EBOOK

privacera

# CxO Guide to Data Security for LLM Use Cases

## How to Ensure Data Privacy and Security Across Internal and External LLM Data

## Excitement around ChatGPT—the large language model (LLM)-based chatbot—

captivated 180.5 million users at an 81% increase from January to August 2023[1]. LLM-based innovations, like ChatGPT, are moving at the speed of light. Their promise of greater flexibility, performance, and accuracy has organizations and users eager to take advantage of the opportunities they provide.

But with these opportunities comes the critical challenge to ensure data privacy and security keep pace. As a CxO or data stakeholder—platform owners, CTOs, enterprise architects, CDOs, CAOs, security and compliance owners, and data stewards—you face the growing dilemma of training and using LLMs responsibly.

In addition to the extreme pressure to deliver on generative AI (GenAI) use cases, you must balance the risks and mandates around privacy, security, and compliance. How do you protect your company's data privacy and security, while ensuring compliance in your LLM?

1. Anna Tong, "Exclusive: ChatGPT traffic slips again for third month in a row." Reuters, Sept. 7, 2023, https://www.reuters.com/technology/chatgpt-traffic-slips-again-third-month-row-2023-09-07/.

As critical as these concerns are, a report from Quantum Black by McKinsey states **only 21% of organizations have established artificial intelligence (AI) governance policies relating to employee use of AI[2].** Even more alarming is that **79% of companies are taking on generative AI and LLMs with no governance policies in place to ensure data privacy, security, or compliance.**

As you embark on leveraging LLMs to produce your GenAI applications, follow this guide's **six vital steps for CxOs** to secure your enterprise's most sensitive data with confidence. Discover how you can safely secure and govern your entire data lifecycle—from raw to training data to how users interact with LLMs.

2. "The State of AI in 2023: Generative AI's Breakout Year." Quantum Black AI by McKinsey, 2023, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

# Inside the Realities of LLMs

The degree of insufficient transparency and explainability that accompany LLMs raises concerns about the risks organizations face without data protections and governance in place. Take a look at the key concerns of using LLMs, the risks they pose, and their potential impact.

## Concerns

- **Toxicity:** Rude, disrespectful, or unreasonable data that might cause a user to abandon a conversation.
- **Training data poisoning:** Injecting polluted data to manipulate a model's behavior and deliver false results.
- **Hallucinations:** Generated text that's erroneous, nonsensical, or detached from reality.
- **Shadow AI:** Applications and tools implemented and used without IT's knowledge or control.

## Risks

- **Bias:** Data that contains or spurs negative stereotypes, prejudices, or discrimination against certain groups or individuals, and could lead to unethical decisions.
- **Intellectual property:** Accidental or intentional disclosure of trade secrets, proprietary information, or confidential data rules.
- **Privacy:** Exposure of sensitive data or personally identifiable information (PII) through third-party prompts when data is stored online and made public when data is released accidentally.
- **Security:** Leaked confidential information, such as sensitive model data, through user responses can lead to unauthorized data access and severe privacy violations.
- **Governance:** Infringement of internal and external governance and security rules, such as GDPR and CCPA, for data access and sharing.

## Impact

The impact of these risks can devastate any organization—from ruined reputation as well as customer and employee relationships to loss of revenue and damaged overall financial wellbeing. All it takes is a single privacy or compliance violation involving PII, sensitive, proprietary, financial, or healthcare data.

Here's where you can expect to incur expenses if your company experiences a data breach:

- **Detection and escalations:** Assessment and auditing, crisis management, and notices to CxOs, upper management, and board of directors.
- **Notifications:** Communication with regulators, outside security experts, and persons with data impacted by the incident.
- **Post-breach activities:** Regulatory fines, legal expenses, credit monitoring, and identity protection services.
- **Business loss:** Operational and business downtime, decreased revenue, customer attrition, new customer acquisition, and reputation recovery.

**The potential risks to data in your LLMs aren't worth the impact to your organization. The simple truth is: you can't afford to let your data go unprotected.**

# €1.2 billion

GDPR fine issued to Meta in May 2023 for transfer of European consumer data to the US without data protections in place[3]

# $1.19 billion

Fine issued to Didi Global in 2022 in violation of Chinese network security, data security, and personal information protection laws[4]

# $392 million

Penalty against Google in 2022 for a lack of data privacy over location tracking[5]

# $250,000 – $1.5 million

Maximum fine per year for a HIPAA violation[6]

3. Adam Satariano. "Meta Fined $1.3 Billion for Violating E.U. Data Privacy Rules." New York Times, May 22, 2023, https://www.nytimes.com/2023/05/22/business/meta-facebook-eu-privacy-fine.html.
4. Kendra Barnett. "Google's $400m penalty and impact of the 5 heftiest data privacy fines on 2023 ad plans." The Drum, November 15, 2022, https://www.thedrum.com/news/2022/11/15/googles-400m-penalty-the-impact-the-5-heftiest-data-privacy-fines-2023-ad-plans.
5. Cecilia Kang. "Google Agrees to $392 Million Privacy Settlement With 40 States." New York Times, November 14, 2022, https://www.nytimes.com/2022/11/14/technology/google-privacy-settlement.html.
6. "HIPAA Violation Fines." The HIPAA Journal. Last accessed October 17, 2023. https://www.hipaajournal.com/hipaa-violation-fines/.

# 6-Step Guide to LLM Data Protection and Governance

To safeguard your data from the potential leaks, exposures, and data breaches LLMs pose, you must apply consistent privacy and governance controls. This level of security gives you peace of mind in knowing your data will stay out of the wrong hands and remain safe even when used in third-party LLMs. Follow these six steps to find out how.

## 1 Define governance and security policies

Define governance and security policies in natural language to enforce desired and expected behaviors across your models and applications. Ensure your policies include:

- **Adaptive access controls** based on application, device, and user role, for example, so you know who gets access to which information.
- **Least privilege access** to ensure your users, systems, and processes access only the necessary resources to perform their tasks.
- **Role-based access control (RBAC) and attribute-based access control (ABAC)** defined by policies according to user roles and attributes such as location.

These policies are fundamental to data privacy and governance to help you gain a sense of who has access to which models and in which context.

## 2 Detect code, PII, and other forms of sensitive data

Before you upload data into your LLM for training, scan, classify, and tag sensitive data, PII, and even proprietary code. By appropriately tagging and inventorying your data, you establish usage guidelines.

If pre-built classifications and rules are available, enable them for immediate use and expand them as needed based on your requirements. After you build your model, continue to scan, classify, and tag data.

## 3 Redact and encrypt data

Going even deeper, use real-time controls to enforce data masking, encryption, and removal of sensitive data elements. Apply fine-grained, sensitive data encryption and masking based on user attributes, data classification, and tags, or at the table, column, or field level.

These controls protect data even if unauthorized access occurs, whether from internal or external users or even malicious actors. Between tagging your data and applying data-level controls, you redact, de-identify, mask, encrypt, or even remove sensitive data and data that could introduce vulnerabilities in the pre-training data pipeline.

## 4 Monitor and analyze user actions and model usage

Apply extensive fine-grained auditing of user interactions and model usage and responses, regardless of the user interface. Use query analytics to monitor and analyze your model and user actions, such as prompts requesting sensitive or private data, or questions that lead to bias or toxic responses.

Through these insights, you can see the percentage of messages that are allowed, denied, or masked for both users and the LLM. You can also identify security and risk patterns, so you can take critical steps to enforce further security and privacy controls.

## 5 Detect and respond to unauthorized or inappropriate prompt injections

Pre-filter questions based on users' security and privacy settings to further safeguard data privacy and information confidentiality. Scan user inputs, queries, and model responses in real-time for sensitive data elements. Allow/deny and redact/de-identify them based on user identity and data access permissions.

If a user prompt or question contains sensitive information, your LLM should respond immediately with an error message indicating that the question isn't allowed. This measure prevents the model from inadvertently disclosing private information or otherwise negative responses.

## 6 Control the rate of incoming requests

Limit the number of requests a user can make within a certain amount of time by implementing API-request throttling. This level of control in your LLM protects it from overuse, abuse, and malicious attacks, such as potential denial-of-service (DOS) attacks.

# What Privacy Controls Mean for LLM Data

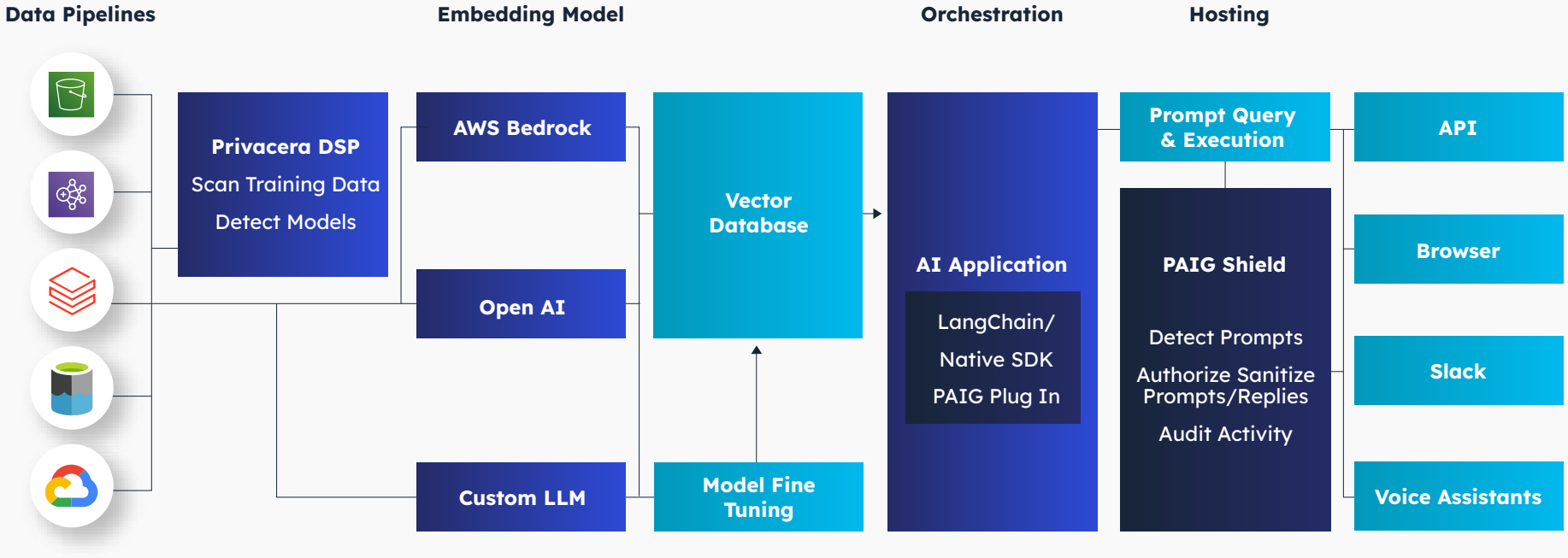
See the difference in your LLM data when you establish privacy controls.

| Without Data Privacy Controls | With Data Privacy Controls |
|---|---|
| **Bias:** The potential release of personal data attributes, such as race, sex, age, and ethnicity, could cause models to produce biased information or unethical decisions. | **Reduce bias:** Scan and redact user inputs and model outputs based on context in real time to keep models from returning biased information or that could result in unethical decisions. |
| **Intellectual property:** Because LLMs use large amounts of data, they could leak trade secrets, proprietary information, or confidential data rules. | **Prevent intellectual property leaks:** Enable access controls that prevent trade secrets, proprietary information, or confidential data rules from getting into the wrong hands. Even if the data is shared through third-party LLMs, it remains protected. |
| **Privacy:** If sensitive, classified, or private information goes unprotected into training models, companies risk costly data breaches and hefty fines for non-compliance. | **Ensure privacy:** Protect sensitive, classified, and private information that goes into training models, and avoid costly data breaches and hefty fines for non-compliance. |
| **Governance and security:** LLMs could infringe on internal and external governance and security rules for data access and sharing. Or third parties could attack them or train them to leak information or cause damage. | **Enable continuous governance and security:** Continuously audit, monitor, and track model usage and user activity. |

privacera | paig

# Protect Data with Privacera AI Governance

The one solution that enables you to follow all six steps in this guide is at your fingertips. Meet Privacera AI Governance (PAIG). Built on Privacera's strengths of responsible data access, PAIG drives dynamic and consistent data security, privacy, and access governance across your purpose-built LLMs and enterprise data landscape.



privacera | paig

With PAIG, you gain control over who can use your models and the context for which they can use them, as well as who can see and request PII and sensitive data:

- **Mask and filter out PII and sensitive data.**
- **See how users are leveraging GenAI.**
- **Identify which models have PII data and who is accessing that data.**
- **View unauthorized requests and responses in real time.**
- **Review prompts, contexts, and responses.**

Plus, PAIG integrates with popular external and local LLMs. It also works with Privacera's policy model and ensures your data meets the latest data privacy regulations and compliance such as GDPR, CCPA, and HIPAA.

## Unified AI Governance with PAIG

- Real-time discovery and tagging.
- Data access controls, masking, and encryption.
- Policy-based allow/deny prompts and responses.
- Redact/de-identify sensitive data.
- AI-powered auditing and monitoring.

KEY BENEFITS

Securely Innovate

Protect PII

Prevent IP Leakage

Track & Audit

PAIG empowers enterprises with confidence to train their LLMs, ensuring their data remains secure, well-governed, and regulatory-compliant. Learn more about how you can leverage PAIG to enhance data security, privacy, compliance, and governance for LLMs—get our PAIG whitepaper.

Fortune 500 enterprises trust Privacera for their universal data security, access control, and governance. Discover how to streamline data security governance with Privacera.

# Take a unified approach to data access, privacy, and security with Privacera.

REQUEST A PAIG DEMO → CONTACT US →

Privacera, based in Fremont, CA, was founded in 2016 by the creators of Apache Ranger™ and Apache Atlas. Delivering trusted and timely access to data consumers, Privacera provides data privacy, security, and governance through its SaaS-based unified data security platform. Privacera's latest innovation, Privacera AI Governance (PAIG), is the industry's first AI data security governance solution. Privacera serves Fortune 500 clients across finance, insurance, life sciences, retail, media, consumer, and government entities. The company achieved AWS Data and Analytics Competency Status, and partners with and supports leading data sources, including AWS, Snowflake, Databricks, Azure and Google. Privacera is recognized as a leader in the 2023 GigaOm Radar for Data Governance; was named a 2022 CISO Choice Awards Finalist; and received the 2022 Digital Innovator Award. The company is also named a "Sample Vendor" for data security platforms in the Gartner® Hype Cycle™ for Data Security, 2023. Learn more about Privacera at privacera.com.

privacera