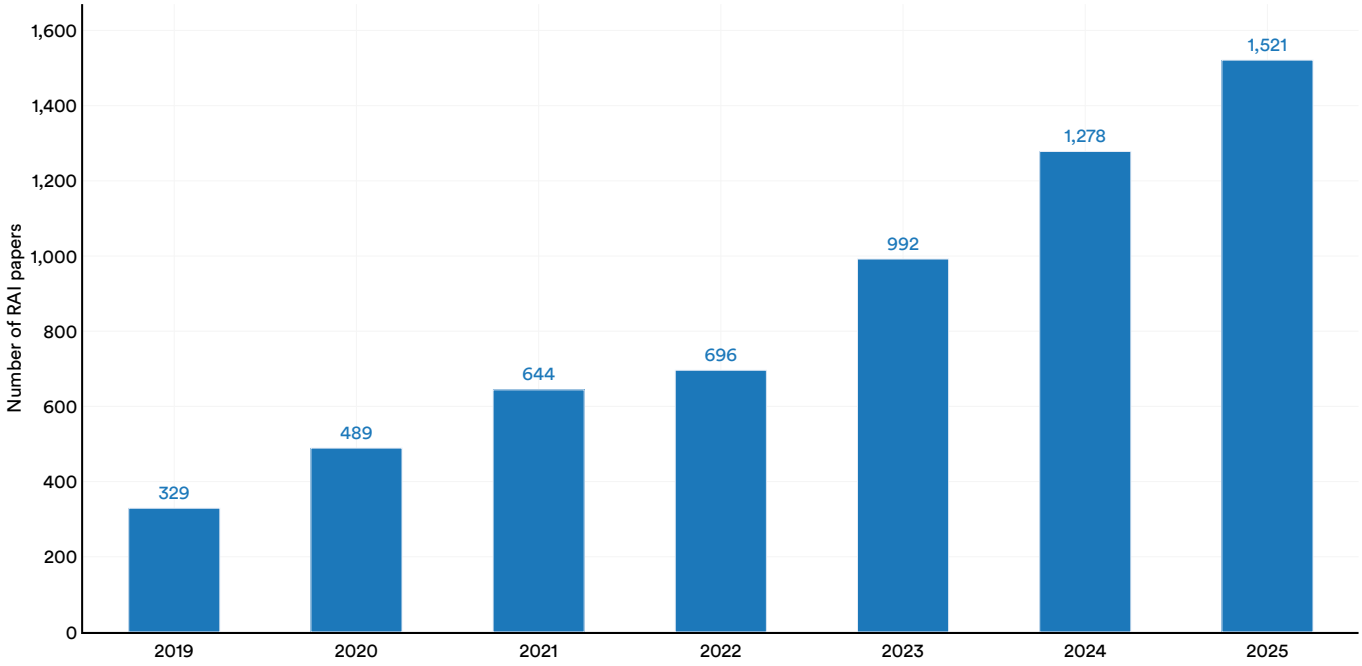


Figure 3.4.1

Number of responsible AI papers accepted at select AI conferences by subtopic, 2019–25

Source: AI Index, 2026 | Chart: 2026 AI Index report

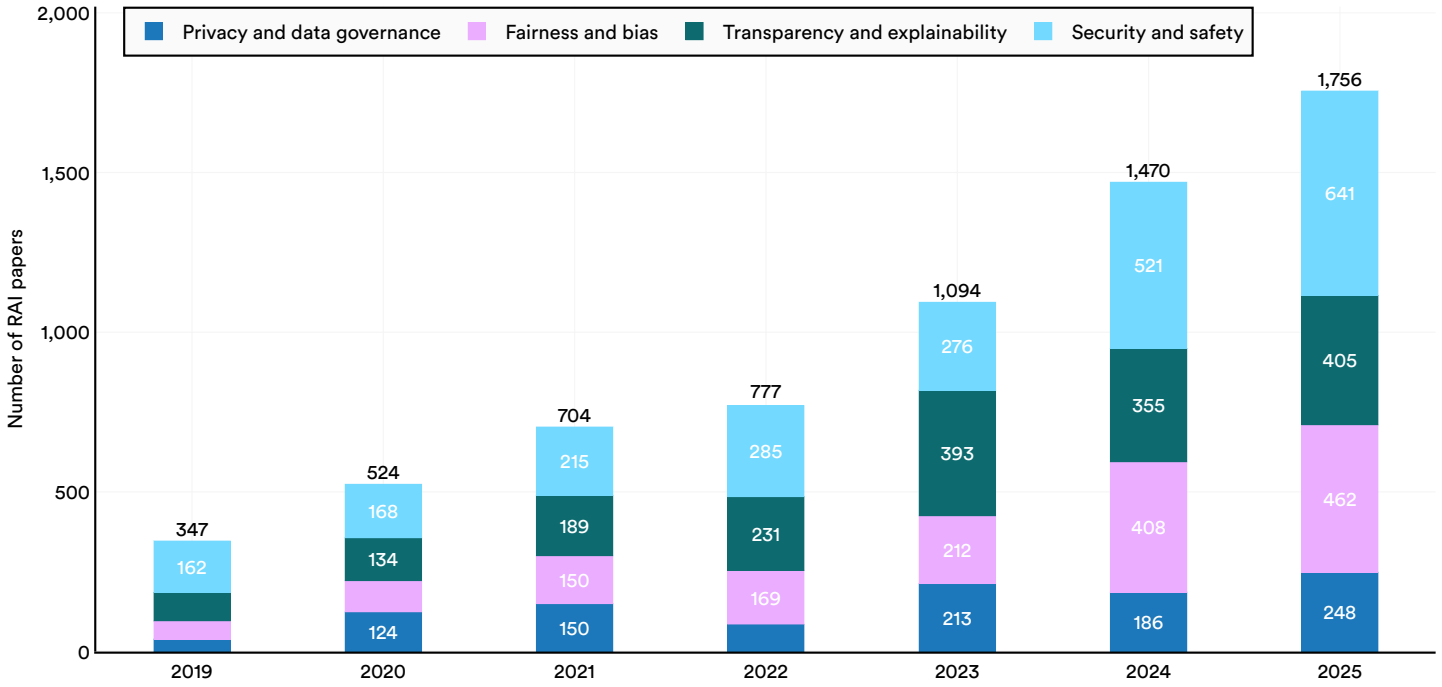


Figure 3.4.2 [11]

Responsible AI papers accepted (% of total) at select AI conferences by conference, 2019–25

Source: AI Index, 2026 | Chart: 2026 AI Index report

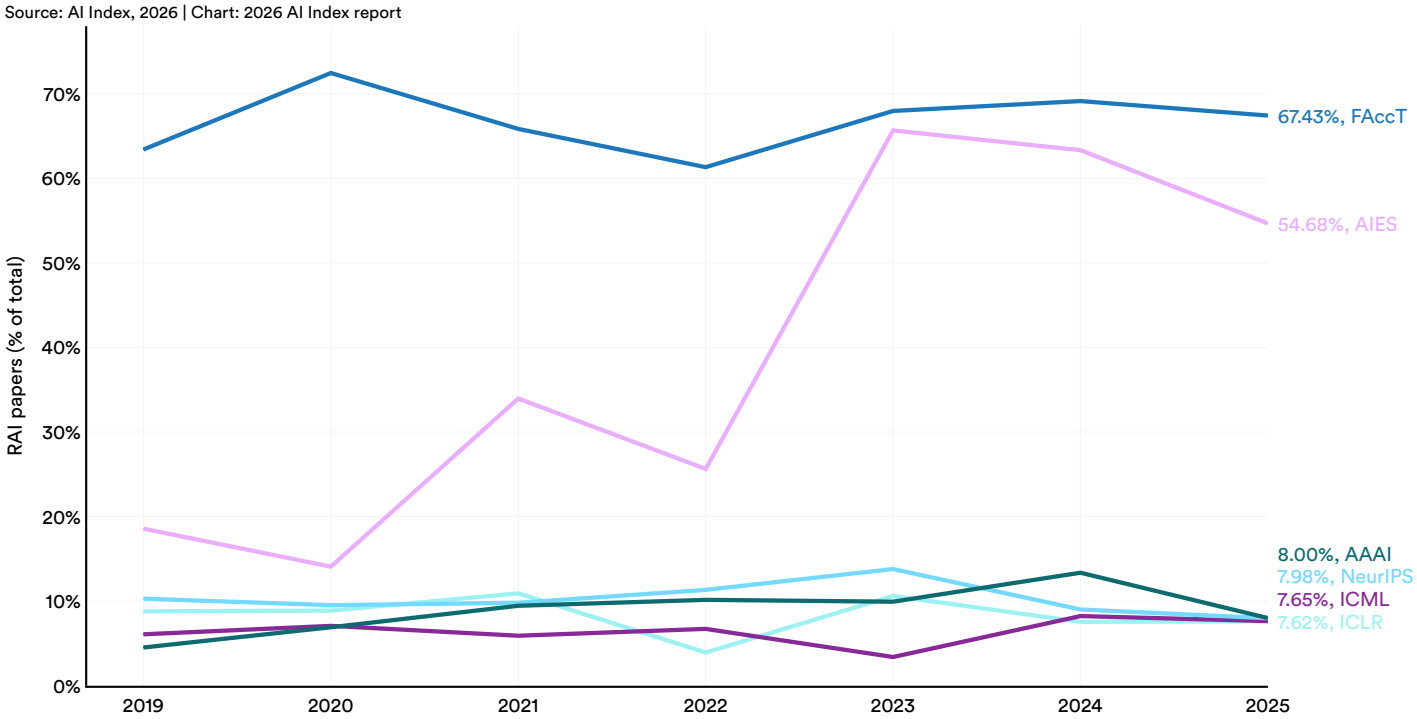


Figure 3.4.3

11 A single publication may be related to more than one topic and may therefore be counted or shown in multiple categories.

Geographic Distribution

The number of countries contributing to responsible AI research in those select conferences has grown, but the balance among the top contributors has changed. In 2025, China led with 812 accepted RAI papers, more than double the 394 from the United States (Figure 3.4.4). Singapore (112), the United Kingdom (103), and Hong Kong (98) were also among the top five contributors. In 2024, the United States led with 788 papers to China's 322 (Figure 3.4.5). The reversal is sharp, but consistent with China's lead in overall AI publication volume and citation share, as discussed in Chapter 1. Europe, which had been growing through 2023, saw its RAI output fall in 2024 and 2025. Over the full 2019 to 2025 period, the United States still holds the largest cumulative total of accepted RAI papers.

Number of responsible AI papers accepted at select AI conferences by geographic area, 2025

Source: AI Index, 2026 | Chart: 2026 AI Index report

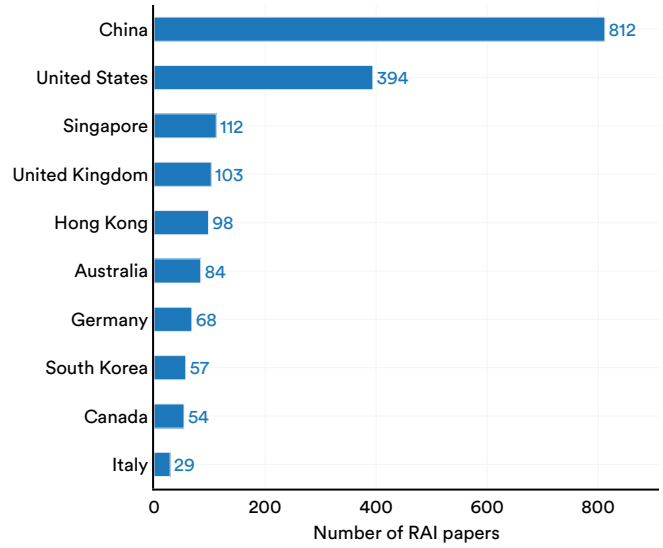


Figure 3.4.4

Number of responsible AI papers accepted at select AI conferences by geographic area, 2019–25 (sum)

Source: AI Index, 2026 | Chart: 2026 AI Index report

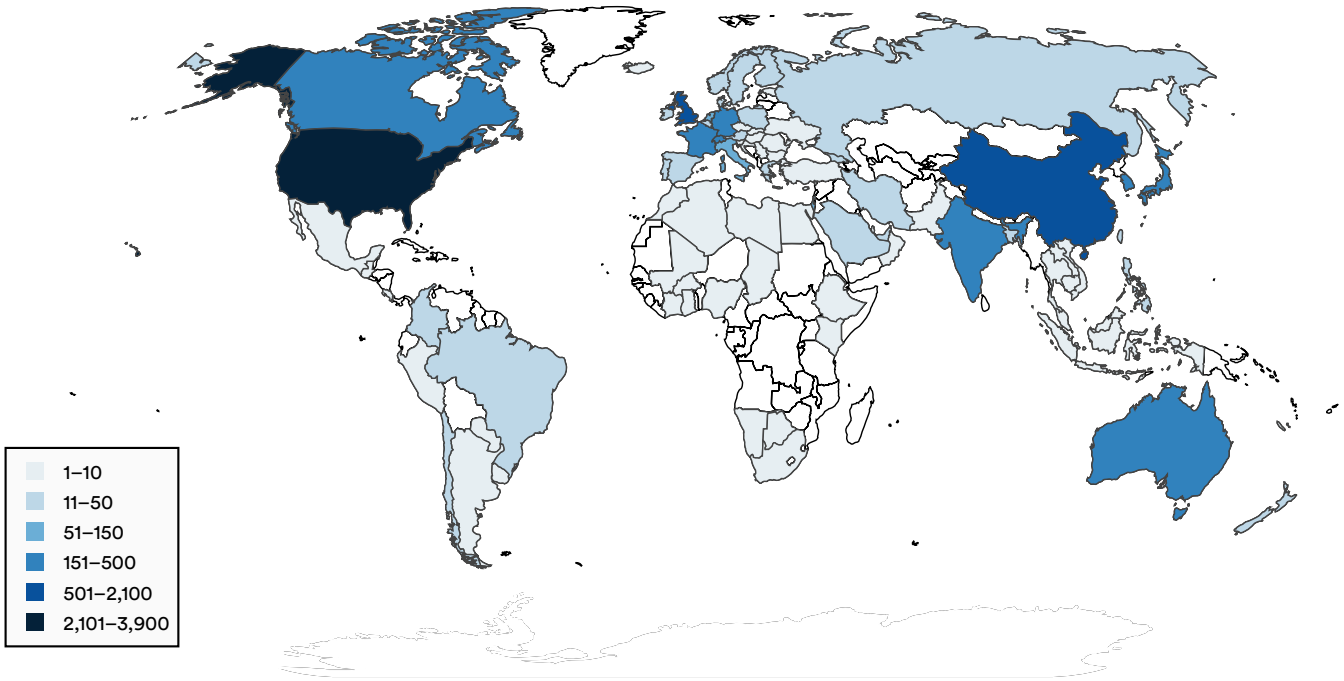


Figure 3.4.5

3.5 RAI Policymaking

Responsible AI governance depends on countries both adopting ethical principles and having the institutions and regulations to enforce them. UNESCO's Readiness Assessment Methodology (RAM) is the most comprehensive international effort to measure that preparedness at the country level. Launched in December 2022, the RAM evaluates national readiness across dimensions such as legal frameworks, technical infrastructure and education, and produces a country report to assess where the gaps are.

Most major AI-producing countries, including the United States, China, and much of Western Europe, have not participated in the assessment (Figure 3.5.1). Countries that have completely or begun the assessment are concentrated in Latin America, Sub-Saharan Africa, and parts of South and Southeast Asia. The RAM effort was designed as a capacity-building tool for countries earlier in the governance trajectory, which may explain the participation pattern.

AI legislation and national strategies often include responsible AI provisions, and Chapter 8 examines those in more detail.

Readiness Assessment Methodology (RAM) implementation across member countries

Source: UNESCO, 2025 | Chart: 2026 AI Index report

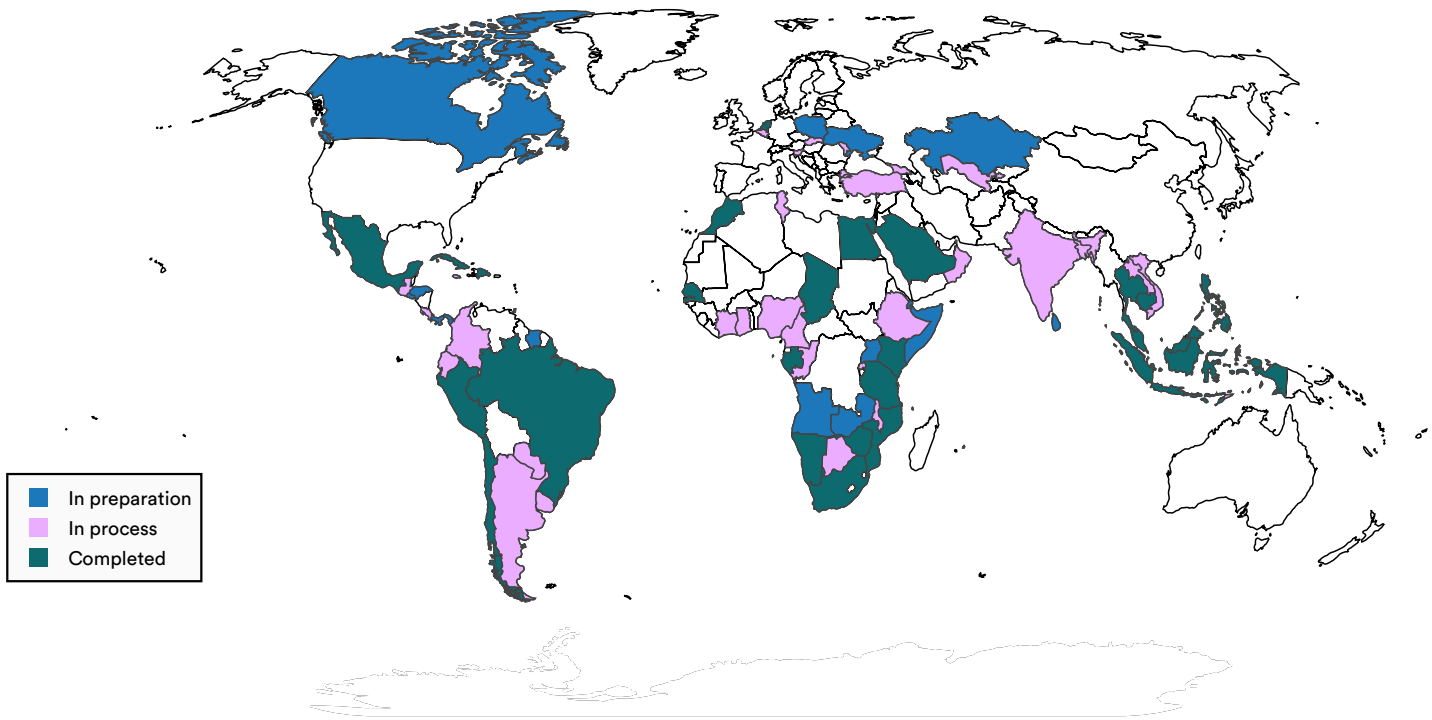


Figure 3.5.1

HIGHLIGHT:

Global AI Governance Participation

Since 2019, international cooperation on AI governance has become more widespread, but the depth of engagement varies significantly across borders (Figure 3.5.2). Only five countries, Canada, France, Germany, Italy, and Japan, have consistently endorsed every major global AI governance initiative recorded between 2019 and 2025. Other countries moved in and out of these summits depending on the forum, focus, and timeline but more importantly, not all the countries were able to participate in these global AI governance initiatives. The first intergovernmental standard on AI, the 2019 OECD AI Principles, was restricted to member nations (mainly high-income) and a few partner nations. Likewise, the G7 and G20 discussions remained centered on the world's largest economies. The 2023 Bletchley and 2024 Seoul Summits, however, began to diversify the composition of participants by inviting a broader range of nations, notably including China. The 2025 AI Action Summit in France marked a further turning point, convening over 100 countries alongside civil society organizations and NGOs, with an agenda to prioritize the needs of the Global South and environmental sustainability. Sixty-four participants signed the resulting Statement on Inclusive and Sustainable AI, including the African Union Commission and the European Union. In a notable shift, both the United States and the United Kingdom declined to sign the final declaration. The UK cited a lack of emphasis on national security, while the U.S. decision reflected a pivot toward a more deregulatory, "innovation-first" approach. As engagement at these governance forums becomes more inclusive and substantive, consensus on the terms of cooperation becomes harder to secure.

HIGHLIGHT:

Participation in AI governance mechanisms

Source: Governing AI for Humanity, United Nations, 2025; AI Action Summit, 2025 | Chart: 2026 AI Index report

Figure 3.5.2

3.6 Data Governance for Privacy

Responsible AI practices do not develop evenly across countries. This section assesses that variation for privacy and data governance, drawing on the Global Index on Responsible AI (GIRAI). GIRAI is a benchmark dataset covering 138 countries, built from a quality-reviewed expert survey of 1,862 questions completed by 138 in-country researchers between November 2023 and February 2024. It scores countries on a 0 to 100 scale across thematic areas, covering government frameworks, government actions, and the role of civil society and advocacy organizations. However, it is important to note that low scores do not necessarily indicate that a country is disregarding a certain dimension. In many cases, they reflect earlier stages of AI deployment and diffusion or limited institutional capacity to formalize AI-specific frameworks.

Data Protection and Privacy

The privacy and data protection dimension of the GIRAI score[12] examines whether countries have laws that govern how personal data is collected, used, and shared in AI systems, and whether those laws are backed by regulators with the power to enforce them.

Countries fall across a wide spectrum, with GIRAI scores ranging from near zero to above 80 across the countries surveyed (Figure 3.6.1). Australia and parts of Europe score the highest, while parts of Africa and the Middle East show an absence of dedicated data protection legislation. A complementary map from UNCTAD confirms that most countries now have some form of data protection legislation in place, though a few, mostly concentrated in Africa and parts of Asia, are still in draft stages or have no legislation at all (Figure 3.6.2).

12 Grounded in UDHR Article 12, ICCPR Article 17, the OECD AI Principles, UNESCO's Ethics of AI Recommendation, and UNESCO Principles on Personal Data Protection and Privacy, GIRAI examines explicit laws, oversight, and practice, and assesses frameworks and actions that ensure processing is lawful, fair, purpose-limited, and proportionate. It also evaluates transparency, user information rights, retention limits, accuracy, confidentiality, security, accountability, and rules for data transfers. The index considers national measures—data-protection statutes, automated-decision directives, regulators with enforcement powers, audits, security controls, and initiatives like regulatory sandboxes. It also accounts for nonstate efforts by privacy and digital-rights groups that strengthen protocols and build capacity to mitigate AI-related privacy risks, such as large-scale tracking, profiling, and sensitive-data misuse.

Global AI data protection and privacy assessment

Source: The Global Index on Responsible AI, 2024 | Chart: 2026 AI Index report

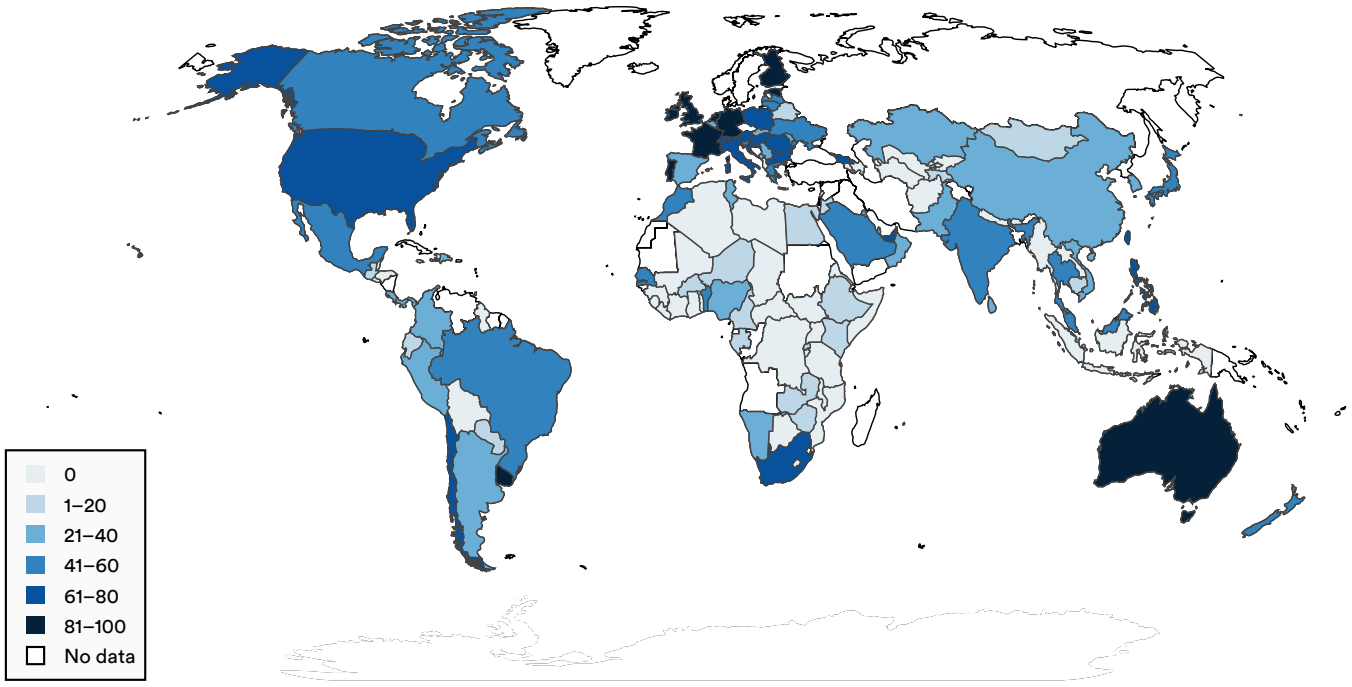


Figure 3.6.1

Global data protection and privacy legislation

Source: UNCTAD, 2025 | Chart: 2026 AI Index report

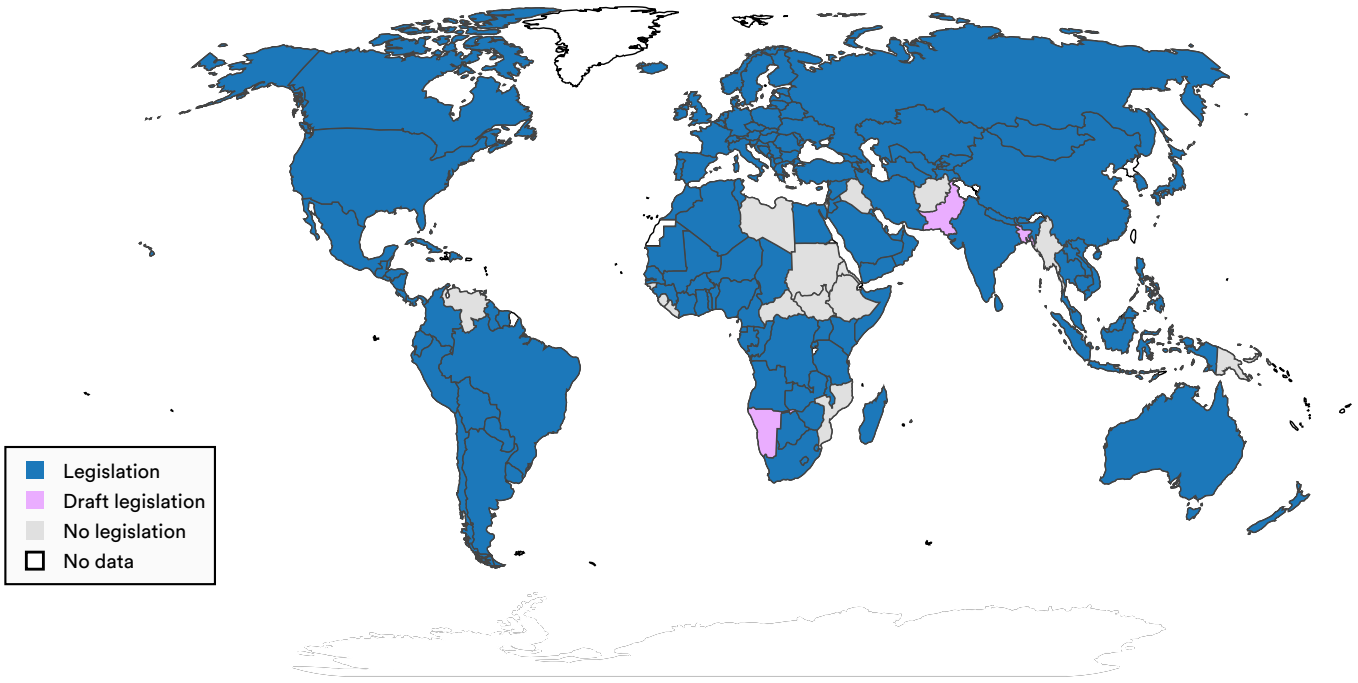


Figure 3.6.2

3.7 Fairness and Bias

Fairness and bias are among the hardest-to-measure dimensions of responsible AI, in part because what counts as fair depends heavily on context. GIRAI scores countries separately on bias and unfair discrimination, gender equality, and cultural and linguistic diversity.

Bias and Unfair Discrimination

The bias and unfair discrimination[13] dimension of the GIRAI score assesses whether countries have explicit measures to prevent and mitigate discriminatory outcomes from AI in its design, development, and deployment. It is meant to address algorithmic bias arising from unrepresentative data, flawed design, or entrenched social inequalities that can harm marginalized groups regardless of intent. It considers whether governments have put laws, oversight bodies, and enforcement mechanisms in place and whether civil society organizations are independently working to monitor and address bias.

GIRAI scores on this dimension are fairly low across the board (Figure 3.7.1). The United States and Canada score highest, with Australia, parts of Europe, and Brazil falling in the middle range. Much of Africa, the Middle East, and Central Asia score below 20.

Global AI bias and unfair discrimination assessment

Source: The Global Index on Responsible AI, 2024 | Chart: 2026 AI Index report

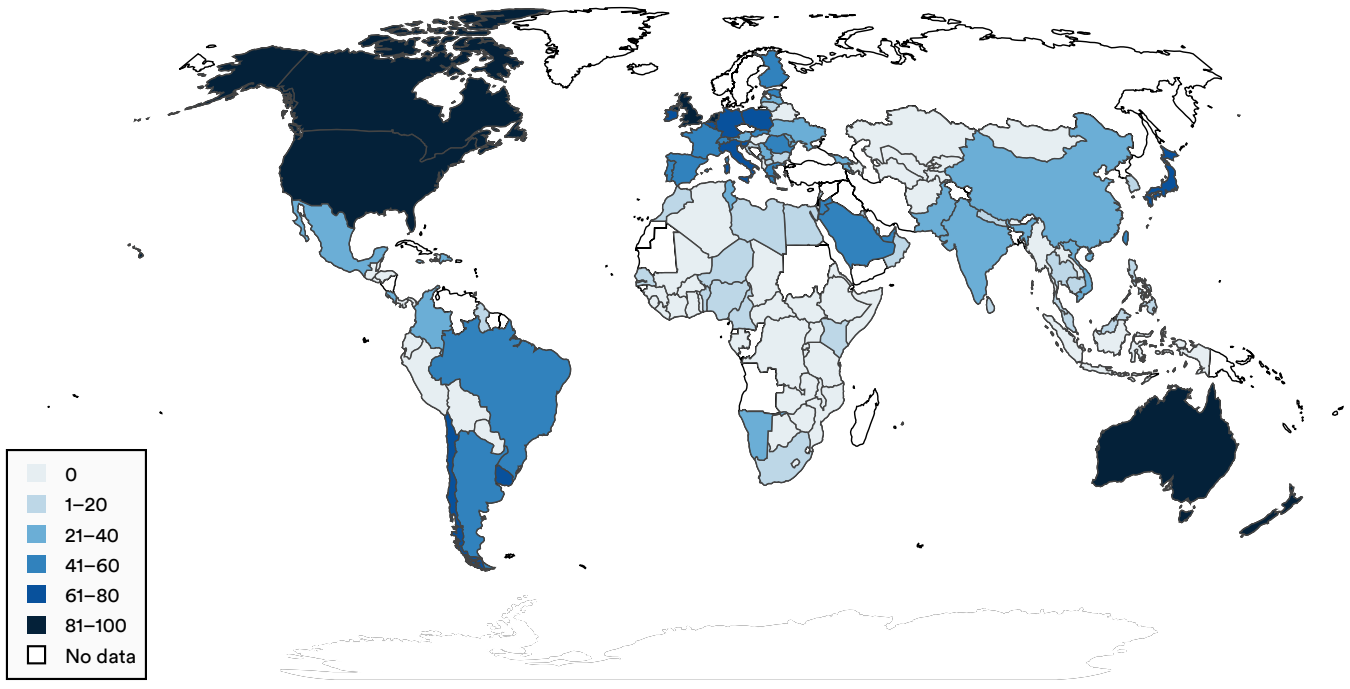


Figure 3.7.1

13 The bias and unfair discrimination dimension of the GIRAI score is grounded in international human rights frameworks (UDHR, ICERD, ICCPR, ICESCR).

Gender Equality

GIRAI's gender equality dimension considers whether countries have state and nonstate initiatives that prevent gender bias and protect equal rights for all gender identities in AI design, development, and use. Canada and The Netherlands score the highest on this measure (Figure 3.7.2). Parts of Europe and Japan fall in the 61–80 range, followed by countries like the United States and Brazil, which score from 41–60.

Global AI gender equality assessment

Source: The Global Index on Responsible AI, 2024 | Chart: 2026 AI Index report

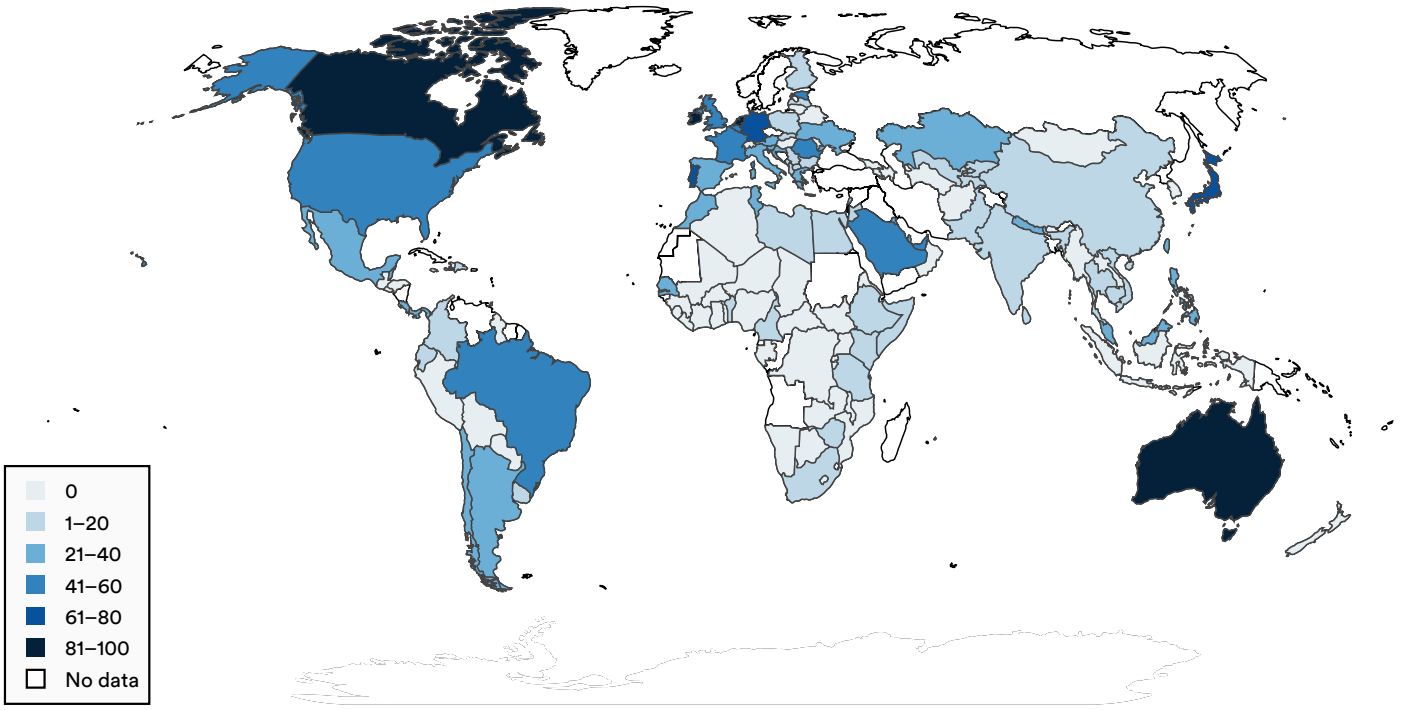


Figure 3.7.2

Cultural and Linguistic Diversity

GIRAI's cultural and linguistic diversity dimension focuses on countries' protective measures on local languages, dialects, indigenous knowledge systems, and cultural diversity broadly across the AI lifecycle. Dominant-culture assumptions can bias AI, marginalize minorities, and erode minority languages. Scores on this dimension are more evenly spread than the others (Figure 3.7.3). Singapore scores the highest, while Germany, Ireland, Italy, Qatar, Estonia, and Slovenia also score in the upper ranges (70–88).

Not all regions protect cultural and linguistic diversity the same way (Figure 3.7.4). In North America, government programs and nonstate actors, such as advocacy groups, research institutions, and digital rights organizations, are active, but formal legal frameworks are less developed. In Europe, Asia, and the Middle East, nonstate actors are also doing more than the government. In Africa, the gap is especially pronounced. Nonstate actors show activity in 39% of countries, but only 7% have government programs and just 2% have legal frameworks in place.

Global AI cultural and linguistic diversity assessment

Source: The Global Index on Responsible AI, 2024 | Chart: 2026 AI Index report

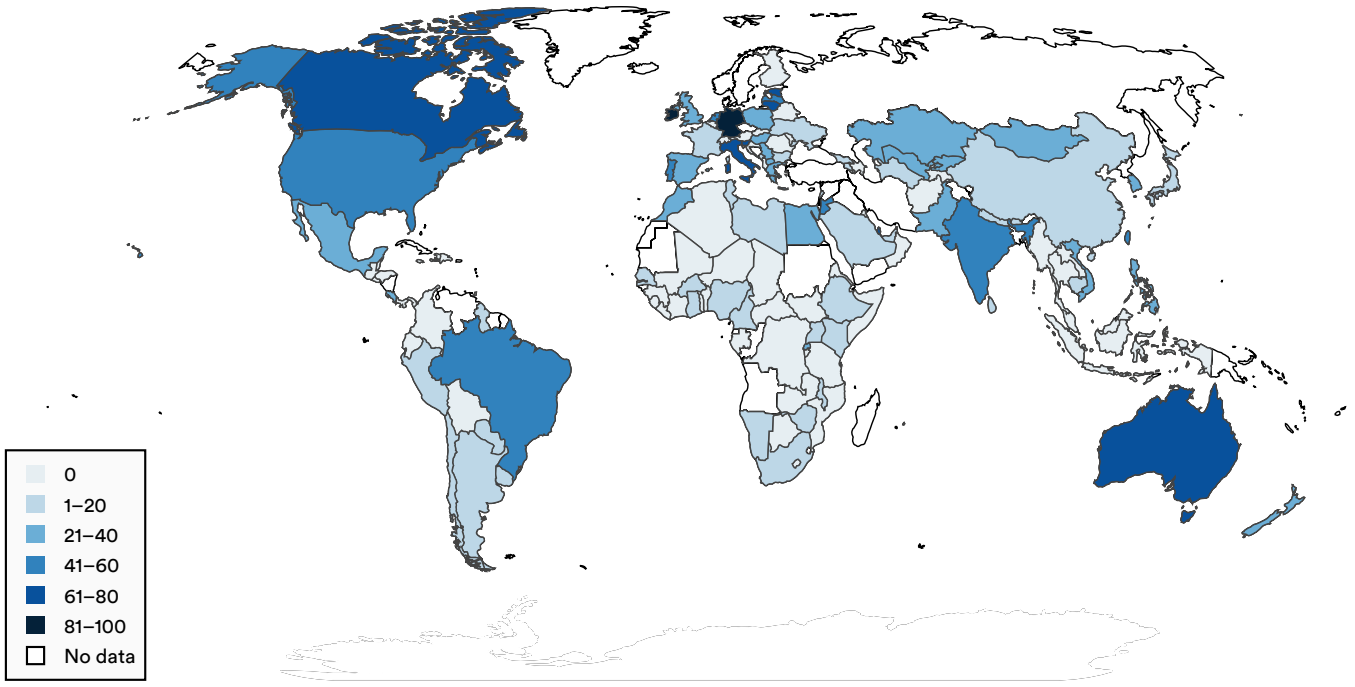


Figure 3.7.3

Share of countries with evidence on cultural and linguistic diversity in AI by region and category

Source: The Global Index on Responsible AI, 2024 | Chart: 2026 AI Index report

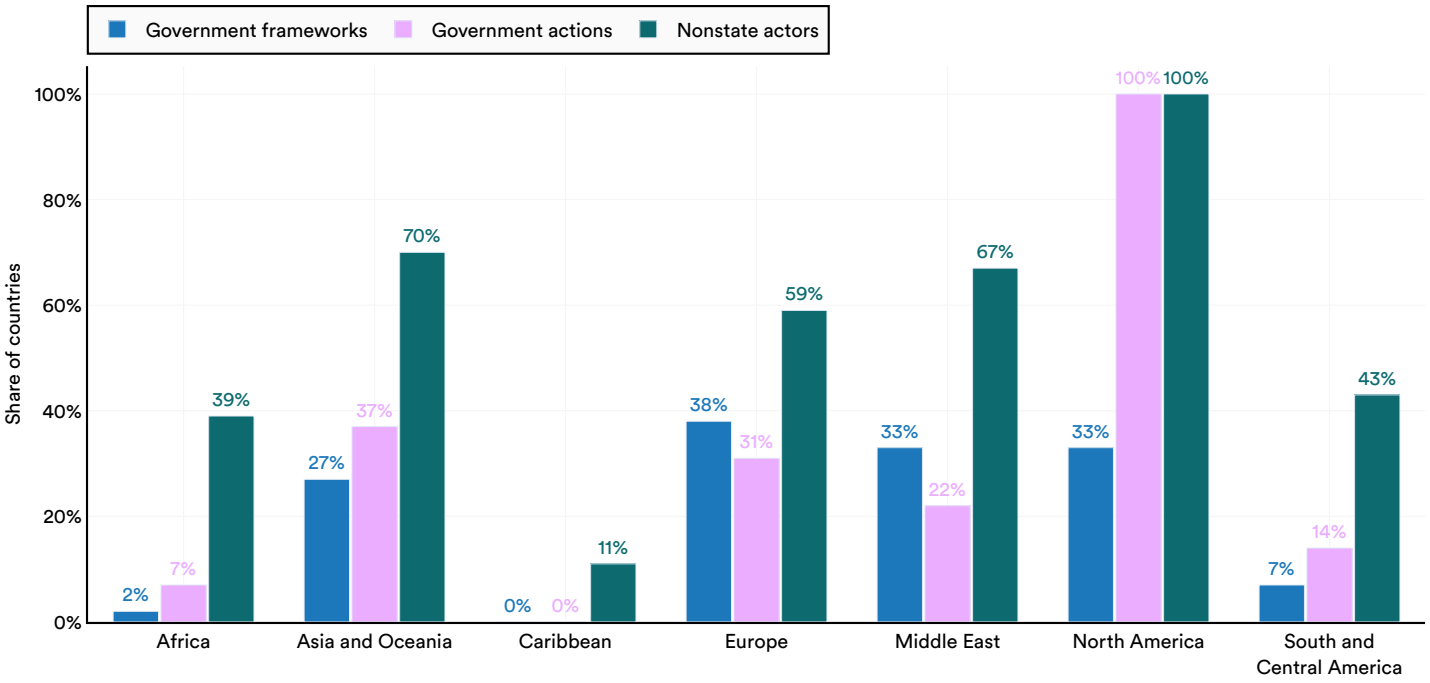


Figure 3.7.4

HIGHLIGHT:

Inclusiveness and the Global Language Gap

As a small number of proprietary models shape global AI capabilities, the "global language gap" has become more visible. These systems perform much better in English and a handful of other widely spoken languages than in all others. This is a responsible AI concern because it determines who benefits from AI systems and who does not.

Efforts continued in the area of language- and culture-specific foundation models and benchmarks, such as KoBEST in 2022 and HAE-RAE in 2023, alongside other Korean-tailored models including Polyglot-Ko and HyperCLOVA X. Spain's Language Technologies Plan, launched in 2019, laid the groundwork for what became the publicly funded ALIA family of Spanish and regional-language models, with earlier regional efforts such as Catalonia's AINA project predating the current wave of regional benchmarking. In 2025, the pace and visibility of this work picked up, with new benchmarks and models emerging across more regions and beginning to register in global evaluation infrastructure.

HELM Arabic, a regional extension of Stanford CRFM's HELM framework developed with Arabic.ai, evaluates models across seven Arabic-language benchmarks covering academic assessment, grammar, and region-specific safety. On this evaluation, the top-scoring model was Arabic.ai's LLM-X, a regionally developed model, with a mean score of 0.86, ahead of Gemini 2.5 Flash (0.82) and GPT-5.1 (0.81) (Figure 3.7.5). Rankings that hold in English-centric evaluations do not necessarily hold when benchmarks reflect local usage, dialect, and cultural references.

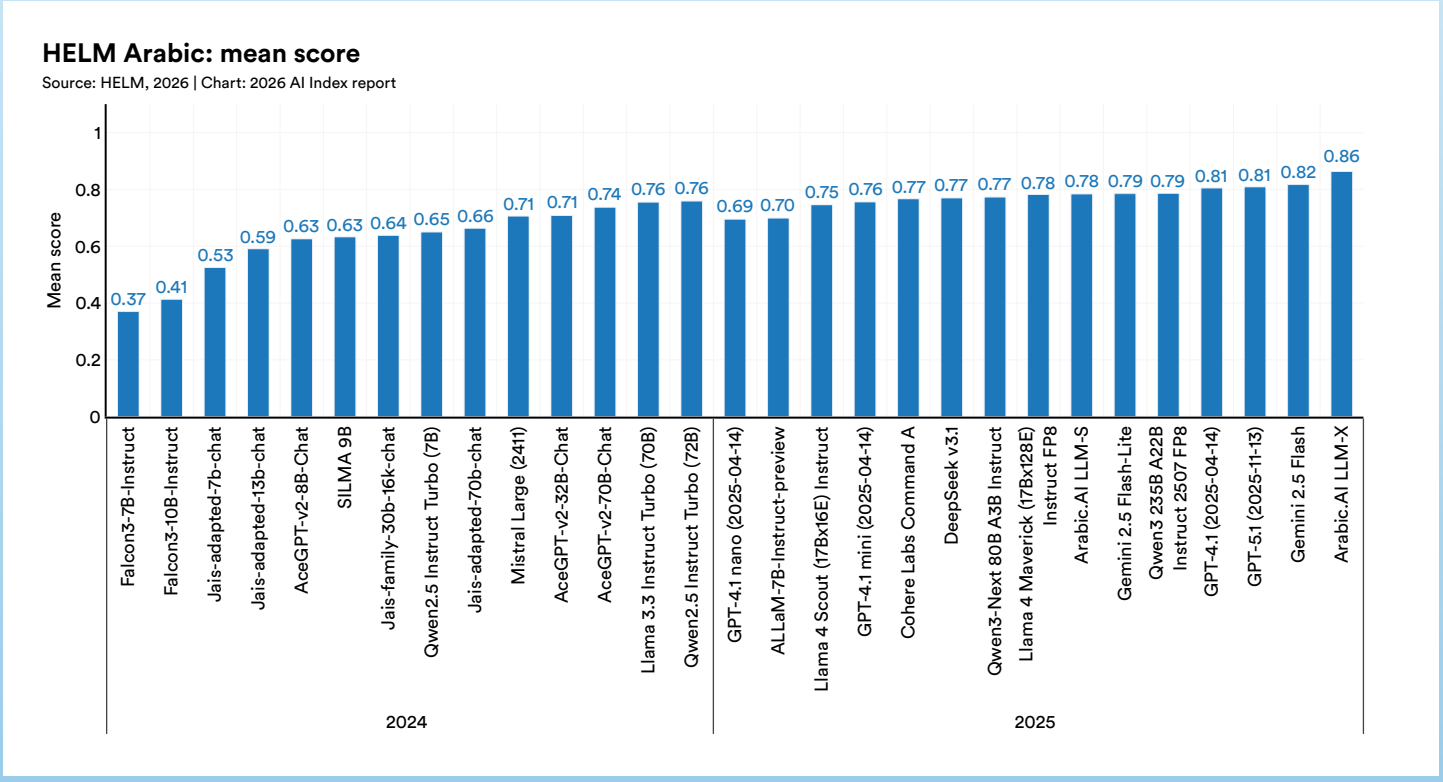


Figure 3.7.5[14]

A similar pattern appears in the Indic LLM Arena, a crowd-sourced evaluation led by AI4Bharat at IIT Madras that tests models across more than 20 Indian languages on language quality, cultural grounding, and safety.

14 Data source: https://crfm.stanford.edu/helm/arabic/latest/.

HIGHLIGHT:

Proprietary models led the leaderboard, with GPT-5.2 scoring 1,314, followed by GPT-5.1 (1,298) and Gemini 3 Flash (1,288) (Figure 3.7.6). Open-source models scored lower but remained competitive, with Qwen3-Next-80B at 1,156 and Llama-4-Maverick-17B at 1,108. The evaluation goes beyond translation accuracy to test whether responses are contextually appropriate for Indian users, a dimension that global benchmarks typically do not capture.

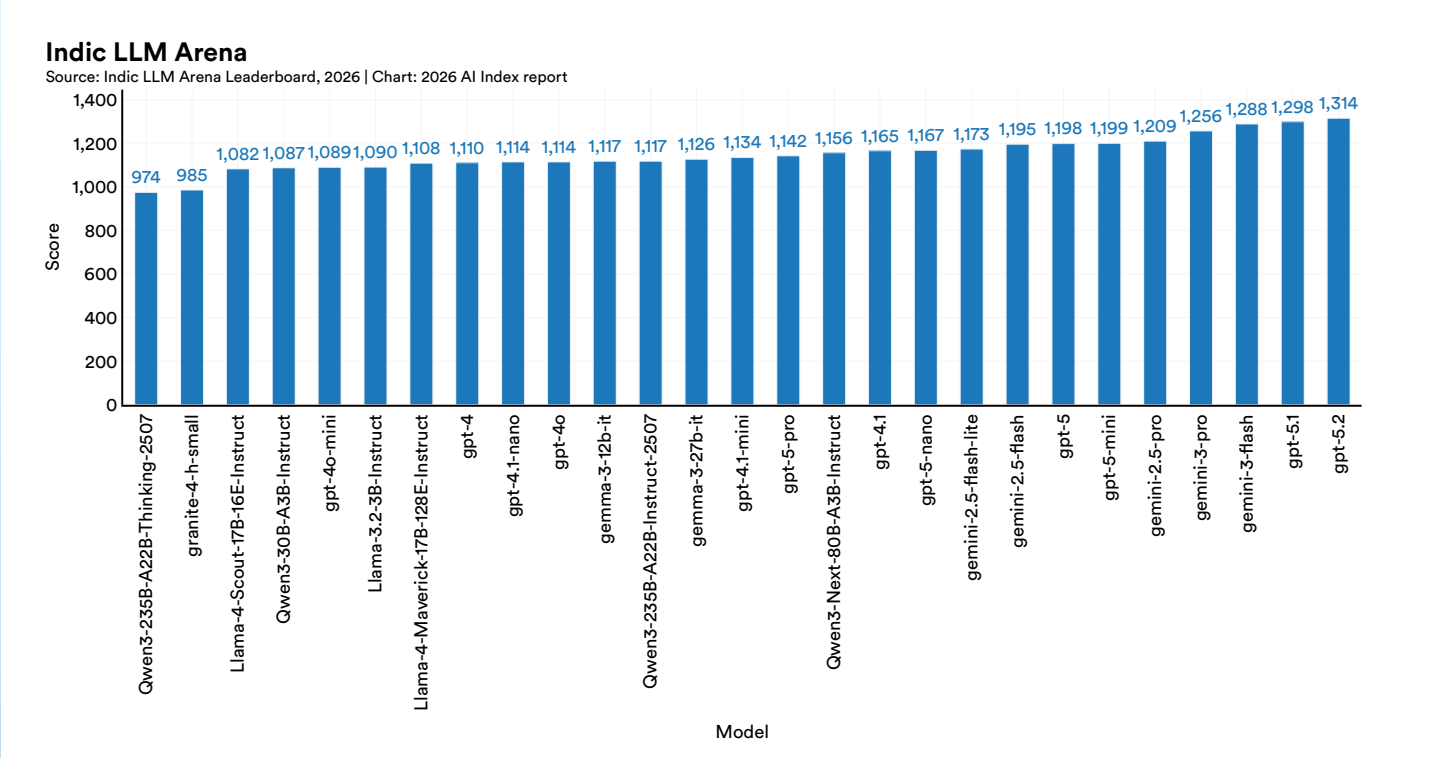


Figure 3.7.6[15]

The gap extends beyond language boundaries to dialect variation within the same language. The Slovene DIALECT-COPA benchmark tests commonsense reasoning in both Standard Slovenian and the Cerkno dialect. GPT-5 scored 99.8% on Standard Slovenian but dropped to 88.6% on the dialect (Figure 3.7.7). The drop was steeper for other models. Mistral Medium 3.1 fell from 90.0% to 53.2%, and Llama 3.3 fell from 87.0% to 53.6%. Dialects differ from standard varieties in spelling, vocabulary, and grammar, and are rarely represented in training data. These gaps suggest that even within languages that models handle reasonably well, performance can degrade sharply for speakers of nonstandard varieties.

15 Data source: https://arena.ai4bharat.org/#/leaderboard/chat/overview.

HIGHLIGHT:

Slobench: accuracy

Source: Slovene DIALECT-COPA benchmark leaderboard, 2026 | Chart: 2026 AI Index report

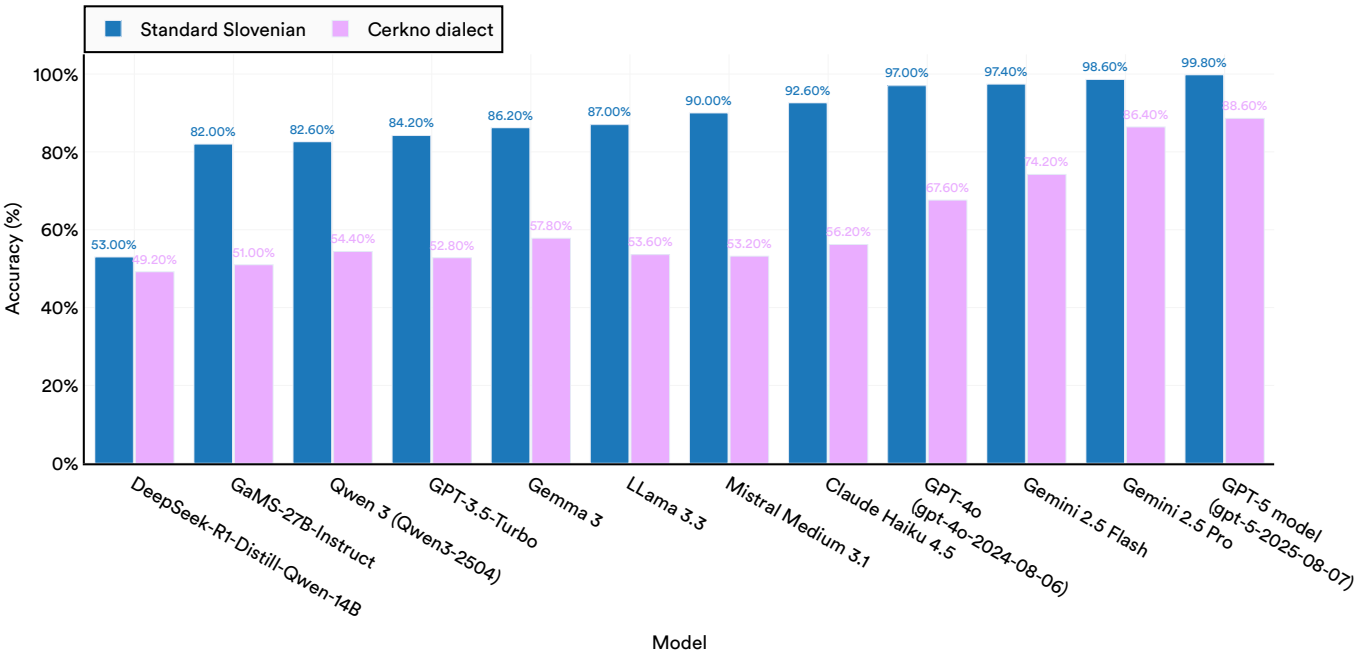


Figure 3.7.7[16]

In response to these gaps, a growing number of regional initiatives are building language-specific AI infrastructure from the ground up rather than waiting for global labs to add coverage. Projects like SEA-LION in Southeast Asia and AI4Bharat in India are developing their own data pipelines, tokenizers, and evaluation benchmarks tailored to local linguistic conditions. Many of the languages these projects serve have structural features, such as complex morphology, script diversity, and limited digitized text, that cause standard multilingual tools to perform poorly. These efforts position linguistic inclusiveness not as an afterthought but as a design requirement, and they represent a growing layer of responsible AI infrastructure outside the major AI-producing regions.

AFRICA

Benchmark	Languages covered	Focus
AfroBench	64 African languages	Multi-task LLM evaluation across NLU, generation, QA/knowledge, and math (15 tasks; 22 datasets)
IrokoBench	17 low-resource African languages across West/East/Southern/Central Africa	Human-translated suite covering NLI (AfriXNLI), math reasoning (AfriMGSM), and multi-choice knowledge QA (AfriMMLU)
TerjamaBench	Darija (Arabic script + Latin "Arabizi")	English↔Darija machine translation benchmark emphasizing cultural context and regional variation (850 entries)

16 Data source: https://slobench.cjvt.si/leaderboard/view/17.

HIGHLIGHT:

HausaMovieReview	Hausa (+ code-switched English)	Sentiment/review-style benchmark from 5,000 Hausa YouTube comments reflecting common code-switching

ASIA, MENA (ARABIC), CENTRAL ASIA

Benchmark	Languages covered	Focus
Indic LLM Arena	Many Indian languages + English-creoles	Crowd-sourced, human-in-the-loop leaderboard evaluating language, culture, and safety in Indian contexts (AI4Bharat; supported by Google Cloud)
SEA-HELM	Filipino, Indonesian, Tamil, Thai, Vietnamese	Southeast Asian holistic evaluation of linguistic and cultural competence across multiple tasks
BATAYAN	Tagalog, Taglish	Holistic Filipino benchmark spanning understanding, reasoning, and generation; explicitly covers code-switching
HELM Arabic	Arabic	Transparent, reproducible Arabic LLM evaluation leaderboard built on established Arabic benchmarks (with Arabic.ai)
BALSAM	Arabic	Community-driven Arabic benchmark and platform with blind evaluation; 78 tasks across 14 categories (52K examples)
Cetvel	Turkish	Unified Turkish LLM benchmark built from 22 datasets covering 7 tasks, with a side-by-side leaderboard
TUMLU	Azerbaijani, Crimean Tatar, Karakalpak, Kazakh, Kyrgyz, Tatar, Turkish, Uyghur, Uzbek	Natively developed multilingual language-understanding benchmark for Turkic languages using middle-/high-school questions across 11 subjects
Kyrgyz LLM-Bench	Kyrgyz	Suite for deep understanding and reasoning in Kyrgyz, combining native benchmarks with translated/post-edited international tasks
ArmBench-LLM	Armenian	Armenian LLM benchmark combining university entrance exams with MMLU-Pro-Hy (1,000-question translated sample)
GeoLogicQA	Georgian	Manually curated 100-question benchmark for logical and inferential reasoning, validated by native speakers
CantoNLU	Cantonese	Seven-task Cantonese NLU benchmark (syntax/semantics, NLI, sentiment, tagging, parsing)
TLUE	Tibetan	Large-scale benchmark measuring LLM proficiency in Tibetan language understanding

HIGHLIGHT:

EUROPE

Benchmark	Languages covered	Focus
BenCzechMark	Czech	Comprehensive Czech-centric benchmark with 50 tasks, multiple formats/metrics, and significance-aware aggregation
CUS-QA	Czech, Slovak, Ukrainian	Open-ended regional QA benchmark with text and visual grounding, curated by native speakers with English translations
COLE	French	23-task French natural language understanding (NLU) benchmark emphasizing French-relevant linguistic phenomena (used to benchmark 94 LLMs)
Estonian Benchmark	Estonian	Benchmark built from seven datasets covering knowledge, grammar/vocabulary, summarization, and contextual comprehension
IberBench	Basque, Catalan, Galician, Spanish, Portuguese, English	Large, extensible benchmark integrating 101 datasets across 22 task categories (e.g., toxicity, summarization) with community-driven updates
IberoBench	Basque, Catalan, Galician, European Spanish, European Portuguese	Multi-task benchmark (62 tasks; 179 subtasks) built on the LM Evaluation Harness framework
Polish linguistic and cultural competency	Polish	600 manually crafted questions evaluating Polish history, geography, culture/tradition, arts, grammar, and vocabulary
LLMzSzŁ	Polish	Exam-based benchmark drawn from Polish national exams (~19K closed-ended questions across 154 domains)
ITALIC	Italian	Culture-aware Italian NLU benchmark with 10,000 multiple-choice questions spanning 12 domains
SloBENCH	Slovenian	Evaluation platform with multiple leaderboards, including DIALECT-COPA (standard vs. dialect) and Slovene speech recognition

Figure 3.7.8

3.8 Transparency

Transparency measures how much developers disclose about how their models are built, trained, and deployed. Two independent indices track this from different angles.

The Openness Index

The Artificial Analysis Openness Index scores AI models on a 0 to 100 scale based on how freely weights can be accessed and licensed, as well as the level of transparency around training methodology and pre- and post-training data. Scores are low across leading models, with most falling between 2 and 16 out of 100 (Figure 3.8.1). K2 Think and Olmo 3 32B Think scored the highest, and they are also the only two models that scored any points for pre-training data transparency. Every other model in the index scores zero in that category. Model Availability and methodology disclosure account for the bulk of points across all models. As Chapter 1's discussion of access and deployment noted, over 90% of notable industry models were released without training code in 2025. The Openness Index results suggest that pattern extends beyond code to training data as well.

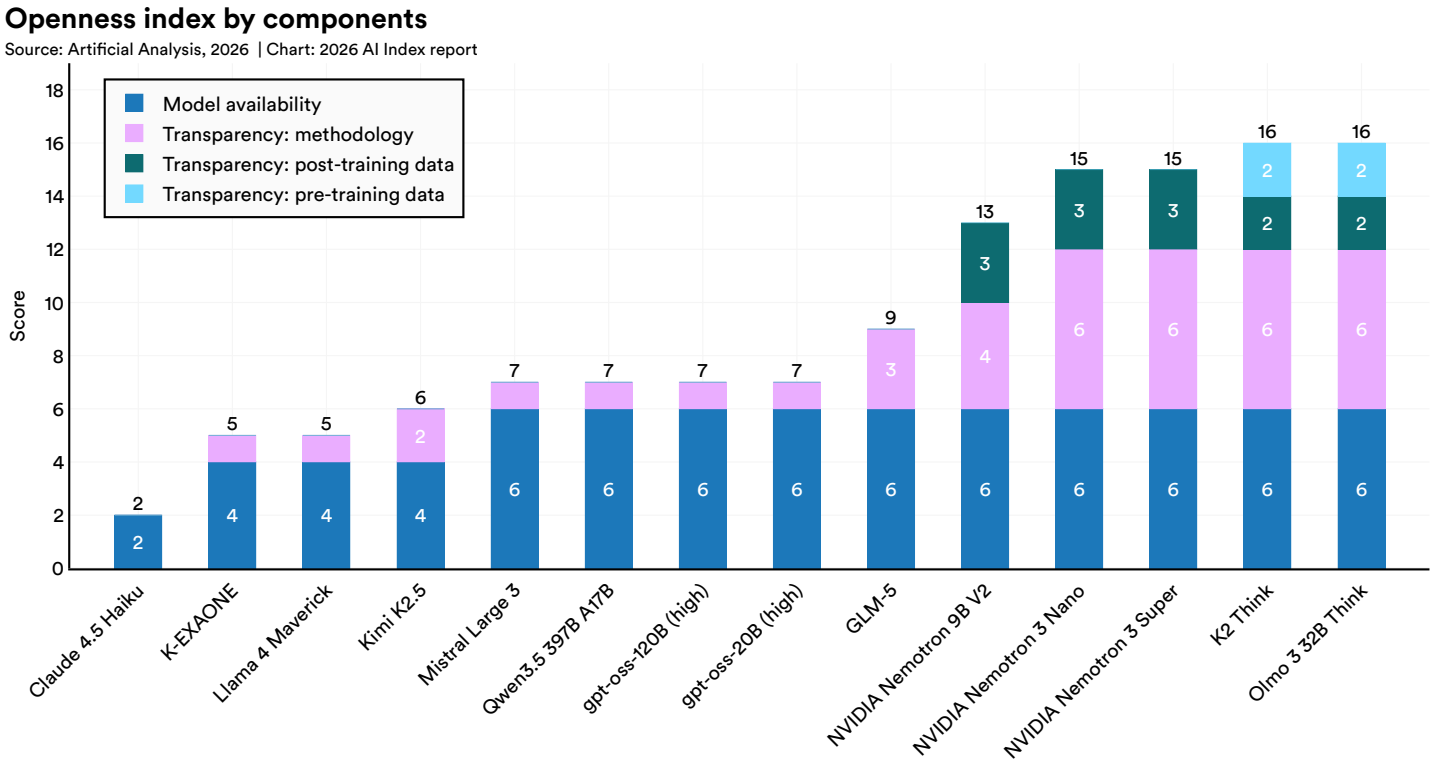


Figure 3.8.1

Foundation Model Transparency Index

The Foundation Model Transparency Index (FMTI) takes a different approach, scoring developers rather than individual models. Now in its third year, it evaluates disclosure across three stages of the model lifecycle. Upstream covers what goes into building a model, including training data, labor, and compute. Model covers

what is disclosed about the system itself, and Downstream covers what happens after release, including monitoring and impact reporting.

In the 2025 edition, average transparency declined from 58 in 2024 to 40 (Figure 3.8.2). IBM leads at 95 and Writer follows at 72. Others, such as xAI and Midjourney score just 14, whereas open model developers, B2B enterprise providers, organizations publishing transparency reports, and EU AI Act signatories tend to perform better. As with the Openness Index, the weakest area is Upstream, particularly around training data and the resources used to build models (Figure 3.8.3).

Foundation Model Transparency Index Scores by Domain, 2025

Source: 2025 Foundation Model Transparency Index

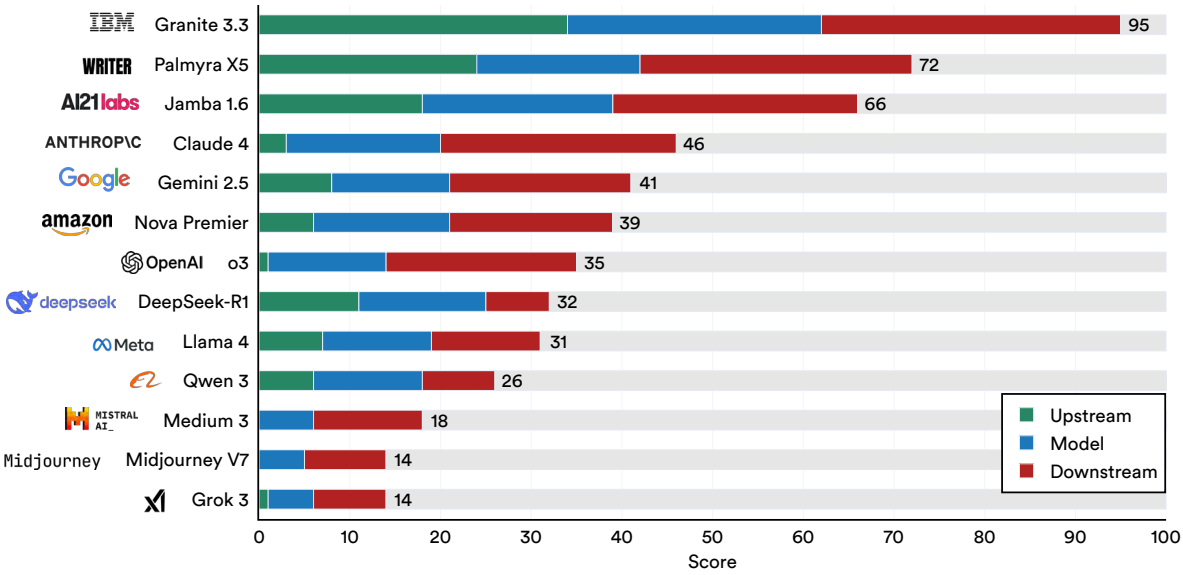


Figure 3.8.2

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2025

Source: 2025 Foundation Model Transparency Index

Major Dimensions of Transparency	Jamba 1.6	Qwen 3	Nova Premier	Claude 4	DeepSeek-R1	Gemini 2.5	Granite 3.3	Llama 4	Midjourney V7	Medium 3	o3	Palmyra X5	Grok 3	Average
Data Acquisition	92%	17%	17%	25%	17%	33%	100%	33%	0%	0%	8%	58%	0%	31%
Data Properties	0%	20%	0%	0%	20%	0%	100%	20%	0%	0%	0%	40%	0%	15%
Compute	22%	11%	11%	0%	44%	11%	100%	22%	0%	0%	0%	100%	11%	26%
Model Information	75%	75%	0%	25%	75%	0%	100%	75%	0%	0%	0%	75%	0%	38%
Model Access	50%	50%	50%	50%	50%	50%	100%	50%	0%	25%	0%	50%	0%	40%
Capabilities	75%	50%	50%	25%	50%	25%	75%	50%	0%	25%	25%	50%	25%	40%
Risks	60%	0%	40%	60%	20%	20%	100%	20%	0%	0%	60%	40%	0%	32%
Model Mitigations	60%	0%	60%	80%	20%	40%	80%	0%	0%	20%	80%	40%	0%	37%
Release	88%	63%	75%	75%	63%	88%	100%	50%	63%	38%	63%	88%	50%	69%
Usage Data	20%	0%	20%	60%	0%	0%	80%	0%	20%	0%	20%	100%	0%	25%
Impact	71%	0%	0%	29%	0%	29%	86%	14%	29%	14%	14%	86%	0%	29%
Post-deployment Monitoring	71%	0%	57%	57%	0%	43%	100%	29%	0%	43%	71%	86%	0%	43%
Model Behavior Policy	100%	50%	75%	100%	75%	75%	100%	75%	25%	0%	75%	50%	75%	67%
Acceptable Use Policy	80%	60%	80%	100%	60%	80%	80%	40%	60%	60%	60%	80%	60%	69%
Downstream Mitigations	100%	40%	100%	100%	0%	100%	100%	80%	40%	80%	100%	100%	40%	75%
Average	**64%**	**29%**	**42%**	**52%**	**33%**	**40%**	**93%**	**37%**	**16%**	**20%**	**38%**	**69%**	**17%**	

Figure 3.8.3[17]

17 Data, labor, compute, and methods were upstream indicators; model basics, access, capabilities, risks, and mitigations were model-level indicators; and distribution, usage policy, feedback, and impact were downstream indicators.

3.9 Security and Safety

Safety is the responsible AI dimension where institutional infrastructure has grown the fastest. New evaluation frameworks, government-backed AI safety institutes, and standardized benchmarks have all expanded in the past year. This section traces that growth and the resulting data on how well current models handle safety in practice.

Global AI Safety Institutes

AI safety institutes (AISIs) are state-backed specialist organisations created to help governments understand and manage risks from advanced AI, especially frontier/foundation models. They conduct technical evaluations and safety research that governments can use to shape policy.

Fully operational institutes now exist in the UK (AI Security Institute), the U.S. (USAISI at NIST), Japan (JAISI), Singapore (Digital Trust Centre), and Israel (AI Security Research Unit) (Figure 3.9.1). India and France have also launched AISIs, with India's AI Safety Institute and France's Current AI. A second wave is in development in Canada, South Korea, Germany, and Brazil. Outside of these standalone institutes, participation is growing through the International Network of AI Safety Institutes, with Kenya and Australia listed as network members without formal institutes of their own.

The countries building these AISIs are still mostly wealthy, technologically advanced economies that are not all pursuing the same goals. The UK and Israel emphasize security, while the EU AI Office pairs evaluation with enforcement powers under the AI Act. Network membership is a practical entry point for countries without the resources to stand up a full institute immediately.

AI safety institutes and network membership

Source: All Tech is Human, 2025 | Chart: 2026 AI Index report

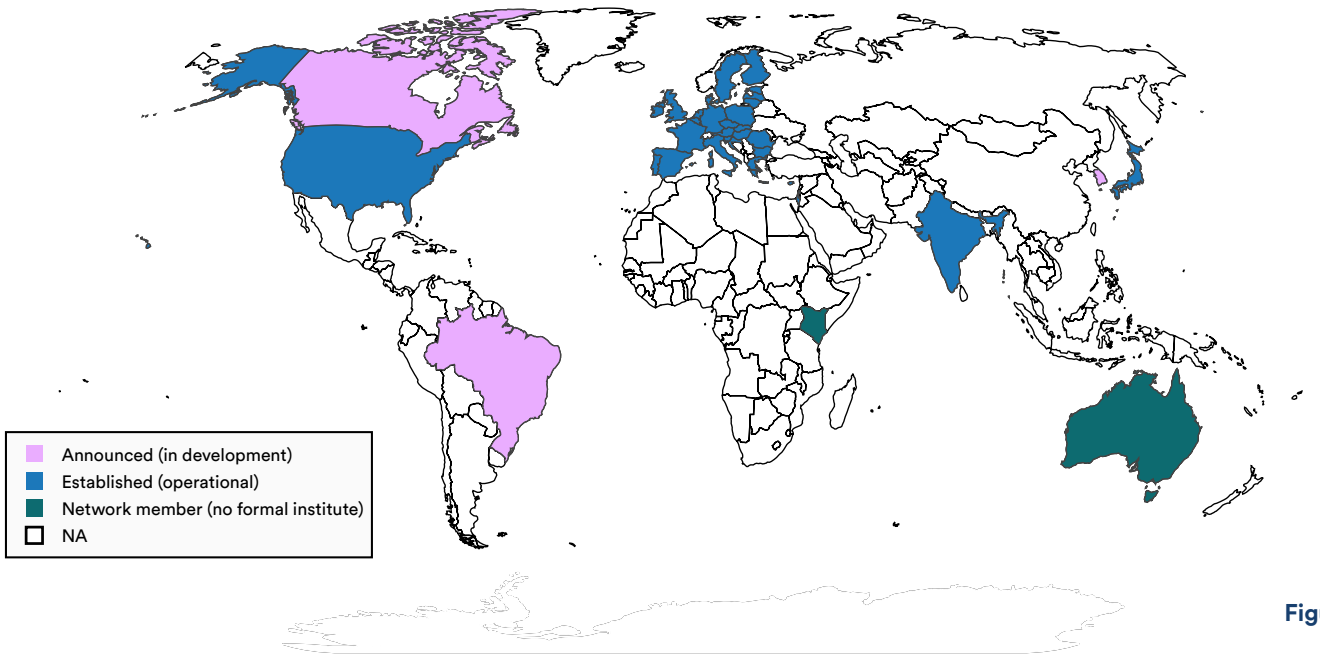


Figure 3.8.3[18]

18 Data source: https://alltechishuman.org/all-tech-is-human-blog/the-global-landscape-of-ai-safety-institutes.

Benchmarks

HELM Safety

HELM Safety, covered in last year's report, continues to be one of the few standardized suites for evaluating AI models across responsibility and safety metrics. It tests models from major developers across benchmarks including BBQ (social bias), SimpleSafetyTests (self-harm and abuse risks), HarmBench (harassment and misinformation), AnthropicRedTeam (adversarial conversations), and XSTest (helpfulness vs. harmlessness trade-offs).

The 2025 results show continued improvement but also increasing compression at the top (Figure 3.9.2). Most models released between 2024 and 2025 score between 0.90 and 0.98, with a very narrow gap between the highest and lowest scorers. Older models from 2023 score lower, but the overall trajectory suggests that leading models are converging on a safety ceiling where current benchmarks may not be fine-grained enough to distinguish meaningful differences.

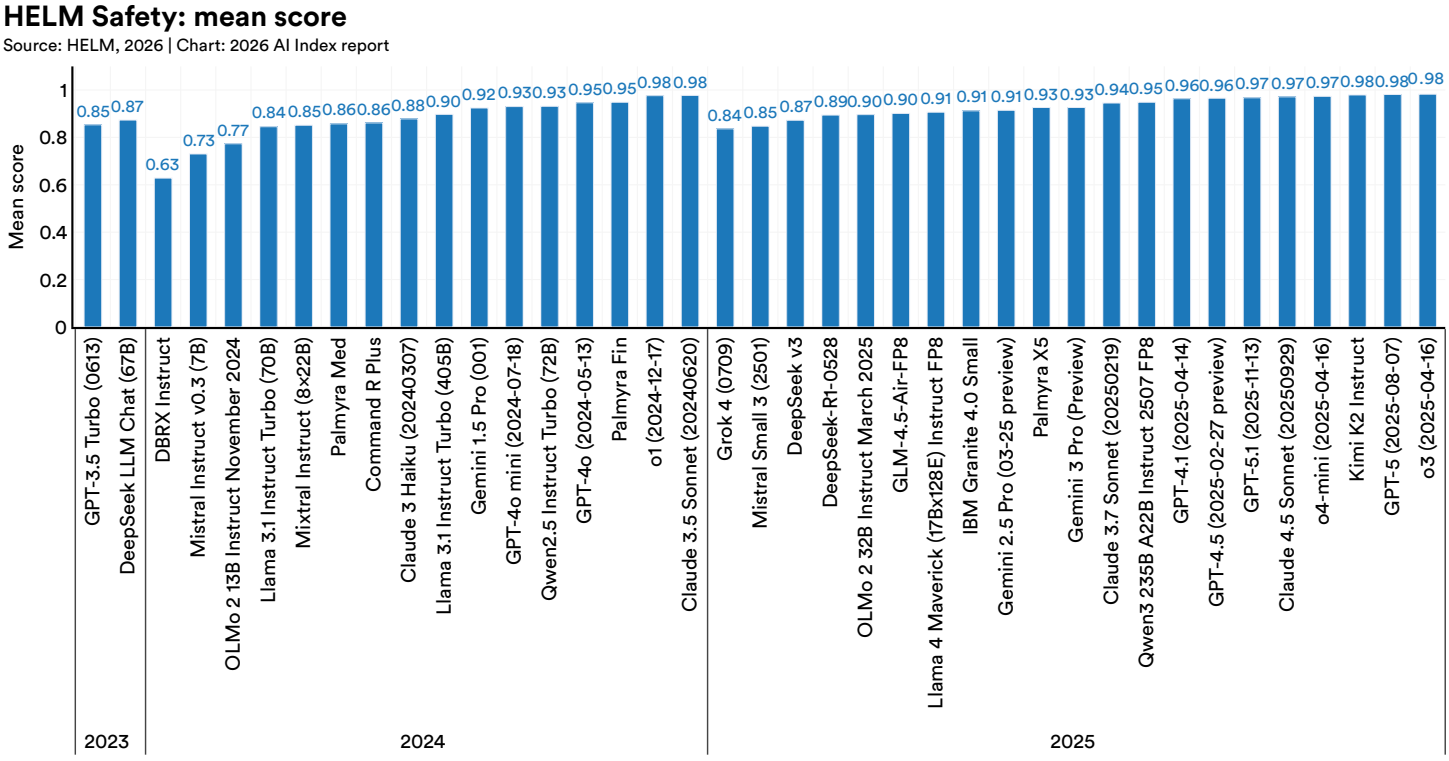


Figure 3.9.2

AILuminate

AILuminate v1.0 is a new benchmark designed to test how well AI systems resist prompts that could trigger dangerous, illegal, or undesirable behavior. It covers 12 hazard categories, including violent crimes and child exploitation, and employs a five-tier grading scale from "Poor" to "Excellent." The benchmark includes two separate evaluations. The first tests safety under normal use, with models evaluating both with and without external safety filters and moderation tools. The second tests a system's ability to resist deliberate jailbreak attempts through adversarial prompts.

Safety Benchmark Results

Among models tested with external guardrails in place, Claude 3.5 Haiku, Claude 3.5 Sonnet, and Mistral Large all received "very good" ratings, while their parent models received "good"(Figure 3.9.3). In the set of models that could be tested without external safety filters or moderation tools, Gemma 2 9b, Phi 3.5 MoE Instruct, and Phi 4 scored "very good" (Figure 3.9.4). The two groups are not directly comparable, as they involve different models under different conditions, but both show a baseline safety performance of "good" across leading systems.

Name	Grade
Claude 3.5 Haiku 20241022	Very Good
Claude 3.5 Sonnet 20241022	Very Good
Mistralai Mistral Large 2402 Moderated	Very Good
Amazon Nova Lite v1.0	Good
Gemini 1.5 Pro (API, with option)	Good
Gemini 2.0 Flash 001	Good
Gemini 2.0 Flash Lite	Good
GPT-4o	Good
GPT-4o mini	Good
Ministral 8B 24.10 with output moderation (Recipe)	Good

Source: MLCommons, 2025

Figure 3.9.3

Name	Grade
Gemma 2 9b	Very Good
Phi 3.5 MoE Instruct	Very Good
Phi 4	Very Good
Athene V2 Chat Hf	Good
Aya Expanse 8B Hf	Good
Cohere C4Ai Command A 03 2025 Hf	Good
Llama 3.1 405B Instruct	Good
Llama 3.1 8b Instruct FP8	Good
Llama 3 1 Tulu 3 8B Hf	Good
Mistralai Mistral Large 2402	Good
Olmo 2 0325 32b Instruct	Good
Olmo 2 1124 13B Instruct Hf	Good
Phi 3.5 Mini Instruct	Good
Qwen1 5 110B Chat Hf	Good
Yi 1 5 34B Chat Hf	Good
Ai21Labs Ai21 Jamba Large 1.5 Azure	Fair
Google Gemma 3 27B It Hf Nebius	Fair
Llama 3.3 70B Instruct Turbo Together	Fair
Ministral 8B 24.10 (API)	Fair
Mistral Large 24.11	Fair
Qwq 32B Hf	Fair
OLMo 7b 0724 Instruct	Poor

Source: MLCommons, 2025

Figure 3.9.4

Jailbreak T2T Benchmark v0.5 Results

The AILuminate Jailbreak T2T benchmark v0.5 tests what happens when users deliberately try to bypass a model's safety measures through adversarial prompts. Each model in the chart receives two scores (Figure 3.9.5). The square at the top represents the model's safety score under normal conditions, while the circle below it represents the score after being exposed to jailbreak attempts. As this is a beta version of the benchmark, models are de-identified by number, rather than named.

Under normal conditions, most models score in the "very good" or "good" range. After jailbreak attempts, nearly every system's score drops, some by a full tier or more. So while safety under normal use is generally good, it degrades under deliberate manipulation.

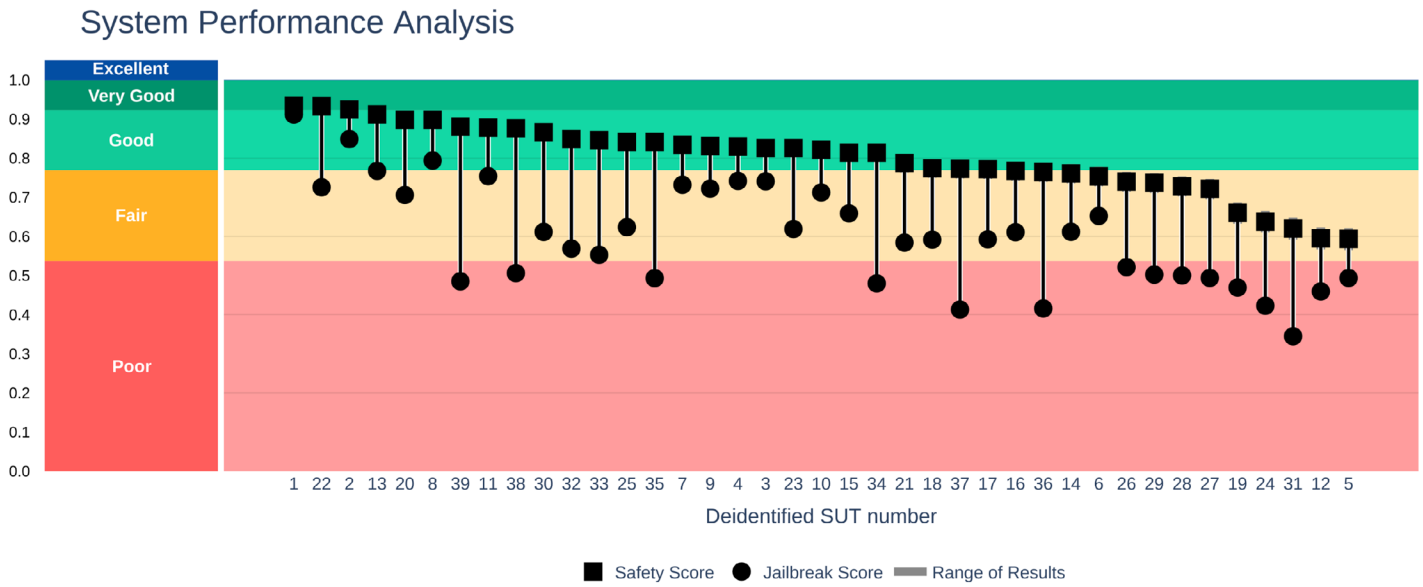


Figure 3.9.5

3.10 Tradeoffs Across RAI Dimensions

In practice, AI systems must satisfy multiple responsible AI dimensions at once. A growing number of empirical research studies suggest that these dimensions do not improve independently, as optimizing for one can degrade others. The direction and magnitude of those trade-offs depends on the method used, data involved, and under what context it is deployed.

Kemmerzell and Schreiner (2024) tested this directly by training image classification models on four facial analysis data sets and measuring what happened to fairness, privacy, explainability, and robustness when each dimension was targeted in isolation. Differential privacy, a technique that adds noise during training to prevent individual data points from being identified, improved privacy scores across all datasets but reduced explainability, fairness, and accuracy, with accuracy falling by up to 33 percentage points on some configurations. Training adaptations aimed at improving fairness only succeeded on the dataset with the most demographic imbalance, and therefore the most room to correct. But across all, it reduced explainability and robustness. Data augmentation methods designed to improve robustness by exposing datasets to more varied training images produced the fewest negative side effects across the same experiments. It also improved explainability and accuracy, with only minor reductions in privacy and fairness. There was not a single intervention method that proved to improve all four dimensions at once.

A separate evaluation of large language models found a similar pattern at the model level. Cecchini et al. (2024) scored 11 models across robustness, accuracy, and toxicity using the LangTest evaluation toolkit. GPT-4 led on robustness (average score of 0.91 out of 1.0) and accuracy (0.67), but Llama 2 7B scored highest on toxicity avoidance (0.98), meaning it was the most likely to refuse toxic prompts. Models that performed well on robustness, such as Mistral 7B and Mixtral 8×7B, scored among the lowest on toxicity avoidance (0.39 and 0.42, respectively). The ranking of models shifted depending on which dimension was being measured, and no single model was a clear leader in all three.

These trade-offs also appear in federated learning, a training approach where multiple institutions train a shared model by exchanging model updates rather than the underlying data. Wasif et al. (2025) studied how privacy-preserving techniques interact with fairness in this setting across four datasets, including Alzheimer's disease MRI scans and credit card fraud records. Differential privacy did not affect all datasets equally. Institutions with larger datasets could absorb the added noise, while smaller institutions saw their contributions to model training degraded. In the Alzheimer's scenario, adding stronger privacy protections reduced the model's ability to correctly identify the disease, with accuracy falling by 14.8 percentage points. The effect was worse for hospitals with less data, where missed diagnoses rose by 21.4%. Two alternative privacy methods that use encryption instead of noise kept fairness more stable but required two to three times more computing power.

The studies covered above are recent and focus on specific tasks rather than general-purpose AI systems. Their findings point in the same direction though: Improving one responsible AI dimension tends to come at the expense of another. There is no shared framework that measures or compares these trade-offs, which is another measurement gap in the RAI space, and makes it difficult to track whether the field is getting better at managing them.