Ethical and regulatory challenges of large language models in medicine

Article *in* The Lancet Digital Health · April 2024

DOI: 10.1016/S2589-7500(24)00061-X

CITATIONS
19

READS
85

11 authors, including:

Jasmine Ong
Singapore General Hospital
29 PUBLICATIONS **212** CITATIONS
SEE PROFILE

Shelley Yin-Hsi Chang
Chang Gung Memorial Hospital
22 PUBLICATIONS **116** CITATIONS
SEE PROFILE

William Wasswa
Mbarara University of Science & Technology (MUST)
29 PUBLICATIONS **489** CITATIONS
SEE PROFILE

Nan Liu
Duke-NUS Medical School
249 PUBLICATIONS **4,050** CITATIONS
SEE PROFILE

Ethical and regulatory challenges of large language models in medicine

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang*, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, Daniel Shu Wei Ting*

With the rapid growth of interest in and use of large language models (LLMs) across various industries, we are facing some crucial and profound ethical concerns, especially in the medical field. The unique technical architecture and purported emergent abilities of LLMs differentiate them substantially from other artificial intelligence (AI) models and natural language processing techniques used, necessitating a nuanced understanding of LLM ethics. In this Viewpoint, we highlight ethical concerns stemming from the perspectives of users, developers, and regulators, notably focusing on data privacy and rights of use, data provenance, intellectual property contamination, and broad applications and plasticity of LLMs. A comprehensive framework and mitigating strategies will be imperative for the responsible integration of LLMs into medical practice, ensuring alignment with ethical principles and safeguarding against potential societal risks.

Lancet Digit Health 2024

Published Online
April 23, 2024
https://doi.org/10.1016/S2589-7500(24)00061-X

*Contributed equally

Division of Pharmacy, Singapore General Hospital, Singapore (J C L Ong PharmD); **Duke–NUS Medical School** (J C L Ong, N Liu PhD, D S W Ting PhD) **and Department of Pharmacy** (L S T Chew MMedSc), **National University of Singapore, Singapore; Department of Ophthalmology, Chang Gung Memorial Hospital, Linkou Medical Center, Taoyuan, Taiwan** (S Y-H Chang MD); **College of Medicine, Chang Gung University, Taoyuan, Taiwan** (S Y-H Chang); **Department of Biomedical Sciences and Engineering, Mbarara University of Science and Technology, Mbarara, Uganda** (W William PhD); **Bakar Computational Health Sciences Institute, and Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA** (Prof A J Butte PhD); **Center for Data-Driven Insights and Innovation, University of California Health, Oakland, CA, USA** (A J Butte); **Stanford Health Care, Palo Alto, CA, USA** (Prof N H Shah PhD); **Department of Medicine, and Clinical Excellence Research Center, School of Medicine, Stanford University, Stanford, CA, USA** (Prof N H Shah); **Singapore Health Services, Pharmacy and Therapeutics Council Office, Singapore** (L S T Chew); **Department of Pharmacy, National Cancer Centre Singapore, Singapore** (L S T Chew); **Harvard Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA** (Prof F Doshi-Velez PhD); **StatNLP Research Group, Singapore University of Technology and Design, Singpore** (W Lu PhD); **Murdoch Children's Research Institute,**

Introduction

In the wake of ChatGPT's public release, over a thousand prominent computer scientists and technology industry experts, including Elon Musk and Steve Wozniak, took the unprecedented step of signing a letter calling for an immediate 6-month pause on AI. They argued that the current trajectory of generative AI development had spiralled "out-of-control", posing "profound risks to society".[1] Despite objections, the medical fraternity continued the pursuit of scaling-up generative AI research and integration into medicine. Archetypal discussions on AI ethics in medicine revolves around poor model accuracy for users not represented in the training data, transparency of models and model-building, accountability for model output, potential model bias, and risk for privacy and confidentiality breaches.[2,3] However, these concerns fail to fully capture distinctive concerns posed by LLMs.

In this context, we find ourselves in a situation that mirrors the classic Collingridge dilemma: "Attempting to control a technology is difficult…because during its early stages, when it can be controlled, not enough can be known about its harmful social consequences to warrant controlling its development; but by the time these consequences are apparent, control has become costly and slow."[4] This dilemma can be viewed as a problem of pacing—although technology development advances rapidly, governance and regulation lags behind. To effectively regulate LLMs, grasping the fundamental ethical issues inherent in their design and use is crucial. 1 year after the release of ChatGPT, we now have a better understanding of the limitations and risks the technology poses. LLMs differ substantially from AI-based technologies that are already regulated, creating unique regulatory hurdles: (1) data privacy and rights of use associated with training on massive datasets sourced from the internet;[5,6] (2) data provenance, intellectual property contamination, and the uncertainty about the data derivatives that could hamper the accuracy of output; and (3) the so-called plastic nature of LLMs that allows for dynamic learning and evolution of LLM applications based on user inputs and changing clinical contexts (table). The broad use of LLM-based models across different industries limit the utility of a single governing framework. Identification and audit of the societal risks posed by LLM-based models becomes challenging because the precise mechanisms of their tuning or modifications remain opaque.[7] In this Viewpoint, we discuss these important peculiarities to position LLMs in the large literature on the ethics of AI.

Data privacy and data rights of use

The development and deployment of LLM models challenge the boundaries of data privacy regulations. When identifiable patient data are used during training, there is potential risk that these models inadvertently memorise and disclose sensitive information in the absence of proper security measures. The use of patient information for LLM pre-training without obtaining explicit informed consent contravenes rights-of-data policies.[8] In addition, data breach of sensitive patient information can occur after adversarial attacks,[7] and the re-identification of even anonymised medical data is now possible with few spatiotemporal datapoints.[9]

A greater effort is needed to enhance data privacy and security of LLM-based models. Clinical LLM models trained with patient information should undergo rigorous cross-examination before implementation[10] as a form of penetration test. Cybersecurity measures, such as the use of pseudonyms implementing differential privacy techniques, could be used to counteract the risks of malicious attacks and data poisoning through deliberate adversarial prompting. Preliminary studies have suggested that LLMs can be taught to shield or protect specific categories of personal information under simulated scenarios.[11] What is currently absent are benchmark approaches that effectively measure the balance between privacy and the utility of LLMs. This benchmarking would help in evaluating the models' ability to maintain confidentiality while still delivering

	Developers and LLM researchers	Regulators and governance bodies
Data privacy and data rights of use	Have a greater focus on: the cross examination of LLM-based models for risk of data breach; penetration tests for adversarial attacks; the development of benchmarks to evaluate the privacy *vs* utility trade-off; and validation frameworks for multimodal LLM evaluation	Use a pragmatic, tiered approach to regulation based on the sensitivity of data used in training and inputs into LLM; evaluate data security measures required in accordance with data risk category
Data provenance and contamination of intellectual property	Promote transparency in training datasets used, including source, quality, and quantity; conduct conceptualisation, testbed, and prospective reviews of new market structures	A generative AI take on fair use doctrine
Broad applications and plasticity of LLMs	Develop benchmarking frameworks and risk-assessment methodologies (eg, quality improvement and failure modes and effects analysis): highlight high-risk areas of harm, including quantification of hallucinations, reproducibility of output, and bias; enforce prospective and continuous stewardship	Create sandbox environments, taking on an iterative approach to the development of regulatory guidance on the basis of new knowledge

Users and consumers (ie, the clinicians and their patients) must be familiarised with the rights of data (ie, right of access, right to rectification, right to erasure, right to restrict processing, right to data portability, right to object, right to not be subject to decisions based solely on automated processing). LLM=large language model.

Table: **Ethical concerns relating to framework and mitigating strategies for responsible development and use of LLMs in medicine**

Melbourne, VIC, Australia (Prof J Savulescu PhD)**; Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore** (Prof J Savulescu)**; Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Oxford, UK** (Prof J Savulescu)**; Artificial Intelligence and Digital Innovation, Singapore Eye Research Institute, Singapore National Eye Center, Singapore Health Service, Singapore** (D S W Ting)**; Byers Eye Institute, Stanford University, Palo Alto, CA, USA** (D S W Ting)

Correspondence to: Assoc Prof Daniel Shu Wei Ting, Artificial Intelligence and Digital Innovation, Singapore Eye Research Institute, Singapore National Eye Center, Singapore Health Service, Singapore 168751
daniel.ting@duke-nus.edu.sg

valuable outputs in a controlled environment. In addition, such benchmarking and evaluation tools will need to take on a multimodal approach, given that a multimodal LLM capable of integrating different inputs of text, images, and audio is fast gaining traction in medical applications.[12]

Moving forward, data protection regulations and guidance will need to take on a more pragmatic approach to avoid placing a hard stop on LLM development and implementation. For example, the use of a tiered approach based on classification of training and input data: public, internal (eg, non-patient data, information, or intellectual property for which there are proprietary interest or contractual obligations), confidential (eg, de-identified or anonymised patient data), and restricted data (eg, identifiable patient data). Data security and information technology infrastructural requirements will differ between different data risk categories (eg, air-gap environment for high-risk tiers). Patients should provide broad informed consent to share data, and should be proactively educated on rights of data, such as right to access, right to erasure, and right to limit data processing. However, in the event whereby the classification of training data is unclear because of an absence of transparency, regulators will need to weigh between the potential risks of data breach and the benefits that the LLM-based model can bring to the general population.

Data provenance and intellectual property contamination

The provenance of data refers to the origins, custody, and ownership of the information used to train these models. When LLMs ingest massive amounts of data from various sources, some of this content might be used without proper licensing despite being protected by intellectual property laws. Additionally, users might inadvertently prompt these models with references to copyrighted or trademarked works, raising questions about the legality and ethics of generating outputs derived from copyrighted or patented inputs. Regulatory rules can be implemented to restrict LLM training on appropriately licensed datasets, but this poses risk to the development, refinement, and maturation of technical standards.

We encourage developers to maintain transparency when describing the training datasets used in developing LLM-based models when possible, including the source, quantity, and diversity of data.[13] One of the key factors contributing to the accuracy of LLMs is their large-scale pre-training on vast amounts of text data from diverse sources. However, the accuracy can also be influenced by biases present in the training data, the quality of the input data, and the inherent limitations of the model architecture. Unlike other AI models, common techniques used to mitigate AI model bias, such as data resampling, prejudice removal, or subgroup modelling, cannot be easily adopted for LLM-based models. There is an absence of robust quantification on the amplification effect of model bias when building fine-tuned models on general-purpose ones. Watermarking techniques seek to address concerns over originality and ownership through embedding a mark into AI-generated content before its release,[14] albeit the robustness of such techniques has been challenged. Work is underway to evaluate feasibility of unlearning in LLMs[15] to enable models to be updated to comply with updated legislation and data protection standards. The techniques present as potential stop-gap measures. However, to comprehensively address this issue, we will most likely need a paradigm shift in current market structures, incentivisation, and reimbursement strategies for LLM-based models or LLM-incorporated medical devices. Segal and colleagues[16] proposed a decentralised or blockchain-based, token economy-based market for medical research and publishing. Purported benefits of this blockchain-based platform include data and workflow transparency, immutability of original work, the minimisation of fraud, and incentivisation of reviewers through token payments. Such endeavours need prospective research and review to evaluate the feasibility of scaling.

The fair use doctrine is a legal framework promoting freedom of expression through permitting unlicensed use of copyright-protected works under specific circumstances.[17] Regulators can apply principles of the fair use doctrine for generative AI-based models developed for medical uses (panel).

Broad applications and plasticity of LLMs

The potential applications of LLMs in medicine can be broad ranging and heterogeneous. LLMs facilitate research by summarising texts and extracting key points from published literature, enhance medical education through data synthesis and interactive learning, and improve clinical tasks by streamlining administrative efforts and supporting decision making.[2] The industry is exploring the performance of medical chatbots in assisting patient care, counselling, expressing empathy, and providing information about health recommendations. These broad and varied applications of LLMs in medicine mean that a single governing framework for their use is impractical. In addition, the plastic nature of LLMs allows for dynamic learning and continuous evolution based on user inputs and changing clinical contexts. Much like neuroplasticity of the human brain,[21] LLMs are capable of changing characteristics of its response to stimuli, such as different prompting strategies or different fine-tuning data inputs. Drawing parallels to human neuroplasticity, structural or functional changes to LLMs can be positive (eg, enhanced personalisation of response through active reinforcement learning) or negative (eg, through propagation of inherent bias and so-called AI hallucinations). The identification and audit of the societal risks posed by LLM-based models becomes challenging as the precise mechanisms of their tuning or modifications remain opaque—known as the black-box nature of LLMs.[22]

There is an urgent need to develop robust frameworks for evaluating LLM-based models for medicine to mitigate the risks discussed in this Viewpoint. Such a framework can incorporate clear assessment methodologies before implementation, such as quality improvement[23] and failure modes and effects analysis,[24] to identify and mitigate potential risks and harms. The evaluation of LLM-based models for medicine in areas of high risk is of utmost importance: the propagation of bias or discrimination, the quantification and reduction of AI hallucinations, and the reproducibility of the model outputs. Techniques such as retrieval-augmented generation can help in minimising harm and bolstering the self-consistency of responses by cross-referencing with reliable data sources.[25] Bias evaluation is another crucial aspect, whereby assessment checklists and benchmarking frameworks are applicable. In one study, published as a preprint, authors developed a generative AI assessment checklist specific to models developed for medical indications.[26] Continuous stewardship after deployment is essential to address any emergent biases or model drifts.[27] Regulatory bodies can take on a proactive role through the creation of sandbox environments to allow exploration, interaction, and evaluation of LLM-based applications without compromising on security and can mitigate risks to patient safety.[28,29]

Panel: Fair use doctrine principles

When evaluating fair use, the fair use doctrine calls for:

(1) Purpose and character of use
Describes the intended use of original material, whether for commercial or not-for-profit use. Highly transformative applications that repurpose the use of the material and cannot be substituted by the original work. Generalist LLMs, such as ChatGPT, are trained on highly diverse datasets,[18] much of which is probably for non-medical intents. LLM-based medical applications with clearly defined attributes might hence be considered as transformative solutions, work that is largely repurposed from its original material.

(2) Nature of the original work
The use of a creative or imaginative work, such as novels or movies, is less likely to support a claim of fair use than the use of factual work. Although creative industries such as art and music encourage imagination and originality, the practice of medicine thrives upon an evidence-based approach grounded on factual information that is more likely to be considered fair use.

(3) Amount and substantiality of original material used
The black-box nature of LLMs (ie, the input and output is known to users, but the internal mechanisms remain unknown) render this evaluation highly challenging. Uncertainty over data derivatives and data provenance pose a barrier to accurate quantification of original material used. Foundation models pre-trained on clinical information[19] present with a clearer definition on data lineage.

(4) Effect of use upon the potential market for, or value of, copyrighted work
The extent to which unlicensed use of the original work harms the existing or future market for copyrighted owners' original work. LLM-based applications that are developed for specific medical purposes might be regulated as medical devices.[20]

Conclusion

The rapid advancement of LLMs in the medical field has ushered in a new era of technological capabilities alongside complex ethical and regulatory considerations. Such challenges are unique to LLM-based models as opposed to conventional machine-learning or deep-learning-based models. Developers and regulatory bodies need to work in tandem to encompass the multifaceted nature of LLMs, ensuring data protection without stifling innovation. As we navigate these challenges, a balanced

Search strategy and selection criteria

We included original papers, reviews, narratives, correspondences, perspectives, and viewpoints identified through searches of PubMed and arXiv from Jan 1, 2020, to Aug 10, 2023, using the search terms ("natural language processing" OR "generative adversarial network" OR "generative model" OR "generative artificial intelligence" OR "generative AI" OR "transformer model" OR "reinforcement learning" OR "large language model" OR "LLM" OR "foundation model" OR "recurrent neural network" OR "RNN" OR "bidirectional encoder representations from transformers" OR "generative pretrained transformer" OR "ChatGPT" OR "Chat Generative Pre-training Transformer" OR "LLaMA" OR "Large Language Model Meta AI" OR "Pathways Language Model") AND ("ethics" OR "bioethics" OR "medical ethics" OR "regulation" OR "regulatory"). We excluded publications that did not mention or discuss ethical issues. 37 of 998 articles from the search on PubMed and 21 of 668 articles on arXiv were relevant to the topics and eligible for our study. Only papers published in English were reviewed.

approach that fosters innovation while upholding ethical standards will be essential for the responsible integration of LLMs into medical practice.

Contributors

DSWT formulated the direction of the article. JCLO and SY-HC led the literature search and manuscript writing. The manuscript was revised and finetuned by WW, AJB, NHS, LSTC, NL, FD-V, WL, JS, DSWT, and finalised by JCLO and SY-HC.

Declaration of interests

DSWT holds patents on a deep-learning system for the detection of retinal diseases. AJB is a cofounder and consultant for Personalis and NuMedii; is a consultant to Mango Tree Corporation; has previously been a consultant for Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, and Vet24seven, and several other non-health related companies and mutual funds; has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical-specific or disease-specific foundations and associations, and health systems; receives royalty payments through Stanford University for several patents and other disclosures licensed to NuMedii and Personalis; has done research funded by NIH, Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, and the Barbara and Gerson Bakar Foundation; and has previously done research funded by the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity. NL is a scientific advisor to TIIM SG. NHS is a cofounder of Prealize Health (a predictive analytics company) and Atropos Health (an on-demand evidence generation company); receives funding from the Gordon and Betty Moore Foundation for developing virtual model deployments; and is a member of working groups of the Coalition for Health AI (CHAI), a consensus-building organisation providing guidelines for the responsible use of artificial intelligence in health care. JS, through his involvement with the Murdoch Children's Research Institute, receives funding from the Victorian State Government through the Operational Infrastructure Support (OIS) programme. JCLO is supported by grants from the National Medical Research Council Singapore (MOH-CIAINV21nov-001) and AI Singapore OTTIC (AISG2-TC-2022-006). All other authors declare no competing interests.

Acknowledgments

This research project is supported by National University of Singapore under the NUS Start-Up grant; NUHSRO/2022/078/Startup/13. This research was funded in part by the Wellcome Trust (WT203132/Z/16/Z).

References

1 Metz C, Schmidt G. Elon Musk and others call for pause on AI, citing profound risks to society. March 29, 2023. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html (accessed July 25, 2023).
2 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023; **29:** 1930–40.
3 Farina M, Lavazza A. ChatGPT in society: emerging issues. *Front Artif Intell* 2023; **6:** 1130913.
4 Genus A, Stirling A. Collingridge and the dilemma of control: towards responsible and accountable innovation. *Res Policy* 2018; **47:** 61–69.
5 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023; **6:** 120.
6 Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA* 2023; **330:** 315–16.
7 Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. *AI Ethics* 2023; published online May 30. https://doi.org/10.1007/s43681-023-00289-2.
8 European Data Protection Supervisor. Rights of the individual. 2023. https://edps.europa.eu/data-protection/our-work/subjects/rights-individual_en (accessed Oct 26, 2023).
9 Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020; **27:** 491–97.
10 He Y, Zamani E, Yevseyeva I, Luo C. Artificial intelligence-based ethical hacking for health information systems: simulation study. *J Med Internet Res* 2023; **25:** e41748.
11 Chen Y, Mendes E, Das S, Xu W, Ritter A. Can language models be instructed to protect personal information? *arXiv* 2023; published online Oct 3. https://doi.org/10.48550/arXiv.2310.02224 (preprint).
12 Meskó B. The impact of multimodal large language models on health care's future. *J Med Internet Res* 2023; **25:** e52865.
13 Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020; **26:** 1320–24.
14 Qiao T, Yuyan M, Ning Z, et al. A novel model watermarking for protecting generative adversarial network. *Comput Secur* 2023; **127:** 103102.
15 Viswanath Y, Jamthe S, Lokiah S, Bianchini E. Machine unlearning for generative AI. *Journal of AI, Robotics & Workplace Automation* 2024; **10:** 37–46.
16 Segal G, Martisiano Y, Markinzon A, Mayer A, Halperin A, Zimlichman E. A blockchain-based computerized network infrastructure for the transparent, immutable calculation and dissemination of quantitative, measurable parameters of academic and medical research publications. *Digit Health* 2023; **9:** 20552076231194851.
17 US Copyright Office. US Copyright Office fair use index. November, 2023. https://www.copyright.gov/fair-use/ (accessed Dec 6, 2023).
18 Gallifant J, Fiske A, Levites Strekalova YA, et al. Peer review of GPT-4 technical report and systems card. *PLOS Digit Health* 2024; **3:** e0000417.

19 Moor M, Banerjee O, Hossein Abad ZS, et al. Foundation models for generalist medical artificial intelligence. *Nature* 2023; **616:** 259–65.

20 Minssen T, Gerke S, Aboy M, Price N, Cohen G. Regulatory responses to medical machine learning. *J Law Biosci* 2020; **7:** lsaa002.

21 Puderbaugh M, Emmady PD. Neuroplasticity. Treasure Island, FL: StatPearls Publishing, 2023.

22 Nature. ChatGPT is a black box: how AI research can break it open. *Nature* 2023; **619:** 671–72.

23 Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med* 2022; **5:** 66.

24 Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health* 2022; **4:** e384–97.

25 Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval augmented generation. *Hepatology* 2024; published online March 7. https://doi.org/10.1097/HEP.0000000000000834.

26 Ning Y, Teixayavong S, Shang Y, et al. Generative artificial intelligence in healthcare: ethical considerations and assessment checklist. *arXiv* 2023; published online Nov 2. https://doi.org/10.48550/arXiv.2311.02107 (preprint).

27 Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA* 2020; **324:** 1397–98.

28 Harvard Gazette. Harvard designs AI sandbox that enables exploration, interaction without compromising security. Sept 26, 2023. https://news.harvard.edu/gazette/story/newsplus/harvard-designs-ai-sandbox-that-enables-exploration-interaction-without-compromising-security/ (accessed Sept 26, 2023).

29 Infocomm Media Development Authority. First of its kind generative AI evaluation sandbox for trusted AI by AI Verify Foundation and IMDA. Oct 31, 2023. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox (accessed Nov 1, 2023).