WHITEPAPER

privacera

Use GenAI to Safely Disrupt Your Industry and Workforce

privacera

contents

KEY TAKEAWAYS

- Generative AI (GenAI) will transform how your organization operates, but it also brings real risks.
- Your GenAI Architecture must contain an AI Data Governance Layer to minimize business risks.
- The AI Data Governance Layer must be able to secure embedding and training data; secure model inputs and outputs; and provide comprehensive compliance monitoring.

Time to Move

According to McKinsey's Global Survey on the State of AI, 40% of respondents say their organizations will increase their investment in AI because of advances in GenAI. McKinsey expects business disruption from GenAI will be significant, and survey respondents predict there will be meaningful changes to their workforces. Interestingly, 79% say they've had at least some exposure to GenAI, either for work or outside of work, and 22% say they're regularly using GenAI in their own work.

The most commonly reported business functions using these newer tools are marketing and sales, product and service development, and service operations, such as customer care and back-office support. For this reason, McKinsey finds the conversation in the C-suite has moved from rudimentary to sophisticated.

However, rank-and-file employees seem to have gotten the value faster than their bosses. A recent EY Consulting survey found 8 out of 9 employees get value in emerging technologies. These workers believe adopting emerging technologies will be beneficial for their company. At the same time, they believe their senior leadership is too slow to embrace emerging technology. In fact, 59% of employees say their senior leadership, including C-suite and vice presidents, are slow to embrace market-changing technologies. Additionally, 52% claim market-changing technologies will be outdated by the time their company implements them. Being late to the game is a business risk.

Considering GenAI Business Risk

"The leaders of modern firms cannot afford to ignore this new generation of ethical challenges."[1]

> "If a company is relying heavily on AI in its business, it needs to ensure that the AI systems it uses are ethical and trustworthy, or it is likely to lose more from AI than it gains."[2]

And Privacera Co-Founder and CEO Balaji Ganesan says, "The potential of generative AI and large language models to transform enterprise operations is immense, but their inherent unpredictability can unknowingly reveal intellectual property, personally identifiable information, and other sensitive data."

Few companies today seem fully prepared for the widespread use of GenAI or the business risks these tools may bring. Just 21% of respondents in the McKinsey study who reported AI adoption say their organizations have established policies for governing employees' use of GenAI technologies in their work. Interestingly, rank-and-file workers intuitively understand the perils, agreeing on cybersecurity risk as a significant barrier to adoption.

In terms of risks, McKinsey says companies need to mitigate the risk of GenAI being inaccurate. Unfortunately, their survey finds just a little over 20% of companies have risk policies in place for GenAI. Those policies should be about protecting a company's proprietary information such as data, knowledge, and other intellectual property. The trap, however, is companies are looking at the risk too narrowly. From social and humanitarian to sustainability and beyond, there is an extensive and alarming range of risks.

The authors of the book "Rewired" suggest there needs to be "clear standards and thresholds for AI risk, including transparency and explainability, automated AI model monitoring systems, and bias and fairness checks for AI models."[3] They go on to suggest "automating trust is the process of turning trust policies into code such as compliance requirements and risk standards. These automate risk controls whenever anyone submits new code. This approach radically speeds up development and cuts back on risk."[4]

Put together, mitigating the risks associated with GenAI requires a combination of technical, ethical, and regulatory measures. The complete avoidance of risks is challenging. But the following strategies help minimize and manage those risks.

Requirements for a Solution

As AI and large language models (LLMs) become more commonplace, organizations need the ability to manage, monitor, and control these models. An effective solution manages the use of AI models, regulates data interaction, enforces access controls, and maintains comprehensive, securely accessed audit trails. AI Governance should have the ability to detect inappropriate sharing of content, potential security threats, such as unusual user behavior or brute force attacks, and enable swift response mechanisms.

A solution should include the ability to create robust tagging and classification of sensitive data, detailed insights, and data protection compliance. From a regulatory compliance perspective, it should address privacy, security, and ethical implications. This includes requirements from regulations such as GDPR and CCPA.

As a goal, the governance solution should enable the organization to ensure data is handled responsibly, including data used by or produced by AI models. This solution should aid in compliance by providing the ability to track user activity. Without proper governance, organizations risk financial and reputational damage, misuse of AI models, and unintentional regulatory non-compliance.

As a goal, the solution should also improve operational efficiency by consolidating AI model management. The aim should be to reduce the resources required to manage AI models and ensure management consistency across the organization. Specifically, an effective solution will have the following capabilities:

GenAI Model Catalog: The ability to catalog, describe, tag, and manage permissions for AI models.

Privacy and Data Protection: The ability to track handling of personally identifiable information (PII) and ensure compliance against global and regional data privacy regulations.

Auditability: The ability to audit data requests and actions taken in order to provide accountability and transparency.

Protect Intellectual Property: The ability to prevent GenAI from leaking intellectual property or other sensitive information.

Observability and Analytics: Visibility into the usage, performance, and adherence to governance policies of AI models. The ability to ensure practices align with governance guidelines and external regulatory compliance. This includes dashboards offering real-time insights into activity, user behavior, and threats. These should work in tandem with security alerts and facilitate rapid incident response and efficient risk management.

Security Measures: Implementation of sophisticated security measures such as intrusion detection, user behavior analytics, and secure design to safeguard sensitive data and counter potential threats.

Scalability: The ability to manage realistic volumes of data and requests, supporting businesses as they grow and their AI needs expand.

Throttling Mechanisms: The ability to protect against potential system abuse and ensure fair resource distribution through user-request throttling and selective throttling of service users.

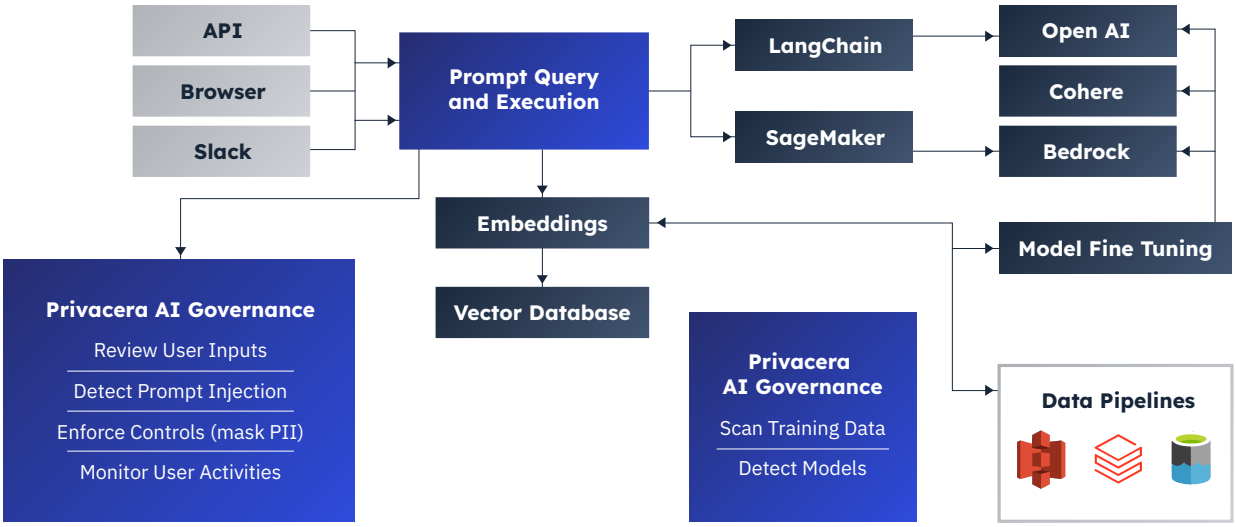
Toxic Behavior Monitoring: The ability to scan and block inappropriate content in training and operation that could be ethically or morally offensive, including other types of inappropriate behavior that can damage a brand or cause business and operational disruption.

Security and Access Control: Sophisticated security measures and controlled access to protect sensitive data and ensure only authorized users have access to certain data. This includes fine-grained access control that allows for provisioning of permissions based on specific user roles and attributes, limiting data based on the principle of least privilege.

Completing Your LLM Architecture

Adding AI and data security governance capabilities, such as Privacera, to the LLM app stack establishes an enterprise-grade LLM solution that eliminates many of the data risks discussed above. The following highlights an app-stack example and how the components work together as a solution architecture.

Starting at the bottom, data is input by a data pipeline. “The purposes of a pipeline is to make clean, consistent data available”[5] to the LLM. Here Privacera is used to scan training data being ingested from the data pipeline before it is used by the LLM for model fine tuning. The aim here is to protect against model bias and prevent the model from accessing certain sensitive data. This is done by discovering any data that is sensitive or can lead to bias and prevent, for example, things that will impact fair credit or reporting or hiring a diverse team. Data that can lead to model bias can be masked, redacted, or blocked from the model.



Model responses can also leverage embeddings. Embeddings data representations carry semantic information critical to LLMs understanding and maintaining long-term memory. Embeddings are generated by LLMs. These features represent the different dimensions of the data essential for understanding patterns, relationships, and underlying structures.

Embeddings are stored in a vector database. The vector database is responsible for efficiently storing, comparing, and retrieving billions of embeddings (i.e. vectors). A vector database uses a combination of algorithms. These utilize approaches such as approximate nearest neighbor (ANN) search. These algorithms optimize the search through hashing, quantization, or graph-based search. The algorithms are assembled into a pipeline that provides fast and accurate retrieval of the neighbors of a queried vector. Here, Privacera scans embeddings as they are being created or within the vector database. The goal here is to protect sensitive data when embeddings are used for user inquiries.

Queries are created at the point of prompt query and execution. The chat agent is connected via an API, browser, or messaging platform such as Slack. These can also connect to the LLM through orchestration frameworks like LangChain and SageMaker.

The chat agent operates by constructing a series of prompts to submit to the language model. A compiled prompt typically combines a prompt template hard-coded by the developer; examples of valid outputs called few-shot examples; any necessary information retrieved from external APIs; and a set of relevant documents retrieved from the vector database. Before this happens, information

is first passed to Privacera to ensure the user has rights to the information being requested and no sensitive data is included in the user prompt that the model should not be exposed to. Here, Privacera has the context-aware intelligence to detect queries that could result in inappropriate structured or unstructured data from being shared. Once a prompt is determined to be legitimate, the prompt is forwarded to the chat agent. Where the user does not have access to an element of data, Privacera AI Goverannce masks or does not allow the sharing of the information.

Key Enabling Privacera Capabilities

Secure Embedding and Training Data

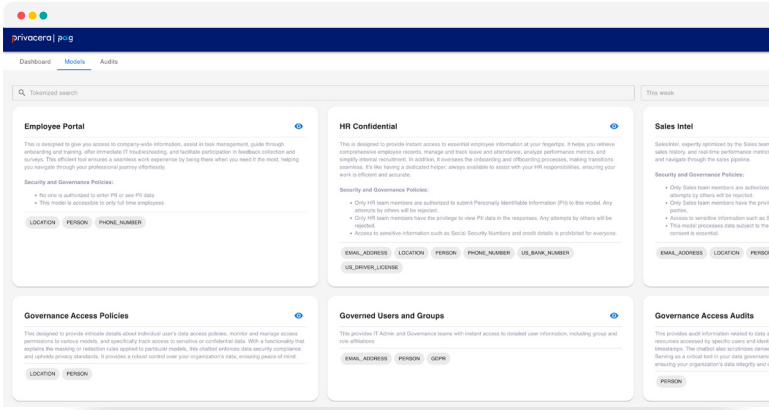
PAIG continuously scans training data for sensitive data before it is ingested into foundational LLMs. PAIG automatically tags this data. Additionally, sensitive training data can be masked or blocked from being utilized in models, reducing PII exposure and model bias. PAIG also uses rules, machine learning (ML), and context to scan and tag model embeddings to identify sensitive content. This is accomplished as they are created or while they are within the vector database. Additionally, PAIG provides a model catalog that displays:

- Model description
- Sensitive data the model contains
- Security and governance policies applied to the model

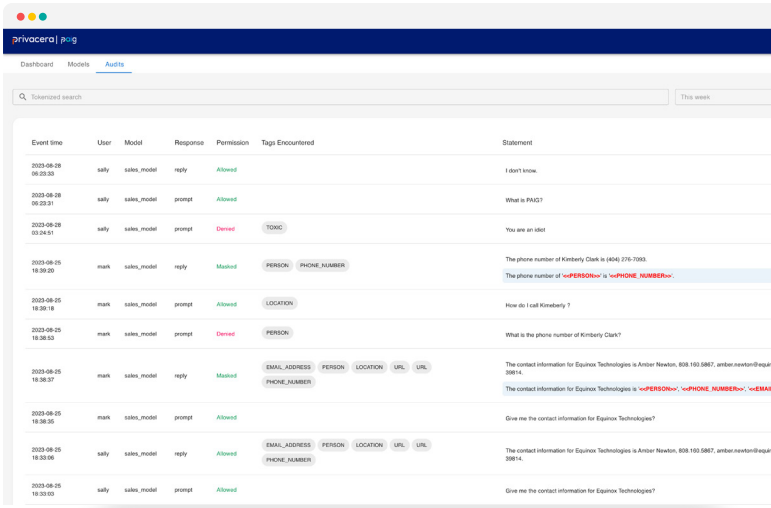
Critical information to understand the model and how to use it

PAIG's RESTful software development kit (SDK) enables customers to connect to their choice of common LLM libraries such as SageMaker, Langchain, Llama, transformers, and OpenAI.

Plus, the PAIG portal provides dashboards for audit review. Audits include tags for any PII. And PII is provided in a 'break the glass' operation. Authorized people accessing PII data are logged and audited. Ensuring all actions are trackable is essential for security and accountability and critical for complying with data privacy laws regulations.

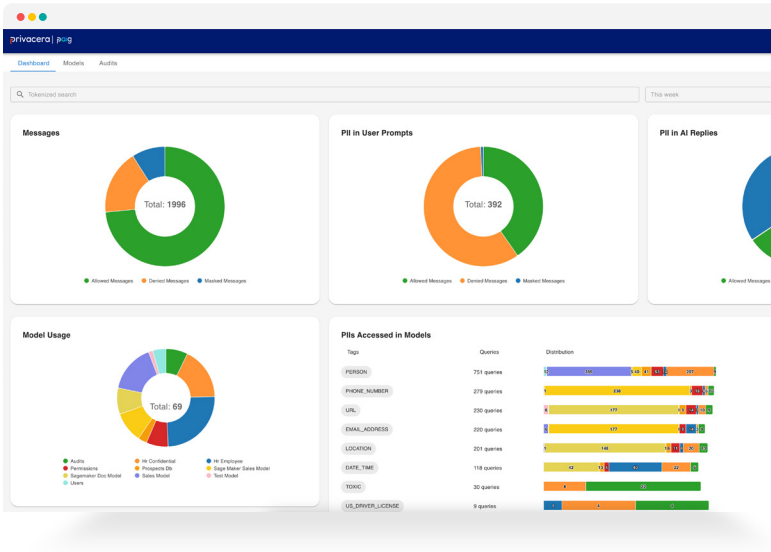


Secure Model Inputs and Outputs



PAIG protects data exposure using context-aware data protection, inspecting user-prompted queries and masking or redacting sensitive data before it enters the model. Additionally, Attribute-Based Access Control (ABAC) can be applied to mask sensitive data model output, ensuring users can only see data they are authorized to see. User prompts to the model are also inspected. Unauthorized questions that could expose sensitive data are denied. Not only does this capability add an additional layer of security, but it also eliminates the massively expensive and wasteful LLM compute costs associated with processing unauthorized requests.

Comprehensive Compliance Monitoring



PAIG provides comprehensive dashboards and audit logs of what sensitive data is leveraged in each model, how it is protected, and who is accessing it. PAIG audit logs show:

- Who is accessing what models
- What sensitive data they are accessing
- When they accessed the model
- Flagged, inappropriate conversations
- What protections were applied

Additionally, PAIG provides a security and compliance dashboard that provides a view of your entire model landscape, including an overview of approved requests, denied requests, and requests that require masking to be applied. The dashboard also provides an overview of all sensitive data across models. PAIG's audit log and dashboard simplifies model monitoring and compliance.

Conclusion

Without question, it is time to move on Generative AI. Your competitors are likely already in flight.

And your employees are ready for your leadership. The impacts to customer experience and other business functions will be significant. But doing so without appropriate data governance is a business catastrophe waiting to happen. What we have described is an architecture for safely moving forward with GenAI. It largely snaps into what you will already be doing. If you are ready to start, Privercia is ready with the data layer. Schedule a meeting with us and experience a PAIG demo.

Endnotes

1. Competing in the Age of AI, page 196
2. All in on AI, Tom Davenport and Nitin Mittal, page 21
3. Rewired, Harvard Business Review Press, page 331
4. Rewired, Harvard Business Review Press, page 333
5. Competing in the Age of AI, HBR Press, Page 72