Usage Governance Advisor: From Intent to AI Governance

Elizabeth M. Daly, Seshu Tirupathi, Sean Rooney, Inge Vejsbjerg, Dhaval Salwala, Christopher Giblin, Frank Bagehorn, Luis Garces-Erice, Peter Urbanetz, Mira L. Wolf-Bauwens

IBM Research - Europe

Abstract

Bringing a new AI system into a production environment involves multiple different stakeholders such as business owners, risk officer, ethics officers approving the AI System for a specific usage. Governance frameworks typically include multiple manual steps, including curating information needed to assess risks and reviewing outcomes to identify appropriate actions and governance strategies. We demo a human-in-the-loop automation system that takes a natural language description of an intended use case for an AI system in order to create semi-structured governance information, recommend the most appropriate model for that use case, prioritise risks to be evaluated, automatically running those evaluations and finally storing these results for auditing, reporting and future recommendations. As a result we increase transparency to stakeholders and provide valuable information to aid in decision making when assessing risks associated with an AI solution.

Introduction

The impressive performance of Large Language Models (LLMs) has catapulted the industry into exploring their potential for a wide range of tasks. However, one barrier for organisations advancing from prototype to full deployment of these AI solutions are the perceived risks. The vast capabilities that make these models attractive solutions mean the potential for risks is also great (Schillaci 2024). In order to enable organisations to embrace AI solutions, governance and transparency must be heavily integrated into the AI life-cycle which involves many different stakeholders (Richards et al. 2020; Cihon, Schuett, and Baum 2021). Given the computational requirements for creating LLMs from scratch, many organisations are choosing to leverage openly available pre-trained models. Questionnaires are a commonly used tool for Trustworthy AI frameworks to ensure risks and mitigation have been taken into account and can serve to provide transparency to the various stakeholders and to demonstrate compliance with regulatory requirements. As highlighted by (Derczynski et al. 2023), risks don't depend just on the model itself but also on the intended use case for that model. This means static questionnaires are not sufficient and additional tooling is required. Identified risks need to be assessed and the information required to understand these risks can include a combination of heterogeneous information ranging from structured benchmark results to model and data cards in natural language detailing the model training pipeline. Assessing the proposed use case of an AI application and collecting all the relevant information can be a time consuming process. The authors in (Ferdaus et al. 2024) provide a comprehensive overview on the opportunities, challenges and limitations of trustworthy AI including existing governance frameworks with particular focus on LLMs. However, despite research and conceptual frameworks that have explored individual facets of governing AI and LLMs, a comprehensive, systemic approach has yet to be undertaken. This demo paper focuses on developing an automated human-in-the-loop methodology to identify and address risks associated with LLM-based applications. `Usage Governance Advisor` assists organisations in creating documentation detailing the expected use case, automatic identification of risks, recommendation of appropriate models and automating the appropriate risk assessment evaluations (for further details please see (Daly et al. 2024)). The solution builds on a knowledge graph which structures the information from heterogeneous data about the models and uses existing risk taxonomies.

System Overview

`Usage Governance Advisor` provides work flows taking the user-intent and multiple sources of information to seamlessly guide the user through the AI governance onboarding process and transparently deploy an LLM model into production for the given use case (Daly et al. 2024). The system assists the end-user fill out compliance questionnaires, identify risks associated with the user intent, generate risk reports for the shortlisted LLM models which includes user preferences and store the results of these risk reports for compliance and auditing. Our solution makes the user intent and context where an LLM will be deployed first-class citizens in the LLM development and deployment pipeline. Figure 1 describes the solution which to aims to provide AI governance while lowering the barrier of entry for the user.

AI Systems Knowledge Graph: We have defined an AI Systems ontology (based on (Golpayegani, Pandit, and Lewis 2022)) which describes pertinent aspects of AI system relevant for guiding stakeholders in their usage. We have used information from Risk taxonomies (IBM 2023;

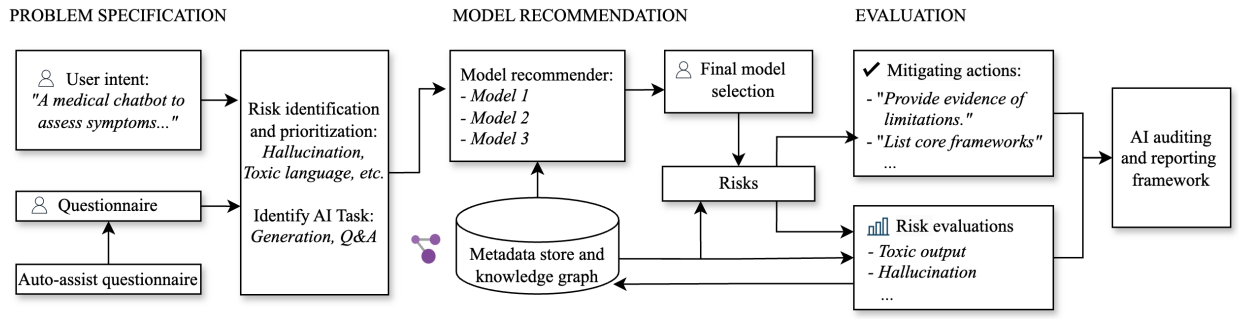

Figure 1: Usage Governance Advisor pipeline

OWASP 2023; NIST 2023; MIT 2024), government regulation (European Parliament and Council of the European Union 2024) and measures of model transparency, e.g the Stanford Transparency Index (Bommasani et al. 2024) to determine the most important aspects to include in this ontology. We materialize this ontology in a Knowledge Graph (KG) that we populate using an automated ingestion process from canonical sources of metadata such as model cards, technical reports etc. This ingestion process uses generative AI to perform entity recognition and linkage. We maintain both a domain graph containing the extracted facts and an evidence graph pointing to the sources of information for those facts and an indication of our confidence in their truth.

Auto-assist questionnaire: Questionnaires are a common tool used by organisations in order to identify risks for AI use cases. Our solution aims to auto-suggest the information required for the questionnaires in order to create a transparent record in a more semi-structured manner to capture the intended use of the AI system. We used few-shot and chain-of-thought (Wei et al. 2022) approaches for auto assist functionality by giving multiple examples for every question on how to answer the question. Some questions require trivial few-shot examples while others require chain-of-thought examples with breakdowns on how answers need to be framed and reasoned to get the right answers.

Use case to risks: The use case definition and responses from the questionnaire can be used to prioritize specific risks associated with the AI tasks for an AI use case. For example, a model used for classification will not be accessible to outside use and its output is heavily processed into one of several possible classes. As a result, the risk of generating toxic language is not a priority. Our solution produces a prioritised list of risks associated with a given task taking into account the answers to the governance questions. For example, the description may explicitly detail human oversight procedures, thus mitigating some risks. An llm-as-a-judge approach is used connecting questions/answer pairs to risks and whether specific answers reduce or amplify a risk (Zheng et al. 2023).

Model Recommender: Given the prioritized list of risks for a given AI task, a risk profile can be created. This risk profile is augmented with a customer policy that defines threshold values for risks (i.e., their tolerance to those risks). The KG is then queried to produce a list of ranked models, given the information available in the KG and the submitted profile. Information that is unavailable in the KG is identified and the user may be prompted with recommended actions, like executing a benchmark or obtaining information from the model vendors. Historic information about models already in use and assessed for similar purposes along with deployment behaviour is available within the KG. While the AI system developer may propose a model, the recommended model can be used as a challenger model to compare with.

Automated Risk Evaluations: Each risk is linked to pre-defined benchmarks aimed to assess the risk profile of the model. These assessments range from social stigma to toxic output (Nagireddy et al. 2024), to hallucinations (Lin, Hilton, and Evans 2021). Our tool takes in a list of prioritised risks and executes the evaluations in order to allow risk officers assess the appropriateness of the model. Risks are also linked to pre-defined actions, in particular, when automatic measurements are not supported, for example in the case of model transparency the mitigation may be to ensure proper documentation associated with the model is available.

Persistence of Governance Information: Finally, use case details, metadata, and risk reports are persisted in order to support transparency and accountability as is best practice (Raji et al. 2020; Arnold et al. 2019). These assets can be used to ensure compliance and also keep the stakeholders informed. The records associated with the AI system are updated with the latest test results, ensuring that model certifications remain current and reflect any changes in model deployment configurations.

The end result is a tool that can be used to assess the risks associated with a given model for a given use case. This information can then be leveraged by not only the system developer but other stake holders to make informed decisions with transparent meta information on approving an AI system. Future work aims to augment the knowledge graph with automated risk mitigation strategies such as guardrails in order to support not only initial deployment governance but prescribe governance for the whole model life cycle.

References

Arnold, M.; Bellamy, R. K.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K. N.; Olteanu, A.; Piorkowski, D.; et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5): 6–1.

Bommasani, R.; Klyman, K.; Kapoor, S.; Longpre, S.; Xiong, B.; Maslej, N.; and Liang, P. 2024. The Foundation Model Transparency Index v1.1: May 2024. arXiv:2407.12929.

Cihon, P.; Schuett, J.; and Baum, S. D. 2021. Corporate governance of artificial intelligence in the public interest. *Information*, 12(7): 275.

Daly, E. M.; Rooney, S.; Tirupathi, S.; Garces-Erice, L.; Vejsbjerg, I.; Bagehorn, F.; Salwala, D.; Giblin, C.; Wolf-Bauwens, M. L.; Giurgiu, I.; Hind, M.; and Urbanetz, P. 2024. Usage Governance Advisor: from Intent to AI Governance. arXiv:2412.01957.

Derczynski, L.; Kirk, H.; Balachandran, V.; Kumar, S.; Tsvetkov, Y.; Leiser, M.; and Mohammad, S. 2023. Assessing language model deployment with risk cards. arXiv.

European Parliament; and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council. *OJ*, (L 2024/1689).

Ferdaus, M. M.; Abdelguerfi, M.; Ioup, E.; Niles, K. N.; Pathak, K.; and Sloan, S. 2024. Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models. *arXiv preprint arXiv:2407.13934*.

Golpayegani, D.; Pandit, H.; and Lewis, D. 2022. *AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards*. ISBN 9781643683201.

IBM. 2023. AI Risk Atlas. https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas. Accessed: 2024-12-09.

Lin, S.; Hilton, J.; and Evans, O. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958.

MIT. 2024. The AI Risk Repository. https://airisk.mit.edu/. Accessed: 2024-12-09.

Nagireddy, M.; Chiazor, L.; Singh, M.; and Baldini, I. 2024. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21454–21462.

NIST. 2023. AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework. Accessed: 2024-12-09.

OWASP. 2023. OWASP Top 10 for LLMs and Generative AI Apps. https://genai.owasp.org/llm-top-10/. Accessed: 2024-12-09.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Richards, J.; Piorkowski, D.; Hind, M.; Houde, S.; and Mojsilović, A. 2020. A methodology for creating AI FactSheets. *arXiv preprint arXiv:2006.13796*.

Schillaci, Z. 2024. LLM Adoption Trends and Associated Risks. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, 121–128. Springer Nature Switzerland Cham.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.