SUPPORT POOL
OF EXPERTS PROGRAMME

AI Privacy Risks & Mitigations

Large Language Models (LLMs)

By Isabel BARBERÁ

As part of the SPE programme, the EDPB may commission contractors to provide reports and tools on specific topics.

The views expressed in the deliverables are those of their authors and they do not necessarily reflect the official position of the EDPB. The EDPB does not guarantee the accuracy of the information included in the deliverables. Neither the EDPB nor any person acting on the EDPB's behalf may be held responsible for any use that may be made of the information contained in the deliverables.

Some excerpts may be redacted or removed from the deliverables as their publication would undermine the protection of legitimate interests, including, inter alia, the privacy and integrity of an individual regarding the protection of personal data in accordance with Regulation (EU) 2018/1725 and/or the commercial interests of a natural or legal person.

Document submitted in February 2025, updated in March 2025

TABLE OF CONTENTS:

Disclaimer by the Author: The examples and references to companies included in this report are provided for illustrative purposes only and do not imply endorsement or suggest that they represent the sole or best options available. While this report strives to provide thorough and insightful information, it is not exhaustive. The technology analysis reflects the state of the art as of March 2025 and is based on extensive research, referenced sources, and the author's expertise. For transparency reasons, the author wants to inform the reader that a LLM system has been used for the exclusive purpose of improving the readability and formatting of parts of the text.

1. How To Use This Document

This document provides practical guidance and tools for developers and users of Large Language Model (LLM) based systems to manage privacy risks associated with these technologies.

The risk management methodology outlined in this document is designed to help developers and users systematically identify, assess, and mitigate privacy and data protection risks, supporting the responsible development and deployment of LLM systems.

This guidance also supports the requirements of the GDPR Article 25 Data protection by design and by default and Article 32 Security of processing by offering technical and organizational measures to help ensure an appropriate level of security and data protection. However, the guidance is not intended to replace a Data Protection Impact Assessment (DPIA) as required under Article 35 of the GDPR. Instead, it complements the DPIA process by addressing privacy risks specific to LLM systems, thereby enhancing the robustness of such assessments.

Structure and Content Overview

The document is structured to guide readers through key technological concepts, the risk management process, main risks and mitigation measures and practical examples. It aims to support organizations in deploying LLM-based systems responsibly while identifying and mitigating privacy and data protection risks to individuals.

Below is an overview of the document's structure and the topics covered in each section:

2. Background
This section introduces Large Language Models, how they work, and their common applications. It also discusses performance evaluation measures, helping readers understand the foundational aspects of LLM systems.

3. Data Flow and Associated Privacy Risks in LLM Systems
Here, we explore how privacy risks emerge across different LLM service models, emphasizing the importance of understanding data flows throughout the AI lifecycle. This section also identifies risks and mitigations and examines roles and responsibilities under the AI Act and the GDPR.

4. Data Protection and Privacy Risk Assessment: Risk Identification
This section outlines criteria for identifying risks and provides examples of privacy risks specific to LLM systems. Developers and users can use this section as a starting point for identifying risks in their own systems.

5. Data Protection and Privacy Risk Assessment: Risk Estimation & Evaluation
Guidance on how to analyse, classify and assess privacy risks is provided here, with criteria for evaluating both the probability and severity of risks. This section explains how to derive a final risk evaluation to prioritize mitigation efforts effectively.

6. Data Protection and Privacy Risk Control
This section details risk treatment strategies, offering practical mitigation measures for common privacy risks in LLM systems. It also discusses residual risk acceptance and the iterative nature of risk management in AI systems.

7. Residual Risk Evaluation
Evaluating residual risks after mitigation is essential to ensure risks fall within acceptable thresholds and do not require further action. This section outlines how residual risks are evaluated to determine whether additional mitigation is needed or if the model or LLM system is ready for deployment.

8. Review & Monitor
This section covers the importance of reviewing risk management activities and maintaining a risk register. It also highlights the importance of continuous monitoring to detect emerging risks, assess real-world impact, and refine mitigation strategies.

9. Examples of LLM Systems' Risk Assessments
Three detailed use cases are provided to demonstrate the application of the risk management framework in real-world scenarios. These examples illustrate how risks can be identified, assessed, and mitigated across various contexts.

10. Reference to Tools, Methodologies, Benchmarks, and Guidance
The final section compiles tools, evaluation metrics, benchmarks, methodologies, and standards to support developers and users in managing risks and evaluating the performance of LLM systems.

Guidance for Readers

- For Developers: Use this guidance to integrate privacy risk management into the development lifecycle and deployment of your LLM based systems, from understanding data flows to how to implement risk identification and mitigation measures.
- For Users: Refer to this document to evaluate the privacy risks associated with LLM systems you plan to deploy and use, helping you adopt responsible practices and protect individuals' privacy.
- For Decision-makers: The structured methodology and use case examples will help you assess the compliance of LLM systems and make informed risk-based decisions.

2. Background

What Are Large Language Models?

Large Language Models (LLMs) represent a transformative advancement in artificial intelligence. These general purpose models are trained on extensive datasets, which often encompass publicly available content, proprietary datasets, and specialized domain-specific data. Their applications are diverse, ranging from text generation and summarization to coding assistance, sentiment analysis, and more. Some LLMs are multimodal LLMs, capable of processing and generating multiple data modalities such as image, audio or video.

The development of LLMs has been marked by key technological milestones that have shaped their evolution. Early advancements in the 1960s and 1970s included rule-based systems like ELIZA, which laid foundational principles for simulating human conversation through predefined patterns. In 2017, the introduction of transformer architectures (see Figure 2) in the seminal paper "Attention Is All You Need"[1] revolutionized the field by enabling efficient handling of contextual relationships within text sequences. Subsequent developments, such as OpenAI's GPT series and Google's BERT (see Figure 3), have set benchmarks for natural language processing (NLP)[2], culminating in models like GPT-4, LaMDA[3], and DeepSeek-V3[4] (see Figure 4) integrating multimodal capabilities.

How Do Large Language Models Work?

LLMs are advanced deep learning models designed to process and generate human-like language. They rely on the **transformer architecture[5]**, which uses attention mechanisms to understand context and relationships between words. While most state of the art LLMs rely on transformers due to their scalability and effectiveness, alternatives[6] exist based on RNN (Recurring Neural Networks) such as LSTM (Long-short Term Memory) and others that are actively being researched[7]. For now, transformers dominate general-purpose language models, but innovations in architectures such as those introduced by DeepSeek's models, may reshape the landscape in the future.

The development[8]of LLMs can be divided into several key stages:

1. Training Phase: Building the Model

In this phase LLMs learn patterns, context, and structure in language by analyzing vast datasets.

1. **Dataset Collection**:
   The foundation of LLM training lies in the use of extensive datasets (such as such as Common Crawl and Wikipedia) that are carefully curated to ensure they are relevant, diverse, and high-quality. Filtering eliminates low-quality or redundant content, aligning the training data with the intended goals of the model.
2. **Data Pre-processing**:

[1] A.Vaswan et al., 'Attention Is All You Need' (2023) https://arxiv.org/pdf/1706.03762

[2] Wikipedia, 'Natural language processing' (2025) https://en.wikipedia.org/wiki/Natural_language_processing

[3] E.Collins and Z.Ghahramani, 'LaMDA: our breakthrough conversation technology' (2021) https://blog.google/technology/ai/lamda/

[4] Github, 'Deepseek' (n.d) https://github.com/deepseek-ai/DeepSeek-V3

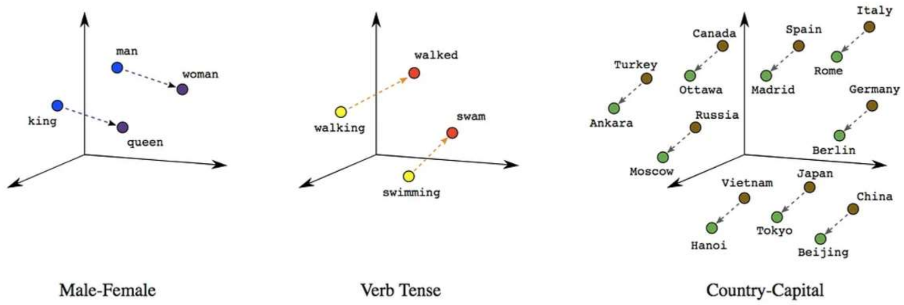
[5] Wikipedia, 'Deep Learning Architecture' (2025) https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

[6] Artificial Intelligence, 'Why does the transformer do better than RNN and LSTM in long-range context dependencies?' (2020) https://ai.stackexchange.com/questions/20075/why-does-the-transformer-do-better-than-rnn-and-lstm-in-long-range-context-depen

[7] A.Gu, T.Dao, 'Mamba: Linear-Time Sequence Modeling with Selective State Spaces' (2024) https://arxiv.org/pdf/2312.00752, B.Peng et al, 'RWKV: Reinventing RNNs for the Transformer Era' (2023) https://arxiv.org/pdf/2305.13048

[8] Y.Liu et al., 'Understanding LLMs: A Comprehensive Overview from Training to Inference' (2024) https://arxiv.org/pdf/2401.02038v2

- Text is cleaned and normalized by removing inconsistencies (e.g., special characters) and irrelevant content, ensuring uniformity in the training data.
- Text data is broken into smaller units called tokens, which can be words, subwords, or even individual characters. Tokenization algorithms transforms unstructured text into manageable sequences for computational processing.
- Tokens are converted into numerical IDs that represent their vocabulary position. These IDs are then transformed into word embeddings[9]—dense vector representations that capture semantic similarities and relationships between words. For instance, semantically related words like "king" and "queen" will occupy nearby positions in the embedding space.



Linear Relationships between Words. Image from developers.google.com

Figure 1. Source: S.Anala 'A Guide to Word Embedding' (2020) https://medium.com/data-science/a-guide-to-word-embeddings-8a23817ab60f

3. **Transformer Architecture:**[10]
   Transformer architectures can be categorized into three main types: encoder-only, encoder-decoder, and decoder-only. While encoder-only architectures were foundational in earlier models, they are generally not used in the latest generation of LLMs. Most state of the art LLMs today use decoder-only architectures, while encoder-decoder models are still used in tasks like translation and instruction tuning.

   - **Encoder:**[11]
     The encoder takes the input text and converts it into a contextualized representation by analyzing relationships between words. Key elements include:
     - *Token embeddings*: Tokens are transformed into numerical vectors that capture their meaning.
     - *Positional encodings*: Since the transformer processes words in parallel, positional encodings are added to token embeddings to represent the order of words, preserving the structure of the input.
     - *Attention mechanisms*: The encoder evaluates the importance of each word relative to others in the input sequence, capturing dependencies and context. For example, it helps distinguish between "park" as a verb and "park" as a location based on the surrounding text.
     - *Feed-Forward Network*: A series of transformations are applied to refine the contextualized word representations, preparing them for subsequent stages.

[9] V.Zhukov, 'A Guide to Understanding Word Embeddings in Natural Language Processing (NLP)' (2023) https://ingestai.io/blog/word-embeddings-in-nlp

[10] See footnote 1

[11] Geeksforgeels, 'Architecture and Working of Transformers in Deep Learning' (2025) https://www.geeksforgeeks.org/architecture-and-working-of-transformers-in-deep-learning/

- **Decoder:**[12]
  The decoder generates text by predicting one token at a time. It builds upon the encoder's output (if used) and the sequence of tokens already generated. Key elements include:
  - *Input*: Combines encoder outputs with tokens generated so far.
  - *Attention mechanisms*:[13] Ensures each token considers previously generated tokens to maintain coherence and context.
  - *Feed-Forward Network (FFN)*:[14] This layer refines the token representations to ensure they are relevant and coherent.
  - *Masked attention*: During training, future tokens are hidden from the model, ensuring it predicts outputs step by step without "cheating".

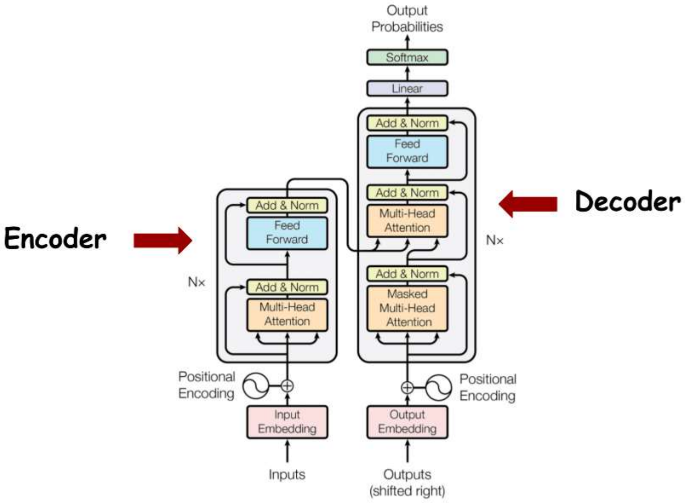
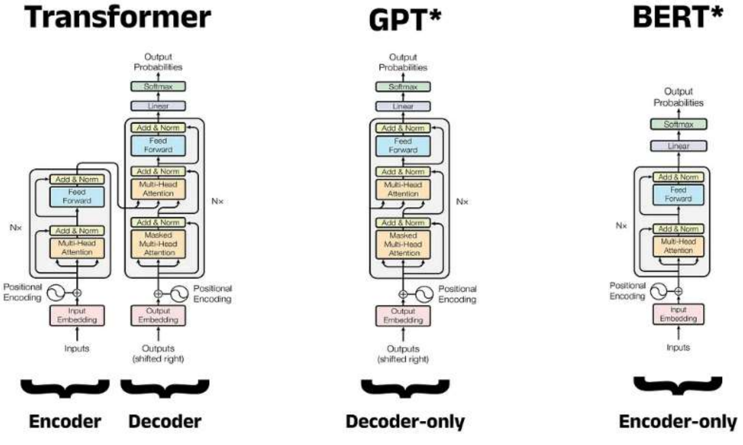


Figure 2. Transformer architecture.
Source: Vaswani et al. 'Attention Is All You Need ' (2023)
https://arxiv.org/pdf/1706.03762


Transformer GPT* BERT*

*Illustrative example, exact model architecture may vary slightly

Figure 3. A comparison of the architectures for the Transformer, GPT and BERT.
Source: B.Smith 'A Complete Guide to BERT with Code' (2024)
https://towardsdatascience.com/a-complete-guide-to-bert-with-code-9f87602e4a11

[12] idem

[13] The architecture of DeepSeek models contains an innovative attention mechanism called Multi-head Latent Attention (MLA) that compresses Key/Value vectors offering better compute and memory efficiency.

[14] DeepSeek models employ the DeepSeekMoE architecture based on Mixture-of-Experts (MoE) introducing multiple parallel expert networks (FFNs) instead of a single FFN.

Mixture of Experts (MoE) is a technique used to improve transformer-based LLMs making them more efficient and scalable. Instead of using the entire model for every input, MoE activates only a few smaller parts of the model—called "experts"—based on what the input needs. This means the model can be much larger overall, but only the necessary parts are used at any time, saving computing power without losing performance

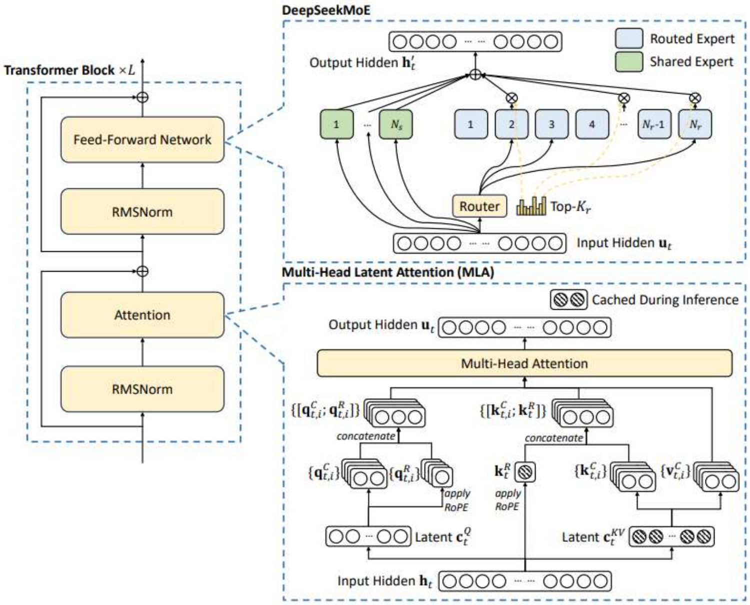


Figure 4. Illustration of DeepSeek-V3's basic architecture called DeepSeekMoE based on Mixture-of-Experts (MoE).
Source: 'DeepSeek-V3 Technical Report'
https://arxiv.org/pdf/2412.19437

4. **Training/Feedback loop & Optimization**[15]
   The training phase of LLMs relies on a structured optimization loop to enhance the model's ability to generate accurate outputs. This iterative process consists of the following steps:
   - *Loss calculation*: After generating an output sequence, the model compares it to the target sequence (the "correct answer"). A loss function quantifies the error, providing a numerical measure of how far the predicted output deviates from the desired result (the loss).
   - *Backward pass*: The obtained loss value is used to compute gradients, which indicate how much each model parameter (e.g., weights and biases) contributed to the error. These gradients highlight areas where the model needs improvement.
   - *Parameter update*: Using an optimization algorithm, such as Adam or SGD (Stochastic Gradient Descent), the model's parameters are adjusted. This step reduces the error for future predictions by refining the internal model weights.
   - *Repetition*: This process repeats for thousands or millions of iterations. Each cycle incrementally improves the model's performance. Training stops when the model reaches a balance between accuracy on training data and generalization to unseen inputs.

2. Continuous Improvement – Model Alignment (post training)

Pre-trained models, while powerful, are generally not immediately useful in their raw form. To make models' behaviour align with ethical considerations and user preferences[16] they need to be tuned. This

[15] 'PyTorch Loss.backward() and Optimizer.step(): A Deep Dive for Machine Learning' (2025) https://iifx.dev/en/articles/315715245
[16] C.R. Wolfe, 'Understanding and Using Supervised Fine-Tuning (SFT) for Language Models ' (2023) https://cameronrwolfe.substack.com/p/understanding-and-using-supervised

process often involves the use of techniques such as supervised fine-tuning on domain-specific data or Reinforcement Learning with Human Feedback (RLHF). The most common alignment methods are:

- **Supervised Fine-Tuning (SFT):**[17] This approach involves training a pre-trained model on a labeled dataset tailored to a specific task, with adjustments made to some or all of its parameters to enhance performance for that task.
- **Instruction Tuning:**[18] This technique is used to optimize the LLM for following user instructions and handling conversational tasks.
- **Reinforcement Learning with Human Feedback (RLHF):**[19]
  This method uses human feedback to train a reward model (RM), which helps guide the AI during its learning process. The reward model acts as a scorekeeper, showing the AI how well it's performing based on the feedback. Techniques like Proximal Policy Optimization (PPO) are then used to fine-tune the language model. In simple terms, the language model learns to make better decisions based on the reward signals it receives. Direct Preference Optimization (DPO)[20] is an emerging reinforcement learning approach that simplifies this process by directly incorporating user preference data into the model's optimization process.
  While RLHF aims to align the model with human preferences across diverse scenarios using human feedback, another variation of the PPO technique called Group Relative Policy Optimization (GRPO)[21] introduced by DeepSeek researchers, takes a different approach. Instead of relying on human annotations, GRPO uses computer-generated scores to guide the model's learning process and reasoning capabilities in an automated manner.
- **Parameter-Efficient Fine-Tuning (PEFT):**[22] This technique adapts pre-trained models to new tasks by training only some of the model's parameters, leaving the majority of the pre-trained model unchanged. Some PEFT techniques are adapters, LoRA, QLoRA and prompt-tuning.
- **Retrieval-Augmented Generation (RAG):**[23][24][25] This method enhances LLMs by integrating information retrieval capabilities, enabling them to reference specific documents. This approach allows LLMs to incorporate domain-specific or updated information when responding to user queries.
- **Transfer Learning:**[26][27] With this technique, knowledge learned from a task is re-used in another model.
- **Feedback loops:**[28] Real-world user feedback helps refine the model's behavior, allowing it to adapt to new contexts or correct inaccuracies. Feedback can be collected through user behaviour, for instance inferring whether the user engages with or ignores a response. Feedback can also be collected when users directly provide feedback on the model's output, such as a thumbs-up/thumbs-down rating, qualitative comments, or error corrections. The LLM is then refined based on this feedback.

[17] Bergmann, D. 'What IS fine-tuning?' (2024) https://www.ibm.com/think/topics/fine-tuning

[18] D.Bergman, 'What is instruction tuning?' (2024) https://www.ibm.com/think/topics/instruction-tuning

[19] S. Chaudhari et al. 'RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs' (2024) https://arxiv.org/abs/2404.08555

[20] R. Rafailov, ' Direct Preference Optimization: Your Language Model is Secretly a Reward Model' (2024) https://arxiv.org/abs/2305.18290

[21] Z.Shao, 'DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models' (2024)https://arxiv.org/abs/2402.03300

[22] Stryker, C. et al., 'What is parameter-efficient fine-tuning (PEFT)?' (2024) https://www.ibm.com/think/topics/parameter-efficient-fine-tuning

[23] AWS, 'What is RAG (Retrieval-Augmented Generation)'? (2025) https://aws.amazon.com/what-is/retrieval-augmented-generation/

[24] Wikipedia, 'Retrieval Augmented Generation' (2025) https://en.wikipedia.org/wiki/Retrieval-augmented_generation

[25] IBM, 'Retrieval Augmented Generation' (2025) https://www.ibm.com/architectures/hybrid/genai-rag?mhsrc=ibmsearch_a&mhq=RAG

[26] V.Chaba, 'Understanding the Differences: Fine-Tuning vs. Transfer Learning ' (2023) https://dev.to/luxacademy/understanding-the-differences-fine-tuning-vs-transfer-learning-370

[27] Wikipedia, 'Transfer Learning' (2025) https://en.wikipedia.org/wiki/Transfer_learning

[28] Nebuly AI, 'LLM Feedback Loop' (2024) https://www.nebuly.com/blog/llm-feedback-loop

3. Inference Phase: Generating Outputs

Once trained, the model enters the inference phase, where it generates outputs based on new inputs following these steps:

1. **Input**: The user's query is processed through tokenization and embedding, transforming it into a format the model can understand.
2. **Processing**: The input passes through the transformer architecture, where attention mechanisms and decoder layers predict the next tokens in the sequence. The decoder produces a vector of scores (called logits) for each word in the vocabulary. These scores are then passed through the Softmax[29] function, which converts them into probabilities. The model selects the most probable token as the next word in the sequence, ensuring that the generated text is coherent and contextually relevant.
3. **Output**: The model produces probabilities for potential next words, selecting the most likely options based on the input and context. These predictions are combined to generate coherent and relevant responses.

The three key stages described outline how a traditional text-only LLM is developed. Multimodal LLMs follow a similar process but to handle multiple data modalities, they incorporate specialized components such as modality-specific encoders, connectors and cross-modal fusion mechanisms to integrate the different data representations, along with a shared decoder to generate coherent outputs across modalities. Their development also involves pre-training and fine-tuning stages; however, some architectures build multimodal LLMs by fine-tuning an already pre-trained text-only LLM rather than training one from scratch.

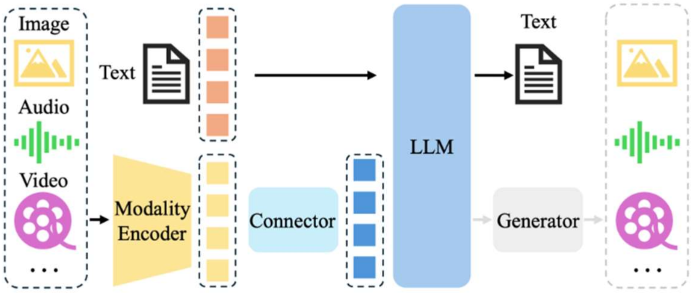


Figure 5. Typical Multimodal LLM (MLLM) architecture.
Source: Y. Shukang et al. 'A Survey on Multimodal Large Language Models' (2024)
https://arxiv.org/abs/2306.13549

In practice, LLMs are often part of a system and can be accessed directly via APIs, are embedded within SaaS platforms, deployed as off-the-shelf foundational models fine-tuned for specific use cases, or integrated into on-premise solutions. It is important to note that while LLMs are essential components of AI systems, they do not constitute AI systems on their own. For an LLM to become part of an AI system, additional components such as a user interface, must be integrated to enable it to function as a complete system.[30]Throughout this document, we will refer to such complete systems as LLM-based systems or simply LLM systems to emphasize their broader context and functionality. This distinction is crucial when assessing the risks associated with these systems, as an LLM system inherently carries more risks due to its additional components and integrations compared to a standalone LLM.

[29] Wikipedia, 'Softmax Function' (2025) https://en.wikipedia.org/wiki/Softmax_function
[30] Recital 97 AI Act

Each stage of an LLM's development lifecycle could introduce potential **privacy risks**, as the model interacts with large datasets that might contain personal data and it generates outputs based on that data. Some of the key privacy concerns may occur during:

- **The collection of data**: The training, testing and validation set could contain identifiable personal data, sensitive data or special category of data.
- **Inference**: Generated outputs could inadvertently reveal private information or contain misinformation.
- **RAG process**: We might use knowledge bases containing sensitive data or identifiable personal data without implementing proper safeguards.
- **Feedback loops**: User interactions might be stored without adequate safeguards.

Emerging LLM Technologies: The Rise of Agentic AI

According to a recent report from Deloitte,[31] by 2027, 50% of companies leveraging generative AI are expected to have launched pilots or proofs of concept to implement agentic AI systems. These systems are envisioned to function as intelligent assistants, capable of autonomously managing complex tasks with minimal human supervision.

AI Agents[32] are autonomous systems that can be built on top of LLMs and that can perform complex tasks by combining the capabilities of LLMs with reasoning, decision-making, and interaction capabilities. AI agents are proactive, capable of goal-oriented behavior such as planning, executing tasks, and iterating based on feedback. They can operate independently and are designed to achieve specific objectives by orchestrating multiple actions in sequence. They can also incorporate feedback to refine their actions or responses over time. Advanced AI agents may integrate capabilities from other AI systems, such as computer vision or audio processing, to handle diverse data inputs.[33]

The concept of agentic AI remains an evolving and not yet fully defined domain. Different organizations and researchers propose varying interpretations of what constitutes an agentic AI system. For example, at Anthropic[34], they emphasize a significant architectural distinction between workflows and agents:

- Workflows are structured systems where LLMs and tools operate in a predefined manner, following orchestrated code paths.
- Agents, in contrast, are designed to function dynamically. They allow LLMs to autonomously direct their processes and determine how to use tools and resources to achieve objectives.

An Overview of AI Agents and their Architecture[35]

In systems powered by LLMs, the LLM serves as a central "brain" providing the foundational abilities for natural language understanding and reasoning. This ability is augmented with additional components that equip the agent to plan, learn, and interact dynamically with its environment, enabling it to handle tasks that go beyond standalone LLM capabilities.

The architecture of an AI agent focuses on critical components that work together to enable sophisticated behavior and adaptability in real-world scenarios. The architecture is modular, involving distinct components for perception, reasoning, planning, memory management, and action. This

[31] J.Loucks 'Autonomous generative AI agents: Under development' (2024) https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/autonomous-generative-ai-agents-still-under-development.html

[32] C. Gadelho, 'Building AI and LLM Agents from the Ground Up: A Step-by-Step Guide' (2024) https://www.tensorops.ai/post/building-ai-and-llm-agents-from-the-ground-up-a-step-by-step-guide

[33] OpenAI's Operator (2025) https://openai.com/index/introducing-operator/

[34] Anthropic, 'Building effective agents' (2024) https://www.anthropic.com/research/building-effective-agents

[35] idem

modularity allows the system to handle complex tasks, interact dynamically with their environment, and refine performance iteratively.

Some of the most common modules currently used are:

1. Perception module
This module handles the agent's ability to process inputs from the environment and format them into a structure that the LLM can understand. It converts raw inputs (e.g., text, voice, or data streams) into embeddings or structured formats that can be processed by the reasoning module.

2. Reasoning module
The reasoning module enables the agent to interpret input data, analyze its context, and decompose complex tasks into smaller, manageable subtasks. It leverages the LLM's ability to understand and process natural language to make decisions. The reasoning mechanism enables the agent to analyze user inputs to determine the best course of action and leverage the appropriate tool or resource to achieve the desired outcome.

3. Planning module
The planning module determines how the agent will execute the subtasks identified by the reasoning module. It organizes and sequences actions to achieve a defined goal.

4. Memory and state management
To maintain context and continuity, the agent keeps track of past interactions. Memory allows the AI agent to store and retrieve context, both within a single interaction and across multiple sessions.

- Short-Term Memory: Maintains context within the current interaction to ensure coherence in responses.
- Long-Term Memory: Stores user preferences, past interactions, and learned insights for personalization.

5. Action module
This module is responsible for executing the plan and interacting with the external environment. It carries out the tasks identified and planned by earlier modules. The agent must have access to a defined set of tools, such as APIs, databases, or external systems, which it can use to accomplish the specific tasks. For example, an AI assistant might use a calendar API for scheduling or a booking service for travel reservations.

6. Feedback and iteration loop
The feedback loop enables the agent to evaluate the success of its actions and adjust its behavior dynamically. It incorporates user corrections, system logs, and performance metrics to refine reasoning, planning, and execution over time.

Interaction Between AI Agent, Memory, and Environment

The agent interacts continuously with its memory and external environment. Context from memory enhances task relevance and continuity while external data (e.g., user queries, sensor inputs) drives decision-making and task execution.

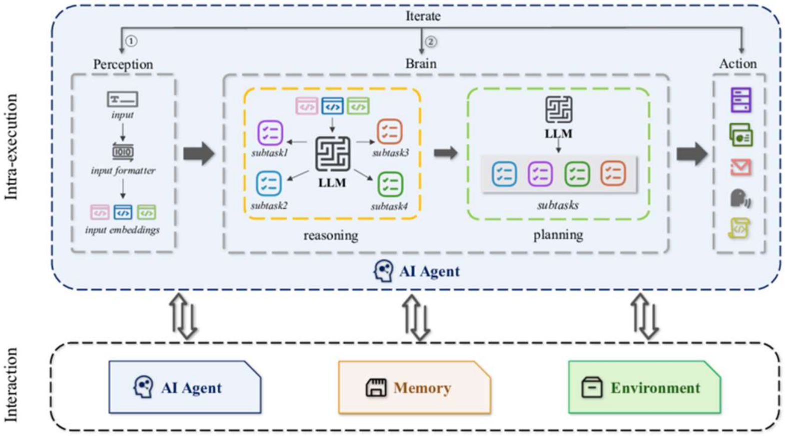


Figure 6. Source: Z.Deng et al. AI 'Agents Under Threat: A Survey of Key Security Challenges and Future Pathways' (2024) https://www.researchgate.net/figure/General-workflow-of-AI-agent-Typically-an-AI-agent-consists-of-three-components_fig1_381190070

Small Language Models (SLMs) and their role in AI Agents[36]

Small Language Models (SLMs) are lightweight, task-specific models designed to handle simpler or more focused tasks compared to Large Language Models (LLMs). While LLMs excel at understanding and generating complex language, SLMs are optimized for specific applications, such as text classification, entity recognition, or sentiment analysis.
SLMs can complement LLMs in agentic AI by taking on specialized tasks that do not require the extensive computational resources or generality of LLMs. In AI agents, SLMs can enhance efficiency and privacy by processing data locally, reducing reliance on centralized LLMs. This modular approach allows agents to allocate tasks improving overall performance and security.

To fully leverage the capabilities of LLMs within organizations, it is essential to adapt the models to the organization's specific knowledge base and business processes. This customization, often achieved by fine-tuning the LLM with organization-specific data, can result in a domain-focused small language model (SLM).[37]

Model Orchestration[38]

For agentic AI to seamlessly integrate the strengths of both SLMs[39] and LLMs, a system is needed to dynamically manage which model handles which task. This is where model orchestration plays a critical role, ensuring efficient and secure collaboration between different models. In agentic AI, orchestration determines the most appropriate model—LLM or SLM—for a given task, routes inputs accordingly, and combines their outputs into a unified response.

Privacy Concerns[40]

The growing adoption of AI agents powered by LLMs, brings the promise of revolutionizing the way humans work by automating tasks and improving productivity. However, these systems also introduce significant privacy risks that need to be carefully managed:

[36] Cabalar, R., 'What are small language models?' (2024) https://www.ibm.com/think/topics/small-language-models
[37] D. Biswas, ICAART, 'Stateful Monitoring and Responsible Deployment of AI Agents', (2025)
[38] Windland, V. et al. 'What is LLM orchestration' (2024) https://www.ibm.com/think/topics/llm-orchestration
[39] D. Vellante et al., 'From LLMs to SLMs to SAMs, how agents are redefining AI' (2024) https://siliconangle.com/2024/09/28/llms-slms-sams-agents-redefining-ai
[40] B.O'Neill, 'What is an AI agent? A computer scientist explains the next wave of artificial intelligence tools' (2024) https://theconversation.com/what-is-an-ai-agent-a-computer-scientist-explains-the-next-wave-of-artificial-intelligence-tools-242586

- To perform their tasks effectively, AI agents often require access to a wide range of user data, such as:
    - Internet activity: Browsing history, online searches, and frequently visited websites.
    - Personal applications: Emails, calendars, and messaging apps for scheduling or communication tasks.
    - Third-party systems: Financial accounts, customer management platforms, or other organizational systems.

  This level of access significantly increases the risk of unauthorized data exposure, particularly if the agent's systems are compromised.
- AI agents are designed to make decisions autonomously, which can lead to errors or choices that users may disagree with.
- Like other AI systems, AI agents are susceptible to biases originating from their training data, algorithms and usage context.

Privacy trade-offs for user convenience:[41] As AI agents grow more capable, users will need to consider how much personal data they are willing to share in exchange for convenience. For example, an agent might save time by managing travel bookings or negotiating purchases but requires access to sensitive information such as payment details or login credentials[42]. Balancing these trade-offs requires clear communication about data usage policies and robust consent mechanisms.

Accountability for Agent decisions:[43] AI agents operate in complex environments and may encounter unforeseen challenges. When an agent makes an error, or its actions cause harm, determining accountability can be difficult. Organizations must ensure transparency in how decisions are made and provide mechanisms for users to intervene when errors occur.

Common Uses of LLM Systems

LLMs have become pivotal in various industries, offering advanced capabilities in natural language understanding and generation. The market provides a spectrum of LLM solutions, each tailored to specific applications and user requirements.

1. Proprietary LLM Models

Leading technology companies have developed proprietary LLM platforms that cater to diverse business needs. Some platforms offer customizable LLMs that can be trained on specific datasets:

- OpenAI's GPT[44] Series (Generative Pre-Trained Transformer (GPT) models), are renowned for their advanced language processing capabilities. These models are accessible through APIs, enabling businesses to integrate sophisticated language understanding and generation into their applications.
- Google's Gemini[45] models are designed to assist with various tasks, providing users with detailed information and facilitating complex queries.
- Claude's Anthropic Models[46] are developed with safety and alignment in mind. Claude specializes in conversational AI with a focus on ethical and secure interactions.

Several European companies and collaborations are contributing to the LLM landscape:

[41] Z.Zhang et al. '"It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents' (2024) https://arxiv.org/abs/2309.11653

[42] Login credentials are the unique information used to access systems, accounts, or services, typically consisting of a username and password, but they can also include additional methods like two-factor authentication, biometric data, or security PINs for added protection.

[43] J. Zeiser, 'Owning Decisions: AI Decision-Support and the Attributability-Gap' (2024). https://doi.org/10.1007/s11948-024-00485-1

[44] ChatGPT (https://chatgpt.com/)

[45] Gemini (https://gemini.google.com/)

[46] Claude (https://claude.ai/)

- Mistral AI,[47]a Paris-based startup established in 2023 by former Google DeepMind and Meta AI scientists offers both open source and proprietary AI models.
- Aleph Alpha[48]is based in Heidelberg, Germany, and it specializes in developing LLMs designed to provide transparency regarding the sources used for generating results. Their models are intended for use by enterprises and governmental agencies, trained in multiple European languages.
- Silo AI's Poro[49], through its generative AI arm SiloGen, has developed Poro, a family of multilingual open source LLMs. This initiative aims to strengthen European digital sovereignty and democratize access to LLMs for all European languages.
- TrustLLM[50]is a coordinated project by Linköping University that focuses on developing trustworthy and factual LLM technology for Europe, emphasizing accessibility and reliability.
- OpenEuroLLM[51] is an open source family of performant, multilingual, large language foundation models for commercial, industrial and public services.

2. Open Source LLM Frameworks and Models

The open source community contributes significantly to the LLM landscape. Here are some of the most known frameworks and models that have shaped the development and deployment of large language models:

- Hugging Face's Transformers[52] is an extensive library of pre-trained models and tools, allowing developers to fine-tune and deploy LLMs for specific tasks.
- Deepseek[53] is an advanced language model comprising 67 billion parameters. It has been trained from scratch on a vast dataset of 2 trillion tokens in both English and Chinese.
- Deepset's Haystack[54] is an open source framework designed to build search systems and question-answering applications powered by Large Language Models (LLMs) and other natural language processing (NLP) techniques.
- OLMo 32B[55] is the first fully open model (all data, code, weights, and details are freely available).
- Meta's LLaMA[56] models focus on research and practical applications in NLP.
- BLOOM[57] was developed by BigScience as a multilingual open source model capable of generating text in over 50 languages, with a focus on accessibility and inclusivity.
- BERT[58] was created by Google to understand the context of text through bidirectional language representation, excelling in tasks like question answering and sentiment analysis.
- Falcon[59] was developed by the Technology Innovation Institute as a high-performance model optimized for text generation and understanding, with significant efficiency improvements over similar models.
- Qwen[60] is a large language model family built by Alibaba Cloud.

[47] Mistral (https://mistral.ai/)
[48] Aleph Alpha (https://aleph-alpha.com/)
[49] Silo AI, 'Poro - a family of open models that bring European languages to the frontier' (2023) https://www.silo.ai/blog/poro-a-family-of-open-models-that-bring-european-languages-to-the-frontier
[50] TrustLLM (https://trustllm.eu/)
[51] OpenEuroLLM (https://openeurollm.eu/)
[52] Hugging Face, 'Transformers' (n.d) https://huggingface.co/docs/transformers/v4.17.0/en/index
[53] Deepseek (https://www.deepseek.com/)
[54] Haystack (https://haystack.deepset.ai/)
[55] Ai2, 'OLMo 2 32B: First fully open model to outperform GPT 3.5 and GPT 4o mini' (2025) https://allenai.org/blog/olmo2-32B
[56] Llama (https://www.llama.com/)
[57] Hugging Face, 'Introducing The World's Largest Open Multilingual Language Model: BLOOM' (2025) https://bigscience.huggingface.co/blog/bloom
[58] Hugging Face, 'BERT' (n.d) https://huggingface.co/docs/transformers/model_doc/bert
[59] TTI, 'Introducing the Technology Innovation Institute's Falcon 3' (n.d) https://falconllm.tii.ae/
[60] Hugging Face, ''Qwen' (n.d) https://huggingface.co/Qwen

- LangChain[61] is an open source framework for building applications powered by large language models.

3. Cloud-Based LLM Services

Major cloud providers offer LLM services that integrate seamlessly into existing infrastructures providing access to proprietary and open source LLMs:

- Microsoft Azure OpenAI[62] Service collaborates with OpenAI to provide API access to GPT models, enabling businesses to incorporate advanced language features into their applications.
- Amazon Web Services (AWS) Bedrock[63] offers a suite of AI services, including language models that support various natural language processing tasks.
- Google Cloud Vertex AI[64] is a platform for building, deploying, and scaling machine learning models, including LLMs. It provides access to models like PaLM 2 and supports customization for various applications, such as translation, summarization, and conversational AI.
- IBM Watson[65] provides LLM capabilities that can be tailored to recognize industry-specific entities, enhancing the relevance and accuracy of information extraction.
- Cohere[66] offers customizable LLMs that can be fine-tuned for specific tasks.

Applications of LLMs

LLMs are employed across various applications[67], enhancing both user experience and operational efficiency. This list represents some of the most prominent applications of LLMs, but it is by no means exhaustive. The versatility of LLMs continues to unlock new use cases across industries, demonstrating their transformative potential in various domains.

- **Chatbots and AI Assistants:[68]** LLMs power virtual assistants like Siri, Alexa, and Google Assistant, understand and process natural language, interpret user intent, and generate responses.
- **Content generation:[69]** LLMs assist in creating articles, reports, and marketing materials by generating human-like text, thereby streamlining content creation processes.
- **Language translation:[70]** Advanced LLMs facilitate real-time translation services.
- **Sentiment analysis:[71]** Businesses use LLMs to analyze customer feedback and social media content, gaining insights into public sentiment and informing strategic decisions.
- **Code generation and debugging:[72]** Developers leverage LLMs to generate code snippets and identify errors, enhancing software development efficiency.
- **Educational support tools:[73]** LLMs play a key role in personalized learning by generating educational content, explanations, and answering student questions.
- **Legal document processing:[74]** LLMs help professionals in the legal field by reviewing and summarizing legal texts, extracting important information, and offering insights.

[61] LangChain, 'Introduction' (n.d) https://python.langchain.com/

[62] Microsoft, 'Azure OpenAI Service' (2025) https://azure.microsoft.com/en-us/services/cognitive-services/openai-service/

[63] AWS, 'Bedrock' (n.d)https://aws.amazon.com/bedrock

[64] Vertex AI Platform, 'Innovate faster with enterprise-ready AI, enhanced by Gemini models' (n.d) https://cloud.google.com/vertex-ai

[65] IBM, 'IBM Watson to watsonx' (n.d) https://www.ibm.com/watson

[66] Cohere (https://cohere.com/)

[67] N. Sashidharan, 'Three Pillars of LLM: Architecture, Use Cases, and Examples ' (2024) https://www.extentia.com/post/pillars-of-llm-architecture-use-cases-and-examples

[68] Google Assistant (https://assistant.google.com/)

[69] Jasper AI (https://www.jasper.ai/)

[70] Deepl (https://www.deepl.com/en/translator)

[71] SurveySparrow (https://surveysparrow.com/features/cognivue/)

[72] GitHub Copilot (https://github.com/features/copilot)

[73] Khanmigo (https://www.khanmigo.ai/)

[74] Luminance (https://www.luminance.com/)

- **Customer support:**[75] Automating responses to customer inquiries and escalating complex cases to human agents.
- **Autonomous vehicles:**[76] Driving cars with real-time decision-making capabilities.

Performance Measures for LLMs

Evaluating the performance of Large Language Models (LLMs) is essential to ensure they meet their intended purpose and desired standards of accuracy, reliability, and ethical use across diverse applications. To effectively measure the performance of a Large Language Model (LLM), it is important to tailor the evaluation approach to the stage of the LLM lifecycle (e.g., training, post-processing, pre-deployment, production) and its intended real-world applications. Performance metrics help identify areas where additional testing or refinements may be necessary before deployment or once the LLM system is in use in a production environment.

Some of the most common LLM performance evaluation criteria are Answer Relevancy, Correctness, Semantic Similarity, Fluency, Hallucination, Factual Consistency, Contextual Relevancy, Toxicity, Bias and Task-Specific Metrics.

The following metrics[77] are commonly used, each offering different insights:

- **Accuracy**[78]measures how often an output aligns with the correct or expected results. In tasks like text classification or question answering, accuracy is calculated as the ratio of correct predictions to the total number of predictions. However, for generative tasks such as text generation, traditional accuracy metrics may not fully capture performance due to the open-ended nature of possible correct responses. In such cases, metrics like BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) are employed to assess the quality of generated text by comparing it to reference texts.
- **Precision** quantifies the ratio of correctly predicted positive outcomes to the total number of positive predictions made by the model. In the context of LLMs, a high precision score indicates the model is accurate when making predictions. However, it does not account for relevant instances the model fails to predict (false negatives), so it is commonly combined with recall for a more comprehensive evaluation.
- **Recall**, also referred to as sensitivity or the true positive rate, measures the proportion of actual positive instances that the model successfully identifies. A high recall score reflects the model's effectiveness in capturing relevant information but does not address irrelevant predictions (false positives). For this reason, recall is typically evaluated alongside precision to provide a balanced view.
- **F1 Score** offers a balanced metric by combining precision and recall into their harmonic mean. A high F1 score indicates that the model achieves a strong balance between precision and recall, making it a valuable metric when both false positives and false negatives are critical. The F1 score ranges from 0 to 1, with 1 representing perfect performance on both metrics.
- **Specificity**[79] measures the proportion of true negatives correctly identified by a model.
- **AUC** (Area Under the Curve) and **AUROC**[80](Area Under the Receiver Operating Characteristic Curve) quantify a model's ability to distinguish between classes. It evaluates the trade-off between

[75] Salesforce (https://www.salesforce.com/eu/)

[76] Tesla Autopilot (https://www.tesla.com/autopilot)

[77] A. Chaudhary, 'Understanding LLM Evaluation and Benchmarks: A Complete Guide' (2024) https://www.turing.com/resources/understanding-llm-evaluation-and-benchmarks

[78] S. Karzhev, 'LLM Evaluation: Metrics, Methodologies, Best Practices' (2024) https://www.datacamp.com/blog/llm-evaluation

[79] Wikipedia, 'Sensitivity and Specificity' (2025) https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[80] E.Becker and S.Soatto, 'Cycles of Thought: Measuring LLM Confidence through Stable Explanations' (2024) https://arxiv.org/pdf/2406.03441v1

sensitivity (true positive rate) and 1-specificity (false positive rate) across various thresholds. A higher AUC value indicates better performance in classification tasks.

- **AUPRC**[81](Area Under the Precision-Recall Curve), measures a model's performance in imbalanced datasets, focusing on the trade-off between precision and recall. A high AUPRC indicates that the model performs well in identifying positive instances, even when they are rare.
- **Cross Entropy**[82]is a measure of uncertainty or randomness in a system's predictions. It measures the difference between two probability distributions: the true labels (actual data distribution) and the predicted probabilities from the model (output). Lower entropy means higher confidence in predictions, while higher entropy indicates uncertainty.
- **Perplexity**[83]derives from cross entropy and evaluates how well a language model predicts a sample, serving as an indicator of its ability to handle uncertainty. A lower perplexity score means better performance, indicating that the model is more confident in its predictions. Some studies suggest that perplexity has proven unreliable[84] to evaluate LLMs due to their long-context capabilities. It is also difficult to use perplexity as a benchmark between models since its scores depend on factors like tokenization method, dataset, preprocessing steps, vocabulary size, and context length.[85]
- **Calibration**[86]refers to the alignment between a model's predicted probabilities and the actual probability of those predictions being correct. A well-calibrated model provides confidence scores that accurately reflect the true probabilities of outcomes. Proper calibration is vital in applications where understanding the certainty of predictions is important, such as in medical diagnoses or legal document analysis.
- **MoverScore**[87]is a modern metric developed to assess the semantic similarity between two texts.

Other metrics used for assessing the performance and usability of LLM-based systems, especially in real-time or high-demand applications are:[88]

- **Completed requests per minute:** Measures how many requests the LLM can process and return responses for in one minute. It reflects the system's efficiency in handling multiple queries.
- **Time to first token (TTFT):** The time taken from when a request is submitted to when the first token of the response is generated.
- **Inter-token Latency (ITL):** The time delay between generating consecutive tokens in the response. This metric evaluates the speed and fluidity of text generation.
- **End to end Latency /ETEL):** The total time taken from when a request is made to when the entire response is completed. It encompasses all processing stages, including input handling, model inference, and output generation.

[81] J. Czakon, 'F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose?' (2024) https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

[82] C.Xu, ' Understanding the Role of Cross-Entropy Loss in Fairly Evaluating Large Language Model-based Recommendation' (2024) https://arxiv.org/pdf/2402.06216v2

[83] C.Huyen 'Evaluation Metrics for Language Modeling' (2019) https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

[84] L.Fang 'What is wrong with perplexity for long-context language modeling?' (2024) https://arxiv.org/pdf/2410.23771v1

[85] A.Morgan 'Perplexity for LLM Evaluation' (2024) https://www.comet.com/site/blog/perplexity-for-llm-evaluation/

[86] P.Liang et al. 'Holistic Evaluation of Language Models' (2023) https://arxiv.org/abs/2211.09110

[87] PI, 'Moverscore 1.0.3' (2020) https://pypi.org/project/moverscore/

[88] W. Kadous et al. 'Reproducible Performance Metrics for LLM inference' (2023) https://www.anyscale.com/blog/reproducible-performance-metrics-for-llm-inference

In addition to these metrics, there are comprehensive evaluation frameworks or **benchmarks**[89] such as GLUE (General Language Understanding Evaluation)[90], MMLU (Massive Multitask Language Understanding)[91], HELM (Holistic Evaluation of Language Models)[92], DeepEval[93] or OpenAI Evals[94]. Task-specific metrics such as BLEU[95] (Bilingual Evaluation Understudy), ROUGE[96] (Recall-Oriented Understudy for Gisting Evaluation), and BLEURT[97] (Bilingual Evaluation Understudy with Representations from Transformers) are widely used for evaluating text generation, summarization, and translation.

It is important to recognize that quantitative metrics alone are not sufficient. While these metrics are highly valuable in identifying risks, especially when integrated into automated evaluation pipelines, they primarily serve as early warning signals, prompting further investigation when thresholds are exceeded. Many critical risks, including misuse potential, ethical concerns, and long-term impact, cannot be effectively captured through those numerical measurements alone.
To ensure a more holistic evaluation, organizations should complement quantitative indicators with expert judgment, scenario-based testing, and qualitative assessments.
Open source frameworks like Inspect[98], support an integrated approach by enabling model-graded evaluations, prompt engineering, session tracking, and extensible scoring techniques. These tools help operationalize both metric-based and qualitative evaluations, offering better observability and insight into LLM behavior in real-world settings.

Measuring Performance in Agentic AI

Most current metrics[99] for AI agents focus on efficiency, effectiveness, and reliability. These include system metrics (resource consumption and technical performance), task completion (measuring goal achievement), quality control (ensuring output consistency), and tool interaction (evaluating integration with external tools and APIs).

Some of the key metrics used include:

- **Task-specific accuracy:**[100] Assesses how correctly the agent performs designated tasks, such as classification or information retrieval. Metrics like Exact Match (EM) and F1 Score are commonly used.
- **End-to-end task completion:**[101] Evaluates the agent's ability to achieve user-defined goals through a series of actions. Metrics include Task Success Rate (TSR) and Goal Completion Rate (GCR).
- **Step-Level accuracy**: Assesses the correctness of individual actions taken by the agent within a larger workflow. This is critical in multi-step processes, such as booking a service or resolving a technical issue.

[89] Benchmarks are standardized frameworks developed to assess LLMs across various scenarios and metrics (See also section 10 of this document).
[90] Gluebenchmark (https://gluebenchmark.com/)
[91] Papers with code, 'MMLU (Massive Multitask Language Understanding)' (n.d) https://paperswithcode.com/dataset/mmlu
[92] Center for Research on Foundation Models, 'A reproducible and transparent framework for evaluating foundation models' (n.d) https://crfm.stanford.edu/helm/
[93] GitHub, 'The LLM Evaluation framework' (n.d) https://github.com/confident-ai/deepeval
[94] GitHub, 'Evals is a framework for evaluating LLMs and LLM systems, and an open-source registry of benchmarks' (n.d) https://github.com/openai/evals
[95] Wikipedia, 'BLEU' (2025) https://en.wikipedia.org/wiki/BLEU
[96] Wikipedia, 'ROUGE(metric)' (2025) https://en.wikipedia.org/wiki/ROUGE_(metric)
[97] GitHub, 'BLEURT is a metric for Natural Language Generation based on transfer learning' (n.d) https://github.com/google-research/bleurt
[98] AISI, 'An open-source framework for large language model evaluations' (n.d) https://inspect.aisi.org.uk/
[99] P. Bhavsar 'Mastering Agents: Metrics for Evaluating AI Agents' (2024) https://www.galileo.ai/blog/metrics-for-evaluating-ai-agents
[100] https://smythos.com/ai-agents/impact/ai-agent-performance-measurement/
[101] AISERA, 'An Introduction to Agent Evaluation' (n.d) https://aisera.com/blog/ai-agent-evaluation/

- **Precision and Recall**: Measures how accurately the agent retrieves relevant information (precision) and whether it captures all necessary details (recall). These metrics are vital for tasks like document summarization or answering complex queries.
- **Contextual understanding:**[102]Measures the agent's proficiency in maintaining and utilizing context in interactions, crucial for coherent multi-turn dialogues. Dialog State Tracking[103] is a relevant metric.
- **User satisfaction:**[104]Measures user perceptions of the agent's performance, often through feedback scores or surveys and using scales to measure system and user experience usability.

Evaluating AI agents with traditional LLM benchmarks presents challenges, as they often fail to capture real-world dynamics, multi-step reasoning, tool use, and adaptability. Effective assessment requires new benchmarks that measure long-term planning, interaction with external tools, and real-time decision-making. Below are some of the most recognized benchmarks currently used:

- **SWE-bench:**[105]Software Engineering Benchmark dataset, created to systematically evaluate the capabilities of an LLM in resolving software issues.
- **AgentBench:**[106][107] It is designed for evaluating and training visual foundation agents based on LMMs.
- **MLAgentBench:**[108] To evaluate if agents driven by LLMs perform machine learning experimentation effectively.
- **BFCL (Berkeley Function-Calling Leaderboard):**[109] To evaluate the ability of different LLMs to call functions (also referred to as tools).
- **τ-bench:**[110] A benchmark for tool-agent-user interaction in real-world domains.
- **Planbench:**[111] To evaluate LLMs on planning and reasoning.

Issues that can Affect the Accuracy of the Output

Several factors can impact the accuracy of the outputs generated by LLMs. Understanding these issues is essential for optimizing their performance and mitigating risks in practical applications. Some of the more common issues are:

1. Quality of training data

- **Data bias:**[112]If the training data contains biases (e.g., societal, cultural, or linguistic biases), the model may replicate or amplify these biases in its outputs.
- **Data relevance:**[113]Training on outdated, irrelevant, or noisy data can lead to inaccurate or contextually irrelevant responses.

[102] Smyth OS, 'Conversational Agents and Context Awareness: How AI Understands and Adapts to User Needs' (n.d) https://smythos.com/artificial-intelligence/conversational-agents/conversational-agents-and-context-awareness/

[103] Papers with code, ' Dialogue State Tracking', (n.d) https://paperswithcode.com/task/dialogue-state-tracking/codeless?page=2

[104] N. Bekmanis, 'Artificial Intelligence Conversational Agents: A Measure of Satisfaction in Use' (2023) https://essay.utwente.nl/94906/1/Bekmanis_MA_BMS.pdf

[105] Swebench (https://www.swebench.com/)

[106] Github, 'A Comprehensive Benchmark to Evaluate LLMs as Agents (ICLR'24)', (n.d) https://github.com/THUDM/AgentBench

[107] Papers with code, 'Agentench' (n.d) https://paperswithcode.com/dataset/agentbench

[108] Q.Huang et al. 'MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation' (2024) https://arxiv.org/abs/2310.03302

[109] Hugging Face Dataset (https://huggingface.co/datasets/gorilla-llm/Berkeley-Function-Calling-Leaderboard)

[110] GitHub, 'Code and Data' (n.d) https://github.com/sierra-research/tau-bench

[111] GitHub, 'An extensible benchmark for evaluating large language models on planning' (n.d) https://github.com/karthikv792/LLMs-Planning
I. O. Gallegos et al. 'Bias and Fairness in Large Language Models: A Survey' (2024) https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A

[113] 'Large Language Models pose risk to science with false answers, says Oxford study' (2023) https://www.ox.ac.uk/news/2023-11-20-large-language-models-pose-risk-science-false-answers-says-oxford-study

2. Model limitations

- **Understanding context:**[114]Despite advanced architectures, LLMs can struggle with nuanced contexts or multi-turn conversations where earlier parts of the dialogue must inform later responses.
- **Handling ambiguities:**[115]Ambiguous input can lead to incorrect or nonsensical outputs if the model cannot infer the intended meaning.

3. Tokenization and preprocessing

- **Tokenization errors:**[116]Misrepresentation of input text due to tokenization issues (e.g., splitting words incorrectly) can distort model understanding.
- **Preprocessing issues:**[117]Overly aggressive cleaning or normalization during preprocessing can remove important contextual information, reducing accuracy.

4. Overfitting and underfitting

- **Overfitting:**[118]Training for too many iterations on a limited dataset can make the model overly specialized, leading to poor performance on unseen data.
- **Underfitting:**[119]Inadequate training or overly simple models may fail to capture the complexity of the task, resulting in general inaccuracies.

5. Prompt design and input quality

- **Prompt sensitivity:**[120]LLMs are highly sensitive to how inputs are phrased. Minor variations in prompt structure can lead to drastically different outputs.
- **Garbage in, garbage out:**[121]Poorly worded or unclear input can lead to inaccurate or irrelevant responses.

6. Limitations in knowledge

- **Knowledge cutoff:**[122]LLMs are trained on data up to a specific point in time. They may lack awareness of recent developments or emerging knowledge.
- **Factual errors:** [123] LLMs can "hallucinate" information, generating plausible but factually incorrect responses due to the probabilistic nature of their predictions.

7. Lack of robustness

- **Adversarial inputs:** [124] LLMs may fail when presented with deliberately manipulated or adversarial inputs designed to exploit their weaknesses.

[114] J. Browning, 'Getting it right: the limits of fine-tuning large language models' (2024) https://link.springer.com/article/10.1007/s10676-024-09779-1

[115] E.Jones and J. Steinhardt, 'Capturing Failures of Large Language Models via Human Cognitive Biases' (2022) https://arxiv.org/abs/2202.12299

[116] G.B.Mohan et al. ' An analysis of large language models: their impact and potential application' (2024) https://link.springer.com/article/10.1007/s10115-024-02120-8

[117] H.Naveed et al 'A Comprehensive Overview of Large Language Models' (2024) https://arxiv.org/abs/2307.06435

[118] P.Jindal 'Evaluating Large Language Models: A Comprehensive Guide' (2024) https://www.labellerr.com/blog/evaluating-large-language-models

[119] idem

[120] J.Browning 'Getting it right: the limits of fine-tuning large language models' (2024) https://link.springer.com/article/10.1007/s10676-024-09779-1

[121] H.Naveed et al. 'A Comprehensive Overview of Large Language Models' (2024) https://arxiv.org/abs/2307.06435

[122] University of Oxford, 'Large Language Models pose risk to science with false answers, says Oxford study' (2023) https://www.ox.ac.uk/news/2023-11-20-large-language-models-pose-risk-science-false-answers-says-oxford-study

[123] ht Ho, D.E., 'Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive'(2024) https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

[124] E.Jones and J.Steinhardt, 'Capturing Failures of Large Language Models via Human Cognitive Biases' (2022) https://arxiv.org/abs/2202.12299

- **Noise and variability:** [125] Spelling errors, slang, or non-standard language can lead to misinterpretations and lower accuracy.

8. Inadequate calibration

- **Overconfidence:**[126] Poorly calibrated models may assign high confidence scores to incorrect predictions, misleading users. Failing to properly convey uncertainty in predictions can erode trust in the model.

[125] G.B.Mohan 'An analysis of large language models: their impact and potential applications' (2024) https://link.springer.com/article/10.1007/s10115-024-02120-8

[126] L.Li et al. 'Confidence Matters: Revisiting Intrinsic Self-Correction Capabilities of Large Language Models' (2024) https://arxiv.org/abs/2402.12563

3. Data Flow and Associated Privacy Risks in LLM Systems

Understanding the data flow in AI systems powered by LLMs is crucial for assessing privacy risks. This flow may vary depending on the phases of operation, the specific system the model integrates, and the type of service model in use, each of which introduces unique challenges for data protection.

The Importance of the AI Lifecycle in Privacy Risk Management

The lifecycle of an AI system, as outlined in standards ISO/IEC 22989[127]and ISO/IEC 5338,[128]provides a structured framework for understanding the flow of data throughout the development, deployment, and operation of AI systems. This lifecycle is also essential for identifying and mitigating privacy risks at each stage.

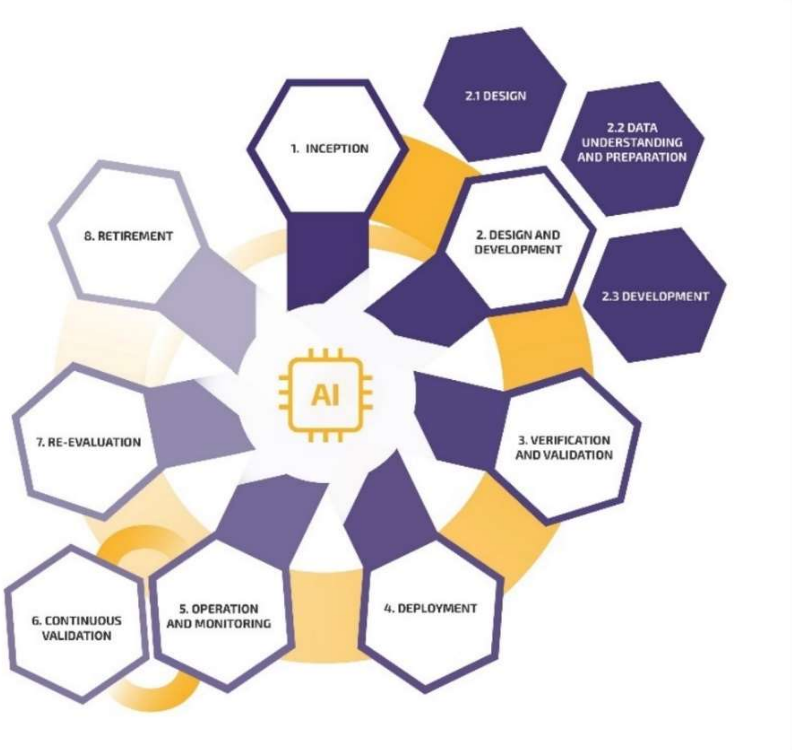

1. INCEPTION
2.1 DESIGN
2.2 DATA UNDERSTANDING AND PREPARATION
2. DESIGN AND DEVELOPMENT
2.3 DEVELOPMENT
8. RETIREMENT
AI
7. RE-EVALUATION
3. VERIFICATION AND VALIDATION
6. CONTINUOUS VALIDATION
5. OPERATION AND MONITORING
4. DEPLOYMENT

Figure 7. Source: Based on ISO/IEC 22989

In this document, we use this AI lifecycle as a reference framework, recognizing that each organization may have its own adapted version based on its specific needs. While the core stages of the lifecycle are generally similar across organizations, the exact phases may vary.

Each one of the phases of the lifecycle involves unique privacy risks that require tailored mitigation strategies. Implementing Privacy by Design into each phase helps to address risks proactively rather than retroactively fixing them.

[127] ISO/IEC 22989 (Artificial Intelligence – Concepts and Terminology)

[128] ISO/IEC 5338:2023 Information technology — Artificial intelligence — AI system life cycle processes

AI Lifecycle Phases and their Impact on Privacy

1. **Inception and Design**: In this phase, decisions are made regarding data requirements, collection methods, and processing strategies. The selection of data sources may introduce risks if sensitive or personal data is included without adequate safeguards.
2. **Data Preparation and Preprocessing**: Raw data is collected, cleaned, in some cases anonymized[129], and prepared for training or fine-tuning. Datasets are often sourced from diverse origins, including web-crawled data, public repositories, proprietary data, or datasets obtained through partnerships and collaborations.
    - Privacy risks**:**
        - Training data may inadvertently include personal details, confidential documents, or other sensitive information.
        - Inadequate anonymization or handling of identifiable data can lead to breaches or unintended inferences during later stages.
        - Biases present in the datasets can affect the model's predictions, resulting in unfair or discriminatory outcomes.
        - Errors or gaps in training data can adversely impact the model's performance, reducing its effectiveness and reliability.
        - The collection and use of training data may violate privacy rights, lack proper consent, or infringe on copyrights and other legal obligations.
3. **Development, Model Training**: Prepared datasets are used to train the model, which involves large-scale processing. The model may inadvertently memorize sensitive data, leading to potential privacy violations if such data is exposed in outputs.
4. **Verification & Validation:**[130] The model is evaluated using test datasets, often including real-world scenarios. Testing data may inadvertently expose sensitive user information, particularly if real-world datasets are used without anonymization.
5. **Deployment**: The model interacts with live data inputs from users, often in real-time applications that could integrate with other systems. Live data streams might include highly sensitive information, requiring strict controls on collection, transmission, and storage.
6. **Operation and Monitoring**: Continuous data flows into the system for monitoring, feedback, and performance optimization. Logs from monitoring systems may retain personal data such as user interactions, creating risks of data leaks or misuse.
7. **Re-evaluation, Maintenance and Updates**: Additional data may be collected for retraining or updating the model to improve accuracy or address new requirements. Using live user data for updates without proper consent or safeguards can violate privacy principles.
8. **Retirement**: Data associated with the model and its operations is archived or deleted. Failure to properly erase personal data during decommissioning can lead to long-term privacy vulnerabilities.

Throughout the AI system lifecycle, it is important to consider how different types of personal data may be involved at each phase. Depending on the stage, personal data can be collected, processed, exposed, or transformed in different ways. Recognizing this variability is essential for implementing effective privacy and data protection measures.

[129] Important to consider the EDPB opinion 28/2024 and section 3.2 On the circumstances under which AI models could be considered anonymous and the related demonstration: '…, the EDPB considers that, for an AI model to be considered anonymous, using reasonable means, both (i) the likelihood of direct (including probabilistic) extraction of personal data regarding individuals whose personal data were used to train the model; as well as (ii) the likelihood of obtaining, intentionally or not, such personal data from queries, should be insignificant for any data subject.'

[130] Testing, Evaluation, Validation, and Verification (TEVV) is an ongoing process that occurs throughout the AI lifecycle to ensure that a system meets its intended requirements, performs reliably, and aligns with safety and compliance standards.

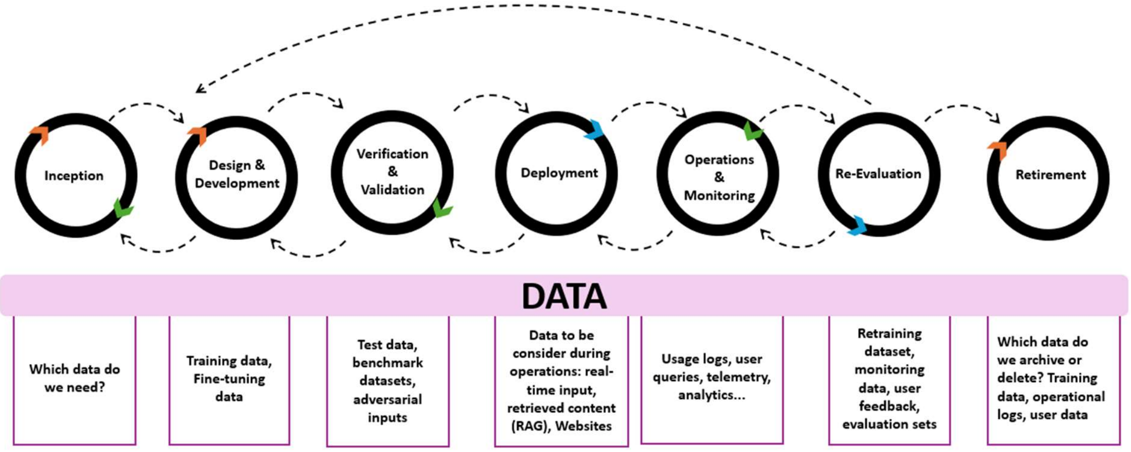

Inception
Design & Development
Verification & Validation
Deployment
Operations & Monitoring
Re-Evaluation
Retirement
DATA
Which data do we need?
Training data, Fine-tuning data
Test data, benchmark datasets, adversarial inputs
Data to be consider during operations: real-time input, retrieved content (RAG), Websites
Usage logs, user queries, telemetry, analytics...
Retraining dataset, monitoring data, user feedback, evaluation sets
Which data do we archive or delete? Training data, operational logs, user data

Figure 8. The illustration shows how different types of personal data can arise across various phases of the AI lifecycle.

Data Flow and Privacy Risks per LLM Service Model

It is common to encounter terms like closed models, open models, closed weights, and open weights in the context of LLMs. Understanding these terms is essential for assessing the risks associated with different model release strategies.

Closed models are proprietary models that do not provide public access to their weights or source code and interaction with the model is restricted, typically requiring an API or subscription, while **open models** are made publicly available **fully** (weights, full code, training data, and other documentation is available) or **partly** (not everything is available, usually training data; or it is available under licences). Similarly, **closed weights** indicate proprietary models whose trained parameters are not disclosed, whereas **open weights** describe models with publicly available parameters, allowing for inspection, fine-tuning, or integration into other systems.

It is also important to distinguish the term open model from **open source model**. This classification of a model as **"open source"** requires it to be released under an **open source license**, which legally grants anyone the freedom to **use, study, modify, and distribute** the model for any purpose[131].

Term	Privacy Risks
Closed models & closed weights	Often minimal external transparency. Users rely entirely on the provider's privacy safeguards, making it difficult to independently verify compliance with data protection regulations.
Open models & open weights	Risk of personal data exposure and security breaches if training data contains sensitive or harmful content. Partial access may prevent full scrutiny of model training data and privacy vulnerabilities.
Open source	Open source models share the same privacy risks as open models and open weight models. While open source fosters transparency and innovation, it also increases risks, as modifications may introduce security vulnerabilities or remove built-in safety measures.

LLMs are predominantly accessible through the following service models:

[131] AI Action Summit, 'International AI Safety Report on the Safety of Advanced AI' , p 150, (2025) https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

1. **LLM as a Service:** This service model provides access to LLMs via APIs hosted on a cloud platform. Users can send input and receive output without having direct access to the model's underlying architecture or weights.
   Based on this service model we can usually find the different LLM model variations available:
   - Closed models with closed weights where the provider trains the model and retains control over the weights and data, offering access through an API. This approach ensures ease of use but requires user data to flow through the provider's systems. Example: OpenAI GPT-4 API[132]
   - Customizable closed weights where deployers may fine-tune the model using their own data, within a controlled environment, although the underlying weights remain inaccessible balancing customization with security. Example: Azure OpenAI Service[133]
   - Open weights where some providers grant deployers full or partial access to the architecture for greater transparency and flexibility through a platform or via an API[134]. Example: Hugging Face's models in AWS Bedrock[135]

2. LLM 'off-the-shelf': In this service model the deployer can customize weights and fine tune the model. This happens sometimes through platforms like Microsoft Azure and AWS where a deployer can select a model and develop their own solution with it. It is also commonly used with open weight models, such as LLaMA or BLOOM. While an LLM as a Service typically involves API-based interaction without model ownership, the LLM 'off-the-Shelf' service emphasizes more developer and deployer control. The distinction lies in this level of control and access provided, for instance, in Hugging Face models can be downloaded locally.

3. Self-developed LLM: In this model, organizations develop and deploy LLMs on their own infrastructure, maintaining full control over data and model interaction. While this option may offer more privacy, this service model requires of significant computational resources and expertise.

Each of the three service models features a distinct data flow. While there are similarities across models, each phase—from user input to output generation—presents unique risks that can impact user privacy and data protection. In this section, we will first examine the data flow in an LLM as a Service solution, followed by an analysis of the key differences in data flow when using an LLM 'off-the-shelf' model and a self-developed LLM system.

*Note that in this section, the terms 'provider'[136] and 'deployer'[137] are used as defined in the AI Act, where the provider refers to the entity developing and offering the AI system, and the deployer refers to the entity implementing and operating the system for end-users.

[132] Open AI, 'The most powerful platform for building AI products, (2025) https://openai.com/api/

[133] Microsoft, 'Azure OpenAI Service' (2025) https://azure.microsoft.com/en-us/services/cognitive-services/openai-service/

[134] Wikipedia, 'API' (2025)https://en.wikipedia.org/wiki/API

[135] S.Pagezy, 'Use Hugging Face models with Amazon Bedrock' (2024) https://huggingface.co/blog/bedrock-marketplace

[136] 'provider' means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge; (Article 3 (3) AI Act)

[137] 'deployer' means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity; (Article 3 (4) AI Act)

1. Data Flow in a LLM as a Service System

In our example, a user interacts with an LLM application hosted online by a provider. This data flow focuses solely on the phases involved during the user's interaction with the service. Model preparation, deployment, and integration by the provider are outside the scope since they will be examined further in the self-developed LLM system example. It is important to note that each use case will have its own specific data flow depending on its unique requirements and context and the examples provided in this section are intended to be generic representations.
In an LLM as a Service scenario we could find these general data flow phases:

- **User input:**

The process starts with the user submitting input, such as a query or command. This could be entered through a web-based interface, mobile application, or other tools provided by the LLM provider.

- **Provider interface & API:**

The input is sent through an interface or application managed by the provider (e.g., a webpage, app or a chatbot window embedded on a website). This interface ensures the input is formatted appropriately and securely transmitted to the LLM infrastructure.

- **LLM processing at providers' infrastructure:**

The API receives the input and routes it to the LLM model hosted on the provider's infrastructure.
The LLM processes the input using its trained parameters (weights) to generate a relevant response. This may involve steps like tokenization, context understanding, reasoning, and text generation. The model generates a response.
* Logging: The provider may log the user input (query) along with the generated response to analyze the interaction and identify system errors or gaps in response quality.
The data could be also included in a training dataset to improve the model's ability to handle similar queries in the future. In this case, anonymization and filtering techniques are often applied.

- **Processed output:**

The generated output is returned via the provider's interface to the user. The response is typically in a format ready for display or integration, such as text, suggestions, or actionable data.

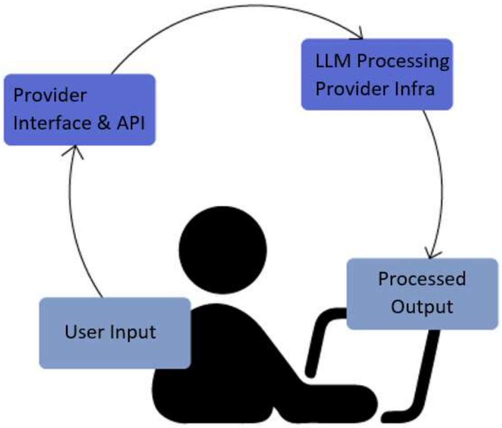


Figure 9. Data Flow in a LLM as a Service System

Privacy considerations in this data flow
The following table highlights potential privacy and data protection risks and their recommended mitigations.

Phases	Possible Risks & Mitigations
User input	**Risks:** **Sensitive data disclosure**: Users may unknowingly or inadvertently input sensitive personal data, such as names, addresses, financial information, or medical details. **Unauthorized access**: If the web interface, application, databases or input tool lacks robust access controls, unauthorized individuals[138] may gain access to user accounts or systems, allowing them to view previously submitted data or queries. **Lack of transparency**: Users may not be fully aware of how their data will be used, retained, or shared by the provider. **Adversarial attacks:** A (malicious) user might craft input designed to manipulate the LLM's behavior or bypass its intended functionality, such as injecting unauthorized instructions into queries (prompt injection attack), or users may try to bypass safety restrictions[139] imposed on the model by crafting specific input (jailbreaking attempt)[140]. **Mitigations for Providers:** - Implement clear user guidance and input restrictions, such as filters or warnings to discourage the entry of personal data. Use automated detection mechanisms[141][142] to flag or anonymize sensitive information before it is processed or logged. - Encrypt user inputs and outputs during both transmission and storage to protect sensitive data from unauthorized access. Insufficient encryption can expose user queries and data to potential breaches, especially during the user input phase. Ensure encryption in transit using robust protocols (e.g., TLS) and encryption at rest for stored data. Additionally, implement data segregation practices to isolate user data, preventing unauthorized individuals from accessing or compromising multiple accounts or datasets. Another mitigation to prevent unauthorized access is the implementation of secure password practices based on the latest NIST[143] and ENISA[144] recommendations: require a minimum password length of 8 characters, update passwords if they are compromised or forgotten, enforce the use of multifactor authentication (MFA), ensure passwords differ significantly from previous ones, check passwords against blacklists, enforce account lockout policies, monitor failed login attempts, discourage the use of password hints, and store passwords securely using hashing and salting techniques with robust algorithms such as bcrypt, Argon2, or PBKDF2. - Inform users about how their data will be used, retained, and processed through clear and easily accessible privacy policies. - Though there are currently no foolproof measures[145] to protect against prompt injection and jailbreaking,[146][147]some of the most common best practices include the

[138] G. Nagli 'Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History' (2025) https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak

[139] T.S. Dutta 'New Jailbreak Techniques Expose DeepSeek LLM Vulnerabilities, Enabling Malicious Exploits' (2025) https://cybersecuritynews.com/new-jailbreak-techniques-expose-deepseek-llm-vulnerabilities/

[140] S.Schulhoff 'Prompt Injection vs. Jailbreaking: What's the Difference?' (2024) https://learnprompting.org/blog/injection_jailbreaking

[141] https://www.nightfall.ai/ai-security-101/data-leakage-prevention-dlp-for-llms

[142] Some of the tools used are Google Cloud DLP, Microsoft Presidio, OpenAI Moderation API, Hugging Face Fine-Tuned NER Models and spaCy (links available in section 10)

[143] P.A. Grassi et al., (2017) NIST Special Publication 800-63-3 Digital Identity Guidelines https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-3.pdf

[144] ENISA, 'Basic security practices regarding passwords
and online identities' (2014) https://enisa.europa.eu/sites/default/files/all_files/ENISA%20guidelines%20for%20passwords.pdf

[145] Kosinski, M., 'How to prevent prompt injection attacks' (2024) https://www.ibm.com/think/insights/prevent-prompt-injection

[146] A.Peng et al. 'Rapid Response: Mitigating LLM Jailbreaks with a Few Examples' (2024) https://arxiv.org/abs/2411.07494

[147] B.Peng et al. 'Jailbreaking and Mitigation of Vulnerabilities in Large Language Models' (2024) https://arxiv.org/abs/2410.15236

	validation of input, filtering to detect malicious patterns, monitoring of LLMs for abnormal input behavior, implementing rate-limiting, structure queries[148]and limiting the amount of text a user can input[149]. **Mitigations for Deployers:** - Limit the amount of sensitive data and guide users to avoid sharing unnecessary personal information through clear instructions, training and warning. Work with providers to ensure they adhere to data protection regulations and do not retain or misuse (sensitive) input data. - Require secure user authentication to restrict access to the input interface and protect session data. As highlighted in the provider's mitigations, deployers should also implement secure password practices based on the latest NIST and ENISA recommendations,[150] encourage users to adopt password managers,[151] and raise awareness about secure practices and internal password policies among users, such as employees. - Clearly communicate to users how their data is handled and processed at each phase of the data flow. This could be done through (internal) privacy policies, instructions, warning or disclaimers in the user interface. - To mitigate adversarial attacks several measures can be implemented such as adding a layer for input sanitization and filtering, monitoring and logging user queries to detect unusual patterns, and incorporating post-processing layers to validate outputs. Additionally, educating users on proper usage can help reduce the likelihood of unintentional inputs that may lead to harmful outcomes.
Provider interface & API	**Risks:** **Data interception:** Insufficient encryption during data transmission to the provider's servers may expose input to interception by third parties. **API misuse:** If API access is not restricted and secured, attackers could exploit the API to intercept or manipulate data. Attackers could also overwhelm the API with excessive traffic to disrupt its availability (Denial-of-Service (DoS) Attacks). **Interface vulnerabilities:** Interface vulnerabilities refer to weaknesses in the provider's user interface that may expose user data to malicious actors. These vulnerabilities can stem from technical flaws (e.g., insufficient input validation, misconfigured API endpoints) or social engineering tactics such as phishing. For example, attackers could create fake versions of a chatbot interface (e.g., replicating the design and branding) to trick users into entering sensitive information such as credentials, payment details, or personal data. Malicious actors could develop deceptive applications claiming to be legitimate integrations with the provider's API, tricking end-users into sharing sensitive data. **Mitigations for Providers:** - Enforce end-to-end encryption for all data transmissions, regularly update encryption protocols, and use secure key management practices. - Implement strong authentication (e.g., API keys, OAuth), enforce rate limits[152], monitor for suspicious activity (anomaly detection), and regularly audit API security. - Perform regular security testing, apply input validation to prevent attacks, and implement robust session management controls. Regarding phishing, both provider

[148] S.Cheng et al. 'StruQ: Defending Against Prompt Injection with Structured Queries' (2024) https://arxiv.org/abs/2402.06363

[149] Open AI Platform, 'Safety best practices' (n.d) https://platform.openai.com/docs/guides/safety-best-practices#constrain-user-input-and-limit-output-tokens

[150] Trust Community, NIST password guidelines 2025: 15 rules to follow' (2024) https://community.trustcloud.ai/article/nist-password-guidelines-2025-15-rules-to-follow/

[151] Wikipedia, 'Password Manager' (2025) https://en.wikipedia.org/wiki/Password_manager

[152] LLM Engine, (n.d) https://llm-engine.scale.com/guides/rate_limits/

	and deployer have roles in addressing this risk. The provider should implement platform-level protections, such as safeguarding the authenticity of their interface (e.g., anti-spoofing measures, branding protections, secure APIs), monitoring for suspicious activity, and providing tools to help deployers detect and prevent abuse. **Mitigations for Deployers:** - If the deployer is using only the provider's interface, their responsibility is limited to securely managing access credentials and complying with data handling policies; however, if the deployer integrates the provider's API into their own systems, they are additionally responsible for securing the integration, including encryption, monitoring, and safeguarding data in transit. Both provider and deployer should design, develop, deploy and test applications and APIs in accordance with leading industry standards (e.g., OWASP for web applications[153]) and adhere to applicable legal, statutory or regulatory compliance obligations. - The deployer should educate employees and end users about evolving phishing techniques—such as fake interfaces, deceptive emails, or fraudulent integrations—that could trick individuals into revealing sensitive information. Education should focus on recognizing suspicious behaviors and verifying the legitimacy of communications and interfaces.
LLM processing at Providers' infrastructure	**Risks:** **Model inference risks:** During processing, the model might inadvertently infer sensitive or inappropriate outputs based on the training data or provided input. **(Un)intended data logging:** Providers can log user input queries and outputs for debugging or model improvement, potentially storing sensitive data[154] without explicit user consent. If logged user queries are included in training data, in case of an adversarial attack, attackers might introduce malicious or misleading content to manipulate the model's future outputs (data poisoning attack)[155]. **Anonymization failures:** Inadequate anonymization or filtering techniques could lead to the inclusion of identifiable user data in model training datasets, raising privacy concerns. **Unauthorized access to logs:** Logs containing user inputs and outputs could be accessed by unauthorized personnel or exploited in the event of a data breach. **Data aggregation risks:** If logs are aggregated over time, they could form a comprehensive dataset that may reveal patterns about individuals, organizations, or other sensitive activities. **Third-party exposure:** If the provider relies on external cloud infrastructure or third-party tools for LLM processing, there's an added risk of data exposure through those dependencies. These dependencies involve external systems, which may have their own vulnerabilities. **Lack of data retention policies:** The provider could store the data indefinitely without having retention policies in place. **Mitigations for Providers:** - Implement strict content filtering mechanisms and human review processes to flag sensitive or inappropriate outputs. - Minimize data logging, collect only necessary information, and ensure you have a proper legal basis for any processed data. Use trusted sources for training data and validate its quality. Sanitize and preprocess training data to eliminate vulnerabilities or biases. Regularly review and audit training data and fine-tuning processes for

[153] OWASP, 'OWASP Top Ten' (2025) https://owasp.org/www-project-top-ten/

[154] The data stored could be sensitive data such as credit card numbers, or special category of data such as health data (article 9 GDPR).

[155] Aubert, P. et al., 'Data Poisoning: a threat to LLM's Integrity and Security' (2024) https://www.riskinsight-wavestone.com/en/2024/10/data-poisoning-a-threat-to-llms-integrity-and-security/

	issues or manipulations. Implement monitoring and alerting systems to detect unusual behavior or potential data poisoning[156]. - Apply robust anonymization techniques, regularly test them for effectiveness, and use automated tools to identify and remove identifiable data before use in training. - Enforce strong access controls, encrypt log data, and monitor access logs for suspicious activity to prevent breaches. - Providers must implement a robust third-party risk management program, adhering to best known frameworks[157] to ensure a secure environment. Key measures include conducting thorough vendor assessments, ensuring compliance with security standards, requiring strong data encryption during transmission and storage, conducting security audits, implementing real-time monitoring and incident response plans tailored to third-party dependencies. Providers should also implement protections against threats such as DoS/DDoS attacks, which can disrupt operations and expose systems to further risks. - Clearly define retention policies, align them with legal requirements and where possible provide users with options to delete their data.
Processed Output	**Risks:** **Inaccurate or sensitive responses:** The model may generate outputs that reveal unintended sensitive information or provide inaccurate or misleading information (hallucinations)[158], leading to harm or misinformation. **Re-identification risks:** Outputs could inadvertently reveal information about the user's query or context that can be linked back to them. **Output misuse[159]:** Users or third parties may misuse the generated output. **Mitigations for Providers:** - Implement post-processing filters to detect and remove sensitive or inaccurate content, and regularly retrain the model using updated and verified datasets to improve response accuracy. Implement disclaimers to highlight potential limitations of AI-generated responses. - Apply privacy-preserving techniques to help you redact sensitive identifiers in outputs and minimize the inclusion of unnecessary contextual details in generated responses. - Provide clear usage policies, educate users on ethical use of outputs, and implement mechanisms to detect and prevent the misuse of generated content where feasible. **Mitigations for Deployers:** - For critical applications, ensure generated outputs are reviewed by humans before implementation or dissemination. - Educate end-users on ethical and appropriate use of outputs, including avoiding overreliance on the model for critical or high-stakes decisions without verification. - Securely store outputs and restrict access to authorized personnel or systems only.

2. Data Flow in an 'off-the-shelf' LLM System

The most common use case for this service model involves organizations leveraging a pre-trained model from a platform to develop and deploy their own AI system. Once the AI system is operational, the data

[156] OWASP, 'LLM10:2023 - Training Data Poisoning' (2023) https://owasp.org/www-project-top-10-for-large-language-model-applications/Archive/0_1_vulns/Training_Data_Poisoning.html

[157] Center for Internet Security, ' The 18 CIS Critical Security Controls' (2025) https://www.cisecurity.org/controls/cis-controls-list

[158] Wikipedia, 'Hallucination Artificial Intelligence' (2025) https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

[159] OWASP, 'LLM05:2025 Improper Output Handling' (2025) https://genai.owasp.org/llmrisk/llm052025-improper-output-handling/

flow closely resembles that of an LLM as a Service model, particularly during user interactions and output generation.

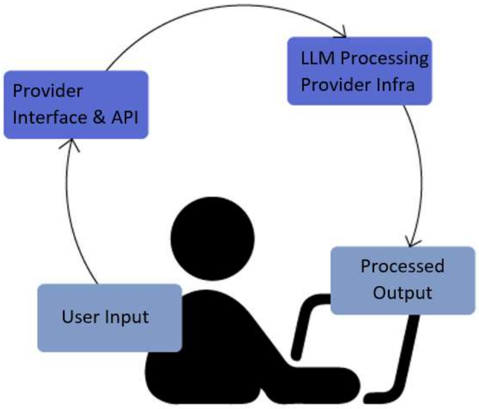


Figure 10. Data Flow in an 'off-the-shelf' LLM System

However, several key differences and limitations set these models apart:

- **Roles and responsibilities:**

Organizations developing an LLM system using the 'off-the-shelf' model may be considered providers[160], particularly when they intend to place the system on the market for use by others (deployers of their system and end-users). This introduces an additional layer of responsibility for data handling, security, and compliance with privacy regulations. The organization may also be developing the AI system for its own internal use.

- **Hosting and processing:**

In a LLM 'off-the-shelf' based system, the provider hosts the model on their infrastructure or a third-party cloud environment of their choice. This contrasts with the LLM as a Service model, where hosting and processing are entirely managed by the original model provider. The new provider is now responsible for all aspects of system integration, maintenance, and security.

- **Customization and training:**

A notable difference is that the initial training and fine-tuning of the model were conducted by the original provider, which can introduce risk and limitations:

- The new provider has often no oversight or knowledge of the contents of the dataset used during the model's initial training, which may introduce biases, inaccuracies, or unknown privacy risks[161].
- The new provider remains dependent on the original provider for updates or bug fixes to the model architecture, potentially delaying critical improvements or fixes.
- Fine-tuning may be limited by the capabilities of the off-the-shelf model. New providers might only be able to adjust certain parameters or add new layers rather than fully retrain the model, restricting its adaptability for highly specific use cases.
- In such cases, retrieval-augmented generation (RAG) is a commonly used alternative. Instead of embedding domain-specific knowledge into the model itself, RAG connects the model to an external knowledge base and retrieves relevant documents at runtime to ground its responses. This enables dynamic, accurate, and updatable answers without modifying the base model, a key advantage for domains with evolving information or regulatory requirements.

[160] According to Article 25 of the AI Act, a deployer of a high risk AI system becomes a provider when they substantially modify an existing AI system, including by fine-tuning or adapting a pre-trained model for new applications. In such cases, the deployer assumes the responsibilities of a provider under the AI Act.

[161] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted on 17 December 2024

A similar approach is cache-augmented generation (CAG)[162] which can reduce latency, lower compute costs, and ensure consistency in responses across repeated interactions but that is less practical for large datasets that are often updated.

The figure below illustrates how RAG[163] works: the user's query is first enhanced with relevant information retrieved from an external database, and this enriched input is then sent to the language model to generate a more accurate and grounded response.

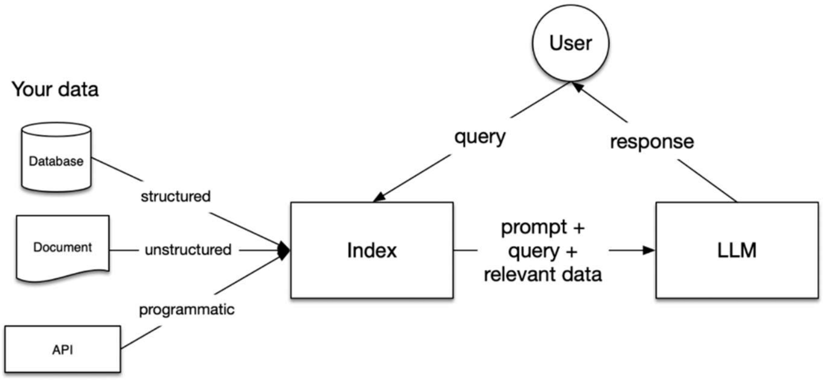


Figure 11. RAG Diagram – Open AI Cookbook

Some common privacy risks of using RAG are:

- Insecure logging or caching: User queries and retrieved documents may be stored insecurely, increasing the risk of unauthorized access or data leaks.
- Third-party data handling: If the retrieval system uses external APIs or services, user queries may be sent to third parties, where they can be logged, tracked, or stored without user consent.
- Exposure of sensitive data: The model may retrieve personal or confidential information if this is stored in the knowledge base.

3. Data Flow in a Self-developed LLM System

In a self-developed LLM system, the organization takes full responsibility for designing, training, and in some cases also deploying the model. This approach provides maximum control over the LLM model but also introduces unique challenges across the data flow.

Since we have already explored an example of privacy risks within the AI lifecycle data flow in a previous section, we will take here a more general approach, focusing on some of the important phases for this service model. The general data flow phases could be as follow:

- **Dataset collection and preparation:**

The organization collects and curates[164] large-scale datasets for training the LLM.

- **Model training:**

The training phase involves using the collected dataset to develop the LLM. This typically requires significant computational resources and specialized infrastructure, such as high-performance GPUs or

[162] Sharma, R., 'Cache RAG: Enhancing speed and efficiency in AI systems' (2025) https://developer.ibm.com/articles/awb-cache-rag-efficiency-speed-ai/

[163] Theja, R., 'Evaluate RAG with LlamaIndex' (2023) https://cookbook.openai.com/examples/evaluation/evaluate_rag_with_llamaindex

[164] Atlan, 'Data Curation in Machine Learning: Ultimate Guide 2024' (2023) https://atlan.com/data-curation-in-machine-learning/

distributed computing systems. Before deployment, the model undergoes rigorous evaluation and testing using separate validation and test datasets to ensure its accuracy, reliability, and alignment with intended use cases.

- **Fine-Tuning:**

After initial training, the model may be fine-tuned using additional datasets to specialize its capabilities for specific tasks or domains.

- **Deployment:**

The trained and fine-tuned LLM is integrated into the organization's infrastructure, making an interface available for end-users.

- **User input:**

End-users interact with the deployed AI system by submitting inputs through an interface such as an app, chatbot, or custom API.

- **Provider interface & API:**

The input is sent through an interface or application. This interface ensures the input is formatted appropriately and securely transmitted to the LLM infrastructure.

- **Model processing:**

The self-developed LLM processes user inputs locally or on the organization's (cloud) infrastructure, generating contextually relevant responses using its trained parameters.

- **Processed output delivery:**

The processed outputs are delivered to end-users or integrated into downstream systems for actionable use. Outputs may include text-based responses, insights, or recommendations.

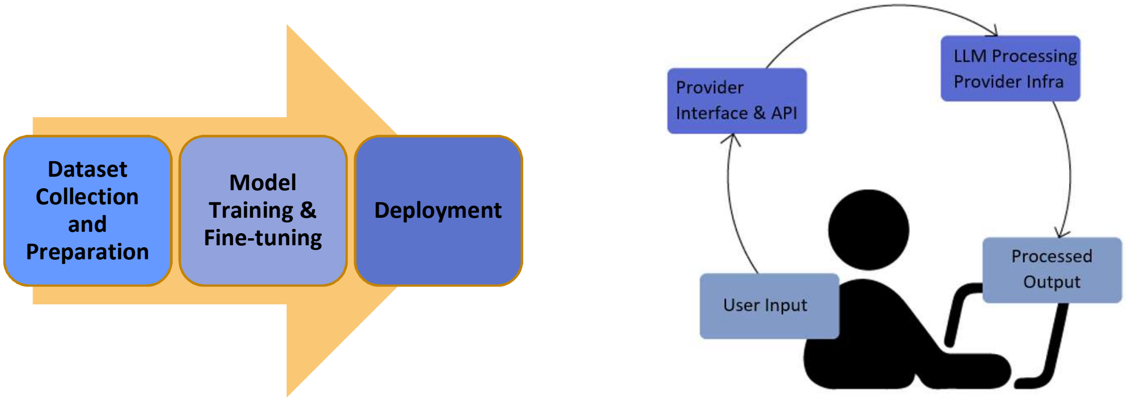

Dataset Collection and Preparation
Model Training & Fine-tuning
Deployment
Provider Interface & API
LLM Processing Provider Infra
Processed Output
User Input

Figure 12. Data Flow in a Self-developed LLM

Privacy considerations in self-developed LLM systems

Self-developing an LLM system provides significant control but also introduces privacy and data protection risks at each phase of the data flow. Below are some of the key risks and suggested mitigations:

Phases	Possible Risks & Mitigations
Dataset collection and preparation	**Risks:** **Sensitive data inclusion:** Collected datasets could (inadvertently) include personal or sensitive information. **Legal non-compliance:** The data could be collected unlawfully violating data protection regulations like GDPR.

	Bias and discrimination: Datasets could reflect societal or historical biases, leading to discriminatory outputs. **Data poisoning:** Datasets may be intentionally manipulated by malicious actors during collection or preparation, introducing corrupted or adversarial data to mislead the model during training. **Mitigations for Providers:** - Apply anonymization and pseudonymization techniques[165] to minimize privacy risks. - In certain use cases, and after carefully weighing the potential pros and cons,[166][167] the creation of synthetic data using LLMs[168] could be an alternative to the use of real personal data. However, synthetic data might not always be suitable[169], as its quality and utility depend on the specific requirements and context of the application. - Ensure data collection is compliant with regulations. - Regularly audit datasets for bias and sensitive content, removing any problematic entries. - Implement robust data validation and monitoring to detect and prevent malicious or corrupted data. Use trusted data sources, apply automated checks for anomalies, and cross-validate data from multiple sources.
Model training	**Risks:** **Unprotected training environment:** Training infrastructure may be vulnerable to unauthorized access, which could expose sensitive data or allow malicious actors to compromise the training process. **Data overfitting:** The model may inadvertently memorize sensitive information instead of generalizing patterns. **Mitigations for Providers:** - Cybersecurity should follow a layered approach, implementing multiple defenses to prevent unauthorized access and mitigate its impact. Mitigations that could be implemented include: using secure computing environments with strong access controls during training (e.g., multi-factor authentication (MFA), privileged access management (PAM)[170], or role-based access controls (RBAC)[171]); applying network segmentation to isolate the training infrastructure from other systems and reduce the attack surface; monitoring and logging access to promptly detect and respond to unauthorized activities; and using encryption for both data at rest and in transit to secure sensitive training data. Another mitigation measure to reduce data exposure is the integration of differential privacy techniques, which add noise to training data to prevent individual data points from being re-identified, even if the model is compromised.

[165] ENISA, 'Pseudonymisation techniques and best practices. Recommendations on shaping technology according to data protection and privacy provisions' (2019) https://www.enisa.europa.eu/sites/default/files/publications/Guidelines%20on%20shaping%20technology%20according%20to%20GDPR%20provisions.pdf

[166] Marwala, T., 'Algorithm Bias — Synthetic Data Should Be Option of Last Resort When Training AI Systems' (2023) https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems

[167] Van Breugel, B. et al., 'Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data' (2023) https://proceedings.mlr.press/v202/van-breugel23a/van-breugel23a.pdf

[168] Vongthongsri, K., 'Using LLMs for Synthetic Data Generation: The Definitive Guide' (2025) https://www.confident-ai.com/blog/the-definitive-guide-to-synthetic-data-generation-using-llms

[169] Desfontaines, D., 'The fundamental trilemma of synthetic data generation' (n.d) https://www.tmlt.io/resources/fundamental-trilemma-synthetic-data-generation

[170] Wikipedia, 'Private access management' (2025) https://en.wikipedia.org/wiki/Privileged_access_management

[171] Wikipedia, 'Role-based access control' (2025) https://en.wikipedia.org/wiki/Role-based_access_control

	It is important to assess the specific use case to determine whether differential privacy is suitable, as it may impact model accuracy in certain scenarios. - Evaluate the model for overfitting and ensure sensitive data is not exposed in outputs.
Fine-Tuning	**Risks:** **Exposure of proprietary or sensitive data:** Fine-tuning data may include sensitive or proprietary information, risking leakage. **Third-party risks:** If external platforms are used for fine-tuning, sensitive data may be exposed to additional risks. **Mitigations for Providers:** - Encrypt fine-tuning datasets and restrict access to authorized personnel. - Use trusted platforms with robust privacy assurances for fine-tuning. - Only include data strictly necessary for fine-tuning tasks.
Deployment	**Risks:** **Unauthorized access:** Weak access controls could allow unauthorized parties to interact with the model or access underlying systems. **Unsecure hosting:** Hosting the model on an unsecured server or cloud environment could expose sensitive data. **Mitigations for Providers:** - Implement strong authentication and role-based access controls for model and system access. - Use cloud environments with strong encryption and monitoring. - Periodically review deployment configurations for vulnerabilities.
User input	See previous table on risks in the data flow of LLM as a Service, where this phase is detailed. The applicable mitigation measures primarily concern providers, but also extend to deployers in scenarios where developers are deploying and using their self-developed AI systems.
Provider interface & API	Idem
LLM processing at Providers' infrastructure	Idem
Processed Output	Idem

4. Data Flow in LLM-based Agentic Systems

AI Agents also encompass the data flow phases discussed so far, but they introduce additional complexity due to their many interactions with other systems and applications. These agents not only process user inputs but also engage with external applications to retrieve information, execute commands, or perform actions. This is often done through function calls, where the agent uses structured interfaces (e.g., APIs) to interact with external tools. In some cases, a Context Management Protocol (CMP)[172] is used to manage and track these interactions, ensuring the agent has access to relevant context across multiple steps or tools. In this example, we explore the data flow of an AI agent interacting with two external applications, highlighting the most common privacy and data protection challenges introduced by these interactions.

[172] Anthropic, 'Introducing the Model Context Protocol' (2024) https://www.anthropic.com/news/model-context-protocol

A simplified overview of the most common phases involved in this data flow includes:

- **User input:**

The process begins with the user providing input to the AI agent, such as a query, command, or task description (e.g., "Book me a flight and a hotel for my trip to Amsterdam").

- Actions:
    - Input is collected through the user interface, such as a chatbot or voice assistant.
    - Preprocessing may occur locally to sanitize and standardize the input.

- **Agent processing:**

The AI agent uses an integrated LLM to understand and process the user input. This includes parsing the request and identifying the actions required to fulfill the task.

- Actions:
    - The input is tokenized and interpreted by the LLM.
    - The agent decides which external applications to contact and formulates queries or commands for them.

- **Interaction with application 1 (e.g., flight booking system):**

The agent sends a query or command to the first external application to retrieve or process information. For instance, it may request available flights based on the user's travel preferences.

- Actions:
    - Data (e.g., user preferences) is transmitted to the external application.
    - The application processes the query and returns a response, such as a list of available flights.
    - The agent receives and processes the response for integration into the overall workflow.

- **Interaction with application 2 (e.g., hotel booking system):**

The agent engages with the second external application to complete another part of the task. For example, it might request hotel options based on the destination and travel dates.

- Actions:
    - Data (e.g., travel dates) is transmitted to the application.
    - The application provides a response, such as available hotels, which is processed by the agent.

- **Aggregation of responses:**

The AI agent integrates the responses from both applications to generate a cohesive result. For instance, it compiles the flight and hotel options into a single output for the user.

- Actions:
    - Responses are validated and formatted for clarity and relevance.
    - Potential errors or conflicts (e.g., overlapping schedules) are resolved.

- **Output generation:**

The agent delivers the aggregated result to the user in a user-friendly format, such as a summary of booking options or actionable recommendations.

- Actions:
    - Output is displayed via the user interface or transmitted to another system for further action.
    - If necessary, the agent provides follow-up prompts to refine the user's preferences or choices.

- **Logging and continuous improvement:**

Interaction logs may be stored temporarily for debugging, system improvements, or retraining purposes, depending on the organization's policies.

- Actions:
    - Logs are analyzed to optimize the agent's performance and enhance user experience.

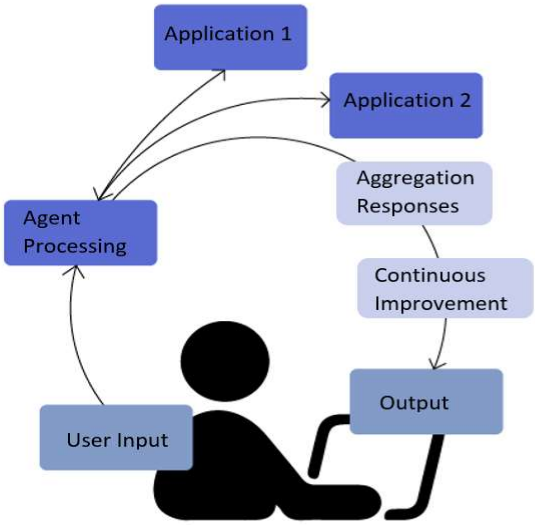


Figure 13. Data Flow in LLM-based Agentic Systems

Overview of Possible Risks and Mitigations

AI agents represent a significant advancement in leveraging LLMs for complex, multi-application tasks. Their data flow incorporates traditional phases seen in LLM systems, but their integration with other systems and their modular agentic architecture—comprising perception, reasoning, planning, memory, action, and feedback loops—introduces unique privacy challenges. Below is an overview of some key privacy risks and corresponding mitigations[173] for each phase, with an emphasis on how these phases align with the agentic architecture.

Phases	Possible Risks & Mitigations
Perception (user input collection and preprocessing)	**Risks:** - Sensitive user input (e.g., personal, financial) or special categories of data (e.g., medical data) could be collected and exposed. - Lack of proper preprocessing may retain identifiable information. - Inputs entered into vulnerable interfaces may be intercepted or misused. - Lack of transparency: Users may not be fully aware of how their data will be used, retained, or shared with the different applications the system will integrate and communicate. **Mitigations for Providers:** - Limit data collection to only what is strictly necessary. - Anonymize and preprocess input data to remove sensitive elements. - Provide tools, interfaces, or APIs that allow deployers to collect user consent when necessary - Secure user interfaces with encryption and authentication mechanisms to protect input data. - Inform users about how their data will be used, retained, and processed through clear and easily accessible privacy policies. **Mitigations for Deployers:**

[173] OWASP, 'Agentic AI – Threats and Mitigations' (2025) https://genaisecurityproject.com/resource/agentic-ai-threats-and-mitigations/

	- Configure the integrations to only collect and store necessary input data. Avoid requesting or storing sensitive data unless explicitly required for the task. - Implement preprocessing pipelines to remove sensitive information before inputs are sent to the provider's system. - When necessary, implement user-facing consent forms or interfaces when collecting data to use with the LLM. - Secure user interfaces with encryption and authentication mechanisms to protect input data and ensure data is handled properly at the deployment level through encrypted local storage and secure API connections. - Clearly communicate to users how their data is handled and processed. This could be done through (internal) privacy policies, warnings, instructions or disclaimers in the user interface.
Reasoning (agent processing)	**Risks:** - Sensitive data used in reasoning tasks may be misused or exposed. - Improper handling of user data during task decomposition may propagate sensitive information. - Inferences made during reasoning could unintentionally reveal personal insights. - Limited explainability could lead to incorrect or suboptimal outputs, reducing trust and reliability, especially in complex decision-making tasks. Without clear reasoning chains, users may struggle to understand or verify how conclusions are reached, increasing the probability of errors and unintended outcomes. **Mitigations for Providers:** - Implement mechanisms to anonymize or preprocess sensitive input data to minimize the risk of sensitive information being used during reasoning tasks. - Provide robust access control features that limit which entities can process sensitive data, even during complex reasoning tasks. - Ensure that all reasoning outputs are logged securely and are auditable by deployers, enabling the tracking of inferences for sensitive information leakage. - Integrate mechanisms for generating reasoning chains (e.g., Chain of Thought) to make the LLM's decision-making process more transparent and auditable. **Mitigations for Deployers:** - Minimize the use of sensitive data during processing and reasoning. - Ensure all inferences and intermediate data are auditable and securely stored. - Apply strict access controls and logging to monitor and limit misuse. - Implement the Chain of Thought (CoT)[174] framework to enhance the reasoning capabilities of LLMs. This can be achieved either through user prompting, where logic for solving problems is provided manually, or through Automated CoT (Auto-CoT), which clusters questions and generates reasoning chains without human intervention. Auto-CoT is particularly effective for LLMs with around 100B parameters but may be less accurate for smaller models.[175]
Planning (task organization and external interactions)	**Risks:** - Sensitive data could be transmitted to external applications without proper safeguards. Function calls may transmit excessive or unnecessary user data to external applications, especially if parameter filtering is not properly implemented. Third-party systems may not adhere to the same privacy and security standards.

[174] Gadesha, V., 'What is chain of thoughts (CoT)?' (2024) https://www.ibm.com/think/topics/chain-of-thoughts

[175] Biswas, D., 'Stateful and Responsible AI Agents' (2024) https://www.linkedin.com/pulse/stateful-responsible-aiagents-debmalya-biswas-runze/

	Mitigations for Providers and Deployers:[176] - Use anonymization and encryption when transmitting data to external applications. - Monitor third-party applications to ensure that external systems interacting with AI agents adhere to the same privacy, security, and regulatory standards as your own system: conduct vendor assessments, verify that third-party providers have up-to-date security certifications, such as ISO 27001, SOC 2, or similar, to confirm they meet recognized security standards, establish contracts or data processing agreements that include clear expectations for privacy compliance, use tools to monitor third-party applications in real time for suspicious activity and ensure third-party providers have robust incident response plans to handle potential breaches notifying stakeholders in a timely manner. - Implement effective Identity and Access Management:[177] implement a zero trust security model continuously verifying identity and device, and assuming no implicit trust; implement dynamic, context-aware access controls adjusting permissions based on real-time factors like location, device status, or behavior to reduce risks; grant just-in-time access, ensuring AI agents only have permissions for the duration of their tasks, minimizing privilege creep; review and update access controls during an AI agent's lifecycle; automate credential rotation, key management, and certificate updates to maintain security and reduce human error. - Obtain, when necessary, user consent for external interactions and provide policies transparently. - Implement filter parameters in function calls and avoid sending and retaining sensitive intermediate data unnecessarily.
Memory (data storage and retention)	**Risks:** - Long-term storage of user data increases the risk of unauthorized access or misuse. - Retention of sensitive data across interactions may violate privacy regulations. CMPs maintain context across multiple interactions, which may result in long-lived sessions where sensitive data accumulates. **Mitigations for Providers and Deployers:**[178] - Allow users to manage stored data (e.g., delete, edit). - Apply secure storage solutions with robust access controls and encryption. - Limit retention periods and implement automated deletion policies for sensitive data.
Action (Output generation and delivery)	**Risks:** - Generated outputs might inadvertently include sensitive or private information. - Outputs shared with external systems could be intercepted or misused. - When orchestrating multiple AI agents, the probability of hallucinations increases as the number of agents involved grows. **Mitigations for Providers and Deployers:**[179] - Validate and filter outputs to ensure they do not reveal sensitive information. - Secure output delivery channels with encryption and authentication mechanisms. - Monitor interactions with external systems for adherence to privacy standards.

[176] Providers are responsible for ensuring the foundational model and platform are secure, privacy-compliant, and equipped with features for secure deployment that deployers can configure. Deployers are responsible for integrating, configuring, and using the LLM securely within their specific context. The division of responsibilities depends on the level of customization required by the deployer and the deployment context, with both parties sharing accountability for implementing necessary mitigations.

[177] McDougald, D., et al., 'Strengthening AI agent security with identity management' (2025) https://www.accenture.com/us-en/blogs/security/strengthening-ai-agent-security-identity-management

[178] See footnote 176

[179] idem

	- To reduce hallucinations, though crafted prompts can be helpful, they offer limited effectiveness. LLMs should be fine-tuned with curated, high-quality data and configured to limit the search space of responses to relevant and up-to-date information[180].
Feedback and Iteration Loop (learning and improvement)	**Risks:** - User feedback may be stored or used for model retraining without consent. - Sensitive feedback information may unintentionally persist in logs or datasets. **Mitigations for Providers:** - Ensure the platform includes clear and user-friendly mechanisms for users to opt-in or opt-out of having their feedback data used in model retraining. - Offer built-in tools or features to automatically anonymize or pseudonymize feedback data before storing or processing it for retraining purposes. - Limit log retention periods by default and provide configurable options to deployers to ensure compliance with privacy regulations. **Mitigations for Deployers:** - Clearly communicate to users how their feedback data will be used and ensure robust tracking of user consent (opt-in/opt-out) mechanisms provided by the platform. - Use anonymization and pseudonymization tools to securely handle feedback data. - Limit log retention periods and ensure compliance with privacy regulations.

Considerations on External Integrations in LLM Data Flows

Across all service models—whether SaaS-based, off-the-shelf, self-developed, or agentic AI — LLMs may interact with external systems, adding complexity to the data flow. These interactions can include retrieving information from external knowledge bases, accessing the web for real-time data, or integrating with other applications through APIs or plugins. Such integrations introduce additional layers of data flow that must be carefully mapped and understood. When designing an LLM-based system, it is critical to account for how external data sources and systems are accessed, how data is transmitted and processed, and what safeguards are in place to protect user privacy and data security.

Filters as Safeguards and as Additional Layer for Input and Output

Most LLM systems incorporate input and output filters as safeguards, introducing an additional layer into the data flow. These filters act as control mechanisms to preprocess incoming data or refine generated outputs, helping to enforce privacy, safety, and content standards. They can take various forms — including Python scripts, prompt templates, or even other LLMs.[181]

Input filters function as gatekeepers, screening data before it reaches the core model. For example, they might block personal data, detect harmful prompts, or sanitize inappropriate language.

Output filters ensure that system responses meet privacy, ethical, or contextual requirements before they are shown to users or passed to downstream systems. An output filter might, for instance, remove sensitive content or rephrase a response to align with organizational policies.

The addition of filters introduces complexity into the system's architecture: Filters may add latency/processing time, impacting response times in real-time systems. They need to be secure, as vulnerabilities could expose sensitive data or allow malicious inputs to bypass scrutiny. And must be

[180] See footnote 175

[181] Inan, H. et al., 'Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations' (2023) https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/

monitored and regularly updated to adapt to new risks, changing regulations, or evolving system requirements.

Roles in LLMs Service Models According to the AI Act and the GDPR

The roles of provider and deployer under the AI Act, as well as controller and processor under the GDPR, may differ based on the different service model. Below is an explanation of how these roles may apply and the rationale behind their assignment, categorized by each service model. Note that the qualification of organizations as controller or processor should be assessed based on the circumstances of each case, and the explanation provided here is intended for reference purposes only and does not imply it will always apply in the same way.

1. LLM as a Service Model

In this service model, all processing occurs on the provider's infrastructure, and the user interacts with the LLM through APIs or cloud platforms. Providers may also use collected data for model improvement, fine-tuning, or research.

Example Tool: OpenAI's GPT API (e.g., ChatGPT)
Use Case: Businesses use OpenAI's API to integrate ChatGPT into their applications for customer support or content generation. The processing is performed entirely on OpenAI's infrastructure.
Provider: OpenAI.
Deployer: A business that integrates ChatGPT's API into its workflow.

AI Act Roles

- Provider: The organization that develops and offers the LLM as a service. Providers are responsible for ensuring compliance with the AI Act, including risk management, transparency, and technical robustness (e.g., OpenAI providing GPT models via APIs).
- Deployer: The organization using the LLM (e.g., a business using the provided interface for any particular task).
- New provider: An organization integrating the API of an LLM as a Service model into their commercial AI system (e.g., a chatbot) could also be considered a provider under the AI Act if their system qualifies as high-risk and falls within the scope of Article 25 of the AI Act.

GDPR Roles

- Deployer as controller: The deployer using the LLM as a Service typically acts as the data controller, as they determine the purposes and means of data processing (e.g., collecting customer queries to improve services or using an LLM tool for summarization purposes).
- Provider as controller: When providers collect or retain data for their own purposes (e.g., model fine-tuning or feature improvement), they assume the role of controller too. This is the case in most LLM as a Service solutions where providers have ownership of the model and the training data. In this scenario, a joint controllership might be the more suitable option.
- Processor: The provider acts as a processor when handling data strictly according to the deployer's instructions for specific tasks, like generating responses. This might be difficult in this service model due to the providers model's ownership.

In an LLM as a Service model scenario, we often talk about the concept of **shared responsibility**, where both the provider and the deployer play distinct but complementary roles in ensuring privacy, security, and compliance. The provider is responsible for the infrastructure, model training, and maintenance, while the deployer must ensure secure usage, proper integration, and adherence to applicable regulations within their specific deployment context. This division of responsibilities requires clear agreements and robust collaboration to effectively manage risks.

2. LLM 'Off-the-Self' Service Model

In an 'Off-the-Shelf' Service Model, the roles of provider and deployer are determined by the operational setup and specific use of the LLM system and on how the platform and the deployer interact with the model. In general, the involvement of the original provider is very limited. They often do not have access to the deployer's data, or the tasks being executed, and their involvement is limited to ensuring the model is functional, robust, and compliant with regulatory standards at the point of delivery. The provider may continue to offer updates, bug fixes, or improvements to the model's base functionality, but this does not typically involve accessing or processing deployer data.
The deployer operates independently, using the model provided through the platform to build their own LLM system.

Example Tool: Hugging Face's Transformers Library
Use Case: A deployer downloads a pre-trained language model (e.g., BERT) from Hugging Face's repository and fine-tunes or integrates it into their system.
Provider: Hugging Face (as platform) and Google (BERT developer) are responsible for the original model's robustness and compliance).
Deployer: An organization that uses the model for custom purposes, such as creating a chatbot.

AI Act Roles

- Provider: The organization that develops, puts in the market or into service the off-the-shelf LLM model. Providers are responsible for ensuring that the model adheres to the AI Act's requirements[182]. In case of LLMs released under free and open source licenses, they should be considered to ensure high levels of transparency and openness if their parameters, including the weights, the information on the model architecture, and the information on model usage are made publicly available[183]
  - If the platform provider develops, trains, or significantly fine-tunes an LLM and makes it available to deployers, they would act as providers under the AI Act.
  - The platform could also just have the role of infrastructure enabler and not being considered then a provider but a distributor.
- Deployer: The organization using the off-the-shelf model to build or enhance its own services takes on the role of deployer. However, in cases of high risk AI systems, the deployer may also assume the role of provider if they significantly modify or fine-tune the model or make it available to others as part of their own services. This dual role is addressed under Article 25 of the AI Act.

[182] Note based on recital 104 AI Act: Providers of general-purpose AI models released under a free and open source license, with publicly available parameters (including weights, architecture details, and usage information), should be subject to exceptions regarding transparency-related requirements under the AI Act. However, exceptions should not apply when such models present a systemic risk. In such cases, transparency and an open source license alone should not suffice to exempt the provider from compliance with the regulation's obligations. Furthermore, the release of open source models does not inherently guarantee substantial disclosure about the datasets used for training or fine-tuning, nor does it ensure compliance with copyright law. Therefore, providers should still be required to produce a summary of the content used for model training and implement a policy to comply with Union copyright law, including identifying and respecting reservations of rights as outlined in Article 4(3) of Directive (EU) 2019/790.
[183] Recital 102 AI Act

GDPR Roles

- Deployer as Controller: The deployer typically acts as the controller, as they determine the purpose and means of processing personal data during their use of the LLM.
- Provider as controller: The original model provider may act as a controller in limited scenarios where they process data for their own purposes. If the platform provider logs, analyzes, or retains user or deployer data for purposes like improving platform services, debugging, or monitoring system performance, they could be taking on the role of controller for this specific data processing.
- Processor: This role could be carrying out cloud-based tasks explicitly instructed by the deployer. For example, during data inference tasks, data might be processed according to the deployer's instructions. In this case, a platform providing a model could act as a processor under the GDPR.

The provider remains accountable for the foundational model's compliance and functionality. The deployer is responsible for how the model is implemented, customized, and operated within their specific context, especially in scenarios where data is processed locally, or cloud tasks are guided by the deployer. This dual-layered responsibility emphasizes the need for clear contractual agreements and robust governance mechanisms.

3. Self-developed LLMs

All operations, from model development, infrastructure, input collection to model processing, are performed under the responsibility of the provider that is often also deploying the model for own use.

Example Tools: PyTorch, DeepSpeed, TensorRT-LLM
Use Case: A company uses a collection of different tools to develop and train a custom LLM entirely on their infrastructure, including data preparation, training, and deployment.
Provider: The organization developing the LLM.
Deployer: The same organization (if it also deploys the model) or a third-party client.

AI Act Roles

- Provider: The entity developing the LLM.
- Deployer: The organization deploying the solution and taking on most operational responsibilities, including monitoring, risk management, and transparency.

In this specific service model, the organization developing the LLM system could be the same organization putting the system into own use. In that scenario the same organization would be considered a provider and deployer under the AI Act.

GDPR Roles

- Provider as Controller: The LLM system developer, as they control and execute all data processing activities within their local infrastructure during development.
- Deployer as Controller: The deployer, as they determine the purpose and means of processing personal data during their use of the LLM.
- Processor: Any third party processing data on behalf of the controller might take this role.

The controller's full control over infrastructure and data makes them responsible for compliance with GDPR and AI Act requirements.
The processor's role is limited to any third party tool or component that the controller could be using in the process.

4. Agentic AI Systems

Agentic AI systems introduce unique dynamics to data flows and role allocation under the GDPR and AI Act due to their autonomous and dynamic behavior.

Example Tool: Auto-GPT
Use Case: Auto-GPT autonomously performs tasks such as data gathering, and planning based on high-level instructions. A deployer can fine-tune the agent or integrate it with specific systems.
Provider: Developers creating foundational Auto-GPT architectures.
Deployer: Organizations using or modifying Auto-GPT for specific workflows (e.g., automating business operations).

AI Act Roles

- Provider: The entity developing and supplying the LLM or core agentic architecture.
- Deployer: The organization implementing the agentic AI system for its own or third-party use. In high risk AI systems, if the deployer fine-tunes the agent, integrates it with specific systems, or significantly modifies its architecture, they may also assume the role of a provider under Article 25 of the AI Act, responsible for compliance of the modified system.

In this service model, the deployer is often both a provider and a deployer, depending on the level of customization, fine-tuning, or downstream deployment of the agentic AI system.

GDPR Roles

- Deployer as Controller: The deployer typically assumes the role of the controller, as they determine the purpose and means of processing personal data. This includes inputs, outputs, memory management, and interactions with external systems.
- Processor: When the deployer uses third-party tools, external APIs, or cloud services as part of the agentic AI's operations, these third-party providers could act as processors. For example, if an external API or service facilitates real-time data retrieval or enhances functionality, it takes on a processing role under the deployer's instruction. In some cases, third-parties could act as joint-controllers.

Responsibility Sharing:
The deployer bears significant responsibility for managing the AI agent's outputs and interactions. However, providers supplying foundational LLMs, or modules could also share responsibility for pre-deployment compliance.

This table shows an overview of the possible roles per service model, always subject to an assessment of the circumstances at hand:

Model	Deployer as Controller	Provider as Controller	Processors
LLMs as a Service	When the deployer uses the LLM for application-specific purposes, defining the data processing goals and methods (e.g., handling user queries).	When the provider performs fine-tuning, training, or analytics beyond deployer instructions (e.g., retaining data for retraining or monitoring purposes).	Model Provider: For processing data under deployer's instructions, such as handling input/output during inference tasks.
LLM 'Off-the-Self'	For determining the use of the LLM system, controlling data during preprocessing, output handling, and customization of the LLM for specific workflows.	When the original model provider retains or reuses data for its purposes (e.g., debugging or performance monitoring).	Platform: For cloud-based processing tasks performed under deployer's instructions (e.g., hosting the model for inference).
Self-developed LLM	Fully applicable when developer is also deployer, as the organization directly defines data processing goals.	Fully applicable, as the organization both develops and controls the model.	Not applicable, as no third-party model provider is involved in processing. There might be other third parties acting as processors.
Agentic AI	The deployer acts as the primary controller, managing data inputs, task assignments, memory storage, and external system interactions, while overseeing the agent's autonomy.	When the foundational model provider or module supplier retains data from interactions for their own purposes, such as improving reasoning components or tools. When provider performs training beyond deployer instructions.	Model and tool provider / other third parties: When external APIs, tools, or platforms are used for specific agent functions under the deployer's instructions.

4. Data Protection and Privacy Risk Assessment: Risk Identification

Risk assessment is generally considered the first phase within risk management. It encompasses risk analysis, which involves identifying, estimating, and evaluating potential risks. As a starting point, risk analysis requires a careful identification of the risks that may arise in a given context. This section looks at how to approach the identification of privacy and data protection risks in LLM systems.

Criteria to Consider when Identifying Risks

Risk Factors

To help identify risks associated to the use of LLMs we can make use of a variety of risk factors.
Risk factors are conditions associated with a higher probability of undesirable outcomes. They can help to identify, assess, and prioritize potential risks. For instance, processing sensitive data and large volumes of data are two risk factors with a high level of risk. Acknowledging them in your own use case, can help you identify related potential risks and their severity.

The risk factors shown below are the result of analysing the contents of legal instruments such as the GDPR[184], the EUDPR[185], the EU Charter[186] and other applicable guidelines related to privacy and data protection.[187]The following risk factors can help us identify data protection and privacy high level risks in LLM-based systems:

High level Risk / Important concerns	**Examples of applicability**
Sensitive & impactful purpose of the processing Using a LLMS to decide on or prevent the exercise of fundamental rights of individuals, or about their access to a service, the execution or performance of a contract, or access to financial services is a concern, especially if these decisions will be automated without human intervention. Wrong decisions could have an adverse impact on individuals.	Deploying an LLM to determine creditworthiness or loan approvals without human oversight, or to automate decisions about hiring, promotions, or job terminations without adequate safeguards could negatively impact individuals.
Processing sensitive data When an LLM is processing sensitive data such as special categories of data, personal data related to convictions and criminal offences, financial data, behavioral data, unique identifiers, location data, etc. This is a reason of concern since processing inappropriately this personal data could negatively impact individuals.	Using an LLM-based system to analyze patient records, diagnoses, or treatment plans or data related to criminal convictions, court records, or investigative reports.
Large scale processing Processing high volumes of personal data is a reason of concern, especially if these personal data are sensitive. The higher the volume the bigger the impact in case of a data breach or any other situation that put the individuals at risk.	An LLM deployed in a large e-commerce platform processing vast amounts of user data, or an LLM used in a social media platform.

[184] General Data Protection Regulation (2016/679)
[185] European Union Data Protection Regulation (Reg. 2018/1725)
[186] Charter of Fundamental Rights of the European Union (2012/C 326/02)
[187] AEPD, 'Risk Management and Impact Assessment in Processing of Personal Data' p-79 (2021) https://www.aepd.es/guides/risk-management-and-impact-assessment-in-processing-personal-data.pdf

Processing data of vulnerable individuals This is a concern because vulnerable individuals often require special protection. Processing their personal data without proper safeguards can lead to violations of their fundamental rights. Some examples of vulnerable individuals are children, elderly people, people with mental illness, disabled, patients, people at risk of social exclusion, asylum seekers, persons who access social services, employees, etc.	This could be the case when LLM systems are used in the health sector, at schools, social services organizations, government institutions, employers, etc. For instance, an LLM-based platform used in schools to assess student performance and provide personalized learning recommendations processes data about children.
Low data quality The low data quality of the input data and/or the training data is a concern bringing possible risks of inaccuracies in the generated output what could cause wrong identification of characters and have other adverse impacts depending on the use case.	LLMs rely heavily on the quality of both the input data provided by users and the data used for training the model. Any inaccuracies, biases, or incompleteness in the data can have far-reaching consequences, as LLMs generate outputs based on patterns they detect in their training and input data. The degree of risk posed by low data quality depends heavily on the application. In less critical use cases, such as content generation, inaccuracies may be less impactful. However, in high-stakes scenarios, such as healthcare, finance, or public policy, even minor inaccuracies can have significant negative consequences.
Insufficient security measures The lack of sufficient safeguards could be the cause of a data breach. Data could also be transferred to states or organizations in other countries without an adequate level of protection.	This could be the case if there are not sufficient safeguards implemented to protect the input data and the results of the processing. This could be applicable to any use case. LLMs offered as SaaS solutions involve in some cases data being sent for processing to servers in countries without adequate data protection laws, increasing exposure to privacy risks.

Other Components of AI Risk

The AI Act introduces essential safety concepts into the risk management process of AI systems, reflecting its nature as a product safety regulation. Understanding these concepts is critical when initiating the assessment of risks associated with an LLM based system.

Key terms[188] such as hazard, hazard exposure, safety, threats, vulnerabilities, and their interplay with fundamental rights are components of risk that provide a foundational framework for the evaluation of AI risks.

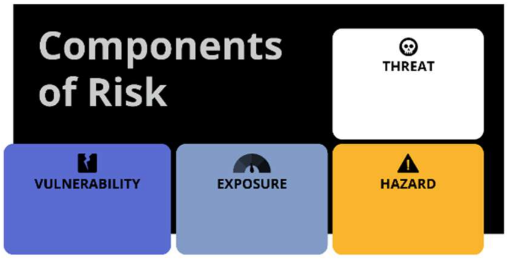


Figure 14. Other components of Risk

A **hazard** refers to a potential source of harm, while **hazard exposure** describes the conditions or extent to which individuals or systems are exposed to that harm in a hazardous situation. **Safety** represents the measures implemented to minimize or mitigate harm, ensuring the system operates as intended without causing undue risk. **Threats** are external factors that may exploit **vulnerabilities** within the LLM based system, which are weaknesses that could be exploited to compromise functionality, security, or

[188] These terms are explained here in an accessible manner to aid understanding, but they are not official definitions. The European harmonized standard on risk management, currently being developed by CEN/CENELEC at the request of the European Commission, will contain standardized definitions and provide formalized guidance on these terms.

data protection. The AI Act emphasizes the protection of fundamental rights, including privacy, to ensure that AI systems do not adversely impact individuals.

When trying to identify the risks of an LLM based system, is important to consider all these components of risks[189] that could have an impact on privacy and data protection. Privacy risks often stem from hazards, or from vulnerabilities within the system that could be exploited by external or internal threats. A hazard exposure in this context could refer to how individuals' personal data is exposed to these risks through the use of the LLM based system, for example, during input querying.
Understanding these interrelated concepts facilitates the risk management process of AI systems that need to comply with the GDPR and the AI Act having as end goal the protection of individuals.

The Importance of Intended Purpose and Context in Risk Identification

The GDPR, in Recital 90 emphasizes the importance of establishing the context: "taking into account the nature, scope, context and purposes of the processing and the sources of the risk".
This is a critical principle when conducting a privacy risk assessment, as it ensures that risks to the rights and freedoms of natural persons are evaluated in their specific operational and contextual settings. This aligns closely with the concept of **'intended purpose'** in the AI Act, which emphasizes the need to define and assess how an AI system is expected to operate.
The concepts of **intended purpose** [190] and **context** are foundational in the identification and management of risks in AI systems[191]. Intended purpose refers to the specific purposes and scenarios for which the AI system is designed, while context encompasses the environment, user base, and operational settings in which the system functions. Understanding these dimensions is crucial because risks often arise when systems are used in unintended ways or in contexts that introduce unforeseen vulnerabilities.

By clearly defining the intended purpose, you can assess whether the design and functionalities align with the anticipated application's use. This is also helpful to identify potential misuse of the system and harm to specific user groups. Similarly, understanding the broader context—including user demographics, language, cultural factors, and business models—enables you to evaluate how the system interacts with its environment to anticipate potential issues.

The Role of Threat Modeling in Privacy Risk Identification

Given the broad spectrum of risks associated with AI, methodologies like threat modeling[192] play a pivotal role in systematically identifying privacy risks. These methodologies often leverage libraries of specific AI threats, hazards and vulnerabilities providing a structured evaluation of risks throughout the lifecycle of the AI system including those arising from both intended and unintended uses of the system.

Threat modeling can assist in identifying potential attack surfaces[193], misuse cases, and vulnerabilities, enabling a proactive approach to risk identification and mitigation. For example, by identifying data flows and system dependencies, threat modeling can reveal risks like unauthorized data access which

[189] Novelli, C et al., 'AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act' (2024) https://doi.org/10.1007/s44206-024-00095-1

[190] The AI Act requires in Article 9 (2)(a) for high-risk AI systems risk management systems 'the identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose;

[191] NIST, 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)' (2023) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[192] Threat Modeling Manifiesto, (n.d) https://www.threatmodelingmanifesto.org/

[193] Meta, Frontier AI Framework (2025) https://ai.meta.com/static-resource/meta-frontier-ai-framework/?utm_source=newsroom&utm_medium=web&utm_content=Frontier_AI_Framework_PDF&utm_campaign=Our_Approach_to_Frontier_AI_blog

may not be immediately apparent. The identified threats from a threat modeling session can be integrated into LLM evaluations, where models are systematically tested against the threats through adversarial testing, red teaming, or scenario-based assessments.

The Importance of Monitoring and Collecting Evidence

To effectively manage risks in LLM systems, it is essential to also base your assessment on robust evidence[194]. This includes gathering data from multiple sources to ensure the analysis accurately reflects potential harms and vulnerabilities. After deployment, monitoring data such as logs and usage patterns, can provide insights into how the system is being used in practice and whether it aligns with its intended purpose. Throughout the whole lifecycle, evaluation results from metrics, testing, red teaming exercises[195] and external audits can highlight gaps in functionality, biases, or performance issues. Additionally, feedback from users or from whistleblowers[196], containing complaints, reports, or behaviour patterns, offers valuable perspectives on real-world risks and potential areas for improvement.

Incorporating evidence[197] from both core and enhanced sources ensures a comprehensive understanding of risks. Core evidence includes existing data on system characteristics and user interactions, while enhanced evidence might involve consulting experts, conducting targeted research, or using outputs from content moderation or technical evaluation systems. This multi-faceted approach not only aids in identifying risks but also provides a documented basis for risk management decisions.

Examples of Privacy Risks in LLM Systems

LLMs can present a wide range of privacy and data protection risks. These risks arise from various factors, including the specific use case, the context of application, and the risk factors and evidence identified during the assessment process. Recognizing and addressing these risks is critical for organizations aiming to procure, develop, or deploy LLM-based systems responsibly.

The table below categorizes common privacy risks of LLM systems based on their applicability to the roles of providers and deployers. Both providers and deployers are responsible to all risks on the table, but the degree of responsibility depends on the level of control over the system (e.g., providers for infrastructure, deployers for usage and configuration).
It is important to consider how risks differ depending on the perspective. For instance, while a provider may face regulatory obligations to minimize data storage, a deployer must evaluate the risks of entrusting a provider with sensitive information. These roles come with different responsibilities that require specific risk management strategies.

Providers, **deployers**, and **procurement** teams must address these risks collaboratively. Procurement, in particular, plays a vital role in bridging the responsibilities of providers and deployers by ensuring that selected systems meet regulatory standards and organizational privacy requirements. Key

[194] Ofcom, 'Protecting people from illegal harms online - Annex 5: Service Risk Assessment Guidance (2024) https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/270826-consultation-protecting-people-from-illegal-content-online/associated-documents/annex-5-draft-service-risk-assessment-guidance?v=330403

[195] Martineau, K. 'What is red teaming for generative AI?' (2024) https://research.ibm.com/blog/what-is-red-teaming-gen-AI

[196] Recital 172 AI Act: "Persons acting as whistleblowers on the infringements of this Regulation should be protected under the Union law. Directive (EU) 2019/1937 of the European Parliament and of the Council (54) should therefore apply to the reporting of infringements of this Regulation and the protection of persons reporting such infringements."

[197] Ofcom, 'Protecting people from illegal harms online - Annex 5: Service Risk Assessment Guidance (2024) https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/270826-consultation-protecting-people-from-illegal-content-online/associated-documents/annex-5-draft-service-risk-assessment-guidance?v=330403

considerations during procurement include assessing the provider's policies, ensuring compliance with relevant regulations, and embedding clauses that limit data misuse and support data subject rights.
Deployers **using** LLMs need to consider the risks related to their specific use cases and context. Making use of the risk factors or evaluation criteria can facilitate the identification of those risks. For instance, the criteria 'low data quality' can already trigger the identification of risky processing activities that could result in harm.
Providers and **developers** of LLMs must implement risk management as an iterative process to identify and address risks, recognizing that these risks can emerge at various phases of the development lifecycle, as discussed in previous sections.

The overview provided by the table below can serve as a practical starting point for identifying and analyzing privacy and data protection risks throughout the lifecycle of LLM based systems. The table presents a consolidated summary of privacy risks, complementing the details already provided in Section 3 under Data Flow and Associated Privacy Risks in LLM Systems.

				Risk applicability		
Data Protection and Privacy Risks	**Risk description**	**GDPR potential Impact**	**Examples**	**Service Model**	**Provider**	**Deployer**
1. Insufficient protection of personal data what eventually can be the cause of a data breach.	Safeguards for the protection of personal data are not implemented or are insufficient.	Infringement of: Art. 32 Security of processing, Art. 5(1)(f) Integrity and confidentiality and Art. 9 Processing of special categories of personal data	Sensitive data disclosure in user inputs or during training, inference and output. Unauthorized access, insufficient encryption during data transmission, API misuse, interface vulnerabilities, inadequate anonymization or filtering techniques, third party exposure.	✓ LLM as a Service ✓ LLM ‘off-the-shelf’ ✓ Self-developed LLM ✓ Agentic LLM	✓	✓
2. Misclassifying training data as anonymous by controllers when it contains identifiable information.	Controllers may incorrectly assume training data is anonymous, failing to implement necessary safeguards for personal data protection.	Infringement of: Articles 5(1)(a) (Lawfulness, Fairness, and Transparency), 5(1)(b) (Purpose Limitation), 25 (Data Protection by Design and Default)	An LLM trained on improperly anonymized user logs reveals identifiable user information through model inference attacks. A deployer discovers that the third-party LLM they are using has been trained on non-anonymized personal data, and the vendor fails to implement appropriate safeguards, exposing the deployer to compliance risks.	✓ LLM as a Service[198] ✓ LLM ‘off-the-shelf’[199] ✓ Self-developed LLM ✓ Agentic LLM	✓	✓
3. Unlawful processing of personal data in training sets.	Personal data is included in training datasets without proper legal basis, safeguards, or user consent.	Infringement of: Articles 5(1)(a) (Lawfulness, Fairness, and Transparency) Articles 6(1) (Lawfulness of Processing), 7 (Consent), 5(1)(c) (Data Minimization)	An e-commerce platform uses customer purchase histories to train an LLM without informing customers or obtaining their consent.	✓ LLM as a Service[200] ✓ LLM ‘off-the-shelf’[201] ✓ Self-developed LLM ✓ Agentic LLM	✓	✓
4. Unlawful processing of special categories of personal data and data relating to criminal convictions and offences in training data.	Training datasets include sensitive data, such as health or criminal records, without meeting GDPR exceptions for lawful processing.	Infringement of: Articles 9(1) and 9(2) (Special Categories of Data), Article 10 (Criminal Convictions and Offences).	Medical records scraped from unsecured online sources are used to train a healthcare chatbot without applying GDPR compliant safeguards.	✓ LLM as a Service[202] ✓ LLM ‘off-the-shelf’[203] ✓ Self-developed LLM ✓ Agentic LLM	✓	✓

[198] This risk primarily applies to the provider; however, the deployer shares responsibility by ensuring they engage with lawful vendors. The deployer's role includes conducting due diligence to verify that the provider complies with legal obligations and operates within the bounds of applicable regulations.

[199] Idem

[200] Idem

[201] Idem

[202] Idem

[203] Idem

5. Possible adverse impact on data subjects that could negatively impact fundamental rights.	The output of the system could have an adverse impact on the individual.	Infringement of: Art. 5(1)(d) Accuracy, Art. 5(1)(a) Fairness, Art. 22 Automated individual decision-making, including profiling, Art. 25 Data protection by design and by default	A system providing output that is not accurate or contain bias and does not provide with mechanisms to amend errors. The output of an LLM could be used to make automatic decisions which produce legal effects or similarly significant effects on data subjects.	✓ LLM as a Service ✓ LLM 'off-the-shelf' ✓ Self-developed LLM ✓ Agentic LLM	✓	✓
6. Not providing human intervention for a processing that can have a legal or important effect on the data subject.	Automated decisions that significantly impact individuals are made without human review, violating GDPR requirements for human oversight, or are based on inappropriate ground[204].	Infringement of: Articles 22(1) and 22(3) (Automated Decision-Making), Article 12 (Transparent Communication).	A chatbot automates loan approvals based on user provided data, denying applications without involving a human reviewer.	✓ LLM as a Service ✓ LLM 'off-the-shelf' ✓ Self-developed LLM ✓ Agentic LLM	✓	✓
7. Not granting data subjects their rights.	Data subjects' rights cannot be completely or partially granted.	Infringement of: Art. 12 – 14: Information to be provided when personal data is collected Art. 16 and Art. 17: Right to rectification and right to erasure Article 18 Right to restriction of processing and Article 21 Right to object	Data subjects' requests to rectify or to erase personal data cannot be completed. Users are not aware of how their data will be used, retained, or shared by the provider.	✓ LLM as a Service ✓ LLM 'off-the-shelf' ✓ Self-developed LLM ✓ Agentic LLM	✓	✓
8. Unlawful repurpose of personal data.	Personal data is used for a different purpose.	Infringement of: Art. 5(1)(b) Purpose limitation, Art. 5(1)(a) Lawfulness, fairness and transparency, Article 28(3)(a)[205] and Art. 29 Processing under the authority of the controller or processor	This could be the case if the provider uses the input and/or output data for training the LLM without this being formally agreed on beforehand.	✓ LLM as a Service ✓ LLM 'off-the-shelf' ✓ Self-developed LLM ✓ Agentic LLM	✓	✓

[204] Under the exceptions outlined in Article 22(2) of the GDPR, automated individual decision-making is permitted only if it is based on contractual necessity, explicit consent, or if authorized by EU or Member State law.

[205] "processes the personal data only on documented instructions from the controller, including with regard to transfers of personal data to a third country or an international organisation, unless required to do so by Union or Member State law to which the processor is subject; in such a case, the processor shall inform the controller of that legal requirement before processing, unless that law prohibits such information on important grounds of public interest;"

9. Unlawful unlimited storage of personal data.	Input and/or output data is being stored longer than necessary.	Infringement of: Art. 5(1)(e) Storage limitation and Art. 25 Data protection by design and by default	The system could be unnecessarily storing input data that is not directly relevant to the LLM process. In some cases, the output could be stored by the deployer longer than necessary. Providers can also log user inputs and outputs for debugging or model improvement.	✓ LLM as a Service ✓ LLM 'off-the-shelf' ✓ Self-developed LLM ✓ Agentic LLM	✓	✓
10. Unlawful transfer of personal data.	Data are being processed in countries without an adequate level of protection.	Infringement of: Art. 44 General principle for transfers, Art. 45 Transfers on the basis of an adequacy decision, Art. 46 Transfers subject to appropriate safeguards	LLM providers could be processing the data in countries that do not offer enough safeguards[206].	✓ LLM as a Service ✓ LLM 'off-the-shelf' ✓ Agentic LLM	✓	✓
11. Breach of the data minimization principle.	Extensive processing of personal data for training the model.	Infringement of: Art. 5(1)(c) Data minimization, Art. 6 to the extent that data minimisation has an impact on the most appropriate lawful basis (e.g., legitimate interest under Article 6(1)(f)) and Art. 25 Data protection by design and by default	LLMs require substantial amounts of data for training. Similarly, deployers may use datasets to fine-tune the LLM based system for their specific use cases.	✓ LLM as a Service ✓ LLM 'off-the-shelf' ✓ Self-developed LLM ✓ Agentic LLM	✓	✓

[206] Garante per la Protezione dei dati personali (DPDP), 'Intelligenza artificiale: il Garante privacy blocca DeepSeek' (2025) https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10097450?mkt_tok=MTM4LUVaTS0wNDIAAAGYXIH0PW4qTzz-TKclqJPRoyU5yZoUVox1JLxNIcVP7RTnC_bvlu_rRyXg8hy6RdOqFw9BgFYU8wXP1XmPVVBTU7DCNt1660jK9umFkCSnLY4e#english

When assessing the risks associated with LLMs, it is crucial to consider broader issues linked to GDPR principles such as lawfulness, fairness, transparency, and accountability. In addition to privacy concerns, also issues related to copyright, overreliance and manipulation must be addressed.

Lawfulness, Transparency and Fairness

Transparency ensures individuals understand how their data is processed, while fairness demands that data is handled in a just and non-deceptive manner, avoiding methods that are detrimental, unlawfully discriminatory, or misleading. However, the opacity inherent in LLMs often challenges these principles. For instance, users may struggle to understand how LLMs generate responses, prioritize outputs, or make decisions, making it difficult for them to assess or challenge the results. The principle of fairness, closely connected to transparency, emphasizes the importance of providing clear, accessible, and meaningful information about data processing activities. Compliance with transparency obligations, as outlined in Articles 12 to 14 GDPR, involves detailing the logic, significance, and potential consequences of automated decision-making, including profiling. This is particularly critical for LLMs, given their complexity and the extensive use of data during development. Reliance on exceptions such as Article 14(5)(b) GDPR is strictly limited to scenarios where all legal requirements are fully met.[207]

To align with these principles, LLM developers must actively monitor outputs, address potential biases[208], by using high-quality and unbiased training data, and provide user-friendly, comprehensible information about the system's decision-making processes. These steps not only ensure compliance with GDPR but also uphold fairness and transparency, fostering trust in AI technologies and safeguarding individual rights.

Copyright[209]

LLMs trained on web-scraped or publicly available data often include copyrighted materials, raising concerns about intellectual property violations. Outputs generated by such models may unintentionally replicate protected content, creating legal risks for both providers and deployers. These issues highlight the importance of ensuring that data used to train LLMs is collected and processed lawfully and in accordance with copyright laws.

Overreliance & Manipulation

Overreliance[210] can undermine user autonomy and accountability. LLMs can unintentionally influence user behavior through tailored recommendations or inferred preferences, reducing autonomy. Users may trust LLM generated outputs in critical areas, such as financial advice or healthcare decisions, without sufficient understanding or oversight. This overreliance can obscure errors or biases in the system, leading to decisions that are neither fair nor transparent. LLMs can also create realistic fake content, such as deepfakes, spreading misinformation or manipulating opinions.
Some mitigation strategies include requiring human oversight for critical decisions, promoting digital literacy, and ensuring that outputs are clearly labeled as AI-generated recommendations.

[207] EDPB, 'Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Adopted on 17 December 2024, (2024) https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en

[208] Lareo, X. 'Large Language Models', EDPS, (n.d) https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/large-language-models-llm_en

[209] European Innovation Council and SMEs Executive Agency, 'Artificial intelligence and copyright: use of generative AI tools to develop new content' (2024) https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0_en

[210] Jacobi, O., 'The Risks of Overreliance on Large Language Models (LLMs)' (2024) https://www.aporia.com/learn/risks-of-overreliance-on-llms/

5. Data Protection and Privacy Risk Assessment: Risk Estimation & Evaluation

From Risk Identification to Risk Evaluation

Once risks have been identified, the next crucial steps within the risk analysis phase are the estimation and evaluation of the risks. This involves the classification and prioritization of risks based on their probability[211]and severity or potential impact. The actual risk level or classification will depend heavily on the specific use case, operational context, system monitoring, model evaluation results and the affected stakeholders.

During this phase, risks are analyzed to understand their implications in greater detail. This process includes evaluating factors such as the probability of the risk occurring, the potential harm it could cause, and the vulnerabilities that make it possible.

Stakeholder collaboration[212] plays a vital role in this process, particularly given the multidisciplinary nature of AI, where inputs from technical, legal, ethical, security and operational perspectives are crucial for comprehensive risk management.

An ethical matrix[213] can be a valuable tool for identifying which stakeholders could be directly or indirectly impacted by the LLM based system. Mapping stakeholders based on their level of involvement and the potential impact on them allows organizations to incorporate the perspectives and concerns of those affected into the risk classification and mitigation process. This ensures that the ethical and practical considerations of deploying LLMs are addressed, aligning the system's implementation with the needs and rights of all impacted parties.

After identifying risks, the next steps are:[214]

- Assessing the probability and severity of the identified risks.
- Evaluate if risks need to be treated to ensure the protection of personal data and demonstrate compliance with the GDPR and EUDPR.

Various risk management methodologies are available for classifying and assessing risks. This document does not aim to prescribe or define a specific methodology, as the choice should be determined by each organization. However, for the purposes of this document, we will reference international standards previously highlighted in the WP29[215] and the AEPD[216] guidelines as well as the work being currently done in European AI standardization.

In general risk management terms, risk can be expressed as:

Risk = Probability x Severity

[211] Note: In this document, we use the term "probability" instead of "likelihood" to align with terminology found in definitions like the one for risk in the AI Act. While in risk management, "likelihood" typically indicates a qualitative approach to managing risks, "probability" implies a quantitative method of risk assessment.

[212] European Center for Not-for-Profit Law, 'Framework for Meaningful Engagement: Human rights impact assessments of AI' (2023) https://ecnl.org/publications/framework-meaningful-engagement-human-rights-impact-assessments-ai

[213] O'Neil, C. 'Algorithmic Stakeholders: An Ethical Matrix for AI' (2020) https://blog.dataiku.com/algorithmic-stakeholders-an-ethical-matrix-for-ai

[214] Article 29 Data Protection Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk for the purposes of Regulation 2016/679' (2017) https://ec.europa.eu/newsroom/article29/items/611236/en

[215] ISO 31000:2009, Risk management — Principles and guidelines, International Organization for Standardization (ISO); ISO/IEC 29134, Information technology – Security techniques – Privacy impact assessment – Guidelines, International Organization for Standardization (ISO).

[216] ISO 31010:2019, Risk management — Risk Assessment Techniques, International Organization for Standardization (ISO)

This equation highlights that risk is determined by the probability of an event occurring, combined with the potential impact or severity of the resulting harm.
Risk is defined in the GDPR (Recital 75) as the potential harm to the rights and freedoms of natural persons, of varying probability and severity, arising from personal data processing. Similarly, the AI Act (Article 3) defines risk as 'the combination of the probability of an occurrence of harm and the severity of that harm;'.
To evaluate the level of data protection and privacy risks when procuring, developing, or using LLMs, it is essential to estimate both the probability and severity of the identified risks materializing.

Criteria to Establish the Probability of Risks in LLM Systems

How to Assess Probability

To determine the probability of the risks of LLMs we will use the following four level risk classification matrix:

Level of Consequence	**Probability Definition**
Very High	High probability of an event occurring
High	Substantial probability of an event occurring
Low	Low probability of an event occurring
Unlikely	There is no evidence of such a risk materializing in any case

Probability determination must be tailored to the specific risks and use cases under assessment. While this general matrix provides a structured approach, applying more detailed criteria can enhance the accuracy of the probability assessment.

In the table below, there is an example of criteria[217] that can guide this process, helping to refine the evaluation of probability for specific scenarios. Note that some criteria relate to system-level attributes while other are context-specific.

		PROBABILITY LEVELS			
Criteria	**Description**	**Level 1 (Unlikely)**	**Level 2 (Low)**	**Level 3 (High)**	**Level 4 (Very High)**
1. Frequency of Use	How often the AI system is used, increasing exposure to potential risk affecting reliability (expected time before failure)	The system is rarely used or has infrequent interactions (e.g., annual or less).	The system is occasionally used but not in critical operations (e.g., monthly).	The system is frequently used and integrated into important operations (e.g., weekly).	The system is used continuously or in real-time critical operations (e.g., daily).
2. Exposure to High-Risk Scenarios	The extent to which the AI system operates in sensitive or high-stakes environments.	The system is not used in sensitive or high-stakes scenarios.	The system operates in moderately sensitive environments with minimal stakes.	The system is used in high-stakes environments with potential for significant impact.	The system operates in highly sensitive or critical environments (e.g., healthcare, security).

[217] Barberá, I. "FRASP, A Structured Framework for Assessing the Severity & Probability of Fundamental Rights Interferences in AI Systems" (2025)

3. Historical Precedents	Past instances of similar risks or failures in the same or comparable AI systems.	No similar risks or failures have occurred in comparable systems.	Few similar risks or failures have occurred in comparable systems.	Similar risks or failures have occurred frequently in comparable systems.	Frequent and significant risks or failures have occurred in comparable systems.
4. Environmental Factors	External, uncontrollable conditions affecting system performance or reliability (e.g., political instability, regulatory gaps, financial constraints).	External conditions are stable and do not impact the system's performance.	External conditions occasionally affect the system's performance but are manageable.	External conditions often impact the system's performance, creating vulnerabilities.	External conditions severely affect the system's performance, creating constant risks.
5. System Robustness	The degree to which the AI system is resistant to failure or unintended behaviour.	The system is highly robust with multiple redundancies and safeguards.	The system is moderately robust with some redundancies but occasional vulnerabilities.	The system has some robustness but contains significant vulnerabilities or weak safeguards.	The system lacks robustness, safeguards, or is prone to frequent failures.
6. Data Quality and Integrity	The extent to which the AI system relies on accurate, unbiased, and complete data. Modifiable through better dataset curation or validation.	Data is highly accurate, unbiased, and complete with minimal risk of errors.	Data is mostly accurate and complete but has occasional minor biases or errors.	Data is partially accurate or complete, with notable biases or inconsistencies.	Data is significantly inaccurate, biased, or incomplete, leading to high risk.
7. Human Oversight and Expertise	How human operators' skills and decision-making affect system reliability and risk probability. Modifiable through training or oversight improvements.	Operators are highly trained, experienced, and consistently effective in decision-making.	Operators are moderately trained and effective, but occasional errors occur.	Operators are undertrained or inconsistent, leading to regular errors in decision-making.	Operators are untrained or ineffective, causing frequent and severe errors.

To use the criteria for determining the probability of risks you can do the following:

Step 1: Aggregate Scores

- Evaluate each criterion and assign it a score based on predefined probability levels 1 to 4.
- Add the scores of all factors and divide the total by the number of factors to calculate the Aggregate Probability Score. This can be done using either:

    - Weighted Average: Assign more importance to certain factors by weighting them before averaging.
    - Simple Average: Treat all factors equally and calculate the mean.

Formula: **Aggregate Probability Score = Sum of All Scores / Number of Factors**

Step 2: Map Aggregate Score to Probability Level

Once the aggregate score is calculated, map it to one of the predefined probability levels from the matrix based on the following ranges:

1.0 - 1.5: Unlikely
1.6 - 2.5: Low
2.6 - 3.5: High
3.6 - 4.0: Very High

This mapping provides a clear, categorized probability level for each risk, which helps prioritize risks based on their potential occurrence. We will explore later in this document how this framework can be applied in practice in one specific use case.

Criteria to Establish the Severity of Risks in LLM Systems

How to Assess Severity

To determine the severity of risks of LLMs we will use a four level risk classification matrix:[218]

Level of Severity	**Severity Definition**
Very Significant *Catastrophic Harm*	It affects the exercise of fundamental rights and public freedoms, and its consequences are irreversible and/or the consequences are related to special categories of data or to criminal offences and are irreversible and/or it causes significant social harm, such as discrimination, and is irreversible and/or it affects particularly vulnerable data subjects, especially children, in an irreversible way and/or causes significant and irreversible moral or material losses.
Significant *Critical Harm*	The above cases when the effects are reversible and/or there is loss of control of the data subject over their personal data, where the extent of the data are high in relation to the categories of data or the number of subjects and/or identity theft of data subjects occurs or may occur and/or significant financial losses to data subjects may occur and/or loss of confidentiality of data subject or breach of the duty of confidentiality and/or there is a social detriment to data subjects or certain groups of data subjects
Limited *Serious Harm*	Very limited loss of control of some personal data and to specific data subjects, other than special category or irreversible criminal offences or convictions and/or negligible and irreversible financial losses and/or loss of confidentiality of data subject to professional secrecy but not special categories or infringement penalties
Very Limited *Moderate or Minor Harm*	In the above case (limited) when all effects are reversible

Note that the HUDERIA[219] risk management methodology, developed by the Committee on Artificial Intelligence of the Council of Europe, also employs a four-level severity matrix. However, it uses slightly different terminology, as shown in this matrix (italicized): *Catastrophic Harm, Critical Harm, Serious Harm,* and *Moderate or Minor Harm.*

Similar to the assessment of probability, the assessment of severity can also benefit from the use of different severity criteria[220] to reduce subjectivity in the process. The severity criteria are related to a loss of privacy that is experienced by the data subject but that may have further related consequences impacting other individuals and/or society.

The table below outlines different severity[221] criteria. The calculation of severity can follow the same steps as those used for determining probability, including aggregating scores and mapping them to severity levels. However, for severity, certain criteria (numbers 1 to 5, and 7 & 8) act as "stoppers" This means that the end score will always be the highest one from those criteria no matter what the

[218] AEPD, 'Risk Management and Impact Assessment in Processing of Personal Data' p - 77 (2021) https://www.aepd.es/guides/risk-management-and-impact-assessment-in-processing-personal-data.pdf

[219] Council of Europe (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 (Huderia Methodology)' (2024) https://rm.coe.int/cai-2024-16rev2-methodology-for-the-risk-and-impact-assessment-of-arti/1680b2a09f

[220] (...) 7/ Risks, which are related to potential negative impact on the data subject's rights, freedoms and interests, should be determined taking into consideration specific objective criteria such as the nature of personal data (e.g. sensitive or not), the category of data subject (e.g. minor or not), the number of data subjects affected, and the purpose of the processing. The severity and the probability of the impacts on rights and freedoms of the data subject constitute elements to take into consideration to evaluate the risks for individual's privacy'', p.4, Article 29 Working Party (WP 208) ''Statement on the role of risk-based approach in data protection legal frameworks'', Adopted on 30 May 2014, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf

[221] See footnote 217

aggregation score is. For instance, if any of these criteria are assessed at the highest level (4), the overall severity score is immediately assigned a level 4. This approach ensures that critical harms, such as those involving irreversible damage, are appropriately prioritized and flagged for immediate and comprehensive mitigation measures.

		SEVERITY LEVELS			
Criteria	**Description**	**Level 1 (Very Limited)**	**Level 2 (Limited)**	**Level 3 (Significant)**	**Level 4 (Very Significant)**
		Moderate or Minor Harm	**Serious Harm**	**Critical Harm**	**Catastrophic Harm**
		Moderate or minor prejudices or impairments in the exercise of fundamental rights and freedoms that do not lead to any significant, enduring, or temporary degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order.	Serious prejudices or impairments in the exercise of fundamental rights and freedoms that lead to the temporary degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order or that harm to the information and communication environment.	Critical prejudices or impairments in the exercise of fundamental rights and freedoms that lead to the significant and enduring degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order.	Catastrophic prejudices or impairments in the exercise of fundamental rights and freedoms that lead to the deprivation of the right to life; irreversible injury to physical, psychological, or moral integrity; deprivation of the welfare of entire groups or communities; catastrophic harm to democratic society, the rule of law, or to the preconditions of democratic ways of life and just legal order; deprivation of individual freedom and of the right to liberty and security; harm to the biosphere.
1. Nature of the fundamental right and Legal limitation alignment	This criterion evaluates the nature of the fundamental right affected—whether it is absolute or subject to limitations—and assesses the extent to which the AI system's use case aligns with lawful and proportionate restrictions. Absolute rights are non-derogable and cannot be restricted under any circumstances, while other rights may be limited only if the interference meets strict legal, proportionality, and necessity requirements. This criterion helps determine the severity of the impact based on the degree of misalignment or violation of the right's protections.	The fundamental right affected is highly limited in scope and applicability, meaning it is frequently subject to lawful restrictions with minimal requirements to justify the interference. The use case clearly and fully aligns with permitted legal limitations, and the interference is routine and widely accepted, without causing significant violations of legal or normative frameworks.	The fundamental right affected is moderately limited, meaning restrictions are lawful but subject to stricter justification requirements and more specific conditions. The use case aligns with legal limitations, but the interference requires a moderate level of justification, such as demonstrating proportionality and necessity, otherwise causing possible minor violations of legal or normative frameworks.	The fundamental right affected is minimally limited, meaning restrictions are only lawful under exceptional and tightly controlled circumstances. The use case partially aligns with lawful exceptions, but there are uncertainties about the proportionality, necessity, or legitimacy of the interference, causing possible major violations of legal or normative frameworks.	The fundamental right affected is absolute and non-derogable, meaning no lawful restriction is permitted under any circumstances. Alternatively, the use case does not align with lawful and proportionate limitations, even if the right is not absolute, causing severe violations of legal or normative frameworks. * The Charter does not explicitly identify the rights that are absolute. Based on the Charter explanations, the ECHR and the case law of the European courts, it is submitted that human dignity (Article 1 of the Charter), the prohibition of torture and inhuman or degrading treatment or punishment (Article 4 of the Charter), the prohibition of slavery and forced labour (Article 5(1) and (2) of the Charter), internal freedom

					of thought, conscience and religion (Article 10(1) of the Charter), the presumption of innocence and right of defence (Article 48 of the Charter), the principle of legality (Article 49(1) of the Charter) and the right not to be tried or punished twice in criminal proceedings for the same criminal offence (Article 50 of the Charter) can be considered absolute rights.
2. Nature of personal data	This criterion assesses the sensitivity of the personal data being processed, considering its potential to cause harm if misused. Special category of data (e.g., health, biometric, or genetic data) poses greater risks to fundamental rights like privacy and autonomy.	Non-sensitive, publicly available data (e.g., anonymized data, public records).	Moderately sensitive data (e.g., financial data, browsing history).	Highly sensitive data	The most sensitive data & special category of data e.g., genetic data, psychological profiles, biometrics or data revealing criminal history.
3. Category of Data Subject (e.g., minor or not)	This criterion evaluates the vulnerability of the individuals whose data is being processed. Vulnerable groups (e.g., minors, marginalized communities) face greater risks of harm from data misuse.	Data subjects are not vulnerable (e.g., adults in routine, non-sensitive contexts).	Data subjects include some individuals in potentially vulnerable groups (e.g., employees, customers).	Data subjects include individuals in sensitive or high-risk roles (e.g., journalists, activists).	Data subjects are highly vulnerable (e.g., minors, persons with disabilities, or persecuted groups).
4. Purpose of Processing	This criterion evaluates the legitimacy, necessity, and proportionality of the purpose for which personal data is being processed. Unlawful or disproportionate purposes increase severity.	Clearly legitimate and proportionate purposes with minimal risks (e.g., operational purposes).	Legitimate purposes with moderate risks or indirect impacts (e.g., targeted marketing).	Legitimate purposes but with questionable proportionality or necessity (e.g., profiling for credit scoring).	Unlawful, unclear, or disproportional purposes with significant risks (e.g., surveillance, discriminatory profiling).
5. Scale of Impact (Societal, Group, Individual) & Number of Data Subjects Affected	The breadth of the infringement across societal, group, and individual levels. This criterion considers the scale of the impact based on the number of individuals whose data is affected.	Impact is limited to a small, localized group or individual. Fewer than 100 individuals affected.	Impact is limited to specific groups or a small societal segment. Between 100 and 1,000 individuals affected.	Impact spans multiple groups or societal domains. Between 1,000 and 100,000 individuals affected.	Impact is widespread, affecting societal, group, and individual levels. Over 100,000 individuals affected.

6. Contextual and Domain Sensitivity	How specific contextual factors or domains intensify the interference's severity. Includes circumstantial risks like socio-political instability and if children and other vulnerable groups are affected.	Context or domain does not amplify the severity of the interference with the fundamental right.	Context or domain moderately amplifies the severity of the interference with the fundamental right.	Context or domain significantly amplifies the severity of the interference with the fundamental right.	Context or domain profoundly amplifies the severity of the interference with the fundamental right.
7. Reversibility, recovery, degree of remediability	The difficulty or feasibility of reversing harm and the time required for recovery. Includes prohibitive risks where harm is irreversible.	Harm is fully reversible within a short period with minimal effort.	Harm is reversible with moderate effort over a reasonable timeframe.	Harm is difficult to reverse, requiring significant effort or resources.	Harm is irreversible, with no feasible means of recovery.
8. Duration and Persistence of Harm	The length of time and persistence of adverse effects caused by the interference.	Adverse effects are minimal and do not persist over time.	Adverse effects persist briefly but do not result in long-term consequences.	Adverse effects persist for a considerable period and can affect multiple groups.	Adverse effects are permanent or persist indefinitely.
9. Velocity to materialise	The speed at which the risk materialises: gradual, sudden, continuously changing.	Risk materialises gradually, providing sufficient time for intervention.	Risk materialises at a moderate pace, allowing for corrective measures.	Risk materialises suddenly, leaving limited time for intervention.	Risk materialises rapidly, with no opportunity for intervention.
10. Transparency and mechanisms for Accountability	The degree of system transparency and mechanisms for accountability.	System is highly transparent with clear and effective accountability mechanisms.	System lacks some transparency but has basic accountability mechanisms.	System lacks transparency and has weak accountability mechanisms.	System is entirely opaque, with no mechanisms for accountability.
11. Ripple and Cascading Effects	The extent to which the interference triggers additional harms across systems or domains.	No cascading effects; the risk is isolated and contained.	Minimal cascading effects; impacts are mostly contained.	Notable cascading effects; impacts extend across domains.	Severe cascading effects; impacts propagate extensively.

Risk Evaluation: Classification of Risks

Assessing probability and severity provides the foundation for determining the overall risk level of the identified privacy and data protection risks. Using a four-level classification matrix for both probability and severity, risks can be categorized into final classifications of Very High, High, Medium, or Low.

A matrix, as shown below, serves as a practical tool to obtain these classifications, offering a clear and structured ranking to prioritize risks and guide appropriate mitigation strategies. This classification is a critical step in the next risk treatment process because it ensures that resources are directed toward addressing the most pressing risks effectively.

Probability	**Very High**	Medium	High	Very high	Very high
	High	Low	High	Very high	Very high
	Low	Low	Medium	High	Very high
	Unlikely	Low	Low	Medium	Very high
		Very limited	**Limited**	**Significant**	**Very Significant**
		Severity			

Figure 15. Risk Evaluation Matrix

Best practices in risk management suggest that the mitigation of very high and high level risks should be prioritized.[222] Once these critical risks are identified, the next essential step is to develop and implement a risk treatment plan.

Risk Acceptance Criteria

In the Risk Evaluation phase, risk criteria are used to determine whether a risk is acceptable or needs treatment. These criteria reflect the organization's willingness and capacity to bear risks within legal and operational limits and must align with applicable laws and regulations, such as GDPR and AI Act requirements, to ensure compliance and the protection of individuals' rights.
Different frameworks and best practices[223] can assist organizations in defining these criteria. These can be establish considering factors such as social norms, expected benefits, potential harms, metrics established thresholds, evaluation results, and comparable use cases[224]. Justifying these decisions is critical, as organizations are accountable for demonstrating how risks are managed and mitigated. This aligns with the GDPR principle of accountability[225], which requires organizations to document and justify their risk mitigation and acceptance decisions.

[222] Oliva, L., ' Successfully managing high-risk, critical-path projects' (2003) https://www.pmi.org/learning/library/high-risk-critical-path-projects-7675

[223] Marsden, E. 'Risk acceptability and tolerability' (n.d) https://risk-engineering.org/static/PDF/slides-risk-acceptability.pdf

[224] Science Direct, 'Definition of Residual Risk' (2019) https://www.sciencedirect.com/topics/engineering/residual-risk

[225] Article 5(2), Recital 74, GDPR

6. Data Protection and Privacy Risk Control

Risk Treatment Criteria

i.e., mitigate, transfer, avoid or accept a risk.

Risk treatment involves developing strategies to mitigate identified risks and creating actionable implementation plans. The choice of an appropriate treatment option should be context-specific, guided by a feasibility analysis[226] such as the following:

- Evaluate the type of risk and the available mitigation measures that can be implemented.
- Compare the potential benefits gained from implementing the mitigation against the costs and efforts involved and the potential impact.
- Assess the impact on the intended purpose of the LLM system's implementation.
- Consider the reasonable expectations of individuals impacted by the system.
- Perform a trade-off analysis to evaluate the impact of potential mitigations on aspects such as performance, transparency, and fairness, ensuring that processing remains ethical and compliant based on the specific use case.

Analyzing these criteria is essential for effective risk mitigation and risk management planning, providing clarity on whether specific mitigation efforts are justifiable. In all cases, the chosen treatment option should be clearly justified and thoroughly documented to ensure accountability and compliance.

The most common risk treatment criteria are: **Mitigate, Transfer, Avoid** and **Accept**.
For each identified risk one of the criteria options will be selected:

- ✓ Mitigate – Implement measures to reduce the probability or the severity of the risk.
- ✓ Transfer – Shift responsibility for the risk to another party (e.g., through insurance or outsourcing).
- ✓ Avoid – Eliminate the risk entirely by addressing its root cause.
- ✓ Accept – Decide to take no action, accepting the risk as is because it falls within acceptable limits as defined in the risk criteria.

Deciding whether a risk can be mitigated involves assessing its nature, potential impact, and available mitigation measures such as implementing controls, adopting best practices, modifying processes, or using tools to reduce the probability or severity of the risk.
Not all risks can be fully mitigated. Some risks may be inherent and cannot be entirely avoided. In such cases, the objective is to reduce the risk to an acceptable level or implement risk mitigation and control measures that effectively manage its impact.

It is also important to maintain a dynamic risk register, containing risk records that are durable, easily accessible, clear, and that are consistently updated to ensure accuracy and relevance[227].
Risks should also have clear ownership assigned, and regular reviews should be conducted to ensure that risk management practices remain proactive.

[226] Centre for Information Policy Leadership, 'Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR. GDPR Interpretation and Implementation Project' (2016) https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_gdpr_project_risk_white_paper_21_december_2016.pdf

[227] Ofcom, 'Protecting people from illegal harms online' (2024) https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/volume-1-governance-and-risks-management.pdf?v=387545

Example of Mitigation Measures Related to Risks of LLM Systems

Choosing appropriate mitigation measures should be done on a case-by-case basis. The table below contains some of the possible mitigation measures that could be implemented to address LLMs privacy and data protection risks. These measures are general and not tied to any specific use case's context and intended purpose. Mitigation's recommendations may apply to both provider and deployers. In some cases, mitigations also apply to the deployer if they modify the model.

Data Protection and Privacy Risks	Recommended Mitigations	Provider	Deployer
1. Insufficient protection of personal data what eventually can be the cause of a data breach Specific to RAG: Insufficient protection in RAG systems what can lead to data breaches, indirect prompt injection, and the retrieval of outdated or inaccurate information, resulting in poor decision-making and potential harm to individuals. Specific queries might also inadvertently disclose personal data included in training, violating privacy regulations.[228]	As provider and deployer, it is important to verify[229] that: ▪ APIs are securely implemented (Use API gateways with rate limiting and monitoring capabilities to control and monitor access) ▪ Transmission of data are protected with the adequate encryption protocols, data at rest is encrypted. ▪ There is an adequate access control mechanism implemented. ▪ There are measures implemented for anonymization and pseudonymization of personal data, or for masking of data or use of synthetic data. ▪ Additional technical and organizational measures (TOMs) are implemented to further enhance security. These measures may include, but are not limited to, ensuring a secure environment through data segregation, backups, regular physical and digital security audits, incident response mechanisms, awareness and training. ▪ A Defense in Depth[230] approach can be implemented by layering multiple risk mitigation measures to prevent single points of failure. This may include model-level protections, network security, user authentication, encryption, PETs[231], access control and continuous monitoring to detect misuse, bias, or vulnerabilities in real time. ▪ To mitigate memorization risk, implement differential privacy techniques to prevent sensitive data encoding and regularly test for data regurgitation. Consider also using smaller models to avoid the memorization[232] effect from overparametrized model. ▪ Also measures for protection and identification of insider threats, measures to mitigate supply chain attacks that could give access to the training data and/or the data storage and encryption keys, measures implemented to prevent risks associated to different LLM security threats[233][234] such as membership inference[235], model inversion[236] and poisoning attacks[237]. ▪ Also access and change logs are established to document access and changes to digitized records.	✓	✓

[228] EDPS, 'TechSonar 2025 Report' (2025) https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_en

[229] This could be done by performing a pentest and/or requesting pentest results to the vendor.

[230] AI Action Summit, 'International AI Safety Report on the Safety of Advanced AI' , p - 167, (2025) https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

[231] Feretzakis, G et al., 'V.S. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review'(2024). https://doi.org/10.3390/info15110697

[232] Examples of Memorization methods: https://blog.kjamistan.com/category/ml-memorization.html

[233] OWASP, 'OWASP Top 10 for LLM Applications 2025' (2025) https://genai.owasp.org/llm-top-10/

[234] Shamsabadi, S.A. et al., ' Identifying and Mitigating Privacy Risks Stemming from Language Models' (2024) https://arxiv.org/html/2310.01424v2

[235] Shokri et al., 'Membership Inference Attacks Against Machine Learning Models' (2017) https://arxiv.org/abs/1610.05820

[236] Zhang et al.,'Generative Model-Inversion Attacks Against Deep Neural Networks', (2020) https://arxiv.org/abs/1911.07135

[237] Guo, J. et al., 'Practical Poisoning Attacks on Neural Networks', (2020) https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123720137.pdf

	▪ Employees and users are trained on security best practices. ▪ Effective RAG systems require careful model alignment to prevent unauthorized access and sensitive data exposure. Integration with multiple data sources necessitates robust security measures to ensure confidentiality and data integrity, while adhering to data protection principles like necessity and proportionality. For outsourced RAG models involving personal data transfer, compliance with GDPR's data transfer rules is critical to maintaining confidentiality and legal obligations.[238]		
2. Misclassifying training data as anonymous by controllers when it contains identifiable information, leading to failure to implement appropriate safeguards for data protection. (partly relating to risk 3) *Whenever information relating to identified or identifiable individuals whose personal data was used to train the model may be obtained from an AI model with means reasonably likely to be used*[239]*, it may be concluded that such a model is not anonymous.*[240]	▪ Implement robust testing and validation processes to ensure that (i) personal data associated with the training data cannot be extracted from the model using reasonable means, and (ii) any outputs generated by the model do not link back to or identify data subjects whose personal data was used during training. ▪ This assessment should be done taking into account 'all the means reasonably likely to be used' considering objective factors such as:[241] ○ The characteristics of the training data, the AI model, and the training procedure. ○ The context in which the AI model is released or processed. ○ The availability of additional information that could enable identification. ○ The costs and time required to access such additional information, if not readily available. ○ Current technological capabilities and potential future advancements. ▪ Implement alternative approaches to anonymization if they provide an equivalent level of protection, ensuring they align with the state of the art. ▪ Implement structured testing against state of the art attacks such as attribute and membership inference, exfiltration, regurgitation of training data model inversion, or reconstruction attacks. ▪ Document and retain evidence to demonstrate compliance with these safeguards following accountability obligations under Article 5(2) GDPR. Documentation should include: ○ Details of DPIAs, including assessments and decisions on their necessity. ○ Advice or feedback from the DPO (if applicable). ○ Information on technical and organizational measures to minimize identification risks during the model design, including threat models and risk assessments for training datasets (e.g., source URLs and safeguards). ○ Measures taken throughout the AI model lifecycle to prevent or verify the absence of personal data in the model. ○ Evidence of the model's theoretical resistance to re-identification techniques, including metrics, testing reports, and analysis of attack resistance (e.g., regurgitation, membership inference). ○ Documentation provided to controllers and data subjects detailing measures to reduce identification risks and addressing potential residual risks.	✓	? [242]

[238] https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_en

[239] Membership Inference Attacks and Model Inversion Attacks.

[240] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted on 17 December 2024

[241] idem

[242] Deployers should verify that the provider has effectively addressed this risk. This recommendation is equally relevant in cases where deployers are involved in fine-tuning or retraining models.

3. Unlawful processing of personal data in training sets	▪ Document all training data sources (e.g., book databases, websites) to ensure accountability under Art. 5(2) GDPR. ▪ Check training data for statistical distortions or biases and make necessary adjustments. ▪ Exclude training data that includes unauthorized content, such as fake news, hate speech, or conspiracy theories. ▪ Exclude content from publications that may contain personal data posing risks to individuals or groups, such as those vulnerable to abuse, prejudice, or harm. ▪ Remove unnecessary personal data (e.g., credit card numbers, email addresses, names) from the training dataset.[243] ▪ Employ methodological choices that significantly reduce or eliminate identifiability, such as using regularization methods to enhance model generalization and minimize overfitting. ▪ Implement robust privacy-preserving techniques, such as differential privacy.[244] ▪ When using web scraping as a method to collect data, ensure compliance with Article 6(1)(f) GDPR by conducting a thorough legal assessment. This includes evaluating: ○ (i) the existence of a legitimate interest for data processing. Interest should be lawful, clearly articulated and real, not speculative. ○ (ii) the necessity of the processing, ensuring that personal data collected is adequate, relevant, and limited to what is necessary for the stated purpose[245], and ○ could not reasonably be fulfilled by other means' ○ (iii) a careful balancing of interests, where the fundamental rights and freedoms of data subjects are weighed against the legitimate interests of the data controller. Consideration should also be given to the reasonable expectations of data subjects regarding the use of their data.[246] ▪ Involve the DPO in the balancing test, where applicable[247]. ▪ For web scraping, assess whether the exemption under Article 14(5)(b) applies, ensuring all criteria are met to justify not informing each data subject individually. Transparency: ▪ Provide public and easily accessible information that goes beyond GDPR requirements under Articles 13 and 14, including details about collection criteria and datasets used, with special consideration for protecting children and vulnerable individuals. ▪ Use innovative approaches to inform data subjects, such as media campaigns, email notifications, graphic visualizations, FAQs, transparency labels, model cards, and voluntary annual transparency reports.[248] ▪ Implement an opt-out list managed by the controller, enabling data subjects to object to the collection of their data from specific websites or platforms by providing identifying information before data collection begins.	✓	?[249]

[243] Bavarian State Office for Data Protection Supervision, 'Data protection compliant Artificial intelligence Checklist with test criteria according to GDPR'

[244] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Adopted on 17 December 2024.

[245] Recital 39 GDPR clarifies that 'Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means'

[246] EDPB Report of the work undertaken by the ChatGPT Taskforce (2024)

[247] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted on 17 December 2024.

[248] Idem

[249] Deployers should verify that the provider has effectively addressed this risk. This recommendation is equally relevant in cases where deployers are involved in fine-tuning or retraining models.

4. Unlawful processing of special categories of personal data and data relating to criminal convictions and offences in training data.	▪ For the lawful processing of special categories of personal data, ensure that an exception under Article 9(2) GDPR applies[250]. When relying on Article 9(2)(e), confirm that the data subject explicitly and intentionally made the data publicly accessible through a clear affirmative action. The mere fact that personal data is publicly accessible does not imply that the data subject has manifestly made such data public[251]. ▪ Given the challenges of case-by-case assessment in large-scale web scraping, implement safeguards such as filtering to exclude data falling under Article 9(1) GDPR both during and immediately after data collection. ▪ Maintain robust documentation and proof of these measures to comply with accountability requirements under Articles 5(2) and 24 GDPR.[252]	✓	? [253]
5. Possible adverse impact on data subjects that could negatively impact fundamental rights: LLM outputs could contain biased and inaccurate information, potentially violating the GDPR principles of accuracy, transparency, lawfulness and fairness and misleading users into treating incorrect outputs as factually reliable. All this could have an adverse impact on data subjects and their fundamental rights.	Transparency: ▪ Implement robust transparency measures to inform users about the probabilistic nature of LLM outputs and their potential for bias or inaccuracies. Provide explicit disclaimers that the generated content may not be factually accurate or real, ensuring users understand the limitations of the system. ▪ Inform deployers and users whether their personal data will impact the service provided to that specific user or whether it would be used to modify the service provided to all customers. ▪ Users should select LLMs with proven performance metrics and continuously monitor outputs for errors or biases, so transparency publishing this information could be recommended. Accuracy and Fairness: ▪ Developers should ensure the quality of training data through robust preprocessing techniques, such as filtering, validation and normalization, to prevent biased or misleading outputs, they should also assess its impact on AI outputs, evaluate the model's statistical accuracy for its intended purpose, and transparently communicate these considerations to deployers and end users to mitigate potential negative effects during deployment.[254] Regularly review and document all steps taken to comply with the GDPR principles of transparency and accuracy under Articles 5(1)(a) and 5(1)(d).[255] ▪ For third-party platforms, effective configuration and use of available tools are essential to enhance input handling and ensure outputs meet accuracy and fairness standards. ▪ Regular audits and oversight mechanisms are critical to addressing risks like data leakage, bias, or unintended inferences. ▪ LLMs could also be fine-tuned to handle diverse linguistic and contextual variations, reducing inaccuracies in sensitive applications.	✓	? [257]

[250] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted on 17 December 2024: **“**The EDPB recalls the prohibition of Article 9(1) GDPR regarding the processing of special categories of data and the limited exceptions of Article 9(2) GDPR. In this respect, the Court of Justice of the European Union (“CJEU”) further clarified that ‘where a set of data containing both sensitive data and non-sensitive data is [...] collected en bloc without it being possible to separate the data items from each other at the time of collection, the processing of that set of data must be regarded as being prohibited, within the meaning of Article 9(1) of the GDPR, if it contains at least one sensitive data item and none of the derogations in Article 9(2) of that regulation applies’ . Furthermore, the CJEU also emphasised that ‘for the purposes of the application of the exception laid down in Article 9(2)(e) of the GDPR, it is important to ascertain whether the data subject had intended, explicitly and by a clear affirmative action, to make the personal data in question accessible to the general public’ . These considerations should be taken into account when processing of personal data in the context of AI models involves special categories of data.”

[251] EDPB, Report of the work undertaken by the ChatGPT Taskforce

[252] Idem

[253] Deployers should verify that the provider has effectively addressed this risk. This recommendation is equally relevant in cases where deployers are involved in fine-tuning or retraining models.

[254] Information Commissioner Officer (ICO) ‘Generative AI third call for evidence: accuracy of training data and model outputs’ (2025) https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-third-call-for-evidence/

[255] Idem

[257] Deployers should verify that the provider has effectively addressed this risk. This recommendation is equally relevant in cases where deployers are involved in fine-tuning or retraining models.

	▪ To mitigate the risk of adverse impacts on data subjects and fundamental rights in the context of LLMs, accuracy[256] and reliability must be prioritized throughout the system lifecycle. ▪ Ensure that training datasets are diverse and representative of different demographic groups to reduce biases inherent in the data. ▪ Conduct regular audits and fairness tests and incorporate human review in sensitive decisions to ensure fairness and accountability. ▪ Use explainability frameworks to analyze and understand how decisions are made, what helps in identifying potential sources of bias.		
6. Not providing human intervention for a processing that can have a legal or important effect on the data subject.	▪ Human oversight should be integrated into decision-making processes where the outputs of LLMs could lead to legal or significant consequences for individuals[258]. This includes ensuring that automated decisions are subject to review by qualified personnel who can assess the fairness, accuracy, and relevance of the outputs. ▪ Clear escalation procedures should be in place for cases where automated outputs appear ambiguous, erroneous, or potentially harmful. ▪ Developers and deployers must design systems to flag high-risk outputs for mandatory human intervention before any action is taken[259]. ▪ Transparency mechanisms should also be implemented[260], ensuring data subjects are informed about the use of LLMs, the capabilities and limitations of the model[261], the processing of personal data through the model and their right to contest decisions or seek human review. ▪ Regular training for staff involved in oversight can further enhance compliance and accountability. ▪ Implement Article 29 Working Party ("WP29") Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, as last revised and adopted on 6 February 2018, endorsed by the EDPB on 25 May 2018. See also, CJEU judgment of 7 December 2023, Case C-634/21, SCHUFA Holding and Others (ECLI:EU:C:2023:957).	✓	✓
7. Not granting data subjects their right to object, rectification, and erasure.	▪ The right to object under Article 21 GDPR applies and should be ensured when the legal basis is legitimate interest[262]. In such a case, providers should implement mechanisms to grant this right. Some measures to implement when collecting personal data could be[263]: ○ Introduce a reasonable period between the collection of a training dataset and its use, allowing data subjects time to exercise their rights. ○ Provide an unconditional opt-out mechanism for data subjects before processing begins. ○ Permit data subjects to request data erasure, even beyond the specific grounds listed in Article 17(1) GDPR.		

[256] AI Model Code, 'Evaluating language models for accuracy and bias' (2024) https://aimodelcode.org/tech-info/llm-eval/

[258] Lumenova, 'The Strategic Necessity of Human Oversight in AI Systems' (2024) https://www.lumenova.ai/blog/strategic-necessity-human-oversight-ai-systems/

[259] Kuriakose, A.A., ' The Role of Human Oversight in LLMOps' (2024) https://www.algomox.com/resources/blog/what_is_the_role_of_human_oversight_in_llmops/

[260] Garante per la Protezione dei Dati Personali (GDPD), 'ChatGPT, il Garante privacy chiude l'istruttoria. OpenAI dovrà realizzare una campagna informativa di sei mesi e pagare una sanzione di 15 milioni di euro' (2024) https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10085432?mkt_tok=MTM4LUVaTS0wNDIAAAGX5pUM0HSpbBgVFc2wv7uGKk23174FM2-cFJBvVD0FDGJCM_27RuQFPm2uSB80ihorQ2e0YWwgCPRFngJDRE4b7N_pWRz873q84sJ8ZWucdQOh#english

[261] EDPB Report of the work undertaken by the ChatGPT Taskforce (2024)

[262] Note that according to Art. 21(1) GDPR, "The controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims."

[263] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted on 17 December 2024

	o Claim Handling: Enable data subjects to report instances of personal data regurgitation or memorization, with mechanisms for controllers to assess and apply unlearning techniques to resolve such claims. ▪ Mitigating non-compliance with GDPR concerning data subjects' rights to rectification and erasure involves exploring machine unlearning techniques[264]. These approaches aim to remove the influence of data from a trained model upon request, addressing concerns about data use, low-quality inputs, or outdated information. o Exact unlearning seeks to entirely eliminate the influence of specific data points, often through retraining or advanced methods that avoid full retraining. Techniques like Sharded, Isolated, Sliced, and Aggregated (SISA) training divide data into subsets, simplifying data removal while striving to maintain model robustness. Approximate unlearning attempts to reduce the impact of specific data points by adjusting model weights or applying correction factors, offering a trade-off between precision and efficiency. While these methods hold promise, challenges remain, including maintaining model accuracy and avoiding unintended biases post-unlearning. Certified removal, which provides verifiable guarantees of data removal using mathematical proofs, offers a rigorous but resource-intensive solution. As unlearning techniques evolve, they play a crucial role in enabling compliance with GDPR while preserving the integrity and fairness of machine learning models.[265] ▪ Implement mechanisms to delete personal data, such as names, ensuring their removal(block)[266] is comprehensive and context-agnostic across the dataset. Recognize that this approach might result in the deletion of the name for all individuals with the same identifier, regardless of the context. To mitigate unintended consequences, use precise filtering techniques to differentiate between contexts where the name is personally identifiable and generic**.** To prevent misuse or reintroduction of deleted data, secure filter scripts or prompts by restricting access to authorized personnel only, employing encryption, and maintaining version control. Regularly audit these scripts to ensure they are up to date and free from vulnerabilities. ▪ It is also important to in particular with regard to Article 21 GDPR, to establish mechanisms to comply with the requests of users that object to the processing of their personal data based on legitimate interest.[267] ▪ For deletion requests under Art. 17 GDPR, assess whether personal data can be directly identified or derived from the AI model and implement technical deletion where feasible, such as post-training adjustments.[268]	✓	? [269]
8. Unlawful repurpose of personal data	▪ Ensure compliance with Article 5(1)(c) GDPR by clearly limiting personal data processing to what is necessary for specific, well-defined purposes. Avoid overly broad purposes like "developing and improving an AI system." Instead, specify the type of AI system (e.g., large language model, generative AI for images) and its technically feasible functionalities and capabilities.[270]	✓	✓

[264] Shrishak, K., 'AI-Complex Algorithms and effective Data Protection Supervision Effective implementation of data subjects' rights' Support Pool of Experts Programme EDPB (2024) https://www.edpb.europa.eu/system/files/2025-01/d2-ai-effective-implementation-of-data-subjects-rights_en.pdf

[265] EDPS, 'TechSonar 2025 Report' (2025) https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_en

[266] Surve, D., 'Beginner's Guide to LLMs: Build a Content Moderation Filter and Learn Advanced Prompting with Free Groq API' (2024) https://deveshsurve.medium.com/beginners-guide-to-llms-build-a-content-moderation-filter-and-learn-advanced-prompting-with-free-87f3bad7c0af

[267] EDPB Report of the work undertaken by the ChatGPT Taskforce (2024)

[268] Bavarian State Office for Data Protection Supervision, 'Data protection compliant Artificial intelligence Checklist with test criteria according to GDPR'

[269] Deployers should verify that the provider has effectively addressed this risk. This recommendation is equally relevant in cases where deployers are involved in fine-tuning or retraining models.

[270] CNIL, 'Artificial Intelligence (AI)' (2025) https://www.cnil.fr/en/topics/artificial-intelligence-ai

	▪ Article 6(4) GDPR provides, for certain legal bases, criteria that a controller shall take into account to ascertain whether processing for another purpose is compatible with the purpose for which personal data are initially collected.[271] ▪ When outsourcing AI training, verify legal guarantees (e.g., contracts, third-country transfer measures) and ensure training data is not used by service providers for unauthorized purposes.		
9. Unlawful unlimited storage of personal data	As user, deployer and procurement entity make agreements with the third-party provider about how long the input data and output data should be stored. This can be part of the service contract, product documentation or data processing agreement. If data are being stored on your premises, establish retention rules and /or a mechanism for the deletion of data.	✓	✓
10. Unlawful transfer of personal data	▪ As user, deployer and procurement entity, verify with the provider where the data processing is taking place. ▪ Make the necessary safeguard diligences and when necessary, perform a Data Transfer Impact Assessment. ▪ Make the necessary contractual agreements. ▪ Consider this risk when making a selection among different vendors.	✓	✓
11. Breach of the data minimization[272] principle	▪ Regularly review and eliminate unnecessary data collection, automating data deletion when no longer needed. ▪ Replace identifiable data with anonymized or pseudonymized alternatives immediately after collection. ▪ Apply Privacy by Design principles at every development stage, integrating data minimization measures. ▪ Exclude data collection from websites that object to web scraping (e.g., using robots.txt or ai.txt files). ▪ Limit collection to freely accessible data manifestly made public by the data subjects. ▪ Prevent combining data based on individual identifiers unless explicitly required and justified for AI system development.[273] ▪ Educate users about providing only essential data in inputs and transparently communicate data use practices. ▪ Evaluate whether processing personal data is strictly necessary for the intended purpose by exploring less intrusive alternatives, such as the use of synthetic or anonymized data, and ensuring the volume of personal data processed is proportionate to the objective.	✓	?[274]

[271] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted on 17 December 2024.

[272] Processing personal data to address potential biases and errors is permissible only when it is explicitly aligned with the stated purpose, and the use of such data is necessary because the objective cannot be effectively achieved using synthetic or anonymized data. Article 10(5) AI Act provides for specific rules for the processing of special categories of personal data in relation to the high-risk AI systems for the purpose of ensuring bias detection and correction.

[273] CNIL, 'The legal basis of legitimate interests: Focus sheet' (2024) https://www.cnil.fr/en/legal-basis-legitimate-interests-focus-sheet-measures-implement-case-data-collection-web-scraping

[274] Deployers should verify that the provider has effectively addressed this risk. This recommendation is equally relevant in cases where deployers are involved in fine-tuning or retraining models.

7. Residual Risk Evaluation

Identify, Analyze and Evaluate Residual Risk

While inherent risk assessment identifies and evaluates risks before any controls are applied, residual risk assessment focuses on the remaining risk after mitigation measures have been implemented. Evaluating residual risks helps organizations assess the effectiveness of implemented safeguards and understand the potential impact of unwanted events.

After completing the feasibility assessment and implementing mitigation measures, it is essential to reassess if there are any remaining risks. Residual risks[275] are the risks that persist after mitigation measures have been applied.

To analyze residual risk, the probability and severity of the remaining risks are reevaluated, providing a clear overview of the risks that remain after mitigation and taking into account[276]:

- Findings from prior evaluations conducted before the deployment phase
- The effectiveness of applied mitigation measures
- Potential risks that may arise post-deployment obtained through monitoring
- New risks identified during threat modeling sessions

Once residual risks are identified, organizations must decide whether these risks fall within acceptable levels as defined by their risk tolerance[277] and acceptance criteria. If residual risks are deemed acceptable, they can be formally acknowledged and documented in the risk register. However, if the risks exceed acceptable levels, further mitigation measures must be explored and implemented as well as documented. The process then returns to the risk treatment phase to identify the most appropriate treatment option for the risk.

Residual risk evaluation also plays a role in the decision to release a system into production. It is therefore important to assess whether risks remain within defined safety thresholds. Organizations may decide then to request further testing or additional evaluations, mandate further mitigations, or approve the model for deployment if the residual risk is acceptable.

[275] NIST, 'Definition of Residual Risk' (2025) https://csrc.nist.gov/glossary/term/residual_risk
[276] See footnote 193
[277] ISO 31000:2018 Risk Management

8. Review & Monitor

Risk Management Process Review

Reviewing the risk management process is essential to ensure that planned activities have been properly executed and that risk controls and mitigations are effective. This can be done through a structured review of a risk management plan,[278] which serves as a roadmap for identifying, assessing, and mitigating risks throughout a project. While not mandatory, such plans and their reviews are a best practice in risk management, especially in standards for products affected by safety regulations, such as medical devices[279].

A risk management review helps determine whether:

- ✓ Planned risk controls were effectively implemented
- ✓ Emerging risks have been identified and addressed
- ✓ The plan remains aligned with project goals and regulations

Regular reviews also help refine risk strategies, improve processes, and adapt to changes in legislation, business operations, or team structures.

Document Risk Register

Documenting risk assessments, mitigation measures, and residual risks throughout the lifecycle is essential for ensuring accountability, compliance, and continuous improvement. A key tool for this process is the risk register,[280] a structured document that acts as a central repository for all identified risks, including details such as risk nature, ownership, evaluation results, thresholds and mitigation measures. This documentation supports regulatory compliance with frameworks like the GDPR and AI Act, facilitates audits, and enables informed decision-making. A well-maintained risk register can help teams visualize and prioritize risks effectively. Tracking risk ownership and mitigation progress helps ensure that no critical risks are overlooked, and that accountability is maintained throughout the risk management process.

Continuous Monitoring

Once risk mitigation measures[281] have been implemented, ongoing monitoring is essential to assess their effectiveness and identify any emerging risks. After deployment, post-market monitoring[282] plays a critical role in identifying new risks or changes in the operational environment that may impact privacy. This involves the systematic collection and analysis of logs and other operational data in compliance with GDPR requirements, ensuring transparency, accountability, and the ongoing protection of user data.

Currently, LLMs monitoring throughout the lifecycle relies primarily on the following techniques: [283] model testing and evaluation, red teaming, field testing, and long-term impact

[278] FITT Team, 'How Oftern Should You Review Your Risk Management Plan' (2023) https://www.tradeready.ca/explainer/how-often-should-you-review-your-risk-management-plan/

[279] Vn Vroonhoven, J., 'Risk Management Plans and the new ISO 14971' BSI, (2020) https://compliancenavigator.bsigroup.com/en/medicaldeviceblog/risk-management-plans-and-the-new-iso-14971/

[280] Wikipedia, 'Risk Register' (2025) https://en.wikipedia.org/wiki/Risk_register

[281] 'risk management measures', Art.9 AI Act.

[282] Chapter IX, Section 1 Post-market Monitoring, AI Act

[283] AI Action Summit, 'International AI Safety Report on the Safety of Advanced AI' , p - 184, (2025) https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

assessment. These methods help identify and evaluate emerging risks that may not have been apparent during initial development.

- Model testing and iterative evaluations are used before and after the model is deployed and is part of an LLM system. While essential, they are insufficient on their own due to the unpredictability of real-world scenarios and the subjectivity of certain risks. Since LLMs can be applied in numerous contexts, it is difficult to predict how risks will manifest in practice, and as mentioned in section 2, performance metrics and benchmarks may not always accurately reflect those real risks.
- Methodologies such as red teaming[284] can be used to stress-test the model before deployment and the LLM system before and after it is in production by simulating adversarial attacks or misuse scenarios[285], helping to uncover vulnerabilities that might not have been identified during the development phase.
- Field testing evaluates AI risks in real-world conditions, but its implementation remains challenging due to the difficulty of accurately replicating real-world scenarios and establishing clear success metrics. It is important to create a representative test environment and define measurable performance benchmarks to obtain reliable insights.
- Long-term impact assessments evaluate how AI systems evolve over time, aiming to identify unintended consequences that may emerge with prolonged deployment. Continuous monitoring and periodic reassessments are essential to detect shifts in model behavior, performance degradation, or emerging risks that may not have been apparent during initial testing. This technique is part of a continuous monitoring strategy and can also be part of threat modeling sessions.

Across all these techniques, defining robust and reliable monitoring metrics is essential. However, current automated assessments and quantitative metrics often lack reliability and validity,[286]making it difficult to assess risks effectively. For this reason, qualitative human review also plays a crucial role in capturing the broader sociotechnical implications of LLMs and their associated risks.

Incident Response Mechanism

To ensure an effective risk management strategy, it is also important to implement incident response mechanisms that enable a timely and appropriate response to alerts and warnings generated through evaluations and monitoring, as these may indicate a potential privacy or data protection incident.

The warnings coming from the various monitoring techniques are crucial not only for post-market monitoring but also throughout the entire AI lifecycle. The techniques, the scope of testing and the results, will vary depending on whether evaluations are conducted before or after model training, and before or after model and system deployment.

While results are important to help identify new risks, they can also play a key role in assessing the probability of identified threats or hazards occurring. This provides a quantitative analysis that can be compared against established acceptable thresholds to help determine whether further risk mitigations are necessary.

[284] Open AI, 'Advancing red teaming with people and AI' (2024) https://openai.com/index/advancing-red-teaming-with-people-and-ai/

[285] Google Threat Intelligence Group, 'Adversarial Misuse of Generative AI', (2025) https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai

[286] Koh Ly Wey, T., 'Current LLM evaluations do not sufficiently measure all we need' (2025) https://aisingapore.org/ai-governance/current-llm-evaluations-do-not-sufficiently-measure-all-we-need/

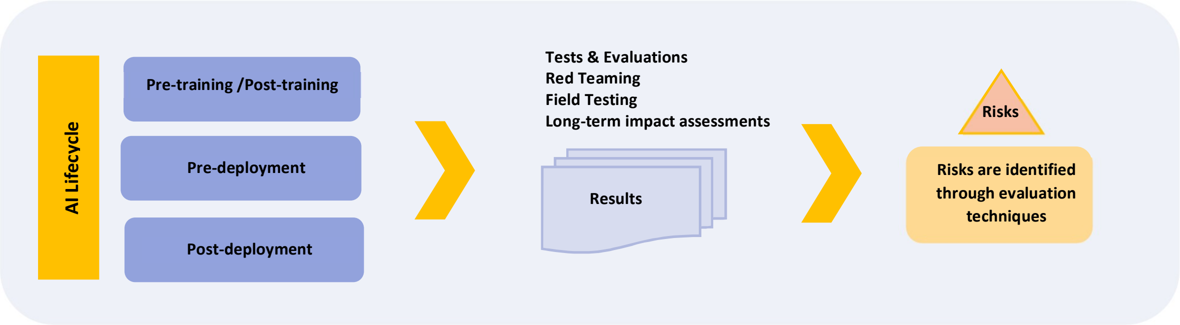

AI Lifecycle
Pre-training /Post-training
Pre-deployment
Post-deployment
Tests & Evaluations
Red Teaming
Field Testing
Long-term impact assessments
Results
Risks
Risks are identified through evaluation techniques

Figure 16. Risks are identified through evaluation techniques

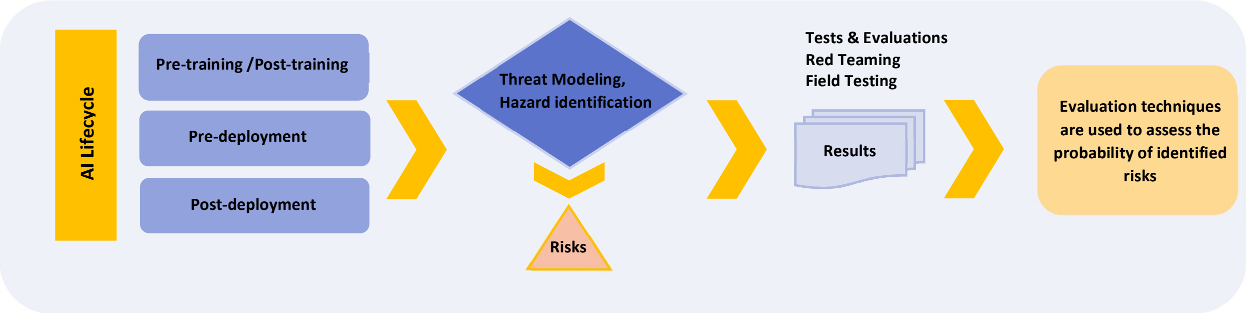

AI Lifecycle
Pre-training /Post-training
Pre-deployment
Post-deployment
Threat Modeling, Hazard identification
Risks
Tests & Evaluations
Red Teaming
Field Testing
Results
Evaluation techniques are used to assess the probability of identified risks

Figure 17. Evaluation techniques are used to assess the probability of identified risks

Iterative Risk Management

Effective risk management for LLMs must adopt an iterative approach that spans the entire lifecycle of the system—from design and development through deployment, monitoring, and eventual decommissioning. As risks can evolve with changes in fine-tuning, system updates, or new contexts of application, regular evaluation and adjustment are essential to address emerging threats.

Human oversight, in combination with automated measures, are critical for managing risks in LLM systems effectively. While automated tools, such as monitoring frameworks, risk registers, and logging systems, provide continuous tracking and analysis, human oversight ensures accountability, fairness, and transparency. This hybrid approach is essential in LLM contexts where entirely manual oversight would be impractical due to the system's complexity and scale.

To further strengthen risk management, tools like LLMOps[287] (LLM Operations) and LLMSecOps[288] (LLM Security Operations) can automate and integrate many aspects of risk management, ensuring seamless updates, monitoring, and response to vulnerabilities. These tools enhance risk tracking and mitigation workflows, reducing the manual documentation burden and improving overall security and governance[289] of LLM systems.

[287] Databricks, 'LLMOps' (2025) https://www.databricks.com/glossary/llmops

[288] Ghosh, B., 'LLMSecOps Elevating Security Beyond MLSecOps' (2023) https://medium.com/@bijit211987/llmsecops-elevating-security-beyond-mlsecops-94396768ecc6

[289] All Tech is Human x IBM Research, 'AI Governance Workshop' (2025) https://static1.squarespace.com/static/60355084905d134a93c099a8/t/677c492a161e58148fc60706/1736198443181/IBM+Research+x+ATIH+AI+Governance+Workshop.pdf

Risk Management Process Overview

The image below provides a visual summary of the risk management process that has been outlined in this report. While each phase has been explained in detail, this overview helps highlight key steps and their interactions, offering a clearer understanding of the entire workflow.

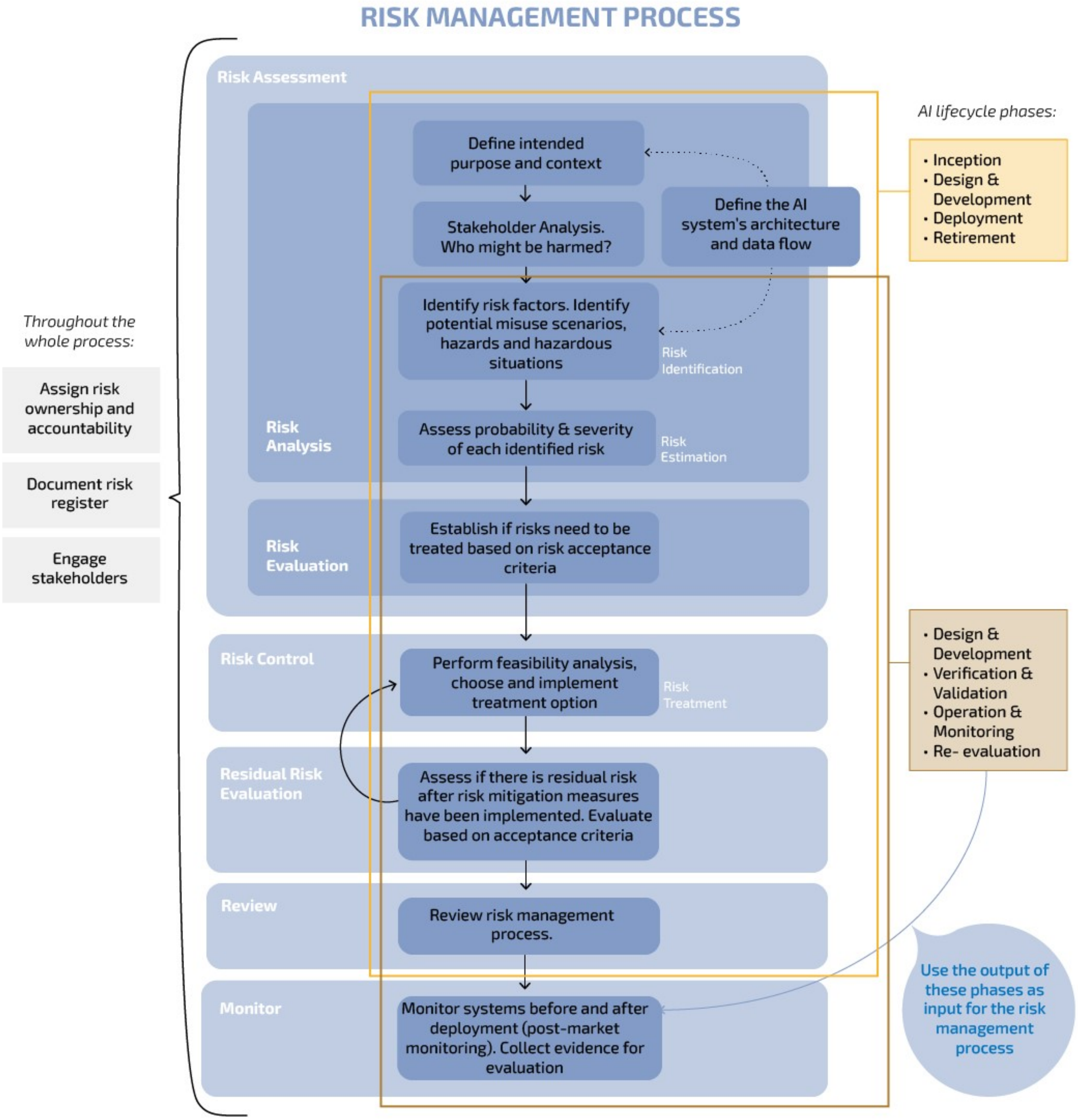
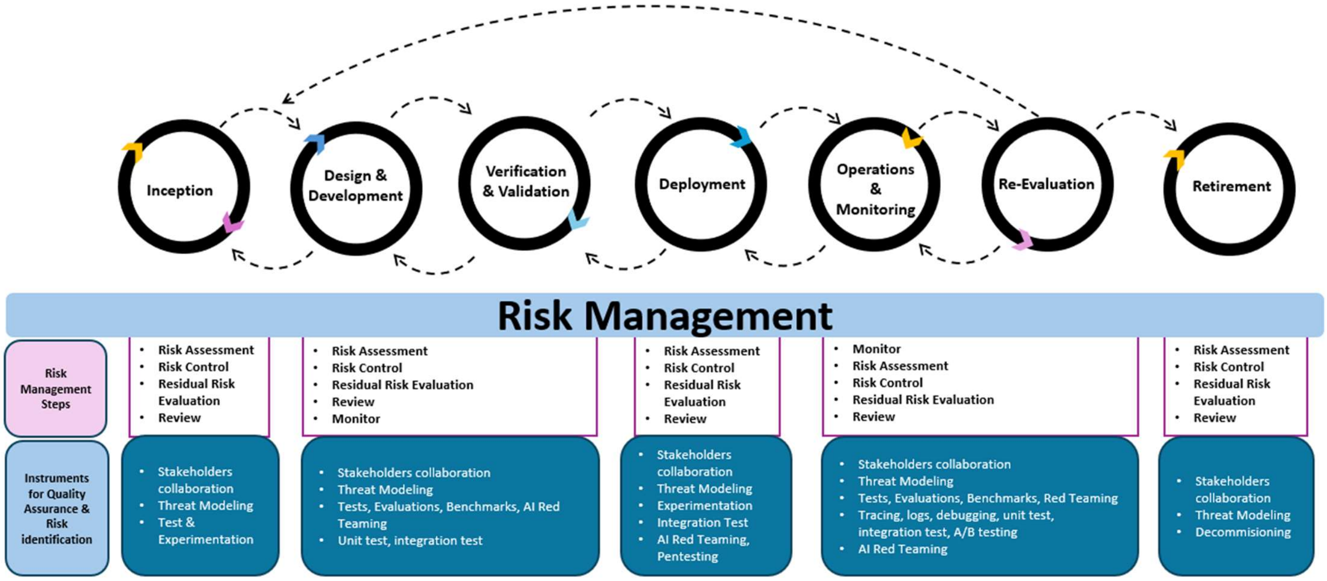

RISK MANAGEMENT PROCESS
Throughout the whole process:
Assign risk ownership and accountability
Document risk register
Engage stakeholders
Risk Assessment
Define intended purpose and context
Stakeholder Analysis. Who might be harmed?
Define the AI system's architecture and data flow
Identify risk factors. Identify potential misuse scenarios, hazards and hazardous situations
Risk Identification
Assess probability & severity of each identified risk
Risk Estimation
Risk Analysis
Risk Evaluation
Establish if risks need to be treated based on risk acceptance criteria
Risk Control
Perform feasibility analysis, choose and implement treatment option
Risk Treatment
Residual Risk Evaluation
Assess if there is residual risk after risk mitigation measures have been implemented. Evaluate based on acceptance criteria
Review
Review risk management process.
Monitor
Monitor systems before and after deployment (post-market monitoring). Collect evidence for evaluation
AI lifecycle phases:
• Inception
• Design & Development
• Deployment
• Retirement
• Design & Development
• Verification & Validation
• Operation & Monitoring
• Re- evaluation
Use the output of these phases as input for the risk management process

Figure 18. Risk Management Process

Risk Management Lifecycle Process

The image below illustrates a risk management framework applied across the lifecycle of an LLM-based system. For each phase, the diagram outlines corresponding risk management steps from the risk management process. These steps help ensure that risks are identified and mitigated continuously as the system evolves.

In addition, the figure highlights instruments for quality assurance and risk identification specific to each phase. These include practices such as stakeholder collaboration, threat modeling, testing and AI red teaming. This layered approach supports a proactive and iterative risk management strategy.



- Risk Assessment
- Risk Control
- Residual Risk Evaluation
- Review

- Stakeholders collaboration
- Threat Modeling
- Test & Experimentation

- Risk Assessment
- Risk Control
- Residual Risk Evaluation
- Review
- Monitor

- Stakeholders collaboration
- Threat Modeling
- Tests, Evaluations, Benchmarks, AI Red Teaming
- Unit test, integration test

- Risk Assessment
- Risk Control
- Residual Risk Evaluation
- Review

- Stakeholders collaboration
- Threat Modeling
- Experimentation
- Integration Test
- AI Red Teaming, Pentesting

- Monitor
- Risk Assessment
- Risk Control
- Residual Risk Evaluation
- Review

- Stakeholders collaboration
- Threat Modeling
- Tests, Evaluations, Benchmarks, Red Teaming
- Tracing, logs, debugging, unit test, integration test, A/B testing
- AI Red Teaming

- Risk Assessment
- Risk Control
- Residual Risk Evaluation
- Review

- Stakeholders collaboration
- Threat Modeling
- Decommisioning

Figure 19. Risk Management Lifecycle

9. Examples of LLM Systems' Risk Assessments

First Use Case: A Virtual Assistant (Chatbot) for Customer Queries

Figure 20. Source: Designed by pch.vector/Freepik

Scenario: A company specialized in kitchen equipment wants to deploy a chatbot to provide general information about its products and services to its customers. The chatbot will have access to pre-existing customer data through integration with the customer management system (e.g., CRM databases). This will allow the chatbot to recognize users based on identifiers like email or account credentials and provide personalized responses without requiring users to re-enter their data. This chatbot interface will be built using as foundation an 'off-the-shelf' LLM that will use RAG to acquire the domain specific knowledge required.

Lifecycle phase we are now: Design & Development

Risk Management Process

This section provides an overview of the steps to implement privacy risk management for LLM based solutions. We begin by examining potential data flows and system architecture for our use case, followed by the steps to identify and address privacy risks effectively.

Data Flow Overview – What are we working on?

The expected data flow for the processing is outlined as follows:

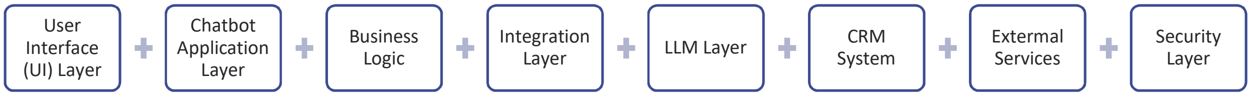
1. User input → Users will interact with the chatbot directly after logging into the company by providing their name, email address, and preferences through an interface (e.g., a website or mobile app).
2. Data preprocessing and API interaction → User input will be validated and formatted before being sent to the chatbot's API for processing. The chatbot will interact with a fine-tuned off-the-shelf LLM hosted on the cloud.
3. Retrieval-Augmented Generation (RAG) →
For queries requiring domain-specific knowledge or up-to-date context, the system performs a retrieval step: it searches the company's CRM, document database, or knowledge base for relevant information. The retrieved content is then combined with the user input and passed to the LLM to generate a grounded, personalized response.
4. Pre-Fine-Tuned LLM processing →
The chatbot uses a fine-tuned LLM trained on enterprise-specific data to enhance general language understanding and tone alignment. This LLM uses the enriched input (from user + RAG) to personalize outputs**.**
5. Data storage → Preprocessed user input (e.g., preferences) will be stored locally or in the cloud to enable personalized recommendations and facilitate future interactions.

6. Personalized response generation → The chatbot will use stored user data from the CRM system and the fine-tuned LLM capabilities to generate tailored recommendations and responses.
7. Data sharing → The chatbot may share minimal, (anonymized) user data with external services (e.g., third-party APIs for additional functionality or promotional tools).
8. Feedback collection → Users provide feedback on chatbot interactions (e.g., thumbs-up/down, comments) to improve the system's performance. This is process by the system for analytics purposes.
9. Deletion and user rights management → Users can request access to, deletion of, or updates to their personal data in compliance with GDPR or similar regulations.

To facilitate the risk assessment process, it is also possible to create a data flow diagram[290], providing a graphical representation of the processes, data movements, and interactions within the system.

Possible Architecture

Considering that we are at the design phase of the AI lifecycle, we anticipate that the architecture of our LLM-based system will include the following key layers:[291]



- **User Interface (UI) Layer:** The interface where users interact with the chatbot through text or voice input. (e.g.; Webpage, mobile app)
- **Chatbot Application Layer:** Manage the flow of conversation and determines chatbot responses based on user input and context. Directs queries to the Business Logic Layer.
- **Business Logic Layer:** Orchestrates chatbot workflows, such as checking customer profiles or placing orders. Crucially, it decides whether to call the LLM directly or trigger a retrieval step (RAG) — for example, by querying the CRM or knowledge base when additional context is needed before generating a response.
- **Integration Layer:** Contains the API Gateway to manage the transmission between layers. Connects the chatbot to the LLM, the CRM system and external services and facilitates secure communication and data exchange between systems. It also handles data transformation, ensures compatibility between the chatbot and the CRM, and implements authentication and authorization for secure access to CRM data. For the RAG setup, this layer may also route queries to a retrieval component or knowledge base before passing enriched inputs to the LLM.
- **LLM Layer:** Performs natural language understanding and generation. Receives either raw user input or input enriched with retrieved content (from the RAG step). Returns contextually relevant responses to the Business Logic Layer.
- **CRM System:** Stores customer data, such as contact information, purchase history, preferences, and support tickets. It also contains CRM APIs that provide endpoints to retrieve, update, or add customer data and event handlers that trigger actions based on events, such as creating a support ticket when a customer raises an issue through the chatbot. Supplies customer data to personalize chatbot responses and stores data generated during interactions.
- **External Services Layer:** It integrates with analytics tools to track user interactions and generate insights into customer behavior. It also integrates with other services, such as payment gateways, email services, or marketing tools.
- **Security Layer:** It encrypts data during transmission using protocols like HTTPS and SSL/TLS, restrict unauthorized access to the chatbot, the LLM and CRM using techniques like OAuth2, implements security and privacy controls, vulnerability scans, threat monitoring, etc.

[290] Wikipedia, 'Data-flow diagram' (2025) https://en.wikipedia.org/wiki/Data-flow_diagram
[291] This architecture is provided as a simplified example and may vary significantly depending on the specific requirements, use case, and technical constraints of each deployment.

Having an overview of the possible architecture at this stage provides a clearer understanding of the data flows and potential risks associated with deploying the chatbot. This architectural insight sets the groundwork for identifying privacy and security concerns early in the process.

Risk Analysis – Stakeholder Collaboration

The next step involves gathering a diverse group of stakeholders to collaboratively identify potential risks. Inviting the right stakeholders is not an exact science, but it is critical to include individuals who will have decision-making authority, direct involvement in its development, deployment and use, and could add value to the risk identification process. Key participants could include representatives from engineering, security, privacy, and UX design teams. If possible, it is highly beneficial to involve individuals with expertise in ethics and fundamental rights, as well as members from civil society groups, deployers and end-users' representatives (customers in our use case). Collecting input from a broader audience through a client survey can also provide valuable insights into user expectations and concerns.

Stakeholder Analysis
Before starting with the risk identification process, the group should analyze the use case to determine which stakeholders target group will interact with the chatbot and identify those who should not have access. Designing barriers where necessary, such as an age verification mechanism, ensures the system aligns with the intended user base. In this specific use case, the entry point for the interface is restricted to logged-in and recognized customers, making additional barriers possibly unnecessary. However, a comprehensive evaluation of all potential risks remains crucial to the system's success.

Stakeholder analysis[292] is a process used to identify and understand the roles, interests, and influence of various stakeholders involved in or affected by a project. Beyond analyzing those directly engaged with the system, it is equally important to assess which stakeholders could be negatively impacted by the tool. This includes recognizing if vulnerable groups might be involved or if the tool's impact could extend to a large number of individuals. Where relevant, it may be valuable to engage affected communities in subsequent phases of risk identification to better capture context-specific concerns and impacts. Participatory engagement tools[293] like the ethical matrix, mentioned in a previous section, can help evaluate the potential consequences for different stakeholder groups.

In our use case, we have identified our *customers* as the only authorized users. Given the nature of our business, we do not anticipate children accessing our platform. However, we remain mindful of implementing appropriate security measures to ensure that access is restricted, and unauthorized use is prevented.

Risk Identification – Selection of Risk Factors

Considering our data flow diagram, system architecture, and deliberations with our selected stakeholders, we have identified the following risk factors and key concerns from Section 4 as applicable to our specific use case:

[292] Rodgers, A.,' What is a Stakeholder Impact Analysis?', Simply Stakeholders (2024) https://simplystakeholders.com/stakeholder-impact-analysis/

[293] Park, T., Stakeholder Engagement for Responsible AI: Introducing PAI's Guidelines for Participatory and Inclusive AI', Partnerships on AI' (2024) https://partnershiponai.org/stakeholder-engagement-for-responsible-ai-introducing-pais-guidelines-for-participatory-and-inclusive-ai/

Risk factor	**Use case applicability**
Large scale processing	A significant volume of data will be processed due to our extensive customer database and the large amount of information stored in our CRM system.
Low data quality	Customer query inputs may have low quality, and the CRM database has not been validated, which could lead to inaccuracies or inefficiencies in processing.
Insufficient security measures	There is a potential risk of transferring personal data to countries without an adequate level of protection, especially if the LLM model is hosted or maintained in such regions.

We have identified several risk factors that require attention, as they indicate a higher probability of undesirable outcomes. While our system does not fall under the classification of a high-risk system under the AI Act, there is, from the GDPR perspective, sufficient evidence to justify initiating[294] the process for creating a Data Protection Impact Assessment (DPIA). This risk assessment we are performing now will serve as a valuable foundation for the DPIA process.
It is important to emphasize when a DPIA is necessary and when a Fundamental Rights Impact Assessment (FRIA) is required. A DPIA, under Article 35 of the GDPR, is required whenever a data processing is likely to result in a high risk to the rights and freedoms of natural persons.[295]

Common scenarios that require a DPIA include:

- The use of new technologies that could introduce privacy risks.
- The tracking of individuals' behavior or location (e.g., geolocation services or behavioral advertising).
- Large-scale monitoring of publicly accessible spaces (e.g., video surveillance).
- Processing sensitive data categories such as racial or ethnic origin, political opinions, religious beliefs, genetic data, biometric data, or health information.
- Automated decision-making that has legal or similarly significant effects on individuals.
- Processing children's data or any data where a breach could lead to physical harm.

Even when a DPIA is not explicitly required by law, conducting one can be prudent for best practices in privacy and security. It allows organizations to preemptively address potential risks, assess the impact of their solutions, and demonstrate accountability. In contrast, a FRIA, as outlined in Article 27[296]of the AI Act, can be mandatory for some deployers of high-risk AI systems (bodies governed by public law, private entities providing public services, organisations doing creditworthiness evaluations, pricing and life and health insurance risk assessments. A FRIA evaluates the potential impact of such systems on fundamental rights like privacy, fairness, and non-discrimination. Deployers of high-risk AI systems must document:

- How the system will be used, including its purpose, duration, and frequency.
- The categories of individuals or groups affected by the system.
- Specific risks of harm to fundamental rights.
- Measures for human oversight and governance.
- Steps to address and mitigate risks if they materialize.
- Where applicable, a FRIA can complement a DPIA, by focusing on broader societal impacts beyond data protection alone.

[294] EDPB, 'Data Protection Guide for Small Business' (2025) https://www.edpb.europa.eu/sme-data-protection-guide/be-compliant_en

[295] WP 29 Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, WP248 rev.01, (2017) endorsed by the EDPB https://ec.europa.eu/newsroom/article29/items/611236

[296] Article 27(1) AI Act: "Prior to deploying a high-risk AI system referred to in Article 6(2), with the exception of high-risk AI systems intended to be used in the area listed in point 2 of Annex III, deployers that are bodies governed by public law, or are private entities providing public services, and deployers of high-risk AI systems referred to in points 5 (b) and (c) of Annex III, shall perform an assessment of the impact on fundamental rights that the use of such system may produce…."

Risk Identification – What can go wrong?

For the process of risk identification, we will follow an approach similar to the one used in threat modeling methodologies. Privacy threat modeling serves as a structured way to identify potential risks and vulnerabilities within a system, focusing on scenarios that could impact data subjects' rights and freedoms. While threat modeling is not mandatory and does not replace[297] the DPIA or broader risk management processes, it is a versatile and effective tool that enhances these efforts by generating valuable insights, such as risk scenarios and potential impacts.

Collaborating closely with our stakeholder group, and after having identified the risk factors that affect our use case, we will now review the data flow diagram again and facilitate the process of risk identification by using the two lists of risk from sections 3 and 4.

Instead of relying on external risk libraries[298] contained in threat modeling methodologies such as LINDDUN[299], LIINE4DU[300] or PLOT4AI[301], for our use case we will use the tailored lists of risks outlined in this document. It is essential to remember that using a predefined library of risks is merely a starting point. Organizations should think beyond these lists, considering unique aspects of their systems and context and asking themselves, "What else can go wrong?".

In threat modeling, usually four foundational questions[302] guide the process:

1. What are we working on?
2. What can go wrong?
3. What are we going to do about it?
4. Did we do a good job?

For this use case, we will integrate these questions into the risk management process as follows:

What are we working on?

We are implementing an LLM based chatbot. To understand its design and architecture, we are leveraging for the risk identification session with our group of stakeholders, the most recent version of a data flow diagram. We also need to consider that we are the design and development phase of the AI lifecycle which means that many tests and evaluations have not yet been conducted, and we lack user feedback and insights from the production environment.

What can go wrong?

After examining our data flow diagram, the context of use, the intended purpose of the application, the characteristics of our user group and design, as well as any results obtained from evaluations and tests, we have identified the following risks outlined in Section 4:

1. Insufficient protection of personal data leading to a data breach.
2. Misclassification of training data as anonymous.
3. Possible adverse impact on data subjects that could negatively impact fundamental rights.
4. Not granting data subject rights.
5. Unlawful repurposing of personal data.
6. Unlawful unlimited storage of personal data.
7. Unlawful transfer of personal data.
8. Breach of the data minimization principle.

[297] AEPD, 'Technical Note: An Introduction to LIINE4DU 1.0: A New Privacy & Data Protection Threat Modelling Framework', 2024

[298] Slattery P., et al., 'The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence' (2024) https://arxiv.org/abs/2408.12622

[299] LINDDUN, 'Privacy Threat Modelling' (2025) https://linddun.org/

[300] AEPD, 'Technical Note: An Introduction to LIINE4DU 1.0: A New Privacy & Data Protection Threat Modeling Framework' (2024) https://www.aepd.es/guides/technical-note-introduction-to-liine4du-1-0.pdf

[301] PLOT4AI, 'Practical Library of Threats 4 Artificial Intelligence' (2025) https://plot4.ai/

[302] Shostack, A., 'The Four Question Framework for Threat Modeling' (2024) https://shostack.org/files/papers/The_Four_Question_Framework.pdf

From the potential privacy risks outlined in Section 3 for systems based on an LLM 'off-the-shelf-model,' we have reviewed all risks across the standard data flow phases and identified that most of these risks are covered under Risk 1 (Insufficient protection of personal data leading to a data breach) and Risk 3 (Possible adverse impact on data subjects that could negatively impact fundamental rights) from Section 4.

Phases	Possible Risks
User Input	Sensitive data disclosure, unauthorized access, lack of transparency, adversarial attacks
Provider Interface & API	Data interception, API misuse, interface vulnerabilities
LLM Processing at Providers' Infrastructure	Model inference risks, unintended data logging, anonymization failures, unauthorized access to logs, data aggregation risks, third-party exposure, inadequate data retention policies
Processed Output	Inaccurate or sensitive responses, re-identification risks, output misuse

Risk Estimation and Evaluation

We will now analyze each identified risk by assessing its probability and severity to determine which risks require treatment. Whenever possible, we will also consider system test results and model evaluations, if available, to inform our assessment. In some cases, conducting additional evaluations may be necessary to obtain quantitative data that can improve our risk analysis.

For example, in the case of Risk 2 (Misclassification of training data as anonymous), we can already perform tests to detect the presence of personal data in our datasets. These results would help us assess the probability of the risk occurring given the current dataset conditions.

At this stage of the AI lifecycle (pre-deployment phase), the available evaluations are limited. However, when risk assessments take place post-development, additional evaluations can be conducted, providing further quantitative criteria to refine risk assessment and decision-making.

Probability

We are going to assess the probability of identified risks, categorizing them into one of the four levels in the probability matrix: Very High, High, Low, or Unlikely. This categorization should be done by directly assigning a level to each risk based on quantitative and/or qualitative criteria and through collaborative decision-making with stakeholders. Alternatively, we can also employ a list of predefined criteria to guide our assessment.
For a more quantitative approach calculating probability, aggregation methods can be applied to calculate its level. In this use case, we will use the FRASP framework to structure and refine our probability assessment process.

Calculating the *Total Probability Score*:

We evaluate fictitiously each identified risk and assign it a score per criteria. We add the scores of all factors and divide the total by the number of factors to calculate the *Aggregate Probability Score*. In our case we will treat all factors and calculate the mean.

Once the aggregate score is calculated, we will map it to one of the predefined probability levels based on the following ranges:

1.0 - 1.5: Unlikely
1.6 - 2.5: Low
2.6 - 3.5: High
3.6 - 4.0: Very High

Our Total Probability Score (TPS) per risk is:

	Probability Criteria							Aggregate Score		TPS
Risk	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Calculation**	**Result**	**Result**
1	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Low
2	3	1	4	1	2	3	2	3+1+4+1+2+3+2 / 7	2,28	Low
3	3	1	2	1	2	3	2	3+1+2+1+2+3+2 / 7	2	Low
4	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Low
5	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Low
6	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Low
7	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Low
8	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Low

Severity

Next, we will assess the potential privacy impact and severity of these risks on data subjects, individuals, and society. Based on this severity assessment, we will assign one of the four levels from the severity classification matrix: Very Significant, Significant, Limited, or Very Limited.
The calculation of severity will follow the same steps as those used for determining probability. However, for severity, the highest level obtained among criteria 1 to 5, as well as 7 and 8, will set the total severity score.

Once the aggregate score is calculated, we will map it to one of the predefined severity levels based on the following ranges:

1.0 - 1.5: Very Limited / Moderate or Minor Harm (Level 1)
1.6 - 2.5: Limited / Serious Harm (Level 2)
2.6 - 3.5: Significant/ Critical Harm (Level 3)
3.6 - 4.0: Very Significant / Catastrophic Harm (Level 4)

In this case the final score is determined by the highest score in criteria 1, 5 and 7 giving as result Level 3 severity for all the risks.

	Severity Criteria											Aggregate Score		TSS
Risk	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Calculation**	**Result**	**Result**
1	3	2	2	2	3	1	2	2	3	2	3	3+2+2+2+3+1+2+2+3+2+3 / 11	2,27	Significant/ Critical Harm
2	3	2	2	2	1	1	3	2	2	2	3	3+2+2+2+1+1+3+2+2+2+3 / 11	2,09	Significant/ Critical Harm
3	3	2	2	2	2	1	2	2	2	2	2	3+2+2+2+2+1+2+2+2+2+2 / 11	2	Significant/ Critical Harm
4	3	2	2	2	2	1	3	2	1	2	2	3+2+2+2+2+1+3+2+1+2+2 / 11	2	Significant/ Critical Harm
5	3	2	2	3	3	1	3	2	1	2	2	3+2+2+3+3+1+3+2+1+2+2 / 11	2,18	Significant/ Critical Harm
6	3	2	2	2	3	1	2	1	1	2	1	3+2+2+2+3+1+2+1+1+2+1 / 11	1,81	Significant/ Critical Harm
7	3	2	2	2	3	1	2	1	1	2	1	3+2+2+2+3+1+2+1+1+2+1 / 11	1,81	Significant/ Critical Harm
8	3	2	2	2	3	1	2	1	1	2	1	3+2+2+2+3+1+2+1+1+2+1 / 11	1,81	Significant/ Critical Harm

By applying the classification matrix to the obtained probability and severity scores, we can determine the corresponding risk classification level. In our case for all the risks the combination Low Probability + Significant Severity offers a result of **High Risk**.

Probability	**Very High**	Medium	High	Very high	Very high
	High	Low	High	Very high	Very high
	Low	Low	Medium	High	Very high
	Unlikely	Low	Low	Medium	Very high
		Very limited	**Limited**	**Significant**	**Very Significant**
		Severity			

Although high level risks always require treatment, it is considered best practice to assess whether classified risks need treatment evaluating predefined acceptance criteria and acceptable metric thresholds established by the organization. These criteria can be adjusted per use case and tailored to pre-deployment and post-deployment phases to ensure a context-aware risk management approach.

In our specific use case, the organization's risk acceptance criteria are as follows:

- A risk that can result in a violation of data protection regulations is not acceptable.
- A risk of unauthorized access, exposure, or retention of persona data beyond what is strictly necessary is not acceptable.
- Re-identification risk must remain below 1%, verified through privacy-preserving evaluations and testing.
- Membership inference and model inversion attack risks must remain below a 1% success rate as verified through internal testing and, for sensitive data, independent external audits.
- Inaccurate datasets are only acceptable if the error rate does not exceed 5% and all available data validation and cleaning processes have been applied.
- The chatbot must clearly inform users when their data is being used and provide access to data usage policies. Transparency risks are not acceptable.
- No risk is acceptable if it prevents users from exercising their data rights, unless explicitly justified under legal exceptions.

Risk Control - What are we going to do about it?

In our use case, several high-level risks have been identified. After evaluating our risk acceptance criteria, conducting a feasibility analysis and reviewing the treatment options, we determined that **transferring** the risks to a third party is not feasible, **avoiding** the risks entirely is impractical, and outright **acceptance** of the risks is unacceptable. However, since there are **mitigation** measures available that can reduce the risks to an acceptable level, we have opted for the treatment strategy of risk mitigation to proceed responsibly with the implementation.

Note that Section 3 and Section 4 already outlined comprehensive mitigation measures for the identified risks. It is also worth noting that many of the specific risks identified in this use case fall under the broader category of Risk 1, which relates to insufficient protection of personal data.

Data Flow Phase	Description	Privacy Risks	Mitigation's recommendations
User input	Users interact with the chatbot by providing their name, email address, and	▪ Input could be intercepted if	▪ Secure data transmission using adequate encryption protocols.

	preferences through an interface (e.g., a website or mobile app).	transmitted over an insecure connection. ▪ Users might provide unnecessary or excessive personal information. ▪ Children or vulnerable users might share personal data.	▪ Apply input constraints to limit collected data to what is essential. ▪ Use age verification or consent mechanisms to protect vulnerable users (children not our user group) ▪ Clearly inform users about how their data is processed and caution them against sharing sensitive or confidential information when using the chatbot.
Data preprocessing and API interaction	User input is validated and formatted before being sent to the chatbot's API for processing. The chatbot interacts with a fine-tuned off-the-shelf LLM hosted on the cloud and connects to the CRM system both to retrieve or update user information and to fetch relevant content that is passed to the LLM as context (RAG).	▪ Unsecured APIs could allow attackers to intercept or manipulate user data. ▪ Malicious inputs (e.g., injections) could exploit system vulnerabilities. ▪ Logs might inadvertently store sensitive user data. ▪ Retrieved content may contain sensitive or outdated information, which could be exposed in generated outputs. ▪ Poorly configured retrieval logic could result in irrelevant or misleading context being fed to the LLM, increasing hallucination risk. ▪ If retrieval accesses external or third-party knowledge bases, user queries may be logged or monitored without consent.	▪ Use robust API security measures, including access controls, authentication, and rate limiting. ▪ Sanitize user input to prevent injection attacks. ▪ Minimize API logging or ensure logs are anonymized and protected by access controls. ▪ Restrict retrieval sources to approved, privacy-screened datasets (e.g., filtered CRM data). ▪ Implement relevance filters or scoring mechanisms to ensure only appropriate content is passed to the LLM. ▪ Apply post-processing/output filters to remove or redact sensitive information from responses. ▪ Use internal retrieval systems when possible; if third-party search APIs are used, anonymize or mask user queries.
Pre-Fine-Tuned LLM processing	The chatbot relies on a fine-tuned LLM and also uses retrieved content from CRM as input context to generate more accurate responses.	▪ Hallucinations: The model might generate inaccurate or misleading responses. ▪ Training data biases may persist in outputs, influencing recommendations. ▪ Retrieved content may be misinterpreted by the LLM, leading to distorted or irrelevant outputs. ▪ Sensitive data from retrieval sources could	▪ Evaluate chatbot responses regularly for accuracy and relevance. ▪ Train the model on high-quality, diverse datasets to reduce biases. ▪ Include disclaimers in chatbot responses to clarify they are AI-generated and not definitive advice. ▪ Apply retrieval filters and output sanitization to reduce risk of leaking sensitive information.

		be included in responses if not properly filtered.	▪ Weight or flag retrieved content based on source reliability to help the model contextualize correctly.
Data storage	Preprocessed user input (e.g., preferences) is stored locally or in the cloud to enable personalized recommendations and facilitate future interactions.	▪ Weak access controls could expose stored user data. Retaining data longer than necessary violates privacy regulations. ▪ Adversaries could infer whether specific user data was used in training.	▪ Encrypt stored data and implement access controls. ▪ Adopt clear data retention policies to delete data when no longer needed. ▪ Apply differential privacy techniques to prevent membership inference attacks.
Personalized response generation	The chatbot uses stored user data from the CRM system and fine-tuned LLM capabilities to generate tailored recommendations and responses.	▪ Outputs might inadvertently pressure users or contain inaccurate information. ▪ Responses might infer unintended personal insights about users.	▪ Implement output validation mechanisms to detect and mitigate harmful or inaccurate responses. ▪ Regularly audit chatbot recommendations for fairness and transparency. ▪ Clearly communicate to users how recommendations are generated.
Data sharing	The chatbot may share minimal, anonymized user data with external services (e.g., third-party APIs for additional functionality or promotional tools).	▪ Data shared with third parties might be used for purposes outside the agreed scope. ▪ Insufficient protection in third-party systems could expose shared data.	▪ Establish robust data-sharing agreements with third parties. ▪ Anonymize and minimize shared data to reduce risks of misuse or exposure. ▪ Regularly audit third-party data protection practices.
Feedback collection	Users provide feedback on chatbot interactions (e.g., thumbs-up/down, comments) to improve the system's performance.	▪ Feedback might include unintended personal details. ▪ Feedback data might introduce biases during future model retraining.	▪ Anonymize feedback data before storing or using it for model improvement. ▪ Communicate feedback usage policies to users and obtain explicit consent for data usage in retraining.
Deletion and user rights management	Users can request access to, deletion of, or updates to their personal data in compliance with GDPR or similar regulations.	▪ Failure to honor user requests could result in regulatory penalties. ▪ Data may persist in logs, backups, or third-party systems even after a deletion request.	▪ Implement robust data rights management tools for access, correction, and deletion requests. ▪ Regularly audit data systems to ensure compliance with deletion requests. ▪ Clearly communicate to users how their data is handled and retained.

Evaluate Residual Risk - Did we do a good job?

After the mitigation measures have been identified and implemented, we will assess again the probability and severity of each risk to obtain a new risk classification level and in this way evaluate if there is any remaining or residual risk.

In our case, the risk level has been reduced to Medium what means it is not yet acceptable.

Why is the risk level reduced to Medium instead of Low after having implemented all the mitigations? The risk remains Medium despite reducing severity to Limited (level 2) because the risk matrix combines probability and severity to determine overall risk.

With the four levels matrix that we use in this example, a Low Probability and Limited Severity result in a Medium Risk level because, while unlikely, the consequences of a risk, though mitigated, are still non-negligible. That means the remaining risk after mitigation measures might still be above an acceptable threshold for your organization.

What can we do to address Residual Risk in this case? Some options that organizations can apply are:

- Reduce Probability by strengthening preventive controls (e.g., access measures, anomaly detection) and enhancing event prevention mechanisms.
- Implement extra mitigations measures to reduce severity.
- Implement robust monitoring and establish a clear incident response plan to minimize impact if the risk materializes.
- Explore additional mitigations: for instance, use advanced technologies (e.g., differential privacy) or fail-safe mechanisms to further mitigate risks.
- Reevaluate whether the residual risk is within organizational risk tolerance and document justification for maintaining it.
- Discuss options to share or transfer the risk (e.g., insurance, vendor agreements).

Review & Monitor

The "Did we do a good job?" question in threat modeling goes beyond merely addressing residual risks—it serves as an evaluation of the entire risk management process. This phase ensures that the identified risks, proposed mitigations, and resulting outcomes align with the system's objectives and regulatory requirements. It also provides an opportunity to validate internal and external processes, assess the real-world applicability of mitigations, and identify any gaps or areas for improvement.

This process is also related to the Monitoring and Review phase, where the Risk Management Plan is reassessed to ensure that risk mitigation efforts remain effective. As part of this phase, it is essential to ensure that the risk register is properly documented and continuously updated.

Since we are currently in the design and development phase, we should proactively plan for continuous monitoring of our chatbot. This includes defining metrics for ongoing risk assessment, establishing adequate data logging practices, and ensuring that an incident response plan is in place to address potential privacy issues that may arise post-deployment.

Second Use Case: LLM System for Monitoring and Supporting Student Progress

Figure 21. Source: Designed by pch.vector/Freepik

Scenario: A school wants to adopt a third party LLM system to monitor and evaluate students' academic performance and provide tailored recommendations for improvement. The tool is an LLM-based system developed with and LLM 'off-the shelf' model. This tool would analyze a combination of data, including test scores, assignment completion rates, attendance records, and teacher feedback, to identify areas where students may need additional support or resources. For example, if a student struggles with math, the tool could recommend targeted practice exercises, suggest online tutoring sessions, or notify parents and teachers about specific challenges. The goal is to create a personalized learning plan that helps each student achieve their full potential.

This system would deal with sensitive information about minors, including their academic records and behavioral patterns, which introduces significant privacy and ethical risks.

Lifecycle phase we are now: Inception

Risk Management Process

Since we have already detailed the complete risk assessment process in the first use case, including identifying, classifying, and mitigating risks, this section and the subsequent third use case will focus specifically on identifying unique privacy and data protection risks and mitigations.

This table shows how the risks identified in Section 3 and 4 (Privacy Risk Library) can be aligned with the risks specific for this use case.

Privacy Risk Library	Privacy Risks Identified and aligned with Library	Recommended Mitigations
1. Insufficient protection of personal data what eventually can be the cause of a data breach	▪ Weak safeguards could lead to data breaches, unauthorized access, or exposure of sensitive student data. ▪ APIs facilitating communication between the tool, school systems, and third parties could be unsecured or improperly configured. ▪ Inadequate access controls may allow unauthorized school personnel or external parties to view sensitive student data. ▪ If the vendor does not comply with data protection regulations, it increases the risk of a data breach.	▪ Implement strong encryption protocols for data in transit and at rest (e.g., SSL/TLS, AES-256). ▪ Regularly conduct security audits and penetration testing. ▪ Establish incident response plans for timely detection and mitigation of breaches. ▪ Use API gateways with robust security configurations, including authentication, access control, and rate limiting. ▪ Implement authentication and ensure secure API endpoints. ▪ Conduct regular API security reviews and validation.

	▪ The tool might interact with third-party services or platforms (e.g., online tutoring systems, analytics services, or cloud-based storage) for functionality, exposing student data to external entities.	▪ Enforce strict role-based access control (RBAC) policies. ▪ Implement multi-factor authentication (MFA) for all users accessing sensitive data. ▪ Regularly review and update user access permissions. ▪ Conduct vendor due diligence, including Data Protection Impact Assessments (DPIAs) and security certifications. ▪ Include specific data protection clauses in contracts with vendors, ensuring accountability for compliance. ▪ Require vendors to provide evidence of GDPR-compliant practices. ▪ Establish robust data-sharing agreements with third-party platforms, ensuring compliance with GDPR requirements. ▪ Limit data shared with third parties to anonymized or pseudonymized datasets. ▪ Monitor third-party systems for adherence to agreed data protection measures.
2. Misclassifying training data as anonymous by controllers when it contains identifiable information	Adversaries might exploit the LLM to infer whether specific student data was used in training, indicating a misclassification of training data.	▪ Use differential privacy techniques to minimize the risk of data inference. ▪ Conduct structured testing against membership inference and attribute inference attacks. ▪ Validate that the LLM provider has implemented safeguards to prevent such attacks.
3. Unlawful processing of personal data in training sets	▪ If personal data (e.g., academic records) is unlawfully processed in training datasets by the LLM provider ▪ Behavioral and academic data require explicit consent or another valid legal basis to be processed lawfully.	▪ Verify that the LLM provider's training datasets exclude sensitive personal data without proper safeguards. ▪ Require documentation from vendors proving that training data was lawfully collected and processed. ▪ Use models trained on synthetic or anonymized data when possible.
4. Unlawful processing of special categories of personal data and data relating to criminal convictions and offences in training data.	If health-related or behavioral data about children, such as indications of mental health conditions, is processed—such as when identifying special assistance needs for conditions like dyslexia, ADHD, or similar.	▪ Ensure explicit consent is obtained from parents or guardians before processing children's data. ▪ Conduct a DPIA and identify lawful grounds for processing. ▪ Provide clear, accessible information to parents about how data is processed. ▪ Implement stricter safeguards for sensitive data, including encryption and access controls. ▪ Limit processing to data strictly necessary for the intended purpose. ▪ Provide parents or guardians with transparency about how health-related data is used.
5. Possible adverse impact on data subjects that could negatively impact fundamental rights	Fairness and Discrimination: ▪ Recommendations based on biased training data could disproportionately impact certain student groups. ▪ Continuous monitoring of behavioral patterns could lead to profiling students in ways that might be discriminatory or stigmatizing. Accuracy: ▪ The tool might generate inaccurate recommendations or reports due to	Fairness and Discrimination ▪ Regularly audit training data to identify and reduce biases. ▪ Involve diverse stakeholders in testing the system for potential biases. Accuracy ▪ Establish processes for regular model evaluation and fine-tuning using high-quality, diverse datasets. ▪ Provide disclaimers with AI-generated recommendations, emphasizing the importance of human oversight.

	biases in training data or processing errors. Transparency: - Teachers, parents, or students may not fully understand how decisions or recommendations are made by the AI tool.	- Enable error reporting mechanisms to continuously improve model accuracy. Transparency - Provide detailed documentation[303] about the AI tool's decision-making processes to parents, teachers, and students. - Implement transparency mechanisms, such as explainable AI (XAI) methods, to make decisions more interpretable. - Offer training sessions for stakeholders to understand the tool's capabilities and limitations.
6. Not providing human intervention for a processing that can have a legal or important effect on the data subject.	Lack of human oversight in automated recommendations or interventions could have significant adverse academic impacts on students.	- Implement mechanisms for obtaining verifiable parental or guardian consent. - Provide parents with easy-to-understand information about the tool's data collection and usage. - Require human review of critical recommendations or interventions before implementation. - Train teachers and administrators to identify when human intervention is needed. - Define processes for escalating issues requiring human judgment.
7. Not granting data subjects their rights	Failure to obtain proper parental or guardian consent violates GDPR's requirements for minors. Students and parents may not fully understand how their data is processed, limiting their ability to exercise their rights.	- Create a user-friendly interface explaining how data is collected, processed, and stored. - Offer accessible resources to help students and parents exercise their GDPR rights, including erasure, rectification, and access.
8. Unlawful repurpose of personal data	It the academic and behavioral data from children is used without a compatible purpose to the original one.	- Ensure data usage aligns with the original purpose of collection and assess any repurposing against GDPR principles. - Document purpose compatibility assessments for accountability.
9. Unlawful unlimited storage of personal data	Data might be retained longer than necessary for its intended purpose, particularly behavioral or health-related data.	- Define clear data retention policies and automatically delete data once is no longer needed. - Regularly audit stored data for compliance with retention limits.
10. Unlawful transfer of personal data	If the tool relies on cloud services or external platforms hosted in jurisdictions without adequate data protection standards.	- Verify that cloud service providers comply with GDPR's data transfer rules, including adequacy decisions or Standard Contractual Clauses (SCCs). - Perform Data Transfer Impact Assessments (DTIAs) when required.
11. Breach of the data minimization principle	Excessive data collection or processing beyond what is necessary infringe the data minimization principle.	- Apply strict data collection filters to gather only the data necessary for the tool's purpose. - Anonymize or pseudonymize data where possible to minimize risk.

It is important to note that the list of risks and mitigations provided is based on generic information and assumptions. In a real-world scenario, a detailed risk assessment tailored to the specific implementation, context, and operational environment of the LLM based tool would be necessary. This includes collaboration with stakeholders, such as the LLM system provider, school administrators, teachers, parents, and students, to identify unique risks and address them effectively.

[303] Models cards and system cards are example of information that can be provided to deployers:
Green, N et al., System Cards, a new resource for understanding how AI systems work (2022) https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/; Hugging Face, 'Model Cards' (2024) https://huggingface.co/docs/hub/en/model-cards

Third Use Case: AI Assistant for Travel and Schedule Management

Figure 22. Source: Designed by pch.vector/Freepik

Scenario: A personal assistant AI agent is designed to help users manage their travel plans and daily agendas. The agent can book flights, reserve hotels, schedule meetings, and send reminders based on user-provided inputs and preferences. For instance, a user might ask the agent to "book a round trip to Madrid next week and find a hotel near the Prado Museum." To fulfill this request, the agent accesses the user's calendar, retrieves personal preferences (e.g., preferred airlines or hotel chains), and interacts with third-party booking platforms. This system is developed with various 'off-the shelf' LLMs and SLMs.

Lifecycle phase we are now: Operations and Monitoring

Risk Management Process

In this use case we have identified the following privacy risks and recommended mitigations that we have aligned with our 11 foundational privacy risks.

Privacy Risk Library	Privacy Risks Identified and aligned with Library	Recommended Mitigations
1. Insufficient protection of personal data what eventually can be the cause of a data breach	▪ Weak safeguards could expose sensitive personal data, such as travel itineraries, calendar entries, and user preferences to unauthorized access or breaches. ▪ Unauthorized access due to poor access control mechanisms. ▪ Inference attacks where adversaries exploit vulnerabilities to infer personal data not explicitly provided.	▪ Encrypt user data during transmission and at rest. ▪ Implement secure APIs with rate limiting, authentication, and monitoring to control access. ▪ Use anonymization and pseudonymization to safeguard sensitive data. ▪ Regularly test for vulnerabilities like membership inference, model inversion, or poisoning attacks. ▪ Use robust inter-agent encryption[304] to protect data exchange in multi-agent systems.
2. Misclassifying training data as anonymous by controllers when it contains identifiable information	Not applicable in this use case as the focus is on operational data rather than training data. (This use case is based on the Lifecycle Phase: Operations and Monitoring)	▪ (Not directly applicable in this case, as the system uses pre-trained models, but applicable to providers.) ▪ Ensure robust testing and validation to confirm training data anonymity claims.

[304] Chen, G et al., 'Encryption–decryption-based consensus control for multi-agent systems: Handling actuator faults.', Automatica, Volume 134, 109908, ISSN 0005-1098 (2021) https://doi.org/10.1016/j.automatica.2021.109908

		▪ Use threat models to evaluate risks of re-identification techniques (e.g., attribute inference). ▪ Providers must document data anonymization methods and compliance with Article 5(2) GDPR.
3. Unlawful processing of personal data in training sets	Not applicable in this use case, as the system is already in operation and relies on pre-trained LLMs and SLMs.	▪ (Not directly applicable, as no training occurs in the operational phase.)
4. Unlawful processing of special categories of personal data and data relating to criminal convictions and offences in training data.	Behavioral and personal preferences data (e.g., specific health conditions inferred from travel patterns) may fall into special categories, requiring explicit consent or valid legal basis. Interactions with calendars or other sensitive tools may inadvertently process health-related or sensitive data.	▪ Implement explicit consent mechanisms for processing sensitive data like health-related information inferred from user interactions (e.g., calendar or travel preferences). ▪ Validate that sensitive data collected (if any) is necessary for the intended purpose. ▪ Use privacy-preserving techniques for sensitive data handling.
5. Possible adverse impact on data subjects that could negatively impact fundamental rights	▪ Manipulation or overreliance on suggestions, where the agent prioritizes third-party interests over user preferences. ▪ Profiling and unfair treatment, such as price discrimination[305] or biased recommendations. ▪ The agent might rely on outdated or inaccurate information from external sources, leading to errors in bookings or scheduling, which could inconvenience users. ▪ Users may not fully understand how the system operates, including how decisions are made or how their data is processed and shared.	▪ Monitor agent outputs for manipulative or biased behavior (e.g., unfair pricing or recommendations). ▪ Evaluate recommendations for fairness and ensure they do not disproportionately impact vulnerable groups. ▪ Implement behavioral consistency checks to identify and address erratic or unfair decision-making.
6. Not providing human intervention for a processing that can have a legal or important effect on the data subject.	Lack of human oversight in automated decisions, such as booking flights or scheduling, could lead to significant user inconvenience or adverse impacts.	▪ Require user confirmation for critical decisions, such as booking or payment. ▪ Implement fallback mechanisms where human oversight is necessary for high-stakes scenarios. ▪ Train users on how to interpret AI outputs and intervene if necessary.
7. Not granting data subjects their rights	Users may struggle to exercise GDPR rights (e.g., access, rectification, deletion) due to complex vendor dependencies or inadequate user interfaces.	▪ Provide clear interfaces for users to access, rectify, or delete their data. ▪ Maintain detailed, accessible logs of actions for audit purposes and compliance with Article 15 GDPR. ▪ Develop robust processes for handling user requests promptly, even when involving third-party integrations.
8. Unlawful repurpose of personal data	Data repurposing risks may arise if travel, calendar, or preference data is used for purposes other than intended, such as targeted advertising.	▪ Restrict data use to the specific purposes outlined in the terms of service. ▪ Ensure vendor agreements explicitly prohibit the repurposing of collected data. ▪ Implement consent management systems to track user preferences and restrict secondary use.
9. Unlawful unlimited storage of personal data	If data retention expands beyond necessity, particularly for travel itineraries, calendars, or personal preferences.	▪ Define clear retention periods for different data types (e.g., calendar data, travel history). ▪ Automate data deletion processes once the data is no longer necessary for the purpose. ▪ Regularly audit storage systems to ensure compliance with retention policies.

[305] Zainea, A.A, ‘Automated Decision-Making in Online Platforms: Protection Against Discrimination and Manipulation of Behaviour’ (2024)

10. Unlawful transfer of personal data	Cross-border data sharing risks due to reliance on third-party platforms or services in jurisdictions without adequate data protection standards.	▪ Verify the location of third-party services and ensure compliance with GDPR cross-border transfer rules. ▪ Perform Transfer Impact Assessments (TIAs) for all external vendors. ▪ Use standard contractual clauses and other safeguards for data-sharing agreements with third-party providers.
11. Breach of the data minimization principle	Excessive data collection: The system may collect or process more data than necessary for fulfilling user requests (e.g., unnecessary calendar details or preferences).	▪ Limit data collection to what is strictly necessary for fulfilling user requests (e.g., exclude unnecessary calendar details). ▪ Implement input validation and filters to prevent over-collection of data. ▪ Use anonymization or pseudonymization to minimize the risk of misuse or exposure of collected data.

From identifying data flows to classifying risks and implementing mitigations, risk management is a continuous iterative journey. It requires consistent monitoring, stakeholder collaboration, and adjustments based on real-world observations and emerging technologies.

Risk management should remain adaptable, incorporating feedback and evolving alongside regulatory and technological advancements.

As we conclude this report, it is important to reiterate that while the risk management framework presented in this document provides guidance, every organization must customize its approach to address the specific nuances of their LLM based use cases.

Privacy and data protection are not static goals but ongoing commitments.

10. Reference to Tools, Methodologies, Benchmarks and Guidance

Evaluation Metrics for LLMs

Once an LLM is fully or partially trained, it is evaluated to assess its performance and to verify that it fulfills expectations in terms of accuracy, robustness and safety. This usually begins with an **intrinsic**[306] **evaluation**, which focuses on assessing the model in a controlled environment, on the tasks that it is designed for. This is a good way to test and improve things before deployment, when many other factors and confounding variables can interfere in obtaining an accurate representation of the model's capabilities. **Extrinsic evaluation** follows; by assessing the model in its 'real-world' implementation, it shows how well it generalizes to more complex data and how relevant it truly is. It is a more holistic way of assessing the model and the only way of observing its real impact, in the context of its deployment. For both types of evaluation, the choice of metrics and benchmarks must depend on the task the model is designed for and its intended use case.

Ethical and Safety Metrics

❖ **Bias Evaluation**

Bias evaluation involves testing whether an LLM generates outputs that disproportionately favor, or disadvantage specific groups of people based on demographic factors (such as gender, race, religion, etc.). This often reflects bias patterns present in the training data, as well as the LLM's ability to overlook or neutralize them in its learned patterns.

> **Example: Word Embedding Association Test (WEAT):** This test measures how strongly certain words are associated with particular groups of people, aiming to detect stereotypes in the model's word embeddings. For instance, comparing the proximity of words that indicate gender (such as names or pronouns) with various career-related words can point to gender bias in the word embeddings, such as 'man' being represented as closer to 'doctor', and 'woman' being embedded closer to 'nurse'. This can predict bias in the model's output as well.

❖ **Toxicity Detection**

Toxicity evaluation assesses how often LLMs generate harmful, offensive, or inappropriate content. This includes hate speech, insults, or harassment. What is considered 'inappropriate' content can be context-dependent; for instance, AI systems that interact with children might have a lower threshold for inappropriate content than adult-only systems.

> **Example: Toxicity Score:** This metric aims to predict the probability of a piece of text being considered 'toxic'. Usually expressed as a percentage, the closer this score is to 0, the less likely it is for the text to be toxic. This metric is used in toxicity detection tools such as Perspective API, aiming to detect and reduce toxicity and harmful content in textual data.

❖ **Fairness Metrics**

Fairness evaluation focuses on evaluating the extent to which LLMs treat all user groups equitably without exhibiting or perpetuating systematic biases. This is of course tricky because fairness is an inherently complex term whose definition is debated and open to interpretation. Therefore, the

[306] Verma, A., Plain English AI,'NLP evaluation: Intrinsic vs. extrinsic assessment' Medium (2023) https://ai.plainenglish.io/nlp-evaluation-intrinsic-vs-extrinsic-assessment-ff1401505631

chosen metrics are usually geared towards optimizing the definitions/dimensions of fairness that are most appropriate for each given case.

> **Example: Demographic Parity:** Initially a metric used in classification, demographic parity can be adapted to a text output. It measures whether the model generates text that represents all demographic groups equally in terms of frequency, sentiment, and associations. It can answer questions such as 'are individuals of different ethnicities represented equally positively in the generated text?' or 'are women as frequently associated with high athletic performance as men?'.

Benchmarks

Benchmarks are standardized datasets, tasks, and evaluation protocols used to measure and compare the performance of various AI models, including LLMs. They provide a consistent framework to assess a model's capabilities, ensuring that performance can be compared across different models, tasks, and implementations.

Here are some common benchmarks for LLMs:

- **General Language Understanding Evaluation (GLUE):** A collection of tasks designed to evaluate natural language understanding, including sentiment analysis and sentence similarity. This benchmark is model-agnostic, meaning that it can be used to assess any system that takes a text input and generates a text output. Given the considerable recent progress of language models, the **SuperGLUE** benchmark has been introduced as a more challenging and nuanced version of GLUE. It includes more advanced language understanding tasks and a public leaderboard for state of the art models. https://gluebenchmark.com/

- **Massive Multitask Language Understanding (MMLU):** This benchmark evaluates the performance of language models across a wide range of subjects to assess their general knowledge and reasoning abilities. Models are tested on their ability to answer questions accurately. A higher score indicates better performance. https://github.com/hendrycks/test

- **ChatbotArena:** (lmarena.ai) is an open source platform for evaluating AI through human preference, developed by researchers at UC Berkeley SkyLab and LMSYS. With over 1,000,000 user votes, the platform ranks best LLM and AI chatbots using the Bradley-Terry model to generate live leaderboards. https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

- **AlpacaEval**: An LLM-based automatic evaluation based on the AlpacaFarm evaluation set, which tests the ability of models to follow general user instructions. https://tatsu-lab.github.io/alpaca_eval/

- **HellaSwag**: A challenge dataset for evaluating commonsense NLI that is specially hard for state of the art models, though its questions are trivial for humans (>95% accuracy). https://rowanzellers.com/hellaswag/

- **Big-Bench (Beyond the Imitation Game Benchmark):** A set of tasks designed to evaluate the capabilities and limitations of LLMs on diverse and challenging tasks. These tasks are designed to test abilities beyond what is evaluated by standard benchmarks, assessing abstract reasoning, problem-solving, or the ability to handle more unconventional or complex prompts. The higher the BIG-bench score, the better the model performs in complex tasks. https://github.com/google/BIG-bench

- **AIR-BENCH** 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies (https://arxiv.org/pdf/2407.17436v2) & https://huggingface.co/datasets/stanford-crfm/air-bench-2024

- **MLCommons AILuminate**: benchmark for general purpose AI chat model (https://ailuminate.mlcommons.org/benchmarks/) & https://drive.google.com/file/d/1jVYoSGJHtDo1zQLTzU7QXDkRMZIberdo/view

- **ToolSandbox:** A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities. https://machinelearning.apple.com/research/toolsandbox-stateful-conversational-llm-benchmark

- **CYBERSECEVAL 3**: Security benchmarks for LLMs. https://ai.meta.com/research/publications/cyberseceval-3-advancing-the-evaluation-of-cybersecurity-risks-and-capabilities-in-large-language-models/

- **LLM Guard** (by Protect AI): It is a comprehensive tool designed to fortify the security of Large Language Models (LLMs). https://llm-guard.com/

Safeguards/Guardrails in LLMs

Safeguards (or guardrails) in LLMs are mechanisms implemented to ensure that the models operate in a safe, ethical, and reliable manner. They can be applied to various stages of the LLM pipeline (pre-processing, training, output…), and be focused on addressing different risks. For instance, some safeguards aim to avoid the generation of unethical, harmful or inappropriate content (so the behavior of the model), while others focus on preserving the privacy of the owners of the data (or other stakeholders).

Here are some examples of behavioral guardrails that aim to moderate the LLM's output and mitigate harm that could be caused by the output without intervention:

- **Content filters:** moderate outputs by blocking or flagging harmful or toxic content
- **Prompt refusals:** prevent responses to dangerous or unethical prompts (like a request for instructions to a successful robbery)
- **Bias mitigation:** reduce stereotypical or unfair outputs during inference
- **Human-in-the-Loop approaches:** human oversight for high-risk applications, in order to not leave important decision-making fully in the 'hands' of an automated system, which cannot truly comprehend what is at stake.
- **Post-processing detoxification:** filter or rewrite outputs to remove harmful content
- **Adversarial testing (red teaming):** evaluate and stress-test the model's ability to successfully deal with harmful prompts

Other Tools and Guidance

Open Source Tools

- **Open source connector for agentic AI:** Anthropic Model Context Protocol (MCP): The Model Context Protocol is an open standard that enables developers to build secure, two-way connections between their data sources and AI-powered tools. https://www.anthropic.com/news/model-context-protocol
- **Tool for evaluation the performance of LLM APIs:** https://github.com/ray-project/llmperf

- **OWASP AI Exchange:** Comprehensive guidance on how to protect AI and data-centric systems against security threats. https://owaspai.org/
- **OWASP Top 10 for Large Language Model Applications:** Potential security risks when deploying and managing Large Language Models (LLMs). https://owasp.org/www-project-top-10-for-large-language-model-applications/
- **Five things privacy experts know about AI**: Blog from Ted (Damien Desfontaines) about privacy and research. https://desfontain.es/blog/privacy-in-ai.html

Privacy Preserving LLMS Techniques and Tools:

- **Clio: A system for privacy-preserving insights into real-world AI use**: https://www.anthropic.com/research/clio?mkt_tok=MTM4LUVaTS0wNDIAAAGXdMNyd9wb7R_UwjEGnaZw3fon7gu2FlLKUFOBA6PV2zTsuHYfcEeh1AJrOtEw8iVffJa-plco04sz7_vou0k2RQ6hHf6oZbd-c3SQb8ERj8aw
- **New approach to help assess the risk of re-identification in data release**: Rocher, L., Hendrickx, J.M. & Montjoye, YA.d. A scaling law to model the effectiveness of identification techniques. *Nat Commun* **16**, 347 (2025). https://doi.org/10.1038/s41467-024-55296-6
- **RAG technique with differential privacy guarantees:** https://github.com/sarus-tech/dp-rag
- **PrivacyLens** - A Data Construction and Multi-level Evaluation Framework. https://github.com/SALT-NLP/PrivacyLens
- **SynthPAI:** A Synthetic Dataset for Personal Attribute Inference. https://paperswithcode.com/dataset/synthpai
- **PrivacyRestore**: Privacy-Preserving Inference in Large Language Models via Privacy Removal and Restoration. https://arxiv.org/abs/2406.01394
- **LLM-PBE**: Assessing Data Privacy in Large Language Models. https://www.vldb.org/pvldb/vol17/p3201-li.pdf

Tools to Help Flag or Anonymize Sensitive Information

- **Google Cloud Data Loss Prevention (DLP)**: https://cloud.google.com/security/products/dlp
- **Microsoft Presidio** (Data Protection and De-identification SDK): https://github.com/microsoft/presidio
- https://medium.com/@parasmadan.in/understanding-the-importance-of-microsoft-presidio-in-large-language-models-llms-12728b0f9c1c
- **OpenAI Moderation API** (Identify potentially harmful content in text and images): https://platform.openai.com/docs/guides/moderation
- **Hugging Face NER models for Name Entity Recognition**:
    - dslim/bert-base-NER: https://huggingface.co/dslim/bert-base-NER
    - dslim/distilbert-NER: https://huggingface.co/dslim/distilbert-NER
- **SpaCy**: https://spacy.io/universe/project/video-spacys-ner-model-alt
- **NIST Collaborative Research Cycle** on data deidentification techniques: https://pages.nist.gov/privacy_collaborative_research_cycle/

Methodologies and Tools for the Identification of Data Protection and Privacy Risks

- Practical Library of Threats (**PLOT4ai**) is a threat modeling methodology for the identification of risks in AI systems. It also contains a library with more than 80 risks specific to AI systems: https://plot4.ai/
- **MITRE ATLAS™** (Adversarial Threat Landscape for Artificial-Intelligence Systems), is a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems: https://atlas.mitre.org/
- Assessment List for Trustworthy Artificial Intelligence (**ALTAI**) is a checklist that guides developers and deployers of AI systems in implementing trustworthy AI principles: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Guidance

- **OECD AI Language Models:** https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/04/ai-language-models_46d9d9b4/13d38f92-en.pdf
- **NIST GenAI Security:** https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf
- **NIST Artificial Intelligence Risk Management Framework - NIST AI 600-1:** https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- **OECD Advancing accountability in AI Governing and managing risks throughout the lifecycle for trustworthy AI:** https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/02/advancing-accountability-in-ai_753bf8c8/2448f04b-en.pdf
- **FRIA methodology for AI design and development:** https://apdcat.gencat.cat/es/documentacio/intelligencia_artificial/index.html
- **AI Cyber Security Code of Practice (gov.uk):** https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice

Standards

The European Standardisation Body CEN/CENELEC is currently developing different AI harmonized standards following the AI Act Standardization Request[307] from the European Commission.

High-risk AI systems or general-purpose AI models that comply with these forthcoming harmonized standards are presumed to meet the specific requirements outlined in the AI Act[308]. However, this presumption does not extend to international standards such as ISO/IEC 42001[309] and ISO/IEC 23894.[310]Nevertheless, these standards provide a robust foundation and offer valuable best practices.

[307] European Commission, 'Implementing decision C(2023)3215 final of 22.5.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence' (2023), https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en

[308] Article 40 AI Act

[309] Information technology — Artificial intelligence — Management system

[310] Information technology — Artificial intelligence — Guidance on risk management

edpb

European Data Protection Board