Strasbourg, 25 February 2026 T-PD(2025)3rev2

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

(CONVENTION 108)

Draft Guidelines on Privacy and Data Protection in the context of LLM-based systems

www.coe.int/dataprotection

Introduction

1. Purpose and Scope

1.1 Context and Purpose

- Assist stakeholders in identifying, assessing, mitigating and monitoring (IAMM approach) privacy risks arising from LLM-based systems under the Articles and principles of Convention 108+.

1.2 The IAMM approach and its steps

1.3 Target audience and stakeholders across the LLM ecosystem

- States and Governments
- LLM Providers (model developers and systems designers)
- Deployers (system integrators and service providers)
- Industry
- Civil society organizations
- Users (i.e. end-users)
- Regulators / Supervisory Authorities

2. Key Concepts and their Definitions

2.1 Key Notions

2.1.1 A Life-cycle approach rooted in the AI Treaty and in the LLM Eco-system

2.1.2 Difference between LLM Model and LLM-based System

2.1.3. Five fundamental steps of LLMs dynamic Lifecycle and Risk management processes

2.2 Types of Privacy Risks Across the AI lifecycle and in the LLM Eco-system

2.2.1 Mapping Privacy Risks in LLM-based Systems Context

2.2.2 Privacy Risks in the LLM Eco-system of Data

2.3 Emerging privacy risks

3. Convention 108+ Principles and Articles Relevant to LLM-based systems

3.1 Understanding the principles of the Convention in the context of the Latest Technological Evolutions in LLMs and agentic frameworks

- Focus on the core Privacy Principles from Convention 108+ [cf. Chapters 2 and 3].
- Explain and contextualize how these principles relate specifically to Large Language Models and LLM-based systems and their lifecycle (e.g., during training, fine-tuning, or deployment.
- Highlight specific tensions and challenges of LLMs and LLM-based systems raised under these principles (e.g., repurposing data, hallucination, inferencing personal data, synthetic data risks).

3.1.1 Data security, accuracy and quality, Transparency and Accountability issues in AI agents

3.1.2 Lawfulness and fairness of use: the case of data proxies leveraged to inferencing personal data and reconstruct individuals' private life

3.1.3 Data minimization and Data subject's rights in the intention economy in the intention economy

3.1.4 Purpose limitation in multimodal data aggregation settings

3.2 Understanding Articles of the Convention in the context of the Latest Technological Evolutions in LLMs and agentic frameworks

- Bipartite structure citing relevant Articles of the Convention and including in each sub-part (a) the interpretation and explanation of the article, and (b) its technological contextualization through examples.

3.2.1 Article 5 : Personal Data Extraction Methods and Memorization
3.2.2 Article 6 : the Processing of Special categories of data
3.2.3 Article 7 & Data Security : when LLMs are leaking personal data
3.2.4 Article 8 & the Transparency of processing
3.2.5 Article 9 & the Rights of data subjects
3.2.6 Article 10 & minimizing the risk of interference with data subjects rights and fundamental freedoms
3.2.7 Article 14 : Transborder flows of personal data.

4. Stakeholder-Specific Guidance

For each stakeholder group (providers, deployers, users, regulators), provide:

4.1 Mapping of Convention 108+ Principles to Actions

- Show how each principle translates into concrete obligations or responsibilities for that group

4.2. Risk Management Responsibilities

- Indicate how each actor is expected to:
  - Identify privacy risks
  - Assess impact
  - Apply mitigation strategies
  - Monitor and review effectiveness

4.3. Mitigation Strategies and Best Practices (per risk and lifecycle stage)

- *Example:*
  - *Providers: Data curation practices, differential privacy, robust evaluation metrics*
  - *Deployers: Use of consent mechanisms, prompt monitoring, and access controls*
  - *Users: Responsible prompt design, privacy-respecting use cases*
  - *Regulators: Oversight mechanisms, capacity-building, and audit criteria*

5. Implementation Considerations

5.1 Highlight governance and accountability mechanisms
5.2 Promote cross-functional collaboration (technical, legal, ethical teams)
5.3 Encourage human rights and fundamental rights impact assessments
5.4 Point to interoperability with other regulatory frameworks (e.g., AI Act, GDPR, DSA)

6. Annexes

Annex I: Risk Management Process for LLMs

- Overview of risk identification, analysis, mitigation, monitoring
- Integration with data protection impact assessments (DPIAs)

Annex II: Lifecycle Phases of LLM Systems

- Detailed description of lifecycle stages relevant for LLMs

Annex III (optional): Case Studies or Examples

- Illustrative examples of privacy risks in real-world LLM deployments
- How different stakeholders addressed them
- Agentic AI implementation

Annex IV (optional): Glossary

- List of key terms with brief definitions for easy reference

Introduction

Emerging technologies create significant opportunities for individuals, organisations and public authorities. At the same time, they may generate serious risks for the effective enjoyment of human rights, the functioning of democracy, and the observance of the rule of law. In recent years, the pace of technological development has increasingly reshaped the conditions under which private life is exercised and personal data are processed, requiring renewed attention to the safeguards that protect individuals in the digital environment. In this

evolving landscape, Large Language Models (LLMs) represent a particularly influential development, combining advanced language-processing capabilities with large-scale automation deployment across sectors thereby amplifying both opportunities for innovation and challenges for rights protection.

LLMs constitute a category of artificial intelligence designed to process and generate human language at scale. Trained on vast amounts of data and embedded in increasingly complex digital infrastructures, they can be used to automate a wide range of tasks, from content generation and summarisation to decision support and automated interaction. [12345] Because their development and operation depend on the large-scale processing of data, which may include personal and sensitive information, LLMs raise significant privacy and data protection considerations. Their technical characteristics, including opacity, probabilistic outputs, and continuous adaptation, may challenge traditional safeguards and complicate the effective application of established data protection principles.

Risks may arise at different points in the lifecycle of LLM-based systems, from model development to system deployment and user interaction. These risks may include unintended retention or reproduction of personal data, insufficient transparency regarding data processing, and difficulties in ensuring meaningful human oversight and accountability. When integrated into automated or semi-automated decision-making environments, such systems may also affect individuals' ability to exercise their rights effectively.

Beyond individual-level impacts, the widespread deployment of LLM-based systems may also have broader societal implications, including effects on autonomy, identity formation, democracy, public discourse and trust in digital environments.

As LLMs continue to expand across sectors such as recruitment, education, healthcare, and public administration, ensuring effective privacy and data protection safeguards becomes increasingly complex. This evolving technological landscape reinforces the urgency of reaffirming and operationalising the principles of Convention 108+ in a manner that remains robust, adaptive, and future proof.

This evolving technological landscape reinforces the need to reaffirm and operationalise the principles of Convention 108+ in a manner that remains robust, adaptive and capable of addressing emerging risks.

These Guidelines are intended to support the application of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108, "Convention 108") and its Amending Protocol CETS No. 223 (referred in this document as "the Convention", "Convention 108+) in the context of LLM-based systems. In doing so, they seek to promote respect for the right to privacy and the protection of personal data, as enshrined in the European Convention on Human Rights (Article 8). The Guidelines build on the previous work of the Consultative Committee of Convention 108 and related Council of Europe guidance on data protection and emerging technologies.[1]

These Guidelines are also grounded in the human-centric approach reflected in the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225) ("the Framework Convention"), which affirms the need to safeguard human rights, democracy and the rule of law throughout the lifecycle of AI systems, including privacy and personal data protection.

Building on the previous work of the Consultative Committee of Convention 108 and related Council of Europe guidance on data protection and artificial intelligence, these Guidelines provide a coherent and forward-looking framework for identifying, assessing and addressing privacy and data protection risks in LLM-based systems.

Section 1 Purpose and Scope

These guidelines address 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 agentic architectures, and have implications not only for data protection, but also for private life, dignity, and autonomy.

The Guidelines present an evidence-based understanding of how LLM-based systems and agentic AI may interfere with individuals' rights and propose a structured methodology to Identify, Assess, Mitigate, and Monitor (IAMM) these risks in real-world settings across diverse stakeholder groups.

1.1 Context and Purpose

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

Persistent governance gaps include difficulties in identifying accountable actors within the value chain in layered AI ecosystems, ensuring transparency in probabilistic outputs, and guaranteeing the effective exercise of data subjects' rights.

The opacity of model training data, inference processes, and downstream system configurations may limit individuals' ability to know whether and how their personal data have been used. This raises questions as to how core principles of Convention 108+, including fairness, transparency, accountability, and effective remedies, can be upheld in practice.

Informing data subjects and identifying accountable data controllers within layered and distributed architectures remains a persistent governance challenge. Where outputs are probabilistic and user-facing configurations opaque, guaranteeing the effective exercise of privacy rights, including the rights to rectification, objection, and meaningful information concerning automated processing, becomes particularly complex.

The absence of observable and accessible data traces and user interfaces further limits individuals' ability to know whether and how their personal data have been used, let alone

from contesting or correcting such processing. This raises serious concerns as to whether the core provisions of Convention 108+ can be upheld in practice without complementary technical and organisational measures. Given the extensive deployment of LLMs-based and agentic systems across diverse contexts, there is an urgent need of preserving and enforcing data subjects' rights under the principles of the Convention 108+ by considering these technological realities.

The purpose of these Guidelines is to provide a comprehensive approach to privacy and data protection risk under Convention 108+ in the context of LLMs and LLM-based systems. They:

- clarify the application of Convention 108+ principles across the LLM lifecycle;
- identify key privacy and data protection risks arising at different lifecycle stages; (Section 2)
- provide a structured methodology for identifying, assessing, mitigating, and monitoring such risks;
- address the roles and responsibilities of relevant stakeholders, including providers, deployers, public authorities, supervisory authorities, and end-users (Section 1); and
- promote coherence with existing Council of Europe instruments and risk-assessment methodologies.

The Guidelines adopt a lifecycle-based methodology reflecting the dynamic and evolving nature of LLMs and LLM-based systems and recognising that privacy and data protection risks may emerge across different stages of the lifecycle (Section 2).
Given the Consultative Committee's role and its previous work in interpreting the Convention 108+ in the context of emerging technologies, these Guidelines inform the foundational legal principles enshrined in the Convention 108+ and their application to LLMs and LLM-based systems throughout their lifecycle (Section 3).
The Guidelines adopt a stakeholder-specific approach when elaborating on concrete obligations or responsibilities for specific groups (providers, deployers, users, regulators). They introduce an evidence-based understanding of how LLMs-based and agentic AI systems may interfere with individuals' privacy and data protection rights and illustrate corresponding mitigation strategies (Section 4).
Overall, the Guidelines provide an up-to-date, structured, research- and risk-informed framework for understanding and governing risks associated with LLMs-based systems. They lay the groundwork for governance tools that operationalize both the principles and the binding obligations of Convention 108+.
A central feature of the operational application of Convention 108+ in this context is the lifecycle-based risk management methodology set out in Section 4. The Guidelines introduce a structured approach to privacy risk governance based on four interrelated steps: "Identify", "Assess", "Mitigate" and "Monitor" ("IAMM approach"). This methodology is designed to support diverse stakeholders in implementing Convention 108+ obligations in real-world settings, ensuring that emerging technologies remain aligned with fundamental rights, democratic values, and the rule of law.
The Guidelines contribute to the standard-setting efforts of the Council of Europe in advancing an integrated approach to data protection and AI governance through promoting a proactive, rights-based approach to innovation while safeguarding the principles of transparency, accountability, and human dignity. The Guidelines provide transversal guidance on governance and accountability mechanisms that promote cross-functional collaboration and

continuous oversight (Section 5). They complement existing evaluation instruments such as Privacy Impact Assessments (under Article 10 of the Convention 108+) and HUDERIA – the Council of Europe's human rights, democracy, and the rule of law impact assessment methodology, by situating privacy risks within a broader systemic perspective and promoting interoperability with other regulatory frameworks (Section 5).

1.2 The IAMM approach and its steps

IAMM Approach

These Guidelines aim to address how organisations can manage privacy risks across the lifecycle of LLM-based systems and agentic architectures, from data collection and model development to deployment, monitoring, and post-market oversight. They respond to widespread uncertainty and inconsistency in current practices and to the pressing need for a harmonised and internationally coherent approach to privacy risk management. In doing so, the Guidelines seek to reduce fragmented internal processes, strengthen the formalisation of privacy-specific assessments, and support the effective application of data protection obligations within complex and rapidly evolving technological architectures.
The approach includes:

- Establishing a common taxonomy (e.g., clarifying the notion of personal data in the context of LLMs) in Section 2.1.
- Analysing both known and emerging privacy risks within the LLM ecosystem, in Section 2.2.
- Analysing how the Principles and Articles enshrined in Convention 108+ and their implementation and safeguards are challenged by latest technological developments, in Section 3.
- Analysing risk management tools, including privacy- and human rights-centred assessment methodologies like the Huderia methodology and COBRA, in Section 2.2 and 4.
- Engaging stakeholders to prioritise piloting and validate the feasibility and relevance of proposed mitigation strategies in Section 4.
- And, grounding best practices in real-world constraints and requirements from businesses, policymakers, and AI practitioners, in Section 5.

The underlying premise is 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 in a dynamic, evolving and incremental manner.
Section 4 outlines 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 distinct phases of development and deployment, a two-tiered assessment of risks at both the model- and composite system-level, and an initial mapping of viable mitigation strategies together with an advanced monitoring approach within a framework named after its four core component: "Identify", "Assess", "Mitigate" and "Monitor" ("IAMM"):

1. Identifying privacy risks and recommendations based on Convention 108+. This first step 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.

2. Grounding these Guidelines in a two–tiered risk assessment methodology, allows a risk identification and assessment approach that distinguishes between model-level and system-level risks and responsibilities for different stakeholders along the technological lifecycle, particularly when considering 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. The aim is to equip data controllers with tools to evaluate risks stemming from different steps of the lifecycle going form 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 and monitoring strategies in real-world applications. Developing a structured approach to Privacy Impact Assessments tailored for LLM-based and agentic 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 will be pointed at as viable mitigation strategies, always by keeping an eye on the applicability, limitations, and governance implications of these tools in practical contexts.

1.3 Target Audience and stakeholders across the LLM ecosystem

Stakeholders Across the LLM Ecosystem

To map the concrete challenges faced by organisations developing or using LLMs, the target audience is constituted by a wide spectrum of stakeholders across the LLM lifecycle and ecosystem, from start-ups, major technology providers, technology auditors, private and public technology deployers, regulatory authorities, and research institutions.

The roles and responsibilities are distributed as well as interdepended across multiple actors and stakeholders. The privacy and data protection risk mitigations are addressed across the entire lifecycle, from model creation to user interaction. This system of governance incorporates multi-stakeholder dimension with distinct legal and operational responsibilities:

– *States and Governments*

The Convention 108+ establishes a technologically neutral, principle-based approach addressing challenges resulting from emerging technologies, the operational strength of which is defined by States at the national level. The primary responsibility for establishing or maintaining legal and institutional frameworks for protecting individuals' right to privacy and data protection is guaranteed by States and Governments. These Guidelines provide general guidance on developing, endorsing, and implementing context-based privacy risk assessment frameworks for LLM-based systems and support regulatory converge to avoid fragmentation across diverse sectors.

– *LLM Providers (Model Developers and System Designers)*

A lifecycle approach grounded in Convention 108+ and the Framework Convention standards guarantees that privacy and data protection principles are embedded throughout the development and use of AI systems, rather than introduced only at later stages as an afterthought. Model developers and system designers should embed privacy by design and by default at the development (model creation) stage, where training data is collected, pre-processed, and models are built. These Guidelines facilitate LLM providers' understanding of key data protection concepts established by the Convention 108+, their legal compliance and risk assessment efforts throughout the lifecycle of LLM-based systems.

– *Deployers (System Integrators and Service Providers)*

As key operational stakeholders, system integrators and service providers remain responsible for the LLM-based systems application in their daily operations. Even in the case of third-party system deployment, they carry legal accountability and responsibility for data processing operations and observe potential privacy threats at post-deployment stage. These Guidelines support deployers' legal compliance efforts, assist in privacy risk assessment prior to deploying the system, and guide in effective adoption of appropriate privacy-enhancing technologies (PETs).

– *Industry*

Industry actors play significant role through developing technical standards and best practices that can facilitate achieving convergence across diverse sectors. In this pursuit, the Guidelines will assist in aligning privacy risk governance and management efforts and implementing lifecycle-based assessments.

– *Civil Society Organisations*

Civil Society Organisations effectively contribute to awareness raising about privacy risks and to challenging legal compliance of system's lifecycles, addressing privacy-related risk awareness and feedback loops between stakeholders and regulators. The Guidelines will offer privacy risk assessment approach and best practices for mitigation strategies to Civil Society Organisations to continue their awareness raising mission and action.

– *Users (End-Users)*

The rapid technological advancement gave rise to potential adverse impacts on human rights and freedoms, including rights to privacy and data protection. These Guidelines inform regarding privacy-related challenges interfering with data subjects' rights and provide structured methodology for "identifying", "assessing", "mitigating", and "monitoring" such risks in real-world settings for a variety of stakeholders.

– *Regulators / Supervisory Authorities*

Data protection supervisory authorities and other competent authorities significantly contribute to overseeing the lawfulness of data processing operations throughout the entire lifecycle of LLM-based systems and legal compliance of data protection frameworks. These Guidelines will help them in setting regulatory convergence thresholds for system's transparency and accountability, effectively exercise supervisory authorities with respect to LLM-based systems and adopt or review privacy risk assessment methodologies for the LLMs lifecycle. The document will assist them in facilitating coordination between different actors.

Section 2
Key Concepts and their definitions

This section provides a technical overview of how personal data may be processed within LLM-based systems, establishing the conceptual foundational layer for the proposed risk assessment framework. It clarifies key terminology and examines the privacy implications of current developments, including issues related to explainability and transparency, and their relevance for AI governance.

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.

2.1 Essential Concepts

2.1.2 A Life-cycle approach rooted in the AI Treaty and in the LLM Eco-system

An introductory privacy risk analysis of the lifecycle of LLM-based systems can identify four main stages based on the types of data processing involved in the training of foundation models, post-training adaptation and alignment, system integration and optimisation; and deployment and user interaction in specific operational contexts.

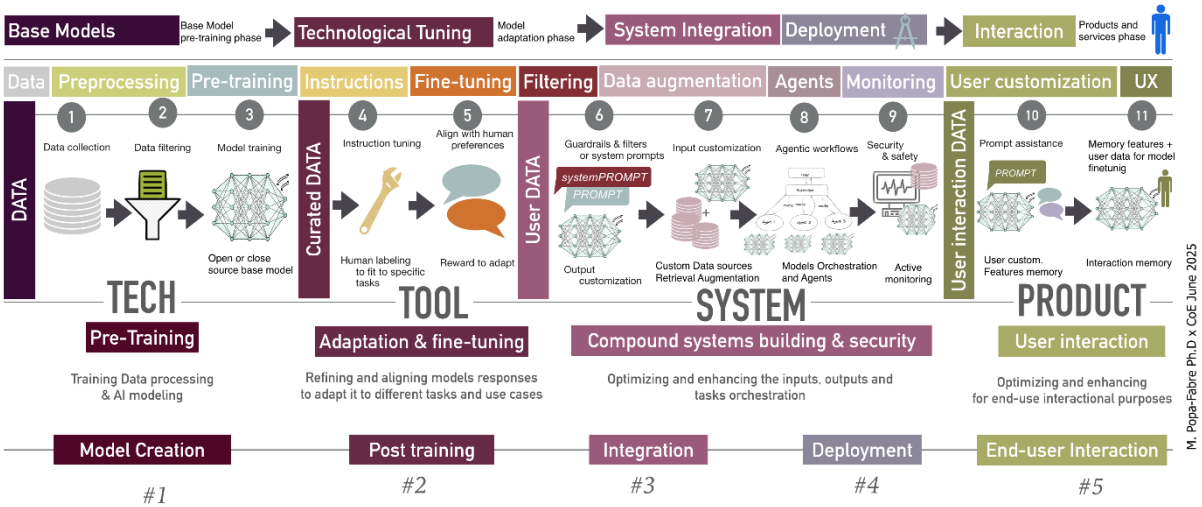

Figure 1: Lifecycle/value-chain LLM-based systems, a broad overview based on the fast-evolving practices of building AI compound systems to optimise LLM-based applications at the "Product" phase.

Importantly, while the foundational phase of LLM development typically relies on the processing of very large and diverse datasets, including data collection and preprocessing (steps 1 to 3), subsequent stages increasingly depend on more curated and context-specific data. During post-training adaptation, models are refined and aligned for particular purposes, often through the processing of targeted datasets and task-specific inputs. This intermediate phase, which includes fine-tuning, system integration, and preparation for deployment (steps 4 to 9), is particularly significant from a privacy perspective. It is at this stage that data protection safeguards can be embedded into system design, governance structures, and

operational configurations. Despite its importance, this phase is often insufficiently examined in privacy discussions.

In later stages of the lifecycle, optimisation and customisation practices further shape the functioning of LLM-based systems in operational environments (steps 10 to 11). The evolving landscape of integration techniques and compound system architectures (steps 7 to 11) has enabled standalone models to operate within increasingly complex and agentic frameworks. These may include, for example:

- Data augmentation techniques such as Retrieval-Augmented Generation (RAG), which allow models to access external or proprietary data sources in response to user queries;
- Agentic workflows that orchestrate multiple models or tools to perform multi-step tasks with varying degrees of autonomy.

At advanced stages of deployment, additional layers of personalisation may be introduced. These frequently rely on the intensive processing of personal data to tailor applications, products, or services to individual users. Such mechanisms may include user intention decoding and prompt customisation features designed to infer user preferences, habits, or contextual cues in order to dynamically adapt system responses. While these developments enhance functionality and adaptability, they may also expand the scope, volume, and sensitivity of personal data processing, thereby increasing privacy and data protection risks.

2.1.1 Difference between LLM Model and LLM-based System

Model vs. System Risks in the LLM Ecosystem

Together with the nascent agent-based personal assistance applications (so-called *Personal Intelligence products)*, 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 and enhanced personalized assistance, raise serious questions about data protection. In particular, they blur the boundaries between training, personalization, and continuous data collection phases, 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 design 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.
- 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 stage, data quality and availability play a critical role. However, it is important to distinguish between the privacy risks associated with foundation LLM models and those arising from LLM-based systems in operational contexts.

2.1.2. Five fundamental steps of LLMs dynamic Lifecycle and Risk management processes

Landscape Tech Ecosystem of LLM-based Systems and Personal Data

A general mapping of the technological ecosystem of LLM-based systems, as illustrated in Figure 1, provides a lifecycle-oriented overview and clarifies the fundamental distinction between foundation Large Language Models and the broader systems built with and around them. Such mapping demonstrates that different technological and economic actors may be involved at each stage, with distinct roles and responsibilities. It also highlights that the nature, scope, and sensitivity of the data processed, including personal data, vary significantly across these phases.

The main stages addressed 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 finetuning 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.

(5) End-user interaction, covering how users interact with the LLM-based systems and how autonomous workflows or agentic functions are managed.

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 model and system 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[1], 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.

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.

2.2 Types of privacy risks Across the AI lifecycle and in the LLM Ecosystem

2.2.1 Mapping Privacy Risks in LLM-based Systems context

LLMs and LLM-based systems 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 risks manifest not only through direct data leakage or memorisation, but more broadly through model behaviours that enable identity manipulation, impersonation, disinformation, or potentially manipulative predictive behaviour. 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 the specific phases of the model and system lifecycle. Each phase presents distinct risk profiles, system exposures, and governance challenges that require context-sensitive assessment:

1. Phase 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

[1] 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.

memorise and later reproduce this content, challenging principles such as data minimisation, purpose limitation, and fairness under Convention 108+.

2. Phase 2: Post-training instruction, Fine-Tuning
   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. Phase 3: System-Level integration and deployment Vulnerabilities in Tasks orchestrations in AI agents 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.

4. Phase 4: Operational Deployment/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.

5. Phase 5: Inference, User Interaction in Apps and Agentic AI
   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.

Yet beyond these structured phases, LLMs also introduce a broader class of systemic and societal-level harms, which require a different kind of scrutiny[2].

2.2.2 Privacy Risks in the LLM Ecosystem of Data

While the previous section mapped privacy risks across the lifecycle phases, this section focuses specifically on the categories of personal data processed within the LLM ecosystem. Understanding what types of data are involved is essential to assessing the scope, sensitivity, and potential impact of privacy risks.

---

[2] 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]}

The data ecosystem surrounding LLM-based systems is layered and dynamic. Different forms of personal data are processed at different stages of model development, integration, and deployment. These categories vary in origin, sensitivity, identifiability, and governance implications.

Phase 1: Training and Pre-Training Data
Training datasets may include publicly accessible online content, licensed corpora, proprietary databases, and other aggregated sources. Such datasets may contain:

- Identifiable personal information, such as names, contact details, or professional affiliations;
- User-generated content, including forum discussions, social media posts, and comments;
- Special categories of data, where scraped or included without adequate filtering;
- Embedded confidential information, such as API keys or personal communications.

At this stage, personal data are often processed at scale and without direct interaction with data subjects. The scale and aggregation of such data increase the potential for unintended retention, inference of sensitive attributes, and reduced transparency regarding provenance and lawful basis.

Phase 1 & 2: Testing and Evaluation Data in Post-training and Fine-tuning
During testing, benchmarking, and red-teaming, additional datasets may be introduced. These may include curated personal data, synthetic data derived from real individuals, adversarial prompts, or evaluation of datasets designed to measure performance. Although often considered technical artefacts, these datasets may still contain identifiable or inferable personal information. Their use raises questions concerning secondary processing, safeguards in experimental environments, and the differentiation between anonymised and pseudonymised data.

Phase 3 & 4: Operational and Interactional Data
Once deployed, LLM-based systems process data generated through user interaction. These may include:

- User queries and prompts;
- Uploaded documents or contextual inputs;
- Retrieved content in architectures such as retrieval-augmented systems;
- Usage logs, telemetry, and behavioural metadata.

Operational data are often highly contextual and may reveal preferences, intentions, habits, or sensitive personal circumstances. Unlike training data, these inputs are directly linked to identifiable users and may be processed continuously or retained for system optimisation.

Phase 5: Monitoring, Feedback, and Retraining Data
Post-deployment improvement mechanisms frequently rely on feedback data, performance logs, and selected interaction records. Over time, these feedback loops may generate new datasets derived from user behaviour. Such iterative data processing may blur the boundaries between initial purposes and subsequent optimisation, raising concerns related to purpose limitation, retention periods, and role allocation among actors in the ecosystem.

Phase 5: Archival and Deletion Data

At later stages, decisions must be taken regarding retention, anonymisation, or deletion of training data, operational logs, and derived datasets. In multi-actor ecosystems, determining responsibility for erasure or rectification may be complex, particularly where personal data are embedded in model parameters, logs, or distributed infrastructures.

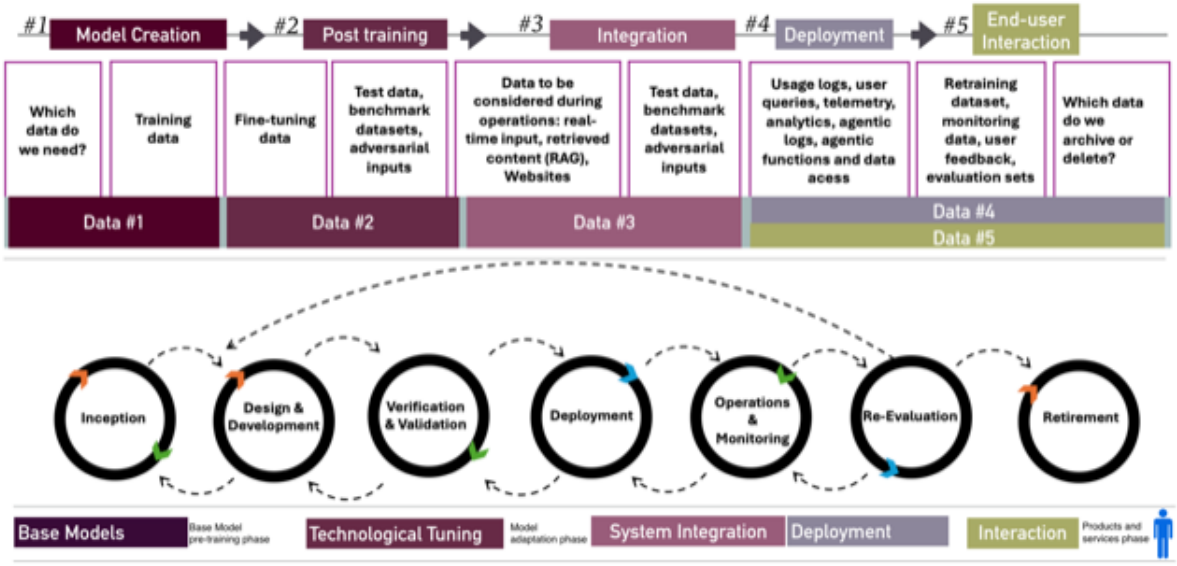

Figure 2: Illustrates the distribution of privacy risks across different lifecycle stages of LLM-based systems with a focus on data access and flows. (Modified from EDPB's report on Privacy Risks & Mitigations in LLMs. The lifecycle phases of ISO/IEC 22989 are also illustrated in the image.

2.3 Emerging Privacy Risks

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, predictions 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 and the composite systems and agentic architectures build upon. 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 individuals 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.