Concept Note on Privacy and Data Protection Risks in Large Language Models (LLMs)

Authors: Isabel Barberá, Murielle Popa-Fabre, Chris Russell

Large Language Models (LLMs) present significant privacy and data protection risks, including personal data exposure, inaccuracies, and misuse of AI-generated outputs. A major concern is data leakage and memorization, where LLMs inadvertently reproduce sensitive information, leading to potential breaches. Successful adversarial attacks allow malicious actors to manipulate outputs or extract confidential data, further undermining the privacy protections in place. Lack of transparency, user control, and consent mechanisms complicates compliance with privacy regulations, making it difficult for users to understand and manage how the data they provide as inputs is processed. But beyond technical privacy concerns, LLMs have the potential to interfere with the broader right to private life, affecting individuals in ways that cannot be fully addressed through standard data protection approaches. This includes risks related to personal autonomy, reputational harm, and the ability to control one's own digital presence. Organizations developing, deploying, or integrating LLMs must adopt a structured approach to identifying and mitigating these risks, with a broader concern for the right to private life as standard privacy related approaches fall short to address the new challenges brought by LLMs.

This document outlines the main aspects of an expert report to inform the elaboration of a normative document by the Committee of Convention 108 on privacy and data protection in LLMs. Such document could also give a guidance to assist organizations in assessing and mitigating privacy risks while promoting awareness and compliance with international human rights standards, through an integrated approach to Artificial Intelligence governance, Privacy and Data Protection.

Scope of Work - A privacy risks framework

The primary goal of the report is to present a privacy risk framework that provides clear, actionable guidance for organizations dealing with LLMs based on Convention 108+.

1-Identifying privacy risks and recommendations based on Convention 108+. A crucial aspect of this work involves determining which privacy risks emerge at different stages of AI development and deployment), how they arise, and identifying the appropriate measures in line with Convention 108+, including where available through best practices to mitigate them. This requires a combination of theoretical research, practical case studies, and engagement with stakeholders from industry, governments, regulatory bodies, and academia.

2- Two–tiered Risk assessment methodology. A foundational component of this report will be the research and development of robust risk identification and assessment methodologies. Data controllers must be equipped with tools to evaluate risks at both the model level—such as training data privacy concerns, hallucinations, and model memorization of sensitive information—and the system level, where risks arise from integration, user interactions, and third-party dependencies.

3-Pointing at viable mitigation strategies in real-world applications. Developing a structured approach to Privacy Impact Assessments tailored for LLM-based applications will be essential, ensuring that privacy risks are systematically analysed and addressed throughout the AI lifecycle. In addition, the incorporation of privacy-enhancing technologies (PETs) like differential privacy, federated learning, and encryption techniques, as well as general machine learning like finetuning, or techniques for the identification of personal data such as mechanical interpretability will be explored as viable mitigation strategies.

Beyond risk management, practical engagement with organizations using LLMs and agentic workflows will be necessary to ground the report in real-world applications. Through case studies, consultations, and hands-on analysis, the report will reflect industry best practices and common challenges. Engaging with key stakeholders—including AI developers, deployers, policymakers, and civil society organizations—will provide valuable insights into the evolving risk landscape and the effectiveness of existing mitigation strategies. This research will inform a dynamic, adaptable methodology that can evolve alongside technological advancements and that is aligned with the current and ongoing work of regulators in line with Convention 108+.

Expected outcome - Methodology Piloting

A critical element of the report will be to analyze comprehensive measures needed for dealing effectively with impact assessment. The application and piloting of a structured risk management lifecycle framework, based on one of the expert's own research, will help define adaptable risk management strategies across the four key phases of LLM-based applications: model development, system integration and deployment, and end-user interaction and possibly in autonomous agentic workflows. Each phase presents unique risks, from ensuring privacy-aware data collection in model development to implementing safeguards when customizing pre-trained models for deployment. At the end-user level, strategies to prevent sensitive data leakage and to enhance user awareness of privacy implications will be outlined.

To ensure the effectiveness of the proposed risk management framework, a piloting phase will be integrated into the project. Selected organizations will be invited to test the methodology in real-world settings, allowing for iterative refinement based on practical feedback. This pilot process will help identify gaps, validate the feasibility of the methodology, its steps and recommendations, and ensure that the proposed measures are actionable for different types of LLM providers and deployers. The piloting phase will also facilitate collaboration with regulatory bodies and sandboxing initiatives, ensuring alignment with existing and emerging legal frameworks. Findings from this phase will be used to enhance the final report, making it more adaptable and applicable across different use cases.

Committee's Support

The committee plays a crucial role in developing international standards on data protection and the use of new technologies. The successful implementation of this initiative will also support the development of clear, standardized guidelines that translate data protection and privacy principles—such as those outlined in Convention 108+ — into LLM-specific guidance and best practices, will provide data controllers and competent authorities with a coherent regulatory framework to manage privacy challenges.

Given the global nature of AI development and deployment, inconsistencies in privacy regulations may create compliance challenges and weaken user protections. International collaboration can help establish common baseline standards for LLM privacy governance. Promoting cross-border regulatory cooperation in line with article 17 of Convention 108+ can be beneficial in ensuring that privacy risk management strategies remain aligned across jurisdictions.

Furthermore, advocating for regulatory clarity in defining the responsibilities of different actors within the ecosystem of LLM-based systems can help raise the awareness globally on the issues at stake, contribute to develop a privacy culture vis a vis the new technologies and improve compliance with privacy laws and propose ways to fill any possible gap (or effectiveness) within existing legal frameworks. Hence, the preliminary step of forging a clear privacy risk management methodology and its piloting are instrumental to inform further legal discussions on clearly delineating the roles of model providers, deployers and end-users to reduce uncertainties around data processing responsibilities, preventing legal ambiguities that might otherwise hinder effective risk management.

Next Steps

To move this initiative forward, several immediate actions are needed.

First, the report should focus on establishing a common taxonomy (e.g., defining personal data in the context of LLMs), identifying both known and emerging privacy threats within the LLM ecosystem, and refining human rights centered and/or privacy impact assessment methodologies. This report will also serve as the foundation for further developing a structured privacy risk assessment framework based on the entire AI lifecycle, incorporating best practices from European and international standardization, Code of Practices, and existing AI governance models.

Engagement with stakeholders should be prioritized to pilot and validate the feasibility and practicality of the proposed risk management approaches. Regulators and organizations currently developing and deploying LLMs will provide invaluable feedback on the challenges and limitations they face in implementing privacy safeguards. This collaborative effort will ensure that the report remains future-proof, relevant and adaptable to evolving technological landscapes.

A guidance document as part of the report containing a privacy risk assessment framework will then be developed and structured around practical implementation steps for stake holders. The document will take into account all considerations coming from businesses, policymakers, and AI practitioners, providing them with proposals and tools to address privacy risks.

Simultaneously, the work carried out by the Committee of Convention 108 could inform other CoE committees, international bodies to align this initiative with existing AI governance efforts and industrial standards. The Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law serves as a pivotal foundation for global cooperation in this domain, therefore it is proposed to develop an integrated approach within the framework of this initiative and possibly during the subsequent normative work on Artificial Intelligence governance and Data Protection with potentially the joint participation of the two conventional committees. Specifically, the implementation of the provisions of Convention 108+ for this new technology, the adherence to Chapter III Article 11 of the Framework Convention, which mandates the protection of individuals' privacy and personal data, and Chapter IV Article 16, which emphasizes the necessity of robust risk and impact assessments for AI systems, are crucial. By providing guidance for that these articles are consistently applied, the two committees could lay down the ground for a unified approach to the issues related to the use of LLMs that upholds human rights, the rule of law and democracy and fosters international collaboration.

Conclusion

As LLMs continue to shape the future of user-facing AI applications, ensuring robust privacy protections must remain a top priority. The development of a comprehensive guidance on the management of privacy and data protection risks based on Convention 108+ will provide data controllers and regulatory authorities with the necessary tools to identify, assess, and mitigate those risks while promoting compliance with privacy and data protection principles and standards. The committee’s role and previous endeavors in providing guidelines on the interpretation of the provisions of Convention 108+ when using new technologies were key in fostering regulatory clarity, international cooperation, and research-driven policymaking. The expert report will gather all scientific evidence and arguments to serve as a basis for a future-proof and comprehensive normative document that will promote a proactive and collaborative approach, and that AI innovation progresses in a manner that upholds privacy, data protection principles and rules, including transparency, and accountability.