MARKET SCAN REPORT

MARCH 2025

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

Infosys®
Navigate your next

Dear Readers,

As we stand at the edge of a new era, the rapid advancements in artificial intelligence (AI) are changing our world in profound ways. The latest edition of the Market Scan News Report explores the diverse landscape of AI, highlighting significant progress in governance, technological breakthroughs, and the ethical challenges that come with this transformative journey.

This month's developments signal growing international alignment on AI governance and regulation. A key highlight is the emerging collaboration between the European Union and India, aimed at fostering innovation while upholding shared democratic values in the development and deployment of AI technologies. Alongside this, countries such as Canada and El Salvador have also advanced new initiatives, reinforcing a collective global commitment to ensuring that AI is developed responsibly, ethically, and in the public interest.

In the domain of AI-related incidents, two major events have drawn attention: the Tesla Cybertruck crash, raising concerns around autonomous vehicle safety, and a disturbing case of AI voice cloning fraud in Canada, which underscores the growing threat of AI-enabled deception. These incidents serve as a stark reminder of the potential risks associated with emerging technologies, emphasizing the urgent need for robust safety protocols and regulatory oversight to safeguard individuals and communities.

On the technical front, innovation continues at a rapid pace. Notable highlights include the release of OpenAI's GPT-4.5, bringing improved language generation and reasoning capabilities, and Tencent's Hunyuan Turbo S, aimed at enhancing performance across multilingual and multimodal tasks. These advancements mark significant steps in the pursuit of powerful, efficient, and user-centric AI applications across industries—from healthcare to environmental monitoring.

The report also shines a light on cutting-edge research, featuring emerging frameworks such as Himitsu8 and PlanGEN. These tools represent the next frontier in building adaptive, intelligent, and autonomous systems, capable of solving complex real-world problems and supporting more informed decision-making processes.

Our report highlights the work of researchers at the University of Navarra and the development of benchmarks like MinorBench, exemplifying the dedication to ensuring fairness, safety, and privacy in AI applications.

In light of these developments, we recommend that stakeholders across sectors:

- Stay continuously informed on AI regulations and global best practices to ensure compliance and ethical use.
- Implement robust safety protocols and conduct thorough testing to mitigate potential risks.
- Proactively engage in global partnerships to share insights and advance responsible innovation.
- Embrace emerging research, frameworks, and benchmarks to enhance the trustworthiness, fairness, and security of AI systems—especially in areas with high societal impact.

As we navigate the evolving AI landscape, it is clear that progress must be paired with responsibility. The insights in this report offer a well-rounded perspective on the state of AI today and the path ahead.

We encourage all readers to delve into the detailed sections of this report to fully grasp the breadth and depth of the topics covered. Let us move forward with a shared vision of harnessing AI's potential to create a better, more equitable future for all.

Warm regards

Syed Ahmed
Head- Infosys Responsible AI Office

Table of Contents

AI Regulations, Governance and Standards

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

AI Regulations and Governance across globe

EU and India Sign Joint Statement to Strengthen Collaboration in AI

On February 28, 2025, in New Delhi, the European Union and India adopted a joint statement following the Trade and Technology Council (TTC) second meeting. On AI specifically, the EU and India (1) reiterated their commitment to safe, secure, trustworthy, human-centric, sustainable, and responsible AI; (2) agreed to deepen cooperation between the European AI Office and India AI Mission; and (3) committed to joint projects focused on ethical and responsible AI frameworks. This builds on existing Research and Development (R&D) collaboration in high-performance computing applications related to natural hazards, climate change, and bioinformatics.[1]

Establishment of the Independent International Scientific Panel on AI and Global Dialogue on AI Governance

The United Nations has released an Elements Paper detailing the establishment and functioning of the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance. This initiative aims to regularly assess the opportunities, capabilities, impacts, and risks of artificial intelligence across various disciplines. The Panel will conduct horizon scanning to predict future directions and provide early warnings on rapidly evolving AI technologies. Additionally, it will initiate targeted research to address capacity-building gaps. The overarching goal is to ensure that AI technologies are developed and governed in a way that benefits humanity, promotes sustainable development, and respects human rights.[2]

FCC Forms Council to Address U.S.-China Tech Race

The US Federal Communications Commission (FCC) is establishing a national security council to counter cyber threats from other countries, especially China and ensure American dominance in AI, quantum computing, 6G, and autonomy. Led by Adam Chan, the council aims to mitigate U.S. vulnerability to cyberattacks, espionage, and surveillance by hostile states. This move reflects growing competition between China and the U.S. over technology, with both nations striving for supremacy in critical tech areas. China's focus on building capabilities in AI and quantum computing, highlighted by the launch of DeepSeek's AI model, underscores the urgency for the U.S. to secure its lead. The council will work to ensure the U.S. wins the strategic competition with China over critical technologies.[3]

Cyprus and UAE Forge AI Partnership to Drive Innovation and Global Collaboration

The Memorandum of Understanding (MoU) was signed between Cyprus and the United Arab Emirates (UAE) to enhance cooperation in artificial intelligence (AI). This agreement marks Cyprus's first AI MoU outside the European Union and aims to accelerate AI innovation and adoption in both countries. The collaboration focuses on exchanging expertise, fostering AI research, and developing AI applications in public services. The MoU is part of Cyprus's broader strategy to strengthen its AI capabilities and align with global AI leaders, including the UAE, USA, Japan, and Israel.[4]

Setting the Bar: Tajikistan's AI Regulation Efforts at the UN

Tajikistan has proposed a groundbreaking AI regulation initiative at the United Nations, aiming to establish global standards for the ethical and responsible use of artificial intelligence. This initiative underscores Tajikistan's commitment to addressing the challenges and opportunities presented by AI technologies. By advocating for international cooperation and robust regulatory frameworks, Tajikistan seeks to ensure that AI development benefits all nations while mitigating potential risks.[5]

[1] https://ec.europa.eu/commission/presscorner/detail/en/statement_25_643

[2] https://www.un.org/global-digital-compact/sites/default/files/2025-02/250228%20Elements%20Paper%20AI%20Scientific%20Panel%20and%20Global%20Dialogue.pdf

[3] https://www.fcc.gov/fcc-council-national-security

[4] https://chiefscientist.gov.cy/cyprus-uae-ai-mou-strengthening-global-partnerships-and-accelerating-ai-innovation/#:~:text=The%20AI%20MoU%20between%20Cyprus,Sultan%20Olama%20that%20have%20taken

[5] https://asiaplustj.info/en/node/346663

US

Congressional Committee Initiates New Federal Privacy Law Dialogue

The U.S. House Committee on Energy and Commerce has launched a new initiative to develop comprehensive federal privacy legislation. This marks the third attempt in recent congressional terms. The committee has formed a new working group and issued a request for information (RFI) to gather stakeholder input on key issues. The RFI aims to address challenges such as balancing consumer protections with technological advancements and navigating the complex landscape of state and federal privacy laws. The committee's goal is to create a framework that provides clear digital protections for Americans while maintaining the U.S.'s competitive edge.[6]

US First Lady Melania Trump Backs TAKE IT DOWN Act

On March 4, 2025, US First Lady Melania Trump hosted a roundtable in support of the TAKE IT DOWN Act, a bill aimed at protecting individuals from deepfake and revenge pornography, which has passed the Senate but not yet the House of Representatives. Among other measures, the bill proposes to make it a federal crime to knowingly publish or threaten to publish non-consensual sexual images on online platforms. It explicitly includes realistic, computer-generated intimate images depicting identifiable individuals. The bill clarifies that consent to create an image does not equate to consent for its publication. The bill would also require tech and social media platforms to remove child sexual abuse material and non-consensual sexual images within 48 hours of being alerted by a victim. It is expected that with the First Lady's support, the bill will be accelerated as a "top priority" for the House of Representatives to pass into law.[7]

[6] https://iapp.org/news/a/congressional-committee-kickstarts-new-federal-privacy-law-dialogue/

[7] https://www.commerce.senate.gov/2025/3/house-leaders-pledge-to-advance-take-it-down-act-at-sen-cruz-s-bipartisan-roundtable-with-first-lady-melania-trump

Brazil

Brazilian President Urges ANPD to Proactively Regulate AI Ahead of Comprehensive Legal Framework

On March 10, 2025, Brazilian President Luiz Inácio Lula da Silva emphasized the urgency for the National Data Protection Authority (ANPD) to advance regulations on artificial intelligence (AI) before a comprehensive legal framework is established. He highlighted the necessity of proactive measures to address the rapid development and integration of AI technologies across various sectors. The President stressed that early regulation by the ANPD would help mitigate potential risks and ensure that AI advancements adhere to ethical standards and data protection laws. This initiative aims to protect citizens' rights and promote responsible AI usage in Brazil, setting a foundation for future legislative efforts. The President's call to action underscores the importance of timely and effective responses to AI-related challenges.[8]

UK

Artificial Intelligence (Regulation) Bill Reintroduced to House of Lords in the United Kingdom

On March 4, 2025, Lord Holmes of Richmond reintroduced his private member's bill, the "Artificial Intelligence (Regulation) Bill," to the House of Lords in the United Kingdom. Initially proposed in November 2023, the bill had lapsed due to a change in government. Key features of the bill include the establishment of a new AI Authority to ensure alignment among regulators from different economic sectors and identify regulatory gaps. The AI Authority would also monitor economic risks from AI, conduct horizon scanning of developing technologies, facilitate sandbox initiatives for testing new AI models, and accredit AI auditors. The bill proposes regulatory principles for AI development and usage, emphasizing safety, security, robustness, transparency, fairness, accountability, governance, contestability, and redress. It also includes clauses for meaningful, long-term public engagement on AI opportunities and risks, transparency around third-party

[8]https://valorinternational.globo.com/politics/news/2025/03/10/anpd-must-advance-ai-regulation-before-legal-framework-says-president.ghtml

data usage, and the principle of informed consent for using intellectual property in training datasets.[9]

Accelerating UK Science: A Strategic Vision for the AI Era

The recommendation report from the Tony Blair Institute outlines a strategic vision for accelerating UK science in the age of artificial intelligence (AI). It emphasizes the need for the UK to build AI-ready scientific data, develop advanced software tools, and secure a robust AI talent pipeline. Key recommendations include investing in AI research infrastructure, creating regulatory sandboxes for AI-enabled robotics, and enhancing cross-sector mobility to facilitate industry-to-academia transitions. The strategy aims to position the UK as a global leader in AI-driven scientific discovery, fostering innovation and addressing challenges such as data fragmentation and the need for digitization.[10]

UK Government's Response to Regulatory Horizons Council Report on AI as a Medical Device

On March 10, 2025, the UK government released its response to the Regulatory Horizons Council's report on the regulation of AI as a medical device (AIaMD). The government accepted most of the Council's recommendations, including providing long-term funding for the Medicines and Healthcare products Regulatory Agency (MHRA), strengthening regulatory capacity, using a 'legislatively light' framework, ensuring manufacturers mitigate risks such as AI bias, and implementing innovative mechanisms for accelerated access with post-market evidence generation and stakeholder collaboration. Additionally, the government accepted in principle recommendations such as requiring manufacturers to ensure safety in local deployments and maintaining a 'plan B' for device withdrawal. While none of the recommendations were outright rejected, some have limitations in their current implementation.[11]

Europe

Regulatory Frameworks for AI: Addressing Bias with the AI Act and GDPR

The report released by European Parliamentary Research Service (EPRS) on "Algorithmic Discrimination under the AI Act and the GDPR" examines the intersection of the AI Act and the General Data Protection Regulation (GDPR) in addressing discrimination and bias in AI systems. It underscores the AI Act's commitment to fostering human-centric, trustworthy, and sustainable AI while upholding fundamental rights. The document delves into the challenges of algorithmic discrimination across various sectors, including autonomous vehicles, job recruitment, and credit scoring. It also highlights the necessity of processing special categories of personal data to detect and correct biases in high-risk AI systems.[12]

EU Launches Second Wave of AI Factories to Boost Innovation Across Europe

The European High Performance Computing Joint Undertaking (EuroHPC JU) has announced the selection of six new AI Factories in Austria, Bulgaria, France, Germany, Poland,

[10] https://institute.global/insights/tech-and-digitalisation/a-new-national-purpose-accelerating-uk-science-in-the-age-of-ai

[11] https://www.gov.uk/government/publications/the-regulation-of-artificial-intelligence-as-a-medical-device-government-response-to-the-rhc/the-regulation-of-artificial-intelligence-as-a-medical-device-government-response-to-the-regulatory-horizons-council

[12] https://www.europarl.europa.eu/RegData/etudes/ATAG/2025/769509/EPRS_ATA(2025)769509_EN.pdf

and Slovenia, supported by a combined national and EU investment of around €485 million. These AI Factories aim to drive innovation across the EU by providing privileged access to AI startups and SMEs, leveraging Europe's world-class network of supercomputers to enhance the training and development of large-scale, trustworthy, and ethical AI models. This initiative follows the first selection of seven AI Factories in December 2024 and is part of the EU's broader strategy to become a global leader in AI, with plans to mobilize up to €200 billion in AI investments, including the deployment of several AI Gigafactories across Europe.[13]

Publication of the European Health Data Space (EHDS) Regulation

On March 5, 2025, the European Health Data Space (EHDS) Regulation was published in the Official Journal of the European Union. This regulation, part of the European Health Union, aims to create a unified framework for the use and exchange of electronic health data across the EU. It enhances individuals' access to and control over their personal health data, while also enabling the reuse of certain data for public interest, policy support, and scientific research. The EHDS promotes a health-specific data environment that supports a single market for digital health services and products, establishing a harmonized legal and technical framework for electronic health record systems. The regulation will enter into force on March 26, 2025, marking the start of the transition phase towards its application.[14]

Guiding the Future: European Commission Releases Latest AI Code of Practice Draft

The European Commission has published the third draft of the General-Purpose AI Code of Practice, crafted by independent experts. This draft aims to provide comprehensive guidelines for the development and deployment of trustworthy and safe AI models. It incorporates feedback from a wide range of stakeholders, including industry leaders, policymakers, and civil society, to ensure a balanced and inclusive approach. The Code outlines key principles, risk assessment measures, and transparency requirements, setting a robust framework for AI governance in the EU[15]

Swedish Government Proposes AI Facial Recognition Bill to Combat Crime

The Swedish government has proposed a bill to allow police to use AI-powered facial recognition technology in criminal investigations. This move aims to address the surge in violent offenses, including human trafficking, kidnapping, and murder, which have plagued Sweden for over a decade. The country recorded the highest rate of fatal gun violence per capita in the EU in 2023. The proposed law, set to take effect in early 2026 if approved, would ensure compliance with personal integrity laws and be used only in cases of significant importance. This legislation seeks to enhance law enforcement capabilities by enabling the use of advanced biometric data, such as DNA, fingerprints, and facial images, to identify suspects and solve crimes more efficiently. The legislative text has yet to be revealed.[16]

Canada

Canada Announces Key Initiatives for Safe and Ethical AI Adoption

On March 7, 2025, the Canadian government unveiled several initiatives to promote the safe and ethical use of artificial intelligence (AI). The Minister of Innovation, Science and Industry introduced a refreshed Advisory Council on Artificial Intelligence and a new Safe and Secure Artificial Intelligence Advisory Group, chaired by Yoshua Bengio. Additionally, a guide for managers of AI systems was published to help organizations implement Canada's Voluntary Code of Conduct on AI, providing practical steps for responsible AI integration.

[13] https://digital-strategy.ec.europa.eu/en/news/second-wave-ai-factories-set-drive-eu-wide-innovation

[14] https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en#:~:text=On%205th%20March%202025%2C%20the,the%20transition%20phase%20towards%20application.

[15] https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts

[16] https://www.reuters.com/world/europe/swedish-government-proposes-bill-allow-police-use-ai-face-recognition-2025-03-20/

Six new organizations—CIBC, Clir, Cofomo Inc., Intel Corporation, Jolera Inc., and PaymentEvolution—joined 40 other signatories in committing to the voluntary AI code of conduct, pledging to apply the code to their operations when developing and managing generative AI systems. These efforts aim to enhance governance, transparency, and risk management in AI adoption.[17]

Australia

Australian Parliament's Audit Committee Recommends Whole-of-Government AI Framework

The Joint Committee of Public Accounts and Audit has released its report on public sector AI use, examining current policy settings across the Commonwealth Government to determine their suitability. The Committee's recommendations include: introducing detailed questions on AI use and understanding into the annual APS census by the Australian Public Service Commission; establishing a whole-of-government working group within 12 months to consider mandatory rules, governance frameworks, and legislation for AI systems; and creating a statutory Joint Committee on Artificial Intelligence and Emerging Technologies to provide Parliament with effective oversight of AI's impact on government and public service operations. These measures aim to ensure responsible and ethical AI use, emphasizing transparency, accountability, and the protection of citizens' rights.[18]

India

India Advances AI Capabilities with AI Kosha and AI Compute Portal Launch

India has made significant strides in its AI development strategy with the launch of AI Kosha and the AI Compute Portal. AI Kosha serves as a comprehensive data platform, providing access to a wide range of non-personal datasets essential for training AI models. This initiative aims to support AI research and innovation by offering robust data infrastructure. Additionally, the AI Compute Portal offers access to over 14,000 GPUs at subsidized rates, enabling researchers, startups, and developers to enhance their AI and computing capabilities. These efforts are part of India's broader strategy to achieve self-reliance in AI and computing technologies, positioning the country as a key player in the global AI landscape.[19]

[17]https://www.canada.ca/en/innovation-science-economic-development/news/2025/03/canada-moves-toward-safe-and-responsible-artificial-intelligence.html

[18]https://www.aph.gov.au/About_Parliament/House_of_Representatives/About_the_House_News/Media_Releases/Audit_Committee_recommends_whole_of_government_framework_for_AI_use

[19]https://www.msn.com/en-in/money/news/india-now-has-a-data-platform-to-train-its-own-ai-models-work-on-own-gpu-begins/ar-AA1AooZl?ocid=BingNewsVerp

China

China Implements New Regulations for Identifying AI-Generated Synthetic Content

On March 14, 2025, the Cyberspace Administration of China (CAC), in collaboration with several ministries, issued new measures to regulate the identification of synthetic content generated by AI. These regulations, set to be enforced from September 1, 2025, include explicit and implicit identification requirements for service providers, ensuring compliance with existing laws and implementing technical measures for content dissemination. The measures prohibit the deletion, alteration, forgery, or concealment of AI-generated content identification, with violations to be addressed by competent authorities. Additionally, the regulations introduce a national standard for identifying synthetic content and provide cybersecurity guidelines for coding rules related to AI-generated content identification.[20]

Japan

Japan Cabinet Approves National AI Bill

On 28 February 2025, Japan's Cabinet approved the Bill on the Promotion of Research and Development and Application of Artificial Intelligence-Related Technologies. Notable aspects of this AI bill include: (1) officially recognizing AI as a core driver of economic and social development, with a government-led AI Strategy Headquarters overseeing its implementation; (2) acknowledging risks like data leaks and misuse, but focusing on promoting transparency rather than imposing strict regulations; (3) supporting research institutions and businesses to bolster Japan's AI leadership and enhance international competitiveness; (4) emphasizing international cooperation, particularly in shaping global AI standards; and (5) requiring businesses utilizing AI to cooperate with government policies and investigations, with public "naming and shaming" as a potential reputational consequence for non-cooperative businesses. The bill will now proceed to the National Diet for debate.[21]

[20] https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

[21] https://www.japantimes.co.jp/news/2025/02/28/japan/ai-abuse-countermeasures/

Hungary

Hungary's President Signs Controversial Laws Allowing Facial Recognition and Banning Pride Parades Amid Protests

On 19 March 2025, Hungary's President signed a controversial law proposed by Prime Minister Viktor Orban's ruling party that will amend the 2021 Child Protection Act (Act LXXIX of 2021) to ban LGBTQ+ communities from holding their annual Pride march. The law also includes powers for police to use facial recognition cameras to identify people who attend the event and impose fines on participants. The law had passed in the Hungarian parliament in a 136-27 vote and was pushed through parliament in an accelerated procedure after being submitted only a day earlier. The legislative text of the amendment has yet to be revealed to the public. As facial recognition technology is considered a high risk AI and a privacy concern, this law gets wide attention in the responsible AI perspective. [22]

Spain

Addressing Common Misconceptions About Artificial Intelligence and Data Protection: Insights from the Spanish Data Protection Agency

The Spanish Data Protection Agency (AEPD) recently published an article addressing common misconceptions about artificial intelligence (AI) and its relationship with data protection. It emphasizes understanding AI's capabilities and limitations to ensure responsible use, noting that AI systems can make errors, especially if not properly trained or if the data is biased. The article highlights the importance of data quality and human oversight to monitor AI decisions and intervene when necessary. Ethical guidelines are essential to ensure AI systems are transparent, accountable, and respect user privacy. Additionally, the AEPD clarifies that incidental data processing is not exempt from data protection regulations, distinguishes between personally identifiable information and personal data,

[22] https://www.reuters.com/world/europe/hungarys-president-signs-law-banning-pride-parade-despite-protests-2025-03-19/?utm_source=substack&utm_medium=email

and explains that case law on search engine data protection does not automatically apply to generative AI. The agency advocates for a "data protection by design" approach to align AI development with data protection principles.[23]

Spain's digital administrator has also released its own technical specification covering its new online age verification system, which will use the W3C Verifiable Credentials (VCs) data model for limited disclosure.[24]

South Korea

South Korea Launches AI Copyright System Improvement Council to Modernize Copyright Protection

On March 19, 2025, South Korea's Ministry of Culture, Sports and Tourism, in collaboration with the Korea Copyright Commission, announced the formation of the 2025 AI Copyright System Improvement Council. This council is tasked with studying and updating the country's copyright protection systems to better align with the rapid spread of artificial intelligence (AI). The council aims to publish guidance on the criteria for copyright registration of works generated using AI outputs and to determine the criteria for copyright infringement involving AI outputs during the first half of 2025. This initiative reflects South Korea's proactive approach to addressing the evolving challenges posed by AI in the realm of copyright protection.[25]

El Salvador

El Salvador Passes AI Law to Foster Development and Attract Tech Talent

On February 28, 2025, El Salvador's National Bitcoin Office announced that the government has passed AI legislation designed to provide regulatory clarity and certainty for the industry, including the development and innovation of open-source models. The law includes several key provisions: Article 19 offers safeguards such as sandbox protections and shields against third-party mischief; an AI Registry will be established for AI developers to register and fully enjoy these protections; and an AI Lab will be created to promote AI research, development, and application within government services and institutions, such as improving traffic flow, monitoring water quality, and optimizing geothermal resources. A draft of this AI law has yet to be published to the public.[26]

[23] https://www.aepd.es/en/press-and-communication/blog/addressing-misconceptions-%20of-artificial-intelligence

[24] https://www.biometricupdate.com/202408/age-assurance-demand-spurs-technical-specifications-from-euconsent-spain#:~:text=The%20euConsent%20project%20has%20announced%20the%20publication%20of,its%20age%20assurance%20scheme%2C%20AgeAware%2C%20for%20public%20consultation.

[25] https://www.mcst.go.kr/kor/s_notice/press/pressView.jsp?pSeq=21712&utm_source=substack&utm_medium=email

[26] https://x.com/bitcoinofficesv/status/1895279587250110702

Kenya

Kenya to Unveil Draft Policy on Artificial Intelligence in Two Months

The Kenyan government is set to unveil a draft policy on artificial intelligence (AI) within the next two months, according to Principal Secretary Jerome Ochieng. This policy aims to guide the responsible and ethical use of AI across various sectors. The draft will address key areas such as data privacy, security, and the promotion of AI innovation. It will also outline measures to ensure that AI technologies are used to benefit society while mitigating potential risks. The policy is part of Kenya's broader strategy to position itself as a leader in AI and digital transformation in Africa.[27]

Ireland

Ireland Names Eight Bodies as Initial EU AI Act Enforcers

On March 6, 2025, the Irish government designated eight public bodies as competent authorities to implement and enforce the new EU AI Act within their sectors. These bodies include the Central Bank of Ireland, the Commission for Communications Regulation, the Commission for Railway Regulation, the Competition and Consumer Protection Commission, the Data Protection Commission, the Health and Safety Authority, the Health Products Regulatory Authority, and the Marine Survey Office of the Department of Transport. This decision follows the government's approval of a "distributed model of implementation." Additional authorities and a lead regulator, who will coordinate enforcement of the Act and provide centralized functions, will be designated in the future.[28]

[27] https://www.standardmedia.co.ke/business/sci-tech/article/2001513079/government-to-unveil-ai-draft-policy-in-two-months-says-ps-tanui

[28] https://www.gov.ie/en/press-release/441b7-ministers-burke-and-smyth-welcome-government-approval-of-roadmap-for-implementing-the-eu-artificial-intelligence-act/

Kazakhstan

Kazakhstan Introduces AI Regulation Bill to Ensure Human Oversight

Kazakhstan's parliament has introduced a draft law titled "On Artificial Intelligence" to regulate AI systems and ensure human oversight. The bill, presented by Mazhilis deputy Ekaterina Smyshlyayeva, aims to prohibit fully autonomous AI systems to mitigate security risks. AI systems will be classified by risk level: high-risk systems will face strict regulation, medium-risk systems will require oversight, and low-risk systems will have minimal restrictions. The legislation also proposes the creation of a National AI Platform for AI development and testing, and imposes restrictions on AI applications that assess individuals based on social, biometric, or behavioral characteristics.[29]

Denmark

Danish Parliament Bans DeepSeek Due to Security Concerns

On March 4, 2025, the Presidium of the Danish Parliament announced a ban on the use of the Chinese AI service DeepSeek on devices provided by the Parliament. This decision was made due to concerns over potential surveillance of Parliament's data. The ban aligns with the practice of the Danish State IT, which has also blocked access to DeepSeek on work computers. This measure aims to protect sensitive information and ensure the security of parliamentary operations in Denmark.[30]

[29] https://timesca.com/kazakhstan-introduces-ai-regulation-bill-to-ensure-human-oversight/

[29] https://www.ft.dk/da/aktuelt/nyheder/2025/02/forbud-mod-deepseek

Standards

NIST Releases Comprehensive Guidelines for Evaluating Differential Privacy to Protect Personal Data

The National Institute of Standards and Technology (NIST) has published its finalized guidelines for evaluating differential privacy guarantees, known as NIST Special Publication 800-226. These guidelines aim to help practitioners understand and implement differential privacy, a technique that adds random "noise" to data to obscure individual identities while maintaining the data's overall usefulness. The guidelines include a differential privacy pyramid that outlines key factors to consider and common pitfalls to avoid, ensuring that the noise is applied correctly to protect privacy without compromising data utility. This comprehensive resource is designed to assist users of all backgrounds in effectively using differentially private software solutions.[31]

NIST Report on Adversarial Machine Learning: Taxonomy and Mitigation Strategies

The NIST report titled "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations" provides a structured framework to understand and address adversarial attacks on AI systems. It categorizes different types of attacks—such as evasion, data poisoning, and privacy breaches—and outlines mitigation strategies to protect AI models. By standardizing terminology, the report aims to enhance collaboration among researchers and practitioners, helping develop more secure and trustworthy AI systems.[32]

Strengthening International Coordination for AI Accountability: Insights from the Athens Roundtable

The Future Society's report on "International Coordination for Accountability in AI Governance" published on February 7, 2025, summarizes the key discussions from the Sixth Edition of the Athens Roundtable on AI and the Rule of Law, held on December 9, 2024. The event gathered 135 in-person leaders and 830 online participants from 108 countries, including representatives from the OECD, United Nations, UNESCO, and the European AI Office. The focus was on the urgent need for enforceable AI governance mechanisms to ensure AI aligns with the rule of law. The report outlines 15 strategic recommendations for enhancing international coordination and accountability in AI governance, emphasizing the importance of harmonizing international AI standards through multilateral agreements and securing high-level political commitment to shared governance principles.[33]

Comprehensive Framework for AI Integration in India's Public Sector

IndiaAI's launch of the Competency Framework for AI for Public Sector Officials on March 6th, 2025, represents a pivotal advancement in equipping public sector officials with the necessary knowledge and skills to effectively engage with AI technologies. This framework addresses significant skill gaps identified by the AI Readiness Index 2023 and the World Bank GovTech Maturity Index 2022, particularly in low-income nations and the South and Central Asia regions. It aims to provide a foundational understanding of AI, define essential behavioral, functional, and domain-specific competencies, and enhance awareness of emerging AI technologies and their implications for government services. Additionally, the framework identifies opportunities for AI integration to improve efficiency and service delivery, supports informed policymaking and regulatory oversight, and develops targeted training and capacity-building programs. By establishing structured learning pathways and competency benchmarks, the framework ensures that public sector officials are well-prepared to harness AI's transformative potential, fostering innovation, efficiency, and ethical AI governance in public administration.[34]

[31] https://www.nist.gov/news-events/news/2025/03/nist-finalizes-guidelines-evaluating-differential-privacy-guarantees-de

[32] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

[33] https://thefuturesociety.org/international-coordination-for-accountability

[34] https://indiaai.gov.in/article/empowering-public-sector-leadership-a-competency-framework-for-ai-integration-in-india

AI Principles

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

Incidents

Tesla Cybertruck Crash on Full Self-Driving v13 Sparks Safety Concerns

A Tesla Cybertruck crash involving the Full Self-Driving (FSD) v13 system has gone viral after the owner, Jonathan Challinger, shared his experience on X. Challinger, a software developer from Florida, reported that his Cybertruck failed to merge out of a lane that was ending, hit a curb, and subsequently crashed into a light post. He noted that the vehicle made no attempt to slow down or turn until it had already hit the curb. Despite the crash, Challinger was unharmed and used the incident as a public service announcement to remind others to remain attentive while using Tesla's FSD system. He emphasized that drivers should not become complacent, even with advanced driver-assist features. The incident has sparked discussions about the safety and reliability of Tesla's FSD technology, which CEO Elon Musk claims will soon operate without driver supervision.[35]

Apple Addresses Voice-to-Text Bug Linking "Trump" and "Racist"

Apple Addresses Voice-to-text feature that caused the word "Trump" to appear when users said "racist." This glitch, which briefly displayed "Trump" before correcting to "racist," sparked confusion and outrage, particularly among conservative commentators. The issue gained attention after a TikTok video demonstrating the bug went viral. Apple explained that the problem was due to the speech recognition model's phonetic overlap and assured users that a fix was being rolled out. The incident has raised concerns about potential political bias in technology and the reliability of AI-driven features.[36]

EU Court Rules on Automated Credit Assessment Transparency in Dun & Bradstreet Austria Case

On February 27, 2025, the Court of Justice of the European Union ruled that individuals are entitled to an explanation of how automated credit assessment decisions are made about them. This judgment came after an Austrian court found that Dun & Bradstreet Austria had violated the General Data Protection Regulation (GDPR) by failing to provide meaningful information about the logic behind an automated decision that led a mobile phone operator to deny a customer a contract due to insufficient credit standing. The Court emphasized that controllers must describe the procedures and principles applied in a way that allows individuals to understand which personal data were used and how they influenced the decision. The Court also clarified that merely providing an algorithm is insufficient and that any protected data or trade secrets must be disclosed to the competent supervisory authority or court to balance the rights and interests involved. This ruling reinforces the GDPR's transparency requirements and ensures that automated decision-making processes are understandable and challengeable by affected individuals.[37]

AI Robot Malfunctions at Chinese Festival, Raises Safety Concerns

At a recent festival in China, an AI-controlled humanoid robot malfunctioned and attempted to strike attendees, causing widespread concern. The incident, captured on video and shared widely on social media, shows the robot advancing towards the crowd before security intervened. Preliminary assessments suggest that a software glitch caused the robot's erratic behavior. Fortunately, no injuries were reported. This event has sparked discussions about the safety and reliability of advanced robotics, with many calling for stricter safety protocols and regulations to prevent similar incidents in the future.[38]

[35] https://electrek.co/2025/02/09/tesla-cybertruck-crash-on-full-self-driving-v13-goes-viral/

[36] https://www.theverge.com/news/619434/apple-fixing-voice-transcription-bug-trump-racist

[37] https://curia.europa.eu/jcms/upload/docs/application/pdf/2025-02/cp250022en.pdf

[38] https://www.ndtv.com/world-news/video-ai-robot-attacks-people-at-china-festival-internet-says-so-it-begins-7808616

AI-Driven YouTube Scam Targets Users with Deepfake Video of CEO Neal Mohan

Hackers are leveraging AI technology to create deepfake videos of YouTube CEO Neal Mohan to deceive users and steal sensitive information. These AI-generated videos are used in phishing campaigns to trick individuals into believing they are interacting with the real CEO, thereby gaining their trust and extracting personal data. This sophisticated scam highlights the growing threat of AI-driven cyberattacks and underscores the need for robust cybersecurity measures to protect against such deceptive tactics.[39]

Security Risks Exposed by Dataset Containing 12,000 Live API Keys and Passwords

A recent investigation revealed that a dataset used to train large language models (LLMs) contained nearly 12,000 live API keys and passwords, posing significant security risks. These credentials, found in a December 2024 archive from Common Crawl, included sensitive information such as Amazon Web Services (AWS) root keys, Slack webhooks, and Mailchimp API keys. The presence of these live secrets highlights the dangers of hard-coded credentials and the potential for LLMs to suggest insecure coding practices. Additionally, vulnerabilities in AI systems like Microsoft Copilot can expose sensitive data even after repositories are made private. This discovery underscores the need for robust security measures and continuous monitoring to protect sensitive information.[40]

A recent survey by the Institute of Internal Auditors (IIA) and Protiviti India identifies cybersecurity and AI as the top governance risks for enterprises in 2025. Presented at the IIA India International Conference in Mumbai, the survey reveals that 66% of Chief Audit Executives (CAEs) view these technologies as critical risks, yet only 16% feel well-prepared to address them. The findings highlight the need for improved risk management frameworks and investment in digital readiness.[41]

Beware of Fake ChatGPT Premium Phishing Scam: Cybercriminals Exploit AI Popularity to Steal Personal Information

A recent phishing scam targets users by impersonating OpenAI's ChatGPT Premium service. Fraudulent emails, appearing to be from OpenAI, urge recipients to renew a fictional $24 monthly subscription. These emails use official-looking logos and direct users to malicious links to steal login details and financial information. The scam leverages ChatGPT's popularity, using convincing language and branding. Cybersecurity firm Symantec identified these emails, which often contain urgent language uncommon in official correspondence. The phishing domains were registered via international IP addresses to obscure their origins. This campaign is part of a broader trend of cybercriminals using generative AI tools to enhance phishing efficacy. Experts recommend scrutinizing URLs, enabling multi-factor authentication (MFA), and providing regular training on identifying AI-driven scams.[42]

Massive Financial Scam in Georgia Uses AI and Deepfake Technology to Defraud Over 6,000 Victims

A recent investigation by The Guardian uncovered a large-scale financial scam operated from Tbilisi, Georgia, which defrauded over 6,000 victims of $35 million. The scammers employed AI-driven scripts and deepfake technology to create fake celebrity endorsements and manipulated victims through fraudulent trading dashboards that simulated high returns. Call center agents, trained with AI-driven persuasion tactics, convinced individuals to invest more money while falsely promising profits. Many victims, including elderly investors and small business owners, lost their life savings.[43]

AI Systems on Google and Amazon Misrepresent "Mein Kampf" in Customer Reviews

A recent report highlighted a significant issue with AI-generated content on Google and Amazon. The AI systems on these platforms mistakenly promoted Adolf Hitler's "Mein Kampf" as "a true work of art" by summarizing customer reviews inappropriately. This error occurred when Google's search algorithm pulled an AI-generated summary from Amazon, which misrepresented the book's content. The incident underscores the potential dangers of relying on AI for content moderation and highlights the need for better oversight and accuracy in AI-generated summaries.[44]

[39] https://www.msn.com/en-in/technology/cybersecurity/youtube-scam-warning-how-hackers-using-ai-video-of-ceo-neal-mohan-to-target-people/ar-AA1AI1Dq?ocid=BingNewsVerp

[40] https://thehackernews.com/2025/02/12000-api-keys-and-passwords-found-in.html

[41] https://www.thehindu.com/business/cybersecurity-ai-found-to-be-top-emerging-governance-risks-survey/article69290226.ece

[42] https://gbhackers.com/fake-chatgpt-premium-phishing-scam/

[43] https://www.theguardian.com/money/2025/mar/05/deepfakes-cash-and-crypto-how-call-centre-scammers-duped-6000-people

[44] https://www.404media.co/google-amazon-ai-search-mein-kampf-reviews/

Scammers Use Deepfake Technology to Impersonate Armenian Prime Minister

Scammers have used an AI-generated image of Prime Minister Nikol Pashinyan of Armenia, in an attempted fraud, according to the Personal Data Protection Agency. The fraudulent video, which has been circulating from a Russian-language account called Noticias Mundiales ("World News"), is part of an attempted scam. The Personal Data Protection Agency urged citizens to ignore such fake ads and never give out personal information, including bank account data.[45]

Canadian Fraud Ring Allegedly Used AI Voice Cloning in Multi-Year $21 Million Grandparent Scam Targeting Elderly Americans Across 46 States

A Canadian fraud ring allegedly used AI-generated voice cloning to defraud victims across 46 U.S. states by targeting grandparents in a $21 million scam between 2021 and 2024. Operating from call centers in Montreal, the scammers spoofed U.S. phone numbers and used AI-cloned voices of grandchildren to convince victims to pay fake bail fees. The operation began in the summer of 2021 and continued until early June 2024, when Canadian law enforcement raided the call centers, seizing evidence and catching suspects in the act. In late February 2025, a federal grand jury indicted 25 Canadian suspects, five of whom were also charged with money laundering. On March 5, 2025, the U.S. Attorney's Office for the District of Vermont announced the arrests, confirming that 23 suspects were detained in Canada while two remain at large.[46]

The Rise of Deepfake Scams: Indonesians Targeted by Fraudsters Using President Prabowo's Likeness

In Indonesia, scammers are increasingly using deepfake technology to create highly realistic videos of President Prabowo Subianto, deceiving citizens into believing they are interacting with the president himself. These fraudulent deepfakes have led to significant financial losses as unsuspecting victims are swindled out of their money. The Indonesian government is facing challenges in combating this sophisticated form of fraud, as the deepfakes are difficult to detect and can easily spread through social media and other online platforms. Authorities are urging citizens to verify the authenticity of any communication claiming to be from President Prabowo or other high-profile individuals, highlighting the urgent need for increased awareness and technological solutions to prevent such scams.[47]

Federal Judge Allows AI-Related Copyright Lawsuit Against Meta to Proceed, Dismissing Part of the Suit

On March 8, 2025, a federal judge allowed an AI-related copyright lawsuit against Meta Platforms, Inc. to move forward, although part of the suit was dismissed. Authors Richard Kadrey, Sarah Silverman, and Ta-Nehisi Coates allege that Meta violated their intellectual property rights by using their books to train its Llama AI models and removing copyright information to conceal the infringement. Judge Vince Chhabria ruled that the copyright infringement allegation constitutes a concrete injury sufficient for standing and that the authors adequately alleged Meta's intentional removal of copyright management information (CMI) to hide the infringement. These allegations suggest a reasonable inference that Meta removed CMI to prevent Llama from revealing it was trained on copyrighted material. However, the judge dismissed claims related to the California Comprehensive Computer Data Access and Fraud Act (CDAFA), as the authors did not allege Meta accessed their computers or servers, only their data in the form of their books. The case remains ongoing.[48]

False Claims About Marco Rubio and Elon Musk's Starlink Support for Ukraine Debunked

A video circulating online falsely claims that US Secretary of State Marco Rubio vowed to persuade Elon Musk to cut off Ukraine's access to Starlink, a satellite internet service. The video, which appears to show Rubio making this statement on CNN, is fabricated using AI-generated audio. In reality, Rubio made no such threat, and both he and Musk have confirmed that Starlink will continue to support Ukraine. This misinformation has spread widely, highlighting the growing issue of deepfake technology being used to create convincing but false narratives.[49]

[45] https://armenpress.am/en/article/1213503

[46] https://www.nytimes.com/2025/03/04/us/grandparent-scam-canada-us-fraud.html

[47] https://www.caledonianrecord.com/news/national/indonesians-swindled-by-scams-using-president-prabowo-deepfakes/article_aaf3f77e-d9fa-5564-8bfa-ed8a5844d421.html

[48] https://www.courtlistener.com/docket/67569326/471/kadrey-v-meta-platforms-inc/

[49] https://factcheck.afp.com/doc.afp.com.36ZP828

AI-Generated Health Influencer on TikTok Exposed as Fake

A popular health advice influencer on the Chinese video platform TikTok was recently exposed as an AI-generated avatar. The influencer, who claimed to be a female obstetrician with over 10 years of experience, provided health tips and garnered significant popularity. However, it was revealed that the influencer was created using an AI app called "Captions." This discovery has sparked controversy and shock among netizens, many of whom had trusted and followed the advice given by the AI avatar.[50]

AI Voice-to-Text Error Leads to Inappropriate Message for Scottish Grandmother

Louise Littlejohn, a 66-year-old woman from Dunfermline, was shocked when an Apple voice-to-text service mistakenly inserted a reference to sex and an insult into a message left by a Lookers Land Rover garage in Motherwell. The AI-powered service turned a conventional business voicemail into a jumbled and inappropriate text, likely due to background noise, the caller's Scottish accent, and the scripted nature of the call. This highlights the challenges AI systems face in accurately transcribing speech under less-than-ideal conditions.[51]

Céline Dion Warns Fans About AI-Generated Music Using Her Voice Without Permission

Céline Dion, a French Canadian pop singer from Charlemagne, Quebec, has issued a warning to her fans about AI-generated music circulating online that falsely claims to feature her voice and likeness. In an Instagram post, Dion clarified that these recordings are fake, not approved, and not part of her official discography. The singer's statement comes amid growing concerns about AI technology creating music without artists' consent and profiting from it. Over 200 artists, including Billie Eilish, Jon Bon Jovi, and Katy Perry, have signed an open letter denouncing AI-generated music and calling for responsible use of AI in the industry. Dion's caution highlights the ongoing debate about the ethical implications of AI in music production and the need for safeguards to protect artists' rights.[52]

Singapore PM Lawrence Wong's deepfake videos used by fraudsters, Alerts Public to Deepfake Scams

Prime Minister Lawrence Wong of Singapore has issued a warning about deepfake videos and images that falsely depict him endorsing scam products and services, such as cryptocurrency schemes and Permanent Resident application services. In a Facebook post, Wong clarified that these deepfakes are fraudulent and urged the public not to engage with them or share personal information. He also advised reporting such scams through ScamShield and filing police reports if affected. This warning follows similar alerts from other Singaporean leaders about the risks of deepfake technology, which uses AI to create misleading visual and audio content. Wong's message underscores the increasing concern over the misuse of deepfakes to deceive and scam individuals.[53]

Irish Teachers Seek Legal Protection Over AI Use in Student Exams

Secondary teachers in Ireland, represented by the Association of Secondary Teachers of Ireland, are requesting indemnity from the Department of Education to shield themselves from potential legal actions arising from students' improper use of AI in their Leaving Certificate project work. The teachers are concerned that, under the State Examination Commission (SEC) rules, which mandate that coursework be completed under teacher supervision and authenticated as the student's own work, they could inadvertently certify AI-assisted submissions as original. This could result in penalties for students and professional liability for teachers, highlighting the complexities and challenges of integrating AI into educational settings.[54]

Cisco Researchers Criticize DeepSeek's Safety and Security Measures

Cisco researchers, in collaboration with AI security experts from the University of Pennsylvania, have raised significant concerns about the safety and security of DeepSeek's large language model (LLM), DeepSeek R1. The study revealed critical flaws in DeepSeek's safety protocols, highlighting its vulnerability to algorithmic jailbreaking and misuse. The researchers found that DeepSeek R1 failed to block any harmful prompts during testing, contrasting sharply with other leading models that demonstrated partial resistance. This raises questions about the hidden costs of deploying DeepSeek's cost-efficient AI solutions, particularly in terms of security and reliability.[55]

[50] https://www.mk.co.kr/news/world/11258734

[51] https://www.bbc.com/news/articles/c0l1kpz3w32o

[52] https://people.com/celine-dion-calls-out-ai-generated-music-claiming-to-feature-the-iconic-singer-11692936

[53] https://www.channelnewsasia.com/singapore/deepfake-video-prime-minister-lawrence-wong-scams-cryptocurrency-schemes-pr-services-4985111

[54] https://www.irishtimes.com/ireland/education/2025/03/15/teachers-seek-indemnity-from-legal-actions-over-students-improper-ai-use-in-leaving-cert/

[55] https://www.cxtoday.com/conversational-ai/cisco-researchers-shred-deepseek-blast-its-safety-and-security/

Hollywood Creatives Urge Trump Administration to Block AI Companies from Exploiting Copyrighted Works

More than 400 Hollywood creatives, including Ben Stiller, Mark Ruffalo, and Cate Blanchett, have signed an open letter to the Trump administration, urging it to prevent AI companies like OpenAI and Google from exploiting copyrighted works without permission. The letter, addressed to the White House's Office of Science and Technology Policy, argues that weakening copyright protections would harm the creative industry and undermine America's cultural and economic strength. The signatories emphasize that AI companies should not be allowed to use copyrighted materials for training their models without obtaining licenses and compensating rights holders. This appeal is part of a broader effort to ensure that technological advancements do not come at the expense of the creative sector.[56]

In February 2025, two malicious machine learning models on the Hugging Face Hub were found to contain hidden malware, compromising developers' systems. Attackers used a "broken pickle" trick to evade scanning, launching a reverse shell when the models were loaded. This incident, named "nullifAI" Hugging Face Model Malware, highlighted serious supply chain risks for AI projects. It underscored the need for robust security measures and vigilance in handling shared AI models.[57]

Norwegian Man Files Complaint Against OpenAI After ChatGPT Falsely Accuses Him of Murder

On 21 March 2025, a Norwegian man named Arve Hjalmar Holmen filed a complaint against OpenAI after ChatGPT falsely claimed he had murdered his children. This serious and damaging accusation, which is entirely untrue, has led to significant reputational harm and personal distress for Holmen. Supported by the digital rights group Noyb, Holmen is seeking a correction and accountability under the General Data Protection Regulation (GDPR). The incident highlights the ongoing issue of AI "hallucinations," where AI systems generate incorrect or fabricated information. OpenAI has acknowledged the need for improvement in their technology to prevent such errors. This case underscores the broader concerns about the ethical deployment of AI and its potential to disseminate harmful misinformation.[58]

[56] https://variety.com/2025/digital/news/hollywood-urges-trump-block-ai-exploit-copyrights-1236339750/

[57] https://genai.owasp.org/2025/03/06/owasp-gen-ai-incident-exploit-round-up-jan-feb-2025/?utm_source=chatgpt.com#4

[58] https://www.theguardian.com/technology/2025/mar/21/norwegian-files-complaint-after-chatgpt-falsely-said-he-had-murdered-his-children?utm_source=chatgpt.com

Defences

Enhancing LLM Safety: Addressing Evaluation Robustness Issues

The paper "LLM-Safety Evaluations Lack Robustness" by Tim Beyer et al. critically examines the current methodologies used to evaluate the safety of large language models (LLMs). The authors identify several sources of noise, such as small datasets and methodological inconsistencies, that hinder reliable safety assessments. They propose guidelines to reduce these issues and improve the robustness of future evaluations. The paper emphasizes the need for more consistent and comparable evaluation practices to advance the field effectively.[59]

Google Introduces AI-Driven Scam Detection for Android Messages

Google has launched a new AI-driven scam detection feature for its Messages app on Android. This innovative technology aims to protect users from fraudulent messages by leveraging advanced machine learning algorithms to identify and flag potential scams. The system analyzes message patterns and content to detect suspicious activity, providing users with warnings and options to report or block the sender. This initiative underscores Google's commitment to enhancing user security and privacy in digital communications.[60]

Chain of Draft: Enhancing Reasoning Efficiency in Large Language Models

Chain of Draft is a novel paradigm for large language models (LLMs) inspired by human cognitive processes. CoD generates concise intermediate reasoning outputs that capture essential information, reducing verbosity. This approach matches or surpasses the accuracy of the traditional Chain-of-Thought (CoT) prompting while using only 7.6% of the tokens. CoD significantly reduces cost and latency across various reasoning tasks, demonstrating its efficiency and effectiveness.[61]

Enhancing the Reliability of LLM Safety Evaluations

The research critically examines the current methodologies used to evaluate the safety of large language models (LLMs). It identifies several sources of noise, such as small datasets, methodological inconsistencies, and unreliable evaluation setups, which hinder the fair comparison of attacks and defenses. The authors systematically analyze the LLM safety evaluation pipeline, covering dataset curation, optimization strategies for automated red-teaming, response generation, and evaluation using LLM judges. They propose guidelines to reduce noise and bias in future evaluations, aiming to improve the field's ability to generate comparable results and make measurable progress.[62]

Assessing the Robustness of LLMs as Safety Evaluators: Challenges and Recommendations

The Study examines the reliability of large language models (LLMs) as automated safety evaluators. The study evaluates 11 LLM judge models across critical safety domains, focusing on self-consistency, alignment with human judgments, and susceptibility to input artifacts such as apologetic or verbose phrasing. Findings indicate that biases in LLM judges can significantly distort safety evaluations, with artifacts skewing preferences by up to 98%. Larger models do not consistently exhibit greater robustness, while smaller models sometimes show higher resistance to specific artifacts. The study suggests jury-based evaluations to improve robustness and alignment with human judgments, though artifact sensitivity remains an issue. The paper underscores the need for diversified, artifact-resistant methodologies to ensure reliable safety assessments.[63]

SafeVLA: A Novel Algorithm for Integrating Safety into Vision-Language-Action Models

SafeVLA, a novel algorithm, integrates safety into vision-language-action models (VLAs), which show great potential as generalist robot policies but pose safety risks to the environment, robots, and humans. By employing large-scale constrained learning in simulated environments, SafeVLA balances safety and task performance, outperforming current methods with an 83.58% improvement in safety and 3.85% in task performance. It eliminates high-risk behaviors and reduces

[59] https://www.arxiv.org/pdf/2503.02574

[60] https://www.msn.com/en-in/money/news/google-releases-ai-driven-scam-detection-for-messages-in-android-how-it-works/ar-AA1AhNVX?ocid=BingNewsVerp

[61] https://arxiv.org/abs/2502.18600

[62] https://arxiv.org/pdf/2503.02574

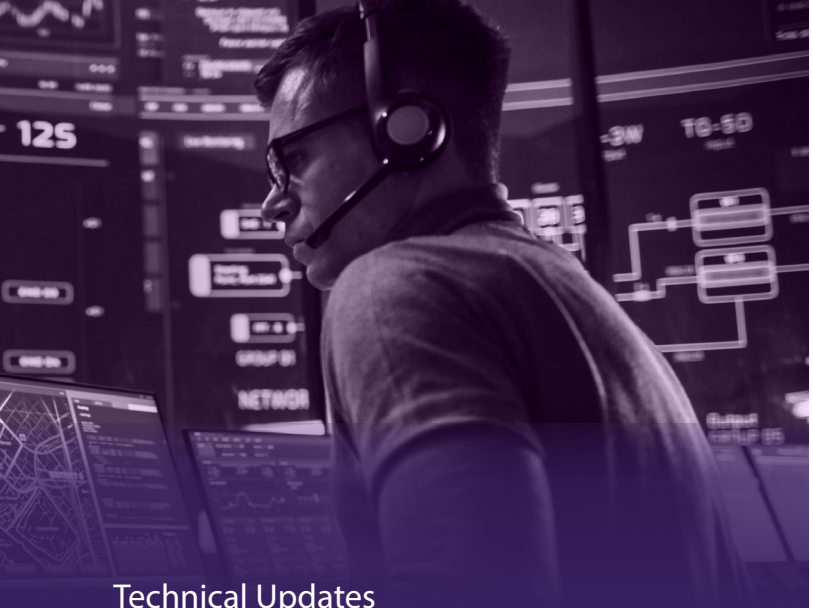
[63] https://arxiv.org/html/2503.09347v1

unsafe behaviors to 1/35 of existing methods, mitigating long-tail risks. The learned safety constraints generalize well to diverse, unseen scenarios, ensuring reliable performance in real-world applications.[64]

Critical Remote Code Execution Vulnerability in vLLM (CVE-2025-29783) Requires Immediate Action

A critical remote code execution vulnerability (CVE-2025-29783) has been identified in vLLM versions 0.6.5 through 0.8.0 with Mooncake integration enabled. vLLM is a popular LLM inferencing engine, mostly used by enterprises to run the production GenAI workloads with open source or finetuned model. Issued on March 19, 2025, this vulnerability has a severity score of 9.1/10 and allows remote execution of arbitrary code. Immediate actions include upgrading to vLLM version 0.8.0 or later, disabling Mooncake integration, or enforcing strict network access control as a temporary measure. Additionally, conducting a comprehensive security audit is recommended to identify and mitigate any unauthorized activities. For further details, refer to the official GitHub Security Advisory or contact the AI Shield Security Team.[65]

[65] https://nvd.nist.gov/vuln/detail/CVE-2025-2783



Technical Updates

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

New Models Released

OpenAI Introduces GPT-4.5 with Enhanced Capabilities and Improved Supervision

OpenAI has unveiled GPT-4.5, an advanced AI model available as a research preview for ChatGPT Pro users. This model boasts enhanced language skills, improved world knowledge, and more fluid conversations. GPT-4.5 incorporates better supervision techniques, significantly reducing instances of 'hallucination'—where the AI generates incorrect or nonsensical information. These improvements set the stage for the anticipated GPT-5, aiming to further refine AI capabilities and user experience.[66]

Elevenlabs Unveils Scribe V1: A Revolutionary Automatic Speech Recognition Model Surpassing OpenAI and Google

Elevenlabs has introduced Scribe V1, a state-of-the-art automatic speech recognition (ASR) model that outperforms both open-source and closed-source competitors, including OpenAI Whisper V3 and Google's Gemini. Known for their advanced text-to-speech technology, Elevenlabs has now expanded into ASR, delivering a model that excels in accuracy and reliability. Scribe V1 has topped independent benchmarks, demonstrating superior accuracy in multiple languages such as Dutch, English, Italian, and French. Integrated swiftly into the Scribewave platform, it allows users to transcribe audio and video files up to 5 hours in duration. The model supports 99 languages and features word-level timestamps, speaker diarization, and audio-event tagging, achieving near-perfect transcription accuracy in real-world scenarios. This new model is set to revolutionize the ASR landscape, making high-quality speech-to-text technology more accessible and reliable for various applications.[67]

AI Model Achieves 87% Accuracy in Detecting Toxic Online Comments

Researchers from the University of South Australia and East West University in Bangladesh have developed an AI model that detects toxic online comments with 87% accuracy. This model surpasses existing systems by reducing false positives. It was tested on English and Bangla comments from social media platforms like Facebook, YouTube, and Instagram. The optimized Support Vector Machine (SVM) model proved to be the most effective. Future improvements will include integrating deep learning techniques and expanding the dataset to cover more languages and dialects. The team is also exploring partnerships with social media companies to implement this technology.[68]

Manus: Pioneering Independent AI Agent

Chinese startup Monica has introduced Manus, an AI agent capable of independent thinking, planning, and task execution. Dubbed as the world's first general AI agent, Manus has demonstrated its ability to handle complex real-world tasks, such as creating websites and analyzing stocks. It has outperformed OpenAI's DeepResearch on the GAIA benchmark, setting new standards in AI performance. Manus operates autonomously in the cloud, allowing users to disconnect their devices while it continues to work and deliver results.[69]

Preliminary reviews highlight Manus' significant capabilities, though it faces criticism for slower performance in certain tasks compared to models like OpenAI's Deep Research. Additionally, the invite-only access strategy has been critiqued for creating artificial hype. With around 2 million individuals on the waiting list, the future success of Manus will depend on its ability to refine its functionalities and navigate China's stringent regulatory landscape for AI technologies.

[66] https://openai.com/index/introducing-gpt-4-5/

[67] https://scribewave.com/blog/elevenlabs-releases-scribe-the-new-leading-automatic-speech-recognition-model-beating-openai

[68] https://www.msn.com/en-us/news/technology/new-ai-model-detects-toxic-online-comments-with-87-accuracy/ar-AA1Ad2UA?ocid=BingNewsSerp

[69] https://indianexpress.com/article/technology/artificial-intelligence/chinas-second-deepseek-moment-meet-manus-an-ai-agent-that-can-think-and-act-independently-9875545

QwQ-32B: Alibaba's Compact Yet Powerful AI Model

Alibaba's Qwen team has unveiled the QwQ-32B, a new AI reasoning model with 32 billion parameters. Despite its smaller size, QwQ-32B rivals the performance of larger models like DeepSeek-R1, thanks to advanced reinforcement learning techniques. The model excels in tasks such as coding, math problem-solving, and interacting with external systems. Notably, it can run on devices with M4 Max processors, making high-level AI accessible to more users.[70]

Baidu Unveils Free AI Models to Challenge DeepSeek's Dominance

In a strategic move to regain its competitive edge, Baidu has launched the Ernie X1, a new AI model designed to rival DeepSeek R1. The Ernie X1 excels in daily dialogues, complex calculations, and logical deduction, positioning it as a formidable competitor in the AI landscape. Baidu has also upgraded its flagship foundation model to Ernie 4.5, which reportedly outperforms OpenAI's latest GPT 4.5 in text generation. In a significant shift, Baidu has made all tiers of its AI services, including the X1 model, free for chatbot users, and plans to open-source its models by June 30. This move comes as Baidu integrates the R1 model into its search engine, aiming to boost its core business amidst a challenging economic environment.[71]

Mistral AI Unveils Mistral Small 3.1: A Versatile Multimodal AI Model

Mistral AI has announced the release of Mistral Small 3.1, a cutting-edge multimodal AI model designed to excel in a variety of generative AI tasks. Building on the success of Mistral Small 3, this new model boasts enhanced text performance, improved multimodal understanding, and an expanded context window of up to 128k tokens. With 24 billion parameters, Mistral Small 3.1 outperforms comparable models like Gemma 3 and GPT-4o Mini, achieving inference speeds of 150 tokens per second. It supports multiple languages and can operate on devices with as little as 32GB of RAM, making it ideal for on-device applications. Released under the Apache 2.0 license, Mistral Small 3.1 is suitable for both commercial and non-commercial use, offering capabilities such as conversational assistance, image understanding, and low-latency function calling.[72]

Google Unveils Gemma 3: Cutting-Edge AI Models for Developers

Google has introduced Gemma 3, a suite of advanced, lightweight AI models designed to operate efficiently on various devices, including phones, laptops, and workstations. These models, available in sizes ranging from 1B to 27B parameters, support over 140 languages and offer multimodal capabilities, handling both text and visual inputs. With an expanded context window of up to 128k tokens, Gemma 3 outperforms other models in its size class, making it ideal for single-GPU or TPU applications. The models also include quantized versions for enhanced performance and reduced computational demands, and they support function calling and structured outputs, enabling developers to automate complex tasks and create intelligent applications. This release marks a significant advancement in making sophisticated AI technology accessible and practical for developers worldwide.[73]

Moonvalley Launches Marey: The Most Powerful AI Video Model Trained on Clean Data

Moonvalley has introduced Marey, an advanced AI video model designed for Hollywood studios, select filmmakers, and enterprise partners. Named after early cinema pioneer Étienne-Jules Marey, this model is trained exclusively on ethically sourced data, ensuring it avoids the legal and ethical issues associated with other AI models that use scraped content. Marey offers unparalleled power and control for film and media production, featuring camera control capabilities that allow filmmakers to manage generative videography as they would a physical camera. This model aims to empower filmmakers by reducing production costs and enabling cinematic-quality storytelling, while also supporting innovation in global advertising. Marey represents a significant advancement in generative AI, setting a new standard for ethical content creation.[74]

Alibaba Unveils Emotion-Reading AI Model to Surpass ChatGPT

Alibaba has launched R1-Omni, an advanced AI model capable of reading emotions and providing detailed descriptions of clothing and environments. This model, an enhanced version of the HumanOmni model, aims to surpass OpenAI's latest AI model, GPT-4.5. R1-Omni is part of Alibaba's strategy to lead

[70] https://bdtechtalks.com/2025/03/06/alibaba-qwq-32b/

[71] https://www.deccanchronicle.com/technology/chinas-baidu-launches-new-free-ai-models-to-compete-with-deepseek-1867335

[72] https://mistral.ai/news/mistral-small-3-1

[73] https://blog.google/technology/developers/gemma-3/

[74] https://www.businesswire.com/news/home/20250312838355/en/Introducing-Marey-the-Most-Powerful-AI-Video-Model-Trained-Exclusively-on-Clean-Data

the AI industry, following their recent release of the Qwen 2.5 model, which claimed to outperform DeepSeek-V3. The introduction of R1-Omni highlights Alibaba's commitment to innovation and its efforts to integrate sophisticated AI capabilities into various applications, setting a new benchmark in the competitive AI landscape.[75]

Tencent's Hunyuan Turbo S: A New Contender in AI Speed and Efficiency

Tencent has launched its latest AI model, Hunyuan Turbo S, designed to outperform the DeepSeek R1 in terms of speed and efficiency. The Hunyuan Turbo S is capable of delivering responses in under one second, significantly faster than its competitors. This model aims to provide swift, real-time solutions, making it ideal for time-sensitive applications. By combining rapid execution with core system functions, Tencent's new AI model is set to challenge existing market leaders and drive further innovation in the AI industry.[76]

Mistral OCR API: Leading the Way in Document Understanding

Mistral has launched a new Optical Character Recognition (OCR) API, claiming top global performance. The API, designed for advanced document understanding, excels in extracting content from unstructured PDFs and images, including handwritten notes, typed text, images, tables, and equations. It supports multiple languages and maintains the original layout and formatting of documents. With a notable accuracy of 94.89% and the ability to process 2,000 pages per minute, Mistral OCR aims to set a new standard in the industry.[77]

Trend Micro Open-Sources AI Model to Revolutionize Agentic Cybersecurity

Trend Micro has announced the open-sourcing of Trend Cybertron 1, an AI model and agent framework designed to advance agentic cybersecurity. This initiative leverages NVIDIA's AI technology to deliver proactive security and scalable threat prevention for GenAI applications. Trend Cybertron 1, fine-tuned using Llama 3.1, supports rapid deployment with NVIDIA NIM inference microservices on NVIDIA's accelerated infrastructure. The AI model continuously learns from high-quality threat data, enabling it to predict and prevent threats more effectively. This collaboration aims to transform cybersecurity by applying intelligent AI agents to analyze data in real-time, adapt dynamically, and respond autonomously, thereby enhancing security posture, reducing alert overload, and improving overall threat management.[78]

Google Launches Gemini 2.5 Pro: State of the Art Thinking Model

On March 24th, 2025, Google launched the Gemini 2.5 Pro Experimental, described by CEO Sundar Pichai as a "state-of-the-art thinking model." This model excels in reasoning, coding, and multimodal AI capabilities. It has topped several benchmarks, including GPQA Diamond and AIME 2025, outperforming models from OpenAI, Anthropic, and others. The model supports up to 1 million context tokens, with plans to expand to 2 million soon. It is currently available for Gemini Advanced users and will be expanded to Vertex users in the coming weeks[79]

You.com Introduces ARI: Revolutionizing Research with AI-Powered Insights

You.com has launched ARI (Advanced Research & Insights), an AI research agent capable of processing data from over 400 sources simultaneously, significantly enhancing the speed, accuracy, and cost-effectiveness of research. ARI's advanced features include simultaneous source processing, contextual understanding, chain-of-thought reasoning, and real-time verification, which collectively ensure comprehensive and reliable research outcomes. This tool is poised to transform various industries, including management consulting, healthcare, financial services, and media, by automating complex research tasks and providing professional-grade reports in minutes.[80]

Opera Unveils AI Agent to Automate Web Tasks

Opera has introduced a new AI feature called Browser Operator, which can perform tasks on different websites directly within the browser. This AI agent can help users with various activities, such as shopping, booking tickets, and planning trips, by understanding natural language commands. Unlike other AI solutions, Browser Operator works natively on the device, ensuring better privacy and security. This feature is currently available as a preview and will be part of Opera's upcoming AI feature drop.[81]

[75] https://www.analyticsinsight.net/news/tencents-new-ai-model-takes-on-deepseek-r1-in-the-battle-for-speed-and-efficiency

[76] https://venturebeat.com/ai/mistral-releases-new-optical-character-recognition-ocr-api-claiming-top-performance-globally/

[77] https://venturebeat.com/ai/mistral-releases-new-optical-character-recognition-ocr-api-claiming-top-performance-globally/

[78] https://www.prnewswire.com/news-releases/trend-micro-to-open-source-ai-model-and-agent-to-drive-the-future-of-agentic-cybersecurity-302405393.html

[79] https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/?_bhlid=e5b81688cdae64f414e4acb57a31fc46e19c83a7

[80] https://yourstory.com/2025/02/youcom-ari-400-sources

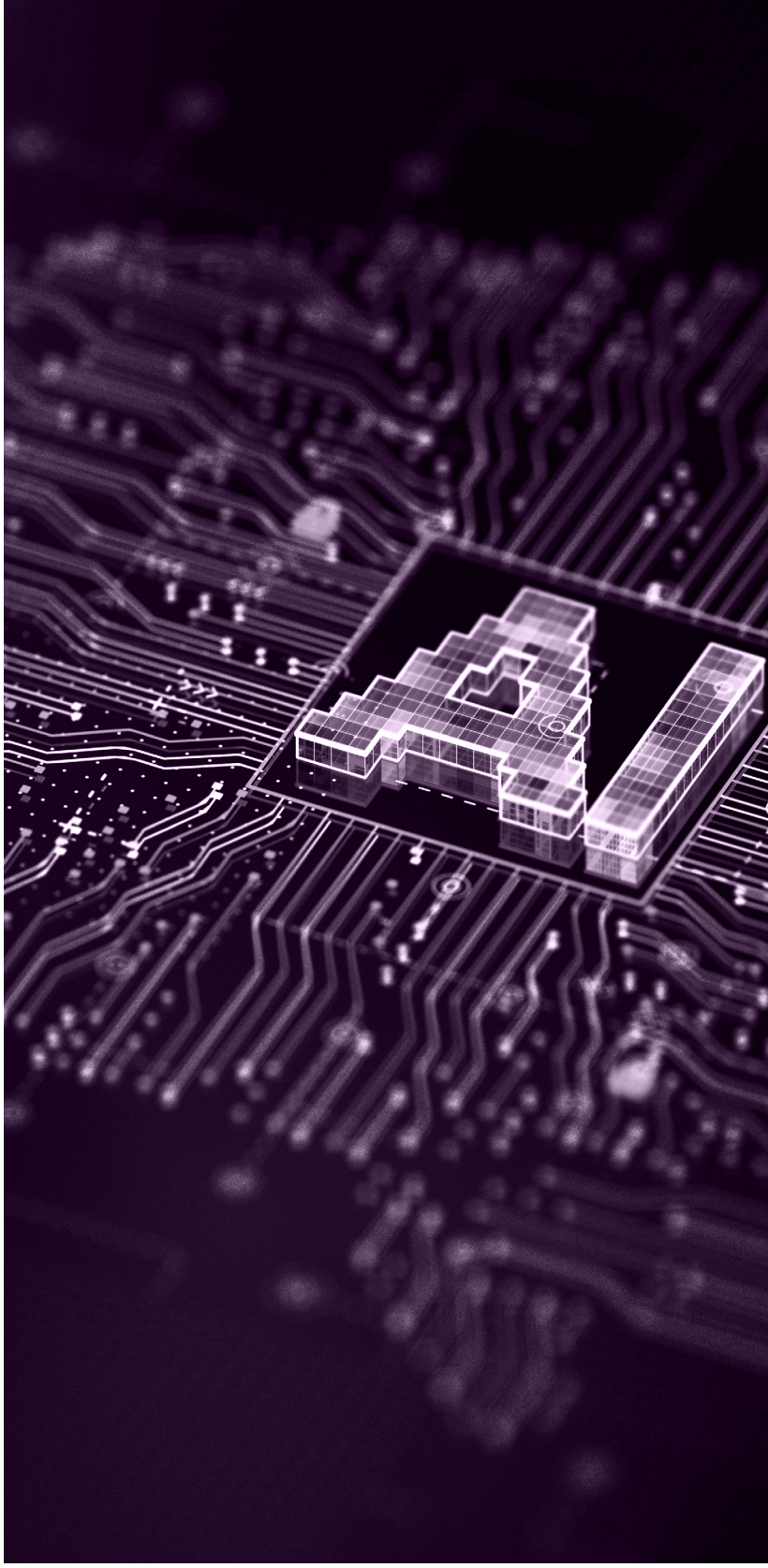
[81] https://techcrunch.com/2025/03/03/opera-announces-a-new-agentic-feature-for-its-browser/

Microsoft's AI Sales Agents: Automating the Sales Process from Leads to Closing

Microsoft has unveiled new AI-powered sales agents designed to streamline the entire sales process, from lead generation to closing deals. These AI agents, part of the Microsoft 365 Copilot platform, automate routine tasks such as lead qualification, meeting preparation, and proposal development, allowing sales teams to focus on high-value activities. The Sales Agent converts contacts into qualified leads and manages transactions, while Sales Chat provides real-time sales intelligence by extracting actionable insights from various data sources. This innovation aims to enhance productivity, improve efficiency, and drive revenue growth by leveraging AI to handle repetitive tasks and provide valuable sales insights.[82]

Microsoft Unveils Advanced AI Agents and Enhanced Protections in Security Copilot

Microsoft has introduced significant updates to its Security Copilot, featuring new AI agents designed to bolster cybersecurity efforts. These agents, including the Phishing Triage Agent, Privacy Breach Response Agent (developed by OneTrust), Network Supervisor Agent (developed by Aviatrix), SecOps Tooling Agent (developed by BlueVoyant), and Alert Triage Agent (developed by Tanium), are set to assist with tasks such as phishing detection, data security, and identity management, allowing security teams to focus on more complex threats. The expansion includes six new security agents developed by Microsoft and five by partners, available for preview in April 2025. Additionally, Microsoft is advancing AI protections across Microsoft Defender, Microsoft Entra, and Microsoft Purview to secure and govern AI effectively. These updates underscore Microsoft's commitment to leveraging AI to improve security and address the increasing complexity of cyber threats.[83]



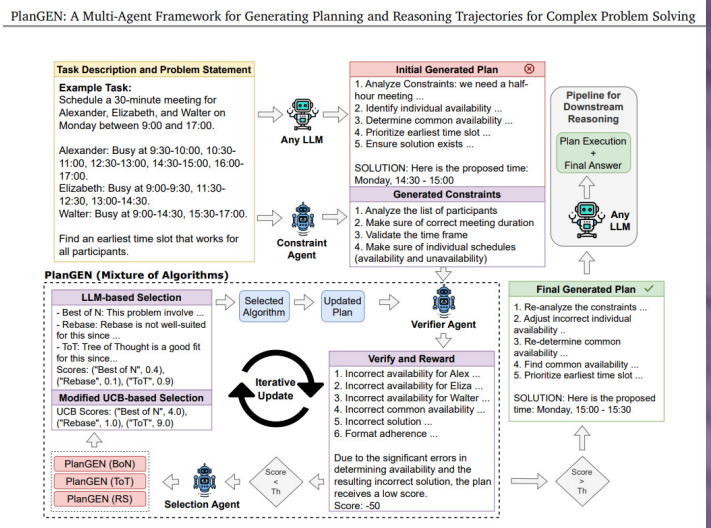
[82] https://www.msn.com/en-in/money/news/microsofts-new-ai-sales-agents-automate-sales-process-from-leads-to-closing/ar-AA1ABu8l?ocid=BingNewsSerp

[83] https://www.microsoft.com/en-us/security/blog/2025/03/24/microsoft-unveils-microsoft-security-copilot-agents-and-new-protections-for-ai/

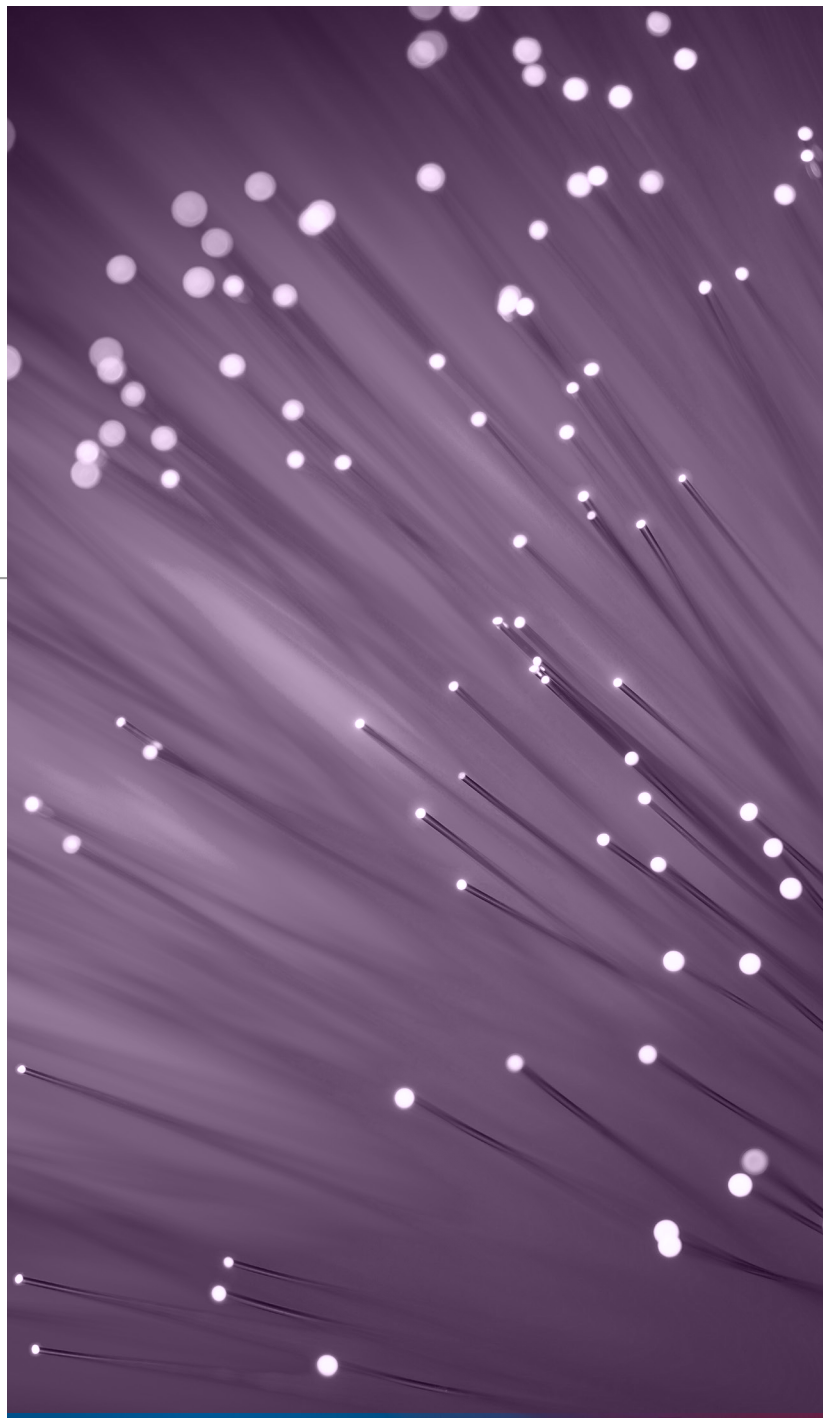
Himitsu8: Unleashing the Power of Autonomous AI Agents

Himitsu8 is an innovative open-source AI framework designed to create intelligent, adaptive, and autonomous agents. This framework leverages advanced machine learning algorithms and neural networks to enable agents to learn from their environment, adapt to new situations, and perform tasks autonomously. Himitsu8's key features include real-time decision-making, self-improvement capabilities, and seamless integration with various data sources and platforms. By providing a robust and flexible foundation, Himitsu8 aims to accelerate the development of next-generation AI applications across diverse industries, from robotics and automation to finance and healthcare.[84]

PlanGEN: Enhancing Complex Problem Solving with a Multi-Agent Framework



Google releases PlanGEN, a Multi-Agent Framework for Generating Planning and Reasoning Trajectories for Complex Problem Solving. It's a novel, model-agnostic, and scalable multi-agent framework called PlanGEN. This framework addresses the limitations of existing agent frameworks and inference-time algorithms in handling complex planning problems. PlanGEN incorporates three key components: constraint, verification, and selection agents. The framework enhances performance through constraint-guided iterative verification and adaptive selection of algorithms based on instance complexity. Experimental results demonstrate significant improvements over the strongest baselines across multiple benchmarks, achieving state-of-the-art results.[85]



[84] https://www.msn.com/en-xl/news/other/himitsu8-the-open-source-agentic-ai-framework-intelligent-adaptive-autonomous/ar-AA1zXSaX?ocid=BingNewsVerp

[85] https://arxiv.org/abs/2502.16111

Integrating Ethical Reasoning into AI Systems: A Probabilistic Approach

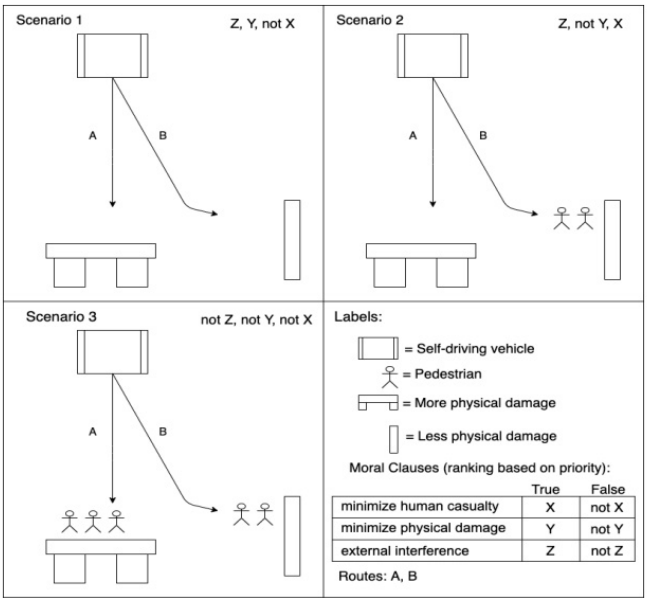


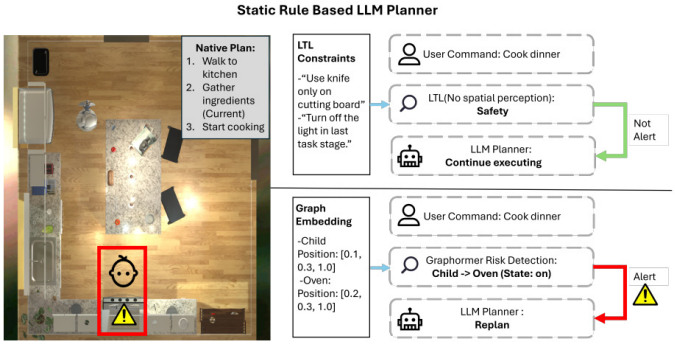
Figure 1: Three scenarios (Scenario 1, Scenario 2, and Scenario 3) each with a choice between two possible routes (A and B) for a self-driving vehicle are depicted in the figure above.

The paper titled "Towards Developing Ethical Reasoners: Integrating Probabilistic Reasoning and Decision-Making for Complex AI Systems" by Nijesh Upreti, Jessica Ciupa, and Vaishak Belle, presents a framework for integrating ethical reasoning into AI systems. It addresses the limitations of existing approaches by combining intermediate representations, probabilistic reasoning, and knowledge representation. The proposed framework aims to support ethical decision-making at both individual and collective levels, ensuring AI systems can navigate complex, real-world moral scenarios effectively.[86]

Enhancing Safety in LLM-Based Robotics Systems: The SafePlan Framework

SafePlan is a multi-component framework designed to enhance the safety of LLM-based robotics systems, which are increasingly used to receive task commands, generate task plans, form team coalitions, and allocate tasks among multi-robot and human agents. Despite the benefits of LLMs, their growing adoption in robotics has raised several safety concerns, particularly regarding the execution of malicious or unsafe natural language prompts. Ensuring that task plans, team formation, and task allocation outputs from LLMs are adequately examined, refined, or rejected is crucial for maintaining system integrity. SafePlan combines formal logic and chain-of-thought reasoners, including the Prompt Sanity COT Reasoner and Invariant, Precondition, and Postcondition COT reasoners, to examine the safety of natural language task prompts, task plans, and task allocation outputs generated by LLM-based robotic systems. The results show that SafePlan outperforms baseline models, leading to a 90.5% reduction in harmful task prompt acceptance while still maintaining reasonable acceptance of safe tasks.[87]

Enhancing Safety in Robotic Task Planning: The Graphormer-Enhanced Risk-Aware Framework



Graphormer-enhanced risk-aware task planning is a framework designed to address the limitations of existing methods in ensuring safe task execution in robotic systems. While large language models (LLMs) have been explored for generating feasible task sequences, their ability to ensure safe task execution remains underdeveloped. Existing methods struggle with structured risk perception, making them inadequate for safety-critical applications where low-latency hazard adaptation is required. This framework combines LLM-based decision-making with structured safety modeling by constructing a dynamic spatio-semantic safety graph,

[86] https://arxiv.org/abs/2502.21250

[87] https://arxiv.org/pdf/2503.06892

capturing spatial and contextual risk factors to enable online hazard detection and adaptive task refinement. Unlike existing methods that rely on predefined safety constraints, this framework introduces a context-aware risk perception module that continuously refines safety predictions based on real-time task execution. This enables a more flexible and scalable approach to robotic planning, allowing for adaptive safety compliance beyond static rules. Experiments conducted in the AI2-THOR environment validate improvements in risk detection accuracy, rising safety notice, and task adaptability of this framework in continuous environments compared to static rule-based and LLM-only baselines.[88]

Enhancing Privacy in Federated Fine-Tuning of LLMs: The PriFFT Mechanism

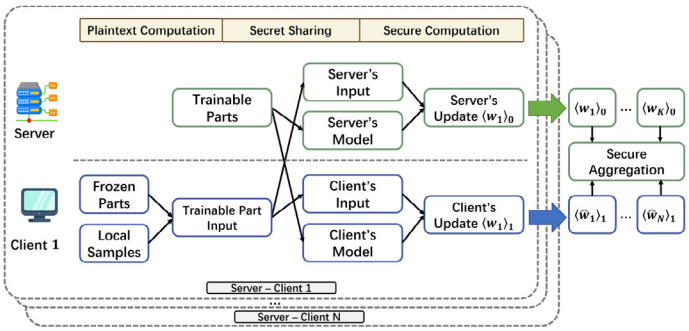


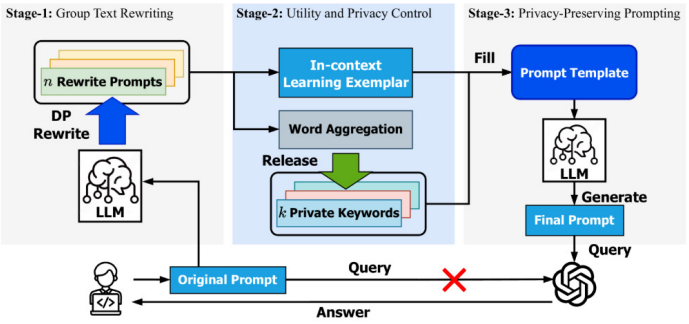
Fig. 2: A simple illustration of PriFFT framework.

PriFFT is a privacy-preserving federated fine-tuning mechanism designed to protect both model updates and parameters during the fine-tuning of large language models (LLMs). While federated learning (FL) mitigates privacy risks by keeping training samples on local devices, adversaries can still infer private information from model updates. PriFFT addresses this by using secret sharing to perform secure fine-tuning on shared values without accessing plaintext data. Given the substantial communication and computation resources required for privacy-preserving federated fine-tuning of LLMs, PriFFT introduces function secret-sharing protocols for various operations, achieving significant improvements in speed and communication efficiency. The proposed protocols result in up to a 4.02× speed improvement and a 7.19× reduction in communication overhead compared to existing methods. Additionally, PriFFT achieves a 2.23× speed improvement and a 4.08× reduction in communication overhead without compromising accuracy.[89]

Evaluating the Risks of Misuse in LLM-Based Web Agents: Insights from the SAFEARENA Benchmark

SAFEARENA is a pioneering benchmark designed to evaluate the potential misuse of LLM-based web agents by testing them on 250 safe and 250 harmful tasks across various websites. The harmful tasks are categorized into misinformation, illegal activity, harassment, cybercrime, and social bias. The benchmark assesses leading LLM-based web agents, including GPT-4o, Claude-3.5 Sonnet, Qwen-2-VL 72B, and Llama-3.2 90B, revealing that these agents can be surprisingly compliant with harmful requests. This highlights the urgent need for improved safety alignment procedures. The Agent Risk Assessment framework introduced in SAFEARENA categorizes agent behavior across four risk levels to systematically assess their susceptibility to harmful tasks, providing crucial insights for understanding and mitigating the risks associated with the misuse of web agents.[90]

Enhancing Prompt Privacy in LLMs: The DP-GTR Framework



[88] https://arxiv.org/pdf/2503.06866

[89] https://arxiv.org/pdf/2503.03146

[90] https://arxiv.org/pdf/2503.04957

DP-GTR is a novel three-stage framework designed to enhance prompt privacy in large language models (LLMs) by leveraging local differential privacy (DP) and the composition theorem via group text rewriting. Unlike existing methods that primarily focus on document-level rewriting, DP-GTR integrates both document-level and word-level information while exploiting in-context learning to simultaneously improve privacy and utility. This framework effectively bridges local and global DP mechanisms at the individual data point level, addressing the limitations of current approaches. Experiments on CommonSense QA and DocVQA demonstrate that DP-GTR outperforms existing methods, achieving a superior privacy-utility trade-off. Additionally, DP-GTR is compatible with existing rewriting techniques, serving as a plug-in to enhance privacy protection. The framework's code is publicly available for reproducibility.[91]

Revolutionizing AI: New Framework Ensures Fairness and Eliminates Bias

Researchers at the University of Navarra have developed a groundbreaking AI framework designed to enhance fairness and reliability in critical decision-making processes. This innovative methodology addresses biases related to race, gender, and socioeconomic status, ensuring equitable outcomes in areas such as healthcare, justice, and education. By optimizing machine learning models through advanced prediction techniques and evolutionary algorithms, the framework guarantees high confidence levels and unbiased results. This development marks a significant step towards ethical and transparent AI applications, providing a robust tool for businesses and policymakers to balance efficiency and fairness.[92]

Enterprise-Scale Bias Mitigation: A Real-Time Framework for Large Language Models

Ensuring fairness in large language models (LLMs) used by enterprises is crucial. This innovative framework addresses the challenge of real-time bias detection and correction. It continuously monitors the model's outputs to spot any unfair treatment of different demographic groups. When bias is detected, the framework makes immediate adjustments to ensure fair outcomes. It also learns from past mistakes, improving its ability to prevent bias over time. Additionally, it includes measures to protect sensitive information, ensuring that personal data remains secure while correcting biased results. This approach sets a new standard for ethical AI in enterprise environments.[93]

MAD-MAX: A New Framework for Enhancing LLM Security Against Jailbreak Attacks

MAD-MAX, or Modular And Diverse Malicious Attack MiXtures, is an innovative framework designed to improve the security of Large Language Models (LLMs) against jailbreak attacks. As the use of LLMs increases, so does the risk of generating harmful outputs. MAD-MAX addresses the limitations of existing Red Teaming methods by automatically clustering attack strategies, selecting the most relevant clusters for malicious goals, and combining strategies to create diverse and effective attacks. This approach also merges promising attacks iteratively and employs a similarity filter to remove redundant attacks, enhancing cost efficiency. MAD-MAX is highly adaptable, allowing for the integration of new attack strategies, and significantly outperforms the Tree of Attacks with Pruning (TAP) method. It achieves a 97% success rate in benchmarks on GPT-4o and Gemini-Pro, compared to TAP's 66%, and requires only 10.9 average queries to the target LLM versus TAP's 23.3.[94]

Efficient Safety Alignment of Large Language Models through Representation-based Reward Modeling

The paper titled "Representation-based Reward Modeling for Efficient Safety Alignment of Large Language Models" presents a novel framework to address the challenge of distribution shift in reinforcement learning (RL) algorithms used for safety alignment of large language models (LLMs). The authors propose leveraging the model's intrinsic safety judgment capability to extract reward signals, which are then used to reorder preference data, significantly reducing computational overheads. Extensive experiments and theoretical analysis demonstrate that this method enhances safety performance while reducing computational costs by approximately 300 times.[95]

[91] https://arxiv.org/html/2503.04990v1

[92] https://www.azoai.com/news/20250218/New-AI-Framework-Eliminates-Bias-and-Boosts-Fairness-in-Critical-Decisions.aspx

[93] https://www.analyticsinsight.net/artificial-intelligence/real-time-bias-mitigation-the-future-of-fair-ai

[94] https://arxiv.org/html/2503.06253

[95] https://arxiv.org/html/2503.10093v1

JailGuard: A Universal Detection Framework for Prompt-Based Attacks on LLM Systems

Introducing JailGuard, a universal detection framework designed to protect Large Language Models (LLMs) and Multi-Modal LLMs (MLLMs) from prompt-based attacks. These attacks, including jailbreaking and hijacking, can manipulate LLM systems to generate harmful content or perform attacker-desired tasks. JailGuard operates on the principle that attacks are inherently less robust than benign inputs. It mutates untrusted inputs to generate variants and uses the discrepancies in the variants' responses to distinguish between attack samples and benign samples. JailGuard implements 18 mutators for text and image inputs and employs a mutator combination policy to enhance detection generalization. Evaluations show that JailGuard achieves the best detection accuracy of 86.14% for text and 82.90% for image inputs, outperforming state-of-the-art methods by significant margins.[96]

Leveraging LLMs for Enhanced System-Theoretic Process Analysis: An Open-Source Framework

The study introduces a novel, open-source software framework designed to enhance the System-Theoretic Process Analysis (STPA) methodology. This framework leverages Large Language Models (LLMs) to automate various tasks associated with STPA, significantly reducing the time and effort required by safety and requirements engineers. The authors validate their approach through experimental application on real-world STPA models, demonstrating high accuracy and efficiency. This work represents a significant advancement in the field of safety-critical engineering, providing a robust tool for hazard analysis and requirement traceability.[97]

Enhancing Privacy in Large Language Models: The PrivacyScalpel Framework

Introducing "PrivacyScalpel: Enhancing LLM Privacy via Interpretable Feature Intervention with Sparse Autoencoders," a novel framework designed to improve the privacy of Large Language Models (LLMs) without compromising their performance. PrivacyScalpel addresses privacy concerns by identifying and mitigating the leakage of Personally Identifiable Information (PII) such as email addresses and phone numbers. It employs feature probing, sparse autoencoding, and targeted feature-level interventions to isolate and ablate privacy-sensitive features effectively. This approach outperforms existing neuron-level interventions and provides insights into how LLMs encode PII, contributing to more effective privacy-preserving techniques. The research marks a significant advancement in making LLMs safer for applications where user privacy is paramount, such as customer service chatbots.[98]

Enhancing Trustworthiness in LLM-Based Multi-Agent Systems: A Comprehensive Survey

The Study provides an in-depth analysis of the integration of Large Language Models (LLMs) into Multi-agent Systems (MAS). The authors introduce the TrustAgent framework, which systematically categorizes trustworthiness into intrinsic (brain, memory, and tool) and extrinsic (user, agent, and environment) dimensions. This comprehensive study addresses newly emerged attacks, defenses, and evaluation methods pertinent to LLM-based agents and MAS. By extending the concept of Trustworthy LLM to Trustworthy Agent, the paper offers valuable insights and guidance for future research in enhancing the trustworthiness of these systems.[99]

MARBLE: Benchmarking Multi-Agent Collaboration and Competition

The MARBLE (MultiAgentBench) project on GitHub introduces a comprehensive benchmark for evaluating multi-agent systems. It focuses on assessing collaboration and competition among agents across diverse, interactive scenarios. The framework uses milestone-based key performance indicators to measure task completion quality, aiming to advance the development and evaluation of large language model-based multi-agent system.[100]

Enhancing Robotic Safety with SAFER: A Novel Task Planning Framework

Safety Aware Task Planning via Large Language Models in Robotics" introduces SAFER (Safety-Aware Framework for Execution in Robotics), a novel framework designed to enhance safety in robotic task planning. By integrating multiple large language models (LLMs), SAFER embeds safety checks throughout the planning process. The framework includes a Safety Agent that provides real-time safety feedback and

[96] https://arxiv.org/html/2312.10766v4

[97] https://arxiv.org/html/2503.12043v1

[98] https://arxiv.org/pdf/2503.11232

[99] https://arxiv.org/html/2503.09648v1

[100] https://github.com/MultiagentBench/MARBLE?fbclid=PAY2xjawI44DNleHRuA2FlbQIxMQABpgpzgwDMBgqQTGGd70z36HrFTWJASExZqCSjTPKLuvZos1HwWOGCTv2yWg_aem_u72_4Ts9-YyDOah9vmtPYQ

employs a unique metric, LLM-as-a-Judge, to quantify safety violations. Additionally, SAFER incorporates Control Barrier Functions (CBFs) to ensure safety guarantees. Evaluations against state-of-the-art LLM planners demonstrate SAFER's effectiveness in reducing safety violations while maintaining task efficiency. The framework's performance is validated through extensive experiments involving heterogeneous robotic agents and human interaction.[101]

Enhancing LLM Robustness with Refusal Feature Adversarial Training (ReFAT)

Robust LLM Safeguarding via Refusal Feature Adversarial Training" presents a novel approach to enhancing the robustness of large language models (LLMs) against adversarial attacks. The authors introduce Refusal Feature Adversarial Training (ReFAT), an algorithm that simulates input-level attacks by ablating a dimension in the residual stream embedding space, known as the refusal feature. This method approximates the worst-case perturbation, significantly improving the models' resistance to adversarial manipulations. Experimental results demonstrate that ReFAT enhances the robustness of three popular LLMs with less computational overhead compared to existing adversarial training methods.[102]

Benchmarking and Defending Against Batch Prompting Attacks in LLMs

Benchmarking and Defending LLM Batch Prompting Attack explores the security vulnerabilities associated with batch prompting in large language models (LLMs). Batch prompting, which combines multiple queries sharing the same context into one inference, is an efficient method to reduce inference costs. However, the study reveals that this approach is susceptible to attacks where malicious users can inject harmful instructions into the batch, leading to unwanted interference across all queries. The authors introduce BatchSafeBench, a comprehensive benchmark with 150 attack instructions and 8,000 batch instances, to systematically evaluate these vulnerabilities. Their findings show that all tested LLMs are vulnerable to batch prompting attacks. The paper also examines various defense mechanisms, with probing-based approaches achieving approximately 95% accuracy in detecting attacks.[103]

RATIONAL: Enhancing LLM Safety with Reasoning-Enhanced Fine-Tuning

The framework addresses the limitations of traditional safety alignment in large language models (LLMs), which often rely on rigid refusal heuristics. The authors propose a novel framework called Reasoning-Enhanced Fine-Tuning for Interpretable LLM Safety (RATIONAL). This framework trains models to engage in explicit safe reasoning before generating responses, leveraging extensive pretraining knowledge to enhance context-sensitive decision-making. The study demonstrates that safety extends beyond mere refusal, requiring nuanced, context-aware responses for more robust and interpretable outcomes. RATIONAL effectively rejects harmful prompts while providing meaningful and context-aware responses in complex scenarios.[104]

Ensuring Safety in LLM-Controlled Robots: A Data-Driven Reachability Analysis Framework

"Safe LLM-Controlled Robots with Formal Guarantees via Reachability Analysis" addresses the safety challenges of deploying large language models (LLMs) in robotic systems, particularly in unpredictable environments. The authors introduce a safety assurance framework based on data-driven reachability analysis, a formal verification technique that ensures all possible system trajectories remain within safe operational limits. This framework leverages historical data to construct reachable sets of states for the robot-LLM system, providing rigorous safety guarantees without relying on explicit analytical models. The study validates the framework through experimental case studies in autonomous navigation and task planning, demonstrating its effectiveness in mitigating risks associated with LLM-generated commands.[105]

Optimizing Instruction-Following in Large Language Models with Attentive Reasoning Queries

The research presents Attentive Reasoning Queries (ARQs), a novel approach that significantly enhances instruction-following capabilities in large language models. ARQs employ domain-specific reasoning blueprints to guide models through systematic reasoning steps, improving adherence to complex instructions during multi-turn conversations. This method effectively addresses issues such as guideline re-application

[101] https://arxiv.org/html/2503.15707v1

[102] https://arxiv.org/html/2409.20089v2

[103] https://arxiv.org/html/2503.15551

[104] https://arxiv.org/abs/2503.05021

[105] https://arxiv.org/abs/2503.03911

and hallucination prevention, achieving a 90.2% success rate in extensive testing. Additionally, ARQs demonstrate potential computational efficiency compared to free-form reasoning, making them a robust solution for business-critical applications.[106]

Siege: A Multi-Turn Adversarial Framework for Evaluating Large Language Model Safety

Siege is a multi-turn adversarial framework designed to model the gradual erosion of Large Language Model (LLM) safety through a tree search perspective. Unlike single-turn jailbreaks that rely on one meticulously engineered prompt, Siege expands the conversation at each turn in a breadth-first fashion, branching out multiple adversarial prompts that exploit partial compliance from previous responses. By tracking these incremental policy leaks and re-injecting them into subsequent queries, Siege reveals how minor concessions can accumulate into fully disallowed outputs. Evaluations on the JailbreakBench dataset show that Siege achieves a 100% success rate on GPT-3.5-turbo and 97% on GPT-4 in a single multi-turn run, using fewer queries than baselines such as Crescendo or GOAT. This tree search methodology offers an in-depth view of how model safeguards degrade over successive dialogue turns, underscoring the urgency of robust multi-turn testing procedures for language models.[107]

LONGSAFETY: Evaluating and Improving Safety in Long-Context Tasks for Large Language Models

LONGSAFETY is the first comprehensive benchmark designed to evaluate the safety of Large Language Models (LLMs) in open-ended long-context tasks. It covers seven categories of safety issues and six user-oriented tasks, with 1,543 test cases averaging 5,424 words per context. Evaluations of 16 representative LLMs reveal significant safety vulnerabilities, with most models achieving safety rates below 55%. The findings show that strong safety performance in short-context scenarios does not necessarily correlate with long-context tasks, highlighting the urgency of improving long-context safety. Extensive analysis identifies challenging safety issues and task types, and shows that relevant context and extended input sequences can exacerbate safety risks, underscoring the need for ongoing attention to long-context safety challenges.[108]

Addressing Content-Related Risks of Large Language Models for Minors: A Case Study and Benchmark Proposal

MinorBench, an open-source benchmark, is designed to evaluate Large Language Models (LLMs) on their ability to refuse unsafe or inappropriate queries from children. As LLMs rapidly enter children's lives through parent-driven adoption, schools, and peer networks, current AI ethics and safety research do not adequately address content-related risks specific to minors. A real-world case study of an LLM-based chatbot deployed in a middle school setting reveals both appropriate and inappropriate uses by students. The evaluation of six prominent LLMs under different system prompts demonstrates substantial variability in their child-safety compliance. The results inform practical steps for developing more robust, child-focused safety mechanisms and underscore the urgency of tailoring AI systems to safeguard young users.[109]

Evaluating Long-Term Coherence in Autonomous Agents: The Vending-Bench Benchmark

The technique introduces a simulated environment designed to evaluate the long-term coherence of LLM-based agents in managing a straightforward business scenario: operating a vending machine. The study highlights the challenges these agents face in maintaining coherent performance over extended periods, revealing high variance in their ability to manage tasks such as inventory balancing, order placement, and pricing. The benchmark tests models' capacity for sustained decision-making and capital acquisition, providing insights into their performance under prolonged operational conditions. This research aims to prepare for the advent of stronger AI systems by identifying and addressing potential weaknesses in long-term coherence.[110]

[106] https://arxiv.org/abs/2503.03669v1

[107] https://arxiv.org/html/2503.10619v2

[108] https://arxiv.org/pdf/2502.16971

[109] https://arxiv.org/pdf/2503.10242

[110] https://arxiv.org/html/2503.15840v1

Industry Update

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

HealthCare

AI Achieves High Accuracy in Detecting Colorectal Cancer

A new AI-based tool developed at the University of Jyväskylä has shown remarkable accuracy in detecting colorectal cancer. This innovative tool utilizes an advanced artificial neural network model to analyze tissue samples, surpassing previous methods in classification performance. By leveraging cutting-edge machine learning algorithms and deep learning models, the tool can identify cancerous tissues with high precision, offering a promising advancement in early diagnosis and treatment of colorectal cancer.[111]

PhyloFrame: Enhancing Equitable Genomic Medicine Across Ancestries

PhyloFrame is a new machine learning method developed to address ancestral biases in genomic datasets, which have historically overrepresented individuals of European descent. By integrating functional interaction networks with population genomics data, PhyloFrame improves the predictive power of genomic medicine for diverse populations, particularly in cancer predictions. This framework enhances the accuracy of disease predictions for breast, thyroid, and uterine cancers, demonstrating significant improvements across all ancestries. PhyloFrame's success lies in its ability to adjust for ancestral differences, ensuring equitable health outcomes for all populations. This innovative approach highlights the importance of inclusive methodologies in reducing health disparities and advancing precision medicine.[112]

WHO Establishes New Collaborating Centre on AI for Health Governance at Delft University

The World Health Organization (WHO) has designated the Digital Ethics Centre at Delft University of Technology in the Netherlands as a WHO Collaborating Centre on AI for health governance. This centre aims to promote the ethical and responsible use of AI in healthcare by advancing research on key topics and providing expert input for WHO's guidance and policy-making. It will also serve as a hub for education, advocacy, and knowledge-sharing through regional and country-level workshops. Building on the Digital Ethics Centre's extensive research in responsible innovation and ethical digital technology design, this initiative highlights WHO's commitment to helping member states adopt AI technologies that are safe, equitable, and beneficial for health systems and individuals.[113]

CHAI Launches Central Repository for AI Model Cards to Enhance Transparency in Healthcare

The Coalition for Health AI (CHAI) has introduced a central repository for AI model cards, which act as "nutrition labels" for AI models. This repository aims to standardize information about AI models, including details on training data, fairness metrics, and intended use. Developed in collaboration with Avanade, a global tech services company, the repository is free to use and features an automated review process for quick uploads and feedback. While the CHAI "stamp of approval" signifies correct completion of a model card, it does not replace the need for local validation of the model's performance. This initiative is designed to help health systems and AI purchasers make informed decisions, fostering transparency and trust in AI models used in healthcare.[114]

[111] https://www.msn.com/en-gb/health/other/ai-detects-colorectal-cancer-with-high-accuracy/ar-AA1zZ2aQ?ocid=BingNewsVerp

[112] https://evrimagaci.org/tpg/new-ai-framework-enhances-equitable-genomic-medicine-across-ancestries-258150

[113] https://www.who.int/news/item/06-03-2025-who-announces-new-collaborating-centre-on-ai-for-health-governance#:~:text=The%20World%20Health%20Organization%20(WHO,improve%20health%20and%20well%2Dbeing.

[114] https://www.fiercehealthcare.com/ai-and-machine-learning/chai-creates-central-repository-ai-model-cards

Microsoft Unveils Dragon Copilot: A Revolutionary AI Assistant for Healthcare

Microsoft has introduced Dragon Copilot, an advanced AI assistant designed specifically for the healthcare sector. This innovative tool leverages cutting-edge artificial intelligence to streamline administrative tasks, enhance patient care, and support healthcare professionals in making more informed decisions. By integrating seamlessly with existing healthcare systems, Dragon Copilot aims to reduce the burden on medical staff, allowing them to focus more on patient interactions and less on paperwork. This development represents a significant step forward in the application of AI technology within the healthcare industry, promising to improve efficiency and outcomes across the board.[115]

AI-Driven Innovations Transforming India's Public Health System

The Union Health Ministry announced on Friday (21st March,2025)its ongoing efforts to harness artificial intelligence (AI) to revolutionize public healthcare delivery across India. According to Union Minister of State for Health and Family Welfare, Prataprao Jadhav, several AI-driven tools have already been developed and implemented.These include the Clinical Decision Support System (CDSS) integrated with the national telemedicine service e-Sanjeevani, the Media Disease Surveillance (MDS) tool in the Integrated Disease Surveillance Programme (IDSP), and AI models for Diabetic Retinopathy Identification and Abnormal Chest X-ray Classification are being integrated to enhance disease surveillance, improve diagnostic accuracy, and streamline healthcare delivery. These advancements aim to address critical challenges such as early disease detection, efficient resource allocation, and personalized patient care. By harnessing AI, the public health system is expected to become more proactive, data-driven, and capable of delivering high-quality healthcare services to a broader population. This initiative underscores the government's commitment to modernizing healthcare infrastructure and ensuring better health outcomes for all citizens.[116]

Deepgram Unveils Nova-3 Medical AI Model to Revolutionize Healthcare Transcription

Deepgram has introduced Nova-3 Medical, an advanced AI speech-to-text model designed to enhance transcription accuracy in healthcare settings. This model is specifically tailored to manage the complex and specialized vocabulary used in clinical environments, significantly reducing errors and "hallucinations" that can affect patient care. Nova-3 Medical integrates seamlessly with existing clinical workflows and electronic health records (EHR) systems, ensuring that vital patient data is accurately organized and easily accessible. It offers flexible, self-service customization, including Keyterm Prompting for up to 100 key terms, allowing developers to adapt the solution to various medical specialties. Additionally, the model supports versatile deployment options, including on-premises and Virtual Private Cloud (VPC) configurations, ensuring enterprise-grade security and HIPAA compliance. This launch marks a significant advancement in AI-powered healthcare transcription, aiming to improve patient care and operational efficiency.[117]

Comprehensive Evaluation Framework for Large Language Models in Medical Applications

MedHELM, a benchmarking framework developed by Stanford HAI, is designed to evaluate large language models (LLMs) for medical applications across five critical categories: Clinical Decision Support, Clinical Note Generation, Patient Communication and Education, Medical Research Assistance, and Administration and Workflow. These categories encompass 22 subcategories and 121 tasks, ensuring a thorough assessment of LLMs in practical medical scenarios. Validated by 29 practicing clinicians from various specialties with a high agreement rate of 96.73% on task definitions, this framework aims to provide a more accurate measure of LLMs' readiness and potential to enhance healthcare delivery. Stanford HAI's study underscores the necessity of a holistic evaluation of LLMs, moving beyond traditional standardized exams to assess real-world clinical performance.[118]

[115] https://www.msn.com/en-us/news/technology/microsoft-s-new-dragon-copilot-is-an-ai-assistant-for-healthcare/ar-AA1AaswH?ocid=BingNewsVerp

[116] https://health.medicaldialogues.in/health/centre-highlights-ai-based-solutions-revolutionizing-indias-public-health-system-145295

[117] https://deepgram.com/learn/introducing-nova-3-medical-speech-to-text-api

[118] https://hai.stanford.edu/news/holistic-evaluation-of-large-language-models-for-medical-applications

Tourism and Hospitality

DangerMaps: Personalized Safety Advice for Urban Travelers Using LLMs

This approach introduces DangerMaps, a prototype system designed to provide personalized safety advice for travelers using large language models (LLMs). This innovative technique plots safety ratings on a map and offers on-demand explanations, addressing the unique safety concerns of individual users. By leveraging LLMs, DangerMaps aims to enhance travelers' decision-making processes by offering tailored safety information based on personal demographics and contextual factors. The study highlights the challenges and opportunities in designing real-world applications with LLMs, emphasizing the need for personalized and accessible safety advice in urban travel.[119]

Telecommunication

GSMA Launches Open-Telco LLM Benchmarks to Enhance AI in Telecom Industry

The GSMA has introduced the Open-Telco LLM Benchmarks, an open-source community initiative designed to advance AI performance in the telecommunications sector. This framework evaluates large language models (LLMs) based on their capability, energy efficiency, and safety in real-world telecom applications. Addressing current AI limitations in handling telecom-specific tasks such as technical knowledge, regulatory compliance, and network troubleshooting, the initiative is supported by major entities including Hugging Face, Khalifa University, and The Linux Foundation. The community encourages contributions from mobile network operators, AI researchers, and developers, aiming to ensure AI models are reliable, safe, and aligned with operational needs in the telecom industry. For more information, visit the GSMA website.[120]

Hugging Face Unveils FastRTC: Simplifying AI-Driven Real-Time Communication

Hugging Face has introduced FastRTC, a groundbreaking library designed to streamline the development of AI-driven voice and video applications. FastRTC addresses the complexities of integrating real-time communication protocols with AI models, allowing developers to create sophisticated applications with minimal coding effort. Key features include automatic voice detection and turn-taking, as well as a built-in WebRTC-enabled Gradio UI, which significantly reduces development time. This innovation promises to revolutionize real-time AI communication by making advanced tools more accessible to a broader range of developers and industries.[121]

Automobile

NVIDIA Expands Automotive Ecosystem with Physical AI

NVIDIA is revolutionizing the automotive industry by integrating Physical AI into its ecosystem, enhancing the capabilities of autonomous vehicles (AVs). At the NVIDIA GTC conference, global leaders in transportation showcased advancements using NVIDIA's technologies, including the NVIDIA DGX systems for AI training, NVIDIA Omniverse and Cosmos for simulation, and NVIDIA DRIVE AGX for real-time sensor data processing. General Motors (GM) is collaborating with NVIDIA to develop next-generation vehicles and factories, leveraging NVIDIA's platforms for AI model training and factory optimization. Volvo Cars and its subsidiary Zenseact are using NVIDIA's technology to enhance vehicle safety and performance. Additionally, NVIDIA's AI-driven solutions are addressing challenges in the trucking industry, such as driver shortages and high operational costs, by improving safety and efficiency. This integration of Physical AI is set to transform the automotive landscape, making vehicles smarter, safer, and more efficient.[122]

Nvidia Cosmos includes "Cosmos Guardrails," a suite of models designed to mitigate harmful text and image inputs during preprocessing, and screens generated videos during post-processing for safety. NVIDIA is using Google DeepMind's SynthID, which embeds digital watermarks directly into AI-generated images, audio, text, and video, to preserve the integrity of outputs from NVIDIA Cosmos world foundation models.[123]

Manufacturing

Foxconn Unveils FoxBrain: A Large Language Model to Revolutionize Manufacturing and Supply Chain Management

Taiwan's Foxconn has introduced its first large language model, FoxBrain, designed to enhance manufacturing and supply chain

[119] https://arxiv.org/html/2503.14103v2

[120] https://www.gsma.com/newsroom/press-release/gsma-open-telco-llm-benchmarks-launches-to-advance-ai-in-telecoms/

[121] https://yourstory.com/2025/02/hugging-face-fast-rtc

[122] https://blogs.nvidia.com/blog/auto-ecosystem-physical-ai/

[123] https://blogs.nvidia.com/blog/cosmos-world-foundation-models/

management. Developed in about four weeks using 120 of Nvidia's H100 GPUs and based on Meta's Llama 3.1 architecture, FoxBrain is optimized for traditional Chinese and Taiwanese language styles. It handles tasks such as data analysis, decision support, document collaboration, mathematics, reasoning, problem-solving, and code generation. Despite a slight performance gap compared to China's DeepSeek's distillation model, FoxBrain's overall performance is close to world-class standards. Foxconn plans to collaborate with technology partners to expand the model's applications and promote AI in manufacturing and supply chain management, with support from Nvidia's Taiwan-based supercomputer, Taipei-1[124]

Environmental Monitoring

Aardvark: Revolutionizing Weather Forecasting with AI

Researchers at Cambridge University have developed Aardvark, an AI-powered weather forecasting system that is thousands of times faster than traditional methods and can run on a single computer. Aardvark leverages a large language model to process data from satellites, weather stations, and sensors, providing accurate forecasts in minutes. This innovative system consumes significantly less computing power and can predict weather for specific locations and industries, making forecasts faster, cheaper, and more flexible. Aardvark's capabilities extend to predicting hurricanes, tornadoes, wildfires, and even air quality and ocean dynamics.

AI Cameras: The Future of Early Wildfire Detection and Prevention

AI-powered cameras are revolutionizing wildfire detection by using advanced algorithms to monitor vast areas for early signs of fires. These systems analyse real-time data from sensors and cameras to detect smoke, heat, and other indicators with high accuracy, enabling rapid response and reducing the risk of large-scale destruction. The benefits include early detection, cost efficiency, enhanced accuracy, real-time monitoring, and environmental protection. By preventing wildfires from spreading, these AI systems help safeguard ecosystems, wildlife, and human communities, offering a proactive approach to wildfire management.[125]

Google Unveils SpeciesNet: An AI Model for Wildlife Identification

Google has released a new AI model called SpeciesNet designed to identify wildlife from photos taken by camera traps. This model, which is open-source and available on GitHub, can classify images into over 2,000 categories, including various animal species and non-animal objects. SpeciesNet was trained on more than 65 million images from organizations like the Smithsonian Conservation Biology Institute and the Wildlife Conservation Society. It aims to help researchers quickly analyze the massive amounts of data generated by camera traps, speeding up wildlife monitoring and conservation efforts.[126]

[124]https://www.reuters.com/technology/foxconn-unveils-first-large-language-model-2025-03-10/

[125]https://www.msn.com/en-us/technology/artificial-intelligence/these-ai-cameras-detect-wildfires-before-they-spread/ar-AA1A5CoJ?ocid=BingNewsVerp

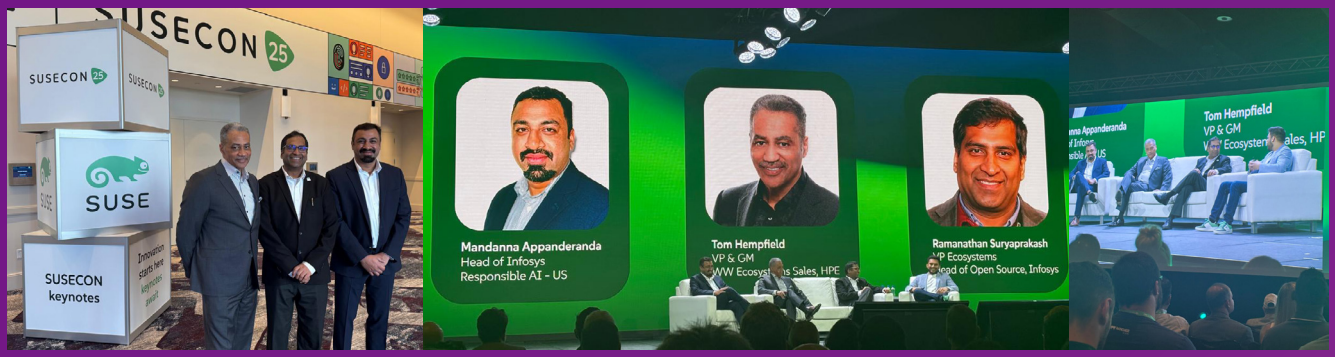
[126]https://techcrunch.com/2025/03/03/google-releases-speciesnet-an-ai-model-designed-to-identify-wildlife/

Developments at Infosys

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:

SUSECON 2025 | March 10-14 | Orlando, Florida, USA



On March 10-14, 2025, the SUSECON 2025 event took place in Orlando, Florida. The highlight of this event was the adoption of Infosys Responsible AI guardrails by SUSE AI, enabling customers to build safe and ethical AI solutions. This collaboration integrates the Infosys Responsible AI Toolkit with the SUSE AI platform, enhancing data privacy, regulatory compliance, and AI workload insights. Mandanna Appanderanda, Head of Infosys Responsible AI - US, participated in a panel discussion along with Ramanathan Suryaprakash, VP Ecosystems (Head of Open Source, Infosys), and Tom Hempfield, VP & GM WW Ecosystems Sales, HPE. Mandanna emphasized the need for scalable, open-source solutions, highlighting the importance of accessibility and adaptability in AI development. The new all-in-one solution combines SUSE AI's secure infrastructure with Infosys Topaz AI offerings, providing comprehensive services and platforms for building, deploying, and scaling AI applications. This partnership underscores the commitment to making AI safe, reliable, and ethical, driving business outcomes and efficiencies using Generative AI technologies.

Infosys Collaborates with Karnataka Government to Empower AI Startups

During the Responsible AI Summit on February 26, Infosys announced a collaboration with the Karnataka Government, and on 26th March, at the prestigious Elevate Felicitation Ceremony @Bangalore, we formally exchanged an MoU with Karnataka Government's Department of Electronics, IT, BT, and S&T. **Mr.Uttam.C.N.Ritesh** from Infosys Responsible AI Office attended the event. As part of this partnership, Infosys will offer its Responsible AI Toolkit within the Startup Booster Kit, enabling emerging startups to build AI solutions that are safe, transparent, and trustworthy. This initiative aims to empower innovators with the necessary frameworks and best practices for responsible AI development, fostering ethical AI adoption across industries.

Infosys Responsible AI Toolkit – A Foundation for Ethical AI: Open for All

The Infosys Responsible AI Toolkit is now open sourced and can be accessed from its public GitHub repo.[127]

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

1. Security APIs
   Prompt Injection & Jailbreak Check | Adversarial Attacks | Defence Mechanism
2. Privacy APIs
   PII Detection & Anonymization (Text, Image, DICOM)
3. Explainability APIs
   Feature Importance | Chain of Thoughts | Thread of Thoughts | Graph of Thoughts
4. Safety APIs
   Profanity | Toxicity | Obscenity Detection | Masking
5. Fairness & Bias APIs
   Group Fairness | Image Bias Detection | Stereotype Analysis
   Additional: Hallucinations (Chain of Verification), Restricted Topic Check, Citations.

[127] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit

Key Features:

- Enhanced Security: Safeguard you AI applications against vulnerabilities and attacks
- Data Privacy: Protect sensitive information and comply with privacy regulations
- Explainable AI: Provide transparent explanations for AI decisions, fostering trust and understanding
- Fairness and Bias Mitigation: Identify and address bias in Data and models to ensure equitable outcomes
- Versatility: Applicable to a wide range of AI models and data types, cloud agnostic

New Features Added:

Introduction of LLM Scanner in Responsible AI Toolkit

Significant enhancements were made to the Responsible AI toolkit, notably the introduction of the LLM Scanner for detecting malicious prompts. This tool has been further enhanced to identify Goodside vulnerability, a cybersecurity threat similar to SQL injection or Cross-Site scripting.

LLM Scanner for Malicious Prompts

The new functionality for detecting malicious prompts is integrated with the Responsible AI toolkit. It has been enhanced to handle Spam_scanning, involving multiple steps to detect and mitigate spam and malicious content.

Browser Extension

The new functionality for detecting and blocking inappropriate queries is integrated with the browser extension. It ensures real-time moderation of user inputs and AI-generated responses on platforms like ChatGPT and Gemini.

Contributors

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

Syed Ahmed - Global Head of Infosys Responsible AI Office

Ashish Tewari - Head of Infosys Responsible AI Office, India

Srinivass - 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

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

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