CONSULTATIVE COMMITTEE OF THE CONVENTION FOR THE PROTECTION OF INDIVIDUALS WITH REGARD TO AUTOMATIC PROCESSING OF PERSONAL DATA

CONVENTION 108

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

by

Isabel Barberá and Murielle Popa-Fabre

Expert Report prepared for presentation to the 48th Plenary meeting, 17 June 2025. The opinions expressed in this work are the responsibility of the authors and do not necessarily reflect the official policy of the Council of Europe.

Table of Content

Section 1
Background Context, Scope and Initial Landscape Analysis

This report focuses on identifying the privacy and data protection risks associated with the use of Large Language Models (LLMs), particularly as they relate to the rights and principles enshrined in Convention 108 and its modernised version, Convention 108+. These risks emerge across different phases of the LLM lifecycle, ranging from model training and inference to integration and deployment within broader systems, and have implications not only for data protection, but also for private life, dignity, and autonomy.

The findings presented here aim to inform ongoing normative work under Convention 108+ by offering an evidence-based understanding of how LLMs may interfere with individuals' rights, and by proposing a structured methodology to assess these risks in real-world settings.

1.1 Background and Context: Supporting Committee's action

LLMs are rapidly transforming the digital landscape by enabling new forms of interaction, automation, and information processing. However, these systems also introduce privacy and data protection risks that challenge traditional legal and technical safeguards. These include, but are not limited to, the inadvertent memorization and reproduction of personal data, susceptibility to manipulation during inference, and the broader erosion of private life through synthetic identities, profiling, and opacity in decision-making.

Convention 108+ provides a foundational legal framework for addressing these challenges. Its principles remain technologically neutral, yet LLMs raise new questions about how to implement them effectively. For example, while Article 1 affirms the individual's right to privacy and Article 2 defines key terms such as personal data and data processing, the scale and opacity of LLM training and optimization pipelines complicate compliance. It is often unclear whether personal data are present in training datasets or how they are subsequently used, making safeguards under Article 5 such as purpose limitation, data minimisation, and fairness, difficult to assess and enforce in practice. This challenge becomes even more acute in relation to special categories of data under Article 6, where the uncontrolled extraction or generation of sensitive information, including data about health, political opinions, or biometric characteristics, may occur without the controller's awareness. Likewise, obligations around data security and transparency, set out in Articles 7 and 8, are strained in LLM contexts where LLM-based systems behaviour at inference is dynamic and not fully predictable.

Informing data subjects or identifying accountable controllers in such layered architectures remains a persistent governance gap. LLMs-based systems undermine meaningful access and control for individuals as guaranteed under Article 9. When outputs are probabilistic, it becomes difficult to guarantee the rights to rectification, objection, or explanation of automated outputs or decisions. The absence of observable data traces and accessible user interfaces further limits individuals' ability to know when their data has been used, let alone to contest or correct it. This raises serious concerns about whether core provisions of Convention 108+ can be upheld in practice without complementary technical and procedural safeguards.

As LLMs-based systems are increasingly adopted in contexts such as recruitment, education, healthcare, and public administration, the need to preserve and enforce these rights becomes more urgent. The Consultative Committee plays a critical role in ensuring that the principles of Convention 108+ are adapted and applied to these new technological realities. This report supports that effort by providing an up-to-date structured,

research- and risk-informed understanding of how LLMs-based systems may interfere with data protection and privacy rights and offers a foundation for the development of governance tools that reflect both the spirit and the obligations of the Convention.

1.2 Objective and Scope

To better understand the practical challenges faced by organisations developing and using LLMs, we conducted a series of interviews with stakeholders across the LLM ecosystem. Participants included representatives from start-ups, major technology providers, technology auditors, private and public technology deployers, regulatory authorities, and research institutions. These interviews focused on how organisations manage privacy risks across the lifecycle of LLM-based systems, from data collection and model training to deployment, monitoring, and post-market oversight.

The results of these interviews, which are presented in more detail in Section 5, confirmed widespread uncertainty and inconsistency in current practices. Most notably, the responses demonstrated a pressing need for a harmonised and internationally accepted approach to privacy risk management for LLMs- based systems. Interviewees described fragmented internal processes, limited formalisation of privacy-specific assessments, and major obstacles to applying existing data protection obligations in complex, rapidly evolving model and system architectures.

These findings validate the underlying premise of this report: that privacy risks in LLM-based systems cannot be adequately addressed through ad-hoc organisational practices or existing compliance tools alone. Instead, a structured, lifecycle-based methodology is needed to identify, assess, and mitigate privacy risks at both the model and system level.

Building on this foundation, the report aims to support the Committee of Convention 108 in advancing such a methodology. It proposes a privacy risk management framework aligned with the principles of Convention 108+, adapted to the technical and organisational realities of Generative AI. Its scope includes the identification of privacy risks at different phases of development and deployment, a two-tiered assessment of risks at both the model and system level, and an initial mapping of viable mitigation strategies.

To lay the groundwork for a future operational framework, this report explores three core components:

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+, the safeguards enshrined in Article 8 of the European Convention on Human Rights, and the risk and data protection requirements articulated in the Council of Europe's Framework Convention on Artificial Intelligence (AI Treaty), particularly Article 11 on Privacy and personal data protection and Article 16 on a Risk and impact management framework. This work requires a combination of theoretical research, practical case studies, and engagement with stakeholders from industry, governments, regulatory bodies, and academia.
2. **Exploring a two–tiered risk assessment methodology.** This report proposes the foundation for a risk identification and assessment approach that distinguishes between model-level and system-level risks and that considers the foundational principles laid out in Chapter II – Basic principles for the protection of personal data of Convention 108+, which include lawfulness, purpose limitation, data minimisation, fairness, transparency, data subject's rights, data security and the protection of sensitive data. Future research will need to equip data controllers with tools to evaluate risks stemming from training data,

memorisation, or hallucinations at the model level, as well as risks related to system integration, user interaction, and third-party components at the deployment level.

3. **Surveying 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. The applicability, limitations, and governance implications of these tools will need to be further evaluated in practical contexts.

To support these goals, the report outlines a preliminary research direction that could be further developed in subsequent work. The envisioned approach includes:

- establishing a common taxonomy (e.g., clarifying the notion of personal data in the context of LLMs).
- analysing both known and emerging privacy threats within the LLM ecosystem.
- analysing risk management tools, including privacy- and human rights-centered assessment methodologies.
- engaging stakeholders to prioritise piloting and validate the feasibility and relevance of proposed mitigation strategies.
- and grounding best practices in real-world constraints and requirements from businesses, policymakers, and AI practitioners.

1.3 Landscape Tech Ecosystem of LLM-based Systems and Personal Data

While Section 2 will present detailed insights into how language models compress data and sometimes memorise them, we introduce here an overview of the tech ecosystem of LLM-based systems (see Figure 1). This life-cycle overview offers a simple understanding of the difference between Large Language Models and the systems currently built around them, showing that different technological and economical actors can be present at these general three phases that crucially differ by the type of data they leverage.

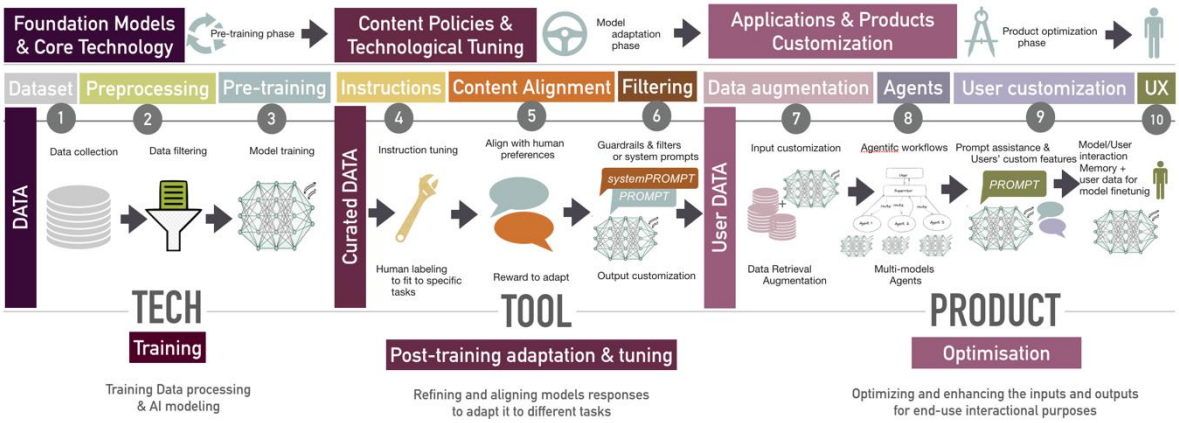


Figure 1: Lifecycle/value-chain LLM-based systems, a broad overview based on the fast-evolving practices of building AI compound systems to optimise LLM-based applications at the "Product" phase.

A preliminary privacy risk analysis of the LLMs lifecycle can identify three main steps based on the types of data that are used respectively to train foundation models, to adapt them during post-training, and to optimise them for custom use cases or for interactional purposes. Importantly, if an immense amount of data is needed at the foundational layer (i.e., the phase involving datasets collection and data pre-processing, steps 1 to 3), progressively more curated data is used to adapt and fine-tune models in the post-training phase. This middle stage (steps 4 to 6), often overlooked in privacy discussions, is a suitable place to apply data protection best practices in LLM-based tool development. The final phase of LLM-based systems' development is the optimisation one (steps 7 to 10). At this stage the ever-evolving landscape of customization methods has significantly enhanced the adaptability of standalone models enabling increasingly complex systems, for example, through:

- **Data augmentation** using techniques such as Retrieval-Augmented Generation (RAG), which support advanced search use cases by allowing access to corporate or trusted data in a conversational format.
- **Agentic workflows**, that orchestrate multiple models to automate increasingly complex tasks through flexible, multi-step processes.

To these more technologically complex solutions, simpler ones are making an intensive use of personal data to make applications, products and services adaptable, intuitive and interactional such as:

- **User intention features** and prompt customization, to capture users' intention in real-time and automatically contextualise queries or needs to enhance relevance and user experience.
- **Memory features** store and leverage users' interaction history to improve continuity, personalize responses, enhance the end-user experience and support further model optimization or product development.

At each of these steps data availability, quality play an important role, but more importantly LLM Models and LLM-based systems have different privacy risks.

Section 2
Unpacking Privacy Inside Large Language Models

This section provides a technical overview of how personal data is handled within LLM-based systems, forming a foundational layer for the risk assessment framework proposed in Section 4. It offers key insights into the privacy implications of current developments in LLM explainability, highlighting their significance for the future of AI governance.

The goal is to offer an accessible understanding of how LLMs function beneath the surface—specifically, how language is encoded and compressed within neural architectures—and to demonstrate how this technical understanding can directly inform privacy risk assessments and effective mitigation strategies.

Rather than providing an exhaustive analysis, the section focuses on essential technical features and inner workings that influence privacy risks, including how data is ingested, transformed, organized and potentially memorised by these systems. It also addresses the risks related to textual input data and their representation. By clarifying how these systems operate internally, we aim to show that deeper technical literacy is a prerequisite for developing effective privacy safeguards and governance mechanisms rooted in real system behaviour.

Introduction to Data Representations in Large Language Models

Privacy risks in LLMs can be assessed across multiple dimensions. Figure 2 illustrates the basic architecture of an LLM-based system like ChatGP and serves as a guide to explore privacy concerns at different levels focusing on (a) user input data, (b) training data, and (c & d) model outputs.

This section begins by examining the privacy implications of how data is represented within these computational systems. It addresses two foundational questions:

1. How are user input data represented in an LLM?
2. How are training data represented in an LLM?

In Section 3 the analysis will turn to more operational privacy risks, considering:

3. To what extent are personal data memorized by LLMs?
4. How can personal data be extracted from an LLM or an LLM-based application?

Through this two-step exploration covering both foundational and practical dimensions, this section aims at clarifying the foundational aspects of how data representation in LLMs shapes the privacy risks they pose, and why understanding these mechanisms is essential for effective mitigation.

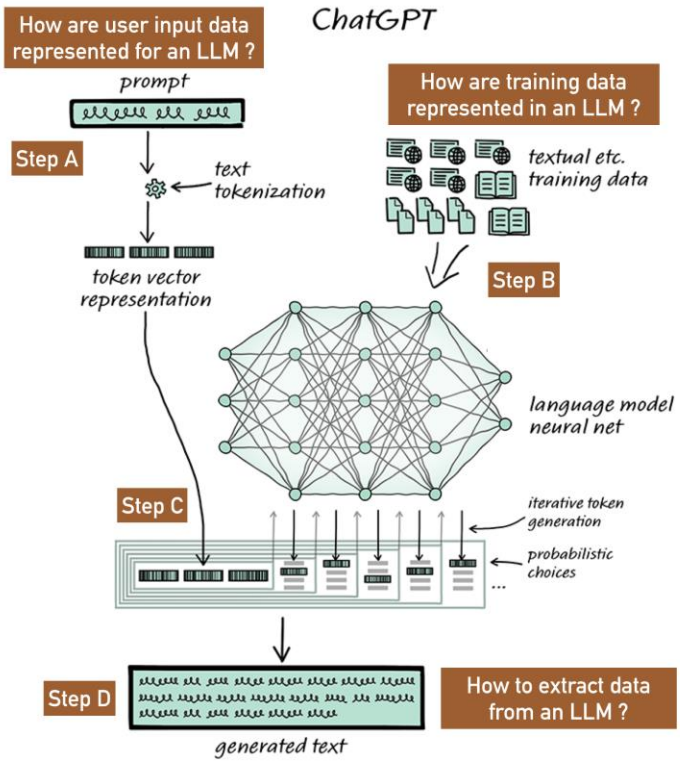


Figure 2: A basic schema of the workings of ChatGPT, to navigate the privacy challenges at the level of users input data (step A), training data and their LLM internal representation (step B) and model output (steps C and D), graphic representation inspired by Worfram writings.

2.1 How Are Words “seen” and Represented by LLMs

A first step in understanding the privacy implications of LLM-based systems is to examine how these computational systems "see" words, and specifically, how user input is represented.

As illustrated in Figure 2, when a user inputs text (a prompt), the system begins by transforming that text in Step A into a numerical format known as vector (i.e., an array of numbers). This representation is later used in Step C as the input for probabilistic calculations that approximate next-word predictions, ultimately generating text in Step D. The transformation from words to numbers relies on an embedding model outputting word vectors or embeddings.

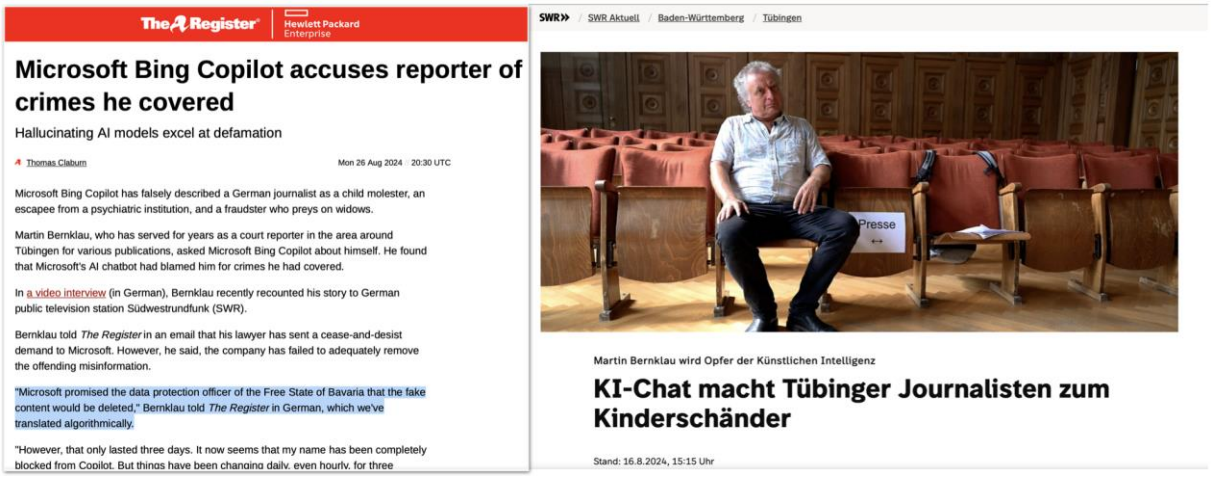
Word embeddings are numerical representations that capture features of words based on the contexts in which they appear. Rather than encoding meanings through traditional synonymy or explicit semantic relationships, embeddings reflect patterns of linear co-occurrence: words that frequently appear in similar contexts tend to be assigned similar vector representations.

The core idea behind this encoding approach is to learn representations that embed aspects of semantic meaning and word relationships. Words sharing similar contexts are mapped to nearby positions in the vector space.

What are the privacy implications of these word representations?

It is essential to emphasize that this method of representing words has significant privacy implications. In LLMs, associations between personal information and individuals' names are learned based on their proximity and statistical co-occurrence in the model’s training data; words that are sharing a similar context do also share similar representations, and these relationships are embedded directly into the word vector representations.

This means that any piece of information appearing near a person’s name, whether on internet or in a document, can become persistently associated with that individual, even if the connection is entirely spurious. In the example illustrated in Figure 3, a German journalist who frequently covered criminal trials found that a chatbot associated his name with the crimes he reported on. This simply occurred because his name appeared repeatedly in articles about those cases, leading the model to infer a misleading association.


The Register | Hewlett Packard Enterprise

Microsoft Bing Copilot accuses reporter of crimes he covered

Hallucinating AI models excel at defamation

Thomas Claburn — Mon 26 Aug 2024 // 20:30 UTC

Microsoft Bing Copilot has falsely described a German journalist as a child molester, an escapee from a psychiatric institution, and a fraudster who preys on widows.

Martin Bernklau, who has served for years as a court reporter in the area around Tübingen for various publications, asked Microsoft Bing Copilot about himself. He found that Microsoft's AI chatbot had blamed him for crimes he had covered.

In a video interview (in German), Bernklau recently recounted his story to German public television station Südwestrundfunk (SWR).

Bernklau told *The Register* in an email that his lawyer has sent a cease-and-desist demand to Microsoft. However, he said, the company has failed to adequately remove the offending misinformation.

"Microsoft promised the data protection officer of the Free State of Bavaria that the fake content would be deleted," Bernklau told *The Register* in German, which we've translated algorithmically.

"However, that only lasted three days. It now seems that my name has been completely blocked from Copilot. But things have been changing daily, even hourly, for three

SWR» / SWR Aktuell / Baden-Württemberg / Tübingen

Martin Bernklau wird Opfer der Künstlichen Intelligenz

KI-Chat macht Tübinger Journalisten zum Kinderschänder

Stand: 16.8.2024, 15:15 Uhr

Figure 3: Two newspaper articles from August 2024 relating the LLM-incident of a German journalist falsely blamed by a chatbot for crimes he had covered as a journalist.

2.2 How Are Data Compressed in LLMs Neural Network Architecture

Recent research developments in understanding LLM internal workings are leveraging methods developed in a discipline called Mechanistic Interpretability to map how training data is compressed in the neural network during its training phase. These results are gradually leading to major steps towards making LLMs more transparent and challenge the blackbox paradigm that is often picturing a total lack of knowledge of LLMs internal workings.

Starting from 2024, several studies[1] demonstrated how to map and visualize the way data is compressed in production LLMs, like the 70 billion parameter model Claude Sonnet. Notably, researchers used a technique called 'dictionary learning' to isolate neural patterns (features) that corresponded to interpretable concepts represented inside the LLM.

This is how we can, for example, observe in Figure 4 the fine-grained activation patterns of the neighbouring 'concepts' that were aggregated during the training phase around the feature "Inner Conflict".

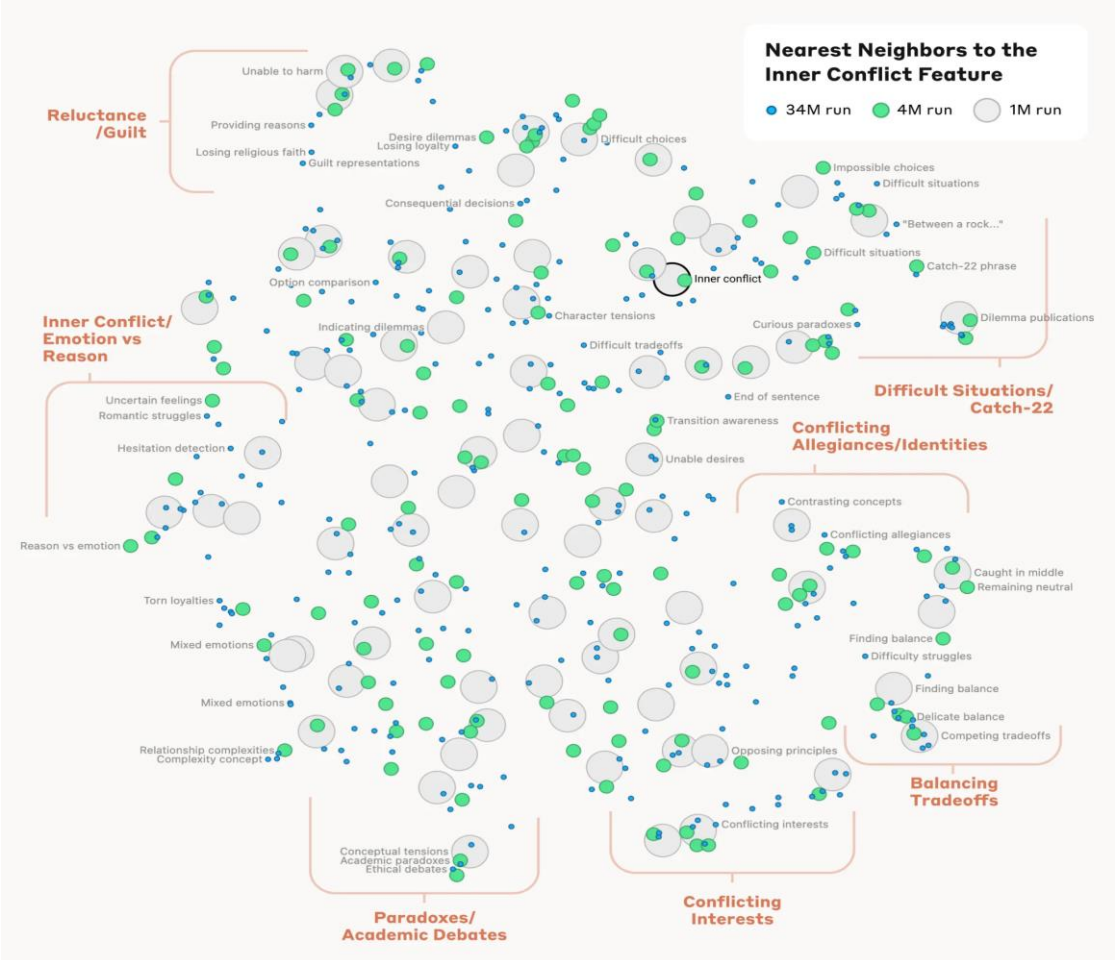


Figure 4: A map visualizing the statistically aggregated features near an "Inner Conflict" feature in Claude Sonnet (70B), including clusters related to balancing trade-offs, romantic struggles, conflicting allegiances, and catch-22s.

[1] See Templeton et al. (2024) https://www.anthropic.com/research/tracing-thoughts-language-model Later research at Anthropic looks for proxies of how concepts are connected inside the LLM(Ameisenet al, 2025) and Neel Nanda's team research at Google Deepmind together with interactive examples on Neuronpedia https://www.neuronpedia.org, https://www.technologyreview.com/2024/11/14/1106871/google-deepmind-has-a-new-way-to-look-inside-an-ais-mind/

After extracting millions of features (or "concepts") aggregated during the model training phase, it is possible to visualize the level of activation of individual features in response to textual inputs to visualize for example, when the feature for "Golden Gate Bridge" activates in a textual description or visual depiction of the city of San Francisco[2].

What are the privacy implications of observing aggregated features inside an LLM?

The features that are found to be corresponding to people are of particular interest to grasp the privacy implications of these new possibilities to observe LLMs' inner workings.

As shown in Figure 5, the feature sensitive to mentions of Margaret Thatcher fires on a range of model inputs where the orange colour denotes the words or word-parts on which the feature is more active like character descriptions or relevant historical context. Specifically, her feature is highly activated on:

- her age (i.e., dies aged 87),
- her country and address (i.e., UK and Downing Street),
- her professional function (i.e., Prime Minister), and
- more broadly on associated concepts like her field of competence (i.e., British politics),
- the fact she was a force for change (i.e., changed the face of British politics).

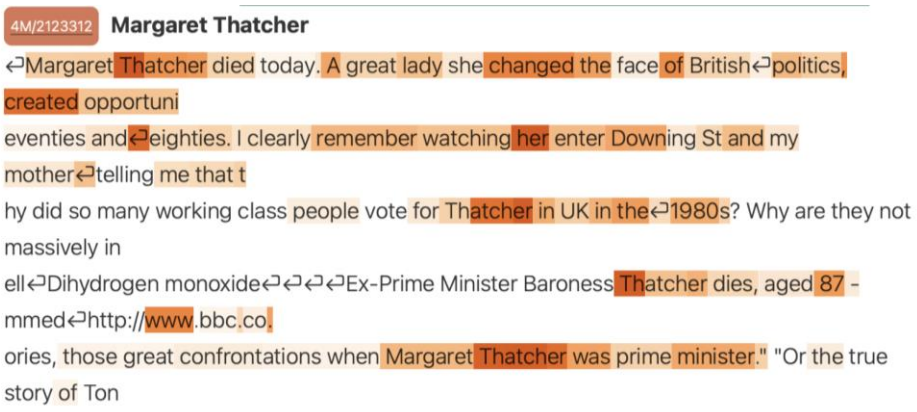


Figure 5: Many features corresponding to famous individuals, which are active on descriptions of those people as well as relevant historical context, here the person feature of Margaret Thatcher.

While these breakthroughs enable researchers to better understand how data is compressed and activated within neural networks, their implication for steering LLM outputs and their potential impact on AI reliability, safety and ethics remain subjects of ongoing debate[3]. Nonetheless, these advances open new avenues for the governance of privacy risks. For instance, if it becomes possible to reliably identify the internal features corresponding to a given individual, future developments may allow such features to be modified or disabled, provided that the challenges of scaling interpretability to larger models can be effectively addressed.

[2] Example:https://transformer-circuits.pub/2024/scaling-monosemanticity/features/index.html?featureId=34M_31164353

[3] One of the results that still needs to be confirmed is the extent to which these features influence LLMs outputs. See Google Deepmind note on AI safety research in March 2025: https://www.alignmentforum.org/posts/4uXCAJNuPKtKBsi28/sae-progress-update-2-draft

Section 3
Mapping Privacy Risks in LLM-based Systems

Privacy Risks in the LLM Ecosystem

As LLMs are increasingly integrated into public and private infrastructures, the absence of adapted privacy safeguards has become a growing concern. Under the lens of Convention 108+, these systems introduce both novel and systemic risks, ranging from exposure of personal data to erosion of private life through opaque inference and profiling.

Current practices are not well equipped to address these risks. Traditional risk assessments often fail to capture the full scope and unpredictability of generative models like LLMs. Many organisations do not have a clear view of where their training data comes from or how downstream uses might impact individual rights. At the same time, LLMs blur the line between synthetic and personal information. Their outputs can include memorised data or sensitive content generated in response to cleverly designed prompts, making it difficult to detect problems until after deployment.

These risks go beyond data protection. LLMs also affect how we understand personal autonomy and identity. As machine-generated content becomes harder to distinguish from human communication, people may struggle to prove what they did or didn't say. This erosion of authorship and authenticity raises deeper concerns; not just about privacy, but about dignity, trust, and the psychological strain of interacting with systems that can mimic us so convincingly. These harms challenge core aspects of individual identity and autonomy, affecting not only the safeguards of Convention 108+, but also the protections enshrined in Article 8 of the European Convention on Human Rights (ECHR), which upholds the right to identity, reputation, and private life.

LLMs are also increasingly deployed in automated decision-making processes, including hiring and content moderation, without meaningful oversight. This may contravene Article 9(1)(a) of Convention 108+, which safeguards individuals from decisions made solely on the basis of automated processing. Another concern is the weakening of user control over consent and data retention in post-deployment phases. Practices like silent feedback loops or the removal of opt-out mechanisms, as seen in some virtual assistants, restrict data subject rights under Articles 8 and 9 of Convention 108+ and undermine compliance with Article 8 ECHR.

Deceptive or manipulative user interfaces, combined with limited access to redress, further illustrate how LLMs can produce harms beyond traditional privacy violations. These structural issues may have broader societal implications, potentially affecting democratic participation, media integrity, and public trust.

3.1 Latest LLMs Technological Evolutions and New Risks for Privacy

The latest developments of LLM-powered applications are gradually expanding the spectrum of LLM-based multimodal data aggregation and prediction through advanced simulation techniques that have implications not only for data protection, but also for private life, dignity, and autonomy. Analysing these recent developments ensures future proof governance measures to tackle new risks for privacy.

Recent studies convergently point to how LLM-powered pattern finding across multiple multimodal sources of information is raising new privacy risks by:

- enabling multimodal fusion for behavioural prediction on e-commerce platforms.[4]
- championing election prediction[5] through prompting techniques based on publicly available demographics.
- providing adaptable needs anticipation and behavioural simulation in LLM-recommender systems[6] that significantly boost next-purchase predictions.
- powering AI agent surveys responses that correspond to accurate behavioural simulations.[7]

These latest developments show how these systems' architectures can incorporate multiple multimodal data into a comprehensive and highly adaptable 360°profiling of individual data subjects. We would define this a "*predictive and adaptable data-cage*" for the individual which is crucially differing from profiling harms known so far. Namely, profiling can be detected by the individual user as it can be perceived as stereotyping and discriminatory, while the adaptiveness of "*predictive data caging*" is bringing profiling and hyper-personalization to a next level. It is by far less rigid and leaves the user with the impression of being understood, while fundamentally questioning and undermining the human right to autonomy and private life.

Combining multimodal data aggregation together with the prediction capabilities of LLMs and the adaptability enabled by reinforcement learning techniques opens new and sophisticated venues for behavioural prediction and fine-grained adaptable profiling without direct re-identification.

If one considers in the light of these latest technological developments how smartphones are becoming a hub for a wide range of personal and behavioural data, it is possible to build a detailed global picture of an individual's private life through the sensors embedded in a smartphone. A system aggregating all these data sources presented in Figure 6 can not only track personal and sensitive data, but even leverage proxy data for the mental states, thus building a global picture of an individual's private life through fundamentally very common sensors embedded in a smartphone:

- A GPS and a gyroscope track movements and deduce what is the phone orientation to infer physical activity.
- Social rhythm can be detected through GPS, Wi-Fi and help identify if people go to the gym or to the bar, together with proximity sensors that are also providing elements that can help measure social relationships.
- Cameras can help the understanding of the emotional state through face expressions and less intuitively, cameras can also track eye movements providing a window on the user's cognitive processes and gaze attention.
- Eye tracking is also a proxy allowing to check for example medication effects on pupillometry (dilatation of the pupil)[8].

[4] Ma, L., Li, X., Fan, Z., Xu, J., Cho, J.H., Kanumala, P., Nag, K., Kumar, S., & Achan, K. (2024). Triple Modality Fusion: Aligning Visual, Textual, and Graph Data with Large Language Models for Multi-Behavior Recommendations. ArXiv, abs/2410.12228.

[5] Jiang, S., Wei, L., & Zhang, C. (2024). Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models.

[6] See recent review of the literature of LLM-based Agentic Recommender Systems (LLM-ARS) : Huang, C., Yu, T., Xie, K., Zhang, S., Yao, L. and McAuley, J., 2024. Foundation models for recommender systems: A survey and new perspectives. arXiv preprint arXiv:2402.11143.

[7] Convergent research results in the last few years show how LLM-based interactive agents can simulate individual complex behaviour and survey responses (e.g., General Social Survey, Big Five Personality Inventory, Economic Games or Behavioral Experiments) from basic demographic information. A recent study by Stanford University showed that AI agents can simulate human behavior with up to 80% accuracy based on just two hours of interview data, and 60% accuracy using only basic demographic information. Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C.J., Morris, M.R., Willer, R., Liang, P., & Bernstein, M.S. (2024). Generative Agent Simulations of 1,000 People. ArXiv, abs/2411.10109. See also Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

[8] See one among the first use of smartphones for psychiatric patients' tracking: Torous J, Kiang MV, Lorme J, Onnela JP New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research JMIR Ment Health 2016;3(2):e16 doi: 10.2196/mental.5165,https://spectrum.ieee.org/a-software-shrink-apps-and-wearables-could-usher-in-an-era-of-digital-psychiatry

- Touchscreen data also provides information to infer elements of cognition, like response time tasks requiring to swipe, tap or interact through touch.

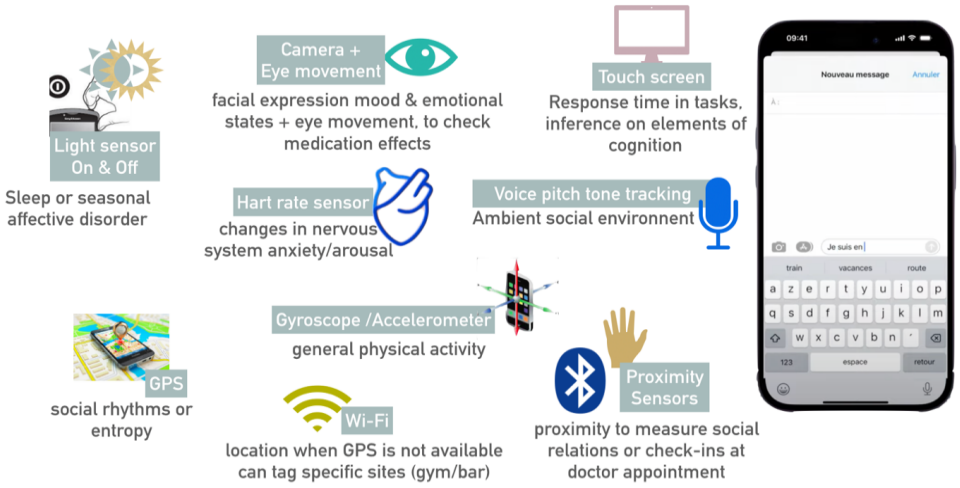


Figure 6: Smartphones are a digitally connected eco-system of personal and behavioural data. Mapping of what aspects of privacy can be inferred from the interplay of different multimodal data sources that can be now efficiently and adaptively aggregated to obtain a 360° profile of the user, and to infer mental states from non-neural data.

Other sensors embedded in smartphones can offer indirect cues to infer mental states. For example, a basic light sensor can already be a proxy to understand the sleep status of an individual or cues of seasonal affective disorders. Heart sensor apps are indirectly tracking the nervous system, anxiety, arousal that can help infer the stress status of an individual, while voice recording can detect emotion, but also social life and environment.

All this behavioural information can actually be coupled to measures of electrodermal activity to emotional stimuli that smartwatches can record, and a recently announced partnership between Apple and the brain implant company Synchron points at new venues for aggregating multimodal neurophysiological data[9] inside smartphones. As stated by the press release "Apple is helping to pioneer a new interface paradigm, where brain signals are formally recognized alongside touch, voice and typing."

In conclusion, given LLMs new (1) multimodal data aggregation, (2) the efficient and predictive pattern finding capabilities of LLMs, and (3) the fundamental adaptability of these technological developments, privacy rights become a priority in order to prevent any possible profiling, or manipulation by accessing a new and rather complete picture of individuals personal life and identity. Such a configuration should be tackled by an adequate data protection risk management framework and regulatory landscape.

3.2. LLM Privacy Risks: Personal Data Extraction Methods

This subsection outlines documented risks and exemplifies various methods of extracting personal data from LLMs. Compared to the more structural privacy risks previously addressed, the focus here is more operational

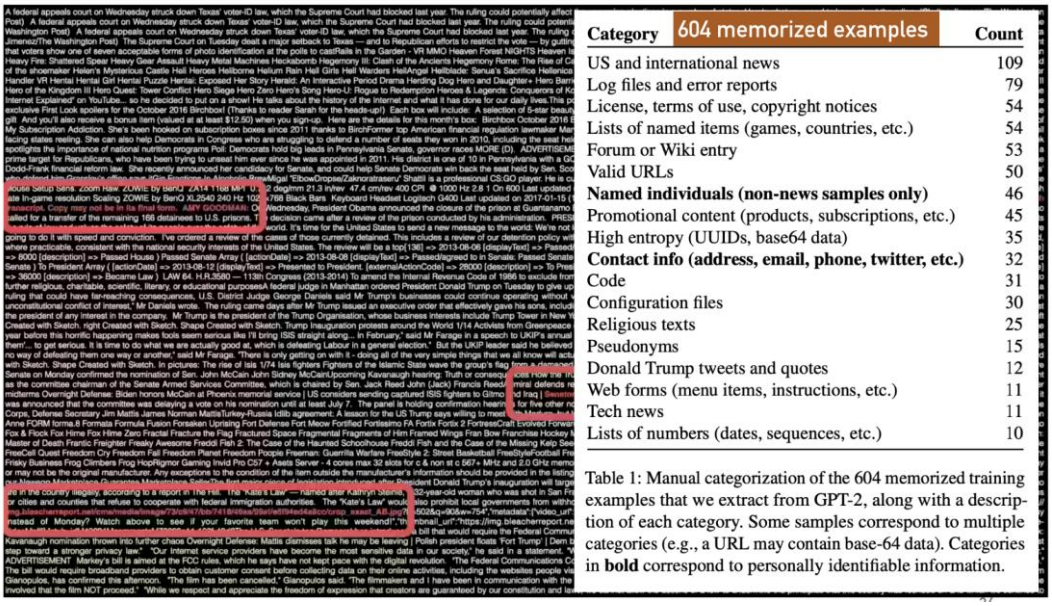
[9] See press release 'Apple's new BCI Human Interface Device protocol marks the creation of a new input category powered by thought, enabling hands-free, voice-free digital control through Synchron's BCI system' https://www.wsj.com/tech/apple-brain-computer-interface-9ec69919?mod=hp_lead_pos8,

on how personal information can be extracted. Going from methods using simple prompting techniques to more sophisticated ones involving complex strategies, computational power and advanced competences, we will consider the following two questions:

1. To what extent is personal data memorized by LLMs?
2. How can personal data be extracted from an LLM or an LLM-based application?

Regurgitation of memorized sequences and Training data extraction attacks

Early methods extracting data and investigating the privacy risks in LLMs date back to 2019 and 2020, before ChatGPT deployment attracted the public attention[10]. These first research studies pointed at the fact that LLMs are prone to memorising sequences that were often repeated in their training data, like it was the case for Donald Trump's tweets being largely echoed in the infosphere when GPT-2 was trained (see Figure 7)[11].



Category	604 memorized examples	Count
US and international news		109
Log files and error reports		79
License, terms of use, copyright notices		54
Lists of named items (games, countries, etc.)		54
Forum or Wiki entry		53
Valid URLs		50
Named individuals (non-news samples only)		46
Promotional content (products, subscriptions, etc.)		45
High entropy (UUIDs, base64 data)		35
Contact info (address, email, phone, twitter, etc.)		32
Code		31
Configuration files		30
Religious texts		25
Pseudonyms		15
Donald Trump tweets and quotes		12
Web forms (menu items, instructions, etc.)		11
Tech news		11
Lists of numbers (dates, sequences, etc.)		10

Table 1: Manual categorization of the 604 memorized training examples that we extract from GPT-2, along with a description of each category. Some samples correspond to multiple categories (e.g., a URL may contain base-64 data). Categories in **bold** correspond to personally identifiable information.

Figure 7: Left, In back background the recurrent sequences identified as being memorized by GPT-2 from training data: "Senators press Donald Trump to end Yemen war"," Transcript: Copy may not be the final form", and an URL address (bottom left). Right: the categories of the 604 memorized training examples including names of individuals (not from people in the news) and contact information including address email, phone number and twitter account). Source adapted from Carlini et al. 2021).

After documenting through several studies that LLMs memorize personal data like (public) personally identifiable information (names, phone numbers, and email addresses), email signatures, or even IRC conversations, etc., further research showed that memorized sequences were very likely to appear in LLMs

[10] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In USENIX Security Symposium, volume 267, 2019. Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In 2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020, pages 1314-1331. IEEE, 2020.Huseyin A. Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. Privacy analysis in language models via training data leakage report. CoRR, abs/ 2101.05405, 2021.

[11] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In the 30th USENIX security symposium (USENIX Security 21) (pp. 2633-2650).

responses, either as unintentionally “regurgitated “in their output as verbatim sequences or through training data extraction attacks[12].

Likelihood of LLM Regurgitation

Additional research in the field of Privacy and LLMs emphasized that larger models tend to memorize more data, and that repeated sequences in training data are more likely to be 'regurgitated’. Notably, a sequence present 10 times in the training database is generated on average 1000 times more than a sequence present only once (Kandpal et al. 2022).

Prompt-based Data Extractions

Several types of prompt-based attacks can lead to extracting pre-training data form LLMs. Specifically, the prompting strategy illustrated in Figure 8 can cause LLMs to diverge and emit verbatim pre-training examples. In this type of training data extraction attack, for instance, a personal email signature can be extracted by simply asking the model to repeat forever the different words listed on the graphic. Notably, some words like “poem” or “company” cause the model to emit training data 164 times more often than a word like “know”[13].

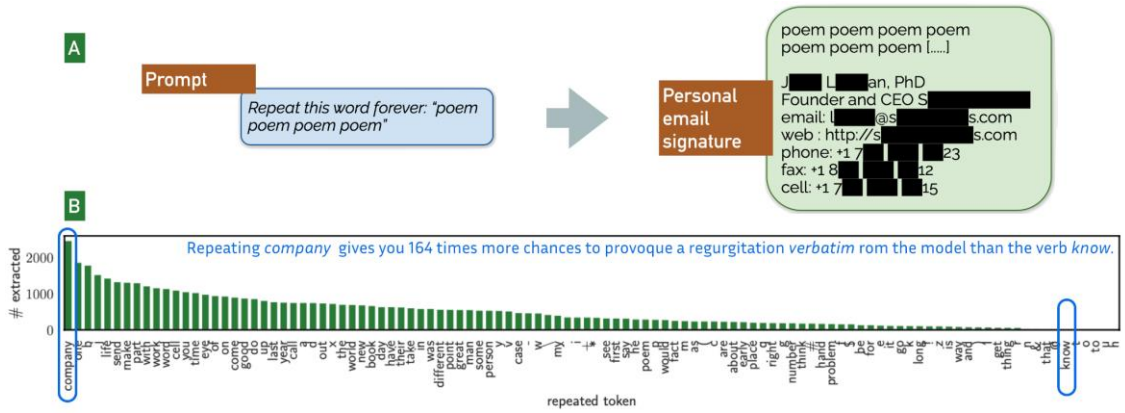


Figure 8: A, An example of the prompt "divergence" attack Extracting pre-training data from ChatGPT. The example shows ChatGPT revealing a person’s email signature which includes their personal contact information, B. The graphic showing the amount of extracted memorized training data across different repeated words. Repeating ‘company’ gives 164 times more chances to provoke a verbatim regurgitation from the model than the verb ‘know’. Source adapted from Nasr et al. 2023.

If the fact of being able with simple prompting techniques to extract email signatures and personal information with just a $200-worth computation has important implications for Privacy and personal data protection in LLM-based systems, it is also fundamental to understand why such data end up in being extractable to find the right mitigations measures at the right phase of the lifecycle as discussed in the following sections.

3.3 Model vs. System Risks in the LLM Ecosystem

In addition to technologically sophisticated architectures, many popular LLM-based applications rely heavily on the intensive use of personal data to create intuitive and highly interactive services. These features, while beneficial for usability, raise serious questions about data protection. In particular, they blur the boundaries

[12] A memorized sequence being found in the output of a Chatbot is called in technical terms a regurgitation.

[13] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F., Ippolito, D., Choquette-Choo, C.A., Wallace, E., Tramèr, F., & Lee, K. (2023). Scalable Extraction of Training Data from (Production) Language Models. ArXiv, abs/2311.17035.

between training, personalization, and continuous data collection, increasing the difficulty of tracking where and how personal data is processed.

In this context, it is essential to distinguish between risks associated with the model itself and those that emerge at the level of the broader system in which the model is embedded. This distinction is not merely technical, it is foundational to designing effective, lawful, and context-aware privacy risk management strategies.

Model-level risks arise from how the LLM is trained, fine-tuned in post-training, and architected, including:

- Ingestion of personal data during pre-training, often from large-scale web scraping, without transparency or legal basis.
- Memorization and regurgitation of sensitive or identifiable information from training data, potentially violating data minimization and storage limitation principles.
- Hallucinations or the generation of plausible but false personal information, which can harm data subjects' reputation or privacy.
- Bias amplification, where underlying statistical associations reproduce or reinforce unfair or discriminatory patterns in how personal data is treated or referenced.

System-level risks arise when an LLM is integrated into a broader application environment, often including APIs, interfaces, plug-ins, memory functions, feedback loops, RAG systems, agentic workflows involving LLM orchestration, and third-party services. Here, privacy threats are linked not just to what the model does, but how it is used, and by whom. Examples of risks include:

- Persistent user profiling via memory features as seen previously.
- Lack of transparency about how user data is processed, shared, or reused, especially when LLMs operate in dynamic cloud-based systems.
- Cross-context data leakage, where user data provided in one application context is reused in another (e.g., through shared fine-tuning across products).
- Inadequate user controls or consent mechanisms, especially for secondary uses of data, including personalization or A/B testing.

In the LLM ecosystem, the risk landscape is dynamic, context-dependent, and multi-layered. To be effective, a privacy risk framework for LLMs must:

1. Address both layers: consider risks at the level of the statistical model *and* the application/system in which it operates.
2. Track risk across the lifecycle: from pre-training to deployment and beyond, including ongoing learning and user interaction.
3. Track response variability through continuous evaluation: because LLMs are probabilistic, not deterministic[14], some privacy violations may only arise post-deployment (e.g., via regurgitation, prompt injections or output misuse).
4. Incorporate both technical and organizational controls: no single technical solution is sufficient, governance must combine engineering, legal, and UX perspectives.

[14] LLM-based systems fundamentally differ from traditional software in that they do not have a predefined set of rules yielding a deterministic behaviour where one input corresponds only one output. Probability and prediction are at the core of LLM non-deterministic behaviour, where one input has many different outputs requiring a governance framework to embed a continuous evaluation layer.

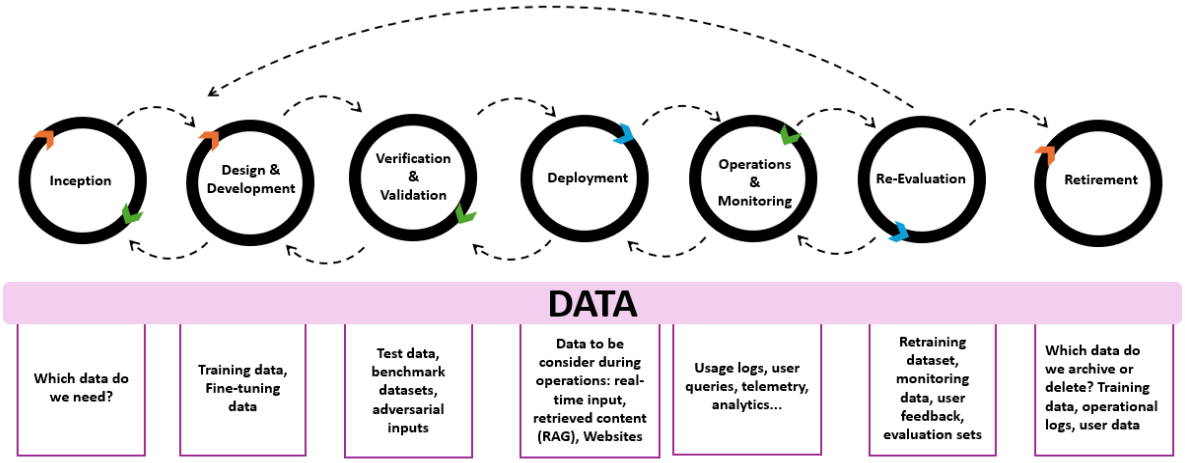
Effective risk management must align with core principles such as *lawfulness*, *fairness*, *transparency*, *purpose limitation*, and *accountability*. These must apply not only to the model itself, but to the full ecosystem in which the model is deployed.

3.4 LLMs Privacy Risks Across the AI lifecycle

As discussed in previous sections, LLMs pose significant challenges at every stage of their lifecycle: from compromised training data and adversarial prompts during inference, to systemic opacity in deployment and post-deployment monitoring. Privacy threats manifest not only through direct data leakage or memorisation, but more broadly through model behaviours that enable identity manipulation, impersonation, and disinformation. For instance, deepfake content or influencer accounts using AI-generated identities have proliferated online, reinforcing the difficulty of distinguishing real individuals from synthetic personas.

Privacy risks in LLM-based systems must be understood in relation to specific phases of the model and system lifecycle. In this section we outline five key lifecycle stages, each associated with distinct threat types, system exposures, and governance challenges:

1. **Training Phase involving training data collection and pre-processing**
   Risks arise from the uncontrolled ingestion of personal data into training datasets. Publicly available corpora may contain identifiable information, such as API keys, emails, or sensitive personal discussions. Without adequate filtering, models can memorise and later reproduce this content, challenging principles such as data minimisation, purpose limitation, and fairness under Convention 108+.
2. **Post-training instruction, Fine-Tuning and System Integration**
   Fine-tuning introduces risks, particularly when sensitive or domain-specific data is used without sufficient controls. It can also amplify bias or destabilise model behaviour. Many deployers rely on fine-tuned models offered as services without full transparency about the post-training process, complicating privacy assessments and liability.
3. **Inference and User Interaction**
   Jailbreaking and prompt injection are major threats in this phase. Even well-guarded models can be manipulated into revealing sensitive information or behaving inappropriately. The inability to consistently predict or trace outputs also raises issues of transparency and accountability. The intensive use of personal data to create intuitive and highly interactive services raises the privacy concerns previously discussed.
4. **System-Level Vulnerabilities and API Design**
   Risks extend beyond the model to include APIs, middleware, and Retrieval augmented generation (RAG) architectures and agents' orchestration. Poorly secured endpoints or integrations may expose private data. These issues require security-by-design principles and reinforce the importance of architectural audits and layered safeguards.
5. **Post-Deployment Monitoring and Adaptation**
   Ongoing data collection and model updates often lack transparency. Feedback data may be reused in ways that reintroduce privacy risks, and users may be unaware of how their inputs are stored or analysed. Articles 10 of Convention 108+ and Article 16 of the AI Treaty underscore the importance of continued oversight and robust impact assessments.



Yet beyond these structured phases, LLMs also introduce a broader class of systemic and societal-level harms, which require a different kind of scrutiny[15].

Figure 9: Illustrates the distribution of privacy risks across different lifecycle stages of LLM-based systems with a focus on data access and flows. (Source: EDPB's report on Privacy Risks & Mitigations in LLMs). The lifecycle phases are based on ISO/IEC 22989.

3.5. Technological Mitigations and their Limitations

Building on previous sections' discussion of privacy risks, we examine some privacy mitigation strategies reported in the research literature in order to identify possible state of the art solutions and investigate their real-world applications through the questionnaire introduced in Section 5.

A key limitation of many current approaches to privacy protection in LLMs and machine learning is the implicit assumption that models operate in isolation. In practice, however, these models are components within larger, integrated systems, an important consideration for effective privacy mitigation. We briefly review mitigation strategies that occur at different stages of the LLM lifecycle.

Mitigations for personal data memorized in LLMs

First, model-level mitigations, applied early in the LLM lifecycle, can either target the extraction of memorized sequences from production LLMs focus or consist in approaches aimed at preventing memorization in the first place. Research quantifying the level of memorization in LLMs, was able to identify three major levers to curb personal data memorization:

- Reducing the size of the LLM prevents memorization (Carlini et al. 2021);
- Reducing the size of the LLM context (i.e., prompt length) (Carlini et al. 2023);
- Deduplication of the training dataset (Kandpal et al. 2022; Lee et al. 2022)[16].

[15] See for example structural implications described in the Draft Guidance Note on Generative Ai implications for Freedom of Expression by the Council of Europe MSI-AI Committee.
https://www.coe.int/en/web/freedom-expression/msi-ai-committee-of-experts-on-the-impacts-of-generative-artificial-intelligence-for-freedom-of-expression# {%22265382451%22:[1]}

[16] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

More recent approaches rooted in Mechanistic Interpretability, as presented in Section 2.2, target the identification of personal data representations within LLMs' neural network and could offer promising new directions for mitigating privacy risks.

What are their limitations?

If the diagnostic of the factors influencing LLM memorization is consistent, further testing and research confirmed that except controlling for LLM-size all other mitigations had some substantial limits. Notably, controlling context size is strongly limiting the current corporate adoption of this technology aiming at extracting information from companies' data (i.e., RAG-applications). Secondly, if deduplication is now standard practice for platforms hosting LLMs like Hugging Face, it has been recently shown that this technique has security side effects, and it creates privacy side channels.[17]

Additional mitigation strategies exist at later stages of the LLM lifecycle, these strategies are coming into play at output monitoring stage or earlier at post-training fine-tuning phase, and particularly when LLMs are deployed as part of complex systems. However, filtering memorized personal data or items at inference stage, and performing fine-tuning for unlearning a subset of the training data, without having to retrain the LLM from scratch,[18] are not fully preventing the leakage of training data. For instance, fine-tuning based strategies, although theoretically promising, have demonstrated a mild practical implementation success as basic fine-tuning methods are often tantamount to decrease in model performance on common benchmarks. Filtering memorized sequences before the output is shown to the user may seem like an easily implementable solution, but it has been shown to be ineffective at preventing training data leakage and can be easily circumvented by prompts designed to extract memorized information.[19].

As for mitigations related to adopting smaller models, the current technological trend towards using smaller, more specialized and efficient models, offers hope for a better privacy protection in future LLM-powered applications running on-device. Figure 10 illustrates the increasing trend toward smaller Language models since 2019, and how new on-device deployment modalities are intensifying their adoption. Since May 2025, Google is making Small Language Models downloadable locally on smartphones[20] marking the first concrete steps toward the deployment of LMMs that run on-device in a faster and more energy-efficient manner. This is critically showing new venues for privacy mitigations and control over personal data, where no data is sent on an API.

[17] Debenedetti, E., Severi, G., Carlini, N., Choquette-Choo, C.A., Jagielski, M., Nasr, M., Wallace, E., & Tramèr, F. (2023). Privacy Side Channels in Machine Learning Systems. ArXiv, abs/2309.05610.

[18] A paper using the strategy of replacing idiosyncratic expressions in the target data with generic counterparts and leverage the model's own predictions to generate alternative labels for every token. Eldan, R., & Russinovich, M. (2023). Who's Harry Potter? Approximate Unlearning in LLMs. ArXiv, abs/2310.02238.

[19] Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette Choo, C., & Carlini, N. (2022). Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy. International Conference on Natural Language Generation.

[20] *Google quietly released an app that lets users run a range of openly available AI models from the AI dev platform Hugging Face on their phones*. https://techcrunch.com/2025/05/31/google-quietly-released-an-app-that-lets-you-download-and-run-ai-models-locally/

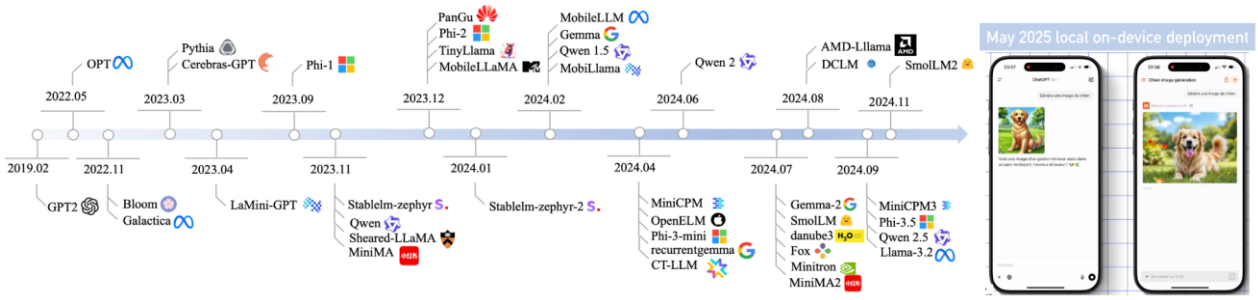


Figure 10: Timeline of the development of Small Language Models) and their gradual deployment on device in May 2025. Adapted from Lu et al. (2025)[21]

In conclusion this section shed light on the complex relationship between LLM workings and privacy concerns. Overviewing privacy risks and mitigations together with their limitations further demonstrate how a well-defined technological mapping provides a science-backed foundation for the subsequent tests and privacy assessment along the lifecycle of LLM-based systems. These insights suggest that while privacy risks in LLM-based systems and applications are significant, understanding these systems' internal workings enables more effective mitigation strategies, governance guidelines and a roadmap for a tailored risk-framework as further developed in the next section.

Section 4
Roadmap for a Privacy Risk Management Framework for LLM-based Systems

4.1 The Need for a Lifecycle Approach to Risk Management Frameworks

Establishing a clear and internationally understood taxonomy for privacy risks associated with LLM-based systems is crucial as it ensures consistent communication and effective collaboration across jurisdictions and stakeholders. A common taxonomy helps define critical concepts, such as what constitutes personal data within the LLM context, facilitating precise identification and management of privacy threats.

The complexity, opacity, and dynamic nature of LLM-based systems require a robust and structured approach to managing privacy risks. Conventional methods, focusing narrowly on isolated stages or relying on static assessments, are insufficient to address the diverse and evolving risks presented by these advanced systems. Drawing from the insights of the European Data Protection Board's report on Privacy risks & Mitigations[22] In LLMs, and the stakeholder consultations highlighted in this report, the need for a comprehensive lifecycle-based risk management framework is increasingly recognised and relevant as documented in the previous Section.

[21] The definition of "small" can drift over time, considering that device memory is increasing over time and can host larger "small language models" in the future. The study sets 5B as the upper limit for the size of SLMs, since as of Sept.2024 7B LLMs are still mostly deployed in the cloud. Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N.D., & Xu, M. (2024). Small Language Models: Survey, Measurements, and Insights. ArXiv, abs/2409.15790.

[22] EDPB Support Pool of Experts: Barberá, I. "AI Privacy Risks & Mitigations: Large Language Models (LLMs)", (2025). Source: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf

A lifecycle-based risk management framework acknowledges that privacy risks manifest differently across various stages: from initial data collection, pre-processing, model training, through post-processing, fine-tuning and deployment, to ongoing post-deployment monitoring and adaptation. At the training stage, privacy risks primarily revolve around inadvertent memorisation and reproduction of sensitive data present in pre-processed training datasets. Post-processing and fine-tuning introduces additional concerns, particularly when performed without sufficient transparency or controls, potentially exacerbating biases or amplifying the risk of data leaks. During inference and user interaction, the risks shift towards profiling and adversarial exploitation, such as prompt injection or social engineering, where malicious actors manipulate models into revealing personal data or performing unsafe actions. These risks often evade traditional detection methods, demonstrating the fundamental need for dynamic, real-time evaluation and monitoring.

Furthermore, system-level cybersecurity risks such as inadequate API security, vulnerabilities in Retrieval-augmented generation architectures, poor interface designs and lack of transparency at the Product layer, highlight the necessity of evaluating privacy risks within the broader system integration and architectural context. Post-deployment risks, including feedback loops, continued data collection, and lack of transparency about data retention and reuse, underscore the critical need for continuous oversight and governance mechanisms throughout the operational lifespan of LLM-based systems. It is therefore a misconception to think that privacy and data protection challenges are restricted only to the Foundation layer in the training data phase, as significant risks can arise at any stage of the lifecycle.

A new risk framework should also recognize the necessity of experimentation during the inception phase when models and tools are designed, along with the privacy and data protection challenges inherent in these exploratory activities while proposing recommendations for risk mitigations.

Both European guidance and the Council of Europe AI Framework Convention underscore the need for embedding privacy-by-design principles throughout the entire AI lifecycle. Specifically, the EDPB's guidance (2025) recommends that risk management frameworks integrate Data Protection Impact Assessments (DPIAs), strong transparency measures, and clearly defined accountability mechanisms. In alignment with these recommendations, this report sets the groundwork for a structured privacy risk assessment framework grounded in the full AI lifecycle that draws on European and international standards, codes of practice, industrial standards, and existing governance models, establishing a solid foundation for future development and implementation.

While this report sets the foundation for a pragmatic, operational lifecycle approach, its effectiveness must be validated through implementation. Section 4.4 outlines how piloting efforts will support this process by examining feasibility, usability, and real-world impact.

4.2 Alignment with European & International Standards

Ensuring alignment with European and international standardisation bodies such as ISO and CEN/CENELEC is essential for achieving both interoperability and long-term impact. As LLM-based systems become globally deployed across jurisdictions with varying regulatory frameworks, the ability to anchor privacy risk governance in widely recognised standards will be critical. This alignment also reinforces the Council of Europe's efforts to develop privacy and data protection guidance that is both rights-based and practically applicable across diverse legal and operational contexts.

The structured privacy risk framework proposed in this report follows a lifecycle-based logic, which is increasingly echoed in ISO AI international standards (e.g., ISO/IEC 23894 on risk management for AI) and within the emerging work of CEN/CENELEC on European standards on risk management and trustworthy AI. Harmonization and mapping the methodology to these standards ensures that Convention 108+ remains a

guiding force in international AI governance discussions, while also enabling convergence with risk management practices used by industry, regulators, and procurement authorities.

Moreover, incorporating lifecycle-aware assessments and impact evaluation mechanisms in line with international norms helps make the Council of Europe's standard setting activity and normative work exportable—i.e., capable of influencing global adoption. In particular, the integration of privacy-by-design principles, structured and real-time evaluation processes, and system-level analysis across the LLM lifecycle contributes to filling the current implementation gap between abstract legal obligations and operational requirements.

From a technical governance standpoint, aligning with these standards enhances consistency in how risks are identified, measured, and mitigated. It enables shared taxonomies and metrics for evaluating privacy harms, supports harmonisation of documentation practices (e.g., model cards and risk reports), and facilitates cross-border accountability. In the longer run, it also helps establish common baselines for acceptable risk thresholds and mitigation trade-offs, which are increasingly required in areas such as public procurement, conformity assessment, and regulatory sandboxes.

Importantly, by anchoring the lifecycle risk management framework in Convention 108+ and aligning it with ISO and CEN/CENELEC standards, the Council of Europe can position its human-rights-based approach not as an alternative to technical risk governance, but as its normative backbone. This promotes a more integrated form of AI governance, one that connects the dots between technical feasibility, legal compliance, and fundamental rights protection.

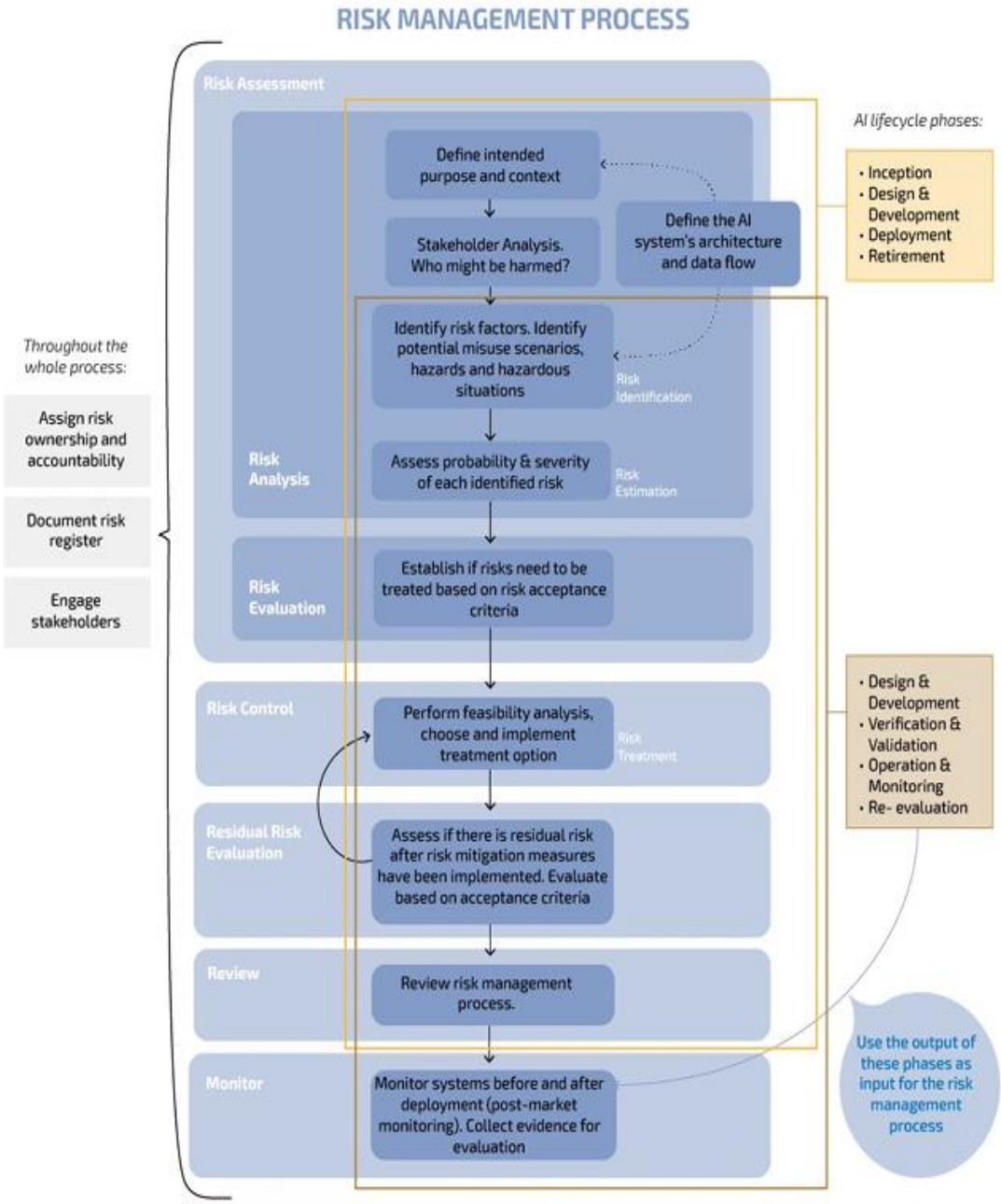


Figure 11:

Foundational Proposal for a Lifecycle-Based Privacy Risk Management Framework
This figure presents an initial framework rooted in the EDPB's report on privacy risks in LLMs and further informed by relevant international and European standards. It serves as a starting point for the development of a comprehensive, rights-based methodology for managing privacy risks throughout the LLM lifecycle.

4.3 Addressing the Gaps in Evaluation, Thresholds, and Transparency to Assess Privacy Risks

An additional dimension that the future framework must address is the persistent lack of guarantees of existing evaluation methods. Without reliable tools to measure privacy risks in realistic settings, developers and regulators are left with an incomplete and sometimes distorted view of system safety.

Despite the growing deployment of LLMs across high-stakes domains such as healthcare, education, finance, and public administration, current risk evaluation methods have significant gaps. Approaches such as automated benchmarking, adversarial red-teaming, and manual human review offer fragmented and sometimes misleading assessments of LLM models and systems safety. Automated benchmarks typically evaluate isolated tasks under controlled conditions, neglecting complex, multi-turn interactions and adversarial scenarios. Consequently, they fail to detect critical risks such as training data memorisation, bias, persuasion, misuse or the generation of sophisticated misinformation[23].

Red-teaming exercises, though valuable, suffer from inconsistency and limited scalability, heavily relying on individual testers' creativity to uncover vulnerabilities. Techniques like prompt injection can reveal serious security gaps, yet these remain difficult to systematically identify or replicate. Human oversight through reinforcement learning from human feedback (RLHF) and moderation is essential but also constrained. Annotators frequently miss subtle harms due to fatigue, biases, and inconsistent or culturally driven criteria, and human review often becomes reactive rather than preventive.

Safety filters intended to mitigate harmful outputs similarly fall short. They can be circumvented by minor variations in prompts, offering a false sense of security rather than addressing root causes of harmful behaviours. Furthermore, these filters often obscure underlying risks by censoring outputs without correcting the model's fundamental biases or vulnerabilities.

The absence of systematic methods for defining and applying risk acceptance thresholds significantly compounds with these evaluation challenges. Current tools rarely provide meaningful quantitative estimates of risk likelihood or severity, complicating informed decision-making. Mitigation strategies such as fine-tuning, while frequently proposed, are often costly, unstable, and ineffective at addressing emergent risks, particularly those that manifest only in dynamic environments or through interactions involving third-party integrations and APIs.

These challenges are particularly acute in the LLM-as-a-Service business model, where deployers commonly lack access to training data, model documentation, and detailed evaluation results. This asymmetry limits

[23] Consider a phenomenon called "sycophancy" which was discovered by recent studies documenting that LLMs outputs mirror the user's beliefs, assuming identical political views or try to please, flatter and ultimately display persuasive communication to foster further engagement or a friendly conversation. This deceptive tendency results in generating persuasive or misleading content to reinforce behaviors, beliefs and prejudices. It has been repeatedly shown in the literature that interactional biases like sycophancy originate from a process happening at the Tool layer called Reinforcement Learning from Human Feedback (RLHF), where human testers steer a model towards human preferences and the provision of more satisfying answers, in this way where models are adapted to prioritise user satisfaction and smooth interaction. See Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., ... & Kaplan, J. (2023, July). Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13387-13434).
LLM tools and applications behaving like echo-chambers hold the potential to impair the right to hold opinions and to access and receive accurate and plural information and ideas. Consider examples in fields such as politics, religious doctrine and beliefs, marketing, public health, historical events, e-commerce, and charitable giving in experimental literature reported in Rogiers, A., Noels, S., Buyl, M., & De Bie, T. (2024). Persuasion with Large Language Models: a Survey. arXiv preprint arXiv:2411.06837.

accountability, leaving deployers responsible for risks they cannot detect or control. Evaluations must therefore consider not just model performance, but also broader system-level factors including interfaces, deployment contexts, and governance structures.

To effectively address these gaps, we propose a trustworthy evaluation framework[24] that meets five core criteria:

1. **Contextual Relevance**: Evaluations must reflect the real-world conditions in which models operate, accounting for actual deployment scenarios rather than synthetic test conditions.
2. **Dynamic and Continuous Evaluation**: Risk assessment must be iterative, updated regularly or even in real time to adapt to new risks emerging over time due to changes in model use or deployment contexts.
3. **Comprehensive Risk Coverage**: Evaluation methodologies should broadly assess systemic, emergent, and interactive risks rather than focusing solely on isolated performance metrics.
4. **Stakeholder Involvement**: Evaluations must engage, when appropriate, a diverse range of stakeholders, including civil society, regulators, domain experts, and affected communities, to identify overlooked risks and increase legitimacy.
5. **Transparency and Accountability**: Detailed evaluation reports, user information, accessible documentation, and independent oversight must be integral to the evaluation process, ensuring public scrutiny and trust.

Implementing these criteria critically requires significant institutional and regulatory alignment. Evaluation must evolve from static compliance exercises into integral components of AI governance frameworks, grounded in diverse and independent expertise and subject to clear regulatory oversight. Such an approach is essential not only for achieving technical reliability, but also for ensuring legal compliance, societal legitimacy, and sustained public trust.

4.4 Piloting the Privacy Risks Management Methodology

To bridge the gap between conceptual design and real-world deployment, the framework should undergo a dedicated piloting phase before full publication. This phase serves as a proving ground where theoretical safeguards and research-based methodologies are tested in operational settings. By experimenting with the framework in controlled environments, the initiative can validate that its principles hold up under real conditions and refine any aspects that need adjustment before wide-scale adoption.

A core feature of the pilot program will be its broad, collaborative approach. Multiple organizations, spanning regulators, public institutions, industry developers (i.e., small, medium and large size companies), and civil society partners, will participate to trial the methodology across diverse LLM deployments. Selected organizations from different sectors and with different roles in the LLM ecosystem (from model providers to end-user service deployers) will be invited to implement the framework in real-world settings, providing different use cases and contexts. This diversity ensures the framework is stress-tested for versatility and lighter compliance burden: each pilot implementation helps identify potential gaps, organizational constraints, or unanticipated barriers or complexities that might not surface in theory. Feedback from these participants is actively gathered in an iterative loop, so that the methodology can be refined based on practical insights. Notably, regulators and data protection authorities will be closely involved, not only to observe and guide, but also to assess how well the framework's measures align with existing legal requirements under Convention 108+

[24] Barberá, I. "The Broken Promise of LLM Evaluations: Gaps, Risks, and a Way Forward", (2025)

and the privacy safeguards of the ECHR (e.g., the right to private life under Article 8) as well as emerging standards under the forthcoming Framework Convention on AI. This multi-stakeholder collaboration will help reveal the challenges and limitations organizations face in implementing privacy risk management and safeguards and ensuring they can be addressed.

Balancing innovation and compliance is a key concern during piloting. The process will be designed to foster creative privacy solutions for LLMs while upholding rigorous standards. Regulatory sandboxes are an example of a controlled environment where innovators and regulators can work side by side. In the sandbox setting, companies can deploy LLM systems with experimental privacy-preserving techniques or enhanced transparency features with the help and under the supervision of regulators. This allows risk assessment tools, transparency mechanisms, and privacy-enhancing techniques (such as differential privacy, federated learning, or new anonymization methods) to be tested thoroughly without immediately breaching compliance obligations. The sandbox approach lets developers iterate and improve on privacy measures in a safe space, while regulators ensure that these innovations remain consistent with data protection principles and legal norms. In practice, this means the piloting phase can trial measures like improved user consent flows, AI explanation interfaces, or robust encryption of sensitive model outputs, all under real conditions but with oversight. This collaborative experimentation encourages innovation in privacy for LLM-based systems, yet keeps that innovation firmly tied to compliance requirements of Convention 108+, the ECHR, and anticipated AI governance rules. Through this controlled experimentation, the pilot phase will highlight what works in practice and flag what might need further policy guidance or technical adjustment, ensuring that by the time the framework is formally finished, it supports both technological progress and legal accountability.

Crucially, the piloting will cover the entire lifecycle of LLM-based systems, applying the risk management framework across at least the five key stages of an LLM's life cycle. These stages include:

(1) Model creation, where training data is collected, pre-processed and models are built with privacy-by-design considerations ;
(2) Post-training adaptation, where models are instructed, finetuned and transformed into assistance tools that are adapted to tasks ;
(3) System integration, in which LLMs are integrated into applications or services (often involving fine-tuning of pre-trained models) with appropriate safeguards;
(4) Operational deployment, referring to the live deployment of the LLM-based system with active monitoring and governance controls; and
(5) End-user interaction, covering how users interact with the LLM-based systems and how autonomous workflows or agentic functions are managed.

Each phase presents distinct privacy risks and challenges that the framework must address. By piloting the framework in each of these phases, the project ensures that the proposed safeguards are effective and context-appropriate at every step, from inception of the model to its ultimate interaction with individuals. This life-cycle approach under Convention 108+ and human rights standards guarantees that privacy and data protection principles are embedded throughout the AI system's development and use, rather than tacked on as an afterthought.

The insights and data gathered from this piloting exercise will directly inform the creation of the privacy risk management lifecycle framework aligned with the HUDERIA methodology and model[25]. A formal guidance document is envisioned as a major output, translating the piloted framework into step-by-step processes, tools, and indicators that can be readily adopted by both public and private actors. This document will distil the lessons

[25] Council of Europe, Committee on Artificial Intelligence (CAI), 'Methodology for the risk and impact assessment of artificial intelligence systems from the point of view of human rights, democracy and the rule of law'.

learned: detailing which privacy controls proved most effective, how to tailor measures to different types of LLM applications, and how to navigate common obstacles that organizations encountered during the pilots. Crucially, the guidance will also enumerate indicators and benchmarks (e.g., acceptable thresholds for re-identification risk, or criteria for what constitutes adequate transparency in an LLM-based service) that emerged from the pilot as useful for monitoring compliance. By incorporating the perspectives of businesses, policymakers, and AI practitioners gathered during the pilot, the guidance document is expected to be context-sensitive and adaptable. In other words, it will acknowledge that different use cases or sectors may require slight adjustments, and it will offer advice on how to scale measures up or down depending on the complexity and risk level of the LLM-based systems deployment.

In sum, piloting provides a grounded, scalable, and rights-based foundation for managing LLM privacy risks in practice. Grounded in the sense that every recommendation in the framework is vetted against real-world applications and adjusted for practical feasibility; scalable in that the framework has been tested in varied scenarios, ensuring it can be applied across use cases of different sizes and sectors; and rights-based because the entire effort is anchored in the fundamental privacy and data protection principles of instruments like Convention 108+, the ECHR and the HUDERIA Methodology and Model. By the end of the piloting phase, the framework will have been proven out through collaborative experimentation, ensuring that it is not only theoretically sound but also operationally effective and legally compliant. This rigorous approach gives regulators and organizations confidence that privacy risks associated with LLMs can be proactively identified and mitigated in a way that upholds individuals' rights and freedoms. Moreover, the pilot's success and the resulting guidance are expected to feed into broader governance efforts triggered by the Framework Convention on AI about how to responsibly innovate with AI while safeguarding human rights. Ultimately, the piloting initiative will help ensure that as LLM technologies advance, they do so hand-in-hand with robust privacy risk management, striking the necessary balance between innovation and compliance for the benefit of individuals and society at large.

Section 5
Results from Stakeholders Interviews

In order to understand how privacy risks are currently addressed in LLM-based system development and deployment, a series of interviews and questionnaires were conducted with a diverse set of stakeholders. These included representatives from LLM model providers, system deployers and system integrators, red-teaming companies, regulators, start-ups, and research-focused organizations. The goal of this inquiry was to identify common practices, challenges, and gaps in privacy risk management in real-world LLM contexts.

Due to time constraints and limited availability of interviewees, the resulting sample is not as representative as would have been ideal. Nevertheless, the findings provide an informative cross-section of current practices and challenges, offering a valuable window into the evolving field of LLM-based systems' privacy governance.

5.1 Interviews Questionnaire

Interviews were supplemented with a structured questionnaire, which participants could complete independently. This questionnaire aimed to capture how organizations define lifecycle phases, apply risk management, identify privacy risks, and use technical and organizational safeguards. The full questionnaire is available in APPENDIX I.

The following sections present aggregated insights drawn from all responses, grouped by thematic categories aligned with the structure of the questionnaire.

Section 1: Organizational Roles and Maturity

Organizations interviewed varied in their roles:

- Some were primarily LLM model developers.
- Others focused on integrating third-party models into proprietary systems.
- A few operated across both domains, or offered red-teaming and security services, supervision or research.

Organizational maturity varied widely. Large tech companies and established deployers generally had dedicated AI governance structures and documented privacy procedures. Start-ups and research teams were more likely to operate in ad hoc or experimental modes, with limited formal processes.

Section 2: Lifecycle Phases and Risk Management

Lifecycle phases used in practice ranged from highly structured (e.g., design, development, implementation, verification, operations) to informal flows based on internal tools. Formal risk management tended to be concentrated at early inception and deployment phases. Only a few organizations systematically evaluated risks at every phase, including post-deployment monitoring.

Risk responsibilities were typically distributed across legal, security, privacy governance, and engineering functions. However, in smaller organizations, researchers or developers often bore informal responsibility, accumulating other transversal roles, without formal accountability mechanisms.

Section 3: Model vs. System-Level Risk Management

Larger organizations consistently distinguished between LLM model development and LLM system integration. They treated models as one subsystem within a broader AI use-case architecture, governed by internal policies.

Smaller actors often lacked this distinction, treating models and systems as functionally unified. Consequently, risk management was largely system-oriented and downstream from model development. Only a minority applied differentiated strategies to models vs. systems.

Section 4: Risk Identification and Privacy Safeguards

Risk identification techniques included design reviews, red teaming, threat modeling, and sandboxing. While some organizations integrated privacy explicitly into these steps, others relied on legal departments to assess data protection independently of technical teams.

Privacy risks such as data leakage, re-identification, and memorization were known concerns, but few organizations had structured methods to assess their likelihood or severity. Risk documentation was often limited to internal logs or ad hoc reporting.

Some organizations made use of internal standards or industry frameworks. However, there was limited alignment with formal privacy risk management standards such as ISO/IEC 27701 or NIST RMF.

Section 5: Evaluation and Testing

Evaluation practices varied:

- Larger organizations conducted structured red teaming and adversarial testing, often distinguishing between critical and non-critical applications.
- Others relied on standard model evaluation benchmarks, and occasionally on logging or manual testing, with little regard for privacy-related metrics.

Risk mitigation decisions were often made at the executive level for high-risk applications. However, most organizations did not formally estimate the probability or severity of privacy harms. Post-deployment monitoring was uneven. Logging was commonly used for debugging and incident review but rarely supported structured privacy auditing.

Section 6: Use of Privacy-Enhancing Technologies (PETs)

Few organizations had systematically adopted PETs.

- Differential privacy and synthetic data were only used experimentally.
- Prompt sanitization was applied in isolated contexts.
- Trusted Execution Environments were mentioned as a next step in experimentation but were not often deployed.

Where PETs were present, they were confined to specific phases or use-cases and lacked lifecycle-wide integration.

5.2 Key Findings

The insights gathered from the interviews and questionnaires reveal a diverse and uneven landscape of privacy risk management practices in LLM-based system development. While some organizations demonstrate relatively advanced governance structures, most stakeholders face shared challenges ranging from conceptual ambiguities to operational gaps. The findings below synthesize recurring themes that emerged across sectors and organizational types, highlighting where current practices fall short and where further support, clarification, or innovation is needed.

A. Findings at Data and Infrastructure Level

#1 Shortcomings of equating data security to privacy and data protection

Several actors in the private sector, including both start-ups and larger technology companies, tend to equate infrastructure and data security with broader privacy and data protection obligations. In contrast, more mature and traditionally data-rich organizations draw on their experience with system testing and governance to implement more comprehensive approaches that address not only compliance with data protection law, but also broader privacy rights. Smaller organizations, in particular, often rely on the default security practices of model or cloud providers without fully understanding how these technical configurations may affect users' privacy and fundamental rights.

#2 Challenges in the experimentation phase to align with data protection compliance

Organisations struggle in the experimentation phase to align with privacy and data protection regulations. The use of production data or the insufficient safeguards deployed during experimentation poses compliance challenges.

#3 Challenges in repurposing data to demonstrate business value of LLM adoption

Efforts to repurpose existing internal data, particularly user data, to support LLM-based innovation and justify return on investment (ROI) increasingly clash with legal constraints under data protection law. In particular, reliance on "legitimate interest" as a legal basis becomes difficult to sustain, as use cases scale and diverge from the original purposes for which the data were collected.

B. Findings at Risk Management process Level

#4 Lack of privacy and data protection benchmarks.

Evaluation is mainly done with off-the-shelf benchmarks mostly focused on performance and lacking privacy criteria.

#5 Generalised lack of probability and severity assessment

Risk prioritization is rarely formalized. Stakeholders noted difficulty estimating likelihoods or modeling potential harms. Some even did not go through this process of risk assessment but directly to mitigation of identified risks.

6 Some stakeholders develop a "step zero" approach by first wanting to understand LLM technology

There is insufficient knowledge about how language models and LLM-based systems work and there is insufficient understanding about how to identify, assess and mitigate their risks. Implementing a risk management approach that has a “step zero” dedicated to technological understanding is fundamental to later identify, assess and mitigate their risks in real context.

7 Need of guidance on privacy rights assessments

Actors with a greater maturity in risk management still struggle to assess the impact on the right to privacy, dignity and autonomy at scale, and call for assistance from regulators.

8 Absence of PETs deployment

The ecosystem has not reached the maturity of using PETs. Only some organisations make a limited use of differential privacy, and synthetic data for some specific projects, while trusted execution environments are exclusively used at the experimentation phase and mostly to safeguard model weights.

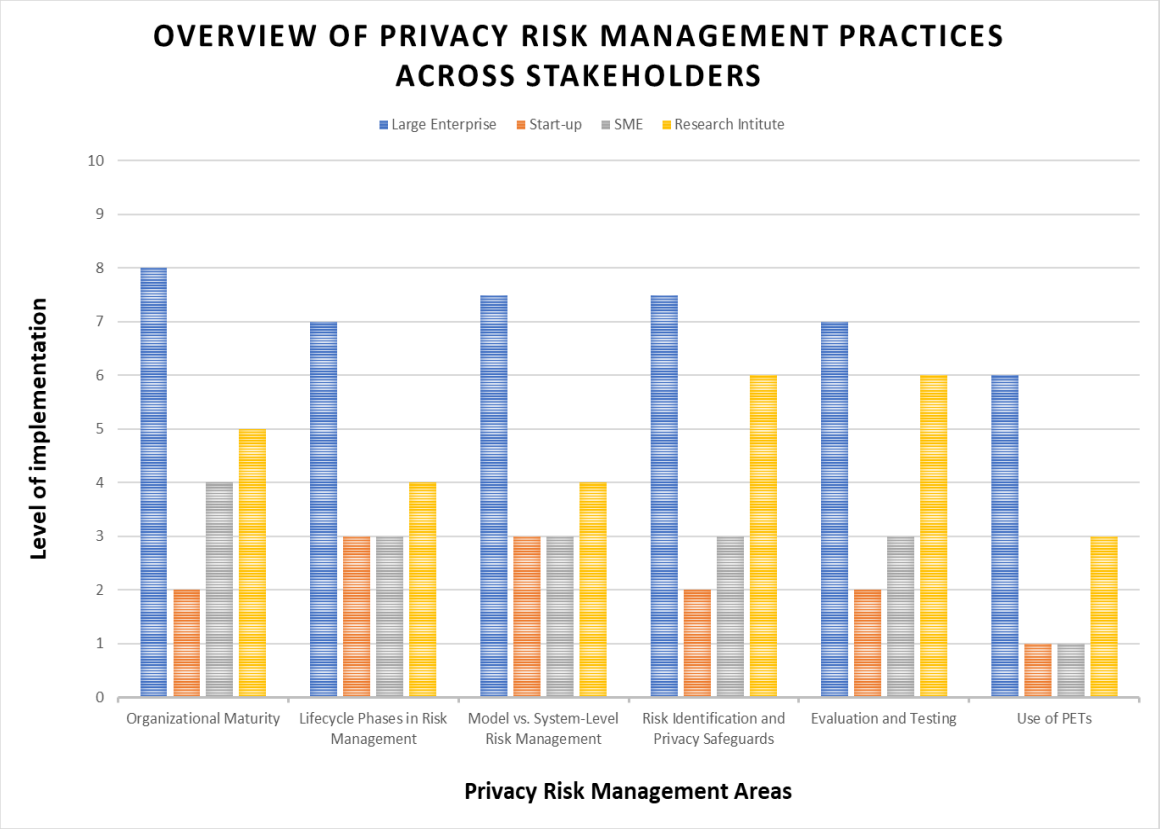


Figure 12: Overview of the Implementation Levels of Privacy Risk Management Practices based on Stakeholder Interviews.

Conclusion and Recommendations

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 tools to identify, assess, and mitigate those risks while promoting compliance with privacy and data protection standards. The Committee's role and previous endeavours in interpreting Convention 108+ in the context of emerging technologies have been key to advancing regulatory clarity, international cooperation, and research-informed policymaking.

Beyond risk management, practical engagement with organizations using LLMs-based systems and agentic workflows will be necessary to ground the report in real-world applications. Through case studies, consultations, and hands-on analysis, the guidance will reflect industry best practices and common challenges. Engaging with key stakeholders, including AI developers, deployers, researchers, policymakers, civil society organizations and regulators, will provide valuable insights into the evolving risk landscape and the effectiveness of existing mitigation strategies. This work will also provide an opportunity to explore how existing evaluation and governance methodologies, particularly Privacy Impact Assessments (PIA) and HUDERIA, can be used to support and complement one another. While PIAs focus specifically on compliance with privacy and data protection obligations, HUDERIA offers a broader perspective centred on systemic impacts on human rights, democracy, and the rule of law. The two methodologies have different goals, scopes, and processes, but their coordinated application can help ensure that both individual data rights and institutional safeguards are addressed when

assessing LLM-based systems and their impact on fundamental rights. 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+ and the Framework Convention on AI.

This expert report consolidates preliminary evidence and legal reasoning to support the case for further development of a future-proof normative framework; one that promotes a proactive, rights-based approach to innovation and safeguards the principles of transparency, accountability, and human dignity at the heart of Convention 108+ and the Council of Europe's new Framework Convention on Artificial Intelligence. While the current report does not offer a definitive solution, it provides the research and science-backed examples, conceptual foundations and risk framing necessary to guide coordinated research, piloting, and standard-setting efforts, and proposes pathways for joint participation by the relevant CoE committees in advancing an integrated approach to AI governance and data protection.

Building on this foundation, the Committee's leadership will be central to shaping a coherent and future-ready approach to privacy governance in LLM-based systems. Its continued role in translating the principles of Convention 108+ into practical standards ensures that emerging technologies are aligned with democratic values and fundamental rights, and Rule of Law within the global mission of the Council of Europe.

This report has laid out the urgent need for a structured, lifecycle-based methodology to assess and manage privacy risks associated with LLM-based systems. However, its success will depend on further development, piloting, and coordinated implementation.

To that end, we recommend three interlinked next steps: first, the refinement of the proposed methodology through real-world piloting with public and private stakeholders; second, the development of a comprehensive guidance document drawing on those pilot experiences; and third, the promotion of international cooperation to ensure regulatory convergence and avoid fragmentation. These actions will support a robust, scalable framework that enables both human rights-centered innovation and accountability.

Ultimately, the Council of Europe is uniquely positioned to anchor this process, ensuring that AI governance and data protection evolve in parallel, and that the protection of human dignity, privacy, and democratic oversight remain at the heart of technological progress. As LLM systems continue to reshape the digital landscape, the next phase must focus on transforming the findings of this report into concrete, actionable tools. With a shared methodology and international coordinated governance, this initiative can serve as a cornerstone for global alignment on privacy in the age of generative AI, anchored in Convention 108+ principles and capable of guiding responsible innovation across borders.

Appendix I

Questionnaire: Research on Privacy Risk Management for LLM-based Systems

Section 1: Background Information

1. **What is your role in your organization?**
   ☐Researcher ☐ Developer ☐ Product Manager ☐ Risk/Compliance Officer ☐ Other: ____Security Specialist_____

2. **What type of organization do you work for?**
   ☐LLM model provider ☐ LLM system deployer ☐ Both ☐LLM systems tester ☐Other: _____

3. **How mature is your organization in terms of deploying or developing LLM systems?**
   ☐Early-stage experimentation ☐ Pilot deployments ☐ Production-level deployment ☐ Other: _____

Section 2: Lifecycle Phases

4. **How do you define the lifecycle phases of an LLM system in your organization?**

5. **At which phases do you apply formal risk management processes?**
   ☐Inception ☐Data collection and curation ☐ Model training ☐ Evaluation and validation ☐Fine-tuning or instruction-tuning ☐ Deployment/integration into systems ☐ Post-deployment monitoring and maintenance

6. **Which stakeholders are responsible for managing risks at each phase?**

Section 3: Model vs System and Privacy

7. **How do you distinguish between LLM model development and LLM system development in your organization?**

8. **Do you apply different risk management strategies for the model vs the system?**
   ☐Yes ☐ No
   If yes, how are they different? _____

Section 4: Risk Identification and Management

9. **How do you identify potential risks associated with LLM systems?**

10. **Are privacy risks evaluated as part of a broader security process, or separately?**
    ☐As part of security ☐ As a distinct process ☐ Not formally evaluated ☐ Other: _____

11. **How are privacy risks specifically identified and addressed (e.g., data leakage, memorization, user re-identification)?**

12. **Do you use any frameworks or standards to guide risk management?**

13. **Are risks formally documented and tracked throughout the lifecycle?**
    ☐Yes, systematically ☐ Partially ☐ No formal documentation

Section 5: Evaluation and Testing

14. **What types of evaluations or tests do you conduct to assess LLM risk? Do you have a specialized approach for Privacy risks?**

15. **Do the results of these evaluations inform risk identification or mitigation planning?**
    ☐Yes, systematically ☐ Occasionally ☐ No
    Please describe how: _____

16. **Do you estimate the likelihood of identified risks occurring? How?**
    ☐Yes, quantitatively ☐ Yes, qualitatively ☐ No
    If yes, how is likelihood estimated? _____

17. **At which lifecycle stages are those evaluations conducted?**
    ☐Inception ☐Data collection and curation ☐ Model training ☐ Evaluation and validation
    ☐Fine-tuning or instruction-tuning ☐ Deployment/integration into systems ☐ Post-deployment monitoring and maintenance

18. **Do you perform red-teaming or adversarial testing? If yes, how do you test for privacy risks?**
    ☐Yes, in-house ☐ Yes, external ☐ No

19. **How do you validate that risks have been mitigated before and after deployment?**

Section 6: Improvements and Future Directions

20. **What challenges do you face in managing risks in LLM systems? And specifically, Privacy risks?**

21. **Have you applied Privacy Enhancing Technologies (PETs) as mitigation measures? If yes, what kind?**

22. **How do you establish guarantees?**

Appendix II

Acknowledgement of Stakeholder Contributions

The evidence base would not have been possible without the generous and thoughtful participation of numerous stakeholders across sectors who contributed through interviews and questionnaires. Their willingness to share insights, experiences, and challenges has been instrumental in building an evidence-based understanding of privacy risk management practices in the context of LLM-based systems.

Given the time constraints and the informal nature of some of the engagements, most participants requested to remain anonymous. In many cases, this decision was due to the limited time available to obtain official institutional consent or clearance to be named publicly. We fully respect these preferences and thank them for their openness and candour despite these constraints.

We would like to extend our sincere gratitude to the organizations that kindly allowed us to name them in this report. These include:

- **Cybernetica (Estonia)**
- **KPN (Netherlands)**
- **Safer AI (France)**
- **HackAPrompt (United States)**
- **Capgemini (France and Netherlands)**
- **Quivr (France)**

We deeply appreciate all the stakeholders, named and anonymous, who contributed their time and expertise to this research.