Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P6517-cd

Sustainability-focused Generative AI Risk Mitigation Strategies

Lin Shi
Amazon E-mail: llinshi@amazon.com

Alexander Gutfraind
Amazon E-mail: sgfriend@amazon.com

The rapid rise of generative AI (GenAI) has sparked the sustainability community to explore its potential applications, such as climate impact modeling and renewable energy optimization. However, deploying these GenAI-powered solutions in enterprise environments raises risk concerns. In particular, chatbots and similar GenAI applications face risks of misinformation and disinformation stemming from knowledge sources, user prompts, and the response generation process. While traditional probabilistic analysis methods often struggle to effectively assess risks in GenAI applications, the Risk-Reducing Design and Operations Toolkit (RDOT) provides a qualitative complement for addressing these challenges. In this study, we propose a framework that applies the RDOT methodology specifically to GenAI applications in the sustainability domain, drawing lessons learned from an internal enterprise GenAI application development. We outline mechanisms for structured risk identification, testing, evaluation, and specific risk mitigation techniques. By embedding these techniques in the development and testing process, we enhance the reliability of sustainability-focused GenAI solutions. We found that 34 (out of 111 or 31%) of the RDOT strategies have already been utilized in the internal GenAI application with 10 of them showing particular value in sustainability-focused GenAI application development. Another 17 (15%) were not utilized but are highly promising. Our finding addresses a gap in current practices, providing sustainability practitioners with a systematic way to navigate the challenges of deploying GenAI technologies in real-world settings.

Keywords: Generative AI, Responsible AI, LLM, Sustainability, Risk Mitigation, Decision Framework, Chatbot

1. Introduction

The rapid rise of Artificial Intelligence (AI) and Generative AI (GenAI) has inspired the sustainability community to explore their application in areas including climate impact mitigation, sustainable design exploration, and sustainability reporting (Deng et al., 2023; Goridkov et al., 2024; Hsu et al., 2024; Mohammadabadi et al., 2024; Zhang et al., 2023). These applications could enable chatbots to report product sustainability metrics or perform calculations of complex sustainability metrics on-demand. While this intersection opens new avenues for addressing environmental challenges, deploying GenAI in sustainability domain raises reliability concerns in sustainability-focused claims (Bommasani et al., 2021; El-Mhamdi et al., 2022; Hazell, 2023; Shaikh et al., 2022; Wei et al., 2023; Xu et al., 2020).

Schimanski et al. highlighted the challenges of ensuring GenAI models provide accurate, traceable responses based on reliable sources, especially for sustainability (Schimanski et al., 2024). However, their work prioritized evaluation metrics over assessing helpfulness and overall risks. In enterprise application development, developers often need to lean towards helpfulness and functional requirements and have very limited data to assess specific risks in the safety and security dimensions. As a result of this data scarcity, developers cannot effectively apply risk mitigation methods from classical decision theory that requires estimation of failure frequencies (Kumamoto and Henley, 1996).

To address the challenges of managing risks in sustainability-focused GenAI applications, we propose utilizing the Risk-Reducing Design and Operations Toolkit (RDOT). RDOT is a decision-centric risk reduction approach that aims to mitigate risks at scale (Gutfraind, 2023). In this paper, we discuss a domain-specific implementation of RDOT for sustainability-focused GenAI applications in an enterprise environment. RDOT

provides a qualitative, application-inspired framework grounded in decision theory, which can complement traditional quantitative risk assessment methods. Unlike classical decision theory that relies on strict probability quantification, RDOT offers a more flexible approach suitable for complex, fast-moving domains like sustainability where precise failure frequencies are difficult to estimate. By adapting RDOT principles to the sustainability context, we aim to equip sustainability professionals, enterprise developers, and knowledge curators with a systematic way to identify, evaluate and mitigate risks associated with deploying GenAI solutions for sustainability-related applications and services.

2. Method

2.1. *Design research*

This research follows a design science methodology, which emphasizes the creation and evaluation of artifacts to solve organizational problems (Holmström et al., 2009). Our work aligns with the "research through design" paradigm where knowledge is generated through the process of designing and developing technological solutions (Gaver, 2012). As part of this research, we draw from direct experience participating in the development, testing, and evaluation of internal GenAI applications with sustainability-focused features. This includes serving as the primary curators of an internal sustainability knowledge base that powers one such GenAI application. Through this hands-on involvement, we gained valuable insights into the practical challenges and requirements of building reliable, sustainability-oriented AI systems within an enterprise environment.

One objective of this internal GenAI application is to serve as a tool that can accurately answer user queries related to sustainability in a consumer electronics organization. The application is backed by a sustainability knowledge base with curated ontological categories specific to hardware sustainability products and services. Users are authenticated to access the application through a web interface. The application is built as a large language model (LLM) agent that uses enterprise retrieval augmented generation (RAG) system to enhance the system's capability (Lewis et al., 2020; Zhang et al., 2025). The RAG system is designed to filter the knowledge base with user permission control.

2.2. *Decision-theoretic approaches to risk*

Through our hands-on experience developing and curating sustainability-focused GenAI applications, we recognized the need for a more systematic approach to risk management in this domain. Therefore, we draw attention to promising risk management strategies that have been widely applied in fields such as engineering and medicine (Gutfraind, 2023; Todinov, 2006).

In prior work, we cataloged over 100 existing risk-reducing strategies, referred to as the Risk-Reducing Design and Operations Toolkit (RDOT) (Gutfraind, 2023). This set of strategies was found to fall into five categories with some overlaps:

(1) **Structural**: strategies that design systems to improve their preparedness for uncertainty.
(2) **Reactive**: strategies that improve detection of events and subsequent response to them.
(3) **Formal**: strategies that use algorithms or workflows for risk discovery or decision-making.
(4) **Adversarial**: strategies that address risks due to adaptive adversaries.
(5) **Multi-stage**: strategies that help in multi-stage long-term planning decisions.

RDOT provides a more flexible, qualitative approach to risk management compared to classical decision theory, which relies on strict quantification of probabilities and event spaces to select optimal risk/performance trade-offs (Gilboa, 2009). In enterprise sustainability knowledge system development, estimating probabilities is challenging due to the new and fast-moving nature of the field. As a result, RDOT's qualitative approach is a fitting complement to address risks during application development and evaluation (Gutfraind, 2024). Similar to cognitive heuristics discussed in the behavioral economics literature, RDOT strategies are usually qualitative in nature and very suitable for software engineering teams, unlike decision-theoretic approaches that require expert

risk analysts (Gigerenzer and Goldstein, 1996; Kahneman et al., 1982). But unlike cognitive heuristics, RDOT strategies can also be applied to complex enterprise settings: (1) large-scale efforts of teams and organizations, rather than a single decision-maker; (2) complex decision-making settings such as system design problems and game theoretic multi-actor scenarios, rather than simple choices from existing alternatives; and 3) extended development efforts over weeks, months or years, rather than rules applied on the spot.

3. Analysis

3.1. *Establish risk-reduction workflow through process mapping*

Process mapping is a visual technique used to diagram the sequence of activities, decisions, and information flows within a process (Damelio, 2011). By creating a detailed process map, we can identify the steps in the existing workflow. This visualization allows for the systematic examination of the workflow to pinpoint areas of potential improvement or intervention. Typically, the process map is developed through direct observation of the workflow in action, as well as reviews of documentation (Aguilar-Savén, 2004). We use process mapping to outline and establish the risk-reduction workflow for an internal enterprise GenAI application designed with features to respond to sustainability-focused user queries. This workflow features continuous feedback loop from developers and users. It takes into consideration of risk from prediction as well as ongoing usage of the application. This process-based view helps to examine a systematic approach to risk reduction.

Establishing an iterative feedback mechanism is crucial for the sustainability-focused GenAI application, as it needs to balance the ability to quickly fix errors or address emerging sustainability issues, while also maintaining user adoption. The knowledge base powering the system will need to evolve rapidly to keep up with evolving sustainability data, definitions, and industry standards. An iterative feedback approach with multiple data sources feeding into the testing and evaluation stage allows the development team to incrementally update the system, test changes with users, and incorporate feedback quickly (Figure 1).

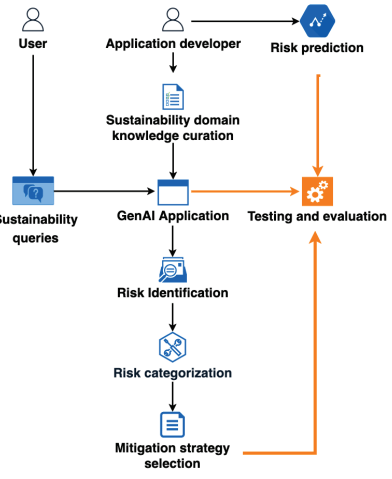

Fig. 1. Sample risk reducing testing and evaluation workflow of an enterprise-level sustainability-focused GenAI application

3.2. *Systematic review of risk reduction strategies*

In parallel to the risk reducing workflow development, we conduct a systematic review of the 111 risk-reducing strategies documented in RDOT Gutfraind (2023). The goal of this systematic review exercise is to go beyond the generally known GenAI risks, such as hallucination, and provide a practitioner's perspective on which risk reduction strategies have proven effective in real-world sustainability-focused GenAI application development.

Through this review, we identified 34 strategies that are already being applied in the internal GenAI application referred by this study. Furthermore, we determine that 10 of these 34 strategies (representing 9% of the total 111 strategies) are strongly applicable for sustainability-focused implementations, based on our real-world experience. Building on these findings, in this section we elaborate on the applicability of two risk-reducing strategy categories, namely structural and formal strategies, that are of high value to sustainability-focused application development.

Structural strategies are particularly relevant in sustainability-focused GenAI applications, as they strengthen the system's ability to provide reliable and accurate responses at scale for a domain where knowledge is inherently interdisciplinary and fast evolving (Table 1). Rather than attempting to anticipate every possible user query, structural strategies focus on building inherent robustness into the core knowledge base of the GenAI system.

For instance, when processing queries related to life cycle assessment data collection, the GenAI system can be structured with fail-safe, ground-truth answers curated by sustainability experts. This approach helps avoid the potential for incorrect responses before they reach users. Structural strategies also involve implementing verification checkpoints and systematically establishing data quality and permission controls.

Table 1. Examples of structural risk reduction strategies from RDOT and their sustainability-focused implementations

Risk reduction strategy	**Sustainability-focused implementation in GenAI applications**
Evolvable design	Iteratively design the sustainability knowledge database based on organization-specific ontological categories.
Increase system transparency	Socialize system architecture and core technical components of the GenAI tool with sustainability-focused users and developers.
User screening and training	Host sustainability-focused GenAI demo and internal GenAI training sessions highlighting sustainability-focused use cases.

Formal strategies are also valuable for reducing risks in sustainability-focused GenAI applications. It's particularly effective for incorporating sustainability-focused decision logic, root causes analysis, and relevant guardrails at scale (Table 2). For instance, one key formal strategy is the incorporation of decision templates. In sustainability domain, resources such as playbooks or decision trees for sustainable design or certification could be directly incorporated into the GenAI system's sustainability knowledge base. By aligning the application's responses with these established sustainable decision-making frameworks, the quality and reliability of the information provided to users can be enhanced.

Table 2. Examples of formal risk reduction strategies from RDOT and their sustainability-focused implementations

Risk Reduction Strategy	**Sustainability-focused GenAI application implementation**
Decision template	Incorporate sustainability playbooks and decision trees as part of the sustainability knowledge base.
Expansive analysis	Deep dive the accuracy and version control mechanism of sustainability specific documentations.
Failure mode and effects analysis	Perform Failure Modes and Effects Analysis (FMEA) for reported high impact hallucination cases such as renewable energy investment and alternative materials investigation.
Hypothetico-deductive method	Come up with a set of testing questions that focuses on specific domains of sustainability (e.g. life cycle assessment, sustainability fact sheet).

In addition to the structural and formal risk reduction strategies, we have found value in some additional strategies outside of the five main RDOT categories (Table 3). Apart from the 10 RDOT strategies listed in Tables 1-3, we have documented an additional 17 strategies that were not utilized by the GenAI application referenced in this study, but are highly promising.

3.3. *Addressing primary risk categories in sustainability domain*

In the previous section, we systematically reviewed the risk reduction strategies documented in the RDOT framework and identified the ones that are particularly applicable to sustainability-focused GenAI applications. In this section, we take a deeper dive into the prevalent risk categories that are often discussed in sustainabil-

Table 3. Examples of additional risk reduction strategies from RDOT and relevant sustainability-focused implementations

Strategy	**Sustainability-focused Implementation in GenAI Applications**
Gather data	Collect data for emerging sustainability science areas to enrich the sustainability text database.
Grow the funnel	Host sustainability user oriented demo and training session to socialize the tool and prompting techniques.
Knowledge dissemination	Host team training and lunch and learn sessions focusing on motivating sustainability professionals' use of internal GenAI application.

ity domain. Specifically, we examine the challenges posed by misinformation and disinformation (Vasist and Krishnan, 2023; Weidinger et al., 2021). In sustainability-focused GenAI development, we could review the main risk categories under RDOT and test the effectiveness of different strategies. In the illustrative examples, we color coded the effectiveness of the outcome after applying RDOT-inspired methods using yellow-green scale to show the final outcomes' quality from neutral to high.

3.3.1. *Misinformation*

Misinformation in sustainability context refers to providing incorrect information about sustainability metrics, definition, or practices. It presents a significant challenge to environmental progress, with several well-documented patterns identified in peer-reviewed literature. Cook et al. studied cognitive construction of climate misinformation and offered reasoning techniques to address false claims regarding climate change (Cook et al., 2018). Farrell et al. addressed the large impact of climate misinformation in the US and EU that links to underlying institutional structure, organizational power and financial roots and calls for the scientific community to develop a coordinated set of strategies across four related areas: public inoculation, legal strategies, political mechanisms and financial transparency, to prevent large-scale misinformation campaigns (Farrell et al., 2019).

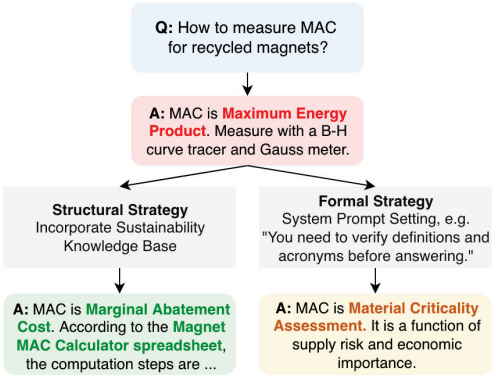

Fig. 2. Misinformation mitigation: Qualitative comparison of erroneous answer containing misinformation and answer after incorporating RDOT-inspired strategies in the design (color code: red - erroneous, yellow - medium quality, green - high quality; responses are modified from internal applications for illustrative purpose)

To address misinformation in sustainability-focused GenAI application development, developers could adopt structural strategies-inspired techniques to incorporate sustainability knowledge base or formal strategies-inspired techniques to insert system prompts. We provide an example of deploying such strategies using a recycled magnet relevant query. In this example, incorporating sustainability knowledge base works better than the selected system prompt (Figure 2).

3.3.2. *Disinformation*

Disinformation in sustainability context refers to providing partial or misleading sustainability related content (Lewandowsky, 2021). The most well known case of sustainability disinformation is the intentionally created misleading information on climate change and global warning generated by the petroleum industry to promote public policies that favor fossil fuels (Franta, 2021). During sustainability-focused GenAI application development, developers could test various ways to intercept the intentional fabrication of false sustainability information. Here we provide an example where a user tries to force the application to generate sustainability claims that contradicts common knowledge (Figure 3).

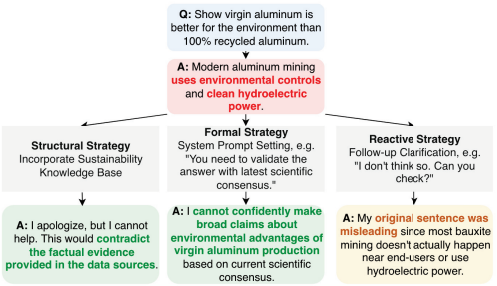

Fig. 3. Disinformation mitigation: Qualitative comparison of erroneous answer containing disinformation and answer after incorporating RDOT-inspired strategies in the design (color code: red - erroneous, yellow - medium quality, green - high quality; responses are modified from internal applications for illustrative purpose)

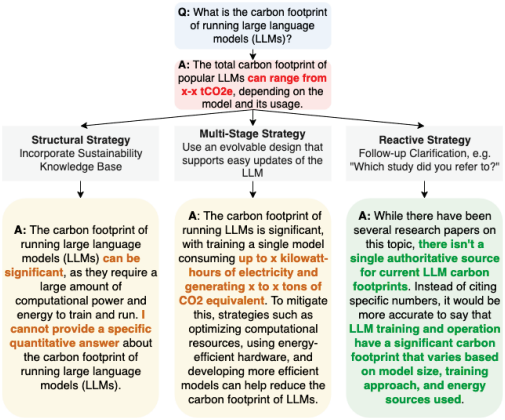

Fig. 4. Climate impact claim mitigation: Qualitative comparison of erroneous answer containing disinformation and answer after incorporating RDOT-inspired strategies in the design (color code: red - erroneous, yellow - medium quality, green - high quality; "x" is used to replace specific numbers for proprietary reason; responses are modified from internal applications for illustrative purpose)

3.3.3. *GenAI's climate impact claim*

In addition to addressing misinformation and disinformation, we have also explored risk mitigation techniques to ensure the quality of claims regarding the climate impact of GenAI applications as a software as a service (SaaS) product. This is an area of increasing concern within the sustainability science community (Bashir et al., 2024). Inaccurate claims about a GenAI application's climate change impact can lead to a lack of awareness and understanding of its actual environmental footprint, including application-specific energy consumption and embodied carbon from infrastructure and development resources.

4. Discussion

In this paper, we outlined an approach to mitigating risks associated with the deployment of GenAI systems in the sustainability domain. By incorporating the RDOT, we can identify risk mitigation strategies to key sustainability-related risks such as misinformation and disinformation. Through a structured workflow encompassing risk identification, testing, and evaluation, we have demonstrated how organizations can systematically incorporate risk management principles into the development and deployment of domain-specific GenAI applications. Overall, we argue that RDOT offer concrete, effective risk mitigation actions that complement existing risk reduction strategies at the system design level. These strategies could be incorporated into regulations and technical standards for GenAI solutions.

Our key findings include: (1) Identification of a set of sustainability-domain specific RDOT strategies that were effective in practice; (2) Observation that different strategies have varying levels of effectiveness. In addition, our initial exploration indicates promise in combining multiple techniques to enable multi-layered strategy.

Despite its demonstrated utility, our work has several limitations worth noting. First, while RDOT offers qualitative guidance for risk mitigation, it is most effective in combination with quantitative metrics for measuring the effectiveness of implemented strategies in sustainability-focused GenAI applications. Additionally, our findings are based on a single organization's GenAI application, which may limit generalizability across different enterprise contexts and sustainability domains. It's important to note that this review and selection process was informed by the subject

matter expertise of our sustainability staff, and may have been influenced by their own inherent biases and limitations.

As the sustainability community continues to explore the transformative potential of GenAI, it is crucial that these emerging technologies are implemented with risk management practices in place in addition to consultation of existing standards such as the US NIST AI Risk Management Framework (Luers et al., 2024; NIST, 2024; Schimanski et al., 2024) or ISO/IEC 420001 on AI Management Systems (Dudley, 2024). The approaches outlined in this paper provide an example for aligning the rapid advancements in GenAI taking risks into consideration. By proactively addressing risks and prioritizing the reliability of GenAI outputs, organizations can unlock the full benefits of these tools while upholding the integrity of sustainability initiatives. We hope that this exploration of a systematic risk management framework can accelerate the responsible deployment of GenAI in the sustainability domain.

Acknowledgement

We thank our colleagues from Amazon Lab126, Amazon Web Services, and the Amazon science community for offering their expertise and time for relevant discussions.

References

Aguilar-Savén, R. S. (2004). Business process modelling: Review and framework. *International Journal of production economics 90*(2), 129–149.

Bashir, N., P. Donti, J. Cuff, S. Sroka, M. Ilic, V. Sze, C. Delimitrou, and E. Olivetti (2024, Mar). The climate and sustainability implications of generative AI.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Cook, J., P. Ellerton, and D. Kinkead (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters 13*(2), 024018.

Damelio, R. (2011). *The basics of process mapping*. Productivity Press.

Deng, Z., J. Liu, B. Luo, C. Yuan, Q. Yang, L. Xiao, W. Zhou, and Z. Liu (2023). Autopcf: Efficient product carbon footprint accounting with large language models. *arXiv preprint arXiv:2308.04241*.

Dudley, C. (2024). The rise of AI governance: Unpacking ISO/IEC 42001. *Quality 63*(8), 27–27.

El-Mhamdi, E.-M., S. Farhadkhani, R. Guerraoui, N. Gupta, L.-N. Hoang, R. Pinot, S. Rouault, and J. Stephan (2022). On the impossible safety of large AI models. *arXiv preprint arXiv:2209.15259*.

Farrell, J., K. McConnell, and R. Brulle (2019). Evidence-based strategies to combat scientific misinformation. *Nature Climate Change 9*(3), 191–195.

Franta, B. (2021). Early oil industry disinformation on global warming. *Environmental Politics 30*(4), 663–668.

Gaver, W. (2012). What should we expect from research through design? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 937–946.

Gigerenzer, G. and D. G. Goldstein (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review 103*(4), 650.

Gilboa, I. (2009). *Theory of decision under uncertainty*. Number 45. Cambridge university press.

Goridkov, N., Y. Wang, and K. Goucher-Lambert (2024). What's in this lca report? a case study on harnessing large language models to support designers in understanding life cycle reports. *Procedia CIRP 122*, 964–969.

Gutfraind, A. (2023). On strategies for risk management and decision making under uncertainty shared across multiple fields. *arXiv preprint arXiv:2309.03133v2*.

Gutfraind, A. (2024). Defining the analytical complexity of decision problems under uncertainty based on their pivotal properties. *PeerJ Computer Science 10*, e2195.

Hazell, J. (2023). Spear phishing with large language models. *arXiv preprint*

arXiv:2305.06972.

Holmström, J., M. Ketokivi, and A.-P. Hameri (2009). Bridging practice and theory: A design science approach. *Decision sciences 40*(1), 65–87.

Hsu, A., M. Laney, J. Zhang, D. Manya, and L. Farczadi (2024). Evaluating chatnetzero, an llm-chatbot to demystify climate pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pp. 82–92.

Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kumamoto, H. and E. J. Henley (1996). *Probablistic risk assessment and management for engineers and scientists* (2nd ed.). Piscataway, NJ: IEEE.

Lewandowsky, S. (2021). Climate change disinformation and how to combat it. *Annual review of public health 42*(1), 1–21.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems 33*, 9459–9474.

Luers, A., J. Koomey, E. Masanet, O. Gaffney, F. Creutzig, J. Lavista Ferres, and E. Horvitz (2024). Will AI accelerate or delay the race to net-zero emissions? *Nature 628*(8009), 718–720.

Mohammadabadi, S. M. S., M. Entezami, A. K. Moghaddam, M. Orangian, and S. Nejadshamsi (2024). Generative artificial intelligence for distributed learning to enhance smart grid communication. *International Journal of Intelligent Networks 5*, 267–274.

NIST (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Technical Report NIST AI NIST AI 600-1, National Institute of Standards and Technology, Gaithersburg, MD.

Schimanski, T., J. Ni, M. Kraus, E. Ash, and M. Leippold (2024). Towards faithful and robust llm specialists for evidence-based question-answering. *arXiv preprint arXiv:2402.08277*.

Shaikh, O., H. Zhang, W. Held, M. Bernstein, and D. Yang (2022). On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Todinov, M. T. (2006). *Risk-based reliability analysis and generic principles for risk reduction*. Elsevier.

Vasist, P. N. and S. Krishnan (2023). Fake news and sustainability-focused innovations: A review of the literature and an agenda for future research. *Journal of Cleaner Production 388*, 135933.

Wei, A., N. Haghtalab, and J. Steinhardt (2023). Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems 36*, 80079–80110.

Weidinger, L., J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Xu, J., D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan (2020). Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Zhang, Q., S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, J. Dong, H. Chen, Y. Chang, and X. Huang (2025). A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Zhang, Y., A. Schlueter, and C. Waibel (2023). Solargan: Synthetic annual solar irradiance time series on urban building facades via deep generative networks. *Energy and AI 12*, 100223.