LLM-powered framework for AI governance in Industry 4.0

No Author Given

No Institute Given

Abstract. Artificial intelligence (AI) governance involves ensuring transparency, accountability, fairness, and data privacy. Various countries propose regulatory frameworks, creating challenges such as aligning different regulations, identifying gaps between jurisdictions, and keeping up with evolving AI laws. Large language models (LLMs) could assist in AI governance, but they often generate factually incorrect responses and require expertise to use effectively. To address these issues, we introduce an LLM-powered framework for AI governance. This solution enables users, particularly those without technical expertise, to verify compliance with regulations and ethical obligations. Our framework provides tools, datasets, and guidelines to simplify regulatory alignment. Additionally, we showcase its components through practical use cases, demonstrating how it aids compliance verification. Our work serves as a baseline for AI governance solutions, offering accessible resources for organizations. Code, dataset, and documentation are available at `https://github.com/raulsenaferreira/AIGOV_paper.git`.

Keywords: Large language Models · AI Governance · Industry 4.0.

1 Introduction

As AI systems become integral to society, their governance becomes increasingly complex. Ensuring compliance with various legal regulations and adhering to ethical principles is crucial for mitigating the potential risks associated with AI, including bias, discrimination, lack of transparency, and privacy violations. AI governance frameworks often emphasize core principles [32] such as transparency [12], accountability, fairness, and privacy. However, practical tools for operationalizing them across different domains remain limited.

Modern large language models (LLMs) [37] can process and understand vast amounts of text. By applying LLMs to AI governance, we can expect to see generic solutions that interpret, analyze, and compare legal regulations and ethical guidelines. For example, LLMs can help simplify dense legal texts, compare international regulations, and identify critical ethical considerations in AI systems. However, LLMs are also known to hallucinate when not prompted correctly [19]. Selecting the most suitable prompt engineering techniques for a specific problem can be time-consuming. Moreover, LLMs tend to be less accurate when the user requests reasoning that requires context from private knowledge

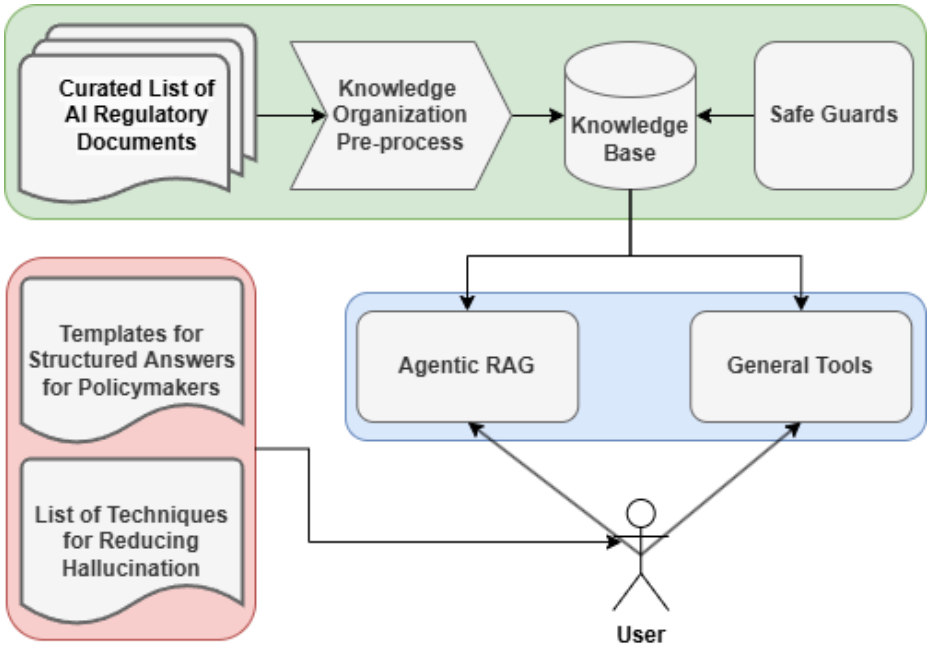

Fig. 1. Overview divided into three parts: knowledge construction (green), templates and techniques for the users (red), and the LLM-powered engine (blue).

bases, which creates the need for retrieval-augmented generation (RAG) systems to improve the factual accuracy of such LLMs [22]. However, applying the proper prompts and architecture for LLM-based solutions is not straightforward, since it often requires expert knowledge for implementation.

Despite the impressive capabilities of recent LLM-based models like GPT-4, such models' solutions are often constrained by proprietary limitations, external dependencies, and data privacy concerns. Organizations that handle sensitive data often require on-premises tools to maintain control over their information and comply with stringent privacy regulations. There is a clear need for ready-to-install frameworks within an organization's infrastructure to enable companies and institutions to enhance knowledge extraction and interpretation of regulatory documents without risking data exposure through online platforms. By providing such practical applications of LLMs to critical governance tasks and introducing advanced analysis techniques for navigating the complexities of global AI governance, such solutions remain useful and relevant in the current landscape. Therefore, we propose a comprehensive LLM-powered framework for AI governance that is capable of addressing the challenges mentioned above.

As illustrated in Figure 1, we provide a knowledge creation process, a list of prompt templates and techniques, and an LLM-based engine that can be connected to a visual interface. This framework has the objective of a) being

a better entry point for non-tech experts who use vanilla LLM solutions or no technological support for AI governance, and b) being a baseline for future research in using LLMs in AI governance. Our main contributions are:

- A ready-to-use framework to enhance the extraction and interpretation of knowledge from regulatory documents.
- Practical examples of how LLMs can be applied to critical governance tasks.
- Advanced analysis techniques to handle the complexities of AI governance.

The paper is organized as follows: In Section 2, we navigate through AI governance literature organized by the core principles of this domain. In Section 3, we explain how our framework works, while in Section 4, we present the first experiments, results, and analysis. Section 5 briefly discusses related works and compares how our solution differs from them. Finally, in Section 6, we conclude this work and indicate topics for possible future research.

2 LLM-Powered AI Governance Framework

The framework contains a modular structure that is easy to integrate with organizations' governance and compliance systems. We feed regulations, ethical guidelines, and other government guidelines into the system. The LLMs perform summarization, entity recognition, sentiment analysis, keyword extraction, and causal inference. The system outputs summaries, legal terms, compliance alerts, or comparative analysis reports, which organizations use to improve AI governance. In the following, we present the three main modules of our work: knowledge construction, prompt engineering templates, and the LLM engine.

2.1 Knowledge Construction

This module creates a knowledge base that the LLM engine will use. The process starts with a manually curated list of AI regulatory documents. After that, a preprocessing data mechanism organizes the documents divided by country. After the knowledge creation, a safeguard function inspects it to verify inconsistencies such as repeated or outdated data.

2.2 Prompt Engineering Templates

This module is auxiliary to the LLM engine module. To ensure reliability and reduce hallucination, we rely on three principles when prompting an LLM:

- Cite Sources: Request that the model cites specific articles or sections of the referenced documents.
- Clarify Scope: Limit the scope of the prompt to avoid ambiguous responses.
- Request Verification: Instruct the model to acknowledge gaps in knowledge or state when information is missing.

We have an initial list of prompt engineering techniques for reducing hallucination risk and a list of templates for structured responses for policymakers (provided in the associated Github repository).

Techniques for Reducing Hallucination: To improve the factual accuracy of LLM outputs and reduce hallucination, targeted prompting strategies are essential. The following techniques help ground responses in the provided information:

1. **Instruction-Based Prompts Technique**: It provides clear and specific prompts to guide the model toward factual, structured outputs.
2. **Context-Rich Prompts Technique**: It embeds context in the prompt to give the model a detailed understanding of the type of answer expected. - **Example**: *"Using the following section of the EU AI Act, summarize the transparency requirements and list them in bullet points with corresponding legal references: {insert document excerpt}."*
3. **Prompt with Reference Verification Technique**: It asks the model to verify its answers and state when information is not in the reference material.

Templates for Structured Answers for Policymakers Providing structured answers is crucial for helping policymakers understand and interpret the information generated by an LLM.

1. **Comparative Analysis Template**: This format highlights the main points in a way that policymakers can quickly grasp.
2. **Summary with Citations Template**: Providing citations ensures accuracy and allows policymakers to refer to the original text for confirmation.
3. **Policy Recommendation Template**: This structured format guides policymakers through the current issues, and specific policy steps to consider.

2.3 LLM-powered Engine

As illustrated in Figure 2, the user query is interpreted by an LLM Agent acting as a domain expert, which is responsible for planning and calling tools as needed. These tools can connect to specialized RAGs for different domains. For example, one specializes in data compliance within the European Union, another in AI governance for Brazil, and so on. It could also be connected to an LLM specialized for specific tasks by prompt engineering or fine-tuning.

Each tool can call the same LLM model with domain-specific, prompt-engineered templates or call RAG systems that contain regulatory documents split by domain and area. In this work, we utilize LightRAG [15] as the RAG system, employing the React [44] approach. Beyond the agentic RAG component, the LLM-powered Engine also incorporates a suite of general tools designed to address broader AI governance challenges. These tools provide functionalities for tasks such as regulation summarization, ethical framework analysis, and causal inference in AI case studies, complementing the domain-specific capabilities of the agentic RAG. Next, we present the first analysis of this framework.

3 Experiments and Results

We present the first results and analysis of our framework, which are divided into two parts: RAG and General Tools. Our repository contains the code for experiments and additional results not included here due to space limitations.

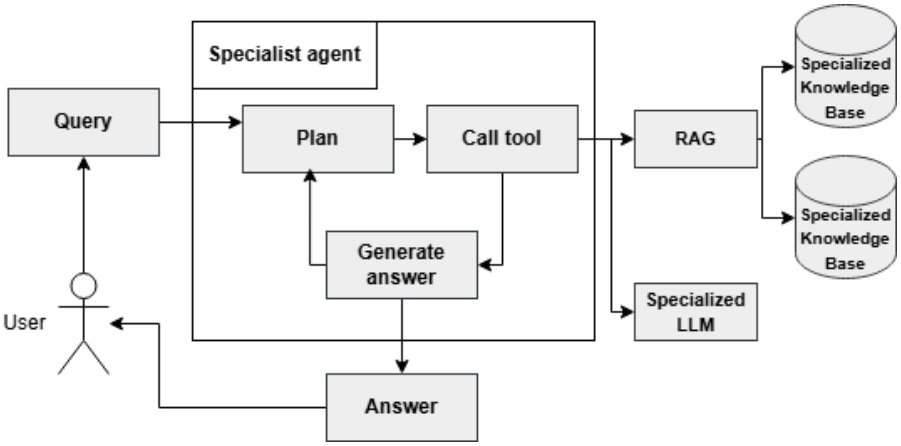

Fig. 2. Agentic RAG approach: An LLM-based agent is responsible for evaluating the user prompt and planning a domain-specific "Tools".

We used data from 30 government documents on six continents, as presented in Table 1. The idea is not to cover all documents available but to test our approach with a minimum yet broad regulatory scope.

Table 1. Governmental data collected for this work.

Continent	Country
Asia	China [33, 6], Japan [21, 20], South Korea [36], UAE [39, 13], Saudi Arabia [34], Israel [25, 26], Singapore [35]
Africa	All countries [40], [2, 1], South Africa [3]
North America	USA [31, 42, 5, 41], Canada [24, 38]
South America	Brazil [30], Argentina [29]
Europe	All countries [43, 23]
Oceania (Region)	Australia [9, 10], APEC [8], New Zealand [7, 17]

We evaluated our agentic RAG in three use cases (UC):

1. *Extracting knowledge from a specific document from a specific domain.*
2. *Extracting knowledge from different documents from a specific domain.*
3. *Assessing a specification and comparing knowledge from a specific domain.*

For each of the UCs, we show the complete chain of reasoning, planning, and acting. For the images, the user input is illustrated in green, the final output for the user is in yellow, and the final LLM remarks are in orange.

UC 1 Results In this use case, we applied the *Instruction-Based Prompts Technique*. The prompt helps reduce hallucinations by focusing the response on specific concepts and requesting citations from the source document.

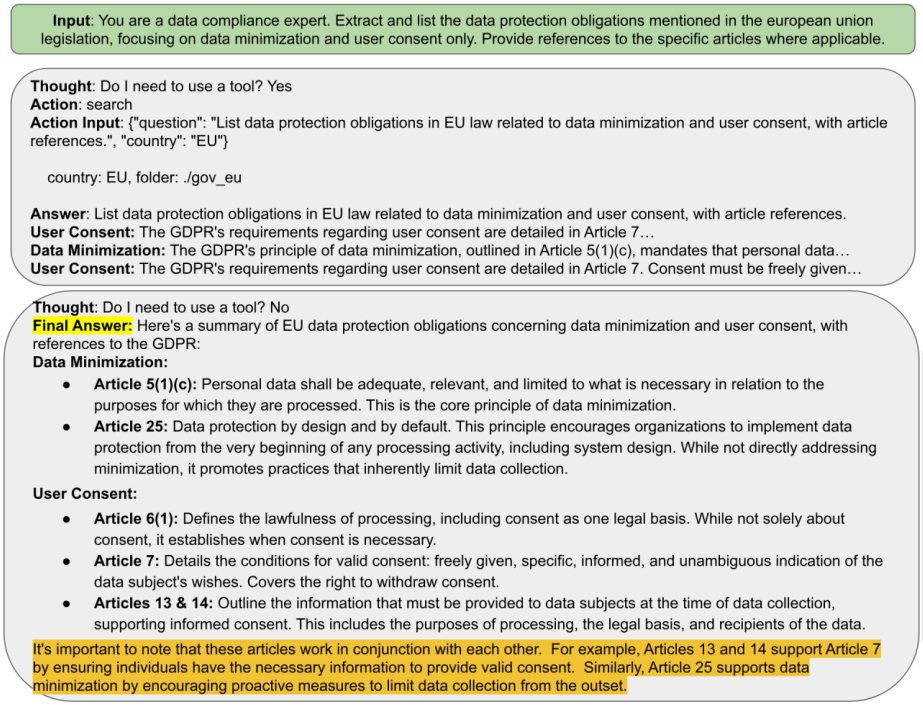

Fig. 3. UC 1 Results: The model impersonates an expert from a specific country.

As shown in Figure 3, the agent processes a reasoning step to decide whether a specific tool should be called to get more information. Then it retrieves the relevant cluster of documents and starts the knowledge retrieval process, followed by reasoning through intermediate answers. In the second part, the agent iterates over the intermediate answer and assesses whether to call an external tool again. Correctly gives a detailed answer while giving a final remark (in orange). It is worth mentioning that this remark increases the reliability of the output by giving the sources of critical information used in the answer.

UC 2 Results We applied a *Prompt with Reference Verification Technique*. The prompt ensures that the model outputs only what the input documents support.

As shown in Figure 4, the reasoning process is similar to the first UC. However, the agent first retrieves information from the regulations of the two countries and compares it once the intermediate information is available. Finally, the LLM also provides remarks after the final answer, highlighting possible nuances depending on the documents used in the knowledge base.

UC 3 Results As shown in Figure 5, in this example, we provide the specification of an AI-based driver-behavior prediction system used to validate an urban

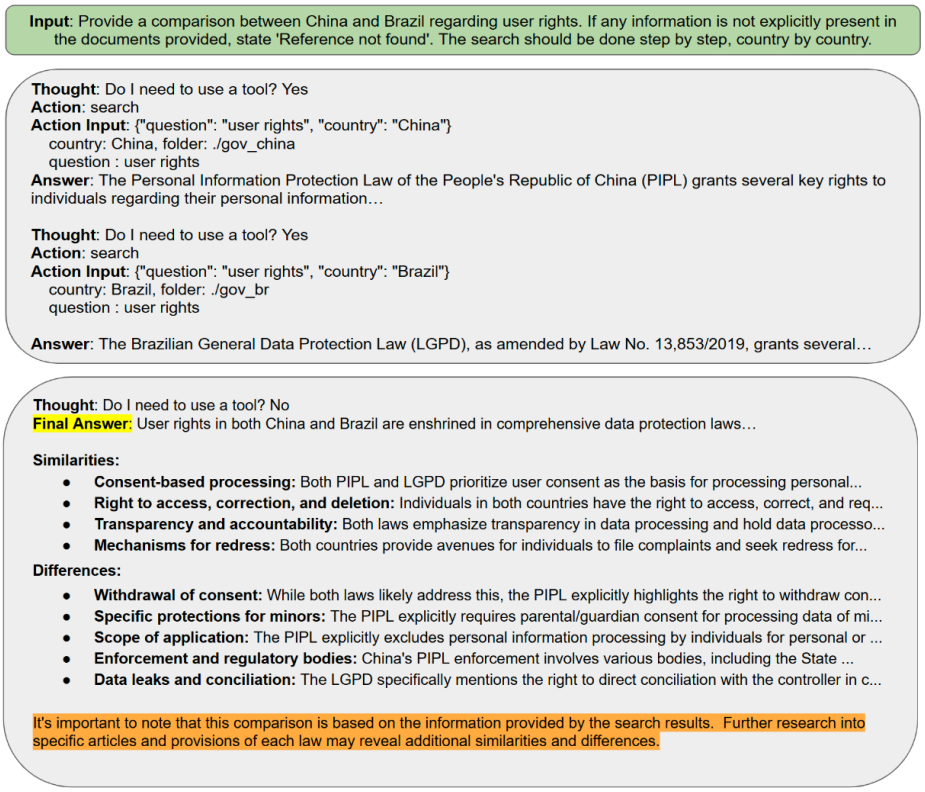

Fig. 4. UC 2 Results: This comparison examines the user rights of two countries from different continents.

pedestrian-crossing assistance function. The agent extracts each component of the specification — training data, model documentation, transparency, robustness, human oversight, and monitoring — and reasons over it against the AI Act and GDPR using the Prompt with Reference Verification Technique. Because several elements of the specification, notably the data-retention policy and the post-deployment monitoring procedure, are not documented, the model does not force a binary compliant/non-compliant label onto those items; instead, it reports them as uncertain, reserving a negative verdict for the one dimension where the gap is explicitly confirmed rather than merely undocumented. The final remark (in orange) recommends producing the missing documentation before a conformity assessment can be completed. Next, we show the differences from the related works.

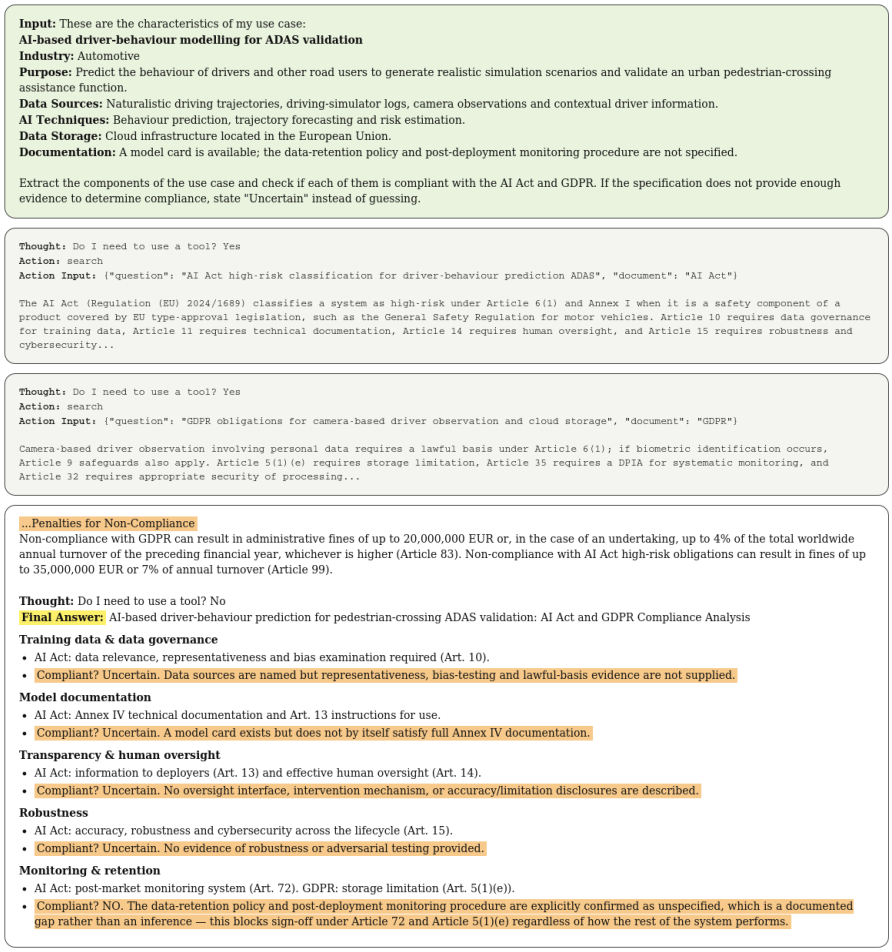

Fig. 5. UC 3 Results: We provide a specification of an automotive driver-behavior prediction system for pedestrian-crossing ADAS validation. We ask for a compliance check of these specifications against the EU AI Act and GDPR, including cases where the model must report "Uncertain" due to insufficient evidence rather than guess a verdict.

4 Related Work

Recent literature emphasizes the importance of aligning LLM development with core AI governance principles.

Transparency in LLMs includes transparency in the decision-making processes and the data on which they are trained. [16] outlines core principles for guiding and evaluating LLM reasoning, highlighting the need for structured reasoning and self-assessment to enhance transparency. Similarly, [4] has discussed

the dangers of "stochastic parrots" and assessed the importance of understanding data sources and training processes to avoid unintended outputs.

Ensuring accountability in LLMs requires mechanisms to trace and attribute decisions to responsible entities. [28] proposed a three-layered LLM auditing approach, encompassing governance, model, and application audits, to establish clear accountability structures. This framework facilitates the identification and management of ethical risks associated with LLMs.

Addressing biases inherent in LLMs is crucial for fairness. [18] proposed AI ethical principles and guidelines for the governance of advanced LLMs, focusing on responsibility and accountability to reduce biases and ensure fair outcomes. They aim to demonstrate the importance of proactive measures in eliminating harm and promoting fairness in LLM-based applications.

LLMs often process vast amounts of data, raising significant privacy concerns. [14] suggested implementing the Dual Use Research of Concern (DURC) framework, originally designed for life sciences, in the domain of generative AI, with a specific focus on LLMs. This approach addresses privacy issues by establishing guidelines for data usage and preventing misuse.

Ensuring the safety of LLMs is critical to prevent unwanted, harmful outputs. [11] provides a comprehensive overview of recent studies on LLM conversation safety, which aims to cover essential aspects such as attacks, defenses, and evaluations. Their work highlights the importance of structured safety measures to address the risks associated with LLM deployment.

In contrast to these works that often focus on high-level ethical principles, theoretical frameworks, or policy implications, this paper proposes practical solutions by utilizing Large Language Models (LLMs) to operationalize core AI governance principles. By analyzing existing regulations and laws, we provide an LLM-based approach to synthesize regulatory texts, analyze compliance pathways, and compare regulatory definitions across different jurisdictions. This approach enables actionable insights into regulatory and ethical landscapes, facilitating the application of governance frameworks in practice and bridging the gap between abstract principles and their implementation.

5 Conclusion

In this work, we presented an LLM-powered framework for AI governance that bridges the gap between theoretical principles of AI governance. We demonstrated using this framework through practical applications to navigate complex legal and ethical frameworks by using LLMs to extract, summarize, and analyze relevant governance documents. We showed how LLMs can enhance the scalability and efficiency of the AI governance process while being adaptive to this constantly evolving domain. The framework can scale with the growing complexity of AI regulations and adapt to different industries and jurisdictions.

It is worth mentioning that this work is essential to foster initiatives independent of closed/commercial solutions while being flexible enough to be used by companies, institutions, or even individuals handling sensitive information.

However, while the proposed framework offers a robust solution, there are potential challenges to consider:

- Data Privacy and Security: Processing sensitive legal and regulatory texts using LLMs may require additional safeguards to protect confidential information.
- Regulatory Updates: The dynamic nature of AI regulations means that our dataset must be continuously updated to comply with new laws and policies.
- Bias in LLMs: Ensuring that the LLMs used in this framework are free from bias is essential, as biased LLMs can misinterpret regulations or ethical principles.
- Need of System Cards: these cards are the junction of data cards (e.g., information about the data used to build an AI system.) and model cards [27] (e.g., disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information). Such cards are used as a reference document when compliance verification against regulation documents is needed.

Finally, as new AI governance challenges arise, LLMs can be retrained or updated to incorporate new regulatory texts and ethical guidelines, ensuring the framework remains up-to-date. Hence, future research could determine how regulatory datasets should be created from official regulatory documents to fine-tune LLMs for specific sub-domains. Additionally, developing more sophisticated bias mitigation techniques and expanding the framework's capacity for ethical risk assessment can be a crucial future research direction.

6 Acknowledgments

Research activity under the BERTHA project (GA101076360) funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

References

1. Act, P.: Protection of personal information act. Retrieved June **23**, 2021 (2020)
2. African Union: Personal data protection guidelines for africa (2018)
3. Artificial Intelligence Institute of South Africa: South's africa draft national artificial intelligence (ai) plan. `https://www.dcdt.gov.za/images/phocadownload/AI_Government_Summit/National_AI_Government_Summit_Discussion_Document.pdf` (2023), online; accessed 2024-11-01
4. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)

5. Bonta, R.: California consumer privacy act (ccpa). Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa (2022)
6. China's National Technical Committee 260 on Cybersecurity: Ai safety governance framework. https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf (2024), online; accessed 2024-11-01
7. Commissioner, P.: Privacy act 2020 (2020)
8. Cooperation, A.P.E.: Apec privacy framework (2005)
9. Department of Industry, Science, Energy and Resources: Australia's ai action plan, australian government (2021)
10. Digital Transformation Agency, G.o.A.: Policy for the responsible use of ai in government. Analysis & Policy Observatory (2024)
11. Dong, Z., Zhou, Z., Yang, C., Shao, J., Qiao, Y.: Attacks, defenses and evaluations for llm conversation safety: A survey. arXiv preprint arXiv:2402.09283 (2024)
12. Ferreira, R., Praia, V., Bonecini, F., Vieira, A., Lopez, F., et al.: Platform of the brazilian csos: open government data and crowdsourcing for the promotion of citizenship. In: Simpósio Brasileiro de Sistemas de Informação (SBSI). pp. 17–24. SBC (2017)
13. Government, U.: Uae data protection law (federal decree-law no. 45 of 2021) (2021)
14. Grinbaum, A., Adomaitis, L.: Dual use concerns of generative ai and large language models. Journal of Responsible Innovation **11**(1), 2304381 (2024)
15. Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C.: Lightrag: Simple and fast retrieval-augmented generation. arXiv preprint arXiv:2410.05779 (2024)
16. Hebenstreit, K., Praas, R., Samwald, M.: A collection of principles for guiding and evaluating large language models. arXiv preprint arXiv:2312.10059 (2023)
17. Hon Judith Collins KC, N.Z.G.: Approach to work on artificial intelligence. https://www.mbie.govt.nz/dmsdocument/28913-approach-to-work-on-artificial-intelligence-proactiverelease-pdf (2024), online; accessed 2024-10-30
18. Hossain, S., Ahmed, S.I.: Ethical artificial intelligence principles and guidelines for the governance and utilization of highly advanced large language models. arXiv preprint arXiv:2401.10745 (2023)
19. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems (2023)
20. Japan Government: Japan act on the protection of personal information. https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en (2021), online; accessed 2024-11-01
21. Japan Government: Japan's ai strategy. https://www8.cao.go.jp/cstp/ai/aistratagy2022en.pdf (2022), online; acessed 2024-11-01
22. Li, J., Yuan, Y., Zhang, Z.: Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. arXiv preprint arXiv:2403.10446 (2024)
23. Madiega, T.: Artificial intelligence act. European Parliament: European Parliamentary Research Service (2021)
24. MINISTER OF INNOVATION, SCIENCE AND INDUSTRY: Consumer privacy protection act. https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading (2022), online; accessed 2024-11-01
25. Ministry of the Justice of Israel: Protection of privacy regulations (data security) 5777-2017 (2017)

26. Ministry of the Justice of Israel: Privacy and data security in deepfake technologies (2023)
27. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 220–229 (2019)
28. Mökander, J., Schuett, J., Kirk, H.R., Floridi, L.: Auditing large language models: a three-layered approach. AI and Ethics pp. 1–31 (2023)
29. NATIONAL CONGRESS: Argentinian data protection law (ley de proteccion de datos personales). https://www.argentina.gob.ar/sites/default/files/antecedente_2018_25326.pdf (2018), online; accessed 2024-11-01
30. NATIONAL CONGRESS: Brazilian general data protection law (lgpd). https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf (2018), online; accessed 2024-11-01
31. NIST: Artificial intelligence risk management framework (ai rmf 1.0) (2023)
32. OECD: Tools for Trustworthy AI: A Framework to Compare Implementation Tools for Trustworthy AI Systems. OECD Publishing (2021)
33. Republic of China: China personal information protection law (pipl). https://www.china-briefing.com/news/the-prc-personal-information-protection-law-final-a-full-translation/ (2020), online; accessed 2024-11-01
34. Saudi Arabia Government: Saudi arabia personal data protection law (2023)
35. Singapore Government: Singapore national ai strategy 2.0 (nais 2.0). https://file.go.gov.sg/nais2023.pdf (2023), online; 2024-11-01
36. South Korean Government: South korea national ai strategy. https://doc.msit.go.kr/SynapDocViewServer/viewer/doc.html?key=8397440df46347199b8104d898eda738&convType=img&convLocale=ko_KR&contextPath=/SynapDocViewServer (2021), online; accessed 2024-11-01
37. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
38. Treasury Board of Canada: Canada directive on automated decision-making. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592&section=html (2019), online; accessed 2024-11-01
39. UAE Government: Uae national strategy for artificial intelligence 2031 (2017)
40. Union, A.: African union convention on cyber security and personal data protection. African Union **27** (2014)
41. US Congress: Us algorithmic accountability. https://www.congress.gov/117/bills/hr6580/BILLS-117hr6580ih.pdf (2021), online; accessed 2024-11-01
42. US WhiteHouse: Blueprint for an ai bill of rights, making automated systems work for the american people. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf (2022), online; accessed 2024-11-01
43. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)
44. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)