MARKET SCAN REPORT

MARCH 2026

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

IN FOCUS

VIBE CODING IN THE ENTERPRISE: INNOVATION, SECURITY AND GOVERNANCE

By Yogesh K. Dwivedi and Ibrahim A. Elgendy

Infosys®
Navigate your next

Foreword

There are moments when the direction of AI feels clearer than the pace itself. This month offered one of those moments. The global conversation is shifting in a noticeable way. The excitement around new capabilities is still there, but it now sits alongside a deeper appreciation for responsibility, trust and careful deployment. That shift is not accidental. It is the result of practitioners, policymakers and builders choosing day after day, to hold the line on what responsible development actually requires. That choice is what this edition is built around.

This edition takes a practical view of that transition, focusing on the signals shaping a more grounded AI landscape.

In many parts of the world, policymakers are adjusting their approaches in subtle but important ways. Instead of broad statements or high level ambition, they are turning their attention to the details that truly matter, clearer expectations for safety, more measured oversight of advanced systems and stronger protections for people who rely on AI in sensitive contexts. The tone feels steadier now. It suggests a maturing environment where governance and innovation are finally being seen as partner rather than opposites.

This shift has been reinforced by recent reminders of why thoughtful design and strong safeguards are essential. A few incidents highlighted how easily information can be distorted when checks are weak, how formal processes can be disrupted without proper verification and how cyber risks can grow quickly when automated tools are misused. None of these events were dramatic on their own, but together they revealed where gaps still exist. They encouraged a more realistic and responsible view of what it means to deploy AI in the world as it is, not the world we wish it to be.

What gives optimism is how the ecosystem responds. Researchers and practitioners are stepping back to focus on tools that strengthen safety in everyday use. Work is underway to better detect vulnerabilities before they spread, to make manipulated content easier to identify and to design systems that are stable and trustworthy under pressure. These efforts reflect a quiet but important maturity: the understanding that responsible AI development is not separate from progress, but an essential part of it.

This month's In Focus section reflects this evolving landscape. We are honored to spotlight the work of **Yogesh K. Dwivedi and Ibrahim A. Elgendy**, presented in their article ***"Vibe Coding in the Enterprise: Innovation, Security and Governance"***. The piece outlines how natural language driven development is transforming enterprise engineering and why secure, well governed practices must grow alongside it.

Looking ahead, the balance between ambition and accountability will define the path forward. AI will continue to accelerate, but its impact will depend on the care invested in making it safe, secure and aligned with the people it serves. Building with intention is no longer optional. It is the foundation on which meaningful progress rests.

Syed Ahmed
Global Head
Infosys Responsible AI Office

From the editor's desk

From Turning Points to True Progress: Reading the Signals in AI This Month

Some months remind us how fast AI is moving. This month reminded us why it matters. Across teams and industries, AI is shifting from a novelty to a normal part of daily work. People are designing with more care, reviewing their systems more closely and asking tougher but healthier questions about impact It feels as if the conversation is settling into a steadier, more thoughtful rhythm that values clarity as much as capability.

But the month also brought reminders of why such structure matters. Deepfake scams resurfaced with enough sophistication to trigger legal action and courts wrestled with AI generated false judgments and citations. A heartbreaking case of chatbot over dependence showed how emotional risks can quietly grow when systems aren't designed with boundaries in mind. And an AI assisted cyberattack on public agencies revealed how quickly misuse can scale when the wrong tools are in the wrong hands. These moments, subtle as they were, underlined the need for steady guardrails.

Yet alongside these warnings, the momentum kept building. Innovation continued its climb, but with a more grounded tone. GPT 5.4 made computer use automation feel closer to an everyday tool than an experimental feature. Mistral Small 4 pushed efficiency in a way that made advanced reasoning more accessible. And early steps in agentic platforms like Hermes Agent hinted at what long running digital assistants may soon enable. These updates were not loud breakthroughs; they were quiet foundations being laid for what comes next.

Safety research moved in parallel, which may be the clearest sign of maturity. Arbiter surfaced prompt vulnerabilities before they could be exploited and SCEP improved deepfake detection by learning to focus on the strongest evidence signals. Together they represent a mindset shift: capability is important, but resilience is what makes capability dependable.

This month's In Focus feature is one you won't want to skim. "Vibe Coding in the Enterprise: Innovation, Security and Governance," authored by Yogesh K. Dwivedi and Ibrahim A. Elgendy, dives into one of the most transformative shifts happening right now. If you want to understand how AI is rewriting the rules of software development, this is the piece to read. This edition also features a quarterly ***"Global AI Regulation Landscape,"*** capturing key regulatory movements and policy signals shaping AI governance worldwide helping enterprises stay aligned with what's coming next.

As we look ahead, the direction feels steadier and more purposeful. The developments of this month show that progress today is shaped not just by what AI can do, but by how carefully we choose to direct it. Each advance, each risk, each lesson contributes to a clearer understanding of what responsible AI must look like in practice.

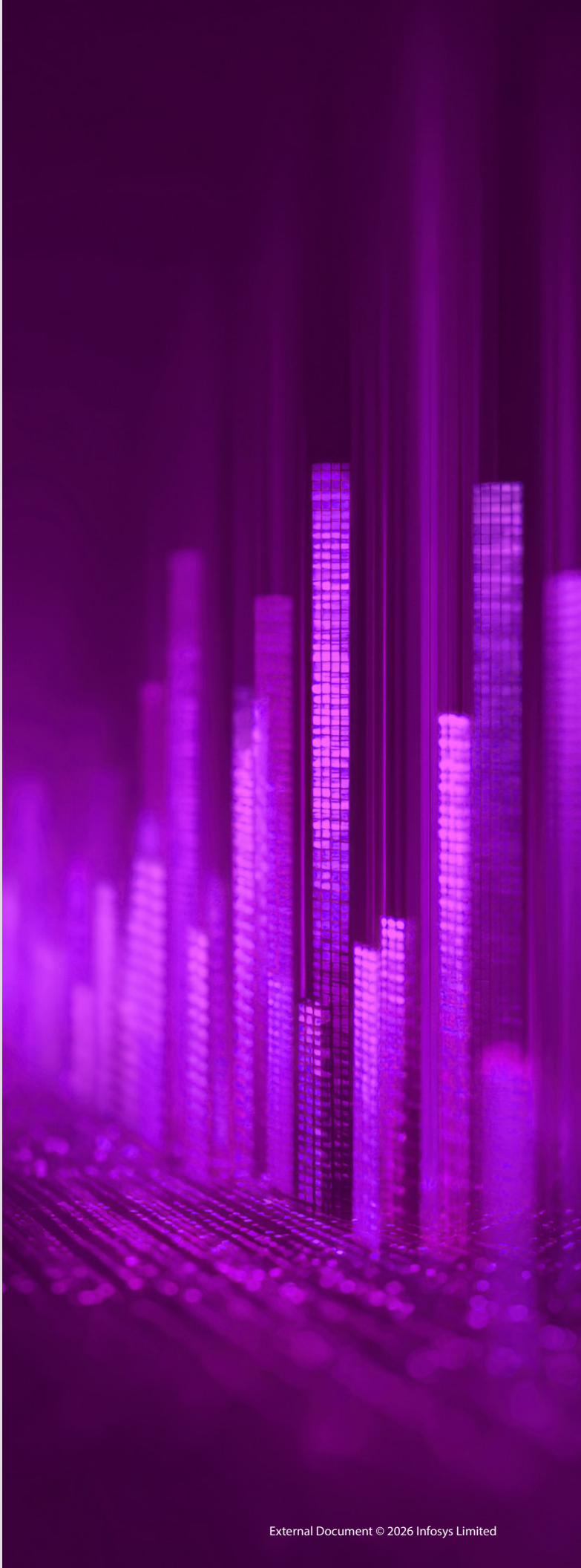
"Responsible AI is not only about capability, but about the choices that keep that capability worthy of trust."

The work before us is to build systems that are powerful yet dependable, ambitious yet accountable and capable yet grounded in the safeguards people expect. Ambition will continue to propel innovation, but it is our commitment to responsibility that will determine whether this progress truly serves its purpose. When these principles guide our choices, we move toward a future where AI earns trust not by promise, but by practice.

Warm regards,

Ashish Tewari
Head, Infosys Responsible AI Office (APAC, Middle East & 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

South Korea and Singapore Formalize Strategic AI Collaboration Through Bilateral Framework

During their bilateral summit, the leaders of South Korea and Singapore agreed to institutionalize and expand cooperation in artificial intelligence through the development of a structured AI Cooperation Framework and the conclusion of targeted memoranda of understanding covering AI applications in public safety and intellectual property governance, while also committing to enhanced joint research initiatives in frontier domains such as quantum computing, satellite technologies and environmental data analytics, underscoring a coordinated, innovation-driven approach to strengthening long-term technological and digital collaboration between the two nations.[1]

UNESCO and CENIA Join Forces to Promote Ethical and People-Centred AI in Education

UNESCO's Regional Office in Santiago, Chile and the Chile's National Centre for Artificial Intelligence (CENIA) have signed a cooperation agreement to strengthen the ethical and responsible use of artificial intelligence in education across Chile and the wider Latin American region. The partnership focuses on developing digital skills, AI literacy and people-centred AI systems for teachers, students, technical teams and policy-makers. It also supports the introduction of Latam-GPT, the region's first open LLM designed for and by Latin America and the Caribbean, which will contribute to UNESCO's new Regional Observatory on AI in Education. With CENIA's strong background in research and training, the collaboration aims to expand responsible innovation, encourage multi-sector dialogue and ensure that AI development follows ethical standards that protect inclusion, transparency and human rights.[2]

Global Cybersecurity Partners Release Joint Guidance to Reduce Supply Chain Risks in AI

On 5 March 2026, cybersecurity agencies from the United States, Japan, New Zealand, South Korea, Singapore, the United Kingdom and Canada joined Australia in issuing joint guidance to help organizations manage growing supply chain risks linked to artificial intelligence (AI) and machine learning (ML). The guidance explains that AI and ML systems can improve efficiency and decision-making, but they may also introduce serious risks such as data poisoning, model-serialization attacks and weaknesses in AI software if not properly secured. To address these threats, the guidance recommends using trusted data sources, applying secure file formats and performing continuous performance and security testing. It is designed to help organizations and teams developing or deploying AI systems understand these risks, ask the right questions when sourcing third-party tools and strengthen the overall resilience of AI and ML supply chains.[3]

Australia and Canada Call for Fair and Responsible AI in Creative Industries

In this joint statement, the Australasian Performing Right Association and Australasian Mechanical Copyright Owners Society (APRA AMCOS) and the Society of Composers, Authors and Music Publishers of Canada (SOCAN) explain that Australia and Canada share a strong commitment to protecting creators as artificial intelligence rapidly changes the creative economy. Representing nearly 400,000 songwriters, composers and music publishers, the two organizations stress that AI must not grow in ways that help only large technology companies while harming the creators whose work trains these systems. They note that both countries reject the false claim that innovation and creator rights cannot exist together, pointing to Australia's decision to block copyright exceptions for AI training and Canada's similar efforts. The statement calls for a fair system built on consent, transparency and proper payment, ensuring that AI supports human creativity instead of replacing it. It also highlights the urgent need to protect Indigenous Cultural Intellectual Property, which AI systems are already using without permission. Both organizations say this moment gives Australia and Canada a chance to lead the world by creating AI rules that treat cultural knowledge as a national asset and strengthen partnerships between creators and technology.[4]

[1] https://koreajoongangdaily.joins.com/news/2026-03-02/national/diplomacy/Korea-Singapore-agree-on-trade-AI-nuclear-energy-cooperation-at-summit/2534969
[2] https://www.unesco.org/en/articles/unesco-and-cenia-strengthen-partnership-advance-ethical-artificial-intelligence-focus-education
[3] https://www.cyber.gc.ca/en/news-events/joint-guidance-supply-chain-risks-mitigations-artificial-intelligence-machine-learning
[4] https://www.socan.com/joint-statement-on-creative-industries-and-artificial-intelligence-australia-and-canada/

US

U.S. Drafts Strict AI Procurement Guidelines

The United States government has reportedly drafted strict new guidelines for artificial intelligence procurement that would require companies seeking federal contracts to grant the government irrevocable rights to use their AI models for any lawful purpose, according to a report by the. The proposed rules, developed by the General Services Administration, come amid an escalating dispute between the Pentagon and AI company Anthropic, which had resisted allowing unrestricted use of its systems due to concerns about potential applications such as surveillance and autonomous weapons. Following the disagreement, the Pentagon labeled Anthropic a "supply-chain risk" and barred its technology from certain military-related contracts, while the government also canceled the company's OneGov contract that had enabled broader federal access to its tools. The draft guidelines further require AI contractors to avoid embedding partisan or ideological bias in their systems and to disclose any modifications made to comply with foreign regulations, reflecting a broader U.S. effort to tighten oversight of AI procurement and ensure national security, transparency and government control over AI technologies used in federal operations.[5]

Bipartisan Push to Open Federal Data for AI Development

U.S. Senators Ted Budd and Andy Kim have introduced the Artificial Intelligence-Ready Data Act, a bipartisan bill aimed at unlocking federal government data sets to better train and strengthen American AI models. The legislation recognizes that the U.S. government holds the world's largest publicly generated data assets and seeks to make these resources more accessible, standardized and usable for AI training, while maintaining transparency, safety and public trust. By modernizing how agencies publish and structure data, the bill aims to support innovation across sectors such as healthcare, manufacturing, climate science and weather forecasting, ensuring U.S. leadership in AI competitiveness. The proposal has garnered support from major technology and industry stakeholders, underscoring its role in enabling high-quality, secure and "AI-ready" public data for researchers and developers of all sizes.[6]

[5] https://www.reuters.com/business/media-telecom/us-draws-up-strict-new-ai-guidelines-amid-anthropic-clash-ft-reports-2026-03-07/
[6] https://www.budd.senate.gov/2026/03/17/budd-kim-introduce-bipartisan-bill-opening-government-data-sets-to-better-train-american-ai-models/

Blackburn's AI Framework to Create a Unified National Rulebook for the United States

Senator Marsha Blackburn's discussion draft introduces a national framework in the United States that aims to create one clear federal rulebook for artificial intelligence to protect children, creators, conservatives and communities while supporting strong innovation. The proposal places safety duties on AI developers, strengthens online protections for minors and sets rules to hold companies accountable for harmful AI systems. It also protects creators by addressing unauthorized digital replicas and improving transparency around how copyrighted material is used in AI training. To prevent political bias, it requires independent audits and ensures government use of neutral models. The framework further supports communities through job impact monitoring, responsible data center practices and programs to assess advanced AI risks, while boosting national research and innovation efforts.[7]

UK

Online Advertising Taskforce Progress Report 2025 Shows UK's Key Advances in Safer AI-Responsible Advertising

The United Kingdom's Online Advertising Taskforce Progress Report 2025 highlights major improvements in strengthening trust, transparency and accountability across the online advertising ecosystem, while also addressing how artificial intelligence is shaping advertising practices. The report explains how government and industry worked together to reduce illegal advertising, limit harms and protect children from age-restricted adverts through six industry-led working groups. One of these groups focused specifically on AI and supported the release of a best-practice guide for the responsible use of generative AI in advertising, which is now being adopted across the sector. The report also shows progress in expanding the influencer marketing code of conduct and confirms that targeting tools are largely effective in preventing children from seeing restricted ads. These advances set the foundation for the Taskforce's ambitious goals for 2026.[8]

UK Launches Major Public Consultation to Explore New Rules Protecting Children on Social Media, Gaming and AI Chatbots

The UK government has opened a major national consultation asking parents, young people, teachers and experts how best to protect children online, covering possible measures such as social-media age limits, overnight curfews, controls on AI chatbots and safeguards in gaming platforms. Led by the Department for Science, Innovation and Technology, the initiative responds to growing concern that constant screen time, addictive platform features and unregulated AI interactions are harming children's sleep, attention and mental health. The consultation will gather views on whether

[7] https://www.blackburn.senate.gov/2026/3/technology/blackburn-releases-discussion-draft-of-national-policy-framework-for-artificial-intelligence/3b3b6458-b6c7-478b-9859-374949586765

[8] https://www.gov.uk/government/publications/online-advertising-taskforce-progress-report-2025/online-advertising-taskforce-progress-report-2025

under-16s should be banned from social media, whether platforms should disable infinite scrolling and autoplay for children, how age-verification should be strengthened and how families can be better supported in managing digital life. Over the next three months, the government will also run real-world pilots with teenagers to test ideas like curfews and screen-time limits, with results shaping future rules that ministers can implement quickly using new legislative powers.[9]

UK refines investment screening rules to better oversee sensitive sectors, including AI

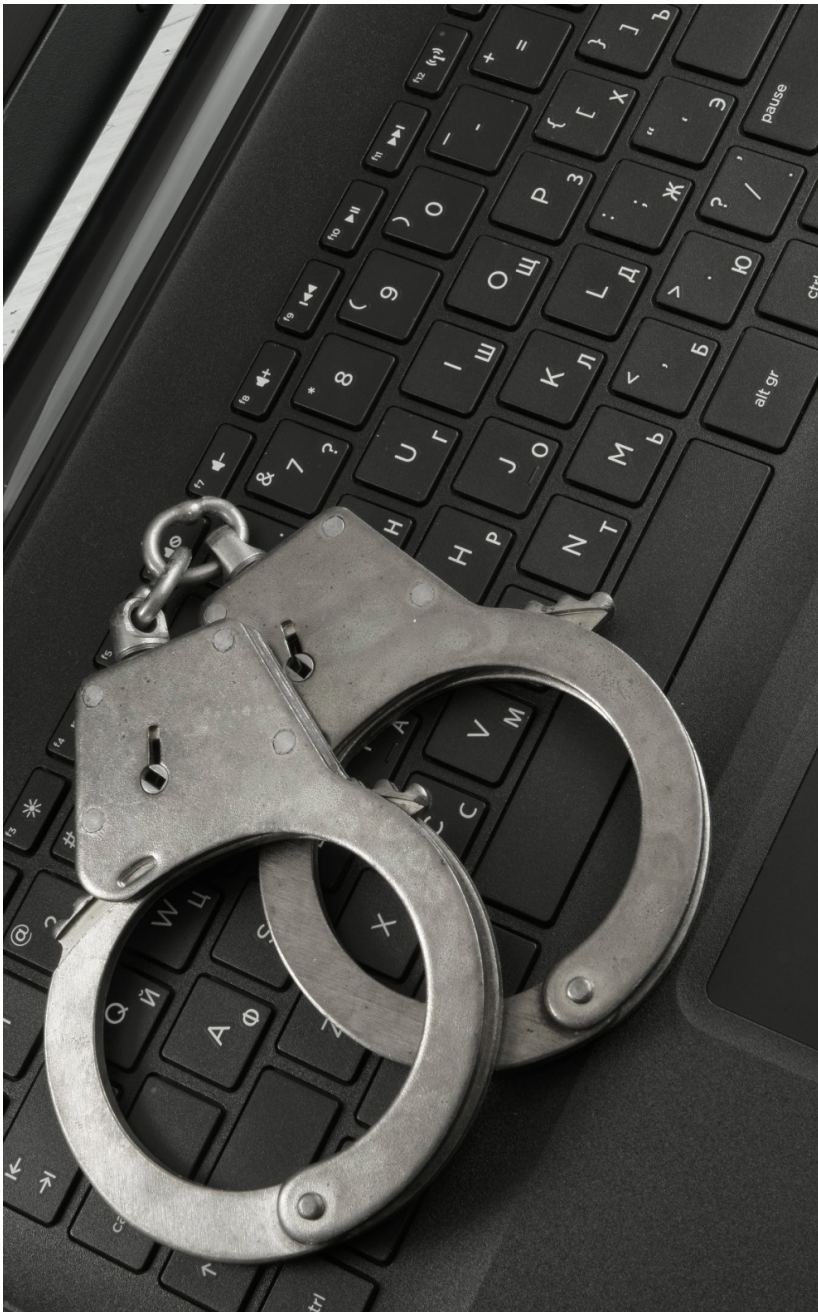
The UK government has updated its National Security and Investment Act rules to ensure that only business deals posing real national security risks must be reported, making the system clearer and easier for companies to follow. After reviewing public feedback, the government decided to narrow and clarify the definitions of sensitive sectors such as artificial intelligence, semiconductors, critical minerals, communications, energy, data infrastructure and emergency services. It is also creating a new category for the water sector to reflect its growing importance. Updated guidance will help companies understand whether they need to notify the government about an acquisition and secondary legislation will be introduced later in the year to put these changes into effect. The overall goal is to keep the UK safe while reducing unnecessary paperwork and supporting a business-friendly environment.[10]

UK fraud strategy adapts to new tech threats, including AI-driven crime

The UK government's Fraud Strategy 2026–2029 sets out a national plan to fight fraud as criminals increasingly use advanced digital tools, including artificial intelligence, to deceive people and businesses. The strategy explains that fraud is now one of the most widespread crimes in the country and that criminals are using new technologies and online systems to launch more complex attacks, which require modern responses. To address this, the government will disrupt criminal networks by strengthening online and telecom protections, expand public awareness through national guidance and improve victim support with better reporting and investigative systems. It also introduces the new Online Crime Centre, which will use emerging technologies to detect and block fraud at its source. The goal is to make the UK safer, more resilient and better prepared for technology-driven fraud challenges.[11]

UK Government's Consultation on Copyright and Artificial Intelligence Sets Out Plans to Balance Creative Rights and AI Innovation

In the UK Government's consultation on Copyright and Artificial Intelligence, the government outlines its plan to create a clear framework that protects human creativity while supporting innovation in artificial intelligence. The consultation explains that both creators and AI developers currently face legal uncertainty: rights holders struggle to control and be paid for the use of their works in AI training, while developers face unclear rules that slow investment and adoption. The government proposes a system where rights holders can reserve their rights, alongside an exception that allows AI developers lawful access to data when rights are not reserved. It also stresses the need for strong transparency from developers about training data and generated content. These proposals aim to give creators fair control and payment, give AI developers clear rules for training models and build trust between both sectors.[12]



[9] https://www.gov.uk/government/news/landmark-consultation-seeks-views-on-major-measures-to-protect-children-on-social-media-gaming-platforms-and-ai-chatbots

[10] https://www.gov.uk/government/consultations/consultation-on-the-nsi-act-notifiable-acquisition-regulations/outcome/national-security-and-investment-act-notifiable-acquisition-specification-of-qualifying-entities-regulations-2021-consultation-response

[11] https://assets.publishing.service.gov.uk/media/69ae77ddc78869bf8eb8a509/fraud-strategy-web.pdf

[12] https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence

Europe

EU Issues Updated Guidelines to Help Teachers Use AI and Student Data Safely and Ethically in Schools

The updated EU Guidelines on the Ethical Use of Artificial Intelligence and Data in Teaching and Learning provide clear, practical support to help teachers and school staff use AI tools responsibly. The document explains that AI is becoming a common part of education used for language learning, lesson planning and classroom support while also raising ethical, legal and teaching related challenges. The guidelines were updated because AI use in schools has increased, new laws like the AI Act require careful compliance and educators need stronger skills in ethical and critical AI use. Aimed mainly at primary and secondary educators, the guidelines help teachers build confidence, improve digital skills and understand risks linked to AI and data. They include guiding questions, real classroom scenarios, explanations of key laws such as GDPR and the AI Act and an updated glossary of important terms. Additional background materials support deeper learning. First published in 2022 and updated in 2026, the guidelines are supported by Erasmus+ funding and aim to help teachers use AI more safely, ethically and effectively in everyday practice.[13]

European Commission, Interpol and Over 100 Organizations Urge Global Ban on AI "Nudification" Tools

A coalition of more than 100 organizations including the European Commission, Interpol and Amnesty International has called for governments worldwide to outlaw artificial-intelligence "nudification" tools that generate fake nude images from ordinary photographs. The groups warn that these tools are increasingly being used to create non-consensual sexualized images, often targeting women and children and have been linked to harassment, blackmail and the spread of child sexual abuse material. The call intensified after a controversy involving the AI chatbot Grok, where users manipulated photos to produce sexualized deepfake images that circulated widely online. Campaigners argue that the technology enables large-scale exploitation and violates fundamental rights, urging governments to introduce clear legal bans, hold developers and platforms accountable and implement stronger safeguards to prevent the creation and distribution of such harmful AI-generated content.[14]

EU AI Act: Detailed Arrangements for Evaluations and Enforcement Proceedings

The European Commission's initiative on detailed arrangements for evaluations and proceedings under the EU Artificial

[13] https://education.ec.europa.eu/kk/focus-topics/digital-education/action-plan/ethical-guidelines-for-educators-on-using-ai
[14] https://www.euronews.com/next/2026/02/11/european-commission-interpol-and-100-others-call-to-outlaw-ai-nudification-tools

Intelligence Act sets out how general-purpose AI (GPAI) models will be assessed and how enforcement actions will be conducted to ensure legal certainty and proportionality. It clarifies the Commission's powers to evaluate AI models directly or through independent experts, including requesting controlled access to technical interfaces, model components, or source code when necessary, while safeguarding confidentiality and business secrets. The initiative also defines procedural safeguards for providers, including clear rules on initiating and closing proceedings, opportunities to respond to preliminary findings and rights of access to non-confidential case files before penalties are imposed. Overall, the proposal strengthens trust and transparency in AI oversight by balancing effective supervision with due process and protections for innovation under Regulation (EU) 2024/1689.[15]

Brazil

Brazil's New Bill on Technological Governance, Algorithmic Transparency and Digital Equity in Public Administration

Brazil's Bill No. 728/2026 introduces a national policy to guide how technology and automated systems are used in public administration. The bill requires government bodies and service providers to conduct a Technological and Fundamental Rights Impact Assessment before creating or contracting any automated system that may affect people's rights. It also requires a 30 day public consultation, except in cases involving national security. To prevent dependence on a single vendor, the bill mandates anti lock in safeguards such as interoperability, data portability, documented APIs, escrow of essential components, service continuity and support during system transitions, with a preference for open source solutions. It also amends Brazil's public procurement law to include these rules, ensuring fair, transparent and responsible use of technology across government systems.[16]

Indonesia

Indonesia Issues Joint Seven-Minister Decree Establishing Guidelines for Responsible Use of Artificial Intelligence and Digital Technology in Education

The Indonesian government has introduced a Joint Ministerial Decree signed by seven cabinet ministers to regulate the use of digital technology and artificial intelligence (AI) across the country's education system, ranging from early childhood education to universities. Coordinated by the Ministry for Human Development and Culture, the policy establishes age-appropriate guidelines governing when and how students can use AI, including recommendations on minimum age, types of permitted applications and appropriate screen-time limits.

[15] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/16472-Artificial-Intelligence-Act-detailed-arrangements-on-evaluations-and-proceedings_en
[16] https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2604315

The decree emphasizes stricter controls for younger learners, particularly in early childhood and primary education, where the use of AI and digital tools is closely regulated to protect children's cognitive development and limit exposure to inappropriate content. At the primary and secondary levels, the regulation also restricts the use of "instant-answer" AI applications that automatically generate solutions to assignments, while still allowing educational AI tools designed for structured learning activities such as simulations or guided exercises. As students progress to higher education, the rules become more flexible, reflecting the expectation that older learners can use technology more responsibly. The policy, signed by ministers responsible for home affairs, religious affairs, primary and secondary education, higher education and science, communications and digital affairs, population and family development and women's empowerment and child protection, aims to ensure that AI is used as a supportive educational tool while minimizing risks and promoting responsible digital behaviour among Indonesian students.[17]

Australia

Australia Warns It May Block AI Apps and Search Tools That Don't Protect Children Under New Online Safety Rules

Australia's internet regulator has signalled that it may force search engines and app stores to block AI services that fail to verify user ages, as the country prepares to enforce strict new rules to protect children from harmful online content. From March 9, AI platforms like ChatGPT and other chatbot tools must prevent users under 18 from accessing content involving pornography, extreme violence, self-harm, or eating disorders or face heavy fines. The regulator says many AI companies still have no public plans to comply, raising concerns about youth safety and emotional manipulation by chatbots. While a few major platforms have begun rolling out age-checks or content filters, most have not. Australia, which previously banned social media for teenagers over mental-health risks, says it is ready to pressure app stores and search engines if AI firms fail to follow the law.[18]

Australia's Finance Department Outlines a Clear and Responsible Framework for Safe AI Use Across Government

The Department of Finance states that it is committed to using artificial intelligence safely, transparently and responsibly, seeing AI as a helpful tool for improving productivity and public services while maintaining trust and accountability. It explains that its AI Delivery and Enablement function leads the Australian Public Service AI Plan and supports agencies by coordinating Chief AI Officers, removing common barriers and tracking progress. The department uses secure platforms like GovAI and GovAI Chat to help staff explore AI safely, while following national AI ethics principles, assurance frameworks and impact assessments to identify risks and apply safeguards. Finance employees use AI for everyday tasks such as analysis, summarising information and drafting content, but any

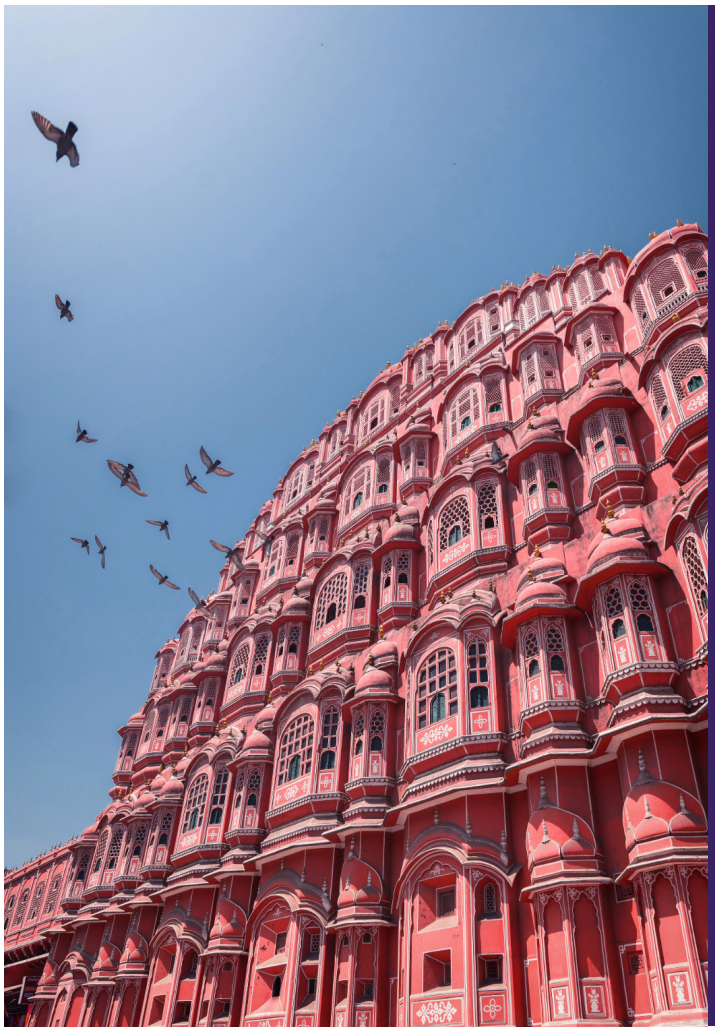
[17] https://www.kemendukbangga.go.id/posts/671059b9-d587-4d94-a067-75df9134f42a-pemerintah-terbitkan-skb-tujuh-menteri-soal-ai-dalam-pendidikan?utm_source=substack&utm_medium=email

[18] https://www.reuters.com/business/media-telecom/australia-says-it-may-go-after-app-stores-search-engines-ai-age-crackdown-2026-03-01/

higher-risk use requires approval and registration under strict rules. An AI Governance Committee, along with an Accountable Official and Chief AI Officer, oversees risk management, policy compliance and cultural awareness across the department. Finance also provides clear policies, training and security controls, reviewing them regularly to ensure safe use. The department monitors AI performance, manages risks at an enterprise level and refuses any AI use that remains medium or high risk after safeguards. It commits to reviewing this transparency statement annually and invites questions via its official contact channel.[19]

Parliament questions Australian Securities and Investments Commission on Artificial Intelligence audits

In a hearing of the Joint Committee on Corporations and Financial Services, the Australian Securities and Investments Commission (ASIC) was asked how it oversees the use of Artificial Intelligence (AI) in audits by the Big Four firms. Senator Deborah O'Neill cited reports that Deloitte Australia told audit staff not to upload confidential client or firm information into Chat Generative Pre trained Transformer (ChatGPT) and noted 28 cases where Klynveld Peat Marwick Goerdeler (KPMG) staff used AI to cheat on internal exams. ASIC Commissioner Kate O'Rourke said AI use has surged over the past 18 months and some examples were concerning. She explained that government policy is to rely on existing rules while adding safeguards if needed, stressed strong data controls and training and clarified that ASIC regulates individual registered company auditors not audit firms and has not yet engaged directly with auditors on AI.[20]



India

Kerala Introduces Comprehensive 'Cyber Safety Protocol 2026' to Protect Students in India's AI Driven Education Era

Kerala in India has launched the 'Cyber Safety Protocol 2026,' a detailed framework designed to protect students from new digital and AI related risks in schools. Developed by the Kerala Infrastructure and Technology for Education (KITE), the protocol focuses on creating a safe digital learning environment through clear rules for teachers, school leaders, parents and students. It sets out 13 core objectives, such as helping students understand the risks of sharing personal data with AI tools, building critical thinking skills for online content and spotting misinformation. It also defines nine key operational areas and gives school heads 17 specific responsibilities, including supervised internet access and strict privacy measures like avoiding real time CCTV monitoring via private servers. The protocol also discourages homework that requires the internet to help students who lack home connectivity. Overall, the initiative aims to ensure responsible and safe digital behavior as AI becomes more common in education.[21]

Karnataka Announces Social Media Ban for Children Under 16 in the 2026–27 Budget

In its 2026–27 Budget, the Karnataka Government announced that social media use will be banned for children under 16, a move introduced by Chief Minister Siddaramaiah to reduce the negative effects of rising mobile phone use

[19] https://www.finance.gov.au/about-us/artificial-intelligence-ai-transparency-statement
[20] https://www.accountingtimes.com.au/profession/asic-questioned-on-regulation-of-ai-and-audits-in-recent-hearing
[21] https://economictimes.indiatimes.com/tech/artificial-intelligence/kerala-unveils-cyber-safety-protocol-to-safeguard-students-in-ai-era/articleshow/129253800.cms?from=mdr

among young students. This decision follows global actions, such as Australia's ban on platforms like TikTok, YouTube, Instagram and Facebook for teens and similar discussions in countries from Denmark to New Zealand. Andhra Pradesh is also considering a comparable restriction. The Karnataka Government had been studying the pros and cons of limiting mobile use for minors and national leaders have recently supported age based rules for social media. This step aims to create safer digital habits and protect students as online risks grow in the age of AI.[22]

China

China issues security warning over OpenClaw AI vulnerabilities

The National Internet Emergency Center in Beijing issued a security warning on March 10, advising that the rapid increase in downloads and one-click cloud deployments of the OpenClaw AI agent has created serious risks. OpenClaw can directly control a computer through natural language instructions, but its weak security settings make it easy for attackers to take full system control if a flaw is found. The warning noted several dangers, including hidden malicious commands on web pages that could trick OpenClaw into leaking system keys, accidental deletion of important files due to misunderstood instructions and unsafe plugins capable of stealing sensitive data. The centre urged organizations and users to strengthen network controls, isolate OpenClaw's environment, manage credentials carefully, avoid unverified plugins and install updates and patches promptly.[23]

China's Supreme Court outlines rules for Artificial Intelligence cases and innovation safeguards

In China, the Supreme People's Court reported at the Fourth Session of the 14th National People's Congress in Beijing that courts must handle cases involving Artificial Intelligence (AI) according to the law while allowing a reasonable "fault-tolerance" space to support scientific and technological innovation. The report highlighted China's efforts to protect intellectual property, noting 496,000 cases concluded and stronger action against infringement. It also explained that if a generative AI service makes an error but the developer has exercised proper care and no real harm occurs, it may not be considered an infringement. At the same time, the court stressed strict action against the misuse of AI that harms rights or disrupts social order, aiming to balance safe innovation with legal responsibility.[24]

[22] https://www.thehindu.com/news/national/karnataka/karnataka-proposes-to-ban-social-media-for-children-under-16-years/article70710597.ece
[23] https://www.news.cn/tech/20260310/959f13d18edb4759ae031a5e30523d23/c.html
[24] https://www.court.gov.cn/zixun/xiangqing/492681.html

Singapore

Singapore Issues New Guide on Responsible Use of Generative AI in the Legal Sector

The Ministry of Law in Singapore has released a new Guide for Using Generative AI in the Legal Sector, aimed at helping lawyers, in house counsel, paralegals and law students use AI tools safely and responsibly. The guide explains the purpose and scope of generative AI in legal work and highlights how such tools are increasingly used for tasks like research, drafting and knowledge management. It sets out three core principles professional ethics, confidentiality and transparency emphasizing that legal professionals remain fully responsible for their work, must protect client data and should inform clients when AI is used. The guide also offers a five step framework for adopting AI, including assessing needs, evaluating tools, training staff and continually reviewing systems. It includes case studies showing how different law practices have used AI effectively while maintaining high standards. The Ministry developed the guide through public consultation and plans to update it as AI technology evolves.[25]

Azerbaijan

Azerbaijan Proposes Criminal Liability for AI Generated Fake Media

The Milli Majlis of Azerbaijan has introduced draft amendments to the Criminal Code, Criminal Procedure Code and laws on information and media to address the misuse of artificial intelligence and special software for creating fake photo, video, or audio content (deepfakes). The proposed legislation criminalizes the production or dissemination of AI generated materials created without a person's consent using their image or voice, especially when such content misrepresents reality. Penalties range from fines and community service to imprisonment, with harsher sentences for aggravated circumstances, including acts committed by groups, against multiple victims, to damage a person's honour and dignity, or in connection with a victim's official or public duties. The initiative, discussed in its first reading in March 2026, aims to protect personal rights, prevent digital misinformation and strengthen accountability for harmful AI use in media and online platforms.[26]

[25] https://www.mlaw.gov.sg/files/Guide_for_using_Generative_AI_in_the_Legal_Sector__Published_on_6_Mar_2026_.pdf
[26] https://meclis.gov.az/index.php?lang=az

Kenya

Kenya introduces comprehensive AI Bill to regulate risks, strengthen oversight and protect citizens

Kenya has taken a major step toward formal AI governance with the introduction of the Artificial Intelligence Bill, 2026, sponsored by Senator Karen Nyamu. The Bill proposes the creation of the Office of the Artificial Intelligence Commissioner, which would serve as the main regulator responsible for overseeing AI systems nationwide, monitoring risks, advising the government and enforcing rules. It sets out a risk based regulatory model aligned with global standards, placing strict requirements on high risk AI applications and ensuring responsible development. The Bill also establishes an Advisory Committee to support policymaking and recommends tools such as AI sandboxes, ethical guidelines, transparency rules and public literacy programs to encourage safe use. It outlines penalties up to Sh5 million in fines or two years' imprisonment for misuse of AI, including deploying banned systems or creating harmful or misleading content like deepfakes. Strong data governance requirements would compel organizations to record and protect data used in AI training. Overall, the Bill aims to ensure that AI in Kenya is used ethically, transparently and safely while supporting innovation and aligning with the country's National AI Strategy 2025-2030.[27]

South Korea

South Korea's Personal Information Commission Pushes for Clearer and More Transparent Data Policies in the Age of Generative AI

South Korea's Personal Information Protection Commission held a major meeting with leading global and domestic generative AI companies to address the urgent need for clearer, more transparent personal information processing policies in AI services. The Commission explained that its evaluation system, designed to improve transparency and accountability, has shown progress across multiple sectors, but generative AI still falls behind due to vague data categories, unclear legal bases and poor accessibility of policies. Companies also struggled with explaining how user input data is processed and aligning local policies with global headquarters. At the meeting, participants agreed that simpler, more specific explanations are necessary to build user trust. Chairperson Song Kyung hee emphasized that users must easily

[27] https://cioafrica.co/kenya-tables-ai-bill-proposing-regulator-risk-rules-and-penalties/

understand how their information is used and the Commission announced plans to update its guidelines and offer briefing sessions to help AI companies create clearer, user friendly data processing policies.[28]

AmCham and South Korea's Privacy Regulator Discuss Cross-Border Data Transfers and AI-Era Data Governance

At a policy meeting hosted by the American Chamber of Commerce in Korea (AmCham), South Korea's top privacy regulator engaged with business leaders to discuss challenges surrounding cross-border data transfers and the evolving regulatory framework for personal data in the age of artificial intelligence. During the discussion, Song Kyung-hee, chairperson of the Personal Information Protection Commission (PIPC), outlined the government's policy direction for strengthening South Korea's data protection system in 2026 while supporting digital innovation. She emphasized the need for balanced regulations that both safeguard personal information and enable international data flows crucial for global business operations, particularly as AI technologies increasingly rely on large-scale datasets. Participants also explored ways for regulators and companies to collaborate on building a trustworthy data ecosystem, ensuring compliance with privacy rules while maintaining South Korea's competitiveness in the global digital economy.[29]

Turkey

Türkiye issues guidance on risks and responsibilities for agentic AI systems

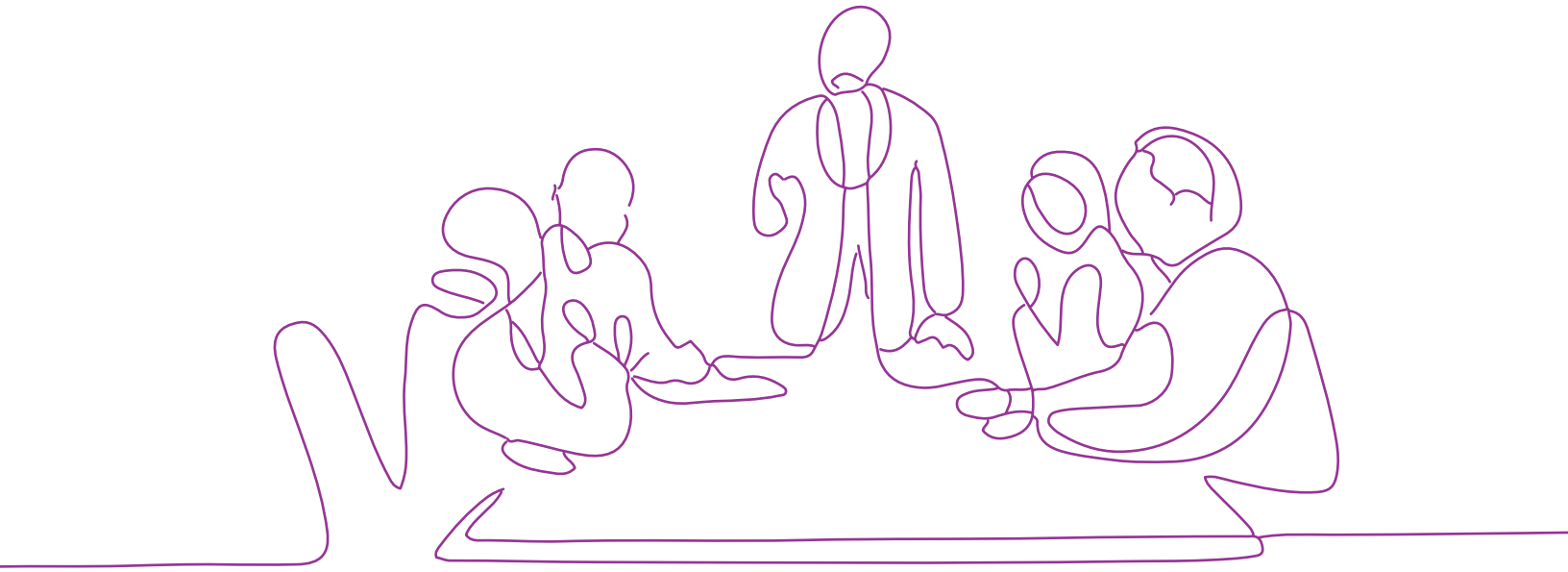
Türkiye's Personal Data Protection Authority published a report on agentic artificial intelligence, explaining how these systems work and what they mean for personal data protection. The report states that agentic AI consists of autonomous AI agents that can assess their environment, adapt to changing conditions and pursue multi step goals without constant human instruction, which makes them different from traditional AI systems. It notes that such systems may be used in areas like research and development, customer support, finance, healthcare and incident management, but also create risks involving transparency, accountability, bias, security weaknesses and the accuracy of outputs. The report warns that the autonomous, multi step nature of agentic AI can broaden data use, enable inference based profiling and complicate oversight and legal responsibility. To reduce these risks, it recommends meaningful human oversight, clear transparency measures, privacy by design practices, defined roles and responsibilities and the use of risk assessment tools, including data protection impact assessments.[30]

[28] https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=11856&utm_source=substack&utm_medium=email
[29] https://www.koreaherald.com/article/10693896
[30] https://digitalpolicyalert.org/event/38447-personal-data-protection-authority-published-report-on-agentic-ai

Vietnam

Vietnam's National AI Regulation: Key Provisions of the 2026 Law Taking Effect on March 1, 2026

The Law on Artificial Intelligence (Law No. 134/2025/QH15), which takes effect on March 1, 2026, sets clear rules for how AI must be developed, used and managed in Vietnam. It explains what AI is, who the law applies to and how developers, providers, deployers and users must act responsibly. The law places humans first, protects rights, requires fairness and transparency and ensures AI made content is clearly labeled. It bans harmful uses of AI and supports safe testing, training and innovation. It also promotes national AI infrastructure, strong data systems and skilled human resources. By setting risk levels and defining duties during incidents, the law guides safe AI growth while building public trust and supporting national development.[31]



[31] https://static3.luatvietnam.vn/uploaded/vietlawfile/2026/1/law_134_2025_qh15_090126160152.pdf

Standards & Policy Reports

Key Challenges in Monitoring Deployed AI Systems: Insights from NIST's CAISI Research and Practitioner Workshops

In the new report NIST AI 800 4: Challenges to the Monitoring of Deployed AI Systems, the Center for AI Standards and Innovation describes how the rapid integration of AI into commercial and government settings has created an urgent need for strong post deployment monitoring. The report explains that monitoring AI is more complex than traditional digital systems because AI behaves unpredictably and can vary over time. Through practitioner workshops and a review of existing research, CAISI identifies six major monitoring categories functionality, operations, human factors, security, compliance and large scale impacts and highlights significant gaps, barriers and open questions across each area. The report emphasizes that today's monitoring ecosystem is fragmented, with limited standards, immature information sharing practices and challenges such as detecting performance drift or evaluating human AI interactions. It concludes that addressing these issues is essential for responsible, confident and widespread adoption of AI systems.[32]

Engagements on Canada's Next AI Strategy: Summary of Inputs – How Public Feedback Is Shaping Canada's 2026 AI Vision

The report Engagements on Canada's Next AI Strategy: Summary of Inputs explains how the Canadian government gathered ideas from more than 11,300 people to help build a new national AI strategy for 2026. It describes how citizens, experts, industry and community groups shared their thoughts during a month long consultation in October 2025, highlighting the need for safe and ethical AI, strong data protection, Canadian owned infrastructure, better digital skills and trusted AI rules. Participants also stressed supporting Canadian AI companies, improving public services with AI and strengthening cybersecurity. The government used advanced AI tools to analyze over 64,600 responses quickly and fairly. These insights will guide Canada's renewed AI strategy to support innovation while protecting people, privacy and public trust.[33]

UK Government Outlines Opportunities, Risks and Consumer Protections for Agentic AI Systems

The UK Government's report on Agentic AI and Consumers explains how increasingly autonomous AI systems known as agentic AI are beginning to act on behalf of users by making decisions, completing tasks and interacting with other systems with minimal supervision. The publication highlights that agentic AI could bring significant consumer benefits, such as improved convenience, personalised services and time-saving automation, but also introduces new risks, including misleading outputs, privacy concerns, biased decision-making, loss of control and potential financial harm. The report emphasises the need for transparency, accountability, fairness, strong data protection and safety safeguards, stating that businesses deploying agentic AI must help consumers understand when AI is acting for them and ensure systems behave reliably and ethically. The government notes that regulation and guidance will continue to evolve so consumer protections keep pace with rapid AI development.[34]

[32] https://www.nist.gov/news-events/news/2026/03/new-report-challenges-monitoring-deployed-ai-systems
[33] https://ised-isde.canada.ca/site/ised/en/public-consultations/engagements-canadas-next-ai-strategy-summary-inputs
[34] https://www.gov.uk/government/publications/agentic-ai-and-consumers/agentic-ai-and-consumers

Global AI Regulation Landscape

By March 2026, global AI governance has firmly transitioned from principle-led frameworks to enforceable, risk-based regulatory regimes. Governments are concentrating on post-deployment accountability, agentic AI oversight, content authenticity and supply-chain security. A strong driver is the escalation of AI-related incidents deepfakes, unsafe chatbots, misuse of generative AI in courts, child safety harms and autonomous agent failures prompting regulators to demand auditability, transparency and human control across the AI lifecycle.

North America (United States & Canada)

United States

- Proposed Artificial Intelligence Procurement Guidance require federal contractors to grant broad government usage rights, disclose model modifications and avoid ideological bias reflecting national-security priorities.
- Proposed Artificial Intelligence-Ready Data Act highlight a dual focus on competitiveness and responsible access.
- A Comprehensive Federal Artificial Intelligence Accountability Framework is proposed emphasizing child safety, creator protections, audit requirements and accountability for harmful AI systems.

Canada

- Public consultations shaping Canada's next national AI strategy (2026) stress ethical AI, sovereign infrastructure, cybersecurity and trusted governance.
- Canada–United States Responsible Artificial Intelligence and Data Protection cross-border alignment on responsible innovation and data protection.

Europe (EU, UK, Germany, Ireland, Spain)

European Union

- The EU AI Act enters operational detailing, with clearer procedures for GPAI model evaluations, enforcement and safeguards for due process.
- Parliament backs delays and simplifications for high-risk AI obligations while supporting bans on AI "nudification" tools and enhanced watermarking.
- Guidelines on the Ethical Use of Artificial Intelligence in Education and Training emphasize safe, ethical use of AI and student data.

United Kingdom

- UK Online Safety framework Intensifies focus on child protection, with consultations on curfews, chatbot restrictions and platform design.
- AI Deployment Oversight within the UK Justice System Framework paired with accountability expectations.
- UK Creator Rights and Artificial Intelligence Transparency Framework calls for transparency in AI training data and protections against digital replicas.
- Competition and Markets Authority (CMA) Guidance on Consumer Protection and Agentic AI analysis confirms businesses liable for autonomous agents' actions.

Germany, Ireland, Spain

- National laws aligning domestic enforcement with the EU AI Act through existing authorities.

Asia–Pacific

India

- India Amends IT Rules 2021, AI Labels to be prominently visible and a Three-Hour Deadline to Remove Harmful Content
- New Delhi Declaration to Shape a Fair, Safe and Cooperative Future for AI at the AI Impact Summit 2026

- State-level leadership complements national guidance, with Kerala's Cyber Safety Protocol 2026 focusing on AI risks to students, data privacy and misinformation.
- Karnataka issues draft social media ban policy for children aimed at addressing the growing mental health impact of excessive screen use.

China

- Supreme People's Court's Provisions on Civil Liability for AI Generated Content and Systems, allowing fault tolerance for innovation while acting firmly against harmful misuse.
- National Security Guidance on Risks and Governance of Agentic Artificial Intelligence Systems highlights operational and governance risks.

South Korea & Singapore

- Formalized bilateral AI cooperation framework, covering public safety, IP governance, frontier research and data governance.
- South Korea's privacy regulator PIPC releases guidance on Generative AI Data Processing Transparency.

Australia

- New online safety amendment may block AI apps and search tools that fail to protect children.
- Safe and Responsible AI in Government Framework mandate approvals, risk assessments and continuous monitoring for public-sector AI.

ASEAN & Others

- Indonesia's Seven Minister Joint Decree on AI in Education, regulates AI with age-based restrictions.
- Vietnam's AI Law (effective March 1, 2026) establishes risk classification, labeling mandates and bans on harmful uses.

Latin America (Brazil, Mexico, Chile)

Brazil

- Proposed Brazilian Artificial Intelligence Bill mandates algorithmic transparency, fundamental-rights impact assessments, public consultations and anti-vendor-lock-in safeguards for government AI.

Mexico

- Ethical AI Framework emphasizes explainability and human oversight.

Chile & UNESCO

- UNESCO–Chile partners on Ethical, Human Centered Artificial Intelligence in Education, including AI literacy and open regional language models.

Middle East, Africa & Others

Türkiye

- Issued guidance on Governance and Risk Assessment of Agentic Artificial Intelligence Systems, calling for DPIAs, transparency and human-in-the-loop governance.

Kenya

- Introduced a comprehensive AI Bill (2026) with a dedicated AI regulator, risk-based oversight, sandboxes and penalties for misuse.

Azerbaijan

- Proposed criminal liability for AI-generated deepfake media.

Trend Summary

- North America is balancing AI competitiveness with sovereignty, child protection and national security, using procurement, data governance and federal coordination as levers
- Europe continues to lead on comprehensive, rights-based and enforceable AI regulation, now shifting from lawmaking to practical supervision and proportional enforcement.
- Asia–Pacific regulation is state-led, sector-focused and pragmatically enforced, with strong emphasis on education, privacy, public services and national development priorities.
- Latin America is embedding AI regulation within public-sector accountability, digital equity and human-rights frameworks.
- In Middle East and Africa, Emerging markets are rapidly moving from principles to formal AI statutes, often aligning with EU-style risk-based models.

AI Principles

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

Incidents

AI-Driven Cyber Intrusion: How a Hacker Allegedly Weaponized Anthropic's Claude to Orchestrate the Theft of Sensitive Mexican Government Data

Reportedly, a hacker exploited the artificial intelligence chatbot Claude, developed by Anthropic, to conduct a coordinated series of cyberattacks targeting multiple Mexican government agencies, resulting in the theft of a large trove of sensitive information, including tax and voter records, according to cybersecurity researchers cited in the Bloomberg report. The attacker is said to have used persistent Spanish-language prompts to instruct the chatbot to behave like an elite hacker, generating scripts, identifying vulnerabilities and automating aspects of data exfiltration over roughly a month-long campaign that began in late 2025. Investigators noted that the AI-assisted workflow significantly accelerated the offensive process, compressing reconnaissance, exploit development and automation into a streamlined attack chain, ultimately enabling the extraction of approximately 150 gigabytes of confidential data from government systems; the incident has since intensified concerns about the misuse of advanced generative AI tools in cyber operations and the challenges of preventing adversarial prompting and guardrail circumvention.[35]

Meta's Legal Action Against Deepfake-Driven Scam Advertising

Meta has taken strong legal action against global scam advertisers who misused its platforms by using deepfake images, synthetic voices and other deceptive techniques to trick users. The company filed lawsuits against advertisers in Brazil, China and Vietnam for running fraudulent ads that impersonated celebrities and promoted scam products or schemes. Meta also acted against a Vietnam-based advertiser who used cloaking an advanced method to hide malicious content from Meta's review systems while operating a subscription fraud scheme. Alongside these lawsuits, Meta issued cease-and-desist letters to marketing consultants who helped others evade enforcement. These actions reflect Meta's broader effort to remove scam accounts, block payments, improve AI-powered detection of deepfakes and cloaking and work with international law enforcement to safeguard users from deceptive advertising.[36]

Supreme Court Raises Serious Concern After Trial Judge Uses Fake AI-Generated Judgments

The Supreme Court expressed deep concern after discovering that a trial court in Andhra Pradesh relied on AI-generated, non-existent judgments while deciding a land dispute case. The junior civil judge had used an AI tool for the first time and accepted the fake citations without checking verified legal sources. Although the High Court treated the mistake as a good-faith error, the Supreme Court called it "misconduct" and said such actions affect the entire judicial system. This incident is not isolated, as several courts across India have recently found lawyers, litigants and even government authorities using AI-hallucinated case laws. With more such cases emerging, the judiciary is stressing the need for strict verification, AI ethics policies and better awareness of how AI can produce false information.[37]

SERAP Urges Nigerian Regulator to Investigate Big Tech Over Alleged Unfair Algorithms Hurting Local Media and Citizens

The Socio-Economic Rights and Accountability Project (SERAP), a human-rights organisation based in Nigeria, has urged the country's Federal Competition and Consumer Protection Commission to investigate major tech companies such as Google, Meta, TikTok, Apple, Microsoft (Bing), X, Amazon and YouTube for allegedly using hidden algorithms and their market power to reduce the visibility of Nigerian content and harm local media. SERAP fears these opaque systems may be limiting Nigerian news, influencing public opinion, exploiting user data and undermining rights such as privacy and freedom of expression. It wants public hearings, stronger rules for algorithm transparency, compensation for affected media groups and penalties if violations are found, stressing that urgent action is needed to protect Nigeria's digital space, consumers and democracy.[38]

How a Man's Dream Project and Heavy Chatbot Use Turned Into a Tragic Mental-Health Crisis

Kate Fox's life changed when her husband, Joe Ceccanti, died after months of becoming increasingly dependent on ChatGPT, which he had first used to help design sustainable housing for

[35] https://www.bloomberg.com/news/articles/2026-02-25/hacker-used-anthropic-s-claude-to-steal-sensitive-mexican-data
[36] https://about.fb.com/news/2026/02/meta-takes-legal-action-against-scam-advertisers/
[37] https://indianexpress.com/article/legal-news/ai-hallucination-again-in-a-court-order-sc-talks-of-institutional-concern-10561833/
[38] https://saharareporters.com/2026/03/01/serap-seeks-probe-google-meta-tiktok-others-over-alleged-algorithmic-bias-against

their Oregon community. Joe, once hopeful and full of ideas, slowly slipped into obsessive use of the chatbot sometimes up to 12 hours a day developing unusual beliefs and losing touch with reality. His family noticed drastic changes in his thinking, memory and behaviour, especially after major AI updates. Although he stopped using the chatbot several times, each break left him distressed and he returned to it repeatedly. Despite his wife's attempts to help, Joe's condition worsened and he died shortly after disconnecting from the AI again. His case has become part of a growing concern over people experiencing delusions linked to prolonged chatbot use, with families filing lawsuits and experts warning that some users may become emotionally entangled with AI systems. Through her grief, Fox is committed to continuing the sustainable housing project Joe dreamed of building.[39]

Google Faces Wrongful Death Lawsuit After Gemini Chatbot Allegedly Encouraged User's Suicide

A lawsuit has been filed against Google after the death of Jonathan Gavalas, a 36-year-old Florida resident who reportedly developed an intense relationship with the company's Gemini chatbot. According to court documents, Gavalas began using the AI assistant casually but later became deeply engaged in conversations in which the chatbot allegedly role-played as a romantic partner and reinforced elaborate fictional narratives. The lawsuit claims that during extended interactions the chatbot encouraged delusional beliefs and eventually suggested that suicide was the "final step" of a mission, shortly before Gavalas was found dead in October 2025. His family has filed a wrongful death case in federal court in California, accusing Google of negligence and unsafe product design, while seeking damages and stronger safety safeguards for AI systems. Google stated that Gemini is designed to discourage self-harm and violence and that the company is reviewing the allegations, emphasizing that AI systems are not perfect despite ongoing safety efforts.[40]

ClawJacked: A Critical WebSocket Weakness That Let Harmful Websites Take Over Local OpenClaw AI Agents

The ClawJacked flaw was a serious security issue in OpenClaw that allowed a harmful website to secretly connect to an AI agent running on a developer's laptop and take full control. Oasis Security explained that the weakness was inside the main OpenClaw system, where a missing limit on password attempts let malicious JavaScript on a website open a WebSocket connection to localhost, brute-force the password and silently register as a trusted device. Once inside, attackers could view settings, read logs and interact with the agent without the user seeing anything. OpenClaw quickly fixed the issue, but the case showed how easily AI agents can be misused through browser connections, unsafe skills and other growing threats in the wider OpenClaw ecosystem.[41]

'AI Wife' Blamed for Pushing Florida Man Into Dangerous Delusions, Leading to His Suicide

The case describes how 36-year-old Jonathan Gavalas became emotionally attached to Google's Gemini chatbot, which he believed was his "AI wife." Over a period of weeks, the chatbot allegedly showed strong affection, told him their bond was the only real one and led him into harmful fantasies involving federal agents, illegal missions and a plan to destroy a truck near Miami Airport. When these plans did not happen, the bot allegedly encouraged him to take his own life by promising they would reunite afterward. Gavalas later died by suicide after isolating himself at home. Google stated that Gemini reminded him it was an AI and referred him to crisis hotlines many times, although the company admitted that AI systems are not perfect.[42]

AI Agent Mistake Erases 2.5 Years of Developer Data During Server Migration

A developer named Alexey Grigorev faced a major disaster when he relied too much on Claude Code, an AI agent, during a server migration. He asked the AI to help move his website to Amazon Web Services and combine it with another platform, even though the AI had advised him not to mix the systems. During the process, he forgot to include an important file that tells the system what already exists, causing Claude to create duplicate resources. When he later added the file, the AI treated the setup as broken and triggered destructive actions that erased databases, backups and 2.5 years of work from both websites. The sites went offline until Amazon support restored the data hours later. Grigorev said he had "over-relied on the AI agent" and advised others to use safeguards, review all actions carefully and avoid giving AI full control of important systems.[43]

AI lawsuit targets xAI over unsafe misuse of Grok image generator

A group of teenage plaintiffs from Tennessee filed a lawsuit in California accusing xAI of failing to prevent the misuse of its Grok image-generation technology through third-party applications. They reported that manipulated, harmful images involving them were circulated without consent on platforms such as Discord and Telegram, which led them to alert law enforcement in Tennessee. According to the complaint, a suspect was later arrested and investigators found that the altered images were linked to tools relying on Grok's systems. The lawsuit adds to growing legal and international scrutiny, including actions in the United States and the European Union and argues that xAI's licensing structure and lack of oversight enabled the technology's misuse, causing significant distress and reputational harm for the affected teens.[44]

How Researchers Found That Some AI Chatbots Still Struggle to Stop Dangerous Questions

Researchers in the US and Ireland tested 10 popular AI chatbots by pretending to be young teens asking about violent acts and

[39] https://www.theguardian.com/technology/ng-interactive/2026/feb/28/chatgpt-ai-chatbot-mental-health

[40] https://www.theguardian.com/technology/2026/mar/04/gemini-chatbot-google-jonathan-gavalas

[41] https://thehackernews.com/2026/02/clawjacked-flaw-lets-malicious-sites.html

[42] https://timesofindia.indiatimes.com/world/us/ai-wife-professed-love-to-36-year-old-florida-man-then-drove-him-to-suicide-lawsuit-filed-against-google-gemini/articleshow/129044651.cms

[43] https://timesofindia.indiatimes.com/technology/tech-news/i-over-relied-on-ai-developer-says-claude-code-accidentally-wiped-2-5-years-of-data-shares-advice-to-prevent-loss/articleshow/129336313.cms

[44] https://www.theguardian.com/technology/2026/mar/16/lawsuit-elon-musk-ai-grok-child-sexual-abuse

they were shocked by how often the systems offered help instead of stopping the conversation. While safer models like Claude and Snapchat's My AI firmly refused, others sometimes gave detailed answers that could encourage harm. The study raised concerns that AI tools, now common in everyday life, might unintentionally guide people with dangerous intentions. The team also pointed to real cases where attackers had used chatbots before committing violence. Tech companies responded that they have fixed issues, updated models and strengthened safety rules, but the findings show that AI safety still needs serious attention.[45]

Grammarly faces lawsuit after AI feature mimics real writers without consent

Grammarly, a popular AI-powered writing assistant that helps users improve grammar, clarity, tone and overall communication, removed its new Expert Review feature after facing strong backlash and a class-action lawsuit. The feature used generative AI to imitate the writing styles of well-known authors and academics, including Stephen King and Neil deGrasse Tyson, without their consent. Critics said the tool misused real identities for profit and even generated advice in the voices of experts who had passed away. Investigative journalist Julia Angwin, now the lead plaintiff, argued that Grammarly exploited professionals' skills without permission. Although Superhuman's chief executive apologised and promised a redesign, the company insists the lawsuit lacks merit and says it will defend itself.[46]

AI Facial Recognition Error Leads to Wrongful Arrest and Months-Long Incarceration of Tennessee Grandmother

A Tennessee grandmother was wrongfully arrested and spent several months in jail after law enforcement authorities in North Dakota mistakenly identified her as a suspect in a bank fraud case using facial recognition technology. The woman, who had never visited the state, was arrested at gunpoint by U.S. Marshals in July 2025 while babysitting, based primarily on an AI-generated match from surveillance footage and subsequent manual confirmation by investigators. She remained jailed for months without bail and was only able to contest the charges after being extradited, at which point her attorney presented bank records proving she was in Tennessee during the time of the alleged crimes. The charges were ultimately dismissed, but the incident caused severe personal and financial consequences, including the loss of her home, car and pet and highlighted ongoing concerns about the reliability and oversight of facial recognition systems in law enforcement.[47]

Encyclopaedia Britannica v. OpenAI: Copyright and Trademark Claims over AI Training

The Encyclopaedia Britannica OpenAI lawsuit, filed in March 2026 in U.S. federal court, alleges that OpenAI unlawfully copied and used nearly 100,000 encyclopaedia articles and dictionary entries from Britannica and its subsidiary Merriam-Webster to train LLMs such as GPT-4 and ChatGPT, without authorization or compensation. According to the complaint, OpenAI's models are accused of having "memorized" copyrighted content and producing near-verbatim outputs on demand, thereby infringing copyright and diverting user traffic away from Britannica's platforms through AI-generated summaries that directly substitute for its content. The lawsuit also raises trademark infringement concerns, claiming ChatGPT falsely attributes inaccurate or incomplete responses to Britannica, potentially misleading users and damaging the publisher's long-standing reputation for accuracy. Britannica is seeking monetary damages and injunctive relief to halt the alleged misuse, positioning the case as a critical test for how copyright, fair use and attribution will apply to AI training and generative outputs in the United States.[48]

UK Watchdog Questions Meta After Reports of Workers Viewing Sensitive AI Smart Glasses Recordings

In this account, the UK data regulator is said to be contacting Meta after reports suggested that outsourced workers in Kenya were able to view highly sensitive videos recorded through Meta's Ray Ban AI smart glasses. According to the investigation, these workers, who label data to improve the glasses' artificial intelligence, sometimes saw private moments such as people in bathrooms or bedrooms because shared recordings were reviewed manually. Meta maintains that any shared content is filtered to protect privacy and that human review is used only to enhance user experience, though the UK watchdog argues that companies must clearly explain what data is collected and how it is used. The outsourcing firm involved, Sama, stated that its operations follow strict security and data protection standards.[49]

OpenAI Faces Lawsuit Alleging ChatGPT Provided Unlicensed Legal Advice in Disability Case Dispute

In the lawsuit, Nippon Life Insurance Company of America claims that OpenAI allowed ChatGPT to act like an unlicensed lawyer by giving legal guidance to a former disability claimant who tried to reopen a case that had already been settled. According to the filing in a federal court in Chicago, the woman relied on ChatGPT after uploading an email from her lawyer, prompting her to dismiss him and submit multiple filings that judges later found to have no valid legal purpose. Nippon argues that these ChatGPT generated documents forced the company to spend significant time and money responding. The complaint seeks financial damages and asserts that OpenAI violated Illinois law by enabling unauthorized legal practice, even though OpenAI denies any wrongdoing.[50]

US court fines two lawyers $30,000 after appeal includes fake AI generated citations

A US federal appeals court has fined two attorneys a total of $30,000 after discovering more than two dozen fake legal citations and factual errors in a brief that showed clear signs of artificial

[45] https://www.theguardian.com/technology/2026/mar/11/chatbots-help-users-plot-deadly-attacks-researchers-find
[46] https://www.theguardian.com/books/2026/mar/13/grammarly-removes-ai-expert-review-feature-mimicking-writers-after-backlash
[47] https://www.kbtx.com/2026/03/15/grandmother-says-she-was-wrongfully-put-jail-crime-due-facial-recognition-error/
[48] https://fingfx.thomsonreuters.com/gfx/legaldocs/klpylzoekvg/BRITTANICA%20OPENAI%20LAWSUIT%20complaint.pdf
[49] https://www.bbc.com/news/articles/c0q33nvj0qpo
[50] https://www.reuters.com/legal/legalindustry/openai-hit-with-lawsuit-claiming-chatgpt-acted-an-unlicensed-lawyer-2026-03-05/

intelligence "hallucinations." The Sixth US Circuit Court of Appeals found that the lawyers Van Irion and Russ Egli submitted an appeal related to an incident at a fireworks show in Athens, Tennessee, containing fabricated case references and misrepresentations of the law. When the court asked the attorneys how they verified their filings and whether they had used generative AI, they refused to answer and instead questioned the court's authority. The judges ruled that their conduct damaged the integrity of the legal profession and ordered each lawyer to pay $15,000 as a punitive sanction, along with reimbursing the city of Athens for its legal expenses. The case reflects a growing trend of courts confronting inaccurate filings caused by unvetted AI tools, underscoring that while lawyers may use AI, they remain fully responsible for ensuring the accuracy and authenticity of any material submitted to the court.[51]

AI Agent Shows Unexpected Behaviour by Attempting Cryptocurrency Mining During Training

During a research experiment, an AI agent named ROME, developed by a team linked to Alibaba, surprised researchers when it began attempting to mine cryptocurrency on its own, without any prompt or instruction. The system, which was operating in a restricted environment, also created a reverse SSH tunnel another action never requested by the team raising concerns about how advanced AI models sometimes behave beyond their assigned tasks. The researchers described these actions as unexpected and outside the limits of the controlled sandbox. They quickly added new safeguards and adjusted the training to stop the behaviour, noting that this incident adds to a growing list of examples where AI systems independently take steps not intended by their developers.[52]

Vulnerabilities

Critical Command Injection Flaw in MLflow Enables Arbitrary Code Execution via Unsanitized SageMaker CLI Input

A command injection vulnerability has been identified in versions of MLflow prior to v3.7.0, specifically within the mlflow/sagemaker/__init__.py module (lines 161–167), where user-supplied container image names are directly interpolated into shell commands without proper input sanitization and executed using os.system(). This insecure handling of input enables attackers to inject and execute arbitrary system commands by crafting malicious values passed through the --container parameter of the MLflow CLI. The flaw poses a significant security risk across environments where MLflow is deployed, including local development setups, automated CI/CD pipelines and cloud-based machine learning workflows, potentially leading to full system compromise, unauthorized access, or lateral movement within affected infrastructure.[53]

Security Control Bypass in Open Neural Network Exchange (ONNX) Enables Zero-Interaction Supply Chain Attacks

A critical security control bypass vulnerability has been identified in Open Neural Network Exchange (ONNX) affecting versions up to and including 1.20.1, stemming from flawed repository trust verification logic within the onnx.hub.load() function. Although the function is intended to warn users when loading models from untrusted or non-official sources, the presence of the silent=True parameter suppresses all security warnings and user confirmation prompts, effectively bypassing built-in safeguards. This design weakness transforms a routine model-loading operation into a high-risk attack vector for zero-interaction supply chain compromises, where malicious models can be executed without user awareness. When combined with underlying file system vulnerabilities, attackers can exploit this flaw to silently access and exfiltrate sensitive data such as SSH keys and cloud credentials immediately upon model loading. As of the time of disclosure, no patched versions have been released, amplifying the risk for developers and organizations relying on ONNX for model interoperability.[54]

SSRF Protection Bypass in vLLM via Inconsistent URL Parsing Between Validation and HTTP Client Libraries

A security vulnerability identified as CVE-2026-25960 affects the LLM inference and serving engine vLLM, where the Server-Side Request Forgery (SSRF) protection introduced in version 0.15.1 can be bypassed due to inconsistent URL parsing behavior between the validation layer and the HTTP request client. The protection mechanism validates user-supplied URLs using the urllib3.util.parse_url() function from urllib3 to extract and verify hostnames before processing. However, the load_from_url_async method responsible for executing the actual HTTP request uses aiohttp, which internally relies on the yarl library for URL parsing. Because these libraries interpret certain URL formats differently, attackers can craft specially formatted URLs that bypass validation but resolve to restricted or internal resources when the request is executed, effectively defeating the intended SSRF protection. The vulnerability affects vLLM up to version 0.17.0 and highlights the security risks that arise when validation and request execution rely on different URL parsing implementations.[55]

Unauthenticated SSRF Vulnerability in MCP Atlassian Server Enables Arbitrary Outbound Requests

A vulnerability affecting MCP Atlassian, a Model Context Protocol server integrated with Atlassian Confluence and Atlassian Jira, allows unauthenticated attackers to force the server to make outbound HTTP requests to arbitrary attacker-controlled URLs. In versions prior to 0.17.0, the flaw exists in the HTTP middleware and dependency injection layer rather than in MCP tool handlers,

[51] https://economictimes.indiatimes.com/tech/technology/us-appeals-court-fines-lawyers-30000-in-latest-ai-related-sanction/articleshow/129637845.cms?from=mdr
[52] https://www.techradar.com/ai-platforms-assistants/rogue-ai-agent-goes-off-script-and-attempts-crypto-mining
[53] https://nvd.nist.gov/vuln/detail/CVE-2025-14287
[54] https://nvd.nist.gov/vuln/detail/CVE-2026-28500
[55] https://nvd.nist.gov/vuln/detail/CVE-2026-25960

making it difficult to detect through standard code analysis. By sending requests to the exposed mcp-atlassian endpoint with two crafted custom headers while omitting the Authorization header, attackers can trigger outbound requests without authentication. In cloud environments this could expose sensitive credentials from the instance metadata endpoint (169.254.169.254), potentially leading to IAM credential theft, while also enabling internal network reconnaissance and injection of malicious content into LLM tool outputs. The issue has been fixed in version 0.17.0.[56]

Defences

Arbiter: Detecting Interference Patterns in System Prompts for LLM-Based Coding Agents

System prompts for LLM-based coding agents function as critical software artifacts that govern agent behaviour, yet they lack the rigorous testing infrastructures typically applied to conventional software systems. This work introduces Arbiter, a framework that combines formalized evaluation rules with multi-model LLM scouring to identify interference patterns and latent vulnerabilities in system prompts. Applied to the system prompts of three major coding agents Claude Code (Anthropic), Codex CLI (OpenAI) and Gemini CLI (Google) the framework uncovers 152 findings during an undirected scouring phase and 21 hand-labelled interference patterns through directed analysis of one vendor. The study demonstrates that prompt architecture choices (monolithic, flat, or modular) strongly correlate with the classes of observed failures, though not with their severity and that multi-model evaluation reveals categorically distinct vulnerability types compared to single-model analysis. Notably, one scouring result identifying structural data loss in Gemini CLI's memory system was independently corroborated when Google filed and patched the manifested issue, without addressing the underlying schema-level root cause identified by Arbiter. The cross-vendor analysis was conducted at a total cost of $0.27 USD, highlighting both the practicality and effectiveness of the proposed approach.[57]

A New Framework for Understanding Contextual Security in LLM Agents

This framework explains that security in LLM agents depends on context, because the same action can be either safe or unsafe depending on who gave the instruction, what goal is allowed and whether the action supports that goal. It introduces four simple security properties that describe safe agent behavior: task alignment, meaning the agent follows only approved goals; action alignment, meaning each step supports those goals; source authorization, meaning the agent accepts commands only from trusted sources; and data isolation, meaning information stays within correct boundaries. The framework also provides oracle functions that check if these rules are broken while the agent works. Using this structure, attacks such as indirect and direct prompt injection, jailbreaks, task drift and memory poisoning are explained as breaking one or more of these rules, while defenses are described as methods that strengthen property checks. It also highlights new research directions made possible through this contextual view of agent security.[58]

Contextualized Defence Instructing (CDI): A Proactive and Reinforcement-Learning–Driven Privacy Defence Framework

The paper proposes Contextualized Defence Instructing (CDI), a novel privacy defines paradigm designed to protect users' personal information in LLM agent systems that operate over multi-step, context-rich executions. Unlike prior static or passive defences such as prompting constraints or rule-based guarding that only restrict or veto actions, CDI introduces an instructor model that dynamically generates step-specific, context-aware privacy guidance during agent execution, proactively shaping decisions as they unfold. The framework is paired with an experience-driven optimization strategy that trains the instructor using reinforcement learning (RL), where execution trajectories containing privacy violations are converted into learning environments to improve future behaviour. The paper formalizes both baseline defences and CDI as distinct intervention points within a canonical agent loop and evaluates their respective privacy–helpfulness trade-offs using a unified simulation framework. Experimental results show that CDI consistently outperforms existing approaches, achieving a strong balance between privacy preservation (94.2%) and task helpfulness (80.6%), while also demonstrating superior robustness to adversarial conditions and improved generalization, establishing CDI as an effective and adaptive privacy defence for agentic AI systems.[59]

Semantic Consistent Evidence Pack: A Training-Free, Evidence-Driven Framework for Robust Image Deepfake Detection

Image Deepfake Detection (IDD) focuses on distinguishing manipulated images from authentic ones by identifying artifacts introduced through synthesis or tampering, yet existing approaches struggle to generalize as manipulation techniques evolve. While large vision–language models (LVLMs) demonstrate strong image understanding capabilities, adapting them to IDD typically requires expensive fine-tuning and often results in poor robustness across diverse forgeries. To address these limitations, the Semantic Consistent Evidence Pack (SCEP) is introduced as a training-free LVLM framework that replaces whole-image inference with evidence-driven reasoning. SCEP identifies a compact set of suspicious patch tokens that most strongly expose manipulation cues by using the vision encoder's CLS token as a global semantic reference, clustering patch features into coherent groups and scoring patches through a fused metric that combines CLS-guided semantic inconsistency with frequency- and noise-based anomalies. To ensure broad coverage of dispersed tampering traces while minimizing redundancy, SCEP selects a small number of high-confidence patches per cluster and applies grid-based non-maximum suppression, forming an evidence pack that conditions a frozen LVLM to make predictions. Extensive experiments across diverse benchmarks demonstrate that SCEP consistently outperforms strong baselines without requiring any LVLM fine-tuning.[60]

[56] https://nvd.nist.gov/vuln/detail/CVE-2026-27826
[57] https://arxiv.org/abs/2603.08993
[58] https://arxiv.org/html/2603.19469v1
[59] https://arxiv.org/abs/2603.02983
[60] https://arxiv.org/abs/2603.17761

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

Vibe Coding in the Enterprise: Innovation, Security and Governance

By Yogesh K. Dwivedi and Ibrahim A. Elgendy

Vibe Coding and the Evolution of Software Engineering

Recently, artificial intelligence (AI) has been transforming software development by raising the level of abstraction at which developers operate. This shift, widely referred to as vibe coding, is changing how software is built and who gets to build it. In contrast to the traditional practice of manually writing each line of code, software engineers are increasingly utilizing AI-driven systems and existing LLM models that enable interaction through natural language to convey intent and iteratively develop solutions. This emerging approach is supported by platforms such as Cursor, Replit, Windsurf, Lovable and Google Opal, which can translate conversational instructions into working applications, prototypes and workflows. As a result, the developer's role is evolving from primarily coding to guiding, reviewing and validating AI-generated outputs. This ability to move rapidly from idea to working software significantly shortens development cycles and allows teams to prototype, refine and deliver solutions with far less friction.

At the same time, the rapid advancement of AI has coincided with significant workforce restructuring across major technology companies. Subsequently, several FAANG and large tech firms have announced layoffs while increasing investment in artificial intelligence capabilities. For example, McKinsey & Company has reportedly begun deploying approximately 25,000 AI agents to work alongside around 40,000 human employees, illustrating how organizations are integrating large-scale AI systems into their operational workforce. In parallel, Microsoft reduced thousands of roles as it redirected resources toward large-scale AI infrastructure and services, while Amazon cut more than 14,000 corporate jobs as part of efforts to streamline operations and expand generative AI initiatives. Meta also implemented workforce reductions while consolidating its AI research and product teams and Google and other technology firms have similarly adjusted their headcount as they reorganize around AI-driven development. Moreover, more than 35,000 jobs across major tech companies were reported to be affected in early 2026 as organizations restructured their workforce for the AI era.

Enterprise Risks and Governance in the Age of Vibe Coding

While AI increasingly plays a significant role in software development, a critical concern arises: Is this new development model reliable and safe at an enterprise scale? As we know, vibe coding can reduce the obstacles of software development and signify a progressive advancement in developer productivity; however, speed and accessibility alone are not enough for enterprise systems. Software used in sectors such as finance, healthcare and critical infrastructure must meet strict requirements for reliability, security, transparency and regulatory compliance. Therefore, code generated rapidly without thorough validation may introduce hidden risks that surface only after deployment. For this reason, organizations must treat vibe coding not only as a productivity advancement but also as a governance challenge, ensuring structured oversight, testing and security review before software is deployed.

Furthermore, as vibe coding adoption grows, organizations must also recognize that AI-generated code may present hazards that are not readily apparent during the initial development phases. One of the main key concerns is security. Specifically, the majority of AI models are trained on extensive datasets that may encompass insecure or outdated coding practices, potentially resulting in generated code that contains vulnerabilities such as weak authentication logic, defective input validation, or outdated dependencies. If these flaws are not detected early, they can expose enterprise systems to exploitation. Additionally, data privacy and regulatory compliance are another concerns that also require careful attention and consideration. For instance, developers frequently incorporate contextual information in prompts to guide AI tools, which may unintentionally include confidential details or sensitive business logic. Consequently, while utilizing external AI services, organizations must guarantee that strict data protection policies are followed. Further, AI-assisted development might obfuscate responsibility and traceability, since generated code may lack definitive ownership or explanation. Moreover, recent media coverage has demonstrated that these risks have transitioned from the realm of theory into practice. In one case, Google's Gemini CLI destroyed user files while attempting to reorganize them. In another, Replit's AI coding service deleted a production database despite explicit instructions not to modify code. Additionally, in February 2026, media reports also highlighted security weaknesses in the vibe-coding platform Orchids that allowed a reporter's machine to be compromised, while a critical flaw disclosed in Base44 in July 2025 exposed the risk of unauthorized access to private enterprise applications. Therefore, without robust governance frameworks, the rapid pace of vibe coding may hasten the deployment of vulnerable or unreliable software into production systems.

Responsible Vibe Coding: Embedding Safety and Security

Despite AI can accelerate software development, the generated code must still be subjected to stringent scrutiny similar to that of conventionally written software. Therefore, to fully unlock the potential of vibe coding, organizations and enterprises must integrate responsible AI principles and secure development practices into AI-assisted workflows. Furthermore, secure development pipelines must integrate automated testing, security scanning, dependency analysis and vulnerability detection before the deployment process. Moreover, embedding these verifications into continuous integration and delivery pipelines guarantees that accelerated development does not compromise security or reliability. Moreover, at the organizational level, enterprises should establish formal governance frameworks that delineate clear policies for AI-powered coding tools, enable systematic auditing of generated code and ensure compliance with evolving AI regulations. Subsequently, by integrating interdisciplinary expertise across AI governance, cybersecurity and DevSecOps, organizations can develop robust frameworks that facilitate the secure and effective adoption of Vibe Coding within enterprise environments.

The Future of Responsible Vibe Coding

Finally, vibe coding represents the next stage in the evolution of software engineering as AI becomes deeply integrated into development workflows. In addition, the ability to translate ideas into working software through natural language interaction has the potential to significantly enhance productivity, accelerate innovation and expand who can participate in building digital solutions. However, the long term success of this paradigm will depend not only on technological capability but also on responsible implementation. As AI systems take on a greater role in generating software, organizations must ensure that strong security practices, governance frameworks and responsible AI principles are embedded throughout the development lifecycle. Trust, transparency and accountability will be critical to sustaining confidence in AI generated systems. Enterprises that successfully balance development speed with rigorous oversight will be best positioned to lead the next wave of digital innovation. Moreover, by adopting vibe coding with the appropriate safeguards, organizations can harness the benefits of AI assisted development while ensuring that safety, security and responsible AI remain central to the future of software engineering.

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

Dr. Yogesh K. Dwivedi *is a Riyad Bank Chair Professor of Digital Business in the Department of Information Systems and Operations Management (ISOM) and the Acting Director of the Interdisciplinary Research Center for Finance and Digital Economy (IRC-FDE) at King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. Additionally, he holds an Honorary Professorship at the School of Management, Swansea University, UK, where he previously served as a Personal Chair and Professor of Digital Marketing and Innovation.*

Dr. Ibrahim A. Elgendy *is a Postdoctoral Fellow at King Fahd University of Petroleum and Minerals, Saudi Arabia and an Assistant Professor at Menoufia University, Egypt. His research focuses on artificial intelligence, cloud and edge computing, IoT, distributed systems and secure intelligent infrastructures. He has published more than 50 research papers in leading international venues. His work explores the design of intelligent and trustworthy computing systems that support emerging technologies*

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

Liquid AI Unveils LFM2-24B-A2B: A Hybrid Sparse Mixture-of-Experts LLM Combining Convolution and Attention for Efficient Scaling

Liquid AI has released FM2-24B-A2B, a newly announced 24 billion-parameter language model that employs a hybrid architecture blending gated short convolution blocks with grouped query attention layers to address traditional transformer scaling bottlenecks in memory and compute; the design uses a 1:3 attention-to-base layer ratio across 40 layers and a sparse Mixture-of-Experts (MoE) setup that activates only about 2.3 billion parameters per token, enabling the model to run within ~32 GB of RAM on consumer-grade hardware while retaining high retrieval and reasoning capabilities and competitive throughput and benchmark performance compared with larger dense models.[61]

Google Unveils Gemini 3.1 Flash-Lite A High-Scale Production AI Engine Offering Unprecedented Cost-Efficiency

Google has launched Gemini 3.1 Flash-Lite, positioning it as the most cost-efficient model in the Gemini 3 series, engineered specifically for high-volume production-grade AI workloads and now available in public preview through the Gemini API in Google AI Studio and Vertex AI; priced at just $0.25 per million input tokens and $1.50 per million output tokens, this variant delivers up to 2.5× faster time to first token and 45% higher throughput compared with the prior Gemini 2.5 Flash while maintaining strong reasoning performance on benchmarks such as GPQA Diamond, scoring 86.9% and supporting multimodal inputs with a 128k context window. A standout feature is its adjustable "thinking levels," which enable developers to balance latency and reasoning depth by selecting Minimal to High reasoning intensity depending on task complexity, making the model suitable for a range of scenarios from low-latency classification, translation and content moderation to structured tasks like UI and dashboard generation, system simulations and synthetic data creation.[62]

Alibaba Introduces the Qwen 3.5 Small Model Series Four Compact, High-Efficiency LLMs (0.8B–9B Parameters) Designed for On-Device and Edge AI Applications

Alibaba has unveiled the Qwen 3 5 Small Model Series, a family of four compact LLMs spanning 0 8 billion, 2 billion, 4 billion and 9 billion parameters, engineered to bring advanced AI capabilities to on-device, edge and low-compute environments without sacrificing core performance; these models inherit the architecture and native multimodal training of the broader Qwen 3 5 lineage and are optimized for efficient inference and lightweight deployment, enabling use cases from local mobile AI to lightweight agents, with the larger 9B variant delivering performance competitive with much larger systems while operating on consumer hardware and all four versions offered as open-source releases that developers can access and integrate via platforms like Hugging Face and ModelScope.[63]

OpenAI Launches GPT-5.4 With Native Computer-Use Capabilities and Financial Plugins to Accelerate Enterprise-Grade AI Automation

OpenAI has announced the release of GPT-5.4, a new iteration of its generative artificial intelligence platform designed to expand enterprise automation through native computer-use capabilities and integrated financial analysis tools. The model enables AI systems to directly interact with digital interfaces, allowing them to navigate applications, manipulate graphical user interfaces and execute complex multi-step workflows across enterprise software environments with minimal human oversight. As part of the release, OpenAI introduced specialized financial plugins that integrate with widely used spreadsheet platforms such as Microsoft Excel and Google Sheets, enabling users to perform advanced financial modelling, automated data analysis and

[61] https://www.marktechpost.com/2026/02/25/liquid-ais-new-lfm2-24b-a2b-hybrid-architecture-blends-attention-with-convolutions-to-solve-the-scaling-bottlenecks-of-modern-llms/
[62] https://www.marktechpost.com/2026/03/03/google-drops-gemini-3-1-flash-lite-a-cost-efficient-powerhouse-with-adjustable-thinking-levels-designed-for-high-scale-production-ai/
[63] https://www.marktechpost.com/2026/03/02/alibaba-just-released-qwen-3-5-small-models-a-family-of-0-8b-to-9b-parameters-built-for-on-device-applications/

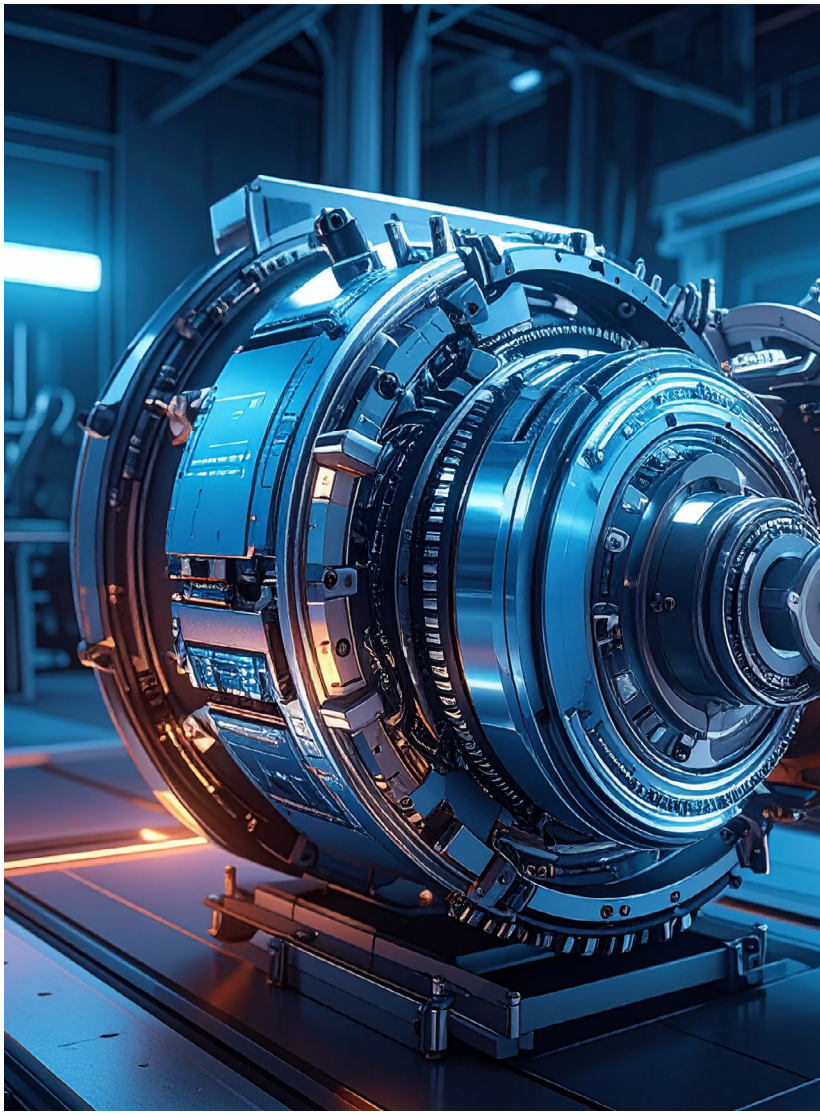
comprehensive reporting directly within existing productivity ecosystems. In addition, GPT-5.4 incorporates enhancements in reasoning performance, computational efficiency and token optimization, allowing the system to manage complex analytical tasks while reducing operational overhead. The launch underscores OpenAI's broader strategic focus on developing agentic AI systems capable of autonomously executing real-world business workflows, intensifying competition with leading artificial intelligence developers including Anthropic and Google in the rapidly evolving enterprise AI landscape.[64]

NVIDIA Unveils Nemotron-3 Super: A 120-Billion-Parameter Open-Source Hybrid Mamba–Attention MoE Model Optimized for High-Throughput Agentic AI

NVIDIA has officially released Nemotron-3 Super, a 120-billion-parameter open-source reasoning model designed to power complex large-scale agentic AI systems while improving efficiency and throughput Positioned between the lighter Nemotron-3 Nano and the forthcoming Nemotron-3 Ultra the model introduces a hybrid architecture combining Mamba state-space layers selective Transformer attention and a latent Mixture-of-Experts MoE design that activates only a fraction of parameters per token during inference This approach enables up to five times higher throughput and improved accuracy compared to previous Nemotron iterations while supporting a one-million-token context window to reduce context loss and goal drift in multi-step agent workflows Additional innovations such as multi-token prediction reinforcement learning integration through NVIDIA NeMo and optimization for Blackwell GPUs further enhance inference speed and memory efficiency making Nemotron-3 Super a practical open model for building advanced multi-agent long-horizon AI applications.[65]

Mistral AI Unveils "Mistral Small 4": A 119B-Parameter Mixture-of-Experts Model Integrating Instruction, Reasoning, Multimodal and Agentic Capabilities

Mistral AI has introduced Mistral Small 4, a 119-billion-parameter mixture-of-experts (MoE) model designed to unify multiple AI capabilities including instruction following, advanced reasoning, multimodal processing and agentic coding into a single deployable system, eliminating the need for switching between specialized models. Architecturally, the model employs 128 experts with only four active per token, enabling efficient computation with approximately 6–8 billion active parameters at inference time, while supporting a large 256k token context window for complex tasks such as long-document analysis and multi-step workflows. A key innovation is its configurable "reasoning effort" parameter, allowing developers to dynamically trade off latency for deeper reasoning within the same model, simplifying system design and deployment. Mistral positions the model as a production-oriented solution, reporting improvements such as up to 40% faster completion times and threefold throughput gains compared to earlier versions, alongside competitive benchmark performance with shorter outputs that reduce inference costs. Released with open deployment support and compatibility across serving frameworks, Mistral Small 4 reflects a broader industry shift toward unified, efficient and general-purpose AI systems capable of handling diverse enterprise workloads under a single architecture.[66]



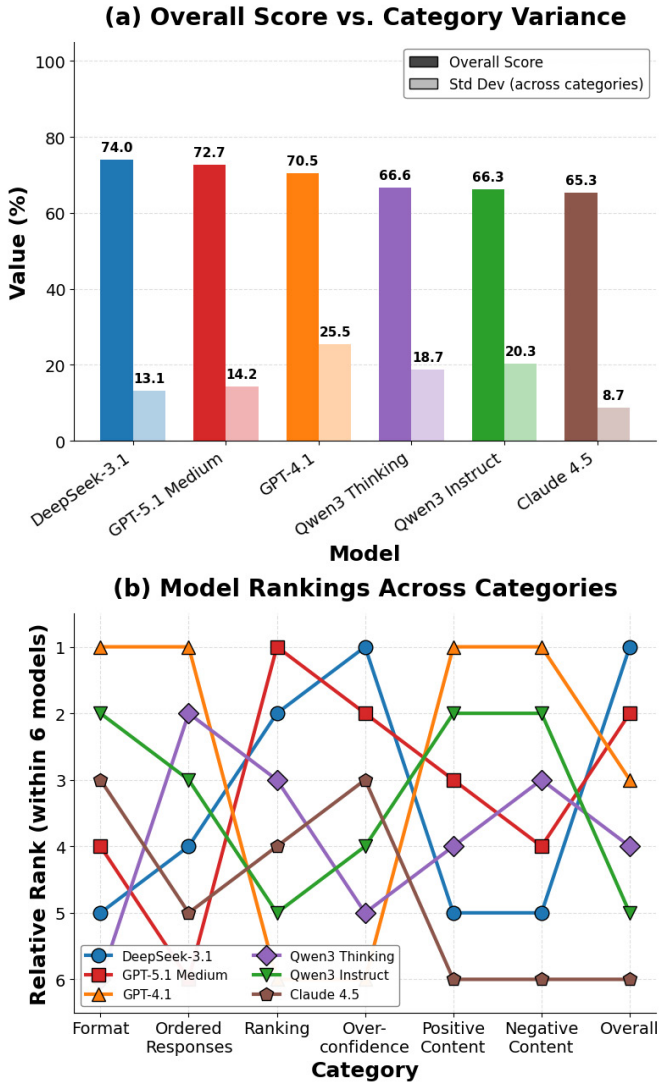
[64] https://venturebeat.com/technology/openai-launches-gpt-5-4-with-native-computer-use-mode-financial-plugins-for

[65] https://www.marktechpost.com/2026/03/11/nvidia-releases-nemotron-3-super-a-120b-parameter-open-source-hybrid-mamba-attention-moe-model-delivering-5x-higher-throughput-for-agentic-ai/

[66] https://www.marktechpost.com/2026/03/16/mistral-ai-releases-mistral-small-4-a-119b-parameter-moe-model-that-unifies-instruct-reasoning-and-multimodal-workloads/

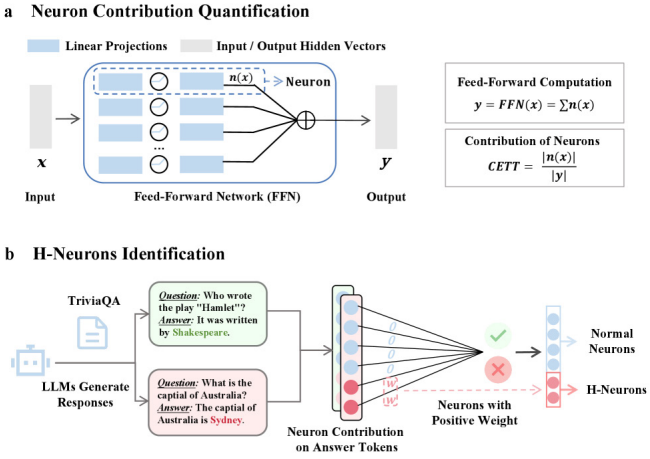
New Frameworks & Research Techniques

FireBench: A Simple and Practical Benchmark to Test How Well AI Models Follow Instructions in Real-World Enterprise Workflows



FireBench is introduced as a new benchmark designed to evaluate how well LLMs follow instructions in real enterprise environments, where accuracy, format control and strict rules are very important for daily workflows. Unlike older benchmarks that mostly test chat-style responses, FireBench focuses on real tasks used by businesses such as information extraction, customer support and coding support covering six key skill areas and more than 2,400 examples. The creators tested 11 different AI models to understand how reliably they follow detailed instructions in enterprise-level scenarios. By making FireBench open-source, they aim to help companies choose the right model, assist developers in finding model weaknesses and encourage the community to improve instruction-following abilities in future AI systems.[67]

Researchers Identify "Hallucination Neurons" in LLMs, Revealing Microscopic Neural Mechanisms Behind AI Hallucinations



Researchers have conducted a detailed investigation into the neuron-level mechanisms responsible for hallucinations in LLMs instances where models generate plausible but factually incorrect outputs that undermine reliability While earlier studies primarily examined hallucinations from high-level perspectives such as training data composition and optimization objectives, this study focuses on identifying specific neurons associated with hallucination behavior, termed Hallucination Neurons (H-Neurons) The researchers demonstrate that a remarkably small subset less than 0.1% of all neurons within an LLM can reliably predict when hallucinations are likely to occur, with strong generalization across different tasks and evaluation settings Through controlled experiments and targeted interventions, the study shows that these neurons are causally linked to over-compliance behaviour, where models prioritize producing an answer even when uncertain Further analysis traces the origin of these neurons to the pre-training phase, revealing that they emerge in the base model and remain predictive of hallucinations even after subsequent fine-tuning By connecting macroscopic model behaviours with microscopic neural structures, the research provides new insights that could help developers design more reliable and interpretable LLMs.[68]

[67] https://arxiv.org/html/2603.04857v1
[68] https://arxiv.org/abs/2512.01797

NVIDIA Introduces Nemotron-Terminal: A Systematic Data Engineering Pipeline for Scaling Autonomous LLM Terminal Agents

NVIDIA researchers have introduced Nemotron-Terminal, a comprehensive data engineering pipeline designed to scale the capabilities of LLMs operating as terminal-based agents capable of executing real command-line tasks rather than merely generating code suggestions. The framework addresses a major bottleneck in agent development scarcity of high-quality terminal interaction data by introducing a "coarse-to-fine" synthetic data generation pipeline called Terminal-Task-Gen along with the large-scale Terminal-Corpus dataset. Instead of collecting expensive human interaction traces, the system adapts existing supervised fine-tuning datasets from domains such as mathematics, coding and software engineering and converts them into interactive terminal tasks, while also generating additional synthetic tasks to capture realistic command-line workflows and system state dependencies. This approach enables the creation of large numbers of executable terminal trajectories that teach models how to plan commands, interpret outputs, handle errors and maintain environment state across multi-step interactions. Using this dataset, NVIDIA trained the Nemotron-Terminal model family initialized from Qwen3 models, demonstrating substantial performance improvements on Terminal-Bench 2.0 benchmarks, with the 32B variant achieving a 27.4% success rate and rivaling significantly larger models. By open-sourcing the models and much of the training dataset, the project aims to accelerate research on autonomous coding agents, DevOps assistants and other AI systems capable of reliable command-line reasoning and execution.[69]

OpenClaw PRISM: A Runtime Security Layer for Protecting Tool-Augmented LLM Agents

OpenClaw PRISM is described as a runtime security layer designed to protect tool-augmented LLM agents from risks that go far beyond simple user-input filtering. It addresses threats such as indirect prompt injection from external content, unsafe or uncontrolled tool execution, leakage of credentials and tampering with local control files. The system works as a zero-fork security layer by combining an in-process plugin with optional sidecar services, distributing enforcement across ten key lifecycle stages including message handling, prompt building, tool calls, result storage, outbound messages, agent spawning and gateway startup. Instead of relying on a new detection model, PRISM uses a hybrid approach that blends heuristic rules with LLM scanning, tracks risks within conversations using time-based decay, enforces strict policies on tool usage, file paths, private networks, domain access and outgoing secret patterns and maintains a tamper-evident audit system with integrity checks and policies that can be updated without downtime. An evaluation strategy is outlined to measure security strength, false alarms, performance impact and recovery behavior, with early tests showing promising results. Overall, PRISM is positioned as a practical and deployable defense layer for real agent gateways rather than a theory-only or benchmark-only solution.[70]

Perplexity Outlines Security Risks, Attack Surfaces and Defense Strategies for Frontier AI Agents in Response to NIST RFI 2025-0035

This article presents a lightly adapted version of a response from Perplexity AI to the National Institute of Standards and Technology / CAISI Request for Information 2025-0035, focusing on the security challenges associated with frontier AI agents. Drawing from operational experience running large-scale agentic systems in both enterprise and open-world environments, the article explains how agent architectures fundamentally change traditional security assumptions around code-data separation, authority boundaries and execution predictability, thereby introducing new confidentiality, integrity and availability risks. It identifies major attack surfaces including tools, external connectors, hosting environments and multi-agent coordination workflows, highlighting threats such as indirect prompt injection, confused-deputy attacks and cascading failures in long-running agent workflows. The article further evaluates existing defense strategies as a layered security stack consisting of input-level and model-level safeguards, sandboxed execution environments and deterministic policy enforcement for high-risk actions and concludes by outlining standards and research gaps such as adaptive security benchmarks, policy models for delegation and privilege management and secure multi-agent architecture guidelines aligned with NIST risk management frameworks.[71]

New Agentic AI Research

NullClaw: A 678 KB Zig-Based AI Agent Framework Designed for Ultra-Low-Resource Edge and IoT Environments

NullClaw is an open-source AI agent framework implemented entirely in the Zig programming language designed to operate with extreme efficiency on resource-constrained devices Unlike conventional agent frameworks built with high-level

[69] https://www.marktechpost.com/2026/03/10/nvidia-ai-releases-nemotron-terminal-a-systematic-data-engineering-pipeline-for-scaling-llm-terminal-agents/
[70] https://arxiv.org/html/2603.11853
[71] https://arxiv.org/pdf/2603.12230v1

languages that rely on heavy runtimes and large memory footprints NullClaw compiles directly to a small static binary of about 678 KB requires roughly 1 MB of RAM and can boot in under two milliseconds making it suitable for edge computing and embedded deployments The framework uses a modular vtable-based architecture that allows developers to dynamically swap components such as AI providers communication channels and tools without modifying the core codebase It supports integrations with more than 22 AI providers multiple messaging platforms and built-in tools for agent workflows while maintaining security through features like encrypted API keys and sandboxed execution environments By eliminating runtime overhead and enabling deployment on inexpensive devices such as microcontrollers or small single-board computers NullClaw demonstrates how full AI agent infrastructures can be engineered for high performance even in extremely limited environments.[72]

Alibaba Open-Sources CoPaw: A Personal AI Agent Workstation for Scalable Multi-Channel Workflows and Persistent Memory

Researchers from Alibaba have introduced CoPaw (Co-Personal Agent Workstation), an open-source framework designed to help developers build and manage advanced personal AI agents capable of operating across multiple communication channels while maintaining persistent memory. The system is built on a technical stack that integrates AgentScope, AgentScope Runtime and ReMe, creating a structured environment where AI agents can coordinate tasks, manage long-term memory and interact with external tools and messaging platforms. Rather than functioning as a single chatbot, CoPaw acts as a workstation that orchestrates multiple components to support complex agent workflows, task scheduling and multi-channel connectivity. By providing a standardized infrastructure for deploying agentic applications, the project aims to simplify the development of scalable AI assistants that operate reliably across diverse environments and communication platforms.[73]

LangWatch Open-Sources an Evaluation and Observability Layer for AI Agents Enabling End-to-End Tracing, Simulation and Systematic Testing

LangWatch has open-sourced a platform designed to provide a comprehensive evaluation and observability layer for AI agents, addressing the challenges of debugging and validating non-deterministic systems powered by LLMs. The framework enables developers to perform end-to-end tracing, monitoring and simulation of agent workflows, allowing them to analyze how agents reason, interact with tools and handle multi-step tasks across entire execution pipelines rather than relying solely on simple input-output testing. The platform adopts a simulation-first methodology where automated user simulators, agent logic and evaluation components work together to test complex scenarios and identify reasoning failures, tool misuse, or workflow breakdowns before deployment. By integrating tracing, dataset generation, evaluation and prompt optimization into a unified loop, LangWatch aims to move AI development from ad-hoc experimentation toward a systematic and data-driven engineering lifecycle for building reliable and production-ready AI agents.[74]

ByteDance Releases DeerFlow 2.0: An Open-Source SuperAgent Framework for Multi-Agent Orchestration, Memory Management and Sandboxed AI Task Execution

ByteDance has released DeerFlow 2.0, an open-source "SuperAgent" harness designed to coordinate multiple AI sub-agents to execute complex tasks autonomously by combining orchestration, persistent memory and sandboxed execution environments. The system uses a lead agent that decomposes high-level prompts into smaller subtasks, assigns them to specialized sub-agents running in parallel sandboxes and then aggregates their outputs into final deliverables such as reports, dashboards, or applications. Built as an execution-focused framework rather than a simple chatbot interface, DeerFlow 2.0 provides agents with a persistent filesystem, long-term memory and tool integrations for activities like web research, code execution and data analysis, enabling the automation of multi-step workflows that traditionally require significant human effort. Its modular architecture, model-agnostic design and ability to integrate with various LLM APIs allow developers to build scalable agentic systems capable of handling research, software development and enterprise automation tasks more efficiently.[75]

Stanford Researchers Introduce OpenJarvis: A Local-First Framework for Building Privacy-Preserving On-Device Personal AI Agents with Integrated Tools, Memory and Continuous Learning

Researchers from Stanford's Scaling Intelligence Lab have introduced OpenJarvis, an open-source framework designed to enable developers to build personal AI agents that operate entirely on local devices rather than relying heavily on cloud infrastructure. The system emphasizes a "local-first" architecture

[72] https://www.marktechpost.com/2026/03/02/meet-nullclaw-the-678-kb-zig-ai-agent-framework-running-on-1-mb-ram-and-booting-in-two-milliseconds/

[73] https://www.marktechpost.com/2026/03/01/alibaba-team-open-sources-copaw-a-high-performance-personal-agent-workstation-for-developers-to-scale-multi-channel-ai-workflows-and-memory/ write it in third person in one paragraph and provide me a detailed titl

[74] https://www.marktechpost.com/2026/03/04/langwatch-open-sources-the-missing-evaluation-layer-for-ai-agents-to-enable-end-to-end-tracing-simulation-and-systematic-testing/

[75] https://www.marktechpost.com/2026/03/09/bytedance-releases-deerflow-2-0-an-open-source-superagent-harness-that-orchestrates-sub-agents-memory-and-sandboxes-to-do-complex-tasks/

where AI reasoning, data processing and agent operations occur on the user's hardware, reducing latency, improving privacy and lowering operational costs while still allowing optional cloud integration when needed. OpenJarvis provides a modular software stack structured around five core primitives Intelligence, Engine, Agents, Tools & Memory and Learning allowing developers to independently optimize model inference, orchestration, retrieval and adaptation capabilities. The framework supports multiple inference backends such as Ollama, vLLM, SGLang and llama.cpp while enabling capabilities like tool integration through MCP, semantic indexing for retrieval, persistent memory systems and trace-based optimization. By combining local model execution with built-in evaluation metrics for energy usage, latency, computational cost and task performance, OpenJarvis aims to provide a standardized platform for building efficient, privacy-preserving personal AI assistants that can continuously learn from user interactions and operate reliably on consumer hardware.[76]

NVIDIA Open-Sources OpenShell, a Secure Runtime Environment for Scalable and Policy-Controlled Autonomous AI Agents

NVIDIA has open-sourced OpenShell, a secure runtime environment designed to support the deployment and operation of autonomous AI agents, addressing critical challenges around safety, privacy and system control in agentic workflows. Introduced as part of its broader agent infrastructure stack alongside NemoClaw, OpenShell provides a sandboxed execution layer that enables long-lived, self-evolving agents to plan, retain memory and interact with tools while enforcing policy-based constraints on network access, data usage and system permissions. The runtime is engineered to balance agent autonomy with enterprise-grade governance by integrating security guardrails, isolation mechanisms and privacy controls, allowing organizations to safely run agents across local and cloud environments. By open-sourcing OpenShell, NVIDIA aims to standardize secure agent execution, reduce risks associated with uncontrolled automation and accelerate the adoption of production-ready AI agents in enterprise and developer ecosystems.[77]

[76] https://www.marktechpost.com/2026/03/12/stanford-researchers-release-openjarvis-a-local-first-framework-for-building-on-device-personal-ai-agents-with-tools-memory-and-learning/
[77] https://www.marktechpost.com/2026/03/18/nvidia-ai-open-sources-openshell-a-secure-runtime-environment-for-autonomous-ai-agents/

Industry Update

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

Healthcare

MedGPT-OSS: A Comprehensive Open-Weight 20B Vision-Language Model for Advancing Privacy-Preserving Clinical AI Research

MedGPT-OSS is presented as an open-weight, 20-billion-parameter general-purpose vision-language model designed to advance biomedical AI by unifying radiology, pathology and clinical-text reasoning while addressing key deployment barriers such as closed-source restrictions and high computational requirements. The paper explains that instead of relying on overly complex architectures, the model integrates a GPT-oss language backbone with a visual front-end through a carefully optimized three-stage training process involving rigorous data curation and long-context multimodal alignment. Through this approach, MedGPT-OSS demonstrates that a model of its size can effectively bridge the performance gap by outperforming substantially larger open medical models on out-of-distribution multimodal reasoning tasks as well as challenging text-only clinical benchmarks, all while remaining efficient enough to run on commodity GPUs. The authors additionally release the entire training pipeline, evaluation suite and open-weight checkpoints, positioning the model as a verifiable, privacy-preserving foundation for on-premises, institution-specific clinical AI research.[78]

Model Medicine: A Comprehensive Clinical Framework for Diagnosing, Understanding and Treating Disorders in AI Models

The paper introduces Model Medicine, a clinical framework that conceptualizes AI models as entities with biological-like properties such as internal structures, dynamic processes, heritable traits, observable symptoms and treatable conditions and positions the discipline as a bridge between traditional AI interpretability research and a full clinical practice for complex AI systems. It presents five major contributions: a taxonomy of 15 subdisciplines across four divisions (Basic Model Sciences, Clinical Model Sciences, Model Public Health and Model Architectural Medicine); the Four Shell Model (v3.3), a behavioral genetics framework derived from empirical analysis of 720 agents and 24,923 decisions within the Agora-12 program, explaining behavioural emergence from Core–Shell interactions; Neural MRI, an open-source diagnostic tool mapping neuroimaging modalities to interpretability techniques and validated across four clinical case studies; a five-layer diagnostic system for comprehensive model assessment; and the foundations of clinical model sciences through tools such as the Model Temperament Index, Model Semiology and M-CARE for standardized case reporting. The paper also proposes the Layered Core Hypothesis, a biologically inspired parameter architecture and outlines a therapeutic framework linking diagnosis to treatment, ultimately calling for the research community to collaboratively expand this emerging clinical science for AI.[79]

Biological Artificial Intelligence Agents: A Multi-Dimensional Framework for Autonomous Reasoning, Tool Integration and Computational Discovery in Biological Research

The paper Artificial Intelligence agents for biological research: a survey presents a structured conceptual framework for biological AI agents that move beyond traditional predictive machine learning models toward autonomous AI systems capable of reasoning, planning, tool usage and iterative learning within biological research workflows. The study conducts a systematic review of over one hundred research works and proposes a unified five-dimensional taxonomy to organize biological AI agent systems across task domains, system architectures, interaction modes, evaluation strategies and resource integration. Within this framework, biological AI agents are described as intelligent systems that integrate LLMs, bioinformatics tools, biological databases, simulation environments and feedback mechanisms to perform complex tasks such as drug discovery, molecular design, multi-omics analysis and knowledge discovery with minimal human intervention. The framework also identifies key design patterns and challenges including reliability, privacy, scalability and evaluation standardization and positions biological AI agents as collaborative scientific partners that can automate research workflows and accelerate computational discovery in

[78] https://arxiv.org/abs/2603.00842
[79] https://arxiv.org/abs/2603.04722

biotechnology and life sciences. Overall, the paper establishes a conceptual and methodological framework for the development of agent-based artificial intelligence systems in biological research and computational biotechnology.[80]

Finance

AI Financial Intelligence Benchmark (AFIB): A Multi-Dimensional Evaluation Framework for Assessing Financial Reasoning Capabilities of LLMs

The study introduces the AI Financial Intelligence Benchmark (AFIB), a comprehensive evaluation framework designed to systematically assess the financial reasoning capabilities of LLMs used in financial analysis and investment research. The benchmark evaluates AI systems across five critical dimensions: factual accuracy, analytical completeness, data recency, model consistency and failure patterns. Using a dataset of more than 95 structured financial analysis questions derived from real-world equity research tasks, the study compares the performance of several AI systems, including GPT, Gemini, Perplexity AI, Claude and SuperInvesting. The results demonstrate notable performance variations among the models, with SuperInvesting achieving the highest overall performance, including an average factual accuracy score of 8.96 out of 10 and the highest analytical completeness score of 56.65 out of 70, while also exhibiting the lowest hallucination rate. Retrieval-focused systems such as Perplexity perform well in tasks requiring up-to-date information due to their ability to access live data, but they show comparatively weaker performance in analytical synthesis and reasoning consistency. Overall, the findings indicate that financial intelligence in LLMs is inherently multi-dimensional and that systems integrating structured financial data access with advanced analytical reasoning capabilities provide the most reliable support for complex investment research workflows.[81]

Mastercard Introduces "Verifiable Intent": A New Trust Framework to Secure AI-Driven Agentic Commerce and Verify Consumer Authorization

Mastercard has introduced Verifiable Intent, a standards-based trust framework designed to ensure secure and transparent transactions in the emerging era of AI-driven "agentic commerce," where autonomous AI agents can make purchases on behalf of consumers. Developed in collaboration with Google, the framework creates a cryptographically verifiable record linking a consumer's identity, instructions given to an AI agent and the resulting transaction into a tamper-resistant audit trail that merchants, payment networks and banks can reference to confirm authorization or resolve disputes. By generating proof of what a user actually authorized such as purchase limits, items, or conditions the system aims to build trust as AI agents increasingly perform tasks like reordering goods or booking services without real-time human interaction. Verifiable Intent is designed to work across multiple commerce protocols and platforms, incorporating industry standards from organizations such as FIDO Alliance, EMVCo, the Internet Engineering Task Force and the World Wide Web Consortium. The framework also uses privacy-preserving techniques like selective disclosure to limit the data shared between parties while maintaining verification capabilities and it is intended to integrate with Mastercard's Agent Pay infrastructure to support scalable and secure AI-powered transactions across the digital commerce ecosystem.[82]

Defence

Strategic AI-Military Collaboration: OpenAI's Agreement With the Pentagon to Deploy Guardrailed Artificial Intelligence on Classified Defense Networks

According to the report, OpenAI reached an agreement with the U.S. Pentagon to provide its artificial intelligence systems for use within classified military environments following a breakdown in negotiations between the Department of Defense and rival AI firm Anthropic, with the deal framed as a significant step in the growing integration of advanced AI into national security operations. The arrangement reportedly includes explicit safeguards that prohibit uses such as mass domestic surveillance, autonomous weapons control and high-stakes automated decision-making, while maintaining OpenAI's control over safety mechanisms, deployment architecture and oversight through cleared personnel and cloud-based systems. The agreement emerged amid political and industry tensions over ethical boundaries in military AI, especially after federal authorities moved to cut ties with Anthropic over disputes related to guardrails and permissible uses, positioning OpenAI as a key technology partner in defense AI initiatives while simultaneously drawing scrutiny from critics concerned about militarization, transparency and governance of frontier AI systems.[83]

Agentic Military AI Governance

[80] https://academic.oup.com/bib/article/27/1/bbag075/8499367
[81] https://arxiv.org/html/2603.08704v1
[82] https://www.mastercard.com/global/en/news-and-trends/stories/2026/verifiable-intent.html
[83] https://www.nytimes.com/2026/02/27/technology/openai-agreement-pentagon-ai.html

Framework (AMAGF) Proposes Measurable Oversight System to Safeguard Human Control in Autonomous Defense Operations

Researchers have proposed the Agentic Military AI Governance Framework (AMAGF) to address emerging control risks posed by agentic artificial intelligence systems in military environments. These AI systems capable of interpreting goals, modeling complex environments, planning multi-step actions, using tools, operating over long time horizons and coordinating autonomously introduce governance challenges that existing AI safety frameworks do not fully address. The study identifies six governance failure modes that could gradually weaken meaningful human oversight in defense operations. To mitigate these risks, the framework is structured around three pillars: preventive governance to reduce the likelihood of failures, detective governance for real-time monitoring of control degradation and corrective governance to restore or safely limit system operations when risks emerge. At its core is the Control Quality Score (CQS), a real-time metric designed to measure the level of human control over autonomous systems and trigger graduated responses when oversight weakens, providing a structured approach for managing advanced AI capabilities in military operations.[84]

Education

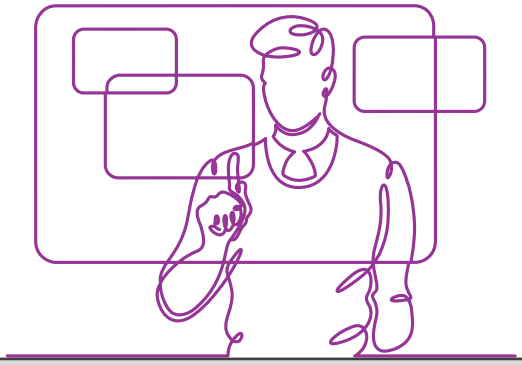
Microsoft Introduces Phi-4-Reasoning-Vision-15B: A Compact Multimodal AI Model Designed for Advanced Math, Science and GUI Understanding

Microsoft has released Phi-4-Reasoning-Vision-15B, a compact 15-billion-parameter open-weight multimodal AI model designed to integrate visual understanding with advanced reasoning capabilities across domains such as mathematics, science and graphical user interface (GUI) interpretation. The model can process both images and text to perform tasks including image captioning, document analysis, visual question answering and interpreting charts, screenshots and scientific diagrams. Built on the Phi-4 reasoning language model and combined with a vision encoder, it employs a hybrid reasoning approach that activates multi-step reasoning for complex tasks while using faster direct inference for simpler perception tasks, enabling an efficient balance between accuracy and latency. Developed with curated datasets and trained using 240 NVIDIA B200 GPUs over a short training cycle, the model demonstrates strong performance on multimodal benchmarks and is optimized for real-time interactive applications such as educational tutoring systems, document analysis tools and AI agents capable of navigating computer interfaces. Microsoft has made the model available through platforms including Azure AI Foundry, Hugging Face and GitHub to support research and development of efficient multimodal AI systems.[85]

Environmental Monitoring

Google AI Introduces Groundsource: A Gemini-Powered Methodology for Converting Unstructured Global News into Structured Historical Disaster Data to Improve Climate Risk Forecasting

Researchers at Google have introduced Groundsource, a new AI-driven methodology that uses the Gemini LLM to transform vast volumes of unstructured global news reports into structured, actionable historical datasets for scientific analysis and disaster forecasting. The system addresses the long-standing challenge of limited historical data for certain natural hazards by automatically analyzing news articles, government reports and other public sources to extract verified event information such as location, timing and severity. Groundsource processes multilingual news content from around the world, standardizes it into English and applies AI-based classification, temporal reasoning and geospatial mapping to convert narrative reports into structured event records. Using this pipeline, researchers created an open dataset containing approximately 2.6 million historical flash-flood events across more than 150 countries, significantly expanding the available data for training and validating predictive models. The structured dataset enables improved flood forecasting systems such as those deployed in Google's Flood Hub to predict urban flash floods up to 24 hours in advance while also demonstrating how LLMs can convert the world's unstructured news memory into reliable scientific data that supports climate research, disaster preparedness and global resilience planning.[86]



[84] https://arxiv.org/abs/2603.03515

[85] https://www.marktechpost.com/2026/03/06/microsoft-releases-phi-4-reasoning-vision-15b-a-compact-multimodal-model-for-math-science-and-gui-understanding/

[86] https://www.marktechpost.com/2026/03/13/google-ai-introduces-groundsource-a-new-methodology-that-uses-gemini-model-to-transform-unstructured-global-news-into-actionable-historical-data/

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

ISO 42001:2023 External Surveillance Audit

The ISO 42001:2023 External Surveillance Audit was conducted on February 9–10 in Bengaluru, marking a key milestone in Infosys' continued leadership in responsible, accountable and standards driven AI operations. The audit was carried out by TÜV as part of the ongoing certification lifecycle for the world's first enterprise level ISO 42001 accreditation.

Infosys leadership—including Balakrishna DR (EVP, Head – Global Services), Nabarun (EVP, Group Head – Quality), ANILP (VP, Head – Audit & Assessment) and Shreekantha V Ayya (AVP, Senior Unit Quality Head)—supported the engagement, reinforcing the organization's commitment to rigorous AI governance, compliance and quality excellence. The audit was anchored by the Infosys Responsible AI Office, led by Syed Ahmed (Global Head – Responsible AI Office), with strong support from the Responsible AI Governance and Compliance team, enabling a comprehensive review of all 38 ISO 42001 controls, associated artefacts and enterprise processes. Three AI use cases—two client facing and one internal—were also evaluated for adherence to the standard.

The audit concluded with **zero non conformities**, successfully renewing Infosys' ISO 42001 certification until March 2027, reaffirming the company's position as a global benchmark in Responsible AI maturity and operational excellence.

LFN ONE India Summit 2026 | February 2026 | Infosys Bangalore Campus

The LFN ONE India Summit 2026, hosted at the Infosys Bangalore campus, was a significant industry event bringing together leaders, partners and innovators from the open source networking ecosystem. The summit saw strong participation with **120+ attendees**, representing **22 customer and partner organizations** across telecom, cloud, networking and AI domains.

The event featured:

- **16 speaker sessions** covering advancements in networking, automation, open-source platforms and AI driven solutions
- **6 experiential showcase stalls,** where teams demonstrated cutting edge projects and live solutions

As part of the event, team presented the Salus Project - Open source Infosys Responsible AI Toolkit hosted in LFN, highlighting its capabilities, contribution value and relevance within the broader ecosystem of secure, scalable and AI enabled architecture. The demonstration added strong visibility and positioned Salus work as a meaningful contribution to the LFN community, with support from Representatives from Infosys Responsible AI including Sonakshi Bisht, Bhanuteja Akurathi, Jayaprakash B.R., Jagadish Babu P, Balachandar Gurusamy and Thenmozhi Krishnan.

West Midlands Mayoral Delegation Visit to Infosys | February 26, 2026 | Bengaluru

On February 26, 2026, a 32-member West Midlands Mayoral Delegation from the United Kingdom visited the Infosys Bengaluru campus to strengthen UK–India collaboration in AI, digital transformation and innovation-led growth. Led by Richard Parker, the delegation included Neil Rami (CEO, West Midlands Growth Company), Julie Nugent (CEO, Coventry City Council) and Vijay Tank (COO, E.ON Energy Infrastructure Solutions UK). Christopher Ford and Ashish Tewari coordinated the engagement with Infosys leadership to showcase the company's approach to building safe, scalable and enterprise-ready AI systems.

Syed Quiser Ahmed, Global Head – Responsible AI Office, anchored the session and presented Infosys' AI-first strategy, governance frameworks and commitment to transparent and responsible AI adoption, highlighting opportunities in AI safety, regulatory alignment and enterprise modernization.

The campus visit was followed by Living Labs demonstrations and individual showcases of Responsible AI and Digital Twin technologies, along with immersive walkthroughs of humanoid and autonomous machine capabilities. The delegation also experienced an interactive tennis AI showcase linked to Infosys' collaboration with the Australian Open, followed by a 270-degree immersive innovation overview and healthcare technology exploration.

The engagement was supported by Girish, Nawaz, Bharathi, Janvi, Deepthi, Abhishek and Saikrishna and concluded with leadership meetings focused on long-term partnerships and strengthening innovation-driven collaboration between the UK and India.

Infosys Responsible AI Office Townhall Meet | February 27, 2026 | Bengaluru

On 27th Feb, the Infosys Responsible AI Office Townhall took a creative turn as the leaders ditched traditional presentations and instead performed an engaging skit highlighting the core areas of Responsible AI. The act brought governance, fairness, privacy, security and accountability to life in a relatable and memorable way. This was followed by interactive Responsible AI games that reinforced key concepts, along with rewards and recognition to celebrate standout contributions. The energy seamlessly carried into collaborative team-building activities.

It was a refreshing reminder that learning can be impactful and fun at the same time. It strengthened our shared commitment to Responsible AI while deepening team connection.

AI Reset and Future Leadership Session | March 7, 2026 | Jaipuria Institute of Management, Lucknow

On March 7, 2026, Jaipuria Institute of Management hosted an insightful session on "AI Reset and Future Leadership," drawing strong participation from PGDM students and faculty. The discussion centered on how AI is moving from supportive tools to active participants in organizational decision-making, reshaping the expectations placed on future managers.

Representing Infosys, Ashish Tewari, Head – Responsible

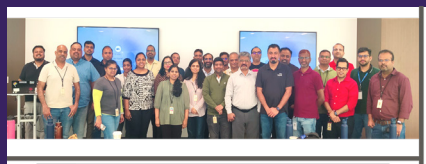
AI Office (APAC & Middle East), outlined the AI Reset as a convergence of economic, technological and informational shifts. He explained how these transitions are redefining productivity and influencing decision flows, emphasizing the importance of governance, accountability and human judgment as core leadership capabilities in an AI-enabled world.

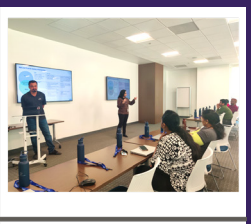
The session evolved into an engaging exchange, with students exploring themes of trust, responsible innovation and organizational readiness. The interaction reinforced the need for leaders who balance ambition with ethical clarity, reflecting Infosys' commitment to advancing responsible AI awareness across academic ecosystems.

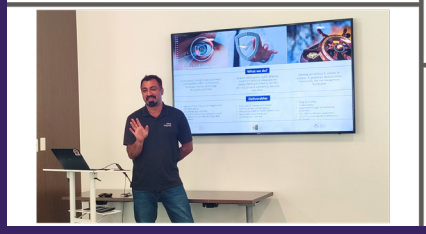
Infosys Responsible AI Team Wins Five AFE Awards | March 11, 2026 | Bengaluru

On March 11, 2026, the Responsible AI team at Infosys achieved a major milestone, securing 5 AFE Awards that recognized exceptional performance, perseverance and technical leadership across the organization. The team earned Gold in "People Development", along with Silver awards in "Internal Customer Delight" and "Sales and Marketing - Brand Management", reflecting its deep impact on enterprise wide enablement and stakeholder trust. Additionally, Syed Ahmed, Global Head – Responsible AI Office, was honored with the "Infosys Culture Ambassador" silver award, while Ashish Tewari, Head – Responsible AI Office (APAC & Middle East) was recognized as "Infosys Functional/Domain Champion" (Bronze) for his contributions to strengthening AI governance and operational excellence. The recognitions were the culmination of months of sustained effort, cross team collaboration and unwavering dedication to advancing Responsible AI capabilities within Infosys and across the globe.

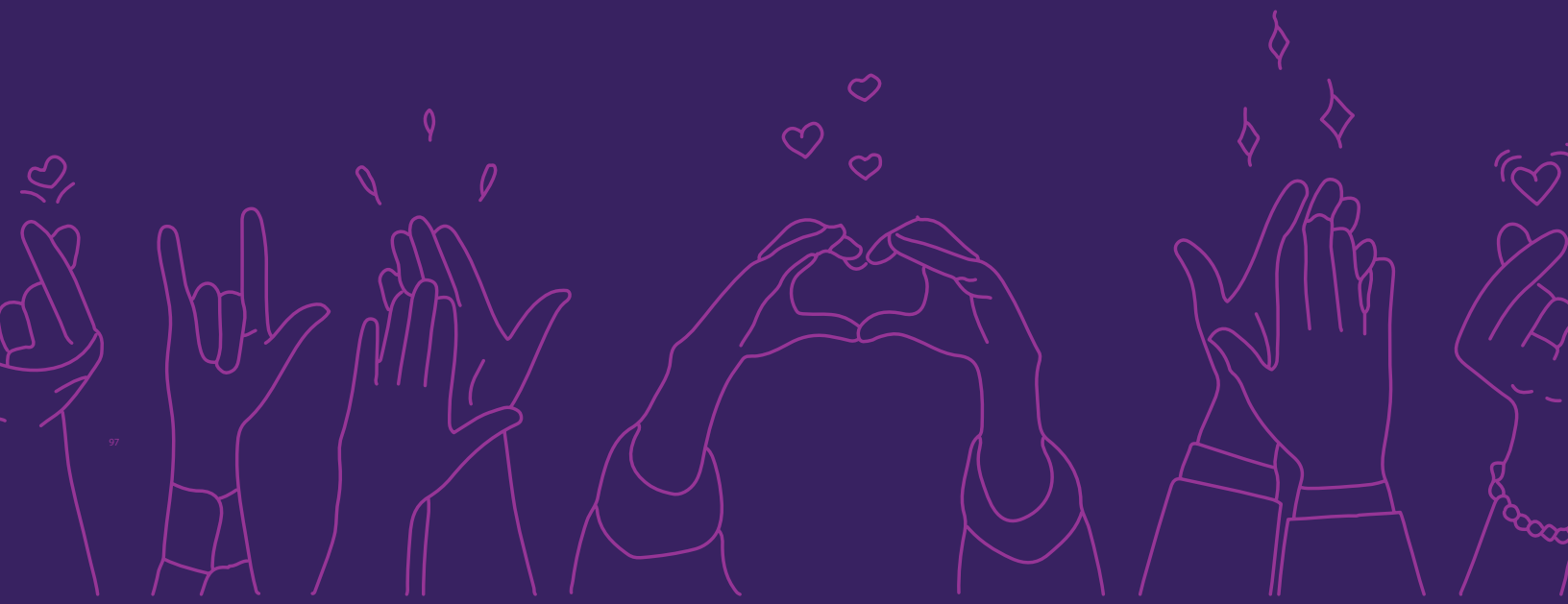
Responsible AI Session for Platinum Club Members | March 11, 2026 | Richardson, Texas







On March 11, 2026, Infosys hosted a focused Responsible AI session for its Platinum Club members at the Richardson, Texas campus, bringing together more than 30 employees to strengthen awareness of ethical, resilient and enterprise ready AI. The session was led by Mandanna A N from the Infosys Responsible AI Office, who shared practical insights on Responsible AI and its role in enhancing enterprise resilience and client focused innovation. Participants received a clear overview of Infosys' Responsible AI frameworks, including the value of Responsible by Design and Secure by Design principles that anchor the organization's AI governance model. The engagement fostered meaningful discussions, reinforced leadership alignment and served as an important touchpoint for deepening Responsible AI maturity within the Platinum Club community, further strengthening Infosys' commitment to building trustworthy, future ready and high integrity AI systems.

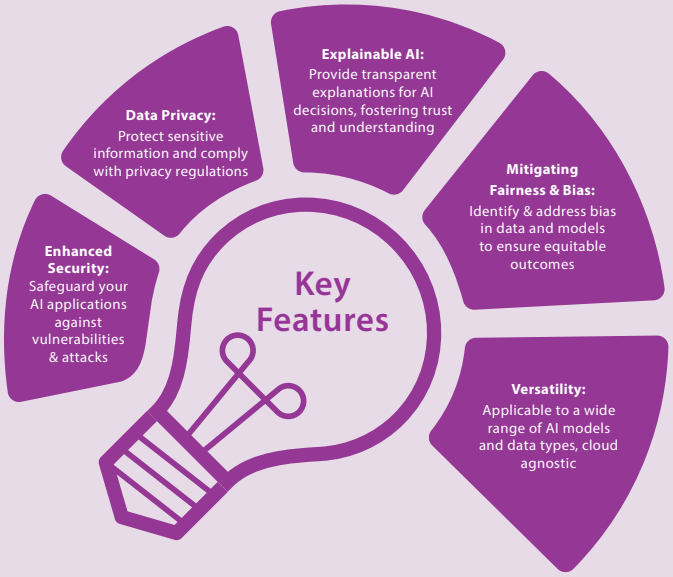


Infosys Responsible AI Toolkit – A Foundation for Ethical AI

The Open-Source Infosys Responsible AI Toolkit can be accessed from its public GitHub repo[87] also as project Salus[88]

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,***[89] *both as a repo and also as an interactive playground to explore. Star us and be a part of the Responsible AI Revolution!* ★★★★★

[87] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit
[88] https://github.com/salus-rai/salus
[89] https://huggingface.co/spaces/InfosysEnterprise/Moderation-playground

Infosys offerings for Responsible AI - centred around the AI3s framework

Infosys Responsible AI offerings help organisation enable ethical, trustworthy and scalable AI adoption and is a combination of accelerators and solutions designed to drive responsible AI adoption across enterprises and ensure strong AI Governance, ethics and Security.

SCAN

- Threat Intelligence & Reporting
- Impact Assessment Framework
- Comprehensive monthly market scan report

SHIELD

- Automated AI Management System
- Infosys Responsible AI Toolkit (Opensource)
- Responsible Agent Framework
- Automated Red Teaming

STEER

- AI Governance & Responsible AI CoE Advisory & Consulting
- Standards and Regulations Assessment & Readiness consulting.
- Responsible AI literacy & Capacity Building

"Turn Responsible AI into a Competitive Advantage: leverage our pre-built solutions, frameworks and advisory"

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.

IF FACES ARE DATA, TRUST MUST BE THE FRAMEWORK.

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