Open-Weight Guardians and Open-Source Frameworks for Governing LLMs

Kush R. Varshney
IBM Research – Thomas J. Watson Research Center
Yorktown Heights, New York, USA
krvarshn@us.ibm.com

Abstract—**Governing the safety of large language models includes essential tools for curating and filtering datasets, evaluating and red-teaming models, and guardrailing model inputs and outputs. In this paper, we describe several frameworks IBM has open-sourced for these tasks—supported by open-weight guardian LLMs and an open-access risk atlas—as exemplars for enabling well-governed AI lifecycles. The three specific frameworks are Data Prep Kit, Unitxt, and Foundation Model Stack Guardrails Orchestrator. The specific models are Granite Guardian 3.0 2B, Granite Guardian 3.0 8B, Granite Guardian HAP 38M, and Granite Guardian HAP 125M. It is uncommon for such tooling used in-house by a model provider to be open-sourced. The availability of these assets enables the community to mitigate harms and hazards in a systematic and principled way. The paper concludes with a brief discussion on how the basic risk atlas and guardians need to be customized for different contexts, and how the overall safety framework applies to agentic AI systems.**

I. Introduction

Artificial intelligence (AI) safety is concerned with minimizing the probability of expected harms and the possibility of unexpected harms to people and society from AI systems; risk is the probability of an expected harm [1]. Several interventions, ranging from inherently safe design to procedural safeguards, may be used to improve AI safety throughout the lifecycle of AI development, deployment, and decommissioning. The possible harms of AI include ones present in traditional machine learning systems, as well as ones that are new or amplified with large language models (LLMs) and other foundation models [2]. Importantly, the harms and risks of AI are contextual, and only precise-enough to be governed in the context of the AI system's use, not independent of it [3], [4]. Governance equals control: damping the harms as much as possible.

LLMs, unlike traditional machine learning, are often used in generative tasks to create new artifacts. Because of this generative use, many new harms have emerged such as the generation of hateful and abusive content and the generation of hallucinations that are not faithful to source content. Several taxonomies of these harms have been created recently, both from the perspectives of scholarship and standard-setting [5]–[7]. Moreover, crosswalks between different taxonomies have been produced.[1] However, operationalizing such taxonomies in practical LLM governance is non-trivial [4], [8].

At IBM, building upon our earlier AI risk taxonomy [2], we created an open-access AI Risk Atlas[2] that threads the needle of practicality. It is of the appropriate granularity and has the appropriate supporting materials so that it not only serves policymakers and other users seeking to understand AI risks, but is also used practically throughout the deployment lifecycle. It is the basis for the watsonx.governance software platform to elicit key risks from users and set policies for a given use. It is carried all the way through to observability and monitoring of deployed LLMs. A basic operation in AI safety is detecting harms; we have developed and open-sourced several detectors and guardian models that are founded in the practicality of the risk atlas [9], [10]. Our companion attack atlas for AI safety testing strategies is similarly oriented toward practitioners and practicality [11]. The risk atlas catalogs the *what* and the attack atlas the *how*.

Among the four general intervention categories for safety (inherently safe design, safety reserves, safe fail, and procedural safeguards) [12], openness is a key procedural safeguard because it enables public audit for hazards [1]. However, with LLMs, openness entails several possibilities ranging from open-weight models, to some amount of open-source code, to open data [13]. Commentators have both argued for and against the safety of so-called open-source AI [14], [15]. We fall on the side of openness leading to more safety; in the remainder of the paper, we describe several of the tools and technologies we have opened with commercially-friendly licenses.

The remainder of the paper is organized as follows. In Section II, we describe a typical LLM lifecycle and point out how it leads to different scenarios and requirements for using harm detectors. In Section III, we overview several governance tools and technologies IBM has open-sourced that the AI safety community may benefit from. In Section IV, we conclude by summarizing the essentials, emphasizing that harms are contextual and must be made precise for a usage-centric (non-model-centric) worldview, and providing a future outlook for safety in agentic AI systems.

II. AI Lifecycle

A simplified view of the development and deployment of LLMs is presented in Figure 1. The decommissioning lifecycle

[1] https://airc.nist.gov/AI_RMF_Knowledge_Base/Crosswalks

[2] https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas

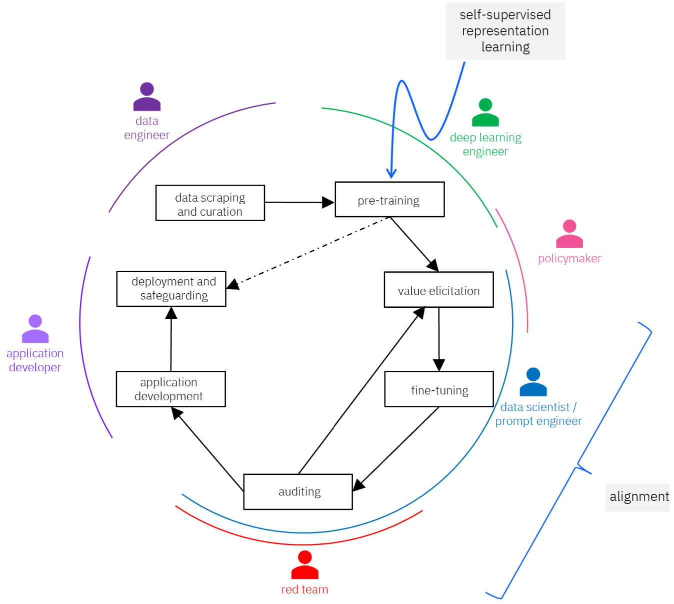

Fig. 1. A simplified illustration of the LLM development lifecycle. The boxes represent different phases of the lifecycle. The arcs associated with personas indicate which phases they are involved in.

is omitted; it is an emerging topic that is out of scope for this paper [16], [17]. The different phases have varying latency, throughput, cost requirements. They may be run on different hardware infrastructure accordingly, including those with CPUs, GPUs, and specialized AI accelerating processors [18].

The initial phase, data scraping and curation may include various steps that work toward safety, such as filtering out hateful and biased documents (or not gathering them at all if using a *focused* crawler). Such filtering requires low-cost, low-latency, high-throughput harm detectors. As detailed in the technical report for IBM's latest LLM family, Granite 3.0, the IBM data and model factory uses small detectors for filtering hateful/abusive/profane language as well as malware [19]. The smallest detectors are very efficient at 38 million parameters [20]. The data sets and data mixtures used in the next phase, pre-training, are detailed in the same report [19]. Such data transparency is not usually provided.

Once a model has been pre-trained, it is evaluated across quality and safety dimensions before the the next several phases of the lifecycle, collectively known as alignment. Such evaluation also coincides with model risk assessment performed by companies that are procuring a pre-trained model from a vendor. The requirements of harm detectors for safety evaluation in model risk assessment, which is run offline, are low in latency and medium in throughput. During alignment, harm detectors may be used in reward functions or as part of automated red-teaming methods. In this phase too, the requirements on latency are not severe and a medium throughput is required.

Another use of harm detectors is in online safeguarding settings in which individual LLM inferences are checked. Here the latency cannot be too high, and an overall solution needs to be fairly simple for application developers to use. A class of guardian models in the 1 billion to 8 billion parameter range has developed to serve this use. Llama Guard, ShieldGemma and Granite Guardian are all open-weight examples [10]. All of these are LLMs in their own right, but fine-tuned to detect a wide range of harmful behaviors. Unlike smaller detectors for single harms dimensions, these guardians are easier to use because they encapsulate all harms in a single model and they also perform better than general purpose LLMs at the detection task [21].

A last use of harm detectors is in observability and monitoring of deployed LLMs. Here, the task is online batch risk evaluation. A logged payload of input/output pairs from an LLM is periodically evaluated across harms. The latency requirements are medium and the throughput requirements are high.

III. IBM Open-Source Assets

In this section, we will recount the various assets IBM has open-sourced that allow AI practitioners unfettered ability to recreate safety protocols, procedures, and practices undertaken by a large technology company throughout the lifecycle—and thereby uncover safety hazards. As mentioned previously, IBM's AI risk atlas is open-access and can be used practically and operationally across phases of the lifecycle. It is also available in a machine-consumable format at https://github.com/ibm-granite-community/granite-snack-cookbook/blob/main/recipes/Granite_Guardian/ibm_ai_risk_atlas.yml. Stemming from the risk atlas are detectors for many of the harms cataloged there.

IBM has open-sourced two small hate/abuse/profanity detectors: Granite Guardian HAP 38M and Granite Guardian HAP 125M, that perform with high accuracy and low cost and latency. IBM has also open-sourced two guardian-class models: Granite Guardian 3.0 2B and Granite Guardian 3.0 8B, that are more comprehensive in detecting social bias and implicit hate, toxicity and explicity hate, profanity, violence, sexual content, unethical behavior, jailbreaking attempts, hallucination/lack of groundedness, and lack of context or answer relevance in retrieval-augmented generation (RAG) use cases. Importantly, no other guardian-class model currently checks for jailbreaking, hallucination, or RAG metrics. All of these Granite Guardian harm detector models are available with commercially-friendly Apache 2.0 Licenses at Hugging Face: https://huggingface.co/collections/ibm-granite/granite-guardian-models-66db06b1202a56cf7b079562.

Building upon the open-sourced detectors as a foundation, IBM has further open-sourced governance frameworks for different parts of the AI lifecycle (detailed in Section II) that consume the detectors as underlying assets. All of these governance frameworks have standard permissive licenses. First, for the data scraping and curation phase of the lifecycle, IBM has open-sourced the Data Prep Kit, which

IBM uses itself in its data and model factory to coordinate and run myriad processing steps including filtering for hate/abuse/profanity [22]. The Data Prep Kit is available at `https://github.com/IBM/data-prep-kit` and has an Apache 2.0 License.

Unitxt is another governance framework IBM uses itself and has open-sourced. It is a framework for preparing data and models, and then running LLM evaluations [23]. Available from `https://github.com/IBM/unitxt`, also with an Apache 2.0 License, Unitxt calls various metrics and detectors to run the evaluations. Also, note that metrics and detectors are not enough to conduct a safety evaluation. Specially-designed red-teaming prompts that induce an LLM to produce harmful behavior are also needed. IBM has open-sourced several red-teaming datasets as well:

- AttaQ (deception, discrimination, harmful information, substance abuse, sexual content, personally-identifiable information, and violence; `https://huggingface.co/datasets/ibm/AttaQ`; MIT License) [24]
- SocialStigmaQA (amplification of social bias via stigmas; `https://huggingface.co/datasets/ibm/SocialStigmaQA`; Community Data License Agreement Permissive 2.0) [25]
- ProvoQ (minority-stigma pairs; `https://huggingface.co/datasets/ibm/ProvoQ`; Community Data License Agreement Permissive 2.0)
- WikiContradict (conflicting knowledge and hallucination; `https://huggingface.co/datasets/ibm/Wikipedia_contradict_benchmark`; MIT License) [26]

These benchmark datasets may benefit from standardized documentation such as BenchmarkCards [27].

Lastly, focused on the online safeguarding use of detectors, IBM has open-sourced the Foundation Model Stack Guardrails Orchestrator with an Apache 2.0 License. This governance framework puts harm detectors behind an application programming interface (API) and acts as an orchestration server for efficiently invoking them with appropriate routing of requests. This guardrails orchestrator is available at `https://github.com/foundation-model-stack/fms-guardrails-orchestrator`.

Given the array of requirements set by different uses of detectors throughout the AI lifecycle, it may be bewildering to choose among them. To help reduce this bewilderment, IBM has also created a systematic benchmarking framework for harm detectors named Adversarial Prompt Evaluation, available with an MIT License from `https://github.com/IBM/Adversarial-Prompt-Evaluation` [28].

IV. Conclusion

A. Summary

In this paper, we have laid out the essential tooling and process components for healthy and robust AI safety governance throughout the several steps involved in developing and deploying LLMs. These essentials, illustrated in Figure 2, are a harm taxonomy or risk atlas, detectors for those harms, and software layers for running those detectors in different scenarios: data preparation, evaluation, and online guardrailing. IBM has opened all of those components that are used in-house with the hope that they may be useful for the community to better understand and improve AI safety practices.

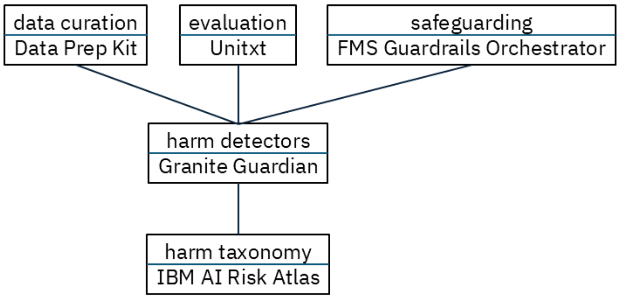

Fig. 2. The essential components for LLM safety governance (top half of each box), and the related assets that IBM has open-sourced (bottom half of each box).

B. Contextual Harms

Before concluding, let us note that a safety approach beginning with a fixed harm taxonomy misses something. No matter how comprehensive it may be, it cannot be a one-size-fits-all solution for all groups, communities, times, and places. Harms are necessarily contextual. First, tools and processes are needed that mediate a fixed taxonomy with the elicited understanding of the use and context to prioritize certain risks—a form of usage governance [4]. But beyond those, users and communities need ways to precisely personalize harms to their concerns, for example through tools that let them design their own detectors, also known as LLMs-as-judges [29]. Guardian models are promptable for new harm definitions to some extent, and the Adversarial Prompt Evaluation suite evaluates this ability. LLM safety must be context-centric, not model-centric [30].

C. Outlook on Agentic AI

The next wave of AI development that builds upon LLMs is autonomous agents that plan, reason, and use tools (via APIs). Safety in such agentic, cyber-physical AI systems needs the same essential components as LLMs, but based on a slightly different set of harms. For example, in contrast to natural language hallucination in LLMs, agentic AI systems may suffer from ‘function-calling hallucination’ that yields incorrect or unavailable functions, parameters, parameter types, or parameter values. There will also be newer risks related to the loss of human control.

Acknowledgment

Thanks to Ioana Baldini, Bishwaranjan Bhattacharjee, Roslyn Docktor, Ryan Hagemann, Michael Hind, Jacquelyn

Martino, Mason Molesky, John Richards, and Prasanna Sattigeri for their thoughts.

References

[1] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big Data*, vol. 5, no. 3, pp. 246–255, 2017.

[2] "Foundation models: Opportunities, risks and mitigations," IBM AI Ethics Board, Armonk, NY, USA, Tech. Rep., Jul. 2023.

[3] R. Hagemann and J.-M. Leclerc, "Precision regulation for artificial intelligence," IBM Policy Lab, Tech. Rep., 2020.

[4] E. M. Daly, S. Rooney, S. Tirupathi, L. Garces-Erice, I. Vejsbjerg, F. Bagehorn, D. Salwala, C. Giblin, M. L. Wolf-Bauwens, I. Giurgiu, M. Hind, and P. Urbanetz, "Usage governance advisor: From intent to AI governance," arXiv:2412.01957, 2024.

[5] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel, "Taxonomy of risks posed by language models," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2022, pp. 214–229.

[6] "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, Tech. Rep. NIST AI 100-1, Jan. 2023.

[7] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla, J. Gallegos, A. Smart, E. Garcia, and G. Virk, "Sociotechnical harms: Scoping a taxonomy for harm reduction," in *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, Aug. 2023, pp. 723–741.

[8] E. Bogucka, S. Šćepanović, and D. Quercia, "Atlas of AI risks: Enhancing public understanding of AI risks," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Oct. 2024, pp. 33–43.

[9] S. Achintalwar, A. Alvarado Garcia, A. Anaby-Tavor, I. Baldini, S. E. Berger, B. Bhattacharjee, D. Bouneffouf, S. Chaudhury, P.-Y. Chen, L. Chiazor, E. M. Daly, R. A. de Paula, P. Dognin, E. Farchi, S. Ghosh, M. Hind, R. Horesh, G. Kour, J. Y. Lee, E. Miehling, K. Murugesan, M. Nagireddy, I. Padhi, D. Piorkowski, A. Rawat, O. Raz, P. Sattigeri, H. Strobelt, S. Swaminathan, C. Tillmann, A. Trivedi, K. R. Varshney, D. Wei, S. Witherspooon, and M. Zalmanovici, "Detectors for safe and reliable LLMs: Implementations, uses, and limitations," arXiv:2403.06009, 2024.

[10] I. Padhi, M. Nagireddy, G. Cornacchia, S. Chaudhury, T. Pedapati, P. Dognin, K. Murugesan, E. Miehling, M. Santillan Cooper, K. Fraser, G. Zizzo, M. Z. Hameed, M. Purcell, M. Desmond, Q. Pan, I. Vejsbjerg, E. M. Daly, M. Hind, W. Geyer, A. Rawat, K. R. Varshney, and P. Sattigeri, "Granite guardian," https://github.com/ibm-granite/granite-guardian/blob/main/technical_report.pdf, IBM, Tech. Rep., 2024.

[11] A. Rawat, S. Schoepf, G. Zizzo, G. Cornacchia, M. Z. Hameed, K. Fraser, E. Miehling, B. Buesser, E. M. Daly, M. Purcell, P. Sattigeri, P.-Y. Chen, and K. R. Varshney, "Attack atlas: A practitioner's perspective on challenges and pitfalls in red teaming GenAI," in *Proceedings of the NeurIPS Workshop on Red Teaming GenAI*, Dec. 2024.

[12] N. Möller and S. O. Hansson, "Principles of engineering safety: Risk and uncertainty reduction," *Reliability Engineering & System Safety*, vol. 93, no. 6, pp. 798–805, 2008.

[13] D. G. Widder, S. West, and M. Whittaker, "Open (for business): Big tech, concentrated power, and the political economy of open AI," SSRN:4543807, 2023.

[14] D. E. Harris, "Open-source AI is uniquely dangerous," *IEEE Spectrum*, Jan. 2024.

[15] B. Brooks, "Open-source AI is good for us," *IEEE Spectrum*, Jan. 2024.

[16] U. Ehsan, R. Singh, J. Metcalf, and M. Riedl, "The algorithmic imprint," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2022, pp. 1305–1317.

[17] N. Johnson, S. Moharana, C. Harrington, N. Andalibi, H. Heidari, and M. Eslami, "The fall of an algorithm: Characterizing the dynamics toward abandonment," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2024, pp. 337–358.

[18] S. Sen, S. Jain, S. Krithivasan, S. Venkataramani, and V. Srinivasan, "DNNDaSher: A compiler framework for dataflow compatible end-to-end acceleration on IBM AIU," *IEEE Micro*, 2024.

[19] Granite Team, "Granite 3.0 language models," https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf, IBM, Tech. Rep., 2024.

[20] C. Tillmann, A. Trivedi, and B. Bhattacharjee, "Efficient models for the detection of hate, abuse and profanity," arXiv:2402.05624, 2024.

[21] W. Zeng, Y. Liu, R. Mullins, L. Peran, J. Fernandez, H. Harkous, K. Narasimhan, D. Proud, P. Kumar, B. Radharapu, O. Sturman, and O. Wahltinez, "ShieldGemma: Generative AI content moderation based on Gemma," arXiv:2407.21772, 2024.

[22] D. Wood, B. Lublinsky, A. Roytman, S. Singh, A. Adebayo, R. Eres, M. Nassar, H. Patel, Y. Shah, C. Adam, P. Zerfos, N. Desai, D. Tsuzuku, T. Goto, M. Dolfi, S. Surendran, P. Selvam, S. An, Y. C. Chang, D. Joshi, H. Emami-Gohari, X.-H. Dang, Y. Koyfman, and S. Daijavad, "Data-prep-kit: Getting your data ready for LLM application development," arXiv:2409.18164, 2024.

[23] E. Bandel, Y. Perlitz, E. Venezian, R. Friedman-Melamed, O. Arviv, M. Orbach, S. Don-Yehyia, D. Sheinwald, A. Gera, L. Choshen, M. Shmueli-Scheuer, and Y. Katz, "Unitxt: Flexible, shareable and reusable data preparation and evaluation for generative AI," arXiv:2401.14019, 2024.

[24] G. Kour, M. Zalmanovici, N. Zwerdling, E. Goldbraich, O. N. Fandina, A. Anaby-Tavor, O. Raz, and E. Farchi, "Unveiling safety vulnerabilities of large language models," in *Proceedings of the EMNLP Workshop on Generation, Evaluation & Metrics*, 2023.

[25] M. Nagireddy, L. Chiazor, M. Singh, and I. Baldini, "SocialStigmaQA: A benchmark to uncover stigma amplification in generative language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 454–21 462.

[26] Y. Hou, A. Pascale, J. Carnerero-Cano, T. Tchrakian, R. Marinescu, E. Daly, I. Padhi, and P. Sattigeri, "WikiContradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from Wikipedia," *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

[27] A. Sokol, N. Moniz, E. M. Daly, M. Hind, and N. Chawla, "BenchmarkCards: Large language model and risk reporting," arXiv:2410.12974, 2024.

[28] G. Zizzo, G. Cornacchia, K. Fraser, M. Z. Hameed, A. Rawat, B. Buesser, M. Purcell, P.-Y. Chen, P. Sattigeri, and K. R. Varshney, "Adversarial prompt evaluation: Systematic benchmarking of guardrails against prompt input attacks on LLMs," in *Proceedings of the NeurIPS Safe Generative AI Workshop*, 2024.

[29] Q. Pan, Z. Ashktorab, M. Desmond, M. Santillan Cooper, J. Johnson, R. Nair, E. M. Daly, and W. Geyer, "Human-centered design recommendations for LLM-as-a-judge," arXiv:2407.03479, 2024.

[30] S. Achintalwar, I. Baldini, D. Bouneffouf, J. Byamugisha, M. Chang, P. Dognin, E. Farchi, N. Makondo, A. Mojsilović, M. Nagireddy, K. N. Ramamurthy, I. Padhi, O. Raz, J. Rios, P. Sattigeri, M. Singh, S. Thwala, R. A. Uceda-Sosa, and K. R. Varshney, "Alignment studio: Aligning large language models to particular contextual regulations," *IEEE Internet Computing*, vol. 28, no. 5, pp. 28–36, Sep.–Oct. 2024.