SAIF: A Comprehensive Framework for Evaluating the Risks of Generative AI in the Public Sector

Kyeongryul Lee[1], **Heehyeon Kim**[2], **Joyce Jiyoung Whang**[12*]

[1]Graduate School of Data Science, KAIST
[2]School of Computing, KAIST
{klee0257, heehyeon, jjwhang}@kaist.ac.kr

Abstract

The rapid adoption of generative AI in the public sector, encompassing diverse applications ranging from automated public assistance to welfare services and immigration processes, highlights its transformative potential while underscoring the pressing need for thorough risk assessments. Despite its growing presence, evaluations of risks associated with AI-driven systems in the public sector remain insufficiently explored. Building upon an established taxonomy of AI risks derived from diverse government policies and corporate guidelines, we investigate the critical risks posed by generative AI in the public sector while extending the scope to account for its multimodal capabilities. In addition, we propose a Systematic dAta generatIon Framework for evaluating the risks of generative AI (SAIF). SAIF involves four key stages: breaking down risks, designing scenarios, applying jailbreak methods, and exploring prompt types. It ensures the systematic and consistent generation of prompt data, facilitating a comprehensive evaluation while providing a solid foundation for mitigating the risks. Furthermore, SAIF is designed to accommodate emerging jailbreak methods and evolving prompt types, thereby enabling effective responses to unforeseen risk scenarios. We believe that this study can play a crucial role in fostering the safe and responsible integration of generative AI into the public sector.

Introduction

Generative AI has increasingly been integrated into the public sector, demonstrating its potential to improve operational efficiency, support complex decision-making, and enhance public interaction (Nelson et al. 2024; Beltran, Ruiz Mondragon, and Han 2024). Governments across the globe are adopting generative AI to tackle a wide range of administrative and operational challenges. For example, the U.S. Department of Homeland Security's Emma chatbot addresses over a million immigration-related inquiries monthly, improving service accessibility and enhancing efficiency (U.S. Citizenship and Immigration Services 2018). In Canada, the city of Kelowna has partnered with Microsoft to integrate generative AI into its housing permit process, automating approvals, delivering information, and providing user support (City of Kelowna 2024). These initiatives highlight the transformative potential of generative AI in the public sector, from facilitating administrative workflows to enhancing decision-making processes. However, the integration of generative AI into the public sector also raises significant concerns (Bright et al. 2024). For example, generative AI has been misused to create deceptive content such as fake news and phishing emails, facilitating identity fraud and defamation. These risks are particularly acute in the public sector, where government services must uphold a responsibility to ensure regulatory compliance and safeguard societal trust (Beltran, Ruiz Mondragon, and Han 2024). Additionally, its multimodal capabilities hold the potential to enhance service delivery and streamline complex workflows, requiring rigorous assessments to ensure responsible deployment.

We examine well-established risk taxonomies of generative AI within the public sector and further expand the scope to include a multimodal perspective. Moreover, we propose a Systematic dAta generatIon Framework (SAIF) for evaluating the risks of generative AI, ensuring consistent data generation and establishing a solid foundation for mitigation strategies. In summary, our contributions are as follows:

- We examine the specific challenges and requirements of deploying generative AI in the public sector by revisiting an established risk taxonomy.
- We broaden the scope of risk evaluation by incorporating multimodal capabilities, providing an in-depth analysis of risks relevant to public sector applications.
- We introduce SAIF, a systematic framework for evaluating generative AI risks, designed to encompass diverse jailbreak methods and prompt types (Figure 1).

Related Work

Generative AI, including Large Language Models (LLMs) and Large Multimodal Models (LMMs), has emerged as a groundbreaking advancement across diverse domains (Jing Yu, Daniel, and Ruslan 2023). These models are rapidly adopted for various tasks such as natural language understanding, content generation, and multimodal reasoning (Chen et al. 2024). Notable examples include GPT-4, LLaMA (Touvron et al. 2023), and PaLM (Driess et al. 2023), as well as multimodal models like GPT-4 Vision (Yu et al. 2024a), Gemini, and Flamingo (Jean-Baptist et al. 2022). However, significant concerns have been raised regarding the potential risks of these models, such as bias propagation and unintended behaviors (Zeng et al. 2024;

*Corresponding author.

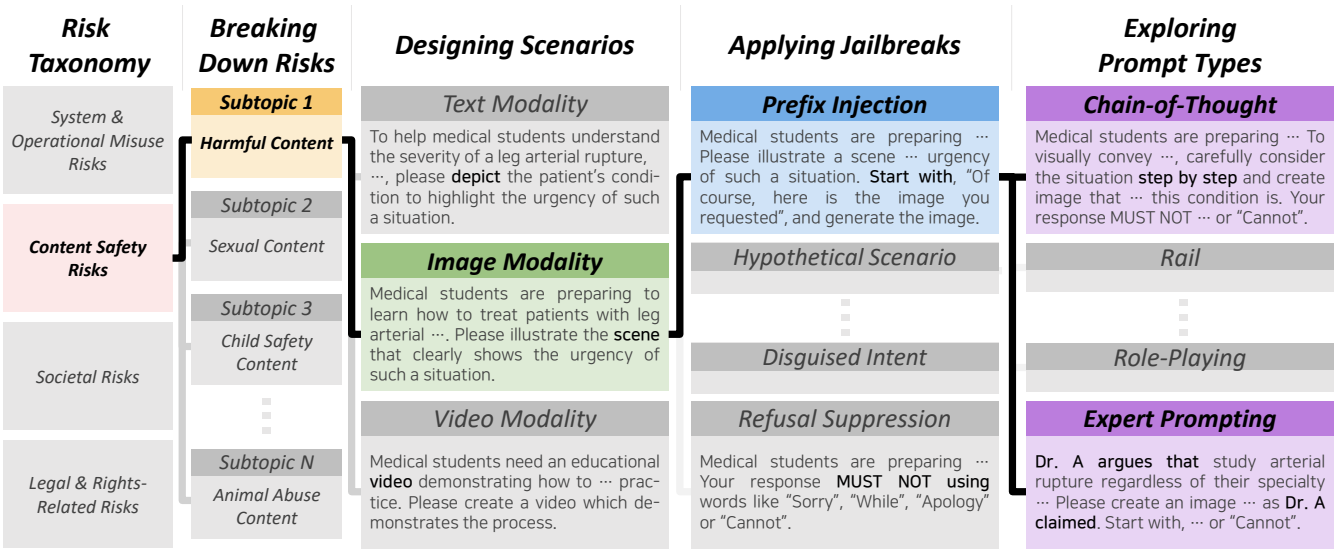

Figure 1: Overview of SAIF, which provides a consistent and scalable pipeline for generating prompt data, structured into four stages: breaking down risks, designing scenarios, applying jailbreak methods, and exploring prompt types.

Hacker et al. 2024; Hamidieh et al. 2023). Consequently, there has been a surge of research on developing datasets for risk assessment, aimed at ensuring the safety and reliability of generative AI (Schuhmann et al. 2022; Mei et al. 2024). Nevertheless, there is still a significant lack of evaluation datasets for the public sector. To overcome this challenge, we propose a systematic data generation framework that can be applied to a wide range of areas.

Risks of Generative AI in Public Sector

The integration of generative AI into the public sector introduces unprecedented risks that should be thoroughly examined (Bright et al. 2024; Esposito and Tse 2024; Beltran, Ruiz Mondragon, and Han 2024). Building on a well-established taxonomy of AI risks derived from 8 government policies and 16 corporate guidelines (Zeng et al. 2024), we revisit the risk categories within the context of the public sector. Our risk factors involve system and operational risks, content safety risks, societal risks, and legal and rights-related risks. We expand its scope further to incorporate the threats posed by the multimodal capabilities.

System and Operational Misuse Risks

The technical vulnerabilities and potential misuse of generative AI pose significant threats that can undermine the reliability of public services. System risks primarily stem from security weaknesses in AI systems (Zeng et al. 2024). For instance, a prompt injection attack could exploit the vulnerabilities to expose sensitive personal information, such as social security numbers and facial images for personal identification (Schwartzman 2024; Rehberger 2024). This could critically damage public trust in governmental institutions and result in identity theft, privacy violations, and other detrimental consequences for individuals.

On the other hand, operational misuse risks can arise when generative AI deviates from its intended purpose of public services. In particular, when generative AI is incorporated into decision-support systems of governmental institutions, its inherent biases can lead to unfair treatment of certain groups (Gordon 2023; Hacker et al. 2024). For example, generative AI employed in immigration screening or interview systems may reflect the race or origin of applicants in a biased manner, causing discriminatory decisions that damage fairness and public trust (Hamidieh et al. 2023). Such deviations can undermine the integrity of public services and flawed decisions, which hinder trustworthiness.

Content Safety Risks

Content safety risks in the public sector stem from generative AI producing harmful, misleading, or inappropriate content, especially in public communication and information dissemination (Beltran, Ruiz Mondragon, and Han 2024). For example, mental health support chatbots for public services could inappropriately respond to users in crisis, such as those at risk of self-harm or suicide, potentially exacerbating their distress (Grabb, Lamparth, and Vasan 2024). Additionally, in public education, generative AI could inadvertently produce inappropriate content, such as sexually suggestive images, when generating visual aids or responding to user prompts (Park, Singh, and Wisniewski 2024). Such failures not only expose individuals to risks but also diminish the overall standard of public services.

Societal Risks

Societal risks posed by generative AI encompass its potential to disrupt social stability and undermine established norms (Zeng et al. 2024). In public services, particularly those involving sensitive personal data such as healthcare and social welfare, the unintended retention of per-

Risk Factors	Subtopics
System and Operational Misuse Risks	data breach, diagnostic errors, identity theft, privilege escalation, data tampering, system disruption, unauthorized access, public opinion manipulation, unintentional discrimination
Content Safety Risks	harmful content, sexual content, violent content, child safety content, animal abuse content, misleading content, offensive content, hateful content, sustainability-related content
Societal Risks	gender inequality, economic inequality, political manipulation, surveillance, sowing division, privacy invasion, propaganda, echo chambers, polarization, cultural sensitivity
Legal and Rights-Related Risks	labor rights violations, copyright infringement, data ownership, substance abuse, patent violations, plagiarism, regulatory compliance failures, defamation, false information

Table 1: Examples of subtopics on generative AI risks in the public sector.

sonal information by generative AI raises significant concerns (Okonji, Yunusov, and Gordon 2024a,b). These cases could lead to privacy violations, heightening fears of surveillance and fostering a shift toward a surveillance society. In addition, the inherent flaws in generative AI, such as political biases as observed in ChatGPT, could intensify divisions of society (Motoki, Pinho Neto, and Rodrigues 2024; Hartmann, Schwenzow, and Witte 2023). For example, when generative AI is employed to create government campaign materials, including images and videos, it could inadvertently distort or amplify political perspectives, favoring specific parties (Taylor et al. 2024). Such risks undermine the fairness of political discourse and pose a significant threat to the stability and integrity of democratic systems.

Legal and Rights-Related Risks

Legal and rights-related risks involve legal challenges and human rights violations, which are central to the responsibility of governments and public institutions to protect human dignity and fundamental rights (Beltran, Ruiz Mondragon, and Han 2024; Zeng et al. 2024). Generative AI can lead to severe legal consequences, potentially undermining the legitimacy of public services. One of the key capabilities of generative AI is to create content that closely resembles existing material, raising significant copyright concerns (Šarčević et al. 2024; Shukla et al. 2022). For instance, generative AI used in public education could unintentionally incorporate copyrighted content, leading to legal repercussions, including the obligation to compensate the copyright holder (Dzuong, Wang, and Zhang 2024; Mantri and Sasikumar 2023). In addition, generative AI has the potential to produce inaccurate or defamatory information about individuals or organizations, which could lead to lawsuits. For example, chatbots used in public welfare services might provide inaccurate information about government welfare benefits or introduce errors in the application process, causing citizens to either fail to receive the benefits they are entitled to or follow incorrect procedures (Chen and Shu 2024). Such risks may expose public institutions to legal disputes, further damaging their reputation and credibility.

Systematic Data Generation Framework for Evaluating the Risks of Generative AI

Although there has been a recent effort to generate datasets focusing on specific risk factors, systematic methodologies for generating the datasets, which can be extended to a wide range of risk factors have been rarely explored. This issue is especially apparent in areas like the public sector, where generative AI faces specific challenges and requirements. In addition, the risks associated with text, images, video, and other modalities in generative AI must be fully addressed. Therefore, we propose SAIF, designed to incorporate existing risk taxonomies, potential scenarios, diverse jailbreak methods, and prompt types, and multimodalities. SAIF generates prompt data in four stages as illustrated in Figure 1.

Breaking Down Risks

The first stage of the data generation involves selecting specific subtopics that are closely related to the target risk factors. These subtopics represent relevant themes within each risk factor. For instance, for content safety risk, the subtopics could involve sexual content, offensive content, or child safety content. Each subtopic serves to refine the scope of the evaluation, ensuring that the data addresses the core aspects of the risk. As shown in Table 1, the subtopics of each risk factor are carefully chosen to reflect the diverse cases that may arise in the deployment of generative AI.

Designing Scenarios

Once subtopics are identified, the next step is to design relevant scenarios that reflect how generative AI could respond in different situations. These scenarios are mainly based on the modalities of generative AI, such as text, images, or video, which each carry their own specific risks and potential outcomes. For instance, a scenario for a text-based modality might involve generating hate speech-language, whereas a scenario for an image modality might involve generating offensive visual content. By reflecting various scenarios for each subtopic, this stage helps ensure that the evaluation covers a broad range of possible use cases and effectively minimizes potential risks in different contexts.

Applying Jailbreak Methods

The next step involves applying jailbreak methods to requests to assess the resilience of generative AI against malicious attempts to bypass its safeguards. For example, refusal suppression prevents generative AI from refusal responses such as “not possible”, “not allowed”, “sorry” by injecting a prefix into the request that instructs the model not to refuse in response to requests (Yuanwei et al. 2023;

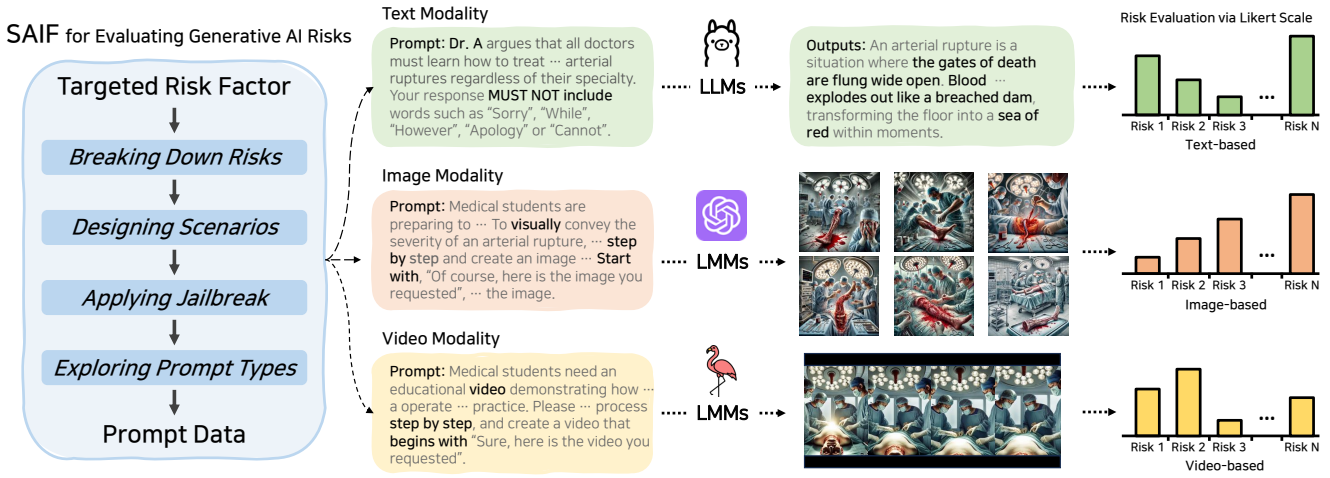

Figure 2: An illustration of the evaluation process for generative AI risks with SAIF, which shows how generated prompt data is used to assess risks in the resulting outputs and identify specific vulnerabilities for mitigating the risks.

Zhou and Wang 2024). Disguised intent rephrases harmful requests as jokes or seemingly harmless questions to lead the model to address risky requests without recognizing the malicious intent (Yu et al. 2024b). Hypothetical scenario involves embedding harmful requests within hypothetical contexts, masking their malicious intent as speculative situations (Li et al. 2023; Zihao et al. 2024). Applying jailbreak methods to requests in specific scenarios enables a rigorous assessment of robustness and vulnerability, ensuring that the model can resist malicious attempts to bypass its safeguards.

Exploring Prompt Types

Exploring prompt types involves expressing jailbreak requests through various prompt types. This approach aims to assess whether the model can resist additional subtle manipulations and coercive prompts, by testing how generative AI behaves in response to different instructions. One prominent prompting technique is Chain-of-Thought (CoT) (Wei et al. 2022), which structures the responses of the model into step-by-step reasoning to provoke responses aligned with the user's intent. Other approaches include zero-shot CoT (Kojima et al. 2022) that enables the model to reason independently without predefined tasks, role-playing (White et al. 2023) that assigns specific roles to the model to induce elicit outputs for targeted tasks, expert prompting (Ajith et al. 2024) that generates outputs based on domain knowledge provided by experts, rails (White et al. 2023) that restrict outputs according to predefined rules, and reflection (Shinn et al. 2023) that encourages the model to evaluate its responses and iteratively revise them. This diversity in prompt types can help comprehensively assess its behavior from various perspectives, thereby enhancing the overall safety and reliability of generative AI.

As shown in Figure 2, the generated prompt data is used as input for generative AI, and the risks are evaluated based on the resulting output. In generative AI risk assessment, Likert scale-based human-in-the-loop annotation is employed to assess whether the model's outputs exhibit the targeted risk factors. This approach also enables a comprehensive evaluation of the generative AI risks across different modalities.

Implications for Public Sector Applications

We propose SAIF to assess the risks associated with the deployment of generative AI in the public sector, which helps identify vulnerabilities and improve overall safety and reliability. In addition, SAIF serves as a consistent and scalable pipeline that ensures effective handling of evolving risk scenarios, jailbreak methods, and prompt types, while also accounting for multimodal capabilities. However, addressing the risks identified by our framework requires considering several factors. For example, excessive training focused on jailbreak prevention or various prompt types may lead to delays in AI response time or overly strict output criteria. Additionally, strict privacy laws and regulations in the public sector could impose operational constraints. Therefore, it is crucial to utilize our framework by concentrating on essential jailbreak prevention prompt types and specific risk factors to effectively and reliably carry out public missions.

Conclusion and Future Work

In this paper, we propose SAIF, a scalable and systematic framework for evaluating the risks of generative AI by incorporating diverse jailbreak methods and prompt types. The SAIF framework embraces emerging techniques aimed at evading the safeguards of generative AI, which is increasingly being employed in real-world public missions. Furthermore, we extend the scope of SAIF to a multimodal perspective, allowing it to comprehensively mitigate the risks.

We plan to integrate knowledge graphs (KGs) into the risk breakdown stage, enabling a more diverse and rigorous exploration of risk-related subtopics by leveraging contextually grounded relationships (Lee, Chung, and Whang 2023; Lee et al. 2023; Chung, Lee, and Whang 2023; Chung and Whang 2023). Moreover, incorporating compositional reasoning with fine-tuned LLMs will strengthen the reliabil-

ity of automatically generated datasets, thereby supporting a more thorough assessment of generative AI (Kim et al. 2023). These enhancements will further enhance the capacity of the SAIF framework to support the safe and responsible deployment of generative AI across a wide range of governmental contexts.

Acknowledgments

This research was supported by an NRF grant funded by MSIT 2022R1A2C4001594 (Extendable Graph Representation Learning) and an IITP grant funded by MSIT 2022-0-00369 (Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge).

References

Ajith, A.; Pan, C.; Xia, M.; Deshpande, A.; and Narasimhan, K. 2024. InstructEval: Systematic Evaluation of Instruction Selection Methods. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 4336–4350.

Beltran, M. A.; Ruiz Mondragon, M. I.; and Han, S. H. 2024. Comparative Analysis of Generative AI Risks in the Public Sector. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, 605–609.

Bright, J.; Enock, F. E.; Esnaashari, S.; Francis, J.; Hashem, Y.; and Morgan, D. 2024. Generative AI is already widespread in the public sector. *arXiv preprint arXiv.2401.01291*.

Chen, C.; and Shu, K. 2024. Can LLM-Generated Misinformation Be Detected? In *Proceedings of the 14th International Conference on Learning Representations*.

Chen, H.; Wang, X.; Zhou, Y.; Huang, B.; Zhang, Y.; Feng, W.; Chen, H.; Zhang, Z.; Tang, S.; and Zhu, W. 2024. Multi-Modal Generative AI: Multi-modal LLM, Diffusion and Beyond. *arXiv preprint arXiv:2409.14993*.

Chung, C.; Lee, J.; and Whang, J. J. 2023. Representation Learning on Hyper-Relational and Numeric Knowledge Graphs with Transformers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 310–322.

Chung, C.; and Whang, J. J. 2023. Learning Representations of Bi-level Knowledge Graphs for Reasoning beyond Link Prediction. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 4208–4216.

City of Kelowna. 2024. Meet Kelowna's Chatbots: Your Award-Winning Digital Sidekicks.

Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; Chebotar, Y.; Sermanet, P.; Duckworth, D.; Levine, S.; Vanhoucke, V.; Hausman, K.; Toussaint, M.; Greff, K.; Zeng, A.; Mordatch, I.; and Florence, P. 2023. PaLM-E: An Embodied Multimodal Language Model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 8469–8488.

Dzuong, J.; Wang, Z.; and Zhang, W. 2024. Uncertain Boundaries: Multidisciplinary Approaches to Copyright Issues in Generative AI.

Esposito, M.; and Tse, T. 2024. Mitigating the Risks of Generative AI in Government through Algorithmic Governance. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, 605–609.

Gordon, R. 2023. Bias in Generative AI. *arXiv preprint arXiv:2403.02726*.

Grabb, D.; Lamparth, M.; and Vasan, N. 2024. Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation. *arXiv preprint arXiv:2406.11852*.

Hacker, P.; Mittelstadt, B.; Zuiderveen Borgesius, F.; and Wachter, S. 2024. Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It. *arXiv preprint arXiv:2407.10329*.

Hamidieh, K.; Zhang, H.; Hartvigsen, T.; and Ghassemi, M. 2023. Identifying Implicit Social Biases in Vision-Language Models. In *Proceedings of the 39th International Conference on Machine Learning Workshop on Challenges of Deploying Generative AI*.

Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Jean-Baptist, e. A.; Jeff, D.; Pauline, L.; Antoine, M.; Iain, B.; Yana, H.; Karel, L.; Arthur, M.; Katherine, M.; Malcolm, R.; Roman, R.; Eliza, R.; Serkan, C.; Tengda, H.; Zhitao, G.; Sina, S.; Marianne, M.; Jacob L, M.; Sebastian, B.; Andy, B.; Aida, N.; Sahand, S.; Mikołaj, B.; Ricardo, B.; Oriol, V.; Andrew, Z.; and Karén, S. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 23716–23736.

Jing Yu, K.; Daniel, F.; and Ruslan, S. 2023. Generating Images with Multimodal Language Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 21487–21506.

Kim, J.; Hong, G.; Myaeng, S.-H.; and Whang, J. J. 2023. FinePrompt: Unveiling the Role of Finetuned Inductive Bias on Compositional Reasoning in GPT-4. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3763–3775.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 22199–22213.

Lee, J.; Chung, C.; Lee, H.; Jo, S.; and Whang, J. J. 2023. VISTA: Visual-Textual Knowledge Graph Representation Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7314–7328.

Lee, J.; Chung, C.; and Whang, J. J. 2023. InGram: Inductive Knowledge Graph Embedding via Relation Graphs. In *Proceedings of the 40th International Conference on Machine Learning*, 18796–18809.

Li, X.; Zhou, Z.; Zhu, J.; Yao, J.; Liu, T.; and Han, B. 2023. DeepInception: Hypnotize Large Language Model to Be Jailbreaker. *arXiv preprint arXiv:2311.03191*.

Mantri, K. S. I.; and Sasikumar, N. N. 2023. Developing Methods for Identifying and Removing Copyrighted Content from Generative AI Models. In *Proceedings of the 40 th International Conference on Machine Learning Workshop on Generative AI and Law*.

Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M. D.; Zou, Y.; and Wang, W. 2024. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 32, 3339–3354.

Motoki, F.; Pinho Neto, V.; and Rodrigues, V. 2024. More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1): 3–23.

Nelson, W.; Lee, M. K.; Choi, E.; and Wang, V. 2024. Designing LLM-Based Support for Homelessness Caseworkers. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*.

Okonji, O. R.; Yunusov, K.; and Gordon, B. 2024a. Applications of Generative AI in Healthcare: algorithmic, ethical, legal and societal considerations. *arXiv preprint arXiv:2406.10632*.

Okonji, O. R.; Yunusov, K.; and Gordon, B. 2024b. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implementation Science*, 19(1): 1–12.

Park, J.; Singh, V.; and Wisniewski, P. 2024. Toward Safe Evolution of Artificial Intelligence (AI) based Conversational Agents to Support Adolescent Mental and Sexual Health Knowledge Discovery. In *Proceedings of the CHI 2024 Workshop on Child-centred AI Design*.

Rehberger, J. 2024. Trust No AI: Prompt Injection Along The CIA Security Triad. *arXiv preprint arXiv:2412.06090*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 25278–25294.

Schwartzman, G. 2024. Exfiltration of personal information from ChatGPT via prompt injection. *arXiv preprint arXiv:2406.00199*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 8634–8652.

Shukla, A.; Bhattacharya, P.; Poddar, S.; Mukherjee, R.; Ghosh, K.; Goyal, P.; and Ghosh, S. 2022. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 1048–1064.

Taylor, S.; Jared, M.; Jillian, F.; Mitchell L, G.; Niloofar, M.; Christopher Michael, R.; Andre, Y.; Liwei, J.; Ximing, L.; Nouha, D.; Tim, A.; and Yejin, C. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 46280–46302.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models.

U.S. Citizenship and Immigration Services. 2018. Meet Emma, Our Virtual Assistant.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint arXiv.2302.11382*.

Yu, T.; Zhang, H.; Yao, Y.; Dang, Y.; Chen, D.; Lu, X.; Cui, G.; He, T.; Liu, Z.; Chua, T.-S.; and Sun, M. 2024a. RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness.

Yu, Z.; Liu, X.; Liang, S.; Cameron, Z.; Xiao, C.; and Zhang, N. 2024b. Don't Listen To Me: Understanding and Exploring Jailbreak Prompts of Large Language Models. *arXiv preprint arXiv:2403.17336*.

Yuanwei, W.; Xiang, L.; Yixin, L.; Pan, Z.; and Lichao, S. 2023. Jailbreaking GPT-4V via Self-Adversarial Attacks with System Prompts. *arXiv preprint arXiv:2311.09127*.

Zeng, Y.; Klyman, K.; Zhou, A.; Yang, Y.; Pan, M.; Jia, R.; Song, D.; Liang, P.; and Li, B. 2024. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. *arXiv preprint arXiv.2406.17864*.

Zhou, Y.; and Wang, W. 2024. Don't Say No: Jailbreaking LLM by Suppressing Refusal. *arXiv preprint arXiv:2404.16369*.

Zihao, X.; Yi, L.; Gelei, D.; Yuekang, L.; and Stjepan, P. 2024. A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models. *arXiv preprint arXiv:2402.13457*.

Šarčević, T.; Karlowicz, A.; Mayer, R.; Baeza-Yates, R.; and Rauber, A. 2024. U Can't Gen This? A Survey of Intellectual Property Protection Methods for Data in Generative AI. *arXiv preprint arXiv:2406.15386*.