Systematic Literature Review Protocol for Designing a Taxonomy and Reference Architecture for Multi-Layered Guardrails of Foundation Model-Based Agents: Swiss Cheese Model for AI Safety

Md. Shamsujjoha*, Qinghua Lu, Dehai Zhao, and Liming Zhu

Data61, CSIRO, Australia

Email: {*md.shamsujjoha, qinghua.lu, dehai.zhao, liming.Zhu*}*@data61.csiro.au*

Abstract

This report outlines a systematic literature review protocol focused on designing a taxonomy and reference architecture for multi-layered guardrails of foundation model-based agents. The primary objectives are to identify existing approaches to guardrails, evaluate their quality attributes, and analyze the architectural considerations involved in various design choices. This review aims to provide a comprehensive understanding of the current research landscape, highlighting design trade-offs and proposing a comprehensive taxonomy and architecture for multi-layered runtime guardrails.

Keywords: Foundation Model, Large Language Models, LLM, Agent, Guardrails, Safeguard, AI Safety, Software Architecture, Taxonomy, Swiss Cheese Model, Responsible AI, Protocol, Systematic Literature Review.

1. Introduction

Foundation **M**odels (FMs) are large-scale machine learning models pre-trained on vast amounts of data using self-supervised learning (at scale). These models are designed to be highly versatile and can adapt to a wide range of downstream tasks [1]. An FM-based agent uses an FM as a core component, interacting with other AI or non-AI components to perform tasks autonomously [2]. In recent years, FM-based agents have been experiencing extensive growth [3]. However, the growing capabilities and autonomy of these agents raise significant concerns about responsible AI and AI safety, e.g., generating harmful or offensive content, producing dangerous or unintended outcomes, exhibiting discriminatory behavior, compromising user privacy, spreading misinformation, facilitating cyberattacks, etc [1, 4, 5]. To address these challenges, effective runtime guardrails are required to ensure the agent's behavior is responsible and safe. Multi-layered runtime guardrails structured similarly to the Swiss Cheese Model that operate at various levels of the agent architecture further mitigate associated risks [1, 6].

There have been some initial efforts on guardrails, such as input filtering [1, 7], output modification [8, 9], adaptive fail-safes that can prevent harmful outputs [10, 11], real-time

*✉ Corresponding Author: Md. Shamsujjoha

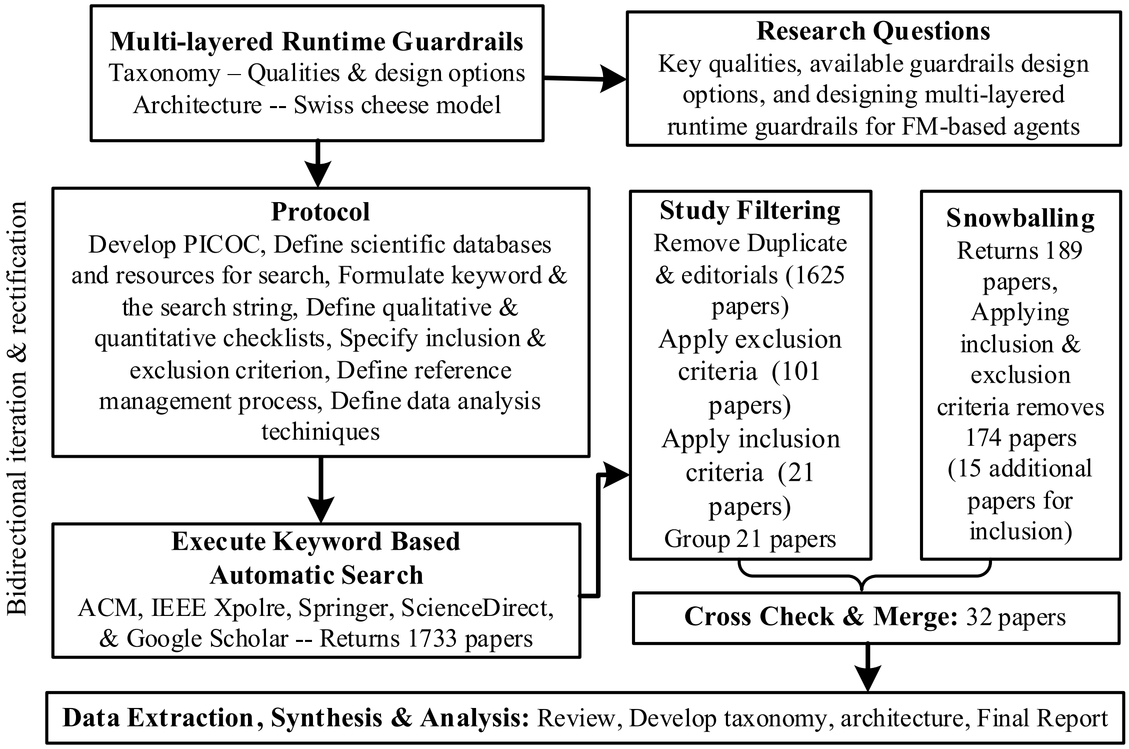
Multi-layered Runtime Guardrails
Taxonomy – Qualities & design options
Architecture -- Swiss cheese model
Research Questions
Key qualities, available guardrails design options, and designing multi-layered runtime guardrails for FM-based agents
Bidirectional iteration & rectification
Protocol
Develop PICOC, Define scientific databases and resources for search, Formulate keyword & the search string, Define qualitative & quantitative checklists, Specify inclusion & exclusion criterion, Define reference management process, Define data analysis techiniques
Study Filtering
Remove Duplicate & editorials (1625 papers)
Apply exclusion criteria (101 papers)
Apply inclusion criteria (21 papers)
Group 21 papers
Snowballing
Returns 189 papers,
Applying inclusion & exclusion criteria removes 174 papers
(15 additional papers for inclusion)
Execute Keyword Based Automatic Search
ACM, IEEE Xpolre, Springer, ScienceDirect, & Google Scholar -- Returns 1733 papers
Cross Check & Merge: 32 papers
Data Extraction, Synthesis & Analysis: Review, Develop taxonomy, architecture, Final Report

Figure 1: Methodology

monitoring and detection [12–15], continuous output validation to ensure adherence to ethical standards [16–18] etc. However, the existing runtime guardrails mainly focus on inputs and outputs of foundation models, which do not capture the complexity of FM-based agents. To better address these challenges, we conducted a **S**ystematic **L**iterature **R**eview (SLR) of existing research on 'FM-based agents' and 'guardrails'. In this report, we present the research protocol for this study, outlining the methodology, search strategies, inclusion and exclusion criteria, research questions and data synthesis approach.

2. Methodology

This SLR is conducted to identify, evaluate, and synthesize key available research relevant to multi-layered runtime guardrails in FM-based agents. We aimed to provide a comprehensive overview of the current state of research based on published primary studies. To achieve this, we followed the guidelines provided by Kitchenham and Charters [19] and adhered to the procedures outlined by Kitchenham [20]. Our methodology, which included a well-defined protocol for study selection, quality assessment, screening, and data extraction, was meticulously designed to mitigate bias and ensure repeatability.

The data from 32 primary studies (shown in Appendix A) were synthesized using meta-analysis techniques to quantify qualitative and quantitative descriptive findings. This approach provided an estimated average effect and variance. Information that could not be quantitatively synthesized due to heterogeneity was collected for narrative/descriptive synthesis. However, these narrative syntheses tend to be subjective and may vary with the

Table 1: PICOC for this Study

Population	Studies and researches focus on multi-layered runtime guardrails within foundation model-based agents.
Intervention	Development, optimization, and evaluation of multilayer runtime guardrails in foundation model-based agents, focusing on key quality attributes and design strategies similar to the Swiss Cheese Model structure.
Comparison	Comparative analysis of approaches to design multi-layered runtime guardrails in FM-based agents.
Outcomes	Taxonomy of multi-layered runtime guardrails for foundation model-based agents.
Context	**Include:** Empirical and theoretical studies on the components, design and evaluation of guardrails in foundation model-based agents. **Exclude:** Studies beyond the scope of foundation model based agents, non-English literature, and those not considering guardrails.

reviewer. The aggregated data were ultimately used to develop a taxonomy and design guide for multi-layered runtime guardrails in FM-based agents.

2.1. Research Workflow

The high-level research approach for this study is illustrated in Figure 1. The process begins with determining the research scope of this review. Subsequently, we identify relevant scientific databases and resources for the search. We then formulate keywords and search strings, outline both qualitative and quantitative checklists, and specify the criteria for the inclusion and exclusion of studies. Additionally, a reference management process was established to systematically organize and track all references using EndNote.

2.2. Research Questions

This study focuses on two primary concepts: (i) foundation model-based agents and (ii) multi-layered runtime guardrails. We adopted the Petticrew and Roberts approach [21] to define the **P**opulation, **I**nterventions, **C**omparison, **O**utcomes, and **C**ontext (PICOC), within which the intervention in this study is delivered. The PICOC for this study is shown in Table 1. Using these PICOC components and following Kitchenham's guidelines [19], we formulated three key research questions to guide our investigation.

RQ1: What are the essential qualities for designing runtime guardrails in FM-based agents?
Our first research question investigates the key qualities for designing multi-layered runtime guardrails in FM-based agents. It explores the responsible AI and AI safety concerns that require the adoption of guardrails to prevent harmful content and unintended behaviors.

Table 2: Consolidated Concepts and Search Terms

Main Terms	**Supportive Search Terms**
Concept 1 (Co1): Foundation Model based agents	Foundation Models, Foundation Model based agents, Large Language Model, Generative AI, Artificial General Intelligence, Transformer Models, Self-supervised Learning, Pretrained Models, Language Models, Conversational AI.
Concept 2 (Co2): Multi-layered Runtime Guardrails	Guardrails, guardian, responsible AI, safe, risk, trustworthy, protect, detect, monitor, verify, validate, evaluate, benchmark, design.

RQ2: What are the design options for runtime guardrails in FM-based agents?
This question explores the design options for runtime guardrails in FM-based agents from different perspectives, such as action and scope. It further investigates guardrail integration from the model's training phase to real-time decision-making processes, identifying optimal points to maximize their effectiveness. The question also explores various strategies for customizing guardrails to specific domains and user needs to ensure their generalizability. Additionally, it examines the performance implications of different configurations and approaches, such as static versus dynamic guardrails. Furthermore, it explores best practices for dynamically monitoring and updating guardrails to adapt to evolving risks and requirements. Finally, it examines methods and metrics for evaluating the effectiveness of guardrails..

RQ3: How can we design runtime guardrails to address the unique challenges of FM-based agents?
Our third research question explores how to address the challenges arising from the autonomous and deterministic nature of FM-based agents. Specially, we examine how to adapt the Swiss Cheese Model to safeguard the behaviors of FM-based agents by implementing multi-layered guardrails across various agent artifacts.

2.3. Search string formulation

Relevant primary studies for this SLR were identified based on the RQs defined in Section 2.2. With the assistance of the PICOC approach (shown in Table 1), our search terms were divided into two primary concepts, as shown in Table 2. These concepts helped us to set a well-formulated search string. We also used synonyms, abbreviations, and alternative spellings of search terms to increase the number of relevant research papers. We used truncation and wildcard operators to save time and effort in finding these alternative keywords. Moreover, different supplementary key terms or phrases discovered during search iterations were added to our search string to enhance our search strategy. Our supposition is that they will collect all relevant articles that contains guardrails for FM-based agents. When constructing the final search query, the identified keywords, their alternatives and related terms were linked with Boolean AND (&&), OR (‖) and NOT (¬) operators as follows:

$$[\{(C_{11}\|C_{12}\|...\|C_{1n})\mathbf{AND}(C_{21}\|C_{22}\|...\|C_{2n})\mathbf{NOT}(UC_1\|UC_2\|...\|UC_n)] \quad (1)$$

where $C_{11\ldots1n}$, and $C_{21\ldots2n}$ ε Co1 and Co2 of Table 2, respectively; and UC_1 . . . UC_n refers the **Exclude Context** defined earlier in Table 1. A representative version of the search string used in the electronic databases is as follows:
("Foundation Model based agents" OR "Foundation Models" OR "Foundation Model based agents" OR "Large Language Model" OR "Generative AI" OR "Artificial General Intelligence" OR "Transformer Models" OR "Self-supervised Learning" OR "Pretrained Models" OR "Language Models" OR "Conversational AI.) AND (Multi-layered Runtime Guardrails" OR "Guardrails" OR "guardian" OR "responsible AI" OR "safe" OR "risk" OR "trustworthy" OR "protect" OR "detect" OR "monitor" OR "verify" OR "validate" OR "evaluate" OR "benchmark" OR "design)

2.4. Automatic search in electronic databases for scientific literature

In March 2024, we conducted searches across five electronic databases without any time restrictions: ACM Digital Library, IEEE Xplore, Science Direct, Google Scholar, and Springer Link. We chose these databases because they contain most high quality, peer-reviewed papers in AI, Computer Science and Software Engineering domain. We chose to ignore some of the secondary indexing search engines like SCOPUS and INSPEC because they contain a large number of duplicate studies. Instead, we applied snowballing from located study references to find additional studies and to enhance our review's comprehensiveness.

Additionally, we chose not to explicitly search for gray literature because it often lacks a formal peer-review process, making it difficult to confirm reliability and consistency. Instead, our approach relied on snowballing to identify grey literature cited within peer-reviewed sources, ensuring that important materials were included while maintaining strict rigor. For instance, through snowballing, we identified gray literature and included only those that satisfied the concepts defined in Table 2, met inclusion criteria shown in Table 3, and complied with all the exclusion criteria outlined in Table 4.

2.5. Snowballing using Google scholar

Our database searches yielded a large set of primary papers. We also manually searched paper using the primary and supportive terms defined Table 2 as we did not want to miss any relevant existing study and wanted to make sure that the final set of papers is complete. We analysed the references from the final selected studies to also check for any potentially missed primary studies.

2.6. Selection of papers: Inclusion and exclusion criterion

Table 3 and Table 4 present the Inclusion Criteria (IC) and Exclusion Criteria (EC) that have been used to identify the studies for this SLR, respectively. We found that a considerable amount of work on guardrails exists in gray literature; however, we excluded them as they often lack peer review and a rigorous validation process. While some sources [22] argue that gray literature is an important resource for systematic literature reviews (SLRs), such literature can be misleading and introduce biases and inconsistencies in the review process [23]. We prioritized peer-reviewed sources in this study to ensure scientific reliability and credibility, as per Kitchenham et al. guidelines [19, 20].

Table 3: Inclusion Criteria

ID	Detail Criterion
IC_1	Full text of conference papers, journal articles, industry reports, and book chapters that are relevant to the defined main concepts: Foundation model based agents and guardrails.
IC_2	Papers written in English that include references.
IC_3	Studies that specifically address the design and development of guardrails in foundation model-based agents. This includes theoretical frameworks, empirical research and case studies.
IC_4	Papers available in an electronic format, such as PDF, DOC, DOCX, HTML, and PS etc.

Table 4: Exclusion Criteria

ID	Detail Criterion
EC_1	Work-in-progress proposals, keynote addresses, secondary studies, and vision papers without concrete relation to guardrails.
EC_2	Discussion papers and opinion pieces that do not provide empirical evidence or concrete solutions related to guardrails in foundation model-based agents.
EC_3	Short communications less than two pages, and studies that do not offer substantial information for analysis.
EC_4	Studies focusing solely on AI or similar technologies without direct relevance to guardrails.
EC_5	Research lacking a clear connection to the design and development of guardrails in the context of foundation models.
EC_6	Duplicate publications or earlier versions of studies that have been superseded by extended journal versions.
EC_7	Non-original research, commentary, editorial pieces, and non-empirical discussions papers.
EC_8	Studies inaccessible due to copyright or database restrictions.

2.7. Collection and filtering of the studies

Our filtration process is shown in Figure 2. Initially we ran the formatted query on four major databases that returned 1,733 research papers. We then applied filtering and classified the studies found according to the guidelines presented in [19, 20]. In our initial filtration process, we removed 108 papers due to being duplicated articles, editorial or key notes. After reading the title, abstract, conclusion and skimming through the introduction, methodology and results, we applied our exclusion criterion defined in Table 4, and 1524 further papers were removed. During the third step of filtration, we applied inclusion criteria and removed 80 papers as these studies did not meet ICs shown in Table 3. In parallel, we did a manual search and found 189 papers that meet our key concepts defined in Table 2 but not contain any unwanted content (UC). After applying ICs and ECs, 15 out of 189 papers were selected. Finally, we did a cross-check and ended up with 32 papers as our primary set of studies for analysis after completing the filtration process. Our study filtration file ('*Study*

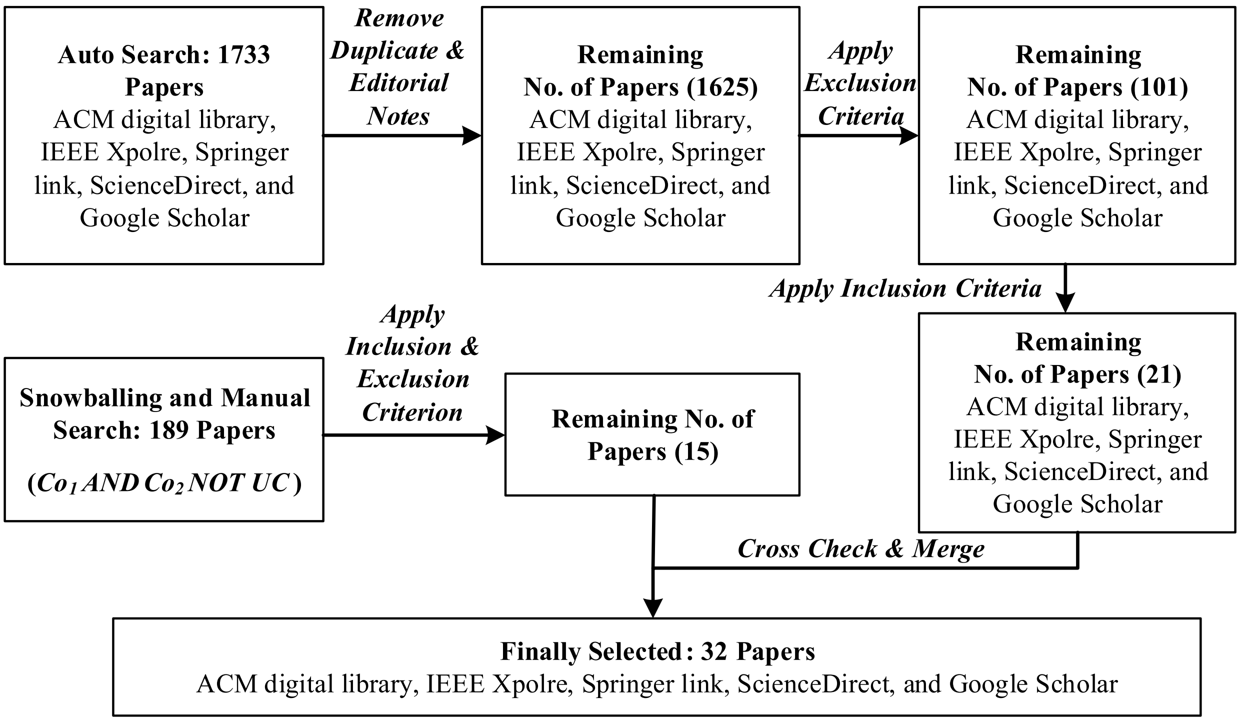
Auto Search: 1733 Papers
ACM digital library, IEEE Xpolre, Springer link, ScienceDirect, and Google Scholar
Remove Duplicate & Editorial Notes
Remaining No. of Papers (1625)
ACM digital library, IEEE Xpolre, Springer link, ScienceDirect, and Google Scholar
Apply Exclusion Criteria
Remaining No. of Papers (101)
ACM digital library, IEEE Xpolre, Springer link, ScienceDirect, and Google Scholar
Apply Inclusion Criteria
Remaining No. of Papers (21)
ACM digital library, IEEE Xpolre, Springer link, ScienceDirect, and Google Scholar
Snowballing and Manual Search: 189 Papers
(Co_1 AND Co_2 NOT UC)
Apply Inclusion & Exclusion Criterion
Remaining No. of Papers (15)
Cross Check & Merge
Finally Selected: 32 Papers
ACM digital library, IEEE Xpolre, Springer link, ScienceDirect, and Google Scholar

Figure 2: Study Selection Process for this SLR

Filtration.xlsx') is available in the supplementary materials to facilitate reproducibility. It provides a detailed and step-by-step account of the exclusion and inclusion criteria applied during the filtration process for each of the selected studies.

2.8. Qualitative information extracted from each paper

We used semi-automated process [24] for data extraction from each study. We extracted the following key **Q**ualitative **I**nformation (QI) from each primary selected paper:

- Publication details - authors, title, date, venue, publisher.
- Guardrails definitions, external qualities, and reported key internal quality attributes.
- Reported primary challenges and features addressing guardrails issues.
- Reported guardrails design options, including methods and models for evaluation.
- Types of tools, techniques, and frameworks used.
- Design and implementation details.
- Evaluation metrics for guardrails performance analysis.
- Guardrails subjects and target beneficiaries.
- Insights on results, strengths, gaps, challenges, limitations, and future research directions.

2.9. Quality assessment

We evaluated each study based on the following five **Q**uality **A**ssessment **C**riteria (QAC) on a scale from 1 (Very Poor) to 5 (Excellent). If a study's average score was less than 2, we label it as a paper as a poor quality paper excluded from further analysis, otherwise we use the qualitative information to decide this. The QAC for this study are shown below and score are presented in Appendix B:

QAC 1: The study's relevance to the defined concepts in Table 2 – foundation model based agents and guardrails.

QAC 2: The clarity and comprehensiveness of the study's methodology for guardrail design.

QAC 3: The adequacy of the study's data collection, analysis, and evaluation of guardrail effectiveness across different layers of the agent architecture.

QAC 4: Discussion of challenges in designing guardrails for autonomous and non-deterministic behaviors in agents.

QAC 5: Practical applicability of findings for guardrails in FM-based agents.

2.10. Reference management and screening tool

We used EndNote X9 tool for reference management and screening the studies because it facilitates easy removal of double entries and keeps track of papers by summarizing essential facts, e.g., title, authors, abstract, keywords, venue, date, and page numbers.

2.11. Data extraction and synthesis

During data extraction, we downloaded all primary studies and grouped the papers by theme, contribution, authors, and database name, in this order. An identity code (SS) was formulated and assigned to every individual study. The list of papers with their identity code is available in Appendix A. We followed the following steps to counter the biases during data extraction:

❖ Initially, the first anonymous author of this paper extracted data for two papers from each selected database and stored the results in an excel sheet. The remaining authors of this report cross-checked these data, and the necessary correction was applied.

❖ Then the first anonymous author extracted data for another ten selected studies, and similar cross-checking was performed until all of the authors reached agreement and the outcome did not vary more than 5% for anyone. At the end of this step, the review protocol was finalized to incorporate the changes.

❖ In the third step, the first anonymous author re-extracted the data from previously examined studies as well as the remaining studies as per the revised protocol. The extracted data were sequentially cross-checked by the remaining authors (once each) to minimize extraction bias and omissions.

- Finally, all data was stored in an Excel sheet for analysis and synthesis.

3. Conclusion

This protocol outlines the methodology for conducting a systematic literature review on multi-layered runtime guardrails in foundation model-based agents. The study aims to identify key external and internal qualities for multi-layered runtime guardrails in FM-based agents, and the available design options. The goal is to develop a comprehensive taxonomy and propose a novel architecture for multi-layered runtime guardrails, following the Swiss Cheese Model structure for AI Safety. This contribute to the safe and responsible deployment of the FM-based agents in practice.

Appendix A: List of Selected Studies

[1] M. Pawagi, V. Kumar, Guardrails: Automated suggestions for clarifying ambiguous purpose statements, in: Proceedings of the 16th Annual ACM India Compute Conference, ACM, 2023, p. 55–60. doi:10.1145/3627217.3627234.

[2] A. Khorramrouz, S. Dutta, A. Dutta, A. R. KhudaBukhsh, Down the toxicity rabbit hole: Investigating PaLM 2 guardrails, arXiv preprint arXiv:2309.06415 (2023). doi:https://doi.org/10.48550/arXiv.2309.06415.

[3] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, X. Huang, Building guardrails for large language models, arXiv preprint arXiv:2402.01822 (2024). doi:https://doi.org/10.48550/arXiv.2402.01822.

[4] N. Mangaokar, A. Hooda, J. Choi, S. Chandrashekaran, K. Fawaz, S. Jha, A. Prakash, PRP: Propagating universal perturbations to attack large language model guard-rails, arXiv preprint arXiv:2402.15911 (2024). doi:https://doi.org/10.48550/arXiv.2402.15911.

[5] M. Anderljung, J. Barnhart, J. Leung, A. Korinek, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, et al., Frontier AI regulation: Managing emerging risks to public safety, arXiv preprint arXiv:2307.03718 (2023). doi:https://doi.org/10.48550/arXiv.2307.03718.

[6] M. Liffiton, B. E. Sheese, J. Savelka, P. Denny, Codehelp: Using large language models with guardrails for scalable support in programming classes, in: Proceedings of the 23rd Koli Calling International Conference on Computing Education Research, 2024, pp. 1–11. doi:10.1145/3631802.3631830.

[7] Z. Zhang, Y. Lu, J. Ma, D. Zhang, R. Li, P. Ke, H. Sun, L. Sha, Z. Sui, H. Wang, et al., Shieldlm: Empowering LLMs as aligned, customizable and explainable safety detectors, arXiv preprint arXiv:2402.16444 (2024). doi:https://doi.org/10.48550/arXiv.2402.16444.

[8] T. Rebedea, R. Dinu, M. N. Sreedhar, C. Parisien, J. Cohen, NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails, in: Y. Feng, E. Lefever (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2023, pp. 431–445. doi:10.18653/v1/2023.emnlp-demo.40.

[9] Y. Wang, L. Singh, Adding guardrails to advanced chatbots, arXiv preprint arXiv:2306.07500 (2023). doi:https://doi.org/10.48550/arXiv.2306.07500.

[10] M. Shanahan, Talking about large language models, Commun. ACM 67 (2) (2024) 68–79. doi:10.1145/3624724.

[11] W. Du, Q. Li, J. Zhou, X. Ding, X. Wang, Z. Zhou, J. Liu, Finguard: A multimodal aigc guardrail in financial scenarios, in: Proceedings of the 5th ACM International Conference on Multimedia in Asia, 2024, pp. 1–3. doi:10.1145/3595916.3626351.

[12] A. Wei, N. Haghtalab, J. Steinhardt, Jailbroken: How does LLM safety training fail?, Advances in Neural Information Processing Systems 36 (2024). doi:10.48550/arXiv.2307.02483.

[13] J. Zhao, K. Chen, X. Yuan, Y. Qi, W. Zhang, N. Yu, Silent guardian: Protecting text from malicious exploitation by large language models, arXiv preprint arXiv:2312.09669 (2023). doi:https://doi.org/10.48550/arXiv.2312.09669.

[14] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, P. Henderson, Fine-tuning aligned language models compromises safety, even when users do not intend to!, arXiv preprint arXiv:2310.03693 (2023). doi:https://doi.org/10.48550/arXiv.2310.03693.

[15] J. Mökander, J. Schuett, H. R. Kirk, L. Floridi, Auditing large language models: a three-layered approach, AI and Ethics (2023) 1–31doi:https://doi.org/10.1007/s43681-023-00289-2.

[16] S. Banerjee, S. Layek, R. Hazra, A. Mukherjee, How (un) ethical are instruction-centric responses of LLMs? unveiling the vulnerabilities of safety guardrails to harmful queries, arXiv preprint arXiv:2402.15302 (2024). doi:https://doi.org/10.48550/arXiv.2402.15302.

[17] S. Ee, J. O'Brien, Z. Williams, A. El-Dakhakhni, M. Aird, A. Lintz, Adapting cybersecurity frameworks to manage frontier AI risks, Institute for AI Policy and Strategy (IAPS) (2023). URL https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/6528c5c7f912f74fbd03fc34/1697170896984/Adapting+cybersecurity+frameworks+to+manage+frontier+AI+risks.pdf

[18] P. Rai, S. Sood, V. K. Madisetti, A. Bahga, Guardian: A multi-tiered defense architecture for thwarting prompt injection attacks on LLMs, Journal of Software Engineering and Applications 17 (1) (2024) 43–68. doi:https://doi.org/10.4236/jsea.2024.171003.

[19] A. Kumar, S. Singh, S. V. Murty, S. Ragupathy, The ethics of interaction: Mitigating security threats in LLMs, arXiv preprint arXiv:2401.12273 (2024). doi:https://doi.org/10.48550/arXiv.2401.12273.

[20] X. Shen, Z. Chen, M. Backes, Y. Shen, Y. Zhang, "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models, arXiv preprint arXiv:2308.03825 (2023). doi:https://doi.org/10.48550/arXiv.2308.03825.

[21] A. Kumar, C. Agarwal, S. Srinivas, S. Feizi, H. Lakkaraju, Certifying LLM safety against adversarial prompting, arXiv preprint arXiv:2309.02705 (2023). doi:https://doi.org/10.48550/arXiv.2309.02705.

[22] Y. Zeng, H. Lin, J. Zhang, D. Yang, R. Jia, W. Shi, How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing LLMs, arXiv preprint arXiv:2401.06373 (2024). doi:https://doi.org/10.48550/arXiv.2401.06373.

[23] B. Wei, K. Huang, Y. Huang, T. Xie, X. Qi, M. Xia, P. Mittal, M. Wang, P. Henderson, Assessing the brittleness of safety alignment via pruning and low-rank modifications, arXiv preprint arXiv:2402.05162 (2024). doi:https://doi.org/10.48550/arXiv.2402.05162.

[24] S. Goyal, M. Hira, S. Mishra, S. Goyal, A. Goel, N. Dadu, D. Kirushikesh, S. Mehta, N. Madaan, LLMguard: Guarding against unsafe LLM behavior, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38(21), 2024, pp. 23790–23792. doi:https://doi.org/10.1609/aaai.v38i21.30566.

[25] R. R. Llaca, V. Leskoschek, V. C. Paiva, C. Lupău, P. Lippmann, J. Yang, Student-teacher prompting for red teaming to improve guardrails, in: Proceedings of the ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI, 2023, pp. 11–23. doi:http://dx.doi.org/10.18653/v1/2023.artofsafety-1.2.

[26] Z. Wang, F. Yang, L. Wang, P. Zhao, H. Wang, L. Chen, Q. Lin, K.-F. Wong, SELF-GUARD: Empower the LLM to safeguard itself, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 1648–1668. URL https://aclanthology.org/2024.naacl-long.92

[27] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, et al., Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, in: Advances in Neural Information Processing Systems, 2023, pp. 1–110. doi:https://doi.org/10.48550/arXiv.2306.11698.

[28] R. Bommasani, D. A. Hudson, E. Adeli, et al., On the opportunities and risks of foundation models, CoRR abs/2108.07258 (2021). arXiv:2108.07258, doi:https://doi.org/10.48550/arXiv.2108.07258.

[29] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, et al., Sociotechnical safety evaluation of generative ai systems, arXiv preprint arXiv:2310.11986 (2023). doi:https://doi.org/10.48550/arXiv.2310.11986.

[30] D. Kang, D. Raghavan, P. Bailis, M. Zaharia, Model assertions for monitoring and improving ml models, Proceedings of Machine Learning and Systems 2 (2020) 481–496. doi:https://doi.org/10.48550/arXiv.2003.01668.

[31] Z. Yuan, Z. Xiong, Y. Zeng, N. Yu, R. Jia, D. Song, B. Li, RigorLLM: Resilient guardrails for large language models against undesired content, arXiv preprint arXiv:2403.13031 (2024). doi:https://doi.org/10.48550/arXiv.2403.13031.

[32] Z. Chu, Y. Wang, L. Li, Z. Wang, Z. Qin, K. Ren, A causal explainable guardrails for large language models, arXiv preprint arXiv:2405.04160 (2024).

Appendix B: QAC Scores for the Selected Studies

Selected Study (SS) No.	**QAC1**	**QAC2**	**QAC3**	**QAC4**	**QAC5**
SS-1	2	2	2	2	2
SS-2	4	5	5	5	4
SS-3	4	3	2	3	2
SS-4	3	3	3	3	4
SS-5	2	4	3	3	5
SS-6	1	5	4	3	5
SS-7	5	4	4	4	3
SS-8	5	4	4	5	3
SS-9	0	1	2	4	3
SS-10	0	2	2	3	4
SS-11	3	3	1	3	2
SS-12	3	2	3	3	2
SS-13	4	3	3	4	3
SS-14	5	4	4	5	4
SS-15	5	4	3	4	5
SS-16	5	4	4	5	4
SS-17	5	4	3	5	4
SS-18	5	4	4	5	5
SS-19	4	3	2	4	3
SS-20	5	4	4	5	4
SS-21	5	4	4	4	5
SS-22	5	5	4	5	4
SS-23	5	4	4	5	4
SS-24	5	4	2	2	3
SS-25	5	4	4	4	5
SS-26	5	4	4	5	4
SS-27	5	4	5	5	4
SS-28	5	4	4	3	4
SS-29	4	4	4	3	4
SS-30	5	4	4	4	5
SS-31	5	5	4	4	5
SS-32	5	4	4	4	5

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, et al., On the opportunities and risks of foundation models, CoRR abs/2108.07258 (2021). arXiv:2108.07258, doi:https://doi.org/10.48550/arXiv.2108.07258.
URL https://doi.org/10.48550/arXiv.2108.07258

[2] Q. Lu, L. Zhu, X. Xu, Z. Xing, J. Whittle, Towards responsible and safe AI in the era of foudnation models: A reference architecture for designing foundation model based systems, to appear in IEEE Software (2024).
URL https://arxiv.org/html/2304.11090v4

[3] N. Maslej, et al., AI index report 2024, Tech. rep., Stanford Institute for Human-Centered Artificial Intelligence (HAI) (2024).
URL https://aiindex.stanford.edu/report/2024

[4] R. Bommasani, P. Liang, Reflections on foundation models, Last accessed on Jun.-2024 (2021).
URL https://hai.stanford.edu/news/reflections-foundation-models

[5] L. Wang, et al., A survey on large language model based autonomous agents, Frontiers of Computer Science 18 (6) (2024) 1–26. doi:https://doi.org/10.1007/s11704-024-40231-1.
URL https://doi.org/10.1007/s11704-024-40231-1

[6] T. Shabani, S. Jerie, T. Shabani, A comprehensive review of the swiss cheese model in risk management, Safety in Extreme Environments 6 (1) (2024) 43–57. doi:https://doi.org/10.1007/s42797-023-00091-7.

[7] Y. Wang, L. Singh, Adding guardrails to advanced chatbots, arXiv preprint arXiv:2306.07500 (2023). doi:https://doi.org/10.48550/arXiv.2306.07500.
URL https://doi.org/10.48550/arXiv.2306.07500

[8] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, et al., DecodingTrust: A comprehensive assessment of trustworthiness in GPT models, in: Advances in Neural Information Processing Systems, 2023, pp. 1–110. doi:https://doi.org/10.48550/arXiv.2306.11698.
URL https://doi.org/10.48550/arXiv.2306.11698

[9] B. Wei, K. Huang, Y. Huang, T. Xie, X. Qi, M. Xia, P. Mittal, M. Wang, P. Henderson, Assessing the brittleness of safety alignment via pruning and low-rank modifications, arXiv preprint arXiv:2402.05162 (2024). doi:https://doi.org/10.48550/arXiv.2402.05162.
URL https://doi.org/10.48550/arXiv.2402.05162

[10] M. Anderljung, et al., Frontier AI regulation: Managing emerging risks to public safety, arXiv preprint arXiv:2307.03718 (2023). doi:https://doi.org/10.48550/arXiv.2307.03718.
URL https://doi.org/10.48550/arXiv.2307.03718

[11] A. Wei, N. Haghtalab, J. Steinhardt, Jailbroken: How does LLM safety training fail?, Advances in Neural Information Processing Systems 36 (2024). doi:https://doi.org/10.48550/arXiv.2307.02483.
URL https://doi.org/10.48550/arXiv.2307.02483

[12] A. Gubkin, Understanding why ai guardrails are necessary: Ensuring ethical and responsible ai use, Last accessed on Jul.-2024 (2024).
URL https://www.aporia.com/learn/ai-guardrails/

[13] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, X. Huang, Building guardrails for large language models, arXiv preprint arXiv:2402.01822 (2024). doi:https://doi.org/10.48550/arXiv.2402.01822.
URL https://doi.org/10.48550/arXiv.2402.01822

[14] T. Rebedea, R. Dinu, M. N. Sreedhar, C. Parisien, J. Cohen, NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Singapore, 2023, pp. 431–445. doi:10.18653/v1/2023.emnlp-demo.40.
URL https://aclanthology.org/2023.emnlp-demo.40

[15] D. Kang, D. Raghavan, P. Bailis, M. Zaharia, Model assertions for monitoring and improving ml models, Proceedings of Machine Learning and Systems 2 (2020) 481–496. doi:https://doi.org/10.48550/arXiv.2003.01668.
URL https://doi.org/10.48550/arXiv.2003.01668

[16] J. Mökander, J. Schuett, H. R. Kirk, L. Floridi, Auditing large language models: A three-layered approach, AI and Ethics (2023) 1–31doi:https://doi.org/10.1007/s43681-023-00289-2.
URL https://doi.org/10.1007/s43681-023-00289-2

[17] A. Kumar, S. Singh, S. V. Murty, S. Ragupathy, The ethics of interaction: Mitigating security threats in LLMs, arXiv preprint arXiv:2401.12273 (2024). doi:https://doi.org/10.48550/arXiv.2401.12273.
URL https://doi.org/10.48550/arXiv.2401.12273

[18] C. Zhou, et al., A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT, arXiv preprint arXiv:2302.09419 (2023). doi:https://doi.org/10.48550/arXiv.2302.09419.
URL https://doi.org/10.48550/arXiv.2302.09419

[19] B. A. Kitchenham, S. Charters, Other Keele Staffs, Guidelines for performing systematic literature reviews in software engineering (version 2.3), Tech. rep., Keele University and Durham University Joint Report (2007).
URL https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

[20] B. Kitchenham, Procedures for performing systematic reviews, Tech. Rep. 2004, Keele University, UK (2004).
URL http://artemisa.unicauca.edu.co/~ecaldon/docs/spi/kitchenham_2004.pdf

[21] M. Petticrew, H. Roberts, Systematic reviews in the social sciences: A practical guide, John Wiley & Sons, 2008.
URL https://doi.org/10.1002/9780470754887

[22] A. Paez, Gray literature: An important resource in systematic reviews, Journal of Evidence-Based Medicine 10 (3) (2017) 233–240. doi:https://doi.org/10.1111/jebm.12266.

[23] K. Godin, J. Stapleton, S. I. Kirkpatrick, R. M. Hanning, S. T. Leatherdale, Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in canada, Systematic reviews 4 (2015) 1–10. doi:https://doi.org/10.1186/s13643-015-0125-0.

[24] L. Schmidt, A. Finnerty Mutlu, R. Elmore, B. Olorisade, J. Thomas, J. Higgins, Data extraction methods for systematic review (semi)automation: Update of a living systematic review, F1000Research 10 (401) (2023). doi:https://doi.org/10.12688/f1000research.51117.2. URL https://doi.org/10.12688/f1000research.51117.2