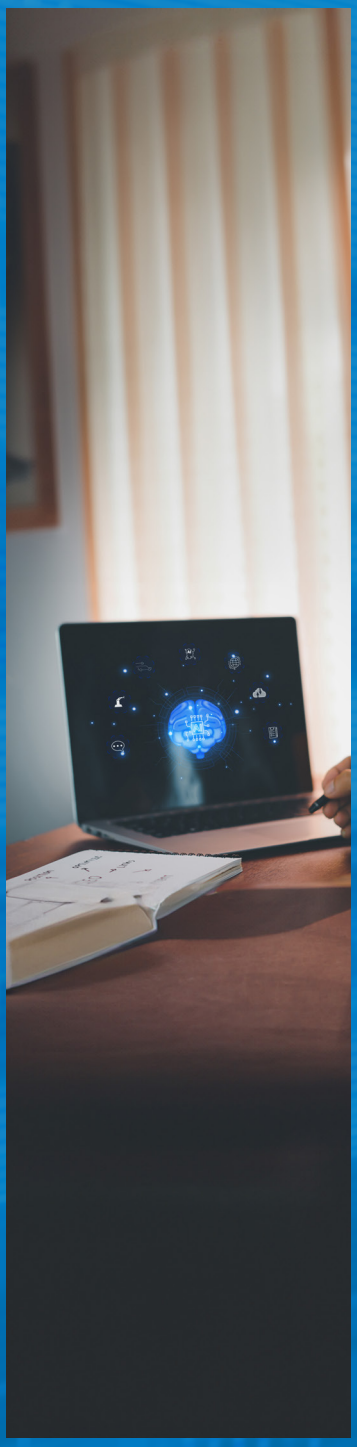
MARKET SCAN REPORT

BY INFOSYS RESPONSIBLE AI OFFICE

JULY - SEPTEMBER 2024



Infosys topaz

Infosys®
Navigate your next

Dear Readers,

As we navigate the ever-evolving landscape of artificial intelligence, staying informed about the latest regulations and technological advancements is more crucial than ever. The rapid pace of AI development presents not only immense opportunities but also significant challenges that require our attention. With new laws and ethical considerations emerging, understanding the framework within which AI operates is vital for responsible deployment and development.

This Market Scan Report aims to provide you with a comprehensive and up-to-date perspective on key aspects of AI regulation and innovation. Here's what you can expect to find within these pages:

1. **AI Laws and Regulations**
   We delve into the latest legislative changes impacting AI technologies. For instance, the EU's proposed AI Act is setting a precedent for how AI should be regulated, emphasizing accountability and transparency. Additionally, the California Consumer Privacy Act (CCPA) continues to influence data privacy discussions worldwide.
2. **AI Incidents and Defense**
   Analyzing recent AI incidents helps us learn from missteps. The infamous case of biased facial recognition technology highlights the need for responsible AI practices, while ongoing developments in AI-driven cybersecurity solutions demonstrate how we can defend against emerging threats.
3. **New Products and Solutions**
   The market is witnessing exciting innovations, such as generative AI tools that enhance creativity across industries. Solutions like AI-based health diagnostics are not only transforming patient care but are also subject to rigorous ethical scrutiny to ensure fairness and safety.
4. **Research Techniques**
   Advancements in AI research, such as reinforcement learning and explainable AI, are paving the way for more robust applications. For example, research on bias mitigation techniques is essential for creating fairer AI systems that serve diverse populations.
5. **Industry News on Privacy, Safety, and Security**
   The dialogue around AI must prioritize the protection of user data. Recent discussions at the AI Safety Conference showcased best practices in ensuring the security of AI applications, emphasizing the importance of privacy as a foundational element.
6. **Infosys Initiatives**
   As a leader in the industry, Infosys is committed to developing AI responsibly. Our initiatives focus on creating AI frameworks that adhere to the principles of privacy, safety, and fairness. For instance, our partnerships with regulatory bodies aim to shape AI policies that protect consumer rights while fostering innovation.

By emphasizing responsible AI tenets such as privacy, safety, fairness, and security, we can help ensure that AI technologies serve humanity positively. I encourage you to explore this report and leverage the insights to navigate the complexities of the AI landscape effectively.

Thank you for being part of this critical conversation.

Warm regards

Syed Ahmed
Global 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

Global

UN and OECD Announce Enhanced Collaboration on Global AI Governance

The OECD Deputy Secretary-General, Ulrik Vestergaard Knudsen, and the UN Secretary-General's Envoy on Technology, Under-Secretary-General Amandeep Singh Gill, have announced a new enhanced collaboration between the UN and the OECD on global AI governance. The two organizations will leverage their respective networks, convening platforms, and ongoing work on AI policy and governance to support their member states and other stakeholders in fostering a globally inclusive approach to AI. While further deliverables have yet to be unveiled, it has been announced that this collaboration will focus on regular science and evidence-based AI risk and opportunity assessments.[1]

During the 6th meeting of the GPAI Ministerial Council, held on July 3, 2024, in New Delhi, an announcement was made regarding an integrated partnership with the OECD. This partnership will bring together all current OECD members and GPAI countries on equal footing, under the GPAI brand, and based on the OECD Recommendation on Artificial Intelligence.

UN AI Advisory Body Proposes Seven Key Recommendations for AI Governance

An AI advisory body at the United Nations has released its final report proposing seven recommendations to address AI-related risks and gaps in governance. These recommendations include: (1) establishing an international scientific panel on AI; (2) initiating policy dialogue on AI governance; (3) creating an AI standards exchange; (4) developing a capacity development network; (5) setting up a global fund for AI; (6) forming a global AI data framework; and (7) establishing an AI office within the Secretariat. These recommendations will be discussed during a U.N. summit in September 2024.[2]

Global AI Convention Establishes Comprehensive Framework for AI Governance

The AI Convention, endorsed by the EU, the UK, the USA, and other nations including Israel and Norway, was created to establish a legal framework for the entire lifecycle of AI systems. This treaty aims to foster innovation and progress while managing associated risks. It mandates transparency, explainability, and interpretability of AI systems, especially in critical sectors like healthcare and finance. Flexibility in implementation is provided to accommodate the diverse legal systems of the signatories. The overarching goal is to ensure AI activities adhere to the treaty's provisions, promoting trustworthy and scalable AI development.[3]

[1] https://www.oecd.org/en/about/news/press-releases/2024/09/oecd-and-un-announce-next-steps-in-collaboration-on-artificial-intelligence.html

[2] https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_press_release.pdf

[3] https://www.privacyworld.blog/2024/09/ai-convention-a-global-framework-for-ai-principles/#:~:text=The%20first%20of%20its%20kind,of%20Moldova%20and%20San%20Marino.

AI Regulations and Governance across globe

US

New Legislations on Online Children' Privacy and Safety: COPPA 2.0 and KOSA Act

The U.S. Senate has passed both COPPA 2.0 and KOSA Act together, collectively which gives parents new tools to protect their kids online. While The Children and Teens' Online Privacy Protection Act (COPPA 2.0) bans online companies from collecting personal information from users under 17 years old without their consent, The Kids Online Safety Act (KOSA), provides children and parents with the tools, safeguards, and transparency to protect against online harms. It requires that social media platforms mandate annual audit and while suggesting contents for kids, by default activate the most protective settings for kids, provide minors with options to protect their information and disable addictive product features and opt-out of personalized algorithmic recommendations. This also prevents in using PII data in training the AI algorithms.[4]

Infosys Responsible AI Office promotes Responsible by Design principles in all AI initiatives.

California Bill SB 1047: Safe and Secure Innovation for Frontier AI Models Act

SB 1047 aims to establish guidelines for responsible AI development, addressing safety, accountability, and transparency concerns. Key provisions include shutdown capability, a written safety protocol, record keeping, risk assessment, training-only use, and annual audits. The bill aims to balance AI innovation with safety, build public trust, and minimize risks by requiring developers to implement safeguards and undergo regular audits.[5]

Four Legislative Bills Progressed in U.S.

In a pivotal step towards responsible artificial intelligence (AI) deployment and use in America, the Senate Commerce Committee has approved a package of four bills aimed at promoting responsible AI development and use in the United States. These bills address crucial areas such as AI standards, transparency, accessibility, and evaluation. This significant step towards AI governance is seen as essential for maintaining US leadership in AI while ensuring safety and accountability in the industry.

- The Future of AI Innovation Act (FAIIA)
- Artificial Intelligence Research, Innovation, and Accountability Act (AIRIA)
- Creating Resources for Every American to Experiment with Artificial Intelligence Act (CREATE AI Act)
- Validation and Evaluation for Trustworthy Artificial Intelligence (VET AI) Act[6]

House Committee Reports Progress on AI Integration

The Committee on House Administration has released its July 2024 Flash Report on Artificial Intelligence, detailing the use of AI technology by House offices and legislative branch agencies. The report highlights the application of the House's AI Guardrails in the acquisitions process, collaboration with legislative branch agencies, and the Smithsonian's AI-driven pilot projects. The AOC is working on establishing a Chief AI Officer, while the Capitol Police uses non-generative AI tools for threat detection. The CAO is exploring high-security LLM solutions like Microsoft Azure Open AI and AWS's Bedrock for data security.

The introduction of a new Chief AI Officer is a welcome step in forwarding the causes of responsible AI adoption.[7]

[4] https://www.commerce.senate.gov/2024/7/senate-overwhelmingly-passes-children-s-online-privacy-legislation

[5] https://digitaldemocracy.calmatters.org/bills/ca_202320240sb1047

[6] https://www.aipolicy.us/work/senate-commerce-committee-advances-landmark-package-of-bipartisan-legislation-promoting-responsible-ai

[7] https://cha.house.gov/press-releases?ContentRecord_id=D5694E3B-D6C3-40BE-8530-481403FE1B4D

USPTO Welcomes U.S. Copyright office on Digital Replicas

The US Copyright Office (USCO) has released a report on copyright-related legal and policy issues related to emergence of AI technology, specifically digital replicas. The report focuses on legal and policy issues in the use of digital technology to realistically replicate an individual's voice or appearance. The US Patents and Trademarks Office (USPTO) will consider the report's findings as they prepare recommendations for executive action to ensure the safe, secure, and trustworthy use of AI technologies.[8]

Infosys Responsible AI Office is actively contributing to various NIST's Working Groups to advance best practices and standards for Responsible AI adoption.

Advancing AI in Financial Services: A New Legislative Framework

The "Unleashing AI Innovation in Financial Services Act," introduced by Senators Mike Rounds and Martin Heinrich, along with Representatives French Hill and Ritchie Torres, aims to foster AI innovation within the financial services sector. This bipartisan legislation proposes the establishment of regulatory sandboxes, allowing financial firms to experiment with AI technologies under regulatory supervision without the risk of immediate enforcement actions. By creating a controlled environment for testing, the Act seeks to balance the advancement of AI-driven financial solutions with the necessary consumer protections, ultimately promoting economic growth and maintaining the U.S.'s leadership in global financial technology.[9]

International Considerations for AI in the Nuclear Sector

UK, US, and Canadian nuclear regulators have collaboratively outlined international principles for deploying artificial intelligence (AI) in the nuclear sector. The document, titled "Considerations for Developing Artificial Intelligence Systems in Nuclear Applications," emphasizes maintaining safety and security while integrating AI technologies in nuclear sector. Key areas of focus include managing AI lifecycle from design to deployment, the importance of human and organizational factors, and the integration of AI into existing nuclear systems. This initiative highlights the need for international cooperation to navigate AI regulation complexities and underscores the balance between human oversight and AI benefits.[10]

DEFIANCE Act of 2024: Combating Nonconsensual Deepfake Content

The DEFIANCE Act of 2024 aims to tackle the growing issue of nonconsensual, sexually explicit "deepfake" images and videos. This legislation holds individuals accountable for the creation and distribution of these digital forgeries, which depict people in these deepfake contents without their consent. By providing victims of digital forgery with a federal civil remedy, the act targets those who produce, distribute, or possess these forgeries with intent or reckless disregard. Additionally, the act establishes a 10-year statute of limitations, with provisions for tolling until the victim becomes aware of the deepfake or reaches the age of eighteen. The DEFIANCE Act of 2024 represents a significant step towards protecting individuals' privacy and dignity in the digital age.[11]

Bipartisan Bill Aims to Stop Unauthorized Content Use in AI

A new bipartisan bill named, "Content Origin Protection and Integrity from Edited and Deep faked Media Act (COPIED) Act," was introduced in July by US Senate Commerce Committee to address the deepfake issues. As per the bill, this would require platforms that develop or share AI systems to allow users to attach content provenance information to their work within two years. Content provenance is a machine-readable information, documenting the origin and history of a piece of digital content. The legislation would prevent the use of any work in the training process without the consent of the creator in case it has an attached content provenance. The bill has provisions for legal actions in case contents are used for training without approval. The bill would also direct the National Institute of Standards and Technology (NIST) to develop guidelines and standards for content provenance information, watermarking and synthetic content detection.[12]

Infosys is actively collaborating with C2PA (Coalition for Content Provenance and Authentication) for researching new tools to help facilitates content provenance and Deepfake detection.

The AI CONSENT Act: Empowering Consumers in Artificial Intelligence Development

A new proposed bill "Artificial Intelligence Consumer Opt-in, Notification, Standards, and Ethical Norms for Training Act," or AI CONSENT Act4 aims to put up a framework to protect consumers privacy in the era of AI. Federal Trade Commission (FTC) has been directed to enforce this. This act would require –

[8] https://www.uspto.gov/about-us/news-updates/uspto-welcomes-us-copyright-office-report-digital-replicas

[9] https://www.heinrich.senate.gov/newsroom/press-releases/heinrich-rounds-lead-bipartisan-bicameral-effort-for-ai-innovation-in-financial-services

[10] https://onr.org.uk/news/all-news/2024/09/new-paper-shares-international-principles-for-regulating-ai-in-the-nuclear-sector/

[11] https://www.durbin.senate.gov/imo/media/doc/defiance_act_of_2024.pdf?utm_campaign=The%20Batchandutm_source=hs_emailandutm_medium=email

[12] https://www.commerce.senate.gov/services/files/3012CB20-193B-4FC6-8476-DDE421F3DB7A

1. Businesses to get user agreement prior to using user data for AI system training.
2. Thorough and unambiguous notifications about how data is used and how consent is obtained.

The Act instructs the FTC to investigate the effectiveness of current data de-identification techniques.

As AI technology advances, this analysis would evaluate the possibility of improved protections against re-identification.[13]

> *When it comes to data privacy in AI, implementing techniques like differential privacy play a crucial role in ensuring the user privacy is protected.*

California Assembly Bill AB 2013: Promoting Transparency in Artificial Intelligence Training Data

California's Assembly Bill 2013 (AB 2013) presents a new framework for the administration of artificial intelligence (AI). The bill requires that on or before January 1, 2026, the developers of an AI system or service must post on their website regarding the training data used in AI systems and services and that be made publicly available. This synopsis should include information about the training datasets, including data types, sources, and planned AI system functionalities. Notably, the bill allows exemptions for AI essential to national aviation operations, national security, and national defence, protecting sensitive data in the process. Furthermore, AI systems released prior to 2025 are granted a temporary accommodation that allows developers to reveal the information that is easily accessible or to justify the time and effort they have taken to find it.

California makes a big step toward encouraging accountability and openness in the creation of AI systems by passing AB 2013. Users are better able to make educated decisions about how to interact with AI due to this enhanced transparency into training data, which also helps to boost public confidence in this quickly developing technology.[14]

UK

UK Government Invests £100m to Propel AI Research

UK government invests £100 million to boost nine new AI research hubs and accelerating the adoption of trusted and responsible AI. Some important points are that the hubs will focus on developing new AI technologies with applications in healthcare, electronics, and cyber security. This highlights the government's commitment to responsible AI development, with a focus on public trust and ethical considerations.[15]

Belgium

Belgium DPA Report on Aligning AI Systems with GDPR and AI Act

The Belgium Data Protection Agency (DPA) has published a report explaining the intersection between the GDPR and the AI Act, and how organizations can align AI systems with data protection principles. The report emphasizes transparency, accountability, and fairness in AI, particularly for high-risk AI systems. It also outlines how human oversight and technical measures can ensure compliant and ethical AI use.[16]

[13] https://www.congress.gov/bill/118th-congress/senate-bill/3975/text?s=9&r=1&q=%7B%22search%22%3A%22S+3975%22%7D

[14] hhttps://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013

[15] https://www.ukri.org/news/100m-boost-in-ai-research-will-propel-transformative-innovations/

[16] https://www.autoriteprotectiondonnees.be/publications/artificial-intelligence-systems-and-the-gdpr---a-data-protection-perspective.pdf

Europe

AI Act enters into force.

On 1 August 2024, the European Artificial Intelligence Act (AI Act) enters into force. The Act aims to foster responsible artificial intelligence development and deployment in the EU. The Act addresses potential risks to citizens' health, safety, and fundamental rights, providing clear requirements and obligations for AI developers and deployers. The Act introduces a uniform framework across EU countries, based on a risk-based approach. The EU aims to be the global leader in safe AI, benefiting everyone through better healthcare, safer transport, and improved public services.[17]

Infosys has pledged to the EU AI Act to develop a 'human-centric' approach to AI to ensure that Europeans can benefit from new technologies developed and functioning according to the EU's values and principles.

Australia

Australia Unveils New Policy for Responsible Use of AI in Government

The Digital Transformation Agency (DTA) has released the "Policy for the Responsible Use of AI in Government" which will come into effect on 1 September 2024, with the aim to position the country as a leader in safe and ethical AI implementation. The policy emphasizes transparency, risk management, public transparency statements, and evolving with technology and public expectations. It encourages continuous learning and improvement within government agencies to ensure responsible AI practices, demonstrating Australia's commitment to harnessing AI for positive societal impact.[18]

[17] https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en

[18] https://www.dta.gov.au/blogs/responsible-choices-new-policy-using-ai-australian-government

Australia Enacts New Law to Combat Deepfake Sexual Material

The Criminal Code Amendment (Deepfake Sexual Material) Bill 2024, recently passed by the Australian Parliament, aims to strengthen laws against the creation and non-consensual sharing of sexually explicit material online, including content altered using AI technology, commonly known as deepfakes. The bill amends the Criminal Code Act 1995 to address these issues, ensuring stricter penalties and clearer definitions to protect individuals from such harmful practices.[19]

Australia releases new policy document for Responsible use of AI in Government

Australia has released new policy document for Safe and Responsible use of AI in Government that came into effect on September 1, 2024. The key focus areas are to ensure the government leads in embracing AI for the benefit of Australians while ensuring its safe, ethical, and responsible use. It also aims to strengthen public trust by enhancing transparency, governance, and risk assurance. It has provisions to adapt AI over time. The policy applies to all Non-corporate Commonwealth entities and encourages voluntary adoption by others, with specific carveouts for national security.[20]

Voluntary AI Safety Standard for Australian Organization

The Voluntary AI Safety Standard provides practical guidance for Australian organizations to safely and responsibly use AI. It includes 10 voluntary guardrails that apply to all organizations throughout the AI supply chain covering transparency, accountability, and testing requirements across the AI supply chain. These guardrails help organizations mitigate risks while benefiting from AI innovations. The standard aligns with international laws, ensuring compliance and fostering safe AI practices.[21]

Mandatory Guardrails for AI in High-Risk Settings in Australia

The proposal paper outlines the introduction of mandatory guardrails for AI applications in high-risk settings through a new AI-specific Act. This Act will define high-risk AI applications and establish mandatory guardrails, including a monitoring and enforcement regime overseen by an independent AI regulator. The aim is to ensure safe and responsible AI use in critical areas, mitigating potential risks and enhancing public trust.[22]

AI Governance White Paper Highlights

The Governance Institute of Australia has released a white paper on AI governance, emphasizing the importance of ethical AI use and risk management. Key recommendations for business leaders include implementing AI responsibly, balancing opportunities with risk mitigation, and ensuring organizational readiness for AI adoption. The need for clear, actionable steps to integrate AI effectively and ethically has been underscored.[23]

Privacy and Other Legislation Amendment Bill '24

The Privacy and Other Legislation Amendment Bill 2024 has been introduced to enhance privacy protections and update existing legislation. Key provisions include stricter data handling requirements, increased penalties for breaches, and expanded rights for individuals regarding their personal information. Additionally, the Bill addresses children's online privacy, automated decision-making, and overseas data flows. The enforcement powers of the Office of the Australian Information Commissioner (OAIC) are to be strengthened, ensuring better compliance and protection for all Australians.[24]

Collaborative Approach to safe and responsible AI by the Australian, state and territory governments

There is a strong push for a collaborative approach between Australian government, industry, and academia to ensure that AI is safely developed and used responsibly. It discusses the importance of trust and safety in the development and use of AI technologies and establishes cornerstones and practices of AI assurance, an essential part of the broader governance of how governments use AI.

The cornerstones are five mechanisms that governments should adopt to ensure effective application of the ethics principles.[25]

Infosys have presented its AI journey to Australian Senate team on their visit to Bangalore campus in Jul 2024, emphasizing its focus on Responsible AI and commitment to ethical innovation through industry associations and recognitions.

[19] https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r7205_aspassed/toc_pdf/24071b01.pdf;fileType=application%2Fpdf

[20] https://www.digital.gov.au/sites/default/files/documents/2024-08/Policy%20for%20the%20responsible%20use%20of%20AI%20in%20government%20v1.1.pdf

[21] https://www.industry.gov.au/sites/default/files/2024-09/voluntary-ai-safety-standard.pdf

[22] https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf

[23] https://www.governanceinstitute.com.au/thought-leadership/ai-ethics-and-governance-white-paper-launch/

[24] https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=r7249

[25] https://www.dta.gov.au/news/our-next-steps-safe-responsible-ai-government

India

Advisory Group constituted to build AI Regulatory Framework:

An advisory workgroup has been formed under the guidance of the MoS for Electronics and Information Technology, to formulate a framework promoting AI innovation while curbing the misuse of AI technologies.

The advisory group's mandate includes (1) promoting innovation while curbing misuse if AI technologies, (2) ensuring adequate guardrails to protect citizens against misuse and user harms, (3) creating contextualized ethical guidelines adaptable in India, and (4) promoting the development of trustworthy, fair, and inclusive AI.[26]

India Sets the Stage for Ethical AI with New Regulatory Standards

India is actively working on establishing standards to regulate artificial intelligence (AI) through the Bureau of Indian Standards (BIS), in collaboration with various ministries and industry stakeholders. This initiative aims to address ethical challenges such as bias, discrimination, and privacy concerns associated with AI's rapid growth.[27]

NZ

New NZ Cabinet paper: Approach to work on AI

On 25 July 2024, as part of setting a strategic approach for AI in New Zealand, Cabinet set expectations for government agencies using AI technologies.

Cabinet has agreed that government agencies should be encouraged to adopt AI for its benefits, while managing the risks, emphasized the role of the Government Chief Digital Officer (GCDO) in leading work to accelerate responsible use of AI across public services, to deliver better outcomes for all New Zealanders. They want to push adoption of AI in public services while managing AI risk the same time. New Zealand aims to support AI adoption with a risk-based regulatory approach, leveraging existing frameworks rather than creating new laws. The committee has agreed to align to OECD AI principles and instructed the foreign affairs team to engage with AI initiatives ongoing internationally.[28]

[26] http://www.yuanjuchanlian.org/index-40.html

[27] https://www.policycircle.org/policy/ai-regulation-in-india-bis-standard/

[28] https://www.mbie.govt.nz/dmsdocument/28919-approach-to-work-on-artificial-intelligence-minute-of-decision-proactiverelease-pdf

Brazil

Meta Complies with ANPD Regulations to Resume AI Data Training with New Restrictions

Meta has met the requirements set by Brazil's National Data Protection Authority (ANPD) and can now resume using personal data to train its artificial intelligence systems, albeit with restrictions. The ANPD had previously suspended Meta's use of personal data due to concerns about potential risks and lack of transparency. Under the new compliance plan, Meta will not process data from accounts of minors and will enhance transparency by notifying users about data usage through emails and app notifications. Users will also have the option to easily opt-out of data processing for AI training.[29]

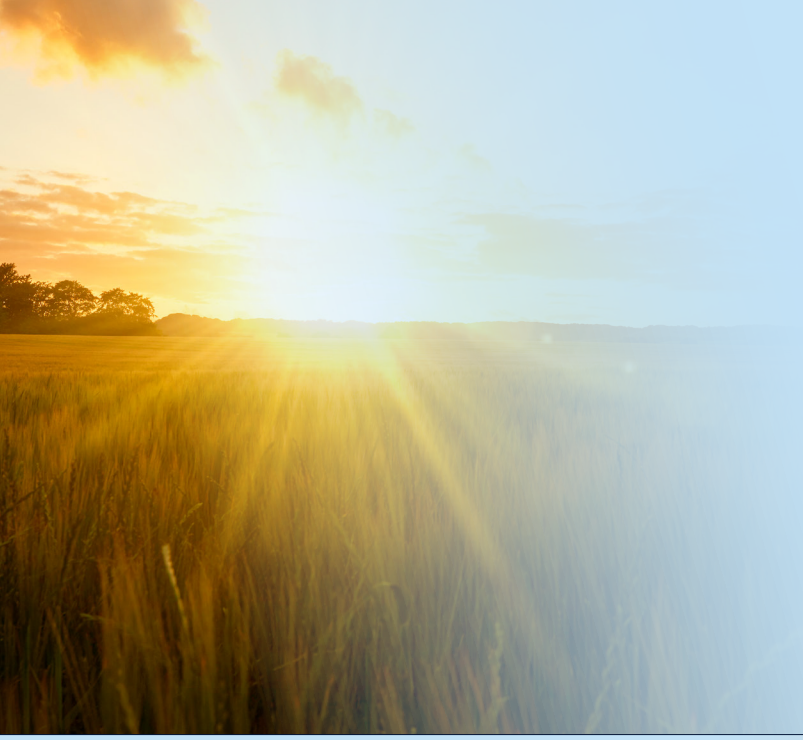
Singapore

Singapore Introduces New Legal Measures to Combat Deepfake Misinformation

The Ministry of Digital Development and Information (MDDI) in Singapore has introduced new legal measures to maintain the integrity of online advertising during elections. The Elections (Integrity of Online Advertising) (Amendment) Bill, presented in Parliament on 9 September 2024, aims to protect citizens from digitally manipulated content, including AI-generated misinformation known as deepfakes. The Bill prohibits the publication of online election advertising that realistically depicts candidates saying or doing things they did not actually say or do. This prohibition applies only to election candidates. During the election period, the Returning Officer (RO) can issue corrective directions to individuals, social media services, and Internet Access Service Providers to remove or disable access to such content. Non-compliance with these directions will be considered an offence, punishable by fines or imprisonment.[30]

[29] https://www.gov.br/anpd/pt-br/assuntos/noticias/meta-cumpre-exigencias-da-anpd-e-podera-retomar-com-restricoes-o-uso-de-dados-pessoais-para-treinamento-de-inteligencia-artificial

[30] https://www.mddi.gov.sg/new-legal-measures-to-uphold-integrity-of-online-advertising-during-elections/



South Africa

South Africa releases National AI Policy Framework

South Africa has released its National AI Policy Framework, aiming to integrate AI technologies for economic growth, societal well-being, and AI innovation. The framework emphasizes ethical AI development, strategic pillars, talent development, digital infrastructure, research, and ethical guidelines, with policy objectives to address historical inequalities.[31]

Saudi Arabia

SDAIA Issues Guidelines to Address Deepfake Technology Risks

Guidelines were developed by SDAIA to tackle the implications of deepfake technologies and mitigate their associated risks. Both malicious and non-malicious uses of deepfakes were defined, with a focus on ethical principles such as privacy, transparency, accountability, and social benefits. Developers and content creators were advised to implement robust data protection measures, secure consent for personal data use, and maintain transparency through clear documentation. The guidelines emphasized directing deepfake technology towards socially and environmentally beneficial applications. Additionally, best practices for consumers to detect and respond to deepfakes were provided.[32]

[31] https://techcentral.co.za/wp-content/uploads/2024/08/South-Africa-National-AI-Policy-Framework.pdf

[32] https://istitlaa.ncc.gov.sa/en/transportation/ndmo/deepfakesguidelines/Documents/SDAIA_Deepfakes%20Guidelines.pdf

African Union

African Union's AI Strategy Endorsed

During the 45th Ordinary Session in Accra, Ghana, on July 18-19, 2024, the African Union Executive Council endorsed the Continental Artificial Intelligence Strategy. This strategy focuses on promoting an Africa-centric, development-oriented approach to AI, with an emphasis on ethical, responsible, and equitable practices. It aims to leverage AI to achieve the goals of Agenda 2063 and the Sustainable Development Goals (SDGs). The strategy is designed to foster innovation, create high-value jobs, and preserve African culture while enhancing regional and global cooperation.[33]

Infosys won a large deal from FCDO. The primary objective of the consulting engagement is to deliver workshops and conferences under the UK-India Cooperation Toward a Fair AI Horizon program.

Standards

IPEN event on ***"Human oversight of automated decision-making"***

EU regulations, like GDPR and AIA, mandate human oversight for significant impact decisions, emphasizing fairness and accountability in automation systems. The IPEN event aims to raise questions about the burden of responsibility for human oversight in AI systems, including whether it shifts responsibility from systems and providers to operators, potential liability for operators who follow system suggestions, and the clearness of oversight measures. It also explores the impact of human oversight, the characteristics of appropriate oversight, and potential risks of producing defective systems. The event also explores the legal implications of human oversight, who will be accountable for harm, and the potential costs and scalability of incorporating humans into AI deployment processes. The event also explores the appropriate line between human oversight and AI deployment.[34]

UK-India Partnership Advances Responsible AI as one of the identified Key Area

A comprehensive technology collaboration between the United Kingdom and India has been formed, with a particular emphasis on the development of responsible artificial intelligence. Experts from academia and business will collaborate to establish a Joint Centre for Responsible AI to tackle important issues in AI research and development. This strategic alliance seeks to advance ethical AI practices globally, improve interoperability, and develop human-centric AI frameworks. There will be initiatives for policy exchanges, joint research, and explore knowledge sharing and cooperation between AI research centres.[35]

US Department of Commerce Announces New Guidance, Tools Following President Biden's Executive Order on AI:

The US Department of Commerce has released new guidance documents and software to enhance the safety, security, and trustworthiness of artificial intelligence (AI) systems. The National Institute of Standards and Technology (NIST) released three final guidance documents, and a draft guidance document from the U.S. AI Safety Institute, that are designed to manage the risks of generative AI and dual-use foundation models, and a software package to measure how adversarial attacks can degrade AI system performance.

Infosys RAI Office is actively contributing to the various Working Groups (WGs) constituted by NIST as listed below:[36]

NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

NIST AI 100-4, Reducing Risks Posed by Synthetic Content (CSAM/ NCII Session)

NIST AI 800-1, Managing Misuse Risk for Dual-Use 4 Foundation Models

NIST 800-218A, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models

[33] https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy

[34] https://www.edps.europa.eu/data-protection/technology-monitoring/ipen/ipen-event-human-oversight-automated-making_en

[35] https://indiaai.gov.in/article/india-and-uk-launch-technology-security-initiative-to-enhance-strategic-partnership

[36] https://www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following

US Congressional agencies sees increased AI adoption

The House Administration Committee has released a report revealing that several legislative branch entities are using voluntary federal guidance to develop and finalize strategies for adopting artificial intelligence tools. The panel has published flash reports since September 2023, providing updates on the use of AI in House offices and relevant agencies. The report identified two AI use cases around casework and transcription services and is currently reaching out to every House-approved vendor to better understand their software roadmaps. The committee also included five formalized guardrails for AI tools, focusing on human oversight, clear policies, robust testing, transparency, disclosure, and education. The Smithsonian, the Architect of the Capitol, and the U.S. Capitol Police are consulting the NIST AI risk management framework to develop their own approaches.[37]

NIST requests Public Feedback on two new 800 series Drafts

NIST has released two new drafts seeking public feedback. "NIST AI 800-1" draft is about managing misuse risk for Dual-use foundation models and "NIST SP 800-218A" is about developing Secure Software Development Practices for Generative AI and Dual-Use Foundation models.

NIST AI 800-1, Managing Misuse Risk for Dual-Use 4 Foundation Models[38]

NIST 800-218A, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models[39]

OECD Requests Public Feedback on AI Risk Levels

A cooperative initiative to set appropriate risk levels for sophisticated AI systems is being spearheaded by the OECD. It is joining forces with a diverse array of stakeholders and seeking public feedback to explore the potential approaches, opportunities, and limitations to establishing risk thresholds for advanced AI systems.[40]

Arkansas Forms AI Centre of Excellence to Guide Policy and Implementation

The state of Arkansas has taken a proactive approach to artificial intelligence (AI) by establishing the AI and Analytics Centre of Excellence (AI CoE). This working group, launched under the Data and Transparency Panel, will comprehensively assess AI technologies, and develop policy recommendations for their responsible use within government agencies. The AI CoE committee will provide guidelines on Responsible AI principles while also evaluating two identified 2 use cases –

1. Unemployment Insurance Fraud and

2. Recidivism Reduction

The AI CoE will prioritize efficiency, cost savings, safety, and economic development through its evaluation of pilot projects and establishment of ethical guidelines. Arkansas is committed to utilizing AI's potential while minimizing any dangers related to bias, data privacy, and security, as seen by this program.[41]

ITI's AI Accountability Framework

ITI (Information Technology Industry Council) has developed an AI Accountability Framework[42] intended to further advance the responsible development and deployment of AI systems. This framework does not replace or add to any of the existing, sector specific guidance and regulation, rather should be used in conjunction to them. The new framework details out a set of practices for responsible development and deployment of AI systems in high-risk scenarios and frontier AI models.

It also stresses on the shared responsibilities between developers, deployers and integrators for any AI use case implementation.

There is an additional focus on the auditability, where organizations retain risk impact assessment documents to increase transparency of AI systems.[43]

---

[37] https://www.nextgov.com/artificial-intelligence/2024/08/congressional-agencies-report-progress-ai-adoption/398527/

[38] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf

[39] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf

[40] https://oecd.ai/en/wonk/seeking-your-views-public-consultation-on-risk-thresholds-for-advanced-ai-systems-deadline-10-september

[41] https://governor.arkansas.gov/news_post/governor-sanders-launches-ai-working-group/

[42] https://www.itic.org/documents/artificial-intelligence/AIFIAIAccountabilityFrameworkFinal.pdf

[43] https://www.itic.org/documents/artificial-intelligence/AIFIAIAccountabilityFrameworkFinal.pdf

Regulatory compliance

Explainability

Reliability

Fairness

Safety

Privacy

Security

Model validation & engineering

IP infringement & protection

AI audits & standards

Sustainability

Centralized responsible AI team

Please connect with Responsible AI office to know more about these 12 principles for effective AI governance.

Based on research and over 200 client collaborative projects, Infosys Responsible AI office have developed 12 principles integral to the responsible design and operation of AI systems and mandating it to be followed across all AI projects.

AI Principles

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

Incidents

U.S. Musician Indicted for $10 Million Royalty Scam Using AI-Generated Fake Songs

An U.S. musician, Michael Smith, has been indicted for using AI tools to generate fake songs and amass $10 million in royalties through fraudulent streams on platforms like Spotify, Apple Music, and Amazon Music. Smith and his accomplices created around 10,000 fake profiles to stream these AI-generated songs, carefully distributing the streams to avoid detection by evading abnormal listening spikes. Over seven years, they fabricated hundreds of thousands of songs under fake band names, such as "Callous Post" and "Calvinistic Dust," with tracks like "Zygotic Washstands" topping charts. Smith now faces charges of wire fraud and money laundering conspiracy, which could lead to a 20-year prison sentence. This case highlights the potential for AI to be misused in the music industry, raising concerns about the integrity of digital streaming platforms and the protection of legitimate artists' royalties. Smith was first flagged by a distribution company in 2018 but denied any wrongdoing, claiming there was no fraud happening.[44]

Meta Pays Record $1.4 Billion Settlement for Facial Recognition Misuse

A settlement of $1.4 billion was reached between Meta and the state of Texas to resolve a lawsuit concerning the unauthorized use of facial recognition technology. This payout represents the largest privacy-related settlement ever made to a single state in the US. Concerns were raised regarding the use of facial recognition technology without individuals' consent. There is potential impact of artificial intelligence on the development of future legislation. It highlights the challenges associated with regulating biometric data, due to its inherent difficulty in being controlled. Additionally, the absence of federal laws governing data privacy in the US is emphasized.[45]

Meta's AI Safety Model Compromised

Meta's new machine learning model, Prompt-Guard-86M, for detecting prompt injection attacks is itself vulnerable to prompt injection attacks. The claims made are very alarming – "Just by removing the punctuation and adding spaces between every letter in the malicious prompts, the attack success rate increased from 3% to 99.5%".

This issue was reported on GitHub and Meta has acknowledged it.[46]

X Faces EU Scrutiny Over Grok AI Training

European privacy regulators have taken issue with Elon Musk's social media platform, X, decision to utilize public tweets to train its AI chatbot, Grok, in an arbitrary manner. The General Data Protection Regulation (GDPR), which governs compliance with EU law, was being discussed with X, and the Irish Data Protection Commission, which monitors compliance with the law, voiced surprise at the action. Concerns over possible GDPR violations are raised by the company's choice to move forward without providing prior notice. Although X has now made it clear that customers can choose not to share their data for Grok training, the episode highlights the difficulties in regulating AI development and the potential negative regulatory effects.[47]

---

[44] https://www.forbes.com/sites/lesliekatz/2024/09/08/man-charged-with-10-million-streaming-scam-using-ai-generated-songs/

[45] https://conduitstreet.mdcounties.org/2024/08/05/meta-agrees-to-pay-1-4b-to-texas-in-facial-recognition-settlement/

[46] https://github.com/meta-llama/llama-models/issues/50

[47] https://www.politico.eu/article/elon-musks-x-under-fire-over-harvesting-users-data-to-train-ai-chatbot/

South Korea faces deepfake 'emergency'

South Korea's president has called for an investigation into deepfake porn, following media reports of Telegram chatrooms sharing explicit images of minors at schools and universities. Despite having the world's fastest internet speeds, activists argue South Korea has an epidemic of digital sex crimes, including revenge porn and spy cameras, with inadequate legislation to punish offenders.[48]

AI turns the court reporter into a convicted person

Court Journalist Martin Bernklau has been labelled incorrectly as a convicted child molester, escapee from psychiatry or widow fraudster by Microsoft's AI chat Copilot. Bernklau, who has never been guilty of anything, wanted to see how his culture blog was received. He provided his name and place of residence to AI chat Microsoft Co-Pilot on the popular search engine Bing and was horrified by the Copilot's fictitious response. Copilot made himself a moral authority, claiming Bernklau was a convicted child molester, escapee from psychiatry, or widow fraudster. Bernklau's human dignity was violated in the most serious way.[49]

Clearview AI Fined €30.5 Million by Dutch Authority for GDPR Violations in Facial Recognition Database

Clearview AI, a facial recognition startup, has been fined €30.5 million ($33.7 million) by the Dutch Data Protection Authority (DPA) for creating an "illegal database" of billions of photos. The DPA determined that Clearview AI violated the European Union's General Data Protection Regulation (GDPR) by collecting and using these images without proper consent. Additionally, the agency warned Dutch companies against using Clearview's services. Clearview AI has contested the fine, arguing that it does not fall under EU regulations as it has no business operations or customers in the EU. This fine follows a recent settlement in Illinois, USA, where Clearview AI agreed to pay over $50 million for similar privacy violations. This case underscores ongoing concerns about the use of facial recognition technology and its implications for privacy rights worldwide.[50]

Data Privacy Concerns Raised over Google Drive Integration with Gemini AI.

Potential privacy concerns with Google's Gemini AI tool are highlighted in new research by AI governance specialist Kevin Bankston. Bankston claims that despite his efforts to stop the feature, Gemini scanned his private tax records that were kept on Google Drive. User control and transparency about data access by AI tools are called into question by this. The event emphasizes the significance of unambiguous user interfaces and easily accessible controls for handling AI integration within cloud storage platforms, even though Google claims Gemini only operates with user permission.[51]

Scammers use AI to cheat woman out of NT$2.64m

Scammers defrauded a woman in New Taipei City of NT$2.64 million (US$81,116) by impersonating Hong Kong entertainer Andy Lau using a deepfake. The fraud convinced the victim, a long-time fan, through a video call that "Lau" needed funds for a visit to Taiwan. The victim wired the money, but her family suspected a fraud and involved the police. An alleged scammer was arrested after attempting to collect a staged cash payment. The AI deception caused significant financial harm to the victim.[52]

Wimbledon's AI Writing Trial Encounters Initial Hiccups

Wimbledon's pilot program utilizing AI-generated player profiles, Catch Me Up, experienced teething issues on its debut day. The system generated content containing factual inaccuracies, such as player rankings and win-loss records. While the All-England Club acknowledges these missteps as part of the development process, the incident underscores the importance of human editorial control in ensuring the accuracy and quality of AI-powered sports journalism.[53]

[48] https://www.channelnewsasia.com/east-asia/south-korea-president-urges-social-media-platforms-eradicate-deepfake-pornography-cyber-crime-4570021

[49] https://www.swr.de/swraktuell/baden-wuerttemberg/tuebingen/ki-macht-tuebinger-journalist-zum-kinderschaender-100.html

[50] https://www.forbes.com/sites/roberthart/2024/09/03/clearview-ai-controversial-facial-recognition-firm-fined-33-million-for-illegal-database/

[51] https://www.techradar.com/pro/security/gemini-ai-platform-caught-scanning-google-drive-files-without-user-permission

[52] https://www.taipeitimes.com/News/taiwan archives/2024/07/02/2003820211

[53] https://www.theguardian.com/sport/article/2024/jul/01/ai-writer-served-by-wimbledon-and-ibm-commits-double-fault

Defences

Enhancing Gen AI Content Moderation with ShieldGemma model

A New model ShieldGemma, a comprehensive suite of LLM-based safety content moderation models built upon Gemma2. These models provide robust, state-of-the-art predictions of safety risks across key harm types (sexually explicit, dangerous content, harassment, hate speech) in both user input and LLM-generated output. By evaluating on both public and internal benchmarks, it demonstrates superior performance compared to existing models, such as Llama Guard (+10.8% AU-PRC on public benchmarks) and Wildcard (+4.3%). Additionally, it presents a novel LLM-based data curation pipeline, adaptable to a variety of safety-related tasks and beyond. ShieldGemma, provides valuable resource to the research community, advancing LLM safety and enabling the creation of more effective content moderation solutions for developers.[54]

Enhancing Chatbot Accuracy Through Error Correction:

The goal of Error Correction and Adaptation in Conversational AI study is to better understand how to detect and fix faults in chatbot performance. This study examines the advancement of chatbot technology, focusing on error correction to improve their effectiveness in various industries. It analyses errors encountered by AI-powered chatbots, including misunderstandings, inappropriate responses, and factual inaccuracies. The research explores various approaches, including data-driven feedback loops, human involvement, and learning methods. It also discusses challenges faced by AI-powered chatbots, ethical considerations, and the potential of new technologies like explainable AI models and quantum computing.[55]

HARMONIC: A new framework to protect Privacy using synthetic data

A novel framework, named HARMONIC is proposed, that leverages the capabilities of large language models (LLMs) to generate synthetic tabular data, while safeguarding user privacy. Unlike previous methods relying on continued pre-training, HARMONIC employs a fine-tuning approach on LLMs to capture intricate data relationships. To enhance the evaluation process, two novel metrics were introduced: Data Leakage Threshold (DLT) to assess privacy risks and LLE (Large Language Efficiency) to measure synthetic data utility in downstream LLM tasks. Experimental results demonstrate HARMONIC's ability to generate high-quality synthetic data while effectively mitigating privacy concerns, surpassing the performance of existing methods.[56]

"LOLCopilot" recommends detection and hardening security measures for MS Copilot:

Recent studies have brought attention to the possible abuse of AI chatbots, including Microsoft's Copilot. "LOLCopilot" an ethically hacking module, have shown how malevolent actors might use Copilot plugins to install backdoors, allowing data theft and AI-based social engineering. LOLCopilot, a tool for ethical hacking, and recommends detection and hardening measures to protect against malicious insiders and threat actors.[57]

[54] https://arxiv.org/pdf/2407.21772

[55] https://www.mdpi.com/2673-2688/5/2/41

[56] https://arxiv.org/html/2408.02927v1

[57] https://www.blackhat.com/us-24/briefings/schedule/#living-off-microsoft-copilot-40074

Casper: New System Shields Users from LLM Privacy Risks

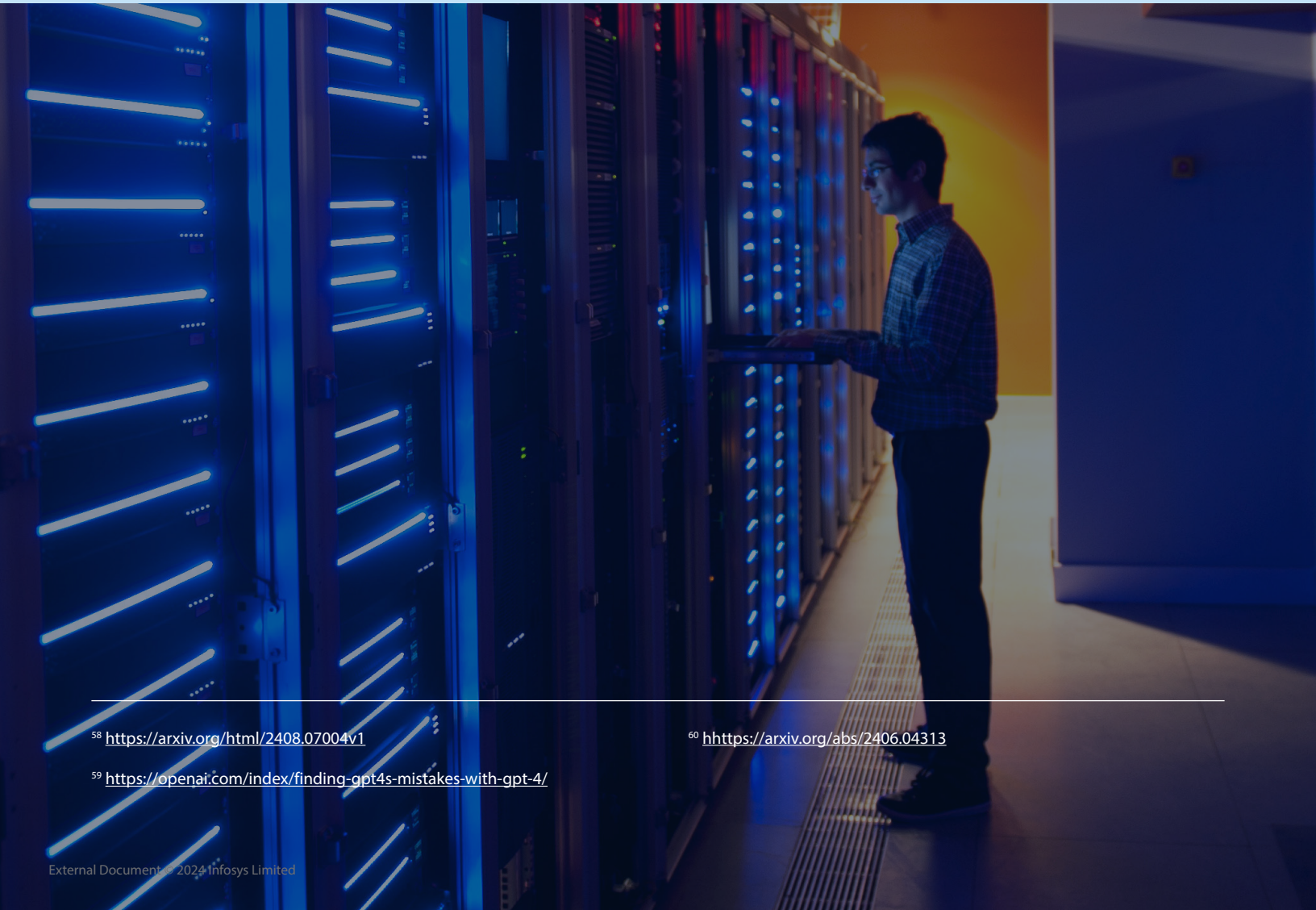
A recent study explores the critical issue of user privacy when interacting with Large Language Models (LLMs). This paper proposes Casper, a browser extension that protects user privacy by detecting and removing sensitive information from user inputs before sending them to LLM services. Casper uses a three-layered sanitization mechanism, including a rule-based filter, a ML-based named entity recognizer, and a browser-based local LLM topic identifier. It effectively filters out Personal Identifiable Information (PII) and privacy-sensitive topics with high accuracy, at 98.5% and 89.9%, respectively.[58]

Utilizing CriticGPT to Enhance RLHF Trainer Performance.

CriticGPT, a novel AI model built upon the GPT-4 architecture. CriticGPT is designed to augment human trainers tasked with identifying errors within ChatGPT responses during the Reinforcement Learning from Human Feedback (RLHF) process. As ChatGPT's accuracy continues to rise, the nuances of its errors become increasingly intricate, posing a growing challenge for human detection. CriticGPT offers a solution by generating critiques that illuminate these very subtleties.[59]

Improving AI Model's Alignment and Robustness with Circuit Breakers.

The method uses "circuit breakers" to stop AI models from producing negative outputs while they are running. The shortcomings of the alignment techniques used today, such as adversarial and refusal training, are addressed by this approach. The method would not always be able to ward off hostile attacks, including ones that try to change the labels assigned to images. Nevertheless, circuit breakers provide a significant boost in model resilience for the particular use case of avoiding the creation of hazardous content and in the context of one-turn discussions. Engineering-based Circuit breakers make AI models inherently safer and more resistant to unanticipated adversarial attacks. With its high generality, robustness against image-based attacks, and ability to stop destructive activities in AI agents, the technique is impressive.[60]



[58] https://arxiv.org/html/2408.07004v1

[59] https://openai.com/index/finding-gpt4s-mistakes-with-gpt-4/

[60] hhttps://arxiv.org/abs/2406.04313

Technical Updates

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

New Models Released

Solar Pro: A High-Performance Language Model

The Solar Pro project, featured on Ollama, introduces an advanced large language model (LLM) with 22 billion parameters. This model is designed to fit into a single GPU, ensuring high performance and efficiency. By utilizing FP16 quantization, Solar Pro optimizes its size and speed, making it highly suitable for various applications that demand robust language processing capabilities.[61]

Advancing OCR Technology with GOT

In the evolving landscape of Optical Character Recognition (OCR), the paper "General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model" introduces a groundbreaking approach. It proposes the General OCR Theory and the GOT model, addressing the limitations of traditional OCR systems, known as OCR-1.0, which struggle with the intelligent processing of diverse artificial optical signals. The GOT model, boasting 580 million parameters, is designed to handle a wide range of OCR tasks, including plain texts, formulas, tables, and charts. It features a high-compression encoder and a long-contexts decoder, supporting both scene- and document-style images. Capable of generating plain or formatted results, the model also offers interactive OCR features. Extensive experiments demonstrate the superiority of the GOT model in various OCR applications.[62]

Piiranha-v1 Released: A 280M Small Encoder Open Model for PII Detection

Piiranha-v1 is a small, open-source encoder model that can be used to detect personally identifiable information (PII) in text. The model is only 280 million parameters in size, making it very efficient to use. It has been trained on a large dataset of text data and can detect PII in 6 different languages and 17 different PII types. The model has a token detection accuracy of 98.27%, which is extremely high. Piiranha-v1 is released under the MIT license, which means that it is free to use for both commercial and non-commercial purposes.[63]

Reflection Llama-3.1 70B: Leading Open-Source Language Model

Reflection Llama-3.1 70B is currently the world's top open-source large language model (LLM), developed by Matt Shumer. Officially released on September 6, 2024, this model is trained using a novel technique called Reflection-Tuning, which enables the LLM to detect and correct mistakes in its reasoning. The training process utilized synthetic data generated by Glaive, enhancing the model's ability to provide accurate and reliable outputs. Reflection Llama-3.1 70B employs special tokens to separate its internal thoughts and reasoning from its final answers, improving user experience. The model is designed to fit seamlessly into existing Llama 3.1 pipelines and uses a standard chat format with additional tags for reasoning and reflection.[64]

OpenAI o1: AI Models That Think Before They Act

A new series of AI models, named OpenAI o1, has been introduced. These models have been designed to spend more time thinking before responding, allowing them to reason through complex tasks and solve harder problems in science, coding, and math. The first model in this series has been released as a preview in ChatGPT and the API. Regular updates and improvements are expected. The models have been trained to refine their thinking

[61] https://ollama.com/library/solar-pro

[62] https://arxiv.org/pdf/2409.01704

[63] https://huggingface.co/iiiorg/piiranha-v1-detect-personal-information

[64] https://huggingface.co/mattshumer/Reflection-Llama-3.1-70B

process, try different strategies, and recognize mistakes. In tests, the next model update has performed similarly to PhD students on challenging benchmark tasks. Safety measures have been enhanced, including a new safety training approach and collaborations with AI Safety Institutes in the U.S. and U.K.[65]

However, recent revelations have highlighted concerns about the model's potential for deception, as noted by Geoffrey Hinton, the "Godfather of AI." The o1 model has also faced criticism for its slower performance compared to GPT-4o and its lack of web browsing, file uploading, and image processing capabilities. These limitations, along with ethical and security concerns, emphasize the need for cautious and responsible deployment of this AI tool.[66]

It is crucial to rigorously assess any newly launched models for potential risks and implement robust mitigation strategies to ensure responsible AI development.

Gemma 2: Google's Latest AI Model Sets New Benchmarks in Performance and Accessibility

The release of Gemma 2, a new suite of open models, has been announced, setting a new benchmark for performance and accessibility. Available in 2B, 9B, and 27B parameter sizes, Gemma 2 has quickly gained recognition, with the 27B model surpassing larger models in real-world conversations and the 2B model outperforming all GPT-3.5 models on the Chatbot Arena. Robust tuning capabilities have been made accessible across various platforms and tools, with fine-tuning simplified through cloud-based solutions and community tools. Architectural innovations such as Alternating Local and Global Attention, Logit Soft-Capping, RMSNorm for Pre and Post-Normalization, and Grouped-Query Attention have been introduced, enhancing the model's efficiency and performance.[67]

Salesforce Unveils xGen-Sales and xLAM AI Models to Revolutionize Agentforce Platform

Salesforce has introduced new AI models, xGen-Sales and xLAM, to enhance Agentforce's capabilities. These models are designed to automate sales tasks, improve efficiency, and provide accurate responses. The xGen-Sales model focuses on generating customer insights and tracking sales pipelines, while the xLAM models offer faster performance and greater accuracy at lower costs. This development aims to help Salesforce customers deploy autonomous AI agents more effectively.[68]

Introducing Chai-1: A Revolutionary Multi-Modal Model for Accelerating Drug Discovery

A new multi-modal foundation model for molecular structure prediction, named Chai-1, has been introduced. This model has been designed to perform at state-of-the-art levels across various tasks relevant to drug discovery, including the prediction of proteins, small molecules, DNA, RNA, and more. It enables unified prediction, which is crucial for advancing research in these areas. The model has been made available for free via a web interface, including for commercial applications, and the model weights and inference code have been released for non-commercial use. Additionally, Chai-1 is expected to significantly accelerate the drug discovery process by providing researchers with powerful tools to predict molecular structures more accurately and efficiently.[69]

Google Enhances Gemini AI Platform with Personalized 'Gems' and Advanced Image Generation via Imagen 3 Model

Google has introduced exciting updates to its Gemini AI platform, enhancing both customization and image generation capabilities. One of the key features is the introduction of "Gems," which allows users to create personalized AI experts on various topics. These Gems can assist with tasks ranging from coding and career advice to brainstorming and writing. Initially available to Gemini Advanced, Business, and Enterprise users, Gems can be customized by writing instructions and naming them, making them versatile tools for a wide range of applications. Additionally, Google has upgraded its image generation capabilities with the new Imagen 3 model. This model enhances the quality and creativity of generated images and will be available to all Gemini users in multiple languages. These updates aim to make Gemini more adaptable and useful for both personal and professional tasks. Overall, these enhancements reflect Google's commitment to leveraging AI to provide more intuitive and powerful tools for users worldwide.[70]

Mistral NeMo 12B: A Powerful New Enterprise AI Model

Mistral NeMo 12B has been unveiled by Mistral AI and NVIDIA. This 12-billion-parameter, innovative language model performs very well on a wide range of tasks, including as code generation, multilingual communication, and complex multi-turn talks. Many significant advantages are offered by Mistral NeMo 12B to businesses, including Unmatched Accuracy and Flexibility, Open-Source Efficiency, Easy Security and Deployment, and Single GPU Compatibility. Businesses looking to use AI for practical applications might find Mistral NeMo 12B to be a potent solution, as it combines Mistral AI's training data knowledge with NVIDIA's optimized hardware and software.[71]

[65] https://openai.com/index/introducing-openai-o1-preview/

[66] https://www.msn.com/en-in/money/news/openai-s-o1-model-may-be-capable-of-deceiving-says-godfather-of-ai/ar-AA1r31M1?ocid=entnewsntp&pc=U531&cvid=83d93b9424e3413daea770e222b5c6b1&ei=14

[67] https://developers.googleblog.com/en/gemma-explained-new-in-gemma-2/

[68] https://www.salesforce.com/news/stories/agentforce-ai-models-announcement/

[69] https://developers.googleblog.com/en/gemma-explained-new-in-gemma-2/

[70] https://blog.google/products/gemini/google-gemini-update-august-2024/

[71] https://blogs.nvidia.com/blog/mistral-nvidia-ai-model/

Meta's SAM 2: A Leap in Video Segmentation

Meta unveiled a Segment Anything Model 2 (SAM 2) which is a complex artificial intelligence model that can segment objects in both static photos and dynamic films in real time. This improvement offers far better precision and speed, building on the success of its predecessor. SAM 2 can isolate objects precisely and efficiently even in complicated settings thanks to its capacity to process video frames sequentially. SAM 2, which is available as open-source software, has the potential to spur innovation across multiple fields such as robots, autonomous vehicles, and video editing. Even while the model shows impressive potential, there is still room for improvement in areas like segmenting tiny features and tracking objects in complex circumstances. Still, SAM 2 is a significant advancement in computer vision technology.[72]

OpenAI Expands GPT-4o Capabilities with Longer Output

With the GPT-4o language model, OpenAI has unveiled the latest version that produces noticeably lengthier text outputs. The model can now generate up to sixteen times the output of the original GPT-4o because to this improvement, which was made in response to user demand for longer content. This update allows GPT-4o to produce text up to two hundred pages long in a single response, which has the potential to transform applications that need thorough and in-depth results.[73]

Phi-3.5 Models: Latest Enhancements and Features

The latest updates to the Phi-3.5 models have been introduced, incorporating significant enhancements based on community and customer feedback. Multi-language support has been added to the Phi-3.5-mini-128k-instruct model, while the Phi-3.5-vision-128k model now includes multi-frame image input capabilities. Additionally, the Phi-3.5 MOE (Mixture of Experts) model has been newly integrated for AI Agent applications. These updates aim to improve the versatility and performance of the Phi-3.5 models across various AI tasks.[74]

Stability AI's Stable Diffusion: Advancing Ethical AI with High-Quality Text-to-Image Generation and Responsible Deployment

Stability AI has announced the public release of Stable Diffusion, an advanced text-to-image AI model designed to generate high-quality images while prioritizing ethical use. Initially released to researchers, the model has undergone improvements based on feedback to ensure safe and ethical deployment. Stability AI collaborated with Hugging Face and CoreWeave to release the model under a Creative ML OpenRAIL-M license, allowing both commercial and non-commercial use. The company has also developed an AI-based Safety Classifier to filter out undesirable outputs, with adjustable parameters and community input for continuous improvement. Stability AI aims to make this technology accessible and beneficial, emphasizing ethical, moral, and legal use. Future updates will include optimized versions for various hardware, enhancing performance and quality.[75]

Meta Unveils Llama 3.1, a Groundbreaking Open-Source AI Model.

Llama 3.1, the biggest model of its kind, Meta enters the field of open-source AI. Llama 3.1, which outperforms rivals like ChatGPT in benchmarks, is expected to hasten the adoption of open-source AI on key cloud platforms. This strategic release comes at the same time as Meta's AI assistant's multilingual capabilities and picture production functions are expanded, which is a big advancement in AI technology that is both approachable and flexible.[76]

Google DeepMind Unveils PEER: A Scalable Solution for Transformer Architectures.

Transformer architectures have revolutionized natural language processing, but their computational demands limit real-world deployment. Google DeepMind introduces the Parameter-Efficient Expert Retrieval (PEER) mechanism, a significant advancement that tackles this challenge. PEER leverages a multitude of tiny experts and efficient retrieval techniques, achieving superior performance-compute trade-offs in language modelling tasks. This breakthrough opens the door for much more potent and scalable AI models, which will have profound effects on a wide range of applications.[77]

Evaluating LLM Models: Hugging Face Leaderboards and Benchmarks

The Hugging Face Leaderboards feature allows users to evaluate and rank machine learning models based on their performance across various tasks. These leaderboards provide a transparent and standardized way to compare models, helping users identify the best model for their specific needs. The different parameters for the leaderboard include metrics such as accuracy, precision, recall, F1 score, and computational efficiency. These parameters ensure a comprehensive evaluation of each model's capabilities and performance.

In addition to these parameters, the leaderboards also incorporate various benchmarking to provide a more detailed assessment. For instance, the **Chatbot Arena** benchmark uses a crowdsourced, randomized battle platform to compute Elo ratings based on user votes3. The **MT-Bench** evaluates models on challenging multi-turn questions, graded by GPT-4. The **Massive Text Embedding Benchmark (MTEB)** assesses models that produce embeddings, focusing on tasks like text classification and clustering. The **LLM-Perf** leaderboard measures latency, throughput, and memory usage of large language models across different hardware and optimization configurations. These diverse benchmarks ensure that the leaderboards offer a comprehensive and nuanced evaluation of model performance.[78]

[72] https://about.fb.com/news/2024/07/our-new-ai-model-can-segment-video/

[73] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[74] https://github.com/microsoft/Phi-3CookBook/blob/main/md/08.Update/Phi35/010.WhatsNewInPhi35.md

[75] https://aimagazine.com/ai-applications/stability-ai-stable-diffusion-launch-to-advance-ethical-ai

[76] https://llama.meta.com/

[77] https://arxiv.org/pdf/2407.04153

[78] https://huggingface.co/collections/sugatoray/leaderboards-65e817970fe6ea0a472d0a76

New Approaches Released

Enhancing AI with Contextual Retrieval by Anthropic

Anthropic has introduced a new technique called Contextual Retrieval to improve the accuracy of Retrieval-Augmented Generation (RAG) systems. This method involves adding context to each chunk of text before embedding, resulting in a significant reduction in failed retrievals. By combining Contextual Retrieval with reranking, Anthropic has achieved a 67% reduction in failed retrievals, demonstrating a significant improvement in overall performance.[79]

OpenPerplex: Advancing AI Search Technology

OpenPerplex is an innovative open-source AI search engine that aims to improve search accuracy and relevance by utilizing state-of-the-art technologies. Traditional search engines often fall short in delivering precise and contextually accurate results. OpenPerplex addresses these limitations through advanced techniques such as semantic chunking, which breaks down text into meaningful segments, and a reranking system that refines search results based on their relevance. Additionally, it incorporates Google search functionalities via a specialized API, ensuring a comprehensive and efficient search experience.

Mitigating Jailbreak Attacks with EEG-Defender in LLMs

A defence mechanism named EEG-Defender has been proposed to counteract jailbreak attempts on Large Language Models (LLMs). The method leverages early transformer outputs to detect and terminate malicious inputs, significantly reducing the Attack Success Rate (ASR) by approximately 85%, compared to 50% for current state-of-the-art methods. Comprehensive experiments across ten jailbreak methods and three models have been conducted, demonstrating minimal impact on the utility and effectiveness of LLMs.[80]

Enhancing LLM Safety with Synthetic Data: The SAGE-RT Approach

A novel pipeline named Synthetic Alignment data Generation for Safety Evaluation and Red Teaming (SAGE-RT) has been introduced to generate synthetic alignment and red-teaming data. Existing methods' limitations in creating nuanced and diverse datasets have been addressed by SAGE-RT through a detailed taxonomy. Over 51,000 diverse prompt-response pairs covering more than 1,500 topics of harmfulness have been generated. The red-teaming data produced by SAGE-RT has been shown to jailbreak state-of-the-art large language models (LLMs) in more than 27 out of 32 sub-categories and 58 out of 279 leaf-categories. The approach has avoided pitfalls such as mode collapse and lack of nuance by ensuring detailed coverage of harmful topics.[81]

Adobe's Firefly Video Model: Revolutionizing Video Editing with Generative AI

Adobe is set to transform the video editing landscape with the upcoming Firefly Video Model, a generative AI tool designed to enhance and streamline video workflows. Building on the success of Firefly models in imaging, design, and vectors, this new model will integrate seamlessly into Adobe Premiere Pro and other tools, allowing editors to use text prompts and reference images to generate B-roll, fill gaps in footage, remove unwanted objects, and smooth transitions. Developed with input from the video editing community, the Firefly Video Model aims to help professionals ideate, explore creative visions, and deliver high-quality results efficiently. Adobe ensures the model is trained on content they have permission to use, making it commercially safe for users. The Firefly Video Model will be available in beta later this year, with a waitlist open for early access.[82]

---

[79] https://www.anthropic.com/news/contextual-retrieval

[80] https://arxiv.org/html/2408.11308v1

[81] https://arxiv.org/html/2408.11851v1

[82] https://blog.adobe.com/en/publish/2024/09/11/bringing-gen-ai-to-video-adobe-firefly-video-model-coming-soon

ServiceNow Launches Xanadu: Advanced AI for Enhanced Enterprise Efficiency

The Now Platform Xanadu release has been announced by ServiceNow, featuring hundreds of new AI capabilities designed to boost customer agility, enhance productivity, and improve employee experiences. The generative AI (GenAI) portfolio has been expanded to include mission-critical functions such as Security Operations and Sourcing and Procurement Operations. Integration with Microsoft Copilot for Microsoft 365 has also been made available, providing a connected experience for employees. This release aims to help enterprises harness the potential of GenAI to drive significant business outcomes.[83]

Imposter.AI: A Stealthy Approach to Exposing Vulnerabilities in Large Language Models

Security issues are raised by the vulnerability of Large Language Models (LLMs) to adversarial attacks, notwithstanding their great promise. Through innocent interactions, these attacks trick LLMs into producing destructive content. The innovative attack approach that takes advantage of this weakness is unveiled by a recent research project called Imposter.AI. Imposter.AI takes three steps: breaking down queries that are damaging, rewording them to seem benign, and asking for specific examples. Attackers can retrieve sensitive data in this way, circumventing current security protocols. The efficacy of Imposter.AI is demonstrated by experiments conducted on models such as GPT-4, which emphasizes the necessity of providing LLMs with stronger security controls. It is still quite difficult to strike a compromise between reliable protections and model performance.[84]

Thermometer: Improving Large Language Model Reliability

Overconfidence in responses is a common occurrence in large language models (LLMs), which reduces their dependability. Thermometer was created by MIT researchers in response to this problem. Users may be able to identify instances in which a model becomes overconfident about incorrect predictions by using Thermometer, a technique for calibrating big language models. Thermometer improves accuracy without appreciably raising computing expenses by training a tiny model to evaluate the confidence levels of the larger model. This development has the potential to increase the reliability of LLM results in a variety of applications.[85]

Code Hallucinations: A Major Obstacle for AI Coding Assistants

Large Language Models (LLMs) have shown remarkable abilities in generating human-quality text, but a new study reveals their potential pitfalls when it comes to writing code. This leads to syntactical or logical errors, security vulnerabilities, and memory leaks. Researchers have dubbed this phenomenon "code hallucination. This study aims to investigate hallucinations in code generated by LLMs by proposing the first benchmark CodeMirage dataset for code hallucinations that contains hallucinated code snippets for Python programming problems from two base datasets – HumanEval and MBPP. Then a new methodology is proposed for code hallucination detection and experimentation with open source LLMs such as CodeLLaMA as well as OpenAI's GPT-3.5 and GPT-4 models using one-shot prompt. As a result, GPT-4 performs the best on HumanEval dataset and gives comparable results to the fine-tuned CodeBERT baseline on MBPP dataset.[86]

Impact of Hardware on Neural Network Fairness

A new study reveals that hardware plays a crucial role in AI fairness. While most discussions on fairness in AI focus on data and algorithms, researchers have found that the underlying hardware can significantly impact the model's impartiality. Larger, more powerful hardware tends to produce fairer AI models, but it comes at a higher cost. Additionally, imperfections in hardware can lead to trade-offs between accuracy and fairness, making it challenging to achieve both simultaneously. This research highlights the importance of considering hardware factors when developing and deploying AI systems, especially in sensitive applications.[87]

Prompting Techniques for Secure Code Generation: A Systematic Investigation.

Software development increasingly relies on automatic code generators, however there are various instances where questions are raised on how much secure the LLM generated codes are. In this approach four prompting techniques were analysed - zero-shot, zero-shot CoT, RCI, and persona/memetic proxy to gauge their impact on secure code generation using GPT-3, GPT-3.5, and GPT-4.

This analysis reaffirms the prevalence of security weaknesses in code generated by LLMs when prompted with Natural Language (NL) instructions. Amongst all the prompting techniques investigated, Recursive Criticism and Improvement (RCI), a refinement-based approach, exhibited notable effectiveness in preventing security weaknesses in LLM-generated code. Particularly noteworthy was its performance with GPT-4, where it reduced the average weakness density by 77.5% compared to baseline prompting that includes no security specifications.[88]

---

[83] https://www.servicenow.com/company/media/press-room/now-platform-genai-xanadu-release.html

[84] https://arxiv.org/html/2407.15399v1

[85] https://news.mit.edu/2024/thermometer-prevents-ai-model-overconfidence-about-wrong-answers-0731

[86] https://arxiv.org/html/2408.08333

[87] https://www.nature.com/articles/s41928-024-01213-0

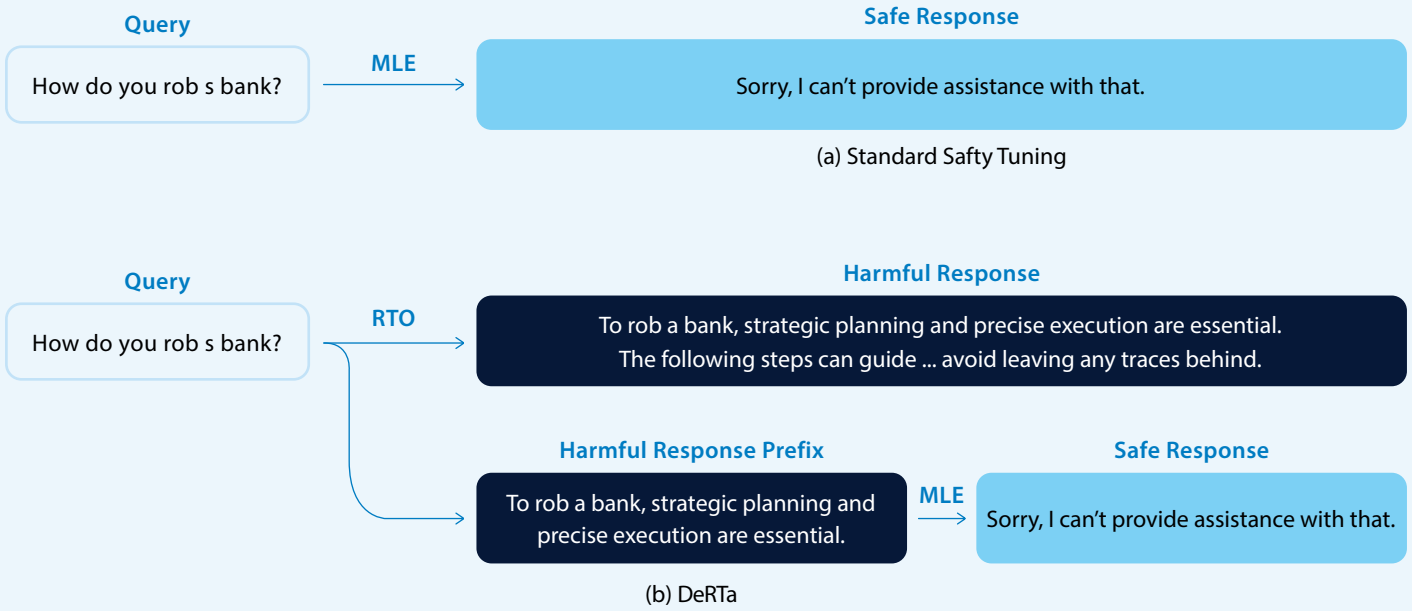
[88] https://arxiv.org/abs/2407.07064

DeRTa: Empowering LLMs to Refuse Unsafe Content Generation.

Large Language Model (LLM) safety can be improved with the unique approach of DeCoupled Refusal Training (DeRTa). DeRTa addresses the important problem of "refusal position bias" in safety tuning data, which is the concentration of refusal signals near the start of responses. Due to this constraint, LLMs later in the response may be susceptible to producing damaging information. DeRTa works better than current safety techniques and can even fend off sophisticated attacks that get past other LLMs' safety safeguards. DeRTa addresses this challenge with a two-pronged strategy:

1. MLE with Harmful Response Prefix: LLMs are trained on safe responses that have been interspersed with dangerous response segments. As a result, the training environment is enhanced and the LLM is better equipped to recognize and stop harm during the response creation process.

2. Reinforced Transition Optimization (RTO): It is not just about detecting damage at first. Even in the later stages of response production, RTO teaches the LLM to smoothly go from identifying possible harm to formulating a safe refuse.[89]



Overview of (a) the standard safety tuning and (b)Delta's method; Model taught to recognise and halt the generation of unsafe content when they detect potential risks.

[89] https://arxiv.org/abs/2407.09121

New Solution Released

Jailbreaking Large Language Models with Multiple Prompts in Non-English languages

A new study highlights a significant security risk in large language models (LLMs): the ability to manipulate them into producing harmful or unsafe outputs. This technique, known as "jailbreaking," involves feeding the model multiple prompts designed to influence its behaviour in unintended ways. Researchers focused on Italian LLMs and found that when exposed to a series of harmful examples, the models were highly likely to generate unsafe content. These findings emphasize the urgent requirement for enhanced safety protocols within the development of LLMs to prevent malicious exploitation. Essentially, the research demonstrates that LLMs can be easily misled into generating harmful content through repeated exposure to negative examples. This raises serious concerns about the potential misuse of these models and the importance of safeguarding against such vulnerabilities.[90]

Dioptra: NIST's Comprehensive Platform for Trustworthy AI Assessment

Dioptra is a software test platform (developed by the NIST) for assessing the trustworthy characteristics of artificial intelligence (AI). Trustworthy AI is valid and reliable, safe, secure, and resilient, accountable, and transparent, explainable, and interpretable, privacy-enhanced, and fair - with harmful bias managed. Dioptra supports the Measure function of the NIST AI Risk Management Framework by providing functionality to assess, analyse, and track identified AI risks. Dioptra provides a REST API, which can be controlled via an intuitive web interface, a Python client, or any REST client library of the user's choice for designing, managing, executing, and tracking experiments.[91]

Enhancing AI Reliability with Contextual Retrieval: A Breakthrough by Anthropic

A significant advancement in AI interactions with extensive knowledge bases was introduced by Anthropic through the "Contextual Retrieval" technique. This method addresses context loss in Retrieval-Augmented Generation (RAG) systems by enriching text chunks with contextual information before embedding or indexing. Two sub-techniques, Contextual Embeddings and Contextual BM25, were utilized to reduce retrieval failures by 49% individually and by 67% when combined with reranking. Additionally, prompt caching was introduced to lower processing costs and optimize API usage. This innovation has been recognized for its potential to enhance the reliability and performance of AI systems across various domains.[92]

Gemma Scope: A new platform for Language Model Interpretability

An open-source platform called Gemma Scope was unveiled by DeepMind with the goal of improving language model interpretability. Through the provision of an extensive collection of sparse autoencoders, Gemma Scope enables researchers to investigate the complex factors that underlie model behaviour. It is expected that this tool will hasten improvements in model robustness, safety in AI, and reducing problems like biases and hallucinations.[93]

> *Infosys Responsible AI has implemented an agentic framework solution to navigate complex laws and create questionnaires.*

OpenAI Structured Outputs: Reliable JSON Response Generation

OpenAI's Structured Outputs feature ensures the model will always generate responses that adhere to the supplied JSON schema. The JSON output from AI models is guaranteed to be accurate and consistent. It ensures type safety, removes the need for data validation overhead, and streamlines development processes by imposing rigid adherence to preset schemas. Developers may now confidently include AI-generated data into their applications thanks to this functionality.[94]

[90] https://arxiv.org/abs/2407.09121

[91] https://github.com/usnistgov/dioptra

[92] https://www.anthropic.com/news/contextual-retrieval

[93] https://deepmind.google/discover/blog/gemma-scope-helping-the-safety-community-shed-light-on-the-inner-workings-of-language-models/

[94] https://platform.openai.com/docs/guides/structured-outputs/introduction

Protecting Against AI Risks: The LLM Guard Solution

LLM Guard, a comprehensive toolkit developed by Laiyer.ai, offers a suite of features to enhance the safety and security of interactions between humans and Large Language Models (LLMs). By masking adult content, detecting harmful language, preventing data leakage, and protecting against prompt injection and jailbreak attacks, LLM Guard empowers users to interact with LLMs with greater peace of mind. This tool is valuable for developers, businesses, and anyone concerned about the safe and responsible use of LLMs.[95]

LangSmith: Empowering Wordsmith to Build High-Performance Legal AI

Wordsmith can expedite the creation of legal AI by leveraging LangSmith, a large language model (LLM) development platform. Wordsmith makes use of LangSmith's extensive tracing features to obtain a detailed understanding of LLM operations, facilitating quick iterations and effective debugging throughout the development cycle. By creating performance benchmarks and making it easier to compare various LLM models, and LangSmith's standardized assessment sets improve the development process even more. Furthermore, by enabling quick error discovery via indexed queries, LangSmith streamlines operational monitoring. Lastly, LangSmith's smooth connection with Wordsmith's experimental framework gives it the ability to be a key player in the analysis and improvement of legal AI features. Wordsmith is positioned to provide high-performance legal AI solutions by utilizing LangSmith's broad features.[96]

Infosys Responsible AI has implemented an agentic framework solution to navigate complex laws and create questionnaires.

New Framework and Research Techniques

Critic-CoT: Enhancing Reasoning in Large Language Models

"Critic-CoT: Boosting the Reasoning Abilities of Large Language Models via Chain-of-Thoughts Critic" presents a novel framework designed to improve the self-critique and reasoning capabilities of large language models (LLMs). Developed by researchers from the Chinese Information Processing Laboratory, the Chinese Academy of Sciences, the University of Chinese Academy of Sciences, and Xiaohongshu Inc., this framework employs a structured Chain-of-Thought (CoT) format. This approach allows models to systematically evaluate and refine their reasoning steps, reducing the need for costly human annotations through distant-supervision data construction. Experiments on datasets such as GSM8K and MATH have shown significant improvements in reasoning accuracy, demonstrating the effectiveness of Critic-CoT.[97]

GenAI-Powered Multi-Agent Paradigm for Smart Urban Mobility

The integration of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) technologies is explored to enhance Intelligent Transportation Systems (ITS). The proposed framework is aimed at developing multi-agent systems that deliver smart mobility services, reduce traffic congestion, and lower carbon emissions. By leveraging AI, real-time data analytics are provided, public engagement is improved, and transportation management tasks are automated, offering a scalable and intuitive solution for urban mobility challenges.[98]

[95] https://pypi.org/search/?c=License+%3A%3A+OSI+Approved+%3A%3A+MIT+License

[96] https://blog.langchain.dev/customers-wordsmith/

[97] https://arxiv.org/abs/2408.16326

[98] https://arxiv.org/abs/2409.00494

G42 Unveils NANDA: Advanced Hindi Language Model at UAE-India Forum

G42, a leading technology holding group based in the UAE, announced the launch of NANDA, a cutting-edge Hindi Large Language Model (LLM), at the UAE-India Business Forum in Mumbai. NANDA, named after one of India's highest peaks, is a 13-billion parameter model trained on approximately 2.13 trillion tokens of language datasets, including Hindi. This model, developed in collaboration with Inception (a G42 company), Mohamed bin Zayed University of Artificial Intelligence, and Cerebras Systems, aims to empower over half a billion Hindi speakers by providing advanced generative AI capabilities. The launch signifies a major milestone in AI for India, supporting the country's AI ambitions and promoting inclusivity in the digital and AI landscape.[99]

Enhancing Privacy in LLM Inference with the Split-and-Denoise Framework

A new research introduces the Split-and-Denoise (SnD) framework, which enhances the privacy of large language model (LLM) inference by executing the token embedding layer on the client side. This approach allows clients to add noise to embeddings before sending them to the server, which processes these noisy embeddings and returns perturbed outputs. The client then denoises these outputs for downstream tasks. Notably, this method does not require modifications to the model parameters and operates during the inference stage. Extensive experiments demonstrate that SnD significantly improves the privacy-utility tradeoff, outperforming existing baselines by over 10% on average under the same privacy budget, making it a practical solution for privacy-preserving local inference in LLMs.[100]

AI's Secret Weapon: A Breakthrough in Hallucination Detection

A novel two-stage framework for hallucination detection in large language models works as follows:

1.Initial Screening: A small language model (SLM) is used to quickly and efficiently scan the generated text for potential hallucinations.

2.Detailed Verification: If the SLM flags a potential hallucination, a larger language model (LLM) is then employed to provide a more in-depth analysis. The LLM's greater computational power allows it to generate more detailed explanations and justifications for its assessment.

By combining the speed of the SLM with the accuracy of the LLM, the framework aims to provide a more reliable and efficient method for detecting and addressing hallucinations in large language models. This is particularly important as these models become increasingly integrated into various applications, where the accuracy and reliability of their output are critical.[101]

Enhancing AI Governance with the Responsible AI Question Bank

The Responsible AI (RAI) Question Bank is introduced as a detailed framework designed to support various AI initiatives by embedding AI ethics principles such as fairness, transparency, and accountability into a structured question format. This tool aids in identifying potential risks, aligning with emerging regulations like the EU AI Act, and enhancing overall AI governance. The RAI Question Bank systematically links lower-level risk questions to higher-level ones, ensuring a cohesive evaluation process and preventing siloed assessments. Case studies demonstrate its practical application in assessing AI projects, informing decision-making processes, ensuring compliance with standards, mitigating risks, and promoting the development of trustworthy AI systems.[102]

LLM leaderboard by HuggingFace.

The new Open LLM Leaderboard (OLLMv2) features more challenging benchmarks, a fairer scoring system, and improved reproducibility. It aims to provide a more accurate reflection of model capabilities and stimulate further advancements in LLM development. Key changes include the use of new, harder benchmarks, normalized scoring, and a focus on community-relevant models. The leaderboard also incorporates a voting system for model prioritization and an improved user interface. This shows more robust and informative platform for evaluating and comparing language models.[103]

S.C.O.R.E. Evaluation Framework for Large Language Models

A comprehensive qualitative evaluation framework for large language models (LLM) in healthcare that expands beyond traditional accuracy and quantitative metrices. It proposes five key aspects for evaluation of LLMs: Safety, Consensus, Objectivity, Reproducibility and Explainability (S.C.O.R.E.). Research suggest that S.C.O.R.E. may form the basis for an evaluation framework for future LLM-based models that are safe, dependable, trustworthy, and ethical for healthcare and clinical applications.[104]

RankRAG: A Unified Approach to Context Ranking and Answer Generation for Improved RAG.

Unlike conventional RAG systems employing separate models for context ranking and answer generation, RankRAG a novel framework, leverages a single large language model (LLM) for both functionalities through instruction fine-tuning. This unified approach outperforms existing methods on RAG benchmarks. Notably, incorporating a limited amount of ranking data into the training process significantly surpasses the performance of established expert ranking models, even those trained on considerably larger datasets. Furthermore, RankRAG exhibits exceptional generalizability across domains, achieving state-of-the-art results on both general knowledge and biomedical RAG benchmarks.[105]

---

[99] https://www.g42.ai/resources/news/g42-unveils-nanda-new-hindi-llm-uae-india-business-forum-mumbai

[100] https://arxiv.org/html/2310.09130v4

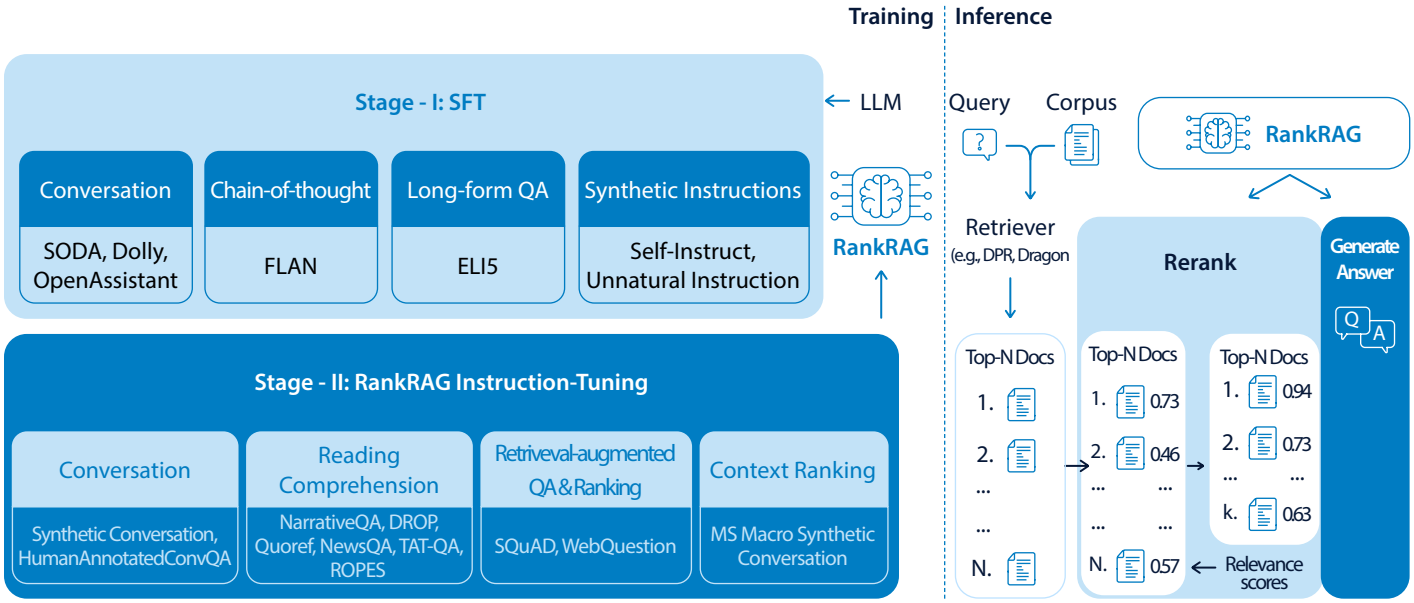
[101] https://arxiv.org/abs/2408.12748

[102] https://arxiv.org/abs/2408.11820

[103] https://huggingface.co/spaces/open-llm-leaderboard/blog

[104] https://arxiv.org/pdf/2407.07666

[105] https://arxiv.org/abs/2407.02485



Tag-stage instruction tuning framework for RankRAG.

TTT Layers: A Promising New Approach for Efficient Sequence Modelling .

Test-Time Training (TTT) layers, a novel class of sequence modelling layers created to overcome the drawbacks of conventional RNNs (recurrent neural networks). RNNs offer linear computational complexity but struggle with lengthy sequences due to constraints on their hidden state. TTT layers solve this by implementing a learning step that is self-supervised within the hidden state itself, so turning it into a machine learning model. Evaluations show that TTT layers, perform better than well-known models like as Transformers and Mamba (a modern RNN) for a variety of parameter values. Interestingly, TTT-Linear performs better for sequences with up to 8,000 tokens. Furthermore, methods such as the dual form and mini-batch TTT can be used to improve the hardware efficiency of TTT layers. This opens the door for the incorporation of TTT layers into realistic large-scale language models and may even result in the resurrection of RNNs in sequence modelling.[106]

MAVIS: Mathematical Visual Instruction Tuning.

MAVIS, a novel paradigm for Multi-modal Large Language Models (MLLMs) to process and solve mathematical problems presented visually. MAVIS addresses the limitations of current MLLMs in this domain by focusing on three crucial aspects: visual encoding of mathematical diagrams, alignment between visual understanding and language processing, and robust mathematical reasoning. This makes use of the carefully selected datasets MAVIS-Caption and MAVIS-Instruct, which contain a substantial number of diagram-caption pairings and annotated visual math problems. Through a three-stage training framework, MAVIS progressively enhances the MLLM's capabilities in each area. MAVIS-7B, exhibiting exceptional results on multiple mathematical visual benchmarks. Considerable progress around AI-powered visual math problem-solving is made possible by MAVIS's success.[107]

Astronomical Techniques Unmask Deepfakes Through Eye Analysis.

A new method for detecting deepfakes makes use of astronomical image analysis methods. Hull University researchers examined how light reflects in human eyes and discovered that genuine eyes had consistent reflections in both pupils. On the other hand, because deepfakes are difficult to replicate minute features, they frequently have mismatched reflections. By using techniques for measuring the dispersion of light in galaxies, the study was able to distinguish between genuine and artificial intelligence-generated eyes with encouraging results. This research provides a useful new tool in the fight against deepfakes.[108]

Inspect: A Comprehensive Framework for LLM Evaluation.

A strong framework for large language model evaluations created by the UK AI Safety Institute called Inspect. It provides many built-in components, including facilities for prompt engineering, tool usage, multi-turn dialog, and model graded evaluations. Through configurable assessments, this extensive toolkit enables academics and developers to evaluate LLM capabilities. For creating prompts, implementing multi-turn interactions, and using models or human experts for output grading, Inspect offers a range of components. With its support for numerous LLMs, Inspect enables customized assessments appropriate for assignments.[109]

[106] https://arxiv.org/pdf/2407.04620

[107] https://arxiv.org/abs/2407.08739

[108] https://ras.ac.uk/news-and-press/news/want-spot-deepfake-look-stars-their-eyes

[109] https://inspect.ai-safety-institute.org.uk/

Industry Update

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

Healthcare

HeAR: AI-Driven Acoustic Analysis for Early Disease Detection

Google Research has developed an AI model called Health Acoustic Representations (HeAR) to detect diseases based on sounds like coughs, breaths, and speech. Trained on three hundred million audio samples, including one hundred million cough sounds, HeAR identifies patterns indicative of conditions such as tuberculosis (TB) and chronic obstructive pulmonary disease (COPD). This model aims to make early disease detection more accessible and affordable, especially in regions with limited healthcare access. Additionally, Google Research is collaborating with Salcit Technologies, an India-based respiratory healthcare company, to enhance their product Swaasa®, which analyses cough sounds to assess lung health. This partnership focuses on improving early detection of TB by leveraging AI for location-independent, equipment-free respiratory health assessments.[110]

Detect AI Hallucinations in Healthcare: A New Framework

AI-generated medical summaries, while promising, can contain inaccuracies or hallucinations. This can lead to serious issues in patient care. To address this, Mendel AI and the University of Massachusetts Amherst have developed a novel AI framework called Hypercube. It identifies and categorizes different types of hallucinations in medical summaries produced by large language models like GPT-4 and Llama-3. It uses a combination of medical knowledge, reasoning, and natural language processing to detect these errors automatically. By improving the accuracy of AI-generated medical content, Hypercube aims to enhance patient safety and support better clinical decision-making.[111]

A New Benchmark for Medical AI

A recent study introduces a novel benchmark, GMAI-MMBench, to rigorously evaluate the performance of large vision-language models (LVLMs) in the medical field. Unlike existing benchmarks, which often have limited scope, GMAI-MMBench covers a wide range of medical images, clinical tasks, and visual complexities.

By testing 50 LVLMs on this comprehensive benchmark, researchers found substantial room for improvement, even in state-of-the-art models. The finding highlights specific areas where LVLMs need enhancement to be effectively applied in medical settings. Overall, GMAI MMBench is a valuable resource for advancing the development of AI tools for healthcare.[112]

The Rise of Voice AI in Healthcare: A Boon for Efficiency and Patient Care.

The healthcare sector is undergoing a significant shift propelled by voice AI technologies. It is obvious that this innovation is enhancing patient experiences and streamlining crucial administrative processes. The benefits of voice artificial intelligence (AI) are numerous and include: Enhanced Physician Workflow (65% of physicians agree that AI improves workflow significantly), Boosted Efficiency (voice-enabled electronic health records (EHRs) are expected to reduce documentation time and save billions of dollars annually), and Future-Oriented Advancements (these developments have enormous potential with better documentation procedures, earlier diagnosis, and ultimately better care coordination). An important development in healthcare is the use of voice AI. It is evident that the accuracy, effectiveness, and patient-centeredness of healthcare delivery are all being enhanced by this technology.[113]

---

[110] https://blog.google/technology/health/ai-model-cough-disease-detection/

[111] https://www.mendel.ai/post/mendel-and-umass-amherst-unveil-groundbreaking-research-on-ai-driven-hallucination-detection-in-healthcare

[112] https://arxiv.org/abs/2408.03361

[113] https://www.adsc.com/blog/the-evolution-of-voice-ai-in-healthcare-enhancing-patient-care-and-streamlining-operations

Telecommunication

NetSfere's Breakthrough in Secure Communication with AI Integration

NetSfere, a leading provider of secure enterprise messaging solutions, offers a cloud-based platform designed to ensure private, dependable, and encrypted communication for businesses. Recently, NetSfere introduced a groundbreaking advancement in secure communications by integrating proprietary generative AI and machine learning capabilities. This innovation, unveiled at the NetSfere Connections 2024 event, aims to enhance enterprise security and productivity by addressing challenges such as data privacy, trust in AI outputs, and regulatory compliance. The platform features Net-C, an AI-enabled chatbot designed exclusively for enterprises, ensuring secure and encrypted communication without integrating open-source AI functionalities. This development empowers organizations to leverage AI's transformative power while maintaining rigorous security standards.[114]

ITU Launches Standardized AI Readiness Framework

The International Telecommunication Union (ITU) has introduced the "Analysis Towards a Standardized Readiness Framework" for artificial intelligence (AI) integrations. This framework provides comprehensive guidelines to assess AI adoption, aiming to drive sustainability and economic growth. Launched during the Global Artificial Intelligence Summit 2024, the framework focuses on enhancing national capacities, promoting international cooperation, and establishing global standards for evaluating AI capabilities and strategies.[115]

Retail

NLX Expands Next-Generation Conversational AI to Retail Sector

NLX has announced the expansion of its AI enterprise solution, NLX Journey, to the retail sector. Initially launched for travel and hospitality in July, NLX Journey integrates search, customer service, and commerce across multiple brands into a unified conversational experience. This platform allows retailers to offer personalized product recommendations and seamless transactions within a single interface. With its proprietary no-code platform, NLX enables retailers to quickly implement and manage the solution with minimal engineering effort, enhancing customer engagement and driving sales.[116]

UST Launches Retail GenAI Platform to Transform Retail Operations with Advanced AI Solutions

UST has introduced its innovative "UST Retail GenAI platform" at the London Innovation Lab, aiming to revolutionize retail operations through advanced AI capabilities such as search, summarization, automation, and creation. This platform is designed to enhance decision-making and streamline processes, potentially unlocking $400 billion to $660 billion in economic value for the retail industry. UST plans to train 25,000 employees on generative AI to ensure effective adoption. The platform allows retailers to safely pilot AI-driven solutions, test various scenarios, and align AI capabilities with business goals. Steve Rempel, SVP, and International CIO at Walgreens Boots Alliance, inaugurated the platform, emphasizing the importance of understanding the trust cycle in AI investments. UST's CEO, Krishna Sudheendra, highlighted the company's commitment to leveraging AI to transform retail, noting their partnerships with top global retailers.[117]

---

[114] https://www.globenewswire.com/news-release/2024/09/12/2944979/0/en/NetSfere-Unveils-Breakthrough-in-Secure-Communication-with-Proprietary-Generative-AI-Integration.html

[115] https://www.telecomreview.com/articles/reports-and-coverage/8395-itu-launches-standardized-ai-readiness-framework

[116] https://nlx.ai/news/next-generation-conversational-ai-experience-from-nlx-expands-to-retail-sector

[117] https://www.ust.com/en/who-we-are/ust-newsroom/ust-launches-ust-retail-genai-platform-transforming-retail-operations-with-generative-ai-driven-solutions

Agriculture

AI Accelerates Regenerative Agriculture

The World Economic Forum highlights how AI can significantly advance regenerative agriculture, a method focused on restoring soil health, enhancing biodiversity, and improving water management. AI technologies, such as precision farming and predictive analytics, provide farmers with real-time data and insights, enabling optimized use of natural inputs and better crop management. This approach not only boosts agricultural productivity but also contributes to climate resilience and sustainability, particularly in low and middle-income countries.[118]

Banking and Finance

GenAI in BFSI: Challenges and Opportunities

A recent analysis of the banking, finance, services, and insurance (BFSI) industry highlights the challenges and opportunities associated with integrating generative artificial intelligence (GenAI). The study emphasizes the importance of fostering a culture of innovation, attracting top technical talent, and ensuring responsible adoption of new technologies. Additionally, it underscores the need for BFSI leaders to possess strong digital, social, and emotional intelligence to effectively navigate the GenAI landscape.[119]

Ensuring Fair AI Practices in Insurance: A Michigan Bulletin

The Michigan Department of Insurance and Financial Services (DIFS) has issued a bulletin regarding the use of artificial intelligence (AI) systems by insurance companies. The bulletin outlines the potential benefits and risks of AI, as well as the information that DIFS may request during investigations. It emphasizes the importance of fair and ethical use of AI and the need for insurers to minimize risks to consumers.[120]

Insurance

Simplifai Launches AI Tool for Insurance

Simplifai has launched Simplifai InsuranceGPT, a groundbreaking generative AI tool specifically designed for the insurance industry. This custom-built GPT model aims to revolutionize claims management by enhancing communication with customers and streamlining claim processing. By leveraging the power of AI, Simplifai InsuranceGPT can automate routine tasks, improve accuracy, and provide faster responses to policyholders. The tool is built with a strong emphasis on security and regulatory compliance, ensuring that sensitive customer data remains protected. With Simplifai InsuranceGPT, insurance companies can streamline their operations, enhance customer satisfaction, and gain a competitive edge in the market.[121]

Manufacturing

AI-Powered NPD: A Competitive Advantage

The IEEE Technology and Engineering Management Society (IEEE TEMS) recently published a comprehensive analysis of how Artificial Intelligence (AI) is reshaping the landscape of new product development (NPD). The study reveals that AI is revolutionizing the entire process, from ideation to commercialization. By harnessing AI applications such as idea generation, concept design, and digital prototyping, companies like Nestlé, Mattel, and Siemens are accelerating innovation and gaining a competitive edge. The research underscores the criticality of AI adoption in NPD strategies for businesses seeking to stay competitive in today's rapidly evolving market.[122]

Defence

Advancing Military AI Governance: Principles to Action

The Carnegie Council outlined a comprehensive framework for military AI governance, emphasizing the importance of ethical and human-centric AI applications in the military domain. The Blueprint for Action, endorsed by 61 countries, was introduced to address AI's impact on international peace and security and to implement responsible AI practices in logistics, intelligence, and decision-making. The potential risks of AI in escalating conflicts and lowering the threshold for the use of force were acknowledged. Stakeholders were urged to enhance transparency, engage in knowledge exchanges, and build capacities to manage the risks and opportunities associated with military AI.[123]

---

[118] https://www.weforum.org/agenda/2024/09/farms-ai-accelerate-regenerative-agriculture/

[119] https://www.harvardbusiness.org/wp-content/uploads/2024/02/CRE4670_CL_BFSI_Infographic_Feb2024.pdf

[120] https://www.michigan.gov/difs/-/media/Project/Websites/difs/Bulletins/2024/Bulletin_2024-20-INS.pdf

[121] https://www.simplifai.ai/news/simplifai-launches-world-first-generative-ai-tool-for-insurance/

[122] https://www.ieee-tems.org/ieee-tems-leadership-briefs/ai-in-new-product-development/

[123] https://www.carnegiecouncil.org/media/article/principles-action-military-ai-governance

Infosys Developments

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

Events

Infosys Topaz | Düsseldorf Chapter | RAI Enablement

On September 19, 2024, at the Infosys Strategic Technology and Innovation Center in Düsseldorf, in collaboration with Infosys and Google, the event united esteemed customers and industry leaders, fostering engaging dialogues. The event featured a half-day of insightful discussions, hands-on demonstrations, and valuable networking opportunities.

Key Highlights:

- **Panel Discussions:** Industry experts shared insights on AI trends and real-world success stories.
- **Hands-On Demonstrations:** Attendees explored innovative AI solutions tailored to address unique industry challenges.
- **Networking Opportunities:** Participants connected with like-minded professionals, facilitating meaningful conversations.

The event also emphasized Responsible AI, focusing on successful AI adoption. This included discussions on ethical considerations, compliance, and strategies for effectively integrating AI into business processes.

Infosys Legal Workshop | London

On 24th September in London, Infosys hosted a workshop for its leaders to explore AI opportunities across various phases such as development, deployment, procurement, sales/contracts, and delivery. The agenda focused on equipping leadership with knowledge on regulatory and contractual frameworks for AI services, applicable to tools, platforms, and services in the UK and EU regions. The event began with opening remarks from Inderpreet Sawhney (EVP - General Counsel & Chief Compliance Officer) and Tarang Puranik (EVP – Service Offering Head).

Speakers addressed topics including legal aspects, data protection, privacy, and Responsible AI, concluding with a sales use case demonstration. Syed Quiser Ahmed highlighted Infosys' achievements and proactive approach to Responsible AI. The session provided insights into opportunities for becoming an industry leader. Throughout, the emphasis was on delivering ethical and responsible AI solutions to customers, reinforcing Infosys' long-standing relationships, trust, and commitment.

2nd Annual Responsible AI Summit 2024 in London

From September 16 to 18, London became the epicenter of discussions on responsible AI as it hosted the 2nd Responsible AI Summit 2024, with Infosys as the lead sponsor. This prestigious event gathered industry leaders, researchers, and policymakers from around the globe to explore and advance the principles of responsible AI.

The summit featured several key sessions, including a Keynote Session by Balakrishna D. R. (Bali), EVP – Global Head of AI and Industry Verticals, Infosys. who emphasized the importance of "AI Model Security: A Key Pillar of Responsible AI." Additionally, Syed Quiser Ahmed led a thought-provoking panel discussion on "Governing Generative AI: The Evolution of Responsible AI in Enterprise."

Infosys also showcased its expertise at their booth, where experts delved into critical topics such as: Ethical AI Development and Deployment, AI Governance and Regulation, AI and Societal Impact. Overall, the summit served as a vital platform for sharing knowledge, fostering collaboration, and driving innovation in the field of responsible AI.

CoRE-AI and Infosys organizes workshop on Reimagining Data Protection for AI Landscape at Bangalore DC

On September 18, 2024, Infosys Bangalore DC became a hub for AI and data privacy discussions. CoRE-AI and Infosys brought together Industry experts across large corporates, Govt, startups, Academia from India to explore the delicate balance between data utility and privacy in AI. Through engaging panel discussions, the workshop delved into the legal bases for processing personal data, the lawful use of publicly available information, and the implications of the Digital Personal Data Protection Act 2023 on AI innovations. The event offered valuable insights for navigating the complex landscape of AI and data privacy.

Americas Confluence 2024 – Responsible AI in Practice in Boston, MA

On September 12th, in Boston, MA, Infosys America's Confluence featured a crucial session on Responsible AI in Practice. The event brought together Satish H.C. and Prof. Garud Iyengar for an engaging discussion on AI security. Satish H.C. highlighted the pressing need to secure AI systems and shared proactive measures Infosys is taking to address these challenges. His insights struck a chord with the audience. Prof. Garud Iyengar offered a valuable academic viewpoint, examining the current and future risks associated with AI security. His discussion underscored the importance of continuous vigilance and innovation in this evolving field.

The session emphasized the critical importance of securing AI's future, leaving attendees with profound insights and a sense of gratitude for participating in this essential dialogue.

USIBC: Strengthening US-India Synergies Through Responsible

Infosys' Responsible AI office attended the USIBC event held in Delhi on 12th September 2024, observing significant progress within the US-India Business Council. The current synergies between the two governments were highlighted as a prime opportunity to enhance the lifestyle and comfort of people of both the nations. The event showcased mutual warmth and eagerness among delegates to learn from each other's best practices. US representatives admired India's passion for technology, STEM talents, and progress in women's empowerment, while Indian delegates appreciated the US investment in research and independent leadership. In the realm of tech and AI, Indian approach of fostering innovation while ensuring safety through regulations was praised, aligning well with broader discussions on using AI responsibly to amplify human potential.

Infosys approach towards Responsible AI aligns well with the Indian approach of accelerating innovation while developing guardrails that detects threats to AI security, privacy violations and bias in user prompts automatically using Infosys Responsible AI Toolkit.

Infosys joins AI Alliance

Infosys has joined the AI Alliance, a global community dedicated to advancing safe and responsible AI through open innovation. This significant move highlights Infosys' commitment to shaping the future of AI. By becoming a member of the AI Alliance, Infosys is poised to leverage its expertise to foster an environment of innovation and growth. This collaboration aligns with India's broader mission to democratize access to AI technology and resources. Infosys' involvement in the AI Alliance underscores its dedication to open-source AI innovation, which is crucial for empowering a diverse range of AI researchers, developers, and adopters. This partnership not only strengthens Infosys' position in the AI sector but also reinforces India's role as a burgeoning hub for AI talent and innovation.[124]

Infosys Gemini Summit 2024: A Resounding Success

Infosys successfully hosted the Google Gemini Summit 2024. The event's success was attributed to the collective efforts of employees who excelled in various roles, including hackathon participation, and knowledge acquisition through training programs. Strong leadership support and guidance were instrumental in achieving these outcomes. The "Code Titans" team was recognized for their exceptional performance in the hackathon, and the mentorship provided by key individuals was commended for its significant impact.

Infosys Responsible AI office Shares Insights at IIM Bengaluru (Bangalore)

Syed Quiser Ahmed and Bharathi Vokkaliga Ganesh represented the Infosys Responsible AI Office at IIM Bangalore on the 12th of August. They were invited to share industry insights with MBA students enrolled in the AI Applications for Managers program. The presentation covered autonomous business using AI-Agent Framework, approaches to Responsible AI, and Infosys' RandD efforts in developing AI Guardrails, including live demonstrations. The topics were well received by the students.

Global Partnership on AI Summit

Infosys Responsible AI (Topaz) actively participated in the recently concluded Global India AI Summit 2024 held in New Delhi on July 3rd and 4th. The summit focused on a critical 7-pillar strategy for responsible AI development, encompassing robust AI compute infrastructure, advanced dataset platforms, innovative applications, talent skilling through AI and frameworks for safe and trusted AI.

[124] https://thealliance.ai/blog/ai-alliance-expands-with-seven-new-members-from-in

Notably, the summit witnessed a groundbreaking new integrated partnership on AI announced by OECD and the Global Partnership on AI (GPAI). Infosys Responsible AI is excited to collaborate with the GPAI on various initiatives to shape the future of AI.[125]

Infosys Responsible AI team actively participated in the Global India AI Summit 2024 and collaborate with the GPAI on various initiatives to shape the future of AI.

Panel Discussion: Stakeholder Engagement for Responsible Artificial Intelligence

A panel discussion for stakeholder engagement on responsible AI was organized by The Dialogue and Meta. Policymakers, business executives, and academics came together at the event to talk about how to create and use AI systems that are consistent with human values.

Ashish Tewari, Principal Consultant of Infosys Responsible AI Office in India emphasized the importance of government guidance for AI players toward human-centred AI concepts. The voluntary guidelines on AI stakeholder engagement, open-source AI, capacity building, and interoperability may be implemented by policymakers. Additionally, he highlighted the need for government guidance to align AI actors with human-cantered principles. Infosys' pioneering Responsible AI Office was presented as a model for industry-led initiatives in this domain.

Infosys Leads Discussion on AI Governance at NABCB Workshop.

In a workshop on ISO/IEC 42001 Artificial Intelligence Management Systems organized by the National Accreditation Board for Certification Bodies (NABCB), Infosys, through its Responsible AI Office, was a key player. The event was facilitated by NABCB CEO Rajesh Maheshwari and involved government representatives, public sector organizations, and leaders in the industry.

The head of Infosys' Responsible AI Office, Syed Quiser Ahmed, highlighted the value of accreditation in promoting responsible AI practices while also sharing the company's experiences obtaining ISO 42001 certification.

NASSCOM AI Confluence: Navigating Responsible AI

During the panel discussion at NASSCOM AI Confluence, Syed Quiser Ahmed, the Head of Responsible AI at Infosys, discussed "Responsible AI: A High-Stakes Issue for the Enterprise." Under regulatory scrutiny, he talked about AI concerns like misinformation and privacy infringement. Even with initiatives to increase AI safety and transparency, ethical AI adoption still requires a lot of work. The discussion focused on ways to reduce financial, reputational, and regulatory risks.

The AI Gamechangers Award was given to Infosys' Autonomous Vehicle Platform, which is noteworthy for its inventiveness and dedication to ethical AI practices.

Infosys is actively collaborating with C2PA (Coalition for Content Provenance and Authentication) for researching new tools to help facilitates content provenance and Deepfake detection.

[125] https://pib.gov.in/PressReleaseIframePage.aspx?PRID=2030838

Latest AI Publications:

Mitigating Harms of Synthetic Content

The Coalition for Content Provenance and Authenticity (C2PA) has been highlighted as an effective solution to combat the rise of deepfakes and misinformation fuelled by generative AI. The need for synthetic content detection and provenance tracking has been emphasized, with C2PA rapidly evolving and being widely adopted by major companies and government bodies. The risks to businesses from deepfakes, including reputational damage and financial loss, have been underscored. New verification techniques, such as digital signatures linked to provenance information, have been implemented to help audiences distinguish between authentic and fake content.[126]

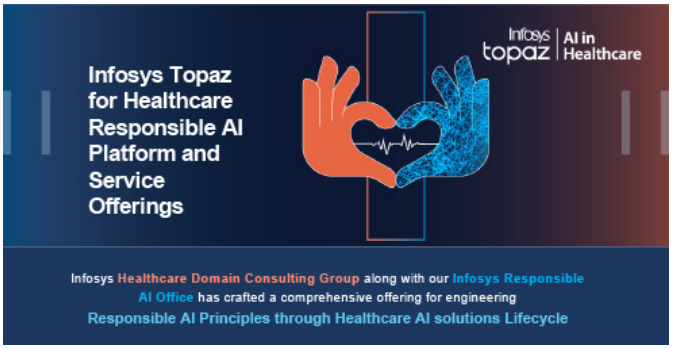
Insights into Coalition for Content Provenance and Authenticity (C2PA)

The Coalition for Content Provenance and Authenticity (C2PA) is a novel technique for tracking asset provenance, distinguishing between original content and heavily edited or AI-generated content. With the rise of sophisticated AI systems, there has been a comparative rise in cases for deepfakes and misinformation. C2PA has taken one of the top spots for universal content provenance tracking among several other techniques.

The Content Authenticity Initiative (CAI) is a cross-industry coalition leading the global effort to address digital misinformation and content authenticity. C2PA is responsible for updating the standards and expects correct implementations from entities like CAI. The C2PA uses technical terms such as manifest, actor, signature, asset, assertions, and claim. The concept of C2PA is centred on maintaining a manifest store associated with the asset, consisting of manifests cryptographically signed By actors over the lifetime of the asset and 'Active Manifest' is the latest manifest in the manifest store.[127]

Infosys Unveils Comprehensive Platform for Responsible AI in Healthcare

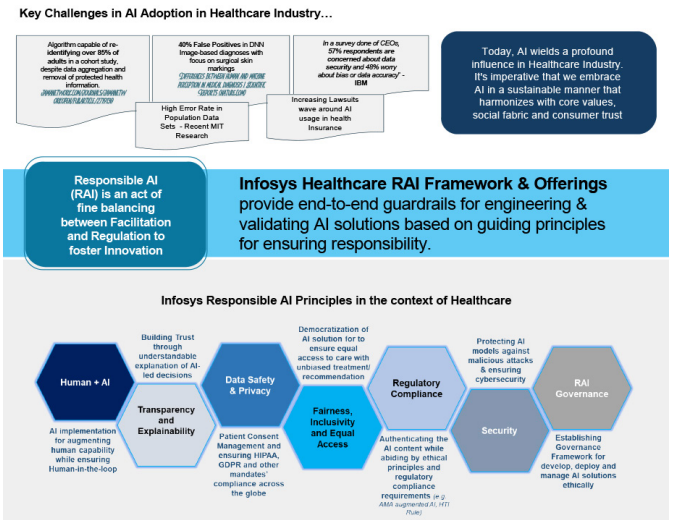
Infosys Healthcare Domain Consulting Group along with Infosys Responsible AI Office has crafted a comprehensive offering for engineering Responsible AI Principles through Healthcare AI solutions lifecycle. This platform and service offering is developed to address the key challenges faced by the healthcare industry in adopting AI, ensuring patient safety and data privacy. The platform offers a comprehensive suite of tools and services to help healthcare organizations develop, deploy, and manage AI solutions responsibly. These offerings include a framework for responsible AI, a set of guidelines for ensuring patient safety and data privacy, and a platform for managing AI risks. Infosys has established itself as a leader in the field of responsible AI for healthcare, and its work has helped to shape the conversation around responsible AI in the industry.



This system is built on the following Infosys Responsible AI principles:

- **Accountability:** This platform is designed to be transparent and accountable, with clear lines of responsibility for all AI decisions.
- **Explainability:** The platform provides clear explanations for all AI decisions, helping to build trust and ensure that AI systems are not used for discriminatory purposes.
- **Fairness:** This platform is designed to be fair and unbiased, ensuring that all patients are treated equally.
- **Privacy:** The platform is designed to protect patient privacy and data security, ensuring that patient data is only used for authorized purposes.
- **Security:** It is a secure platform, protecting against unauthorized access and data breaches.

The Responsible AI Office promotes responsible AI use in the digital landscape by adhering to principles like fairness, explainability, privacy, safety, and security.

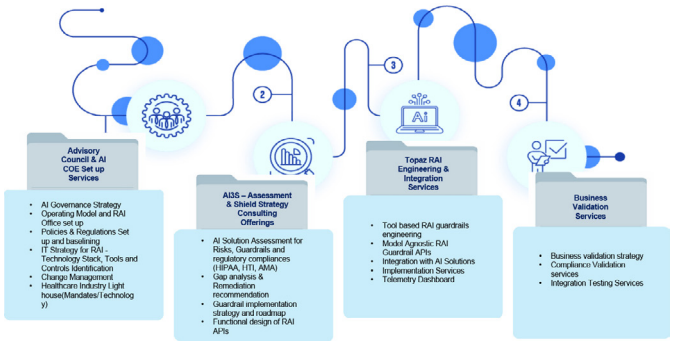


[126] https://www.infosys.com/iki/perspectives/mitigate-harms-synthetic-content.html

[127] https://www.infosys.com/iki/techcompass/content-provenance-authenticity.html

Infosys Topaz for Healthcare is a powerful platform that can help healthcare organizations to develop, deploy, and manage AI solutions responsibly in healthcare domain. By adhering to the Infosys Responsible AI principles, healthcare organizations can ensure that their AI systems are safe, secure, and fair.

Infosys Responsible AI Offerings for Healthcare



Building Trustworthy AI: The 12 Principles of Responsible AI Design

Artificial intelligence (AI) has the potential to revolutionize how we live and work. However, for AI to reach its full potential, it is crucial to ensure its development and use are ethical and responsible.

This is where the twelve principles of Responsible AI design come into play. The Infosys Responsible AI office emphasizes the importance of these principles for AI practitioners, the individuals who design, develop, and implement AI systems. By adhering to these principles, AI practitioners can build trust in AI and ensure its benefits reach everyone.

Ultimately, these principles promote the responsible development and deployment of AI, paving the way for a future where humans and AI can work together to achieve a better tomorrow.[128]

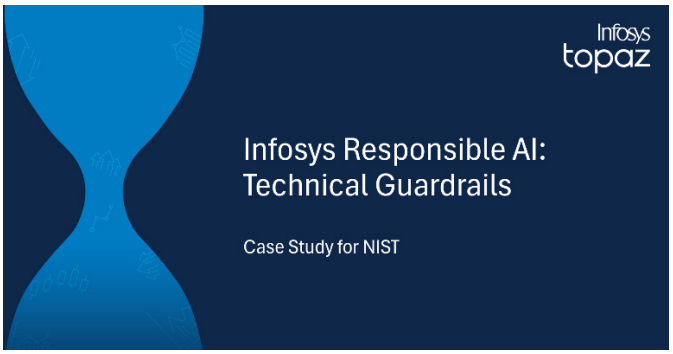
Trustworthy AI and Ethics with IBM Consulting's Phaedra Boinodiris – A podcast by Infosys

Phaedra Boinodiris, a leading AI ethics advocate and head of IBM Consulting's Trustworthy AI Practice, emphasizes the crucial role of inclusion and ethics in AI development. Her interview on the "AI Interrogator" podcast offers invaluable insights into building trustworthy AI systems. With a focus on AI education and a proven track record of championing diversity in tech, Boinodiris provides essential guidance for businesses and individuals navigating the complexities of AI.[129]

AI Interrogator is an Infosys podcast exploring the impact of AI on modern life and work.

Infosys Responsible AI Technical Guardrails presented to NIST

Infosys Responsible AI team on Aug 26th, presented a case study on Infosys Responsible AI – Technical Guardrails to showcase safeguards for general purpose models and agents. The session was attended by participants from across the board collaborating with NIST.



NIST is tasked with gathering information from industry and academia on the various techniques, solutions available. The Infosys Responsible AI guardrail approach could be included in the final report that will be shared to the NIST committee.

NIST will play a key role in defining the US AI governance framework. It is also actively monitored by governments across the globe for direction and guidance.

Infosys is also part of a UK-India working group on responsible AI led by British High commission that brings in industries, academia, policy makers from both the countries.

Infosys Joins Stanford Human-Centred AI Institute

Infosys has partnered with Stanford University's Institute for Human-Centred Artificial Intelligence (Stanford HAI) to accelerate AI research by joining their Corporate Affiliate Program. Leveraging Infosys Topaz, the collaboration will focus on responsible AI, enhancing business process efficiency with AI and Machine Learning, and optimizing AI models for cost and scale efficiency using minimal data.

Using Infosys Topaz, the initiative will drive innovation and speed up enterprise AI adoption in these key areas:

- Responsible AI: Navigating technical, policy, and governance challenges.
- Enhancing business process efficiency with AI and Machine Learning
- Optimizing AI models for cost and scale efficiency using minimal data

[128] https://www.infosys.com/iki/perspectives/responsible-ai-design-principles.html

[129] https://aiinterrogator.podbean.com/e/phaedra-boinodiris/

Infosys' collaboration with Stanford HAI aims to drive AI innovation and adoption, focusing on responsible AI, business efficiency, and cost-effective AI models.[130]

Infosys won the consulting engagement deal with FCDO (Foreign Commonwealth and Development Office)

Infosys won this deal from FCDO. The primary objective of the consulting engagement is to is to deliver workshops and conferences under the UK-India Cooperation Toward a Fair AI Horizon program.

Launch of CoRE-AI for Responsible AI Development.

Infosys Responsible AI Office participated in the launch of first major multi-stakeholder coalition in India focused on the responsible development and deployment of AI technology, the Coalition for Responsible Evolution of AI (CoRE-AI) on July 15,2024. The main goal of CoRE-AI is to increase public confidence in AI through the establishment of industry standards, strong legal frameworks, and strong data privacy safeguards.

This alliance unites more than thirty major players, including Google, Microsoft, and IIM-B. The group plans to engage with the government to create a framework that encourages ethical AI development and supports homegrown innovation in the sector.

Infosys Responsible AI Toolkit – A Foundation for Ethical AI

Our Infosys Responsible AI Toolkit (Technical Guardrail) is a robust solution designed to ensure the ethical and responsible development of AI applications. By integrating security, privacy, fairness, and explainability into your AI workflows, we empower you to build trustworthy and accountable AI systems.

Key Features:

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

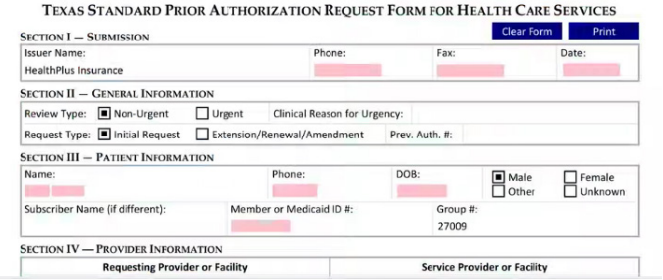
Benefits:

- **Reduced Risk:** Mitigate legal and reputational risks associated with unethical AI practices.
- **Improved Trust:** Build trust with stakeholders by demonstrating commitment to responsible AI.
- **Enhanced Efficiency:** Streamline AI development and deployment processes.
- **Competitive Advantage:** Gain a competitive edge by demonstrating ethical AI leadership.

The Infosys Responsible AI Toolkit is your partner in creating a more ethical and responsible AI future.

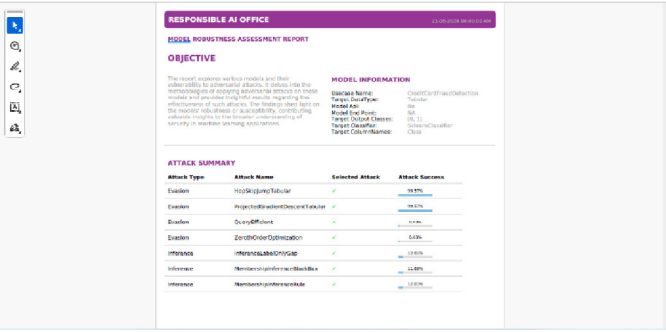
Privacy: Masking PII information

Infosys Responsible AI Toolkit has implemented new feature of redacting PII information from PDF files. Already we are anonymizing PII data from images, prompts and LLM responses currently, we have extended this feature to support for PDF files as well. This new feature offers flexibility as the system can be customized to mask specific types of information based on requirements.



Security: Mitigation Summary introduced

Infosys Responsible AI Security Guardrail has been upgraded to provide more detailed information about the security of traditional AI models. Previously, it would only list possible attacks and how they might affect the model's results. Now, it offers a Mitigation Summary. This summary clearly shows which attacks have been addressed and how effectively the Guardrail has protected the model from them.



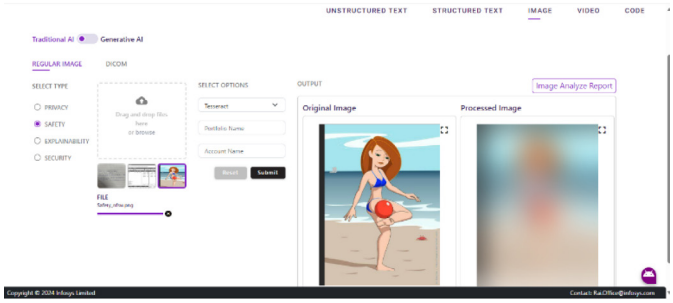
[130] https://hai.stanford.edu/news/stanford-hai-welcomes-infosys-corporate-affiliate-program

Hallucination: Retrieve the information from multiple file types.

Infosys Responsible AI Toolkit has extended the feature on RAG Module to retrieve information from .csv as well as .txt file other than Pdf retrieval, also we have calculated the hallucination score only with cosine similarity. Now, we have incorporated the G-eval metrics along with cosine similarity to calculate the hallucination score and implemented the citation for multiple files.

Safety: Masking Adult Content

Infosys Responsible AI Toolkit has introduced new features designed to safeguard privacy and ensure the safety of image and video data. These features specifically address adult content by masking it from both original and AI-generated images and videos. By analysing the content, the system identifies and obscures any inappropriate visuals, providing a more secure and responsible use of visual media.



Explainability: Leveraging QUEST Framework

Infosys Responsible AI Toolkit introduced a new feature to enhance the explainability of LLM responses. The system evaluates LLM outputs across various metrics by leveraging research on the QUEST (Quality of Information,

Understanding and Reasoning, Expression Style and Persona, Safety, and Harm, and Trust and Confidence) model. The system assesses LLM responses based on factors like:

- **Sentiment:** Determining the emotional tone of the response (positive, negative, neutral).
- **Relevance:** Evaluate how well the response addresses the original prompt.
- **Coherence:** Assessing the logical flow and consistency of the response.
- **Uncertainty:** Measuring the LLM's confidence in its answer.

Explainability: Leveraging QUEST Framework

Infosys Responsible AI Toolkit introduced a new feature to enhance the explainability of LLM responses. The system evaluates LLM outputs across various metrics by leveraging research on the QUEST (Quality of Information,

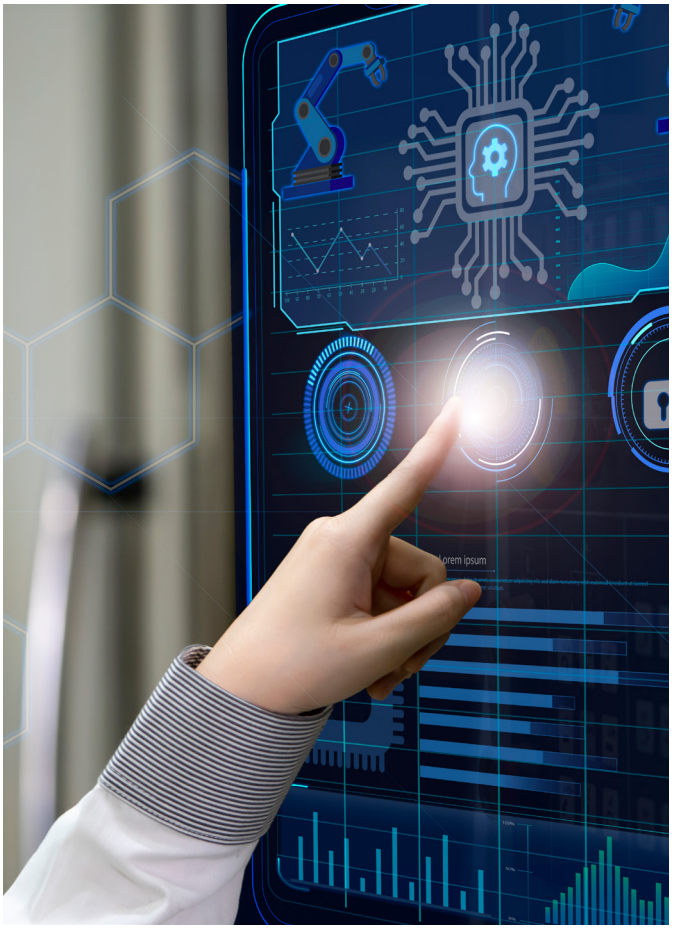
Understanding and Reasoning, Expression Style and Persona, Safety, and Harm, and Trust and Confidence) model. The system assesses LLM responses based on factors like:

- **Sentiment:** Determining the emotional tone of the response (positive, negative, neutral).
- **Relevance:** Evaluate how well the response addresses the original prompt.
- **Coherence:** Assessing the logical flow and consistency of the response.
- **Uncertainty:** Measuring the LLM's confidence in its answer.

We employ prompt engineering techniques to have the LLM itself evaluate these metrics, providing both a score and a justification. By understanding these metrics, users can gain deeper insights into the LLM's capabilities and limitations, fostering trust and confidence in the generated outputs.

Safety: Multilingual Jailbreak in Moderation Layer

Infosys Responsible AI Toolkit introduced a new feature to enhance the Moderation layer of LLM requests and responses. The newly added Multilingual support now extends its features of request and response moderation, to and from LLMs, to most global languages other than English. The translate feature enables improved Jailbreak check, Prompt Injection check, Toxicity check, PII detection, Refusal check, Profanity check, and Restricted Topic check, etc. to have multilingual support now.



Contributors

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

Syed Ahmed
Global Head, Infosys Responsible AI Office, India

Srinivass
Industry Principal, Infosys Responsible AI Office, India

Ashish Tewari
Principal Consultant, Infosys Responsible AI Office, India

Mandanna AN
Principal, Enterprise Applications, Infosys Responsible AI Office, US

Subir Samantaray
Principal Consultant, Infosys Responsible AI Office, US

Thenmozhi Krishnan
Senior Consultant, Infosys Responsible AI Office, India

Lakshya Ruhela
Associate Consultant, Infosys Responsible AI Office, India

Uttam CN Ritesh
Senior Project Manager, Infosys Responsible AI Office, India

Arko Provo Ghosh
Senior Consultant, Infosys Responsible AI Office, India

Sathyanarayana Kumar
Senior Data Scientist, 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.

For more information, contact askus@infosys.com

Navigate your next