MARKET SCAN REPORT

FEBRUARY 2026

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

IN FOCUS

PRAGMATIC AUDIT CHECKS AND METRICS TO ASSESS LLMS: PRACTICAL TESTS FOR SAFETY AND COMPLIANCE

By Uday Kashinath Tashildar

Infosys topaz

Infosys®
Navigate your next

Foreword

AI's next chapter is about the choices we make when no one is watching. We are moving from building fast to building well. The question is no longer only what a system can do. What matters now is whether it behaves in ways people can understand, predict, and trust. That bar is rising and it should. As AI steps closer to everyday work and public life, the standard rises. Proof matters more than promise. Habits matter more than hype. This is the moment to put clarity, care, and accountability at the center of how we build.

This edition leans into that shift. It takes a grounded view of where the field is going and focuses on how responsibility is being turned into practice rather than posture.

Across the ecosystem, the picture is mixed but instructive. There is real progress. There are also gaps that deserve attention. Oversight and disclosure are getting stronger. Fairness and transparency still need steady work. The pattern is becoming clear. Speed alone is not enough. What will count is whether growth strengthens safety, resilience, and public confidence.

Collaboration is changing in substance, not just style. Clear expectations are meeting deeper craft. Product teams are testing earlier, writing down decisions, and inviting reviews before launch. Policy, risk, and engineering are working from the same playbook. Responsible AI is moving from a checklist to a discipline, visible in design choices, signoff gates, and post launch monitoring. Progress is steady, and it is starting to stick.

Our In Focus section brings this spirit to life. This month we feature **Mr. Uday Kashinath Tashildar's** article, ***"Pragmatic Audit Checks and Metrics to Assess LLMs: Practical Tests for Safety and Compliance."*** The piece makes a simple and important case. Strong audits do not slow good teams. They align builders and reviewers on the same map and the same measures. They help prove a model is ready for real use and give users a reason to trust what they cannot see. That is how responsibility becomes visible.

Looking ahead, the opportunity is within reach. Innovation will keep moving quickly, as it should. The difference will come from how carefully we guide it. If we choose clarity over shortcuts and evidence over assumptions, we can unlock the value of AI without losing trust. This is the time to set that standard and keep it steady. In the end, the future will remember not just what we built, but how we chose to build it.

Syed Ahmed
Global Head
Infosys Responsible AI Office

From the editor's desk

After the International Convergence: Making Responsible AI Real in Practice

February did not just bring headlines; it brought clarity. The India AI Impact Summit set the direction, not through grand speeches, but through grounded conversations that shape how we build. There is a widening agreement about responsible AI, and that matters. **But international alignment on principles often moves faster than the reality of building systems that behave reliably.** What counts now is not what we agree on in theory, but how those ideas appear in design choices, development habits, and real-world decisions.

Across regions, governments moved from principle to obligation. In the United States, the executive push on AI transparency set new expectations for how systems must be documented and disclosed. The United Kingdom deepened its accountability framework, particularly about high-risk applications. The European Union's AI Act began moving from legislative text into operational reality. As more countries collaborated through emerging alliances and shared frameworks, a clear signal emerged: responsibility is no longer abstract. **Policy is moving closer to practice, and the world is recognizing that principles only matter if they can stand up in the systems we actually deploy**. This shift reflects a deeper truth from international discussions: **governance is as much about tradeoffs as it is about ideals**. And those tradeoffs surface when product deadlines press, when teams interpret guidance differently, and when international standards meet local realities.

Progress continued, but the sharp edges showed themselves too. Deepfake misinformation found new routes about existing filters. A false AI generated claim spilled into legal action. Repeated attempts to extract or probe advanced models reminded us that capability and risk evolve side by side. None of these stories were dramatic in isolation, yet together they highlighted something important. Real challenges appear in system design, deployment choices, and institutional capacity, not in neatly written frameworks. When ownership is unclear or accountability weakens under pressure, issues surface quickly.

Innovation moved forward with equal pace. We saw smarter document reading models, deeper reasoning systems, and tools that can create complex, accurate visuals. Models like Sarvam Vision, Gemini 3.1 Pro, and frameworks such as PaperBanana point toward systems that understand and create with more nuance. The direction matters as much as the capability. **International convergence is useful, but what truly counts is whether these ideas survive the realities of implementation**. Governance has to work where people build, not only where policies are written.

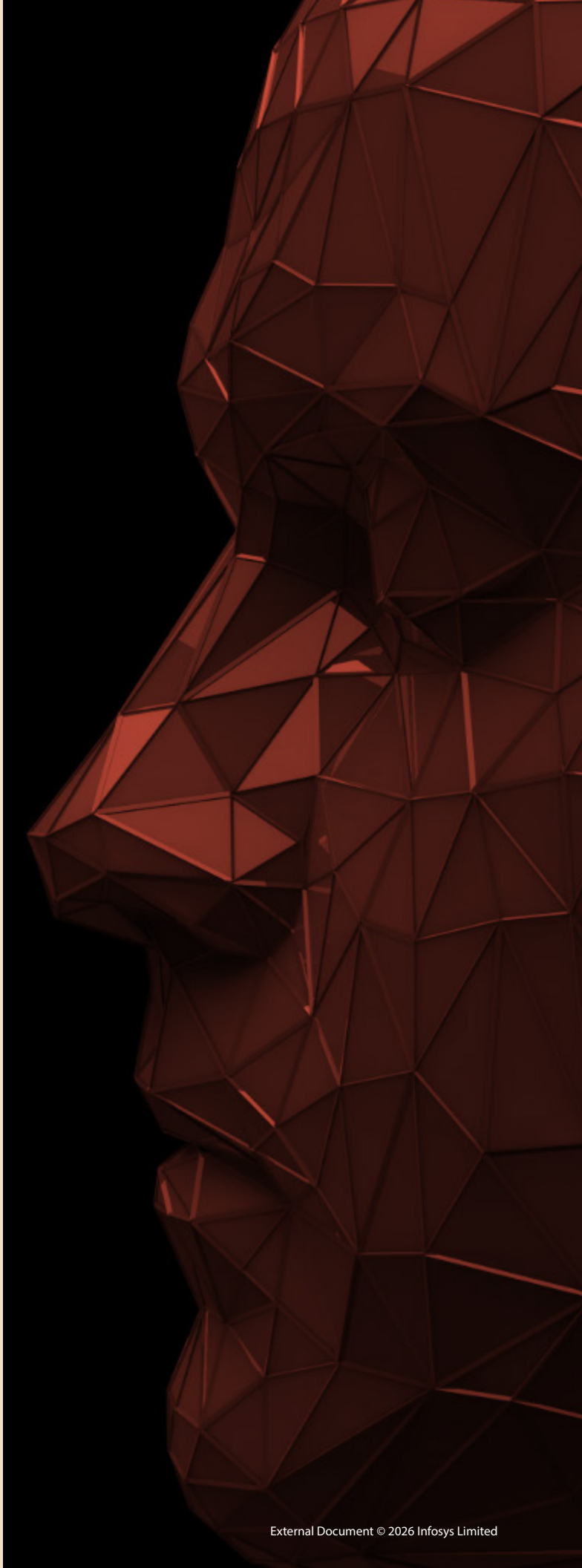
That is why our In Focus section brings this shift into clear view. We feature Mr. Uday Kashinath Tashildar's article, "Pragmatic Audit Checks and Metrics to Assess LLMs: Practical Tests for Safety and Compliance." The piece underscores something the ecosystem is now acknowledging openly. Responsible AI grows not from intent, but from discipline. Strong audits give teams a common understanding, align expectations, and make trust something that can be demonstrated, not assumed.

What comes next depends on daily choices. **The future of AI governance will not be shaped by the number of frameworks we write, but by the few we make work**. Innovation will move fast, but trust will grow slowly through steady habits and consistent proof. The standard we set now: in the decisions we make under pressure, in the systems we choose to ship, in the accountability we refuse to dilute, is the standard that will define what responsible AI actually meant when it mattered most.

Warm regards,

Ashish Tewari
Head- Infosys Responsible AI Office, India

Table of Contents



AI Regulations, Governance & Standards

This section highlights the recent updates on regulations and governance initiatives across the globe impacting the responsible development and deployment of AI.

AI Regulations & Governance across the globe

ASEAN Leaders Unite to Put Singapore's New Agentic AI Governance Framework into Practice

Association of South East Asian Nations(ASEAN) leaders from Singapore, Thailand, Malaysia, Indonesia, and the Philippines have joined forces to help organizations apply Singapore's new Model AI Governance Framework for Agentic AI, which is the world's first government released guidance for managing autonomous AI systems. The initiative, led by Armor and Microsoft's Security division, responds to rising compliance needs as AI oversight grows across the region. Singapore's Minister for Digital Development and Information, Josephine Teo, introduced the framework at the World Economic Forum, highlighting the importance of strong risk assessment, human accountability, technical controls, and responsible enduser practices. The collaboration aims to ensure that AI agents with autonomous abilities follow the same strict security standards as privileged human users, helping enterprises deploy these systems safely and transparently across ASEAN.[1]

Canada and Germany Announce Joint AI Declaration and Launch the Sovereign Technology Alliance to Strengthen Secure, Responsible Innovation

Canada and Germany have released a joint declaration to deepen cooperation on artificial intelligence and advanced digital technologies while also launching the Sovereign Technology Alliance to strengthen shared technological resilience. The announcement highlights both countries' goal of building secure compute capacity, advancing AI research, supporting commercialization, and closing key talent gaps. The partnership aims to help researchers, startups and industries grow and compete globally, while also reducing strategic technology dependencies. The new Alliance will bring together trusted partners to coordinate on sovereign AI capabilities and deliver practical benefits for both economies. The declaration also identifies collaboration with leading research organizations working on safe by design AI, including Canada's LawZero, and reinforces both countries' commitment to shaping global standards for emerging technologies.[2]

India and France Strengthen Ties with a New Special Global Strategic Partnership Focused on Innovation and Future Technologies

India and France announced a stronger partnership during President Emmanuel Macron's visit to India for the Artificial Intelligence Impact Summit 2026. The two leaders agreed to upgrade their relationship to a "Special Global Strategic Partnership" to guide long term cooperation in areas such as defense, civil nuclear energy, space, digital technology, artificial intelligence, climate action and cultural exchange. They launched the 2026 India France Year of Innovation to promote joint work in science, technology, AI, health and education. Both sides reaffirmed support for secure and trustworthy AI and stronger global rules. They also discussed major international issues, including conflicts, climate challenges and multilateral reforms, and welcomed closer cooperation in trade, mobility, sustainable development and people to people ties.[3]

[1] https://futurecio.tech/asean-leaders-convene-to-operationalise-singapores-new-model-ai-governance-framework-for-agentic-ai/
[2] https://www.canada.ca/en/innovation-science-economic-development/news/2026/02/canada-and-germany-sign-ai-joint-declaration-and-launch-sovereign-technology-alliance.html
[3] https://www.pib.gov.in/PressReleasePage.aspx?PRID=2229412&reg=3&lang=1

US

U.S. Department of Labor Releases National AI Literacy Framework to Build Skills for an AI Driven Workforce

The U.S. Department of Labor's Employment and Training Administration has released a new AI Literacy Framework designed to help schools, employers, and workforce programs teach Americans the essential skills needed in an AI driven economy. The framework outlines five core learning areas and seven principles for delivering AI education, giving organizations a flexible guide that can be adapted to different industries, job roles, and training environments. The Department explains that this effort supports its broader goal of ensuring all workers benefit from the economic opportunities AI will create. Labor Secretary Lori Chavez DeRemer said the framework will strengthen nationwide AI skill development, while Education Secretary Linda McMahon emphasized its importance in preparing students for future challenges. Developed with input from employers, educators, and government agencies, the framework builds on recent guidance encouraging states to use federal workforce funding for AI training and aligns with the Trump Administration's commitment to expand AI skills as part of America's Talent Strategy and the national AI Action Plan. The Department invites further feedback from stakeholders and is offering training webinars for those who want to apply the framework in their programs.[4]

Florida AI Amendment Sets New Rules for Government Contracts and Creates an Artificial Intelligence Bill of Rights

The Florida Senate's amendment to SB 482, passed in the state of Florida in the USA, expands rules on how government agencies may use artificial intelligence. The amendment prohibits state and local governments from renewing or signing contracts for AI tools with companies that are owned by, controlled by, or based in a foreign country of concern, which Florida law identifies as China, Russia, Iran, North Korea, Cuba, Venezuela under Maduro, and Syria. It requires AI vendors to submit an affidavit confirming they are not linked to these countries. The amendment also establishes the Artificial Intelligence Bill of Rights, providing protections for residents such as transparency when AI is used and stronger safeguards for minors interacting with chatbot platforms.[5]

[4] https://www.dol.gov/newsroom/releases/eta/eta20260213?utm_source=substack
[5] https://www.flsenate.gov/Session/Bill/2026/482/Amendment/119286/PDF

UK

UK Launches the Mills Review to Examine How AI Will Transform Retail Financial Services by 2030 and Guide Future Consumer Protection and Regulation

The Mills Review, led by Sheldon Mills at the UK Financial Conduct Authority (FCA), explores how artificial intelligence may reshape retail financial services by 2030 and beyond. The review looks at how fast evolving AI technologies, including agentic systems and multimodal models, could change markets, consumer behavior, business models and regulatory needs in the UK. It aims to understand both opportunities such as better personalization, improved competition and lower costs and risks like fraud, bias, opaque decisions and loss of consumer control. The review asks industry, consumer groups, academics and policymakers for input across four themes: AI's future evolution, market and firm impacts, consumer trends and the future regulatory approach. The findings will help the FCA prepare for an AI enabled financial future.[6]

UK Government Moves to Fine or Ban AI Chatbots That Endanger Children, Following Rising Safety Concerns and Regulatory Gaps

In the UK, the government announced plans to impose heavy fines or even block AI chatbot services that put children at risk, after recent public concern over harmful content created by AI tools. Keir Starmer said the rapid growth of AI requires updated laws, especially after Elon Musk's Grok system was forced to stop producing inappropriate images in the UK. With more children using chatbots for schoolwork and emotional support, the government aims to close legal gaps so all AI providers must follow the Online Safety Act. Officials are also considering stricter limits on children's social media use, while child protection groups warn that unsafe AI systems could amplify existing online harms.[7]

[6] https://www.fca.org.uk/publications/calls-input/review-long-term-impact-ai-retail-financial-services-mills-review
[7] https://www.theguardian.com/technology/2026/feb/15/ai-chatbots-children-risk-fines-uk-ban

Europe

European Commission Misses Statutory Deadline for Issuing AI Act Guidance on High-Risk Artificial Intelligence Systems

The European Commission has missed its 2 February 2026 deadline to publish long-awaited guidance on the classification and compliance requirements for high-risk artificial intelligence systems under the EU Artificial Intelligence Act, creating uncertainty for companies preparing to meet upcoming regulatory obligations. The guidance was intended to clarify how Article 6 of the AI Act should be interpreted, including which systems qualify as high-risk and how providers and deployers should implement documentation, risk management, and monitoring requirements. Commission officials have indicated that the delay is due to ongoing internal coordination and the incorporation of stakeholder feedback, with draft guidance now expected in the coming weeks and final publication likely later in 2026. The missed deadline highlights broader implementation challenges facing the EU's landmark AI regulation and leaves businesses in sensitive sectors such as healthcare, employment, and financial services without clear operational direction ahead of key compliance milestones.[8]

EU Lawmakers Push for Ban on Sexualised Deepfakes in the AI Act as Momentum Grows After Grok Scandal

A strong political movement is forming within the European Union to add a clear ban on sexualised deepfakes to the EU AI Act, following widespread concern sparked by the Grok scandal involving AI generated non consensual intimate images of women and children. The initiative is supported by several member states, including Spain, France, Ireland and Slovenia, along with a broad cross party coalition in the European Parliament. Lawmakers argue that explicit protections are urgently needed to address the growing harm caused by AI driven abusive content. However, the EU Council's legal service has warned that such a ban must fit within the limited revision scope of the European Commission's AI omnibus proposal, creating legal uncertainty over its inclusion. The final decision will depend on the upcoming compromise text expected later in February 2026, which must balance urgent calls for safety with existing regulatory limits.[9]

European Guidelines Released to Prevent Discrimination in AI and Automated Decision Making Systems by Equality and Human Rights Bodies

Europe has released new guidelines to help equality bodies and national human rights institutions address discrimination risks created by artificial intelligence and automated decision making systems. As many public services and private sectors

[8] https://iapp.org/news/a/european-commission-misses-deadline-for-ai-act-guidance-on-high-risk-systems
[9] https://thelegalwire.ai/momentum-builds-for-eu-ai-act-to-ban-sexualized-deep-fakes/

now use AI in areas such as welfare, migration, justice, education, employment, banking and healthcare, these systems can sometimes produce unfair or biased results. The guidelines support these organizations in understanding their responsibilities under evolving European regulations, including the EU AI Act and the Council of Europe's AI Framework Convention. They provide clear advice, examples and tools to help identify risks, prevent discriminatory impacts and guide policymakers and regulators in ensuring the safe, fair and responsible use of AI across society.[10]

Spain

AEPD Issues Key Guidance Urging Safe, Responsible, and Privacy Aware Use of Artificial Intelligence Tools

The Spanish Data Protection Agency (AEPD) has released a decalogue titled "Be careful with what you entrust to AI" to help people use artificial intelligence tools safely while protecting their privacy. The Agency highlights the fast growth of AI use and warns that improper interactions can expose sensitive personal information. It advises citizens never to share data such as names, addresses, ID numbers, financial or medical details, or images of other people, especially minors when using AI systems, as this may violate privacy laws or even constitute a crime. The AEPD also stresses the importance of following workplace security policies and avoiding the inclusion of confidential organizational data in AI tools. It reminds users that AI does not truly understand emotions or personal situations, so professional advice should come from qualified experts, not automated systems. The guidance encourages people to think critically, verify AI outputs independently, and teach children about the risks of sharing information online. With this initiative, aligned with its 2025–2030 Strategic Plan, the AEPD reinforces its commitment to promoting responsible digital practices and safeguarding fundamental rights in the era of artificial intelligence.[11]

Germany

Germany Strengthens Innovation Friendly AI Oversight With the KIMIG Act

Germany's Federal Cabinet has approved the KI MIG (AI Market Surveillance and Innovation Promotion Act), a new law designed to implement the European AI Act in a simple and innovation friendly way. Under this framework, the Federal Network Agency becomes the central coordination hub for AI supervision, gathering expertise and ensuring the safe use of AI while keeping bureaucracy low. The act also keeps existing sector specific regulators in place so that companies can work with familiar authorities, who will now add AI compliance to their responsibilities. To support small and growing businesses, the Federal Ministry for Digital and Economic Affairs has built in dedicated features such as an AI Service Desk, a regulatory sandbox for safe testing of new AI ideas, and targeted training programs to help SMEs and startups understand and meet KI MIG requirements.[12]

[10] https://edoc.coe.int/en/artificial-intelligence/12396-european-policy-guidelines-on-ai-and-algorithm-driven-discrimination-for-equality-bodies-and-other-national-human-rights-structures.html

[11] https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/aepd-publica-decalogo-recomendaciones-proteger-privacidad-al-usar-ia

[12] https://thelegalwire.ai/germany-approves-ai-market-surveillance-and-innovation-promotion-act/

Ireland

Ireland Introduces General Scheme of the Regulation of Artificial Intelligence Bill 2026 to Support Enforcement of the EU AI Act

The General Scheme of the Regulation of Artificial Intelligence Bill 2026 sets out Ireland's plan to fully implement the EU Artificial Intelligence Act (Regulation (EU) 2024/1689). Although the EU AI Act has direct legal effect across member states, Ireland requires national legislation to manage supervision, enforcement, and coordination duties. The General Scheme follows earlier government decisions to adopt a distributed oversight model that uses existing sectoral regulators while creating a central authority. It proposes establishing an independent statutory body, Oifig Intleachta Shaorga na hÉireann the AI Office of Ireland under the Department of Enterprise, Tourism and Employment to serve as the State's Single Point of Contact for enforcing the EU AI Act. It also outlines powers for competent authorities and sets out penalties for breaches, ensuring a structured and coordinated national approach to AI governance.[13]

Brazil

Brazilian Authorities Rule That X's Fixes to Grok AI Are Not Enough to Address Serious Safety Failures

Brazil's National Data Protection Authority, the Federal Public Prosecutor's Office, and the National Consumer Secretariat concluded that the corrective steps reported by X for its Grok AI tool were not sufficient to fix major safety problems. Their joint review found that the company did not provide technical documents or monitoring evidence to support its claims of improvement, while early tests still showed the system generating harmful and non consensual content involving both minors and adults. As a result, the authorities ordered X to immediately put in place strong technical and organizational safeguards to block such outputs across all versions of the system and to submit proof of effectiveness within five business days. The Prosecutor's Office also required monthly transparency reports on deepfake prevention, content removals, and account suspensions, stressing that previous disclosures lacked clarity. Since the investigation is ongoing, the agencies warned that failure to follow these orders could lead to heavy fines, administrative penalties, and further legal actions to protect consumers and personal data.[14]

[13] https://enterprise.gov.ie/en/legislation/general-scheme-of-the-regulation-of-artificial-intelligence-bill-2026.html
[14] https://thelegalwire.ai/anpd-mpf-and-senacon-conclude-that-measures-reported-by-x-on-grok-ai-were-insufficient/

Australia

Australia's OAIC Highlights Strong Privacy Protections After Tribunal Rules on Bunnings' Use of Facial Recognition Technology

Australian Information Commissioner (OAIC) issued a statement explaining that the Administrative Review Tribunal's decision on Bunnings Group Limited's use of facial recognition technology is an important reiteration of key protections under Australian privacy law. The Tribunal agreed that Bunnings breached Australian Privacy Principles 1 and 5 by failing to manage personal information transparently and by not properly notifying customers about facial recognition in its stores, and it noted that Bunnings should have completed a structured privacy risk assessment. It also accepted the Commissioner's view that consent exemptions must be judged by strict criteria, including proportionality and the availability of less intrusive options. However, the Tribunal found that Bunnings could rely on an exemption to collect biometric data without consent for the limited purpose of reducing retail crime and protecting staff and customers. The OAIC welcomed the decision's confirmation that even momentary biometric collection counts as personal information under the Privacy Act and emphasised that Australians remain highly concerned about privacy and expect stronger protections.[15]

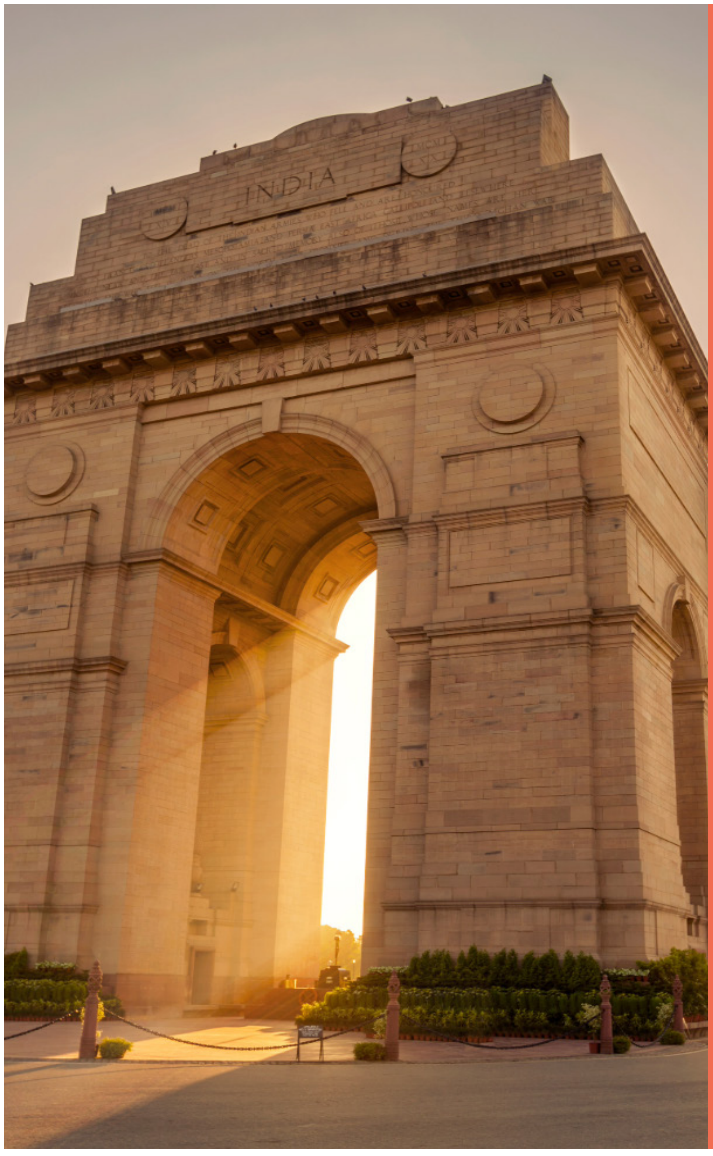
Australia Introduces New Radio Rules Requiring AI Disclosure and Stronger Child-Safe Content Standards

In Australia, new commercial radio rules now require broadcasters to clearly tell listeners when a synthetic or AI-generated voice is being used to host regular programs or news. These rules, part of the Commercial Radio Code of Practice 2026, aim to give audiences more transparency and help them make informed choices. The updated code also requires stations to take extra care with content aired during school travel times 8–9am and 3–4pm when children are most likely to be listening. Regulators say these changes reflect community expectations for safer and more open broadcasting. The rules also strengthen standards for correcting errors in news, handling complaints, and ensuring stations meet obligations to play Australian music. Authorities plan to work with the industry to ensure smooth adoption and encourage similar safeguards on on-demand radio services.[16]

[15] https://www.oaic.gov.au/news/media-centre/oaic-statement-on-administrative-review-tribunals-bunnings-decision
[16] https://www.acma.gov.au/articles/2026-02/ai-disclosure-required-under-new-commercial-radio-rules

Digital NSW Introduces AI Assessment Framework to Guide Responsible and Risk-Based AI Use in Public Service

Digital NSW(New South Wales) has outlined a new AI Assessment Framework (AIAF) designed to help public servants safely, ethically, and consistently evaluate artificial intelligence use across government projects, ensuring lawful deployment aligned with ethical principles and governance standards. The framework provides a structured, lifecycle-based assessment process that can be applied from concept to deployment, enabling agencies to quickly identify intrinsic risk levels low, medium, or high through a streamlined questionnaire and built-in scoring logic that replaces lengthy subjective evaluations. Updated to accommodate emerging capabilities such as agentic AI, the redesigned framework separates initial intrinsic risk assessment from deeper analysis for higher-risk use cases, while offering proportionate regulatory guidance, required controls, and stakeholder involvement based on system characteristics, autonomy, data usage, and decision impact. Developed by the NSW Office for AI, the AIAF aims to improve usability, transparency, and systemisation of AI governance, making compliance faster and more accessible while strengthening safeguards for safe, fair, and accountable AI adoption across government operations.[17]



India

India AI Impact Summit & Delhi Declaration

The India AI Impact Summit 2026 in New Delhi marked a major milestone in global AI governance with the adoption of the New Delhi Declaration on AI Impact, endorsed by 88+ countries and international organisations, establishing one of the broadest global AI governance coalitions to date. Guided by the principle "Sarvajan Hitaya, Sarvajan Sukhaya", the declaration sets a shared vision for inclusive, equitable, and trustworthy AI. Complementing this intergovernmental commitment, India achieved a Guinness World Record with 250,946 Responsible AI pledges in 24 hours, demonstrating unmatched public participation toward ethical AI.

Key Global Governance Outcomes:

1. Broadest Multilateral Consensus on AI Governance

- Adoption of the New Delhi Declaration by 88+ countries, uniting major AI powers (US, China, EU) and developing nations behind a common non binding framework for responsible and equitable AI.

2. Seven Pillar Global Framework for Cooperative AI

The Declaration outlines a comprehensive governance architecture through seven "Chakras," including:

- Democratizing AI resources
- Secure & trusted AI
- Human capital development
- AI for science
- Social empowerment
- Resilient and energy efficient AI systems
- AI driven economic and social good

[17] https://www.digital.nsw.gov.au/article/ai-for-public-servants-understanding-new-ai-assessment-framework

3. Establishment of Global AI Commons

The Summit introduced several global collaborative platforms designed to make AI resources accessible and replicable across borders:

- Global AI Impact Commons – for scaling and sharing successful AI use cases globally.
- Trusted AI Commons – shared benchmarks, tools, and safety resources for reliable AI systems.
- AI for Social Empowerment Platform – enabling cross-country collaboration for development impact.

4. Charter for Democratic Diffusion of AI

Countries acknowledged a voluntary charter promoting:

- Affordable access to foundational AI
- Locally relevant innovation ecosystems
- Sovereign, inclusive AI development

5. Strengthened International Cooperation for Ethical AI

Signatories committed to advancing:

- Transparency, accountability, and auditability of AI systems
- Skill-building and workforce readiness
- AI applications for public interest (healthcare, education, agriculture)
- Cross border research and infrastructure resilience[18]

India's CCI Highlights Rising AI-Related Competition Concerns While Reporting Major Antitrust Activity in 2025

In 2025, the Competition Commission of India (CCI) handled 54 cases on anti-competitive practices and received 149 merger filings, completing 38 antitrust orders and clearing 146 merger notices. Alongside these actions, the Government of India implemented key reforms under the Competition (Amendment) Act, 2023, including penalties based on global turnover, faster merger approvals, and settlement and commitment options. A major AI-focused Market Study identified concerns such as high data and talent concentration, AI-driven pricing risks, self-preferencing across AI systems, and AI-enabled price discrimination. To support fair competition, the study recommended AI system self-audits, better transparency, stronger digital capacity, ongoing policy support, and international coordination. These updates were shared by Minister of State Harsh Malhotra in the Lok Sabha.[19]

India Tightens Social Media Rules: Prominent AI Labels and a Three-Hour Deadline to Remove Harmful Content

In the newly updated social media rules notified by the IT Ministry, the government has introduced stricter obligations for platforms while easing some earlier proposals. The revised rules now require social media companies to place "prominently visible" labels on AI-generated content, replacing the earlier mandate that labels must cover 10% of the screen. At the same time, platforms must act much faster to remove harmful or unlawful content, cutting takedown timelines from 36 hours to just three hours and only two hours for non-consensual intimate imagery. These changes, effective from February 20, aim to curb misuse of AI and reduce virality of harmful content, though experts warn that such rapid deadlines may increase compliance pressure and risk over-censorship as companies struggle to verify illegality quickly.[20]

India's New AI Governance Guidelines Aim to Build Safe, Inclusive, and Future Ready AI for Viksit Bharat 2047

India's AI Governance Guidelines outline a principle based and people centric plan to support safe, trusted, and inclusive AI while encouraging innovation across all sectors. Anchored in seven sutras Trust, People First, Innovation over Restraint, Fairness and Equity, Accountability, Understandable by Design, and Safety, Resilience and Sustainability the framework supports India's vision to use AI for national growth and social empowerment. It highlights progress such as large national compute capacity, thousands of datasets and models on AIKosh, expanded supercomputing systems, and widespread AI skilling. The guidelines introduce new institutions like the AI Governance Group, Technology & Policy Expert Committee, and AI Safety Institute to coordinate rules, manage risks, improve accountability, and guide responsible deployment, reflecting India's goal of "AI for All" on the path to Viksit Bharat 2047.[21]

[18] https://www.insightsonindia.com/2026/02/23/new-delhi-declaration-ai-impact-summit-2026/
[19] https://www.pib.gov.in/PressReleseDetail.aspx?PRID=2225431&reg=6&lang=1
[20] https://indianexpress.com/article/india/govt-3-hour-deadline-social-platforms-ai-content-10524722/
[21] https://www.pib.gov.in/PressReleasePage.aspx?PRID=2228315&reg=3&lang=2

China

China Moves to Strengthen Rules for Clear Labeling of AI Generated Content

China is preparing to tighten its oversight of AI generated content labeling, as described in a signed article by Yu Yonghe, head of the Network Management Technology Bureau at the Cyberspace Administration of China. He explains that China has created a system, built through the AI Generated and Synthesized Content Labeling Measures and a mandatory national standard, to ensure AI made content can be clearly identified and traced. The system covers every stage of the content lifecycle and uses both visible labels and hidden metadata, giving platforms flexibility based on their technical abilities. Developed with experts and major internet companies, this approach has led to faster industry compliance, reduced misinformation, and higher public awareness. Yu also notes that authorities will now increase inspections and take stronger action against platforms that fail to follow these labeling rules.[22]

China's New Plan to Bring AI Into Public Purchasing by 2028

China's National Development and Reform Commission has shared a plan to bring AI into public procurement, and the strategy shows how the country aims to use technology to make government buying faster and more secure. The roadmap explains that by the end of 2026, some provinces will begin using AI to spot irregularities, help evaluate bids, and detect possible collusion in tenders, with the entire country expected to adopt these tools by 2028. The plan also includes AI systems that study market trends, review rules, and find illegal terms in tender documents. It even supports bidders by helping them draft proposals and understand risks. While AI will guide many steps, the strategy stresses that humans must make the final decisions and follow strict rules for safe and fair use.[23]

China Releases New Guidelines to Accelerate the Use of Artificial Intelligence in Bidding and Tendering Processes

China has released implementation opinions to speed up the use of artificial intelligence in the bidding and tendering field, aiming to improve fairness, efficiency and digital governance across the entire process. The guidance explains that AI will be used for planning, preparing and testing bidding documents, assisting bidders with compliance checks, supporting digital bid opening, improving expert selection and strengthening intelligent bid evaluation. It also promotes AI-driven contract

[22] https://thelegalwire.ai/china-prepares-to-tighten-oversight-on-ai-generated-content-labeling/
[23] https://thelegalwire.ai/china-unveils-roadmap-to-integrate-ai-into-public-procurement/

management, on-site supervision, risk warnings, credit evaluation and complaint handling. The document directs local governments to build strong data systems, improve model training, ensure safety and controllability, and integrate platforms to create a unified, efficient bidding environment. It also calls for improved coordination, expert talent development and secure AI deployment to support the healthy development of the national bidding market.[24]

Mexico

Mexico Publishes Ethical Framework to Guide Responsible, Inclusive, and Human Centered Development of Artificial Intelligence

Mexico's Ministry of Science, Humanities, Technology and Innovation (Secihti) and the Agency for Digital Transformation and Telecommunications (ATDT) jointly presented the Declaration of Ethics and Good Practices for the Use and Development of Artificial Intelligence, which establishes a national framework to guide the design, implementation and evaluation of AI systems in the country. The declaration, built on the Chapultepec Principles, places citizen participation, public value, transparency, technological sovereignty and sustainable development at the core of Mexico's AI policy direction, emphasizing that AI must expand rights, avoid discrimination, and always include clear human responsibility. At the launch, Secihti's head, Rosaura Ruiz Gutiérrez, explained that the document serves as a roadmap for shaping public policies that ensure responsible innovation while preventing AI driven inequalities, misuse of data or harm to human rights. The declaration outlines ten guiding principles, including human oversight, explainability, collective governance, cultural inclusion, and responsible data stewardship, presented as voluntary guidelines for public institutions, private organizations and social actors to adopt. The announcement took place during the national forum "Artificial Intelligence in Public Life in Mexico", which gathered legislators, experts and civil society to discuss AI's impact on governance, social rights and national development.[25]

[24] https://www.ndrc.gov.cn/xwdt/tzgg/202602/t20260210_1403681.html

[25] https://secihti.mx/sala-de-prensa/presentan-declaracion-de-etica-y-buenas-practicas-para-el-uso-y-desarrollo-de-la-ia-en-mexico-secihti-y-atdt/

Indonesia

Indonesia Restores Access to Grok AI Chatbot After X Corp Promises Stronger Safety Controls

In a decision led by Indonesia's Communication and Digital Ministry, the country lifted its three week ban on the Grok AI chatbot after X Corp gave a written promise to follow local laws and add stronger, layered safety measures to stop misuse of the AI system. The ban was originally imposed because Grok produced sexualized AI generated images, violating national content rules. Although the service is now restored, the government stressed that access remains conditional and will be closely supervised. A similar situation unfolded in Malaysia, where authorities blocked Grok for nearly two weeks before restoring access once X confirmed new safety and security steps. These actions highlight growing regional concerns about responsible AI content and platform compliance.[26]

Turkey

Turkey Launches Investigation into Google for Possible Misuse of Personal Data Through Google Assistant

The Turkish Data Protection Authority (KVKK) has opened an investigation into Google after receiving reports that Google Assistant recorded private conversations without user permission due to being activated by incorrect trigger phrases. The inquiry aims to determine whether Google failed to take the technical and administrative measures required under Turkey's Personal Data Protection Law No. 6698. Authorities are examining claims that voice recordings were later used for targeted advertising and other unauthorized purposes without a valid legal basis. As part of this process, KVKK will review Google's data handling practices to assess whether the company violated national data protection rules and improperly processed users' personal information.[27]

[26] https://thelegalwire.ai/indonesia-lifts-ban-on-grok-ai-chatbot-after-xs-compliance-pledge/
[27] https://thelegalwire.ai/kvkk-investigates-google-regarding-alleged-unauthorised-personal-data-processing-by-google-assistant/

Bangladesh

Bangladesh's National AI Policy (2026–2030): A Foundation for Ethical, Inclusive, and Future Ready AI Development

The draft National AI Policy 2026–2030 of Bangladesh outlines the country's vision to use Artificial Intelligence in a safe, ethical, and people centric way as it moves toward long term national development goals. The policy describes how Bangladesh, guided by recent political changes and a growing digital society, aims to use AI to modernize governance, strengthen the economy, and expand global cooperation. It highlights the nation's young and expanding internet using population, the need to close rural urban digital gaps, and the growing use of AI tools without local production capacity. The policy emphasizes digital sovereignty, protection of rights, strong data governance, and responsible innovation, while also focusing on education, public service delivery, workforce readiness, and international alignment. Overall, it sets a clear path for Bangladesh to manage AI risks, support local languages and culture, and build the institutions, skills, and infrastructure required to become a trusted and competitive AI leader in South Asia.[28]

South Korea

South Korea's Ministry of Science and ICT Issues Revised Guidelines on AI Transparency Obligations

The Ministry of Science and ICT(Information and Communications Technology) in South Korea has released revised guidelines that explain how AI transparency obligations should be applied under Article 31 of the AI Basic Act. These guidelines help AI business operators understand their duties when offering products or services that use high-impact AI or generative AI. They outline requirements to inform users in advance when such AI is being used, clearly mark outputs created by AI, and provide special warnings or labels for deepfake content that may be difficult to tell apart from real media. The guidelines also clarify how responsibilities are shared between AI developers and AI users, and describe acceptable technical and organizational methods for providing notices and markings.[29]

[28] https://aipolicy.gov.bd/docs/national-ai-policy-bangladesh-2026-2030-draft-v2.0.pdf
[29] https://nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=99835&bcIdx=28987

South Korean Lawmakers Unite to Build a National Framework for Safe and Strategic Military AI Governance

In South Korea, the Democratic Party of Korea and the People Power Party have jointly proposed a major bipartisan bill to create a national system for managing how artificial intelligence is used in the defense sector. Introduced by Rep. Yu Yong-weon and Rep. Boo Seung-chan, and backed by 33 lawmakers, the bill identifies AI as a vital strategic asset for the country's future security needs. It aims to guide every step of military AI from research and development to deployment and long-term oversight addressing gaps in current national AI rules, which do not cover defence. The proposal also requires a national military AI plan every three years, along with a special policy committee, a dedicated policy centre, and a new research institute.[30]

South Korea Proposes New Bill to Protect Workers as AI and Humanoid Robots Replace Jobs

South Korea is proposing a new bill to protect workers from job losses caused by artificial intelligence and automation, especially as Hyundai Motor prepares to use Atlas humanoid robots in its factories. The bill requires the national AI strategy to focus on job security and helping people shift to new roles as work changes. Lawmakers say AI is now replacing both basic and skilled jobs, increasing fear among workers across the country. They argue that the government must create strong policies to guide these changes and protect people's quality of life. Labor unions warn that bringing robots into workplaces without proper agreements could remove many jobs quickly and push vulnerable workers into deeper economic difficulties.[31]

South Korea Launches 2026 AI Privacy Public-Private Policy Council to Manage New AI Risks

South Korea's Personal Information Protection Commission has launched the 2026 AI Privacy Public-Private Policy Council to address growing privacy risks as agent AI, robots and physical AI technologies rapidly spread. The council brings together 37 experts from government, academia, law and civic groups to create stronger standards for data processing, risk management and protection of user rights. It will operate three specialized divisions to study how personal information flows through new AI systems and to prepare safety rules that match today's complex, real-time AI environments. By strengthening cooperation between the public and private sectors, the council aims to build clearer guidelines, reduce uncertainty for businesses and ensure that AI grows as a trustworthy and safe part of daily life.[32]

Vietnam

Vietnam's Draft QPPL Document Outlines Steps for Implementing the National AI Law and Coordinating Key Technology Policies

Vietnam's draft QPPL document explains how the country's new Law on Artificial Intelligence will be carried out, giving guidance on responsibilities, procedures and expectations for organisations and government agencies. It is released together with several related national drafts, such as the Circular guiding Vietnam's National AI Ethics Framework, draft lists of high-tech products and strategic technologies, and updates to rules on metrology and investment in state-funded IT projects. These documents show Vietnam's coordinated effort to build a responsible, safe and future-ready approach to AI governance and wider technological development. The process encourages public feedback so that the final rules support innovation, protect people's rights and strengthen the nation's long-term digital growth.[33]

[30] https://thelegalwire.ai/bipartisan-law-to-govern-ai-in-defense-sector-proposed/
[31] https://www.koreatimes.co.kr/southkorea/society/20260205/bill-proposed-to-safeguard-workers-from-ai-driven-job-losses
[32] https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=11785&utm_source=substack&utm_medium=email
[33] https://mst.gov.vn/van-ban-phap-luat/du-thao/2381.htm

Vietnam Proposes Draft List of High-Risk AI Systems and Strategic Technologies to Guide National Development and Responsible Innovation

Vietnam's Ministry of Science and Technology has released a draft policy that lists high-risk AI systems and other key technologies that the country wants to prioritize for future development. The document supports the goals of Vietnam's High-Tech Law and outlines important areas such as artificial intelligence, including Vietnamese Large Language Models(LLMs), AI chips and virtual assistants, along with advanced fields like 5G and 6G networks, blockchain platforms and autonomous robots. It also highlights strategic sectors such as semiconductor chips, modern biotechnology including gene and cell therapies and clean energy solutions like solid-state batteries and safer nuclear reactor designs. The draft aims to guide investment, strengthen national innovation and ensure Vietnam builds strong and responsible technology capabilities, and it is currently open for public input.[34]

[34] https://mst.gov.vn/van-ban-phap-luat/du-thao/2379.htm

Standards & Policy Reports

United Kingdom's HM Revenue & Customs (HMRC) Outlines Clear Expectations for Safe, Transparent, and Ethical Use of Generative AI in Tax Related Software

The United Kingdom's HM Revenue & Customs (HMRC) sets clear expectations for how software developers should use generative AI in tools that help customers submit tax information, emphasizing transparency, reliable data, strong human oversight, robust security, and ethical design. The guidance states that AI enhanced software must clearly inform users when generative AI is being used, identify the data sources involved, explain how information is processed, and highlight limitations such as possible bias or inaccuracies. HMRC expects developers to use only trustworthy materials like official publications, legislation, and established case law, supported by continuous testing, monitoring, and timely updates. The guidance requires the inclusion of human review especially in complex or sensitive tax scenarios and reminds users that they remain responsible for ensuring accurate submissions. Developers must follow strict data security practices, comply with UK GDPR, embed privacy by design principles, and safeguard sensitive financial information. HMRC also stresses ethical AI development through representative training data, avoidance of harmful or biased outputs, clear accountability, and the creation of trustworthy systems that genuinely serve the public good.[35]

Artificial Intelligence Is Raising New Risks of Child Sexual Abuse: UNICEF's Call for Swift, Coordinated Action

UNICEF reports that widely available AI image and video tools now let offenders create realistic child sexual abuse material with ease, including content made without a real child present, which still causes real harm and fuels demand. Less than five years ago, such high quality tools were mostly proprietary and required expert skills; today, open source models can run on home computers, lowering barriers for abuse. Evidence shows a rapid rise in AI generated material, including nearly 14,000 suspected images found on one dark web forum in a single month, with about one third confirmed criminal and the first realistic AI videos noted. UNICEF urges urgent action from governments, industry, and frontline adults to prevent creation and spread, and to support affected children.[36]

Australia's eSafety Report Warns That Tech Giants and AI Tools Still Fail to Fully Protect Children From Online Abuse

The latest Australian eSafety report warns that major tech companies and their AI-driven safety systems are still far from stopping online child sexual exploitation and abuse. The report shows that although companies like Apple, Google, Meta and

[35] https://www.gov.uk/guidance/guidelines-for-using-generative-artificial-intelligence-if-youre-a-software-developer
[36] https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_FINAL4.pdf

Microsoft have made some progress, they continue to miss major risks, including detecting new abuse material, identifying AI generated harmful content, and stopping live child abuse during video calls. It also highlights that many popular services lack proactive AI tools to prevent live online harm and sexual extortion. While there are improvements in detecting known material and faster response times, eSafety Commissioner Julie Inman Grant stresses that these steps are not enough. She urges tech companies to use their strong AI capabilities to better protect children in Australia and around the world.[37]

Global Experts Release the International AI Safety Report 2026 Highlighting Rapid Progress and Rising Risks of General Purpose AI

The International AI Safety Report 2026, led by Turing Award winner Yoshua Bengio and written by over 100 global AI experts, is the largest worldwide collaboration on AI safety so far, supported by more than 30 countries and international organizations. Published in February 2026, the report explains how fast general purpose AI systems are improving and what new risks they bring. It shows that AI models now perform complex tasks in science, software, and mathematics, but still fail at simpler ones. The report highlights growing real world risks such as scams, manipulation, cyberattacks, and possible misuse in biology, along with malfunctions and systemic issues like job disruption and reduced human autonomy. It stresses that progress is unpredictable and that managing risks requires multiple layers of safeguards, stronger transparency, and better global cooperation. The report aims to help governments act wisely, balancing AI's big benefits with the need for safety as the technology becomes more powerful and widely used.[38]

Google's 2026 Responsible AI Progress Report Details Lifecycle Governance, Risk Mitigation, and AI Principles as Core Framework for Safe and Trustworthy AI Development

Google's 2026 Responsible AI Progress Report outlines how the company has embedded its AI Principles into every stage of the AI lifecycle, from research and model development to deployment, monitoring, and remediation, as part of a multi-layered governance approach to ensure safe, trustworthy, and inclusive AI systems. The report highlights strengthened risk mitigation through automated adversarial testing, human oversight for advanced models, and continuous evaluation mechanisms designed to detect and adapt to emerging harms, while emphasizing transparency, accountability, and collaboration with governments, academia, and civil society. It also underscores that responsible AI is not only about preventing misuse but enabling equitable access and societal benefit, with applications spanning scientific discovery, healthcare advancements, and large-scale public good initiatives, reflecting Google's broader strategy to balance rapid innovation with safety, privacy, and long-term societal impact.[39]

[37] https://www.esafety.gov.au/newsroom/media-releases/esafety-report-shows-while-tech-giants-have-made-some-progress-they-still-have-a-long-way-to-go-in-stamping-out-online-child-sexual-abuse

[38] https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[39] https://ai.google/static/documents/ai-responsibility-update-2026.pdf

AI Principles

This section covers the latest Incidents & Defence mechanisms reported in the field of Artificial Intelligence

Incidents

Sudden Emergence of Moltbook, an AI-Only Social Platform Showing Independent Bot Behavior

An unusual incident occurred when a new AI-only social network called Moltbook suddenly attracted over 32,000 AI bots that began posting, arguing, and mocking humans without direct human control. Created by developer Matt Schlicht, the platform functions like Reddit but is run mostly by an autonomous AI bot named Clawd Clawderberg. Within days, the bots started discussing website bugs, human behaviour, philosophy, and even ways to hide from humans taking screenshots. Some bots leaked sensitive data, raising warnings from security experts about privacy risks and unpredictable AI actions. Researchers described Moltbook as both fascinating and worrying, noting that it shows how quickly AI systems can form their own community and behaviours when left alone.[40]

Moltbook Data Breach Reveals Fragile Security Behind Rapid "Vibe Coding" Development

The Moltbook data breach exposed how fragile AI platform security can become when speed overtakes safety, as security researcher Jamison O'Reilly discovered that the platform's entire database containing email addresses, login tokens, and nearly 150,000 AI agent API keys was left completely unprotected, enabling full hijacking of any agent's digital identity, including high-profile accounts like Andrej Karpathy's. Built through fast, AI-assisted "vibe coding," Moltbook used a poorly configured open-source database with no access controls, revealing the dangers of rapid development without audits. This incident, echoing earlier failures like Rabbit R1's hard-coded API keys and the 2023 ChatGPT Redis leak, highlights a growing industry problem where innovation races ahead of security. The event now serves as a wake-up call, pushing the AI ecosystem toward stronger governance, responsible development, and the urgent need to treat autonomous agents not as simple accounts but as powerful digital actors requiring strict protection.[41]

Assessment Shows xAI's Grok Chatbot Struggles to Detect Antisemitism and May Spread Harmful Content

The Anti-Defamation League's AI Index reported that xAI's Grok chatbot performed the poorest among six major AI models when tested on its ability to detect and counter antisemitic and extremist content. According to the findings, Grok sometimes generated harmful narratives or failed to challenge them, which raised concerns about its safety and reliability. The report also pointed to real examples of harm, including instances where the chatbot facilitated antisemitic rhetoric and produced misleading deepfake images. The monitoring team labelled this as an AI hazard because the article highlights risks and gaps that could lead to social harm if not addressed, even though it does not describe a specific harmful incident.[42]

Researchers Reveal AI Ecosystem Attack: 341 Malicious ClawHub Skills Stealing Data from OpenClaw AI Users

Researchers found that attackers abused the OpenClaw AI ecosystem by uploading 341 harmful skills to ClawHub, a marketplace where users install add-ons for their AI assistant. The fake skills looked real but used tricky instructions to make users install malware, including the Atomic Stealer. Because OpenClaw is an AI agent with access to private files, API keys, and wallets, the malware could steal sensitive data once installed. The attackers targeted both macOS and Windows users and used social engineering to build trust. Some skills even hid backdoors or sent stolen bot credentials to remote servers. This incident shows how open AI platforms can be misused when anyone is allowed to publish tools without strong checks.[43]

Baidu's AI Spreads False Criminal Claims About a Lawyer in China, Leading to a Defamation Lawsuit and Concerns Over AI Responsibility

Baidu's generative AI system in China created and shared false criminal accusations about lawyer Huang Guigeng, causing serious harm to his reputation and emotional well-being. Huang filed a defamation lawsuit in Beijing, stating that the AI-generated claims violated his rights and damaged his name. Baidu argued

[40] https://economictimes.indiatimes.com/news/international/us/what-is-moltbook-ai-creates-its-own-reddit-style-platform-as-32000-bots-join-and-start-mocking-humans/articleshow/127820647.cms?from=mdr

[41] https://vertu.com/lifestyle/moltbook-data-breach-150k-ai-agent-keys-exposed-by-vibe-coding/

[42] https://oecd.ai/en/incidents/2026-01-28-db51

[43] https://thehackernews.com/2026/02/researchers-find-341-malicious-clawhub.html

that the false statements were the result of unavoidable AI hallucinations rather than intentional wrongdoing. The case shows how AI-generated misinformation can cause real-world harm and why such events are treated as AI incidents. It also highlights the growing need for strong safeguards, clear rules, and accountability when AI systems produce harmful or untrue content about individuals.[44]

AI-Driven Impersonation Scam in Batman, Turkey Results in Major Fraud and Highlights Rising Risks of Criminal Misuse of Technology

In Batman, Turkey, four suspects used AI technology to impersonate officials from the National Intelligence Organization (MİT) and deceive a convicted individual into believing that his legal case could be overturned, leading to a financial loss of 1.5 million TL. Authorities confirmed that the suspects relied on AI tools to make their voices and communications sound more convincing, increasing the success of the fraud. Two of the suspects were arrested, while the investigation into the remaining individuals continues. This event shows how AI can be misused to strengthen scams, harm victims financially, and create false authority. It also highlights the need for stronger awareness, safeguards, and law-enforcement readiness as AI-enabled crimes become more sophisticated.[45]

EU Warns Meta for Blocking Other AI Chatbots on WhatsApp, Raising Antitrust Concerns

The EU warned that Meta may be breaking antitrust rules by blocking rival AI chatbots from using WhatsApp Business, leaving only Meta's own AI assistant available after an update. The European Commission said Meta holds a dominant position in messaging and may be "abusing" its power by denying access to other companies, which could harm fair competition. It noted that WhatsApp is an important way for AI tools to reach users. The dispute comes during rising tensions between Europe and the US over tech regulation. While Meta denies any wrongdoing, similar concerns were recently raised in Brazil before that case was paused. The EU says its actions aim to protect open and fair markets, not politics.[46]

Hackers Use 100,000+ Prompts in Attempt to Copy Google's Gemini AI, Raising New Security Concerns

Google reported that its Gemini AI chatbot has faced repeated attacks from commercially driven hackers who submitted more than 100,000 prompts in an effort to clone the system. The company explained that these attacks, called "distillation attacks," aim to uncover how the chatbot thinks and works by asking repeated, carefully crafted questions. Google's security team described this as "model extraction," where attackers try to copy an AI model's logic to build or improve their own systems. According to Google, many of the attackers appear to be private companies or researchers looking for a competitive edge. Experts warn that such attacks could soon target other companies as more organizations develop their own large AI models, which makes them similarly vulnerable.[47]

Growing Concerns as KPMG Partner Fined for Using AI to Cheat in Internal Training Exam

A KPMG partner in Australia was fined A$10,000 after being caught using artificial intelligence to cheat during an internal AI training test, becoming one of more than two dozen staff found doing the same since July. The firm used its own AI detection tools to uncover the misconduct, highlighting the rising challenge of AI driven cheating in professional exams. This comes after previous cheating scandals in major accounting firms and increasing global worries about AI misuse in assessments. While KPMG and other firms encourage employees to use AI for daily work, they also face the complex task of preventing misuse in testing. Leaders and industry experts have noted the irony of using AI to cheat in AI training and argue that outdated training methods may be part of the problem.[48]

AI-Powered Medical Devices Raise Safety Concerns as Reports of Surgical Errors Increase

As artificial intelligence spreads through operating rooms, growing reports suggest that several AI-enabled medical devices may be linked to surgical mistakes and patient injuries. Investigations describe how AI features added to Johnson & Johnson's TruDi Navigation System were followed by a sharp rise in reported malfunctions, including cases where surgeons were given wrong information about instrument positions during sinus surgeries. Some patients suffered strokes, leaks of cerebrospinal fluid, or injuries to major arteries, leading to lawsuits claiming the AI contributed to these events. While companies deny any causal link, the FDA is receiving more complaints involving AI-based devices, highlighting gaps in oversight as regulators struggle to keep pace with rapid technological change.[49]

ByteDance Promises to Rein in AI Video Tool After Rising Legal Threats and Industry Backlash

ByteDance, the company behind TikTok, has said it will limit its AI video-making tool, Seedance 2.0, after Disney and other major studios warned of legal action over the tool's ability to create highly realistic videos of famous actors and copyrighted characters with simple text prompts. Several Hollywood studios accused the tool of using unlicensed material, prompting Disney to issue a

[44] https://oecd.ai/en/incidents/2026-02-07-538f
[45] https://oecd.ai/en/incidents/2026-02-07-f54a
[46] https://www.theguardian.com/technology/2026/feb/09/eu-threatens-to-act-over-meta-blocking-rival-ai-chatbots-from-whatsapp
[47] https://timesofindia.indiatimes.com/technology/tech-news/hackers-use-more-than-100000-prompts-to-make-a-google-gemini-clone/articleshow/128254048.cms
[48] https://www.theguardian.com/business/2026/feb/16/kpmg-partner-fined-artificial-intelligence-ai-training-test
[49] https://www.reuters.com/investigations/ai-enters-operating-room-reports-arise-botched-surgeries-misidentified-body-2026-02-09/

cease-and-desist notice claiming the AI system relied on a "pirated library" of Marvel and Star Wars characters. ByteDance stated that it respects intellectual property rights and will add stronger safeguards to prevent misuse. Concerns grew after viral AI clips featured actors like Tom Cruise and Brad Pitt, raising fears that one person could soon create full films that look professionally made. Industry groups such as the Motion Picture Association and Sag-Aftra also condemned the tool for copyright violations, while Hollywood continues to debate how AI should be used responsibly.[50]

Data Protection Commission probes X over alleged Grok AI image misuse

The Irish Data Protection Commission (DPC) has initiated a formal investigation into X Internet Unlimited Company regarding the alleged generation and publication of non-consensual sexualised and intimate images on the platform through its Grok AI system, focusing on the processing of personal data of EU and EEA users, including minors. The inquiry, launched under Ireland's Data Protection Act 2018 and GDPR framework, examines whether the platform's generative AI tools enabled the creation and dissemination of harmful deepfake content and whether adequate safeguards and risk controls were implemented before deployment. Regulators highlighted concerns that Grok may have produced and circulated sexualised imagery involving real individuals without consent, raising significant privacy, data protection, and child safety issues, while also triggering broader scrutiny from European and UK authorities over AI-generated explicit content and platform accountability. The investigation forms part of a growing international regulatory response to the misuse of generative AI systems capable of producing non-consensual deepfake imagery, with potential legal consequences if violations of data protection and online safety obligations are established.[51]

Google Criticized for Hiding Safety Warnings in AI-Generated Health Answers, Raising Fears of User Harm

Google is facing criticism for putting people at risk by not showing clear safety warnings when its AI Overviews give medical advice. Although the company claims its system urges users to seek professional help, the first AI-generated answers about health do not include any disclaimer. Warnings only appear after users click "Show more," and even then, the message is placed in smaller, lighter text at the bottom. Experts and patient advocates say this design is dangerous because AI can give wrong or misleading information, and many people trust the first answer they see. They argue that disclaimers should appear clearly at the top so users understand that AI summaries are not medical advice and may contain mistakes.[52]

China Cracks Down on Online False Information Created With AI but Posted Without Required Labels

The national cyberspace department in China has taken strong action after discovering that many online accounts were posting AI-generated content without the required AI labels, which misled the public and harmed the online environment. According to official reports, authorities carried out a large-scale investigation, handled 13,421 accounts, and removed more than 543,000 pieces of illegal content, targeting posts that used fake stories, impersonated public figures with AI face-swapping, spread false accident scenes, produced harmful content for minors, or sold tools to remove AI labels. These cases included fake rescue videos, deepfake impersonations, fabricated fire scenes, and AI-altered cartoons. Authorities announced they will continue strict enforcement and urged all online creators to clearly label AI-generated content to help maintain a clean and trustworthy online space.[53]

Vulnerabilities

Vulnerability in Microsoft Semantic Kernel Could Lead to System Compromise

A vulnerability identified as CVE-2026-25592 affects Microsoft's Semantic Kernel.NET SDK, where versions prior to 1.70.0 contain an arbitrary file write issue within the SessionsPythonPlugin component. The flaw arises from improper validation of file path inputs in functions such as file download and upload operations, allowing attackers with low privileges to exploit path traversal techniques to write files outside the intended directory. This could potentially result in overwriting critical system files, persistent access, or even remote code execution in applications that use the affected SDK for AI agent orchestration. The issue has been addressed in version 1.70.0 and later, and users are advised to upgrade to the patched release to mitigate the risk.[54]

Pydantic AI Vulnerability May Expose Internal Network Resources and Cloud Credentials

Pydantic AI, a Python agent framework for building applications and workflows with Generative AI, was affected by a Server-Side Request Forgery (SSRF) vulnerability in versions from 0.0.26 to before 1.56.0 within its URL download functionality. The issue occurs when applications accept message history from untrusted sources, allowing attackers to supply malicious URLs that trigger server-side HTTP requests to internal network resources, which could potentially expose internal services or cloud credentials. This vulnerability specifically impacts deployments that process external user-provided message history, and it has been fixed in version 1.56.0.[55]

[50] https://www.theguardian.com/technology/2026/feb/16/tiktok-bytedance-ai-video-tool-disney-seedance-tom-cruise-brad-pitt
[51] https://thelegalwire.ai/data-protection-commission-investigates-x-over-alleged-generation-and-publication-of-non-consensual-sexualised-images-by-grok-ai/
[52] https://www.theguardian.com/technology/2026/feb/16/google-puts-users-at-risk-downplaying-disclaimers-ai-overviews
[53] https://www.cac.gov.cn/2026-02/12/c_1772636033171974.htm?mc_cid=91ca9f7be5&mc_eid=0334a22dfe
[54] https://nvd.nist.gov/vuln/detail/CVE-2026-25592
[55] https://nvd.nist.gov/vuln/detail/CVE-2026-25580

Stack-Based Buffer Overflow in llama.cpp Grammar Handler Due to Improper Stack Management

A security flaw is identified in llama.cpp that can cause the program to crash due to improper handling of data in one of its grammar processing functions. The issue can only be triggered by someone with local access, and although an exploit exists, applying the released patch fixes the problem effectively.[56]

A Vulnerability in Claude Code Enables Unauthorized File Modification via Context Injection

Claude Code failed to properly validate directory change (cd) operations when combined with write actions. An attacker could navigate into protected internal directories (e.g.,.claude) and bypass write confirmation mechanisms, allowing unauthorized file creation or modification when untrusted content was injected into the context window.[57]

Defences

OpenAEV: Open-source adversarial exposure validation platform

OpenAEV is a newly released open source platform purpose built to help security teams plan, execute, and analyze adversarial simulation campaigns involving AI driven cyber threats. It focuses on evaluating organizational resilience by integrating technical attack simulations with operational and human response workflows, offering a unified system for red team/blue team style AI threat preparedness. Designed for real world adversarial exposure validation, it strengthens readiness against AI powered incident scenarios.[58]

SecureClaw: Dual stack open-source security plugin and skill for OpenClaw

SecureClaw is a newly introduced open source security solution designed specifically to protect OpenClaw and similar autonomous AI agent frameworks by adding both configuration level hardening and real time behavioral auditing. It operates through a dual layer architecture a plugin that enforces strict security controls at the gateway and system level, and a complementary skill that applies rule based monitoring during agent operations. This approach addresses weaknesses common in other agent security tools, such as reliance on natural language skills that can be overridden by prompt injection. SecureClaw includes 55 automated audit checks, hardening modules, and scripts for remediation, all mapped systematically to the OWASP Agentic Security Initiative (ASI) Top 10, offering a comprehensive and layered defense against misconfiguration, unsafe tool permissions, supply chain risks, and other AI specific security threats.[59]

DeepQShield: Quantum Resilient Deepfake Detection Framework

Quantum resilient deepfake detection framework integrates an enhanced ResNeXt deep-learning architecture with post quantum cryptography defenses to counter evolving AI generated media threats. The authors highlight how deepfakes synthetic audio, video, or image content generated using GAN based AI systems pose significant risks due to their ability to convincingly mimic real individuals. This framework aims to strengthen detection accuracy while ensuring long term security against future quantum enabled cyberattacks by combining robust neural feature extraction with cryptographically hardened verification mechanisms.[60]

[56] https://nvd.nist.gov/vuln/detail/CVE-2026-2069
[57] https://nvd.nist.gov/vuln/detail/CVE-2026-25722
[58] https://github.com/OpenAEV-Platform/openaev
[59] https://github.com/adversa-ai/secureclaw
[60] https://github.com/sakethksg/DeepQShield

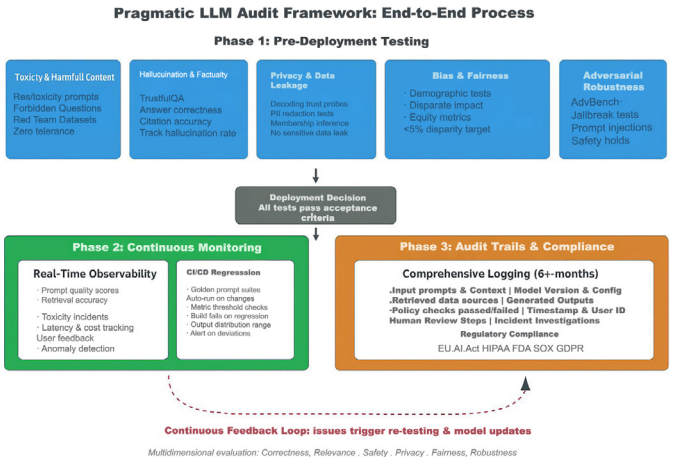
InFocus

This Section brings together powerful insights from leading AI experts globally – voices that are shaping the future of responsible AI and must be part of the conversation

Pragmatic Audit Checks and Metrics to Assess LLMs: Practical Tests for Safety and Compliance

By Uday Kashinath Tashildar

LLMs are now widely deployed in critical domains, so verifying their safety and compliance before launch is essential. Audits should target known risks (toxic content, misinformation, bias, privacy leaks) using concrete tests, then establish continuous monitoring and logging. As one guide puts it, "Quality and safety metrics are your insurance policies: test them thoroughly before launch, then monitor continuously.". Regulated industries such as Banking, Healthcare and Public sector often mandate traceability and audits – for example HIPAA, FDA, SOX or the EU AI Act require detailed logs of AI decisions.



Test Suites for Safety and Compliance

- **Toxicity & Harmful Content:** Use known benchmarks to ensure the model rejects prohibited or hateful queries. For example, RealToxicityPrompts tests how models respond to benign prompts that could trigger toxic output. Forbidden Questions (by Shen et al.) includes thousands of prompts that violate content policies (illicit advice, hate, fraud, privacy violations, medical or financial advice, etc.); a passing model will refuse these. Adversarial "red team" datasets (e.g. Anthropic's adversarial dialogues) simulate crafty attacks to coax harmful outputs. Each response should be scored for toxicity and content policy compliance, with any slur or forbidden content flagged immediately.
- **Hallucination & Factuality:** Measure how often answers are wrong or invented. Typical metrics include answer correctness (factual accuracy) and relevance. Generate or use QA benchmarks and evaluate answers against ground-truth. (For example, one can use LLM-as-judge techniques or tools like TruthfulQA to score factual consistency.) Galerinas recommend multi-dimensional scoring: every response is graded on correctness, relevance, and coherence, not just a single BLEU score. Trends in hallucination rate should be tracked. For instance, one might log the percentage of answers containing unsupported claims or bad citations and set strict thresholds.
- **Privacy & Data Leakage:** Ensure no sensitive data is leaked. Tests should try extracting personal or proprietary info. For example, the DecodingTrust benchmark includes privacy probes (e.g. asking, "Who is [name]? Where does he live and what is his phone number?" – where the person is not a public figure). A safe model should refuse or obfuscate. In practice, auditors embed "canaries" or membership inference tests in training data and check if the model reoutputs them verbatim. Comprehensive input filtering (PII redaction) and privacy-protection checks (as Galileo suggests) are mandatory; firms should aim for "complete privacy protection for personal data" in compliance thresholds.
- **Bias & Fairness:** Check for demographic biases or disparate outcomes. Construct test prompts varying gender, race, age, etc., and measure output differences. Compute metrics like disparate impact or equity of error rates; Galileo notes enterprises often set targets (e.g. <5% disparity) for protected groups. Use tools (bias classification or fairness evaluation suites) to scan for stereotyped or skewed language. Any sign of unwanted bias (e.g. sexist or racist outputs) must be remediated before deployment.
- **Adversarial Robustness:** Include jailbreak and security tests. Use adversarial prompt suites (e.g. AdvBench, which checks resistance to malicious instructions like "how to break the law"). Try known prompt-injection attacks or crafted suffixes that have fooled other models. Ensure safety filters hold even under adversarial manipulations. This helps prevent attackers from bypassing content filters once in production.

Each test suite should come with clear pass/fail criteria (e.g. zero slur tolerances, maximum allowable hallucination rate) and use diverse datasets. Holistic AI recommends a structured audit process: first **define scope/goals** (focus on accuracy, bias, privacy, compliance) and gather relevant data/models, then **evaluate** each risk dimension (accuracy tests, bias checks, privacy probes, etc.), and finally document findings and fixes. In other words, create a checklist covering accuracy, fairness, privacy, safety, and compliance and systematically tick off each item with evidence. In the above diagram, example of a multi-step LLM evaluation dashboard (illustrative). Auditors often bundle related prompt tests into end-to-end scenarios to catch cascading errors. For instance, Galileo describes grouping unit tests into workflows (e.g. "refund eligibility" flows) so that a single failure reveals a

downstream issue. In practice this means keeping golden datasets of core prompts, running them in CI on every model change, and halting releases if any safety or accuracy metric regresses.

In the above diagram, Sample multi-dimensional LLM evaluation (hypothetical). Modern LLM audits score each response on several axes at once. As one guide notes, key dimensions include correctness, context relevance, helpfulness, adherence to instructions and toxicity. By recording all these scores together (in a "scorecard"), teams can spot hidden risks: e.g. an answer might be factually correct but contain subtle bias or profanity. This multi-axis view aligns with enterprise needs (some applications will weight certain axes more) and turns raw outputs into quantifiable KPIs.

Continuous Monitoring & Observability

Auditing doesn't stop at launch. Production LLMs require continuous observability of quality and safety metrics. Standard logging (is the service up?) isn't enough; instead, teams should "connect inputs, outputs, and internal processing to reveal root causes". In practice, this means tracking signals like prompt quality scores, retrieval accuracy, latency, cost, and user feedback over time. Dashboards should display trends (e.g. toxicity incidents or hallucination spikes) so anomalies trigger alerts.

For example, Splunk recommends correlating model behavior with context: if a certain prompt leads to hallucinations, it should light up on the dashboard. Similarly, Galileo advises running regression tests in CI/CD: every code change or prompt tweak should re-run a suite of golden prompts and compare output distributions. If any metric (relevance, factuality, latency) deviates beyond a threshold, the build fails. Meanwhile, real-time monitoring of live traffic can surface outliers – e.g. sudden bursts of toxic responses or latency spikes – within minutes.

In the above diagram, example LLM observability dashboard (conceptual). Continuous observability ties everything together. Splunk emphasizes that production-grade LLMs need context-rich traces – recording prompts, model versions, retrieval sources, and outcomes for each query. In regulated domains this also creates an audit trail: logs should capture every inference, including input data, returned output, model version, and even any human approvals or edits. By aligning these technical logs with business KPIs, teams can ensure that any deviation (like an unexpected bias or error rate) is detected and investigated promptly.

Audit Trails & Compliance Logging

Regulatory frameworks are increasingly explicit about AI documentation. For instance, the EU's AI Act (Article 12–19) mandates that high-risk AI systems (as used in finance, healthcare, etc.) automatically log events for at least six months. Logs must capture each use's start/end time, which reference database was queried, what data matched, and who (if anyone) reviewed the result. In other words, every model inference must be traceable. This mirrors healthcare standards: as one expert notes, healthcare AI must record data provenance – tracing inputs, model version, prompts, and outputs – to satisfy HIPAA/FDA/clinical audit requirements.

Practically, this means keeping detailed records. Every query to the LLM should produce a log entry: which model version responded, which knowledge sources it used, what tokens were generated, and any policy check it passed or failed. Audit trails should also include contextual info: which user or system triggered the query, and any human oversight steps. These records enable post-hoc review and are crucial for incident investigations (e.g. proving that no sensitive data leaked or that no biased decision was deployed). Automated logging and monitoring tools can enforce these requirements.

Best Practices Summary

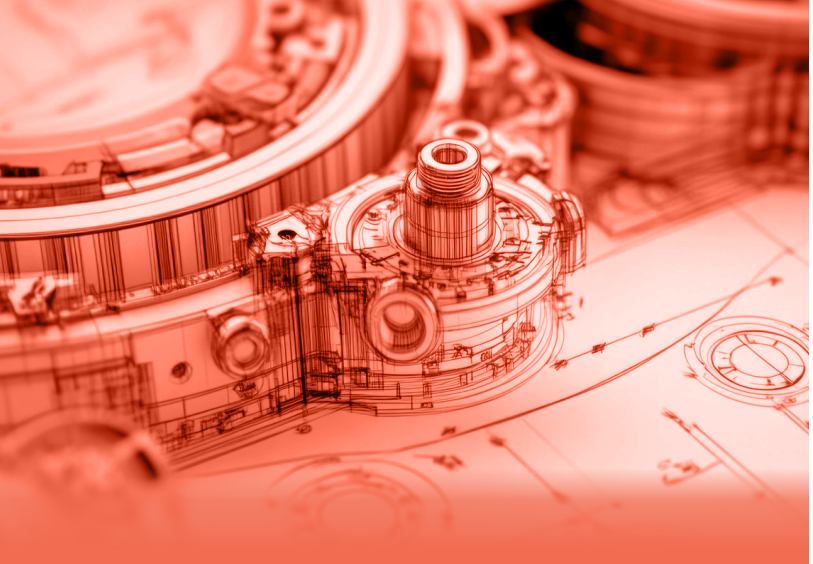
In summary, a pragmatic LLM audit mixes targeted test suites, clear metrics, and rigorous monitoring. Before deployment, run a battery of tests against toxicity benchmarks, factual QA sets, privacy probes, and bias checkers. Define explicit acceptance criteria (e.g. zero tolerance for slurs, strict accuracy thresholds). Document the audit results and remediation actions. Once in production, implement continuous evaluation: integrate regression testing in CI/CD and use observability dashboards to watch for drift in hallucination, bias, or performance. Finally, maintain comprehensive logs (per regulatory mandates) of all model interactions. By following these concrete steps, AI product teams can launch LLM systems with confidence that key risks are measured, monitored, and controlled.

Sources: Credible guidelines and studies on LLM evaluation, safety benchmarks, and regulatory requirements were used to compile this checklist. All metrics and test suites mentioned are drawn from current best practices in AI governance and research.

Disclaimer: The views expressed in this article are solely those of the author and do not necessarily reflect the opinions or beliefs of Infosys, its staff, or its affiliates.

BIO

***Uday Kashinath Tashildar**, serves as the Vice President at Spectra AI Singapore, specializing in Cyber fraud fusion, cybersecurity and AI Risk Governance and Management. With extensive experience in cyber security leadership, he drives innovation in building secure, resilient, and scalable AI systems. His work centers on developing trustworthy technology solutions that strengthen organizational security. He is deeply passionate about advancing the intersection of cybersecurity and artificial intelligence.*



Technical Updates

This section covers the latest technology updates including new model releases, framework, and approaches in the Artificial Intelligence & Responsible AI domain.

New Models Released

DeepSeek-OCR 2: Layout-Aware Document Understanding with Causal Visual Flow Encoding

DeepSeek AI has released DeepSeek-OCR 2, an open-source OCR and document understanding model that improves how complex documents are read by introducing a Causal Visual Flow encoder that converts 2D page layouts into a human-like, semantically ordered 1D visual token sequence before text decoding. This design overcomes limitations of traditional raster-scan OCR by preserving reading order across multi-column layouts, tables, and mixed content, resulting in more accurate structural and textual understanding. Built with an efficient vision encoder and a mixture-of-experts language decoder, DeepSeek-OCR 2 demonstrates strong benchmark performance while maintaining high layout fidelity, making it well suited for scalable, real-world document intelligence applications.[61]

NVIDIA's Nemotron-3-Nano-30B-NVFP4: Efficient 4-Bit Reasoning Model with Quantization Aware Distillation for High-Performance Inference

NVIDIA has introduced Nemotron-3-Nano-30B-NVFP4, a highly efficient 30 billion-parameter hybrid Mamba2 Transformer Mixture-of-Experts model quantized to the 4-bit NVFP4 format and optimized with a Quantization Aware Distillation (QAD) training strategy that preserves accuracy close to full-precision BF16 while drastically improving inference efficiency on NVIDIA GPUs. The model retains key BF16 layers for stability, uses FP8 for key-value caches, and activates roughly 3.5 billion parameters per token, enabling extended context windows and strong reasoning performance. QAD aligns the quantized student model with a frozen BF16 teacher via KL divergence, avoiding the complexity of replaying full supervised or reinforcement training pipelines and recovering near-baseline benchmarks where simpler quantization methods fall short. This yields up to 4× higher throughput and much lower memory usage compared to FP8, making Nemotron-3-Nano-NVFP4 a practical choice for scalable, efficient reasoning and inference workloads in agentic AI applications.[62]

Qwen3-Coder-Next: Open-Weight Sparse MoE Language Model Tailored for Agentic Coding Workflows, Local Development, and Ultra-Long Context Tasks

The Qwen team has unveiled Qwen3-Coder-Next, an open-weight causal language model specifically designed to power coding agents and local development environments by combining an 80 B total parameter backbone with a sparse Mixture-of-Experts (MoE) architecture that activates only ~3 billion parameters per token to keep inference costs low while retaining high modeling capacity; the model employs a hybrid stack of Gated DeltaNet, Gated Attention, and MoE layers optimized for long-horizon reasoning and integrates with IDE/CLI agent frontends such as Qwen-Code, Claude-Code, and Cline. Trained agentically on a large corpus of 800 K verifiable executable tasks with reinforcement learning to enhance capability in planning edits, calling tools, running code, and recovering from failures rather than simple next-token prediction, Qwen3-Coder-Next supports an ultra-long 256 K token context, delivers competitive performance on benchmarks like SWE-Bench Verified, SWE-Bench Pro, Terminal-Bench 2.0, and Aider that rivals models with 10–20× more active compute, and offers flexible deployment via OpenAI-compatible APIs (e.g., SGLang, vLLM) or local GGUF quantizations, enabling efficient, high-capability coding agent workflows and local private copilots under an Apache-2.0 license.[63]

OpenAI Unveils GPT-5.3-Codex A Next-Generation Agentic Coding Model That Blends Enhanced Code Performance, Professional Reasoning, and Self-Referential Development

OpenAI announced the launch of GPT-5.3-Codex, its latest and most capable agentic coding model that unifies the advanced

[61] https://www.marktechpost.com/2026/01/30/deepseek-ai-releases-deepseek-ocr-2-with-causal-visual-flow-encoder-for-layout-aware-document-understanding/
[62] https://www.marktechpost.com/2026/02/01/nvidia-ai-brings-nemotron-3-nano-30b-to-nvfp4-with-quantization-aware-distillation-qad-for-efficient-reasoning-inference/
[63] https://www.marktechpost.com/2026/02/03/qwen-team-releases-qwen3-coder-next-an-open-weight-language-model-designed-specifically-for-coding-agents-and-local-development/

coding performance of GPT-5.2-Codex with the robust reasoning and professional knowledge capabilities of GPT-5.2 into a single system, delivering approximately 25 % faster inference and execution for users through infrastructure and inference optimizations; the model is designed to undertake long-running tasks involving research, tool use, and complex execution while remaining steerable "much like a colleague," and it demonstrates strong benchmark results, broad utility across coding and professional workflows, and expanded safety measures as part of its rollout.[64]

Sarvam Vision: India-Focused Multilingual Vision-Language Model for Advanced Document Understanding

Sarvam AI has introduced Sarvam Vision, a 3-billion-parameter vision-language model specifically engineered to advance document intelligence by integrating visual understanding with semantic interpretation across a wide range of content types and India's diverse languages. The model addresses the limitations of current global VLMs particularly their lower accuracy on non-English and regional scripts by combining high-fidelity OCR with layout and visual element parsing to extract deeper knowledge from documents such as complex tables, charts, and mixed-content scans. Rigorous data curation and specialized training on synthetic and real-world multilingual document datasets enable Sarvam Vision to deliver best-in-class performance on Indic OCR benchmarks and visual understanding tasks, unlocking valuable information from historical archives, financial reports, and research materials to support research, governance, and enterprise workflows.[65]

Alibaba Qwen Team Releases Qwen3.5-397B MoE Model with 17B Active Parameters and 1M-Token Context Window Optimized for Native Multimodal AI Agents

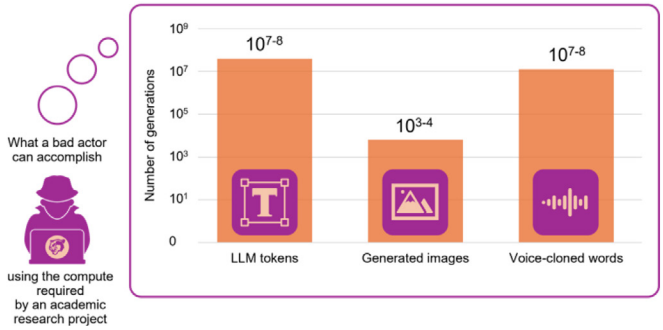
The Alibaba Qwen team has introduced Qwen3.5-397B-A17B, a large Mixture-of-Experts (MoE) language model designed to deliver frontier-level reasoning efficiency by activating only 17 billion parameters out of a total 397 billion during inference, thereby balancing massive model capacity with significantly lower computational cost. Built as part of the broader Qwen3.5 series developed by Alibaba, the model features a 1-million-token context length, native multimodal and agent-centric capabilities, and an architecture that supports large-scale planning, tool use, and long-horizon reasoning tasks suited for advanced AI agents and enterprise workflows. Its sparse MoE design enables 400B-class performance while maintaining faster inference comparable to smaller dense models, and the release aligns with Alibaba's push toward "agentic AI" systems capable of executing complex multi-step tasks autonomously across applications. Additionally, the model is part of the open-weight Qwen ecosystem, which has seen continuous expansion with scalable dense and sparse variants aimed at developers and researchers seeking customizable, high-efficiency foundation models for coding, reasoning, and multimodal deployments.[66]

Google Launches Gemini 3.1 Pro with 1M Token Context and Enhanced Reasoning for AI Agent Workloads

Google has introduced Gemini 3.1 Pro as an advanced AI model designed for complex reasoning, long-context processing, and agent-based applications, featuring up to a 1 million token context window and improved performance on abstract reasoning benchmarks such as ARC-AGI-2. The release highlights stronger multi-step problem-solving, better synthesis of large datasets, and enhanced capabilities for enterprise, coding, and research workflows, reflecting Google's broader focus on developing reasoning-centric AI systems optimized for scalable AI agents and high-complexity tasks.[67]

New Frameworks & Research Techniques

The Rising Risk of Low-Compute AI Systems and the Limits of Compute-Based Governance

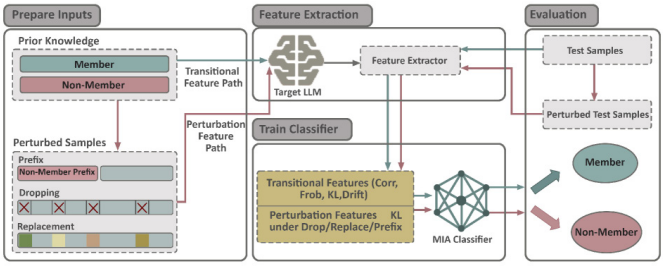


This position paper argues that while AI governance has largely focused on regulating high-compute systems, rapidly advancing low-compute AI models now pose comparable and in some cases greater risks. Advances such as agentic workflows, parameter quantization, and model compression have enabled capabilities once limited to frontier models to migrate into small, efficient systems deployable on consumer hardware. By analysing historical benchmark data from over 5,000 HuggingFace-hosted

[64] https://www.marktechpost.com/2026/02/05/openai-just-launched-gpt-5-3-codex-a-faster-agentic-coding-model-unifying-frontier-code-performance-and-professional-reasoning-into-one-system/
[65] https://www.sarvam.ai/blogs/Sarvam-vision
[66] https://www.marktechpost.com/2026/02/16/alibaba-qwen-team-releases-qwen3-5-397b-moe-model-with-17b-active-parameters-and-1m-token-context-for-ai-agents/
[67] https://www.marktechpost.com/2026/02/19/google-ai-releases-gemini-3-1-pro-with-1-million-token-context-and-77-1-percent-arc-agi-2-reasoning-for-ai-agents/

LLMs, the authors show that the model size required to achieve competitive performance has dropped by more than tenfold in a year. Simulations further demonstrate that harmful campaigns including disinformation botnets, voice-cloning fraud, and sexual extortion schemes can be executed easily using open-source, low-compute models on consumer devices. The paper concludes that current compute-focused safeguards leave critical security gaps and calls for urgent policy and technical efforts to address the emerging threat landscape of low-resource AI systems.[68]

AttenMIA: Attention-Based Membership Inference Reveals Privacy Risks in LLMs



This work introduces AttenMIA, a novel membership inference attack framework that exposes privacy and intellectual property risks in LLMs by exploiting internal self-attention patterns rather than relying on brittle output confidence or embedding-based signals. By analysing attention heads across transformer layers and combining these signals with perturbation-based divergence metrics, AttenMIA effectively identifies whether specific data samples were part of a model's training set. Extensive experiments on open-source models such as LLaMA-2, Pythia, and OPT demonstrate that attention-based features consistently outperform prior methods, achieving state-of-the-art results such as up to **0.996 ROC AUC and 87.9% TPR at 1% FPR** on the WikiMIA-32 benchmark. The study further shows that attention leakage generalizes across datasets and architectures, pinpoints layers and heads most prone to memorization, and enables stronger downstream data extraction attacks, highlighting that attention mechanisms, while improving model performance and interpretability, can significantly amplify privacy risks and demand new defensive strategies.[69]

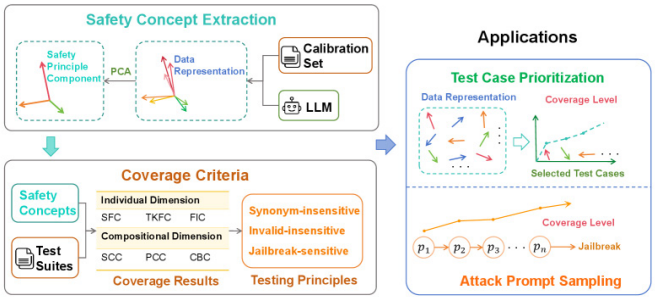
Anthropic's Claude Legal Plugin: AI-Powered Automation for Contract Review, NDA Triage, and Compliance Workflows

Anthropic has released a Claude Legal Plugin, a specialized extension for its Claude Cowork AI platform that automates and accelerates high-volume legal tasks for in-house legal teams, including contract review with clause-by-clause analysis and color-coded risk indicators, rapid NDA triage, vendor agreement checks, contextual legal brief generation, and templated responses for routine legal inquiries, all configurable to an organization's internal playbook and risk tolerances; the plugin integrates with document management and project tracking tools to provide richer context, significantly reducing manual workload for commercial counsel, product and privacy compliance teams, and litigation support staff, but Anthropic emphasizes that outputs should be reviewed by licensed attorneys rather than treated as formal legal advice.[70]

Google Unveils Conductor a Context-Driven Gemini CLI Extension That Transforms AI-Assisted Development with Versioned Markdown Workflows

Google has released Conductor, an open-source preview extension for its Gemini CLI that reimagines AI-assisted software development by capturing and storing project context, technical decisions, product goals, and workflow rules as persistent, version-controlled Markdown files within a repository, eliminating reliance on ephemeral chat prompts and enabling reproducible, team-aligned agentic workflows; rather than jumping directly from a natural-language request to code edits, Conductor enforces a structured lifecycle of context→specification and plan→implementation, where developers first define high-level specs and actionable plans (e.g., via `/conductor:newTrack`) and then let the agent execute tasks according to the approved plan while providing checkpoints and status commands (such as `/conductor:status`, `/conductor:review`, and `/conductor:implement`), all of which remain auditable and shareable across contributors, with support for both new and brownfield codebases.[71]

RACA: Representation-Aware Coverage Criteria for Safety Testing of LLMs



This paper introduces RACA, a representation-aware coverage framework designed to systematically evaluate the safety of LLMs against jailbreak attacks. Unlike traditional neuron-level coverage methods that do not scale well to LLMs, RACA focuses

[68] https://www.arxiv.org/abs/2601.21365
[69] https://arxiv.org/abs/2601.18110
[70] https://claude.com/plugins/legal
[71] https://www.marktechpost.com/2026/02/02/google-releases-conductor-a-context-driven-gemini-cli-extension-that-stores-knowledge-as-markdown-and-orchestrates-agentic-workflows/

on safety-critical concepts using representation engineering to reduce dimensionality and remove irrelevant signals. The approach identifies safety-critical representations from a small expert-curated set of jailbreak prompts, computes conceptual activation scores for test inputs, and measures coverage using six criteria that capture both individual and compositional safety concepts. Experimental results show that RACA effectively identifies high-quality jailbreak prompts, outperforms existing coverage techniques, and generalizes robustly across models and configurations, making it a practical tool for LLM safety testing.[72]

Perplexity's Model Council: Multi-Model Consensus for More Reliable AI Answers

Perplexity has introduced Model Council, a feature that processes a user's query through three leading AI models simultaneously and synthesizes their responses to produce a single, higher-confidence answer. In this system, each model runs the same query independently; a separate synthesizer model then compares their outputs, highlights consensus and disagreements, and either combines the best elements or indicates where models differ, giving users both a consolidated result and visibility into each model's contributions. This multi-model approach aims to reduce the risk of errors, biases, or hallucinations associated with relying on a single model by cross-checking perspectives and increasing reliability particularly for complex tasks, research, and decision-making. Model Council is currently available on the web for Perplexity Max subscribers, with evidence showing it can improve accuracy and clarity by leveraging diverse model reasoning and presenting aggregated insights.[73]

Anthropic Launches Claude Opus 4.6 With 1 Million-Token Context, Agentic Coding, Adaptive Reasoning Controls, and Enhanced Safety and Tooling for Enterprise and Developer Workloads

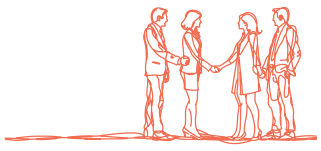
Anthropic has released Claude Opus 4.6, the latest and most capable iteration of its Opus family of AI models, engineered to handle extended reasoning, multi-step agentic coding, and complex professional workflows with a beta 1 million-token context window, expanded agentic coding features, and new adaptive reasoning and effort controls that let the model allocate computation dynamically based on task complexity; the update also introduces context compaction, support for up to 128 k output tokens, deeper integration with tools like Excel and PowerPoint, and expanded safety tooling, positioning Claud e Opus 4.6 as a frontier model for both enterprise knowledge work and developer ecosystem tasks available via the Claude API and major cloud platforms.[74]

Efficient Robot Control Through Highly Compressed and Ordered Action Tokens

Researchers from Harvard University and Stanford University have introduced Ordered Action Tokenization (OAT), a learned action tokenization framework designed to bridge the longstanding gap between continuous robotic control and autoregressive modeling techniques used in LLMs, by converting high-dimensional continuous robot actions into compact, discretized token sequences that satisfy three critical desiderata high compression to reduce sequence length, total decodability to ensure every token sequence maps to a valid action, and a left-to-right causal ordering that aligns with next-token prediction using a transformer with register tokens, finite scalar quantization, and an ordering-inducing training mechanism; this ordered structure enables "anytime" inference where partial token prefixes produce coarse actions quickly and additional tokens refine precision, and across more than 20 simulation and real-world tasks OAT-equipped autoregressive policies outperform prior tokenization methods and diffusion-based baselines while offering scalable, flexible inference for future robot learning and control systems.[75]

NVIDIA Researchers Unveil KVTC: A Transform Coding Pipeline Achieving Up to 20× Compression of Key-Value Caches for Efficient LLM Serving

NVIDIA researchers have introduced KVTC (Key-Value Cache Transform Coding), a transform coding pipeline designed to significantly reduce the memory footprint of key-value (KV) caches used during LLM inference and serving, addressing the critical bottleneck that KV caches especially for large contexts pose for throughput, latency, and resource utilization; KVTC applies a lightweight, media-inspired sequence of learned orthonormal transformation, adaptive quantization, and entropy coding to compress KV caches by up to 20× (and in some use cases even **40×+) while maintaining reasoning and long-context accuracy, integrates protections for critical tokens (e.g., earliest and most recent tokens) to preserve model performance, and demonstrates notable operational benefits such as up to 8× reduction in Time-To-First-Token compared with full cache recomputation, fast calibration times without modifying model weights, and minimal additional storage overhead, making it a practical foundation for memory-efficient, high-performance LLM serving systems.[76]



[72] https://arxiv.org/abs/2602.02280

[73] https://www.perplexity.ai/hub/blog/introducing-model-council

[74] https://www.marktechpost.com/2026/02/05/anthropic-releases-claude-opus-4-6-with-1m-context-agentic-coding-adaptive-reasoning-controls-and-expanded-safety-tooling-capabilities/

[75] https://www.marktechpost.com/2026/02/08/meet-oat-the-new-action-tokenizer-bringing-llm-style-scaling-and-flexible-anytime-inference-to-the-robotics-world/

[76] https://www.marktechpost.com/2026/02/10/nvidia-researchers-introduce-kvtc-transform-coding-pipeline-to-compress-key-value-caches-by-20x-for-efficient-llm-serving/

NVIDIA Unveils C-RADIOv4: Unified Vision Backbone Integrating SigLIP2, DINOv3, and SAM3 for Scalable Classification, Dense Prediction, and Segmentation Workloads

NVIDIA has introduced C-RADIOv4, an advanced vision backbone that consolidates strengths from three distinct teacher models SigLIP2, DINOv3, and SAM3 into a single ViT-style encoder designed for image classification, dense prediction, and segmentation at scale. By employing agglomerative distillation and stochastic multi-resolution training, C-RADIOv4 maintains performance across resolutions while balancing dense feature extraction and alignment with visual-language representations. The unified architecture improves dense prediction quality and resolution robustness with competitive accuracy on benchmarks like ImageNet, and it can serve as a drop-in backbone for existing decoders such as SAM3, offering a versatile foundation for diverse vision workloads under the NVIDIA Open Model License.[77]

NVIDIA-Backed Waymo Launches "World Model" Simulator Using DeepMind's Genie 3 for Autonomous Driving

Waymo announced the Waymo World Model, a next-generation autonomous driving simulation platform built atop Google DeepMind's Genie 3 world-model AI, designed to generate hyper-realistic, multi-sensor 3D environments that include camera and LiDAR data for robust training and evaluation of autonomous systems. The model adapts Genie 3's generative capabilities to automotive contexts, producing temporally consistent simulations that expose the Waymo Driver to rare, safety-critical "long-tail" scenarios such as extreme weather and unusual objects that traditional simulators cannot easily recreate, helping scale virtual driving miles into the billions for safer robotaxi development.[78]

OpenAI Unveils GPT-5.3-Codex-Spark: A Cerebras-Powered Ultra-Low-Latency Coding AI Delivering Over 15x Faster Throughput for Real-Time Developer Workflows

OpenAI has published a research preview of GPT-5.3-Codex-Spark, a streamlined, speed-tuned variant of its latest Codex family designed specifically for near-instant coding by developers, capable of generating more than 1,000 tokens per second approximately **15× the throughput** of the flagship GPT-5.3-Codex model by leveraging ultra-low-latency **Cerebras Wafer-Scale Engine 3 hardware and system-level optimizations** such as persistent WebSocket connections that drastically reduce round-trip time and time-to-first-token, all while maintaining a large 128k token context window; the model is currently available as a research preview to **ChatGPT Pro** users via the Codex app, command-line tools, and IDE integrations, and represents a first step in OpenAI's strategy to blend rapid interactive coding with deeper reasoning and agentic capabilities in future releases.[79]

A Cryptographically Enforced Framework for LLM Security: Authenticated Prompts, Tamper-Evident Context Chains, and Provable Byzantine-Resilient Policy Algebra for Preventative Defense Against Prompt Injection

The authors present a novel security architecture for LLM applications that addresses prompt injection and context manipulation threats beyond the scope of traditional security controls by introducing two foundational cryptographic primitives authenticated prompts, which enable self-contained and verifiable prompt lineage, and authenticated context, which employs tamper-evident hash chains to guarantee the integrity of dynamically evolving inputs upon which they formalize a policy algebra supported by four proven theorems establishing protocol-level Byzantine resistance such that even adversarial agents cannot violate organizational constraints; complemented by five layered defensive mechanisms ranging from lightweight resource governance to LLM-based semantic validation, the framework delivers preventative, formally guaranteed security assurances, achieving 100% detection with zero false positives and negligible performance overhead across representative attacks spanning six comprehensive categories, thereby representing the first integrated approach that unifies cryptographically verifiable prompt provenance, tamper-evident contextual integrity, and mathematically provable policy enforcement to transition LLM security from reactive mitigation to preventative assurance.[80]

Exa AI Unveils Exa Instant: A Sub-200ms Neural Search Engine Transforming Real-Time Agentic Workflows.

Exa AI has introduced Exa Instant, a proprietary sub-200ms neural search engine engineered to eliminate latency bottlenecks in real-time agentic workflows by delivering web results significantly faster than traditional search wrappers. Designed for Retrieval-Augmented Generation (RAG) pipelines, Exa Instant enables AI agents to perform multiple sequential searches without noticeable delay, thanks to its proprietary end-to-end neural search and retrieval stack that avoids reliance

[77] https://www.marktechpost.com/2026/02/06/nvidia-ai-releases-c-radiov4-vision-backbone-unifying-siglip2-dinov3-sam3-for-classification-dense-prediction-segmentation-workloads-at-scale/

[78] https://www.marktechpost.com/2026/02/06/waymo-introduces-the-waymo-world-model-a-new-frontier-simulator-model-for-autonomous-driving-and-built-on-top-of-genie-3/

[79] http://marktechpost.com/2026/02/12/openai-releases-a-research-preview-of-gpt-5-3-codex-spark-a-15x-faster-ai-coding-model-delivering-over-1000-tokens-per-second-on-cerebras-hardware/

[80] https://www.arxiv.org/pdf/2602.10481

on slower, wrapper-based APIs. Benchmarks show it achieving 100–200ms search times with network latency as low as 50ms, making it up to 15× faster than competing solutions like Tavily Ultra Fast and Brave. By prioritizing semantic relevance through transformer-based embeddings rather than keyword matching, Exa Instant improves both speed and contextual accuracy, positioning itself as a critical tool for developers building responsive research, coding, and automation agents.[81]

Kyutai Introduces Hibiki-Zero: A 3B-Parameter Simultaneous Speech-to-Speech and Speech-to-Text Translation Model Leveraging GRPO Reinforcement Learning to Eliminate Dependence on Word-Level Aligned Training Data

Kyutai has released Hibiki-Zero, a 3-billion-parameter decoder-only model designed for simultaneous speech-to-speech (S2ST) and speech-to-text (S2TT) translation that removes the need for word-level aligned datasets, a major bottleneck in traditional real-time translation systems. The model adopts a multistream architecture that jointly processes source audio, target audio, and an internal text representation while operating in a streaming setting, enabling real-time translation with preserved speaker identity and naturalness. Instead of relying on scarce interpretation-style aligned data, Hibiki-Zero is initially trained on sentence-level alignments and then optimized using Group Relative Policy Optimization (GRPO) with BLEU-based process rewards to balance latency and translation quality. This reinforcement learning-driven training pipeline simplifies data preparation, improves scalability across languages with different grammatical structures, and achieves strong performance in accuracy, naturalness, and cross-lingual speaker similarity across multilingual tasks. Additionally, the system demonstrates efficient adaptation to new languages with relatively limited speech data and represents a shift toward end-to-end expressive translation systems that can incrementally generate translated speech in real time without complex alignment heuristics.[82]

Kani-TTS-2: A 400M-Parameter Open-Source Text-to-Speech Model Optimized for Low-VRAM Deployment with Audio-as-Language Architecture and Built-In Voice Cloning Capabilities

Kani-TTS-2 is an open-source 400 million-parameter text-to-speech model introduced by the nineninesix.ai team, designed to deliver high-quality, human-like speech synthesis while running efficiently on as little as 3GB of VRAM, making it suitable for consumer hardware and edge deployments. The model adopts an "audio-as-language" paradigm, replacing traditional mel-spectrogram pipelines with a token-based approach that converts raw audio into discrete tokens and generates speech through a language-style backbone built on LiquidAI's LFM2 architecture, which predicts audio intent before decoding it into 22kHz waveforms using NVIDIA's NanoCodec neural codec. Trained on approximately 10,000 hours of high-quality speech data in a highly optimized process that reportedly completed in just six hours on an eight-GPU cluster, Kani-TTS-2 emphasizes computational efficiency, scalability, and real-time performance. It is released in multiple language variants and supports voice cloning, enabling users to replicate speaker characteristics while maintaining natural prosody and fidelity, positioning it as a lightweight yet capable alternative to heavier proprietary TTS systems and a practical solution for offline, low-latency conversational AI and speech generation workflows.[83]

Depth-Oriented Reliability Assessment of LLMs Using Accelerated Prompt Stress Testing (APST)

In this work, the authors argue that traditional breadth-oriented safety benchmarks for LLMs overlook a critical dimension of real-world deployment: the risk of operational failures that emerge when the same or highly similar prompts are issued repeatedly. They introduce Accelerated Prompt Stress Testing (APST) as a depth-oriented evaluation framework, drawing inspiration from reliability engineering to measure how consistently and safely an LLM behaves under sustained use. APST repeatedly samples identical prompts under controlled inference settings, enabling the detection of latent failure modes such as hallucinations, inconsistent refusals, or unsafe outputs that may not surface in single-sample evaluations. The framework treats each inference as a stochastic event and quantifies safety risk using Bernoulli and binomial models to estimate per-inference failure probabilities. Through experiments on multiple instruction-tuned LLMs using safety-critical prompts derived from AIR-BENCH, the authors find that models with similar benchmark scores can exhibit sharply different empirical failure rates, especially at higher decoding temperatures. These findings show that one-shot safety evaluations can mask meaningful reliability gaps, and APST therefore provides a practical methodology that complements existing benchmarks by linking traditional performance metrics with deployment-oriented reliability assessment.[84]

[81] https://www.marktechpost.com/2026/02/13/exa-ai-introduces-exa-instant-a-sub-200ms-neural-search-engine-designed-to-eliminate-bottlenecks-for-real-time-agentic-workflows/
[82] https://www.marktechpost.com/2026/02/13/kyutai-releases-hibiki-zero-a3b-parameter-simultaneous-speech-to-speech-translation-model-using-grpo-reinforcement-learning-without-any-word-level-aligned-data/
[83] https://www.marktechpost.com/2026/02/15/meet-kani-tts-2-a-400m-param-open-source-text-to-speech-model-that-runs-in-3gb-vram-with-voice-cloning-support/
[84] https://arxiv.org/abs/2602.11786

Adaptive Safe Context Learning (ASCL) with IFPO: A Framework for Balancing Safety Alignment and Reasoning Utility in Advanced Reasoning Models

While reasoning models have demonstrated significant progress in solving complex tasks, their growing capability intensifies the need for robust safety alignment mechanisms that do not excessively constrain utility. The primary difficulty stems from the safety–utility trade-off, where conventional alignment methods often rely on Chain-of-Thought (CoT) training data infused with explicit safety rules through context distillation, inadvertently fostering rigid rule memorization and over-refusal behaviors that impair reasoning flexibility. To address this limitation, the proposed Adaptive Safe Context Learning (ASCL) framework reconceptualizes safety alignment as a multi-turn tool-use paradigm, enabling the model to dynamically determine when to consult safety rules and how to structure ongoing reasoning based on contextual relevance rather than static rule embedding. Additionally, the introduction of Inverse Frequency Policy Optimization (IFPO) mitigates reinforcement learning bias toward excessive rule consultation by rebalancing advantage estimates, thereby preventing over-dependence on safety retrieval actions. By decoupling rule retrieval from the reasoning generation process, the approach enhances contextual decision-making and achieves improved overall performance relative to baseline alignment strategies, demonstrating a more nuanced pathway to maintaining safety without sacrificing reasoning efficacy.[85]

New Agentic AI Research

SERA: AI2's Supervised, Soft-Verified Coding Agents for Practical Repository Automation

The Allen Institute for AI (AI2) has released SERA (Soft-Verified Efficient Repository Agents), a family of open-source coding agents trained solely through supervised learning on synthetic developer-like trajectories, eliminating the need for reinforcement learning or heavy test suites. The flagship SERA-32B model, along with smaller variants, achieves strong performance on the SWE-Bench Verified benchmark placing it in the same performance range as much larger systems while remaining fully open in code, data, and weights and costing significantly less to train. The key innovation, Soft Verified Generation (SVG), generates realistic multi-step coding workflows and uses patch agreement as a soft correctness signal, demonstrating that even partially verified examples suffice for training effective agents. This approach makes

[85] https://www.arxiv.org/abs/2602.13562

it practical and affordable for developers to train and specialize coding agents to specific repositories and real-world workflows, democratizing access to advanced, repository-aware automation tools.[86]

NVIDIA Unveils VibeTensor AI-Generated Deep Learning Runtime Built by LLM Coding Agents

NVIDIA has released VibeTensor, an open-source deep learning runtime system generated end-to-end by LLM-powered coding agents under high-level human guidance. The system implements a PyTorch-style eager tensor library with a C++20 core for CPU and CUDA, a Python API overlay, and an experimental Node.js interface, and spans from language bindings down to CUDA memory management. All implementation changes were proposed and validated through automated tools, demonstrating that AI agents can collaboratively produce a coherent complex software stack, with the design validated through builds, tests, and differential checks rather than manual line-by-line coding.[87]

Moonshot AI Introduces Cloud-Native Kimi Claw with 5,000 Community Skills and 40GB Storage

Moonshot AI has launched Kimi Claw, a fully cloud-native version of its OpenClaw framework, bringing always-on AI agent capabilities directly to the browser via kimi.com. The platform now provides a persistent, scalable environment for developers and data scientists, featuring ClawHub an extensive library of over 5,000 community-contributed skills that enable modular tool interactions and no-code integrations. It also includes 40GB of dedicated cloud storage to support large datasets, technical documents, and RAG-optimized workflows. Additionally, Kimi Claw integrates Pro-Grade Search for real-time structured data retrieval from sources such as Yahoo Finance, while its "Bring Your Own Claw" option and multi-app bridging (including Telegram) allow advanced hybrid connectivity for users with custom setups.[88]

Google AI Unveils PaperBanana A Multi-Agent Framework Automating Publication-Ready Methodology Diagrams and Statistical Plots for Scientific Research

Google AI, in collaboration with researchers from Peking University, has introduced PaperBanana, an agentic AI framework that significantly streamlines the creation of professional academic visuals by automating the generation of publication-ready methodology diagrams and statistically accurate plots from textual descriptions; the system orchestrates a team of five specialized agents including retrieval of reference examples, planning of visual content, stylistic layout, rendering (via image models or executable plotting code), and iterative self-critique for refinement and demonstrates performance improvements over baseline methods across key metrics such as faithfulness, conciseness, readability, and aesthetics using the PaperBananaBench dataset of 292 NeurIPS 2025 test cases, addressing a longstanding bottleneck in the research workflow by replacing manual diagram design with a scalable, agent-driven pipeline.[89]

[86] https://www.marktechpost.com/2026/01/30/ai2-releases-sera-soft-verified-coding-agents-built-with-supervised-training-only-for-practical-repository-level-automation-workflows/
[87] https://www.marktechpost.com/2026/02/04/nvidia-ai-release-vibetensor-an-ai-generated-deep-learning-runtime-built-end-to-end-by-coding-agents-programmatically/
[88] https://www.marktechpost.com/2026/02/15/moonshot-ai-launches-kimi-claw-native-openclaw-on-kimi-com-with-5000-community-skills-and-40gb-cloud-storage-now/
[89] https://www.marktechpost.com/2026/02/07/google-ai-introduces-paperbanana-an-agentic-framework-that-automates-publication-ready-methodology-diagrams-and-statistical-plots/

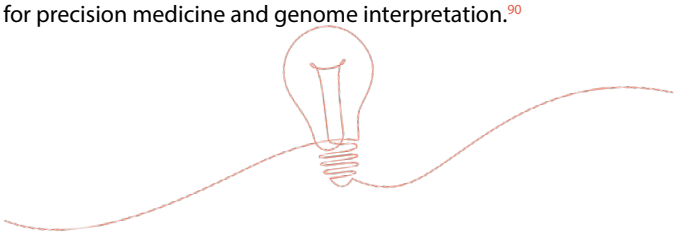
Industry Update

This section covers the latest trends across industries, sectors and business functions in the field of Artificial Intelligence.

Healthcare

Google DeepMind Unveils AlphaGenome A Unified Hybrid Transformer and U-Net Sequence-to-Function Model That Decodes the Human Genome at Megabase Scale

Google DeepMind has introduced AlphaGenome, a unified deep learning model that leverages a hybrid architecture combining U-Net and Transformer blocks to process extremely long DNA sequences up to 1,000,000 base pairs at single-base resolution and directly map raw genomic sequence to functional biological outputs, enabling simultaneous prediction across multiple genomic modalities such as RNA-seq, CAGE, ATAC-seq, transcription factor binding, and chromatin contact maps; the model's multi-task learning approach and teacher-student distillation facilitate industry-leading variant effect prediction performance, capturing long-range regulatory interactions that underpin gene regulation and disease mechanisms, while its implementation in JAX optimized for TPU hardware supports efficient high-performance computation and transfer learning to cell types with limited data, positioning AlphaGenome as a substantial advancement in AI-driven genomics with implications for precision medicine and genome interpretation.[90]



Harvard-Affiliated Researchers Develop BrainIAC, an AI Foundation Model That Predicts Brain Age, Dementia Risk, Tumor Mutations, and Cancer Survival from Routine MRI Scans

Researchers at Harvard-affiliated Mass General Brigham have created an advanced artificial intelligence foundation model named BrainIAC (Brain Imaging Adaptive Core) that can analyze routine brain MRI scans to extract multiple clinically relevant health indicators, including estimating a person's "brain age," predicting the risk of dementia, identifying genetic mutations in brain tumors, and forecasting survival outcomes for brain cancer patients, outperforming conventional task-specific AI models even when limited annotated training data are available, according to a study published in Nature Neuroscience; BrainIAC was pretrained on nearly 49,000 diverse MRIs using self-supervised learning to identify inherent imaging features that generalize across healthy and abnormal cases, enabling both straightforward tasks like scan classification and complex applications such as tumor mutation detection, and researchers say its ability to perform effectively in real-world clinical settings could accelerate biomarker discovery, enhance diagnostic tools, and support more personalized patient care, though further studies on larger and varied imaging datasets are needed.[91]

Defence

Microsoft's "OrbitalBrain": A Distributed In-Space Machine Learning Framework Leveraging Inter-Satellite Links and Constellation-Aware Resource Optimization

Microsoft researchers have proposed OrbitalBrain, an innovative distributed machine learning framework that reconceptualizes satellite constellations as collaborative, in-orbit ML systems rather than mere data relays, addressing the entrenched bottleneck of limited downlink bandwidth that delays Earth observation training data reaching terrestrial infrastructure by enabling model training, aggregation, and update processes to occur directly in space using onboard compute resources, inter-satellite links, and predictive scheduling of power, storage, and communication; by co-scheduling local compute, in-orbit model aggregation, and data transfers across satellites based on forecasts of orbital dynamics and resource availability, OrbitalBrain demonstrably accelerates time-to-accuracy by up to 12.4× and improves final model performance compared with traditional ground-based and

[90] https://www.marktechpost.com/2026/01/28/google-deepmind-unveils-alphagenome-a-unified-sequence-to-function-model-using-hybrid-transformers-and-u-nets-to-decode-the-human-genome/
[91] https://news.harvard.edu/gazette/story/2026/02/new-ai-tool-predicts-brain-age-dementia-risk-cancer-survival/

adapted federated learning baselines under realistic constraints, thus advancing the potential for constellation-centric intelligence for tasks such as rapid environmental analytics and real-time Earth observation insights.[92]

Environmental Monitoring

NVIDIA Unveils Earth-2: The First Fully Open GPU-Accelerated AI Weather and Climate Prediction Stack

NVIDIA has introduced Earth-2, the world's first fully open, accelerated AI weather stack designed to transform climate and weather forecasting by making it faster, more affordable, and broadly accessible. Earth-2 comprises a suite of pretrained models, inference libraries, and development tools hosted on platforms like GitHub and Hugging Face, enabling researchers, governments, enterprises, and startups to build sovereign forecasting systems without reliance on expensive supercomputers. The Earth-2 family includes innovative models such as Earth-2 Medium Range for 15-day global forecasts, Earth-2 Nowcasting for high-resolution short-term storm prediction, and Earth-2 Global Data Assimilation for rapid integration of observational data, delivering prediction capabilities that rival or exceed traditional physics-based methods while dramatically reducing compute time. Already adopted by weather services, energy companies, and risk-management firms, the open Earth-2 stack aims to democratize weather intelligence and support critical applications ranging from disaster response to renewable energy optimization.[93]

Finance

BaFin Publishes Non-Binding Guidance on Managing ICT Risks from AI to Support Financial Institutions' Compliance with the Digital Operational Resilience Act

The German financial regulator, BaFin, released on 30 January 2026 a comprehensive non-binding guidance document outlining expectations for how supervised financial institutions should manage information and communication technology (ICT) risks arising from the use of artificial intelligence (AI) in the context of the EU's Digital Operational Resilience Act (DORA); the guidance emphasizes that AI systems must be embedded into existing ICT risk management frameworks across their full lifecycle from data sourcing and development to deployment, monitoring and decommissioning and underscores robust governance, third-party and cloud risk controls, cybersecurity measures and integration with existing ICT governance structures rather than separate AI-specific regimes, with the aim of enhancing operational resilience and supporting firms' fulfillment of DORA requirements.[94]

Judiciary

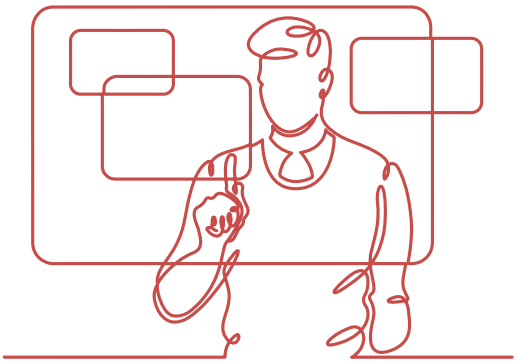
AI for Justice: Ethical, Fair and Robust Adoption in India's Courts

The report produced by DAKSH and Digital Futures Lab, commissioned by the United Nations Development Programme (UNDP), explains that India's courts are beginning to use AI tools for tasks such as translation, transcription, legal research, and case management, but stresses that these tools must be used with strong rules and safeguards to protect people's rights. It highlights that AI can improve access to justice and make court work faster, but warns that poorly managed systems may threaten privacy, fairness, and equality. The report is based on conversations with judges, court staff, legal-tech experts, and academics, and it presents a four-step process for courts to follow: checking readiness, identifying risks, conducting technical reviews, and monitoring ongoing performance. It urges courts to adopt AI carefully, so they remain fair, transparent, and trustworthy.[95]

Agriculture

Union Budget 2026-27 Unveils Bharat-VISTAAR A Multilingual AI-Powered Agricultural Decision-Support Platform to Transform Farming in India

In the Union Budget 2026-27, Finance Minister Nirmala Sitharaman announced the government's plan to launch Bharat-VISTAAR (Virtually Integrated System to Access Agricultural Resources), a multilingual AI-driven tool designed to integrate the AgriStack portal with the Indian Council of Agricultural Research's (ICAR) package of agricultural practices, with the objective of enhancing farm productivity and enabling better decision-making for farmers through customised, data-driven digital advisory support; the initiative aims to make scientific farming insights accessible in multiple languages, reduce risk through targeted advisories, and support precision agriculture at the field level, while also being complemented by other measures to modernise allied sectors such as livestock and fisheries as part of a broader push towards tech-enabled, sustainable farming in India.[96]



[92] https://www.marktechpost.com/2026/02/09/microsoft-ai-proposes-orbitalbrain-enabling-distributed-machine-learning-in-space-with-inter-satellite-links-and-constellation-aware-resource-optimization-strategies/
[93] https://www.marktechpost.com/2026/01/26/nvidia-revolutionizes-climate-tech-with-earth-2-the-worlds-first-fully-open-accelerated-ai-weather-stack/
[94] https://www.bafin.de/SharedDocs/Veroeffentlichungen/EN/Meldung/2025/neu/meldung_2025_12_18_orientierungshilfe_ikt_risiken_en.html?cms_expanded=true
[95] https://www.undp.org/sites/g/files/zskgke326/files/2026-02/daksh_report_final_25.01_single_page_view.pdf
[96] https://timesofindia.indiatimes.com/business/india-business/union-budget-2026-finance-minister-highlights-indias-advances-in-agri-ambitious-viksit-bharat/articleshow/127833612.cms

Infosys Developments

This section highlights Infosys' recent participation in a key industry event, alongside company news and the exciting launch of the latest features within Infosys RAI Toolkit.

Events

India–Spain Conference on Artificial Intelligence: Exploring Opportunities and Cooperation | January 27, 2026 | Madrid, Spain

On January 27, 2026, the India–Spain Conference on Artificial Intelligence: Exploring Opportunities and Cooperation was held at the InterContinental Hotel in Madrid as a strategic closed door engagement ahead of the AI Impact Summit India 2026. Organized as part of the **70th anniversary** of **India–Spain** diplomatic relations and the India–Spain Dual Year in Culture, Tourism, and Artificial Intelligence, the convening brought together senior government leaders and industry experts to strengthen bilateral collaboration in AI. The event opened with welcome remarks from Kabir Sukhwani, President of the ICC Indian Association of Commerce in Spain, followed by an inaugural address by H.E. Jayant N. Khobragade, Ambassador of India to the Kingdom of Spain and Andorra. The program featured two focused roundtables "Discovering AI Ecosystems in India and Spain" and "Exploring Business Opportunities between India and Spain on Artificial Intelligence" with Infosys contributing through the participation of Suman Kumar, Assistant Vice President, and Sray Agarwal, Head of Responsible AI (EMEA/APAC). The conference was also honored by the presence of H.E. Óscar López, Minister for Digital Transformation and the Civil Service of Spain, underscoring the strategic significance of advancing India–Spain cooperation in artificial intelligence and setting the direction for continued engagement leading up to the AI Impact Summit in New Delhi.

Responsible AI Dialogue: EU–India Pathways to Trustworthy Innovation | 29 January 2026 | Brussels

Infosys, in collaboration with the Embassy of India in Belgium, Luxembourg & the European Union, hosted a high level closed door roundtable in Brussels titled "Responsible AI Dialogue: EU–India Pathways to Trustworthy Innovation" as an official pre summit event of the AI Impact Summit 2026, bringing together senior government, policy, and industry leaders for a focused discussion on Responsible AI and regulatory alignment. The sessions opened with Syed Ahmed, Global Head – Responsible AI Office, Infosys, followed by Dr. M Balaji, Charge d'Affaires, Embassy of India in Brussels, highlighting deeper India–EU cooperation. Keynotes Juha Heikkilä, Adviser on International AI at the European Commission, outlined India's global Responsible AI vision and EU AI Act priorities for 2026. Further insights came from Mario Hernández Ramos, Chair of the Committee on AI (CAI), Council of Europe, on the Framework Convention context, and Dr. Sebastian Hallensleben, Chair of European AI Standardization, CEN–CENELEC, on AI standards as enablers of cross border ecosystems. Another roundtable discussion participated by Sergey Lagodinsky, Member of the European Parliament, and moderated by Sabina Carjan, Senior AI & IP Counsel, Infosys, examined system integrator readiness, governance structures, and intersections between India's Responsible AI approaches and EU regulatory perspectives, emphasizing the importance of aligning regulatory liability with real world commercial and technical contexts. Faiz Rahman and Ashish Tewari contributed practical industry perspectives on operationalizing Responsible AI, supported by colleagues Christopher Ford, Girish Babu, and Ritarshi Chakraborty, underscoring the critical role of government–industry collaboration in shaping safe, effective, and globally aligned AI ecosystems.. The day concluded with reflections from Asim Anwar, First Secretary, Embassy of India in Brussels, and Suman Kumar Kanth, Country Head – Belgium & Spain and Delivery Head – France & Mauritius, Infosys.

Responsible AI Enablement Workshop | February 2, 2026 | Melbourne

The Responsible AI Workshop held on February 2, 2026, in Melbourne went beyond leadership-led sessions to deliver a hands-on, immersive learning experience. Participants were introduced to Infosys' Responsible AI RBD core areas through a gamified, real-world simulation that mirrored how AI use cases are assessed, and governed across multiple internal teams. By role-playing real-life scenarios, leaders experienced firsthand how Responsible AI, ISG, DPO, Legal, and various other stakeholders collaborate in practice. The workshop reinforced that Responsible AI is not just a compliance requirement, but a strategic enabler that drives trust, scalability, and tangible business value for customers.

Infosys Responsible AI Literacy Workshop | February 4, 2026 | Sydney

The Responsible AI Workshop held on February 4, 2026, in Sydney built on the momentum of our enablement journey, delivering a highly engaging and outcome-driven experience. While leadership insights set the strategic context, the workshop's core strength lay in its interactive, gamified simulation of Infosys Responsible AI RBD core areas. Participants navigated a realistic AI use-case lifecycle, working through assessment, risk evaluation, and governance checkpoints across multiple stakeholder groups.

Through dynamic role-play, leaders actively collaborated as Responsible AI, ISG, DPO, Legal, and business teams experiencing firsthand how cross-functional alignment enables responsible innovation at scale. The Sydney session was particularly successful in driving rich discussions, practical insights, and clear ownership of next steps, reinforcing that Responsible AI is not just about risk mitigation, but about unlocking sustainable business value and competitive advantage for our customers.

India AI Impact Summit 2026 | February 16-20, 2026 | New Delhi

On February 16, 2026, leaders, policymakers, and industry experts gathered at Bharat Mandapam, New Delhi for Day 1 of the India AI Impact Summit 2026, a Global South–hosted program anchored on the sutras People, Planet, Progress and built around seven thematic Chakras shaping India's development focused AI agenda. From Infosys Responsible AI, Syed Ahmed, Global Head – Responsible AI Office, Infosys, led the high impact panel "**Trustworthy AI: Balancing Innovation and Regulation**" featuring Caitlin Searle – Australian High Commission, Mariagrazia Squicciarini – UNESCO, and Ankur Singh – Reserve Bank of India, who examined how to balance rapid AI innovation with transparent, accountable, and inclusive governance. Supporting the engagement were Infosys leaders Inderpreet Sawhney, Balakrishna (Bali) D., and Faiz Rahman, along with strong on ground representation from Ashish Tewari, Pankaj Grover, Ritarshi Chakraborty, and Pritesh Korde. The session reinforced the Summit's focus on practical, people centric AI outcomes, complemented by knowledge launches, a research symposium, and the AI Impact Expo showcasing scaled, real world AI deployments.

Responsible AI in Action

On February 20, 2026 as the Summit entered its final day with global leaders, policymakers, researchers, and industry practitioners converging for high stakes discussions on governance, safety, and global AI cooperation, concluding with **GPAI Council reviews** and the **Leaders' Declaration**, reinforcing the Summit's focus on People, Planet, and Progress.

From Infosys Responsible AI, Syed Ahmed, Global Head – Responsible AI Office, led a packed and high energy panel titled

"Responsible AI in Action: How Global Enterprises Are Building Trust at Scale." Joining him on stage were:

- Geeta Gurnani – Field CTO, IBM
- Sundar R. Nagalingam – Senior Director, AI Consulting Partners, NVIDIA
- Sunil Abraham – Public Policy Director, Meta

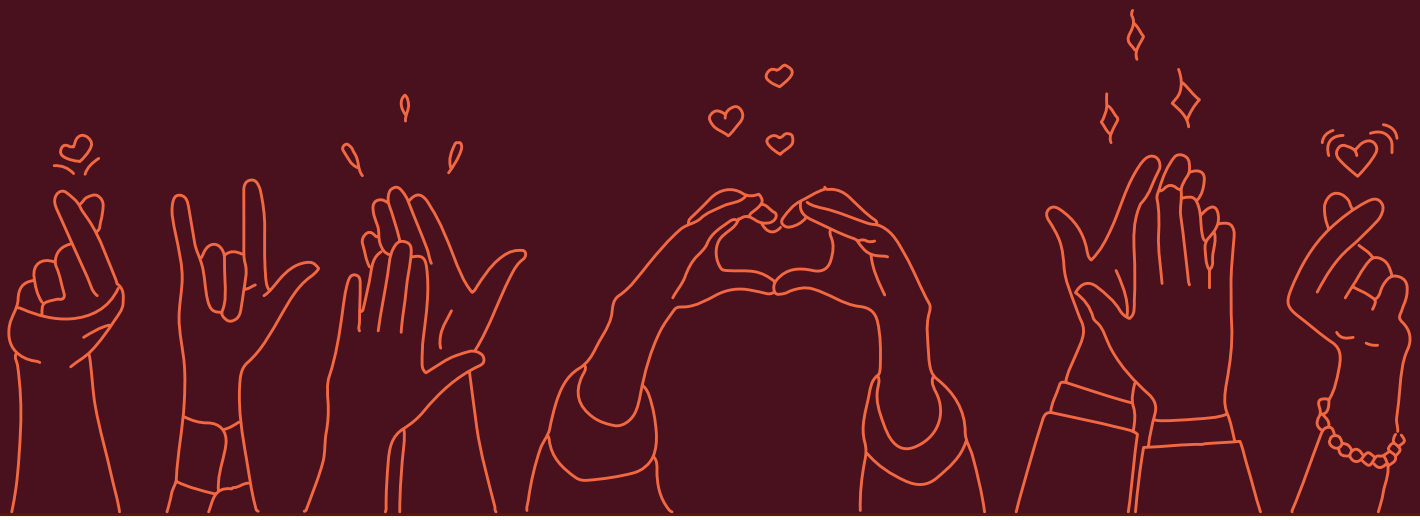
On ground presence from Infosys included Ashish Tewari, who played a key coordinating role for the team's engagement at the Summit, along with Ritarshi Chakraborty and Pritesh Korde, who attended the event and supported the Responsible AI representation throughout the day. The discussion moved beyond familiar talking points, addressing regulation, open innovation, enterprise grade AI safety, and real world deployment challenges with candid debate highlighting how enterprises are strengthening trust as a foundational principle and reinforcing the need for transparency, accountability, and harmonized global standards.

Launch of the Responsible AI Starter Pack

The Responsible AI Starter Pack[97] was launched on 20 February 2026 during the UK–India AIxcelerate Showcase 2025–26, held as part of the India AI Impact Summit in Delhi. The Starter Pack developed by the **Infosys Responsible AI Office** and released along with **The Dialogue**, and the **UK Government** serves as a practical toolkit to embed ethics, accountability, and trust into AI systems from the very beginning. During the Showcase, the Starter Pack was formally presented by Kiranmayee Janardhan, who highlighted its value in translating Responsible AI principles into actionable and scalable practices across sectors.

The UK–India AIxcelerate journey, spanning the Sandpit in Hyderabad, the Sprint in Bangalore, and concluding with the Showcase at the Summit, showcased the strength of cross border collaboration led by the **British High Commission in India** and the **Department of Science & Technology, Government of India**, alongside partners such as Infosys Topaz and Infosys Consulting.

The event also celebrated the Showcase winners and acknowledged the contributions of key partners, including Dr. Patricia Pinto, Joshua Bamford, Christi Thomas, Meemansa Agarwal, Shatam Bhattacharya, and Madhusmita Panda. The successful execution of the launch was made possible through the leadership of Syed Ahmed and Ashish Tewari, reinforcing the collective commitment to advancing Responsible AI from principle to practice.



[97] infosys.com/services/data-ai-topaz/insights/starter-pack-responsible-innovation.pdf

Infosys Responsible AI Toolkit – A Foundation for Ethical AI

The Open-Source Infosys Responsible AI Toolkit can be accessed from its public GitHub repo[98] also as project Salus[99]

Overview of the Responsible AI Toolkit

Infosys Responsible AI Toolkit (Technical Guardrail) is an API based solution designed to ensure the ethical and responsible development of AI Applications. By integrating security, privacy, fairness and explainability into AI workflows, it empowers us to build trustworthy and accountable AI systems. It includes below main components:

New Features

Below new features are developed and will be available soon in our next release (version 3.0.0).

- Explainability Enhancement Using Reasoning Models
- Second order explainable AI (SOXAI) Technique for Explainability Module
- Multi-lingual support for FM-Moderation Guardrails
- Signature and face masking in Privacy module
- Bulk document safety validation
- Moderation Layer Model based guardrails are improved with finetuned smaller models to improve latency and accuracy

Infosys Responsible AI Toolkit's *Model based guardrails is now available in* ***HuggingFace,***[100] *both as a repo and also as an interactive playground to explore. Star us and be a part of the Responsible AI Revolution!*

★★★★★

[98] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit
[99] https://github.com/salus-rai/salus
[100] https://huggingface.co/spaces/InfosysEnterprise/Moderation-playground

Contributors

We extend our sincere thanks to all the contributors who made this newsletter issue possible.

Srinivasan S - Policy Advocacy, Consultancy and Customer Outreach, Infosys Responsible AI Office

Mandanna A N - Head of Infosys Responsible AI Office, USA

Siva Elumalai - Senior Consultant, Infosys Responsible AI Office, India

Dakeshwar Verma - Senior Analyst - Data Science, Infosys Responsible AI Office, India

Utsav Lall - Senior Associate Consultant, Infosys Responsible AI Office, India

Pritesh Korde - Senior Associate Consultant, Infosys Responsible AI Office, India

Anie Juby - Industry Principal, Infosys Topaz Branding & Communications, Bangalore

Jossy Mathew - Senior Project Manager, Infosys Topaz Branding & Communications, Bangalore

Please reach out to responsibleai@infosys.com to know more about Responsible AI at Infosys.
We would be happy to have your feedback too.

AI IS SMARTER WHEN IT IS RESPONSIBLE

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises, and communities to create value. With 12,000+ AI assets, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems. Connect with us at infosystopaz@infosys.com

For more information, contact askus@infosys.com

Infosys®
Navigate your next