The Risks of Multimodal Generative AI Models and Misinformation: Policy Strategies for the High-Risk Domains of Healthcare and Journalism

Yifeng Liu
Computer Science Department, UCLA
liuyifeng@ucla.edu

Bruce Xu
Computer Science Department, UCLA
brucexu0222@ucla.edu

Aman Ganapathy Manvattira
Computer Science Department, UCLA
amanvatt02@ucla.edu

Laleh Jalilian
UCLA Health
ljalilian@mednet.ucla.edu

Abstract

The emergence of multimodal generative AI models has transformed the creation of digital content, enabling the simultaneous creation of realistic content. Although multimodal AI models are popular to the lay public and continue to advance rapidly, they also pose a substantial risk of spreading misinformation at scale. There exists a lack of a comprehensive policy framework that guides their development and mitigates this risk. In this article, we examine the challenges posed by misinformation from multimodal AI models, focusing on case studies in medicine and journalism. We also propose structured policy frameworks that consider the definition of technical standards, the development of ethical AI, and regulatory safeguards in both fields. By proactively addressing these issues, we hope to foster trust and accountability in the deployment of these transformative technologies in real-world settings.

1 Introduction

The field of digital content creation has seen a paradigm shift with the advent of multimodal generative AI models (Liu et al., 2024). Although models like OpenAI's Sora (OpenAI, 2024) represent significant technological advancement, the potential for misuse through misinformation propagation (Goldstein et al., 2023) is substantial and already causing harm from this technology (Mogavi et al., 2024). The commercial use of automated output from these models is increasing, but the absence of comprehensive governance frameworks poses a threat to public confidence in the authenticity of digital content and the safe use of this technology (De Almeida et al., 2021). Given the rapid technological advances and the increasing use of these technologies already, there is a pressing need to create a policy framework for the development and deployment of multimodal AI models, with a specific focus on misinformation prevention. In this paper, we consider the risks of misinformation as it pertains to two high-stakes fields, journalism and medicine, and propose a framework that considers technical controls, transparency protocols, and regulatory measures to reduce the risk of misinformation.

2 Motivation

2.1 The Impact of Misinformation in Healthcare and Journalism

Misinformation can include inaccurate, incomplete, misleading, or false information, as well as selective or half-truths. In the context of multimodal generative AI models, we can alter this definition to include the generation and dissemination of false, misleading, or inaccurate content across multiple modalities (e.g., text, images, audio, video) through AI-driven models. In medicine, misinformation can refer to the generation and propagation of incorrect biomedical facts that can be mistaken for valid medical guidance or the omission of information that is needed for clinical decision-making. This can lead to incorrect diagnoses and incorrect treatment decisions, which could impact care quality and patient safety. Although there is excitement that multimodal models can increase access to medical information and improve public health messages, there is little evidence of their safety and effectiveness in healthcare (Laranjo et al., 2018) and instead a substantial concern that these tools can facilitate the intentional spread of targeted health misinformation (Menz et al., 2024). A recent example of this occurred with the 'infodemic' observed during the COVID-19 pandemic, which caused confusion, panic, and mistrust and serves as an example of the potential harm that misinformation can inflict (Nogara et al., 2022; Zarocostas, 2020).

In the field of journalism, misinformation refers to false, misleading, or inaccurate information that is shared, regardless of intent, through the news

media or journalistic sources. The risks of using generative AI models in journalism are concerning in eroding public trust in the media and impacting democratic institutions (Badminton, 2023). The most concerning aspect about the use of these models in any domain is that they can automate the intentional creation and dissemination of misinformation on an extraordinary scale (Marcus, 2022; Goldstein et al., 2023). With the generation of deepfakes and misinformation, including coercive text, visual content, and audio content, the intent to engineer social attacks, swaying social discord, and influencing elections is real (Federspiel et al., 2023; COM, 2021), which brings us a pressing need to address these challenges and regain fairness and public trust.

3 Related Works

In healthcare, Yang et al. (2024) introduced a novel way to enhance an LLM's medical reasoning capabilities by giving it the ability to query a knowledge graph. A knowledge graph consists of medical terms as nodes, and the edges between them indicate a relationship between those terms. By extracting medical terminology from queries made, querying the graph, and providing a more informed answer, there is great potential to reduce the risk of misinformation. However, the existing policy literature in this sphere has had limited technical specifications, which mark it as unsuited to deal with the myriad challenges present.

Generative AI has brought about an unprecedented explosion in information in journalism; however, the development of AI tools in journalism also caused a boom of misinformation (Loth et al., 2024). With this issue in mind, some researchers exploited a number of possible approaches. The DSA of the EU (Commission, 2024) regularized that governments should ensure explainability of misinformation, and platforms have the responsibility to avoid spread of misinformation. Denais (2024) proposed to motivate AI tool developers to embed security mechanisms, prioritize the use of watermarking technology, and develop common standards based on collaborative frameworks between AI tool developers and online platforms. In journalism, De Angelis et al. (2023) also recommended that AI generation be clearly disclosed and labeled. However, a more systemized framework is needed for better regulation of misinformation in journalism.

4 Towards AI-generated Misinformation in Healthcare

4.1 The Risks of Altered AI Models for Spreading Misinformation

4.1.1 Targeted Misinformation Attacks

Targeted misinformation attacks are one way multimodal generative AI models can be altered to spread misinformation. Encoded knowledge about the medical field is represented in several of the layers of the transformer-based architecture on which current multimodal generative AI models are built. These associations can be built up during the training process or during fine-tuning. By targeting these associations, an attacker could change the weights in such a way that a certain drug or treatment is more likely to be recommended than another (Han et al., 2024). This is a targeted misinformation attack and the goal of such an attack is to make a multimodal generative AI model more likely to recommend a particular drug or treatment.

4.1.2 Data Poisoning Attacks

We draw special attention to data poisoning attacks, a subset of targeted misinformation attacks. Multimodal generative AI models are trained on an ever-increasing dataset, often involving many publicly available websites and repositories. As these models grow in scale, the need for data increases. Consequently, this makes them especially vulnerable to a type of attack called a data poisoning attack. In this attack, these generative models end up encountering 'poisoned' data or deliberately planted misinformation (Alber et al., 2025). This ends up affecting the weights of the model, making it more likely to spread misinformation.

4.1.3 Hallucinations

A hallucination can be defined as the content generated by a model that is not considered reasonable or incorrect (Huang et al., 2025). Although the exact sources of hallucinations are not fully understood (McKenna et al., 2023), they are part of the interaction with multimodal models. The risk of hallucinations lies in the potential of users of multimodal models to get inaccurate information from the said models and then to work off of them.

4.2 Moving towards Health-System Centered Policy

In this section, we specifically focus on the concerns that health systems have about the integration

of multimodal generative models into their clinical, research, and operational workflows. The success of any new technological tool in a healthcare context is evaluated by the so-called 'quintuple performance goals' of improving health outcomes, patient experience, provider well-being, health equity and cost effectiveness (Coleman et al., 2016). The potential for generative model misinformation to affect all these goals exists and poses substantial ethical and legal risks (Monteith et al., 2024) for health systems. Reasons for this include biased training data and incomplete data.

There is a growing consensus in medicine that robust quality control measures are essential to ensure the accuracy and reliability of AI-generated information, to minimize the risk of misinformation and ensure patient safety and trust in generative AI. What is considered accurate must also be defined, but broadly, accuracy is considered to be closeness of agreement between the measured value and the true value of what is intended to be measured), complete (the presence of all necessary data), and traceable (permits an understanding of the relationships between the output and the source data). With this in mind, we consider a comprehensive framework that incorporates multiple levels of intervention to mitigate misinformatino, including preventive measures, mitigating measures, and reducing measures (Coleman et al., 2016), and the summarized policy details are shown in Table 1 (in the appendix).

Measures that could help prevent the use of misinformation include a mandatory disclaimer to all staff using GenAI models that the output can be inaccurate. This would provide a warning to staff to use their judgment as clinicians when considering the output of a model. In addition, ensuring judicious use and audit of training data helps avoid data poisoning attacks.

A focus should also be on measures that mitigate the risk of misinformation. Specifically, technical frameworks that incorporate retrieval-augmented generation (RAG) and human-in-the-loop frameworks should be designed for user-friendly AI-clinical decision support systems that reduce the incidence of hallucinations/ misinformation. RAG is a technical method that augments the model's knowledge with medical data, allowing health systems to guide the model's output with medically accepted reference documents containing the most up-to-date literature and clinical guidelines to improve the output accuracy. RAG combines external knowledge retrieval with AI-generated generative models to provide up-to-date, evidence-based, and contextual recommendations. Additionally, we propose a novel solution that integrates "Knowledge Graphs" with large language models. As referenced in the "Related Works" section, there is existing technical literature on teaching an LLM to query a knowledge graph to establish proper relations between medical terms, thus making medical inference more logical and less misinformation prone. We propose mandating this for any models deployed in clinical contexts, and to augment this system by adding routine audits of the source databases used to build the graph to protect against targeted misinformation attacks.

Finally, ways to reduce the risk of misinformation would incorporate human-in-the-loop feedback systems, which would enable clinicians to ensure that their clinical judgment is increased when verifying the veracity of the model output. Gen AI models trained on older data may not know the most accurate medical information, and ensuring HITL allows end users to fill in any knowledge gaps and verify accurate information from the model.

5 Risks and Mitigation of AI-generated Misinformation in Journalism

Although generative AI applications in journalism seem not to be so harmful as those in healthcare industry, propagation of AI-generated misinformation may also caused serious consequences.

5.1 Misinformation Risk with Generative AI in Journalism

Generative AI models can be used to generate highly convincing content, having become one of the primary sources for fake news or misleading narratives (Neumann et al., 2024). The decrease in the credibility of information can do great harm to the reputation of journalists. Furthermore, these models may produce false stories that are difficult to distinguish from real news, posing a significant threat to public trust in the media. Some researchers believe that hallucination is inevitable (Xu et al., 2024) because of the stochastic nature of the probabilistic model in AI models (Yao et al., 2023). And risks related to misinformation in journalism include the generation of false content that can manipulate public opinion, amplify biases, impact journalistic integrity, etc.

5.1.1 Manipulation of Public Opinion

Multimodal generative AI models can be elaborated to appeal to specific audiences, potentially manipulating public opinion on various issues (Li et al., 2025). This risk is exacerbated when social media has become a great part of our lives, where such content can be easily spread with unbelievable speed. Moreover, advanced multimodal generative AI models can effectively influence individuals' beliefs by mixing misinformation with true information (Mahony and Chen, 2024). In this way, as false information is mixed with true information, people would find it harder to distinguish the overall validity of a piece of information.

5.1.2 Amplification of Biases

Generative AI models may be greatly affected by biases within human beings, and such biases can be unintentionally or deliberately passed on to the generated contents. This can lead to the advancement of "biased" stereotypes and the spread of misinformation (Critch and Russell, 2023). For example, biases in training data can result in biased news narratives, and this could further provide misleading figures and characteristics to the public. As people absorb such information, the training data of multimodal generative AI models, which are learned from people's opinions, become more biased, creating a "biased" loop of information exchange. Such phenomena reinforce the bias and discrimination within the training datasets and within people's opinions at the same time (Vock, 2022).

5.1.3 Impact on Journalistic Integrity

Although multimodal generative AI models can improve the efficiency of journalism works, the use of these models may lead to a decline in the quality of news coverage (Tseng et al., 2025). Simply replacing human journalists with AI models would result in a loss of critical thinking and investigative depth, which are significant components of quality journalism. Also, we take the diversity of opinions in journalism to a serious extent. If we do not regulate the usage of generative AI models in journalism, this diversity of opinions and thinking could diminish. This could further bring a great decline in public trust as people recognize the decreasing quality of journalistic literature (Cheng, 2025). Although there could be an increase in the content published with the help of generative AI models, it cannot remedy the decrease in the quality of journalism works, especially in critical thinking and investigative depth as we mentioned above.

5.1.4 Ethical and Legal Confusion

Misuse of multimodal generative AI models in journalism would pose ethical risks related to authorship, accountability, and transparency (Mueller et al., 2024; Novelli et al., 2024). For example, it is still unclear how to attribute responsibility for content generated by these models. Should the responsibility of the content be tied to the person who utilizes multimodal generative AI models or the multimodal generative AI model developers, or the models themselves? This is an important unanswered question. Additionally, there are legal implications related to the potential misuse of multimodal generative AI models for spreading misinformation (Yao et al., 2023). For instance, if a person follows the misleaded journalism information from multimodal generative AI models and harm/kill another person physically, how should we decide the liabilities? This also remains an unanswered question.

5.2 A Novel Policy Framework for Using Multi-modal Generative AI in Journalism

We propose a comprehensive policy framework that integrates regulatory, technical and ethical considerations to mitigate risks of misinformation from generative models in the field of journalism.

With respect to regulatory measures, we would require the disclosure of AI. For journalism, news organizations must disclose AI-generated content and provide clear labeling to distinguish it from human-written articles (De Angelis et al., 2023). It is also necessary for governments to enforce accountability for misinformation, where platforms should be legally responsible for AI-generated misinformation, similar to the DSA of the EU (Commission, 2024). Moreover, we want to ensure that all generative AI systems used in journalism must undergo independent audits to assess their reliability and biases. This could be achieved by establishing a new subdepartment in government agencies to manage multimodal generative AI-related issues.

Second, we want to add technical safeguards to ensure that the content of generative AI is appropriate before reaching the users. Although it is not necessarily the sole responsibility of the government to establish technical safeguards, we value the efforts of creating technical safeguards from the governments, preferably the joint efforts from governments and corresponding companies in the

industry, where the government provides the legal power and the stakeholder companies provide the technical power. Regarding technical details, we encourage all AI-generated journalism to be cross-verified using independent fact-checking systems. Bias mitigation algorithms should be one of the most promising targets. When such algorithms reach a mature state, governments should ensure continuous efforts in auditing in multimodal generative AI model training datasets from stakeholder companies to reduce misinformation to the greatest extent. With the rapidly developing power of AI, governments around the world should take advantage of it by developing AI tools that can trace the origins of false narratives and detect coordinated disinformation campaigns.

Finally, we want to propose ethical and organizational guidelines to make the regulations easier to execute. In terms of institutional structures, there are also a few actions that governments can take. For AI ethics governance committees, establishing ethics boards within media organizations to oversee AI-generated content policies would be a great action. By transferring the responsibility of regulating multimodal generative AI models to a specific board, the government could increase the processing efficiency of related issues. Moreover, under government regulations, newsrooms should implement AI literacy training programs to ensure responsible AI use in journalism before publishing or announcing any content to the public. To prevent misleading AI-generated information from being spread, as another layer of protection, governments and educational institutions should promote digital literacy to help audiences critically assess AI-generated news with an educational value.

6 Conclusion

The use of multi-modal generative AI in healthcare and journalism is rapidly growing. Although there are substantial beneficial uses for this technology, the negative impacts of AI-generated misinformation already pose a real risk to the general public. Misinformation created by generative AI models in the fields of healthcare care and journalism includes factual errors, omitted information, fabricated sources, and dangerous advice, among others. We believe that domain-specific policy frameworks should focus on regulatory, ethical, and technical ways that AI models or AI decision support systems could reduce harm from misinformation. As we move forward, we anticipate that policy frameworks for generative AI tools will evolve, but so will the strategies of malicious actors seeking to exploit these systems. The establishment of appropriate misinformation mitigation frameworks is imperative in maintaining an AI ecosystem with an acceptable level of risk.

7 Embedded Ethics Discussion

To effectively integrate the ethical considerations raised in this paper into an AI course, we would design a module for the social impacts and risks of generative AI models in real life. The module would begin with some examples in real life, such as the medical "infodemic" and the decrease in public trust in journalism, to indicate that the rapid development of generative AI models has already led to the potential for widespread misinformation.

After that, we will introduce the risks in generative AI models and the potential solutions for these risks within a series of discussion lectures and coding assignments. In the lectures, we discuss specific examples of how misinformation can have serious risks in different downstream applications. In addition, we plan to discuss the differences between the policy framework proposed in our paper and other similar frameworks, figuring out their potential usages in the coming AI-era. For coding assignments, students will implement basic components for the proposed policy framework, such as the bias detection algorithm. These exercises would not only offer students the opportunity to practice writing codes for AI systems but also strengthen their sense of responsibility for proper use of AI models. By integrating ethics into the learning process, students are expected to become AI developers who are not only skilled in AI systems construction, but also capable of addressing the challenges posed by AI models in the future.

8 Contribution Statement

The introduction is written by Bruce Xu and Laleh Jalilian. The motivation and related work are written by Laleh Jalilian, Aman Ganapathy and Yifeng Liu. The healthcare case study was written by Laleh Jalilian and Aman Ganapathy, and the journalism case study is done by Yifeng Liu, Bruce Xu and Laleh Jalilian. Laleh Jalilian wrote the conclusion and future directions. Yifeng Liu completed the final part for the embedded ethics discussion. All authors reviewed the paper in its entirety.

References

Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, pages 1–9.

Nikolas Badminton. 2023. Chatgpt discusses the risks of large language ai models in journalism. *Futurist.com*. Accessed: 2025-03-11.

Sophia Cheng. 2025. When journalism meets ai: Risk or opportunity? *Digital Government: Research and Practice*, 6(1):1–12.

Katie Coleman, Edward Wagner, Judith Schaefer, Robert Reid, and Lisa LeRoy. 2016. Redefining primary care for the 21st century. *Rockville, MD: Agency for Healthcare Research and Quality*, 16(20):1–20.

EU COM. 2021. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *Proposal for a regulation of the European parliament and of the council*.

European Commission. 2024. Digital services act (dsa). https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package. Accessed: 2025-03-11.

Andrew Critch and Stuart Russell. 2023. Tasra: a taxonomy and analysis of societal-scale risks from ai. *arXiv preprint arXiv:2306.06924*.

Patricia Gomes Rêgo De Almeida, Carlos Denner dos Santos, and Josivania Silva Farias. 2021. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, 23(3):505–525.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in public health*, 11:1166120.

Anna Denais. 2024. Mitigating disinformation risks in the run-up to 2024 european elections: A three-step action plan against ai-generated disinformation. *Natolin Policy Papers Series*, 2024(1).

Frederik Federspiel, Ruth Mitchell, Asha Asokan, Carlos Umana, and David McCoy. 2023. Threats by artificial intelligence to human health and human existence. *BMJ global health*, 8(5):e010435.

U.S. Food and Drug Administration. 2024. International medical device regulators forum (imdrf). International Medical Device Regulators Forum; U.S. Food and Drug Administration. Accessed: 2025-03-11.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K Bressem, et al. 2024. Medical large language models are susceptible to targeted misinformation attacks. *NPJ digital medicine*, 7(1):288.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Isaac S Kohane. 2024. Injecting artificial intelligence into medicine.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Zhi Li, Wenyi Zhang, Hengtian Zhang, Ran Gao, and Xingdong Fang. 2025. Global digital compact: A mechanism for the governance of online discriminatory and misleading content generation. *International Journal of Human–Computer Interaction*, 41(2):1381–1396.

Moyang Liu, Kaiying Yan, Yukun Liu, Ruibo Fu, Zhengqi Wen, Xuefei Liu, and Chenxing Li. 2024. Misd-moe: A multimodal misinformation detection framework with adaptive feature selection. In *NeurIPS Efficient Natural Language and Speech Processing Workshop*, pages 114–122. PMLR.

Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or curse? a survey on the impact of generative ai on fake news. *arXiv preprint arXiv:2404.03021*.

Simon Mahony and Qing Chen. 2024. Concerns about the role of artificial intelligence in journalism, and media manipulation. *Journalism*, page 14648849241263293.

Gary Marcus. 2022. Ai platforms like chatgpt are easy to use but also potentially dangerous. *Scientific American*, 19.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Bradley D Menz, Natansh D Modi, Michael J Sorich, and Ashley M Hopkins. 2024. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. *JAMA internal medicine*, 184(1):92–96.

Reza Hadi Mogavi, Derrick Wang, Joseph Tu, Hilda Hadan, Sabrina A Sgandurra, Pan Hui, and Lennart E Nacke. 2024. Sora openai's prelude: Social media perspectives on sora openai and the future of ai video generation. *arXiv preprint arXiv:2403.14665*.

Scott Monteith, Tasha Glenn, John R. Geddes, Peter C. Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2):33–35.

Felix B Mueller, Rebekka Görge, Anna K Bernzen, Janna C Pirk, and Maximilian Poretschkin. 2024. Llms and memorization: On quality and specificity of copyright compliance. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 984–996.

Terrence Neumann, Sooyong Lee, Maria De-Arteaga, Sina Fazelpour, and Matthew Lease. 2024. Diverse, but divisive: Llms can exaggerate gender differences in opinion related to harms of misinformation. *arXiv preprint arXiv:2401.16558*.

Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. 2022. The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter. In *Proceedings of the 14th ACM web science conference 2022*, pages 348–358.

Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*, 55:106066.

OpenAI. 2024. Sora: Creating video from text. https://openai.com/sora.

Emily Tseng, Meg Young, Marianne Aubin Le Quéré, Aimee Rinehart, and Harini Suresh. 2025. " ownership, not just happy talk": Co-designing a participatory large language model for journalism. *arXiv preprint arXiv:2501.17299*.

Ido Vock. 2022. Chatgpt proves that ai still has a racism problem. *New Statesman*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, et al. 2024. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

John Zarocostas. 2020. How to fight an infodemic. *The lancet*, 395(10225):676.

A Future Work

Our work leaves open many doors to go through in the future. It has been a primarily reactive framework, developed in response to identified and articulated risks. Future works could try to address preemptive risk prevention in the design process. Additionally, our work is set up as a one-size-fits-all policy, but owing to the different ways misinformation risk can manifest in different cultural contexts, there is merit in developing more locally scoped solutions. On a closing note, we caution readers about the rapid pace of generative AI development, and urge policy makers and deployers to reevaluate our framework in light of the evolving technological scene.

B Summary of Proposed Policy Framework for AI-generated Misinformation Mitigation in Healthcare

We summarize the policy framework for AI-generated misinformation mitigation in healthcare in Table 1.

Table 1: Proposed Policy Framework for Misinformation Mitigation from Generative AI for Healthcare Operations.

Topic	Policy Considerations and Details
Bias and Accuracy	**Bias**: Users must verify information provided by the model before using it in their work and should consider the potential for bias in each use case. Users must be educated on the limitations of GenAI technology in this regard. **Incompleteness**: End users must be informed that outputs may be incomplete due to knowledge cutoffs. Prompt engineering strategies must be developed and disseminated for end users to follow-up with the model if they perceive the result of the output is inaccurate. **Responsibility**: End users are ultimately responsible for any decision made with support from a Generative AI model. Users should only make decisions with Generative AI if it falls within the scope of their expertise and what they are able to verify.
Suitability of Use Cases	**Direct clinical care**: Appropriate use cases for Generative AI must enable measurable outcome measurements. Frameworks similar to those used for new drugs and medical devices, as defined by the FDA/IMDRF Software as a Medical Device risk-categorization framework, should be used for assessing the quality of AI-driven interventions (Food and Administration, 2024; Kohane, 2024). **Administrative Tasks**: Low risk administrative tasks, both for healthcare operations and research, that combine information extraction may provide low hanging fruit options for implementation of GenAI models that do not directly impact patient care.
Technical Frameworks	**Output Reliability and Definitions**: What constitutes reliable outputs should be defined. Outputs should be **accurate** (closeness of agreement between the measured value and the true value of what is intended to be measured), **complete** (the presence of all necessary data), and **traceable** (permits an understanding of the relationships between the output and the source data). **Retrieval Augmented Generation**: RAG methods have demonstrated improvements in clinical accuracy of model outputs. Institutionally accepted databases, guidelines, and documents should be provided to Generative models to increase output reliability. **Human in the Loop**: Domain experts using AI-CDSS need to be able to validate and provide feedback on the model output. Domain experts must be actively involved in the development, validation, deployment, and feedback of outputs from these CDSS systems.
Policy Development	**Iterative Policy Development**: Health system policies pertaining to the use of Generative AI must evolve as its capabilities change and as real world experience and challenges are encountered from real world implementation. Domain experts and health systems should work collaboratively to develop regulations and monitoring systems.