Securing Generative AI in the Enterprise

In a race to adopt LLMs, companies need to overcome the last hurdle: Data privacy

Raluca Ada Popa, PhD, MIT
Associate Professor of Computer Science, UC Berkeley

Rishabh Poddar, PhD, UC Berkeley
CTO, Opaque Systems

White Paper
Published February 2024

Contents

As generative AI, particularly LLMs, continues to disrupt industries and alter the way organizations work with data, it's also introducing data privacy challenges that are crucial to address. Between security breaches and evolving privacy regulations, enterprises must adopt strategies to protect sensitive data if they want to maximize the benefits of LLMs.

Introduction

Artificial intelligence (AI) has come a long way over the past decade, transforming the way we interact with technology. Machine learning (ML) has been radically changing the way data can be processed and analyzed, enabling businesses across various industries to make data-driven decisions. Now, in addition to generating predictions and uncovering meaningful insights from existing data, AI systems can generate new, original content in human-like ways.

The advent of generative AI (GenAI) has captured the attention of businesses and industry leaders, scientific communities and tech enthusiasts, as well as the public in recent years. Large language models (LLMs), most notably OpenAI's ChatGPT, have emerged as one of the most widely used applications of generative AI. Because they are extensively trained on large datasets, LLMs have broad applicability across industries, from retail and marketing to finance and healthcare.

According to the Cisco 2024 Data Privacy Benchmark Study[1]:

- A significant majority of organizations, nearly four out of five, report getting considerable value from their use of GenAI.
- An overwhelming 92% of respondents perceive GenAI as a distinct innovation that necessitates new methods for handling data and associated risks.
- Nearly half of the organizations that were surveyed are incorporating non-public company details into their GenAI applications.
- More than two-thirds express concern that GenAI might negatively impact their company's legal standing and the protection of intellectual property.

Value versus risk

Nearly 80% of organizations report considerable value from GenAI use, yet there's a high level of concern over its potential to negatively impact legal standing and intellectual property protection.[1]

The benefits of generative AI and LLMs are immense, but such technologies introduce complications around data security and privacy. This has led to various workplaces banning the use of AI tools that can leak sensitive company information. A case in point is Samsung, which enforced strict rules regarding company use of generative AI systems after discovering that its employees accidentally leaked sensitive code by uploading it to ChatGPT.[2] Even countries are considering or implementing bans on generative AI tools. After Italy became the first country to ban ChatGPT following a data breach that affected OpenAI and temporarily exposed the titles of some users' chat histories to others,[3] the Italian Data Protection Authority notified OpenAI of potential violations of the European Union (EU)'s General Data Protection Regulation (GDPR).[4]

It's no surprise that such events, along with the uncertainty associated with generative AI, can cause the general population to be wary of AI applications. A recent survey found that more than half of respondents supported federal regulation of AI.[5]

Despite the explosive popularity of generative AI, this field is still in its infancy and can pose threats to data privacy without proper guidelines and safety measures in place. Understanding the complex and evolving landscape of generative AI is essential for businesses and enterprise organizations to successfully leverage this technology. This white paper explores the impact and challenges of generative AI and the privacy-preserving techniques available for securely adopting LLMs.

The impact of LLMs on privacy

Privacy concerns are growing, and the number of privacy-related security incidents are on the rise. GenAI adoption is also driving these concerns, especially when data includes sensitive information and when models are proprietary. The involvement of multiple stakeholders—most notably LLM providers and developers, data providers, and end users—naturally leads to concerns about privacy.

LLMs are revolutionizing the way machines comprehend and generate human language, offering numerous benefits across industries and organizations. Users and businesses can leverage LLM applications to write code, generate text for various forms of content ranging from newsletters and articles to product descriptions and marketing copy, summarize papers and reports, and so on. Businesses can also take advantage of AI-powered chatbots to provide immediate support to customers. These chatbots can answer questions and offer personalized recommendations in a conversational manner, improving the customer experience and optimizing business operations.

Insider threats on the rise

Companies are experiencing a 32% monthly increase in insider-related incidents, highlighting the urgent need for robust data protection strategies.[7]

However, the rapid development and adoption of AI technologies also raises concerns, particularly when it comes to trusting AI systems and the current safeguards in place for protecting privacy.[6] For example, using AI chatbots requires sending prompt and context data to a third party for processing. This creates reliance on the security practices and guarantees of service providers, who might not prioritize privacy and security, and leads to various privacy concerns for users and data owners.

According to a study conducted in 2023 by Vanson Bourne for Code42,[7] the perception about data privacy concerns was very real:

- **On average, companies are witnessing a 32% rise in insider-related incidents each month**, translating to about 300 such events annually. This uptick heightens the risk of data leaks and security breaches.
- **Chief information security officers (CISOs) identify the risk from insiders (27%) as the most challenging threat to pinpoint**, ranking it higher than issues related to cloud data exposures (26%) and malware or ransomware attacks (22%).
- **A significant 76% of CISOs anticipate an upsurge in data losses due to insider threats over the coming year**, attributing this to the ineffectiveness of current technologies and strategies in identifying and countering these risks.
- **With the average financial impact of an insider threat incident estimated at $16 million**, addressing insider risk is crucial to prevent severe financial, reputational, and compliance consequences.

Service providers also have their own set of concerns with protecting the intellectual property of their AI systems and algorithms. Given that much is left to be learned about generative AI models, heuristics-based techniques might not be enough to ensure model safety and privacy. The following sections discuss privacy-related challenges posed by generative AI and LLMs.

Data and model sharing

LLMs that generate relevant, accurate, reliable content in response to user queries rely heavily on learning from large amounts of quality data. To acquire such comprehensive, high-quality data, LLM providers likely need to collect data from multiple teams and organizations. Figure 1 shows how different types of data are involved in model training and inference.

Sensitive data during model training. Model training can be broadly divided into two stages: pretraining and fine-tuning.

- **Pretraining:** LLMs are initially trained on extensive datasets that undergo preprocessing to ensure the quality and effectiveness of the training data. This preprocessing typically includes the anonymization or redaction of personal information. However, given the sheet volume of data and the variety of sources from which it is collected, LLMs can sometimes be

Sensitive data exposure in LLMs

Both the fine-tuning and inference stages of LLMs pose risks of exposing sensitive data, whether through inadvertent inclusion in training datasets or through user inputs during model interaction.

exposed to sensitive data. Without robust measures in place to protect or remove such data, it could inadvertently become part of the training dataset and risk exposure.

- **Fine-tuning:** To perform specific tasks more effectively, a pretrained model needs to be retrained (in other words, fine-tuned) on data that's relevant to the specific task. Task-specific datasets might include sensitive data collected from organizations in a relevant domain. Even if personal identifiers and confidential information in the datasets are removed, the model can still learn new patterns and memorize aspects of the data that might lead to generating content that inadvertently reveals sensitive information. For example, an LLM trained on de-identified medical records could potentially leak details of an individual's medical history. If the model was trained on a corpus of data points that overlap with those in the medical records, it could create data linkages and enable re-identification of that individual.[8]

Sensitive data during inference. LLMs are designed to adapt their responses in real-time based on user inputs, which might contain sensitive or confidential information such as users' personally identifiable information (PII) or proprietary business data. In some cases, LLM providers use the input content as training data to improve their models. Even if LLM providers don't use customer or user data for training, all data undergoes a period of retention, during which the providers can access and view all user prompts and queries.

Proprietary models. On the other hand, LLM providers can face privacy challenges when companies want to fine-tune the provider's proprietary models on company data. If companies are unwilling to send their data to an LLM provider for fine-tuning, then the provider must deploy their proprietary models within those companies' environments. In this case, model owners depend on the security practices and guarantees of their customer organizations. This leads to concerns with potential competitors stealing and copying their models.

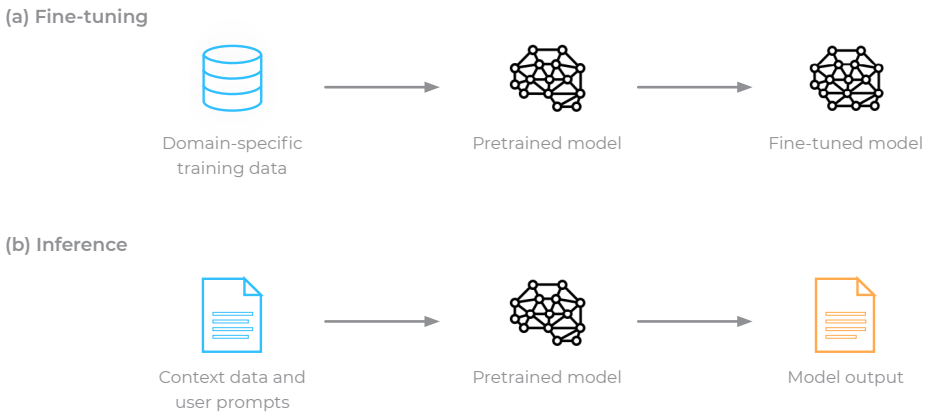


Figure 1: *(a) Fine-tuning an LLM, which might be proprietary, often involves the use of sensitive training data. (b) During inference, user inputs might contain sensitive data, and the model can potentially reveal sensitive details through its output.*

LLMs are susceptible to attacks

The accumulation of vast amounts of data in GenAI models makes them prime targets for cyberattacks, with inference attacks posing a significant threat to data privacy.

Collaborative AI/ML efforts, whether they involve combining data between multiple parties for training or sending data to a third party for processing, increase the risk of unauthorized access and can compromise the privacy of all parties involved. Organizations and users need a secure way to collaborate on generative AI while maintaining confidentiality and privacy of their respective data and AI models, even when they don't necessarily trust each other.

Security breaches and attacks

As more data gets fed into AI models, the risk of security breaches and attacks increases. AI systems that regularly accumulate vast amounts of data are particularly ideal targets for cyberattacks, which can result in identity theft, financial fraud, and other undesirable consequences for individuals and organizations.

Security breaches revealing input data. Any user data that's retained for a period of time by AI providers leaves the data vulnerable to leaks and attacks. User queries might contain sensitive information, such as PII or proprietary code, and can be potentially stolen. Even if input data is encrypted or protected in some way, today's LLMs still need to operate on plaintext. At some point in the process, models need to operate on the data in its unencrypted form, which risks data leakage and therefore requires an additional level of protection.

Inference attacks against ML models. Inference attacks are techniques used by adversaries to gain insights into sensitive or proprietary information about a model's training data or parameters by observing the model's outputs. Multiple types of inference attacks exist:

- **Attacks targeting training data:** Attackers exploit the fact that LLMs learn patterns from their training data to gain information about the data that a model was trained on. Even when the training data has been discarded, an attacker can submit carefully crafted inputs to the model and then analyze its responses to infer information about its training data. For example, *membership inference attacks* attempt to determine whether a particular data sample was part of the original training dataset for a model, while *model inversion attacks* aim to reconstruct data samples that a model was trained on. These attacks could lead to privacy breaches and violate individuals' privacy by exposing private and sensitive data.

- **Attacks targeting text embeddings:** Instead of training models on extensive data, organizations can use retrieval-augmented generation (RAG), a technique that fetches contextually relevant information from an external database to optimize the response to a query. However, like other ML systems, RAG-based systems can potentially be vulnerable to inference attacks. The text embeddings used in RAG workflows can reveal a significant portion of the original text as a result of inversion attacks.[9]

Evolving privacy regulations

GenAI systems are subject to existing and evolving privacy laws. Organizations need to stay informed and compliant to avoid legal and financial penalties.

Privacy regulations

The broad applicability of generative AI makes LLMs subject to existing privacy laws. Even if safeguards are implemented to prevent data leaks or attacks, collecting and processing certain types of data beyond their original purpose could be considered a privacy violation in certain jurisdictions. For example, the GDPR mandates, among other requirements, obtaining explicit and informed consent and having a lawful basis for processing personal data of EU citizens or residents.

As the world of generative AI continues to expand, so will privacy laws and regulations. This is already evidenced by California's proposed regulations governing the use of automated decision tools,[10] Canada's proposed Artificial Intelligence and Data Act (AIDA),[11] and the European Parliament's vote to pass the AI Act.[12] To prevent legal ramifications and avoid losing customers' trust, AI providers, developers, and researchers must continually keep up with regulatory requirements and ensure that they develop and deploy generative AI models that adequately protect user privacy.

In addition, we are witnessing an increase in the number of record-breaking fines imposed on companies worldwide for breaching the trust of their customers. For instance:

- In September 2022, Instagram was fined $403 million by Ireland's Data Protection Commissioner (DPC) for violating children's privacy under the GDPR.
- China's ride-hailing conglomerate, Didi Global Inc. (Didi), was fined RMB 8.026 billion (approximately USD 1.18 billion) for violating cybersecurity and data-related laws.
- In the summer of 2021, the financial records of retail giant Amazon disclosed that the Luxembourg authorities had imposed a €746 million ($877 million) fine for GDPR breaches.

Not only is compliance necessary for your business, but the penalties on top of losing customer trust are real dollars extracted through steep fines for breaching the trust and safety of the public.

Complying with privacy regulations is undoubtedly essential for protecting data and upholding privacy standards, yet doing so can limit the full potential and value of generative AI, as these systems rely on large amounts of data to operate effectively. Restricting how LLM providers collect, retain, and process certain types of data could hinder the progress and advancement of LLMs. This also prevents organizations that work with sensitive data from maximizing the benefits of LLMs. Harnessing the power of LLMs while mitigating data risks and ensuring privacy is a critical challenge that all stakeholders must navigate and address.

Confidential computing: A privacy-preserving AI solution

Organizations need a privacy-preserving AI solution that bridges the gap between protecting enterprise data and realizing the full potential of LLMs.

Despite recent advancements, few AI-based applications have successfully been leveraged by organizations to securely operate on confidential and sensitive data. To protect privacy throughout the stages of a generative AI lifecycle, strict data security techniques must be implemented to securely and efficiently perform all security-critical operations that directly touch a model and all confidential data used for training and inference.

Privacy-enhancing technologies

Implementing privacy-enhancing technologies (PETs) can address the challenges associated with data sharing, data breaches, and privacy regulations. PETs enhance privacy by securing the processing of confidential data while enabling a system to perform its intended functionalities and services. Some of the more well-known PETs are described below.

Homomorphic encryption enables computing on data in its encrypted form. The resulting data can be decrypted to output a result that would have been produced if the computations were performed on the original data in its unencrypted form. Although this approach enables third parties to access and operate on user data, it is limited in computational power and functionality. Because it operates directly on encrypted data, its speed is orders of magnitude slower compared to workloads that process plaintext data. And while attackers cannot decrypt the underlying data, they could still alter the data or the computation, thus violating integrity.

Secure multi-party computation (MPC) is a similar approach to homomorphic encryption, but it also enables multiple parties to operate on multiple encrypted datasets while protecting each party's data from one another in addition to outside users and adversaries. Although some MPC protocols can also protect against malicious attackers, its performance for training, fine-tuning, or inferencing LLMs remains orders of magnitude slower than regular computation.

Differential privacy protects user privacy by adding noise (in other words, making a controlled amount of random changes) to the original data, either during data collection or before the output data gets released. This technique helps prevent re-identification of users without significantly impacting the overall analysis of the aggregated data. Because the amount of data leakage is related to how much noise was added to the original data, some differential privacy implementations might not sufficiently protect privacy, while others might substantially impact the

Privacy protection with confidential computing

Among the privacy-enhancing technologies (PETs) available today, confidential computing stands out as the optimal choice for enterprises aiming to secure their GenAI initiatives. It offers robust security without the performance drawbacks of other PETs like homomorphic encryption and secure multi-party computation.

accuracy of the analysis because of the added noise. Additionally, while differential privacy can add noise to data or a proprietary model, it cannot protect the data or the model itself during the computation process, such as during training or inference. This leaves the data and model potentially exposed to attackers. For these reasons, differential privacy is often used in combination with another PET.

While each of the described techniques offer ways to protect sensitive data, none of them alone can ensure the functionality and efficiency that is required of generative AI models. An emerging PET approach, *confidential computing*, can offer a powerful solution by isolating data in a hardware-based trusted execution environment (TEE) as the data is being computed on.[13] Such hardware-based technologies provide secure environments to prevent unauthorized entities—the host operating system (OS), system administrators, service providers, the infrastructure owner, or anyone else with physical access to the hardware—from viewing and changing the data or altering the code within the environment.

Confidential computing technology

Confidential computing is an emerging technology that focuses on protecting data during its use. This concept extends the data protection beyond data at rest and in transit to include data in use, which is particularly relevant in today's computing environment that spans multiple platforms—from on-premises to cloud and edge computing. This technology is crucial for organizations handling sensitive data, such as PII, financial data, or health information, where threats targeting the confidentiality and integrity of data in system memory are a significant concern.

The Confidential Computing Consortium (CCC), a project community at the Linux Foundation, plays a central role in defining and accelerating the adoption of confidential computing. The CCC brings together hardware vendors, cloud providers, and software developers to foster the development of TEE technologies and standards.

This cross-industry effort is essential due to the complex nature of confidential computing, which involves significant hardware changes and how programs, operating systems, and virtual machines are structured. Various projects under the CCC umbrella are advancing the field by developing open-source software and standards, which are crucial for developers working on securing data in use.

Confidential computing can be implemented in different environments, including public clouds, on-premises data centers, and distributed edge locations. This technology is vital for data privacy and security, multi-party analytics, regulatory compliance, data localization, sovereignty, and residency. It ensures that sensitive data remains protected and compliant with local laws, even in a multi-tenant cloud environment.

Your data is for your eyes only

With systems that leverage confidential computing, your sensitive information is visible to no one but you, ensuring unparalleled privacy and security in AI interactions.

How confidential computing works

Two types of hardware-based TEEs are an *enclave* and a *confidential virtual machine (CVM)*, both of which are secure regions on an otherwise untrusted machine. The following security features of TEEs make confidential computing an ideal approach for privacy-preserving generative AI:

- **Remote attestation:** To establish trust with a remote party, a TEE uses an attestation process to cryptographically prove that it was built by an expected entity and is running expected code. The remote party can then establish a secure connection with the TEE and share sensitive information with it.
- **Isolated execution:** Each enclave has access to a restricted subset of a machine's memory, while each CVM has access to a virtual machine's memory. Any data or software placed within the TEE is encrypted and isolated from the rest of the system. The hypervisor and other processes running on the same machine cannot access the encrypted TEE memory.
- **Memory encryption:** Data that exits the TEE is encrypted within the processor by a memory encryption engine (MEE) to ensure that the main memory stores only encrypted TEE data. The host OS or a hacker outside of the TEE can see only encrypted data in main memory. When encrypted data returns to the TEE from main memory, the MEE decrypts the data to enable the central processing unit (CPU) to process unencrypted data, resulting in high speed for computation—a distinct contrast to purely cryptographic computational techniques such as homomorphic encryption and secure MPC.

Figure 2 depicts the use of two types of TEE technologies: Intel SGX and AMD Secure Encrypted Virtualization-Secure Nested Paging (SEV-SNP).

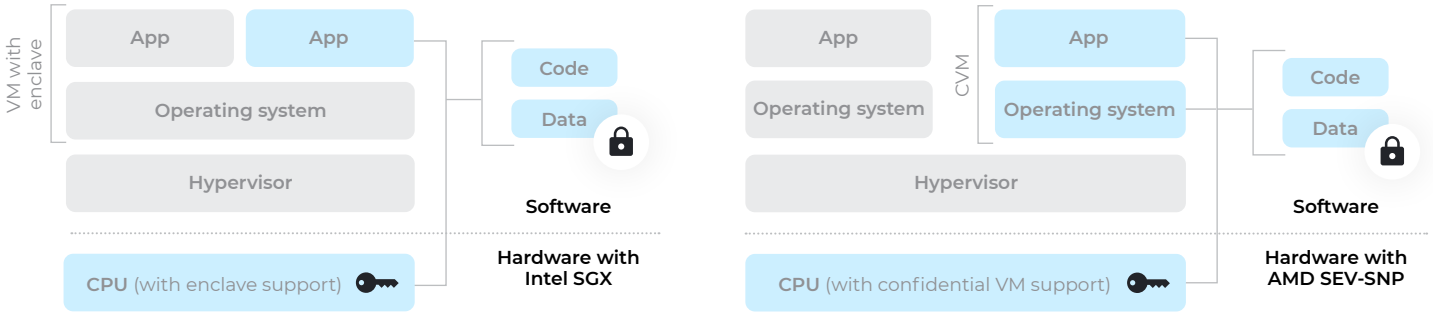


Figure 2: *TEE technologies offer different levels of security and usability trade-offs. The code and data in use within a TEE are protected from other software on the machine. Data is decrypted only within that TEE, and is always encrypted in main memory.*

Securing data during LLM fine-tuning

Opaque enables privacy-preserving fine-tuning of LLMs on encrypted data, ensuring that sensitive training data remains confidential and inaccessible to other parties. Data is encrypted before being uploaded to the training environment and decrypted only within an attested TEE, maintaining data confidentiality throughout the model fine-tuning process.

Confidential computing for trusted AI

A trusted AI solution uses hardware-based TEEs to enable the secure training and operation of models on sensitive data. Training, fine-tuning, and inferencing can all take place within TEEs, enabling multiple parties to collaborate while assuring that their sensitive data and proprietary models remain protected, even from each other. Data owners and users can leverage LLMs on their data without revealing any confidential information to model providers and other unauthorized parties. Likewise, model owners can train their models while protecting the training data and the architecture and parameters of their models. If a data breach were to occur, hackers can see only encrypted data and not the sensitive data protected within TEEs.

Confidential computing alone does not resolve the issue of models potentially memorizing and accidentally revealing details about the data that they've been trained on. A way to reduce this risk is combining confidential computing technology with differential privacy.[14] Using this strategy, data is computed on within TEEs and then applied with a differential privacy update (prior to getting released) to reduce the risk of leakage from inferencing.

A trusted AI platform also enables LLM providers and data providers to comply with privacy laws and regulations. By protecting all sensitive and proprietary data with advanced encryption and secure TEE technology, model builders and providers have less concerns surrounding the amount and the type of user data that they can collect.

Opaque's trusted AI solutions

Opaque's suite of privacy-preserving solutions enable organizations to safely and securely use generative AI to protect confidential data.

The Opaque platform makes confidential data useful by enabling secure machine learning on encrypted data within TEEs. The platform leverages a cluster of CVMs powered by CPUs featuring confidential computing capabilities. The platform manages cluster scalability and ensures high availability, providing a secure and stable infrastructure for AI applications to run on top of.

Secure fine-tuning

Opaque's confidential ML solution facilitates privacy-preserving model fine-tuning, enabling model providers to train LLMs on confidential data. This approach supports fine-tuning models without compromising the privacy of the training data, enhancing collaboration between LLM providers and data providers.

Protecting user queries

Opaque's confidential AI solution delivers a trusted environment for inferencing, where users can interact with LLMs without the LLM provider, Opaque, or other entities accessing their sensitive data.

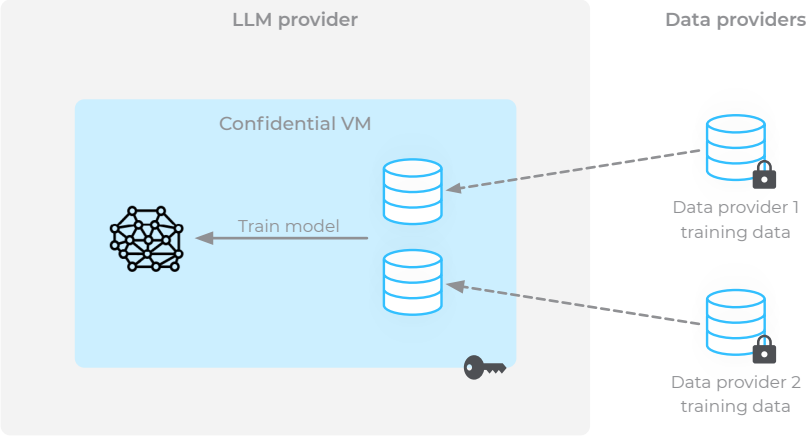


Figure 3: *Confidential ML enables LLM providers to fine-tune their models without exposing the sensitive training data.*

To ensure that all training data remains confidential, the data is encrypted before being uploaded to the Opaque platform and, as shown in Figure 3, is decrypted only within an attested TEE. The process works as follows:

1. The confidential computing environment is initialized and attested.
2. One or more data providers send encrypted data to the TEE.
3. Within the TEE, the LLM undergoes fine-tuning using the encrypted data.

Each data provider's data remains inaccessible to the LLM provider, other data providers, or any external entities during this process, enabling secure collaboration between model owners and data providers.

Secure inferencing

Opaque's model-serving application performs secure inference on confidential data, providing a solution for users who need to include private information in their prompts without exposing it to the model provider. Inferencing takes place in a secure environment where users can interact with the LLM while knowing that their confidential data is safeguarded against unauthorized access and potential breaches.

The application consists of a pretrained LLM and enables collaboration between the model owner and a user providing the prompt and context data. The platform exposes a set of application programming interfaces (APIs), allowing developers to build confidential LLM applications.

To ensure authenticity, the model must be attested before it can be used. The model is guaranteed to not change after attestation, ensuring that no malicious party can tamper with it. The input, which consists of prompt and contextual information, is encrypted before being uploaded to the Opaque platform, thus

Guarding user prompts

Opaque's LLM Gateway includes mechanisms to filter and sanitize user inputs prior to submitting them to an external LLM service, reducing the risk of sensitive data being inadvertently included in user queries.

guaranteeing its confidentiality. It only gets decrypted within the TEE, which is secured by Opaque's confidential computing technology.

The application works as follows:

1. The confidential computing environment is initialized.
2. The model owner loads and attests the model.
3. A policy manager can provide policies to be applied to the model and input. For example, a policy can filter certain types of data.
4. The end user provides the context and the input prompt or question to ask the generative AI model.
5. The application runs inference and provides the answer.

Steps 4 and 5 can be repeated as needed. The user query, context, and results all remain private throughout the process, enabling privacy-preserving collaboration between model owners, context providers, and end users.

LLM Gateway

In the race to launch LLM solutions to production, enterprises must implement innovative solutions to leverage the vast potential of LLMs while safeguarding their proprietary and confidential data. Opaque's LLM Gateway offers a suite of features designed to enhance data privacy, reduce operational costs, and provide valuable insights into usage patterns. This gateway acts as a security layer between enterprises and external LLM services, enabling enterprises to maintain confidentiality of their data while using any LLM.

The LLM Gateway service runs on top of a CVM cluster and can be integrated with existing LLM ecosystems. It dynamically sanitizes personal information in user prompts to prevent the exposure of sensitive data. Additionally, it provides prompt-level compression to optimize data transmission and minimize costs, further complementing its privacy-preserving capabilities. Beyond these functionalities, LLM Gateway equips organizations with monitoring and reporting services, delivering critical insights into prompt usage and interactions with LLMs.

Input filtering and sanitization

Enterprises possess a wealth of proprietary and confidential data, which they seek to analyze using external LLMs. However, ensuring this data's confidentiality poses a significant challenge during interactions with the external LLM services. Opaque's privacy-preserving prompting tool is designed to address this very concern.

Our LLM Gateway, as shown in Figure 4, acts as a protective layer between your enterprise and the LLM service, dynamically sanitizing sensitive and confidential information in user prompts. This means that your organization can continue to use LLMs to generate valuable insights from private data, without exposing confidential information to the LLM service or even Opaque.

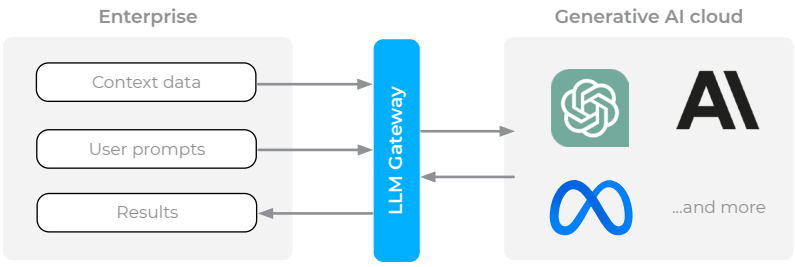


Figure 4: *LLM Gateway enables organizations to maintain confidentiality of their data while using any LLM.*

With LLM Gateway, developers can build LLM applications for users that want to query LLMs with prompts that might contain sensitive information. The platform exposes an API that sanitizes a prompt by encrypting and redacting all personal information, and then desanitizes the LLM-provided output to restore any previously encrypted personal information in the response returned to the user. Opaque cannot view or access the user prompts and underlying data, and third-party LLM providers can view only the redacted data.

Figures 5 and 6 show how sanitization and desanitization both occur within an attested TEE to ensure privacy of sensitive user data. The sanitization/desanitization process can be automated so that end users interact seamlessly with LLM applications without having to manually sanitize their prompts before submitting them to the LLM application. This process works as follows:

1. The user submits a query to the LLM application.
2. The LLM application augments the user's query with context, and then passes the prompt to LLM Gateway.
    a. Using retrieval-augmented generation (RAG), the application fetches relevant data from an external database and packages the data with the user's query.
    b. The application encrypts the augmented prompt and sends it to LLM Gateway.
3. The LLM Gateway API identifies and encrypts sensitive data, and produces a sanitized version of the prompt.
4. A third-party LLM API receives the sanitized prompt.
5. The LLM API returns an output based on the sanitized prompt.
6. The LLM Gateway API desanitizes the output and returns the result, with sensitive information restored, to the LLM application.
7. The LLM application returns the desanitized response to the user.

Any PII data in the user prompt remains private and never gets shared with the LLM provider, alleviating concerns around the LLM provider's collection and retention of sensitive and personal data.

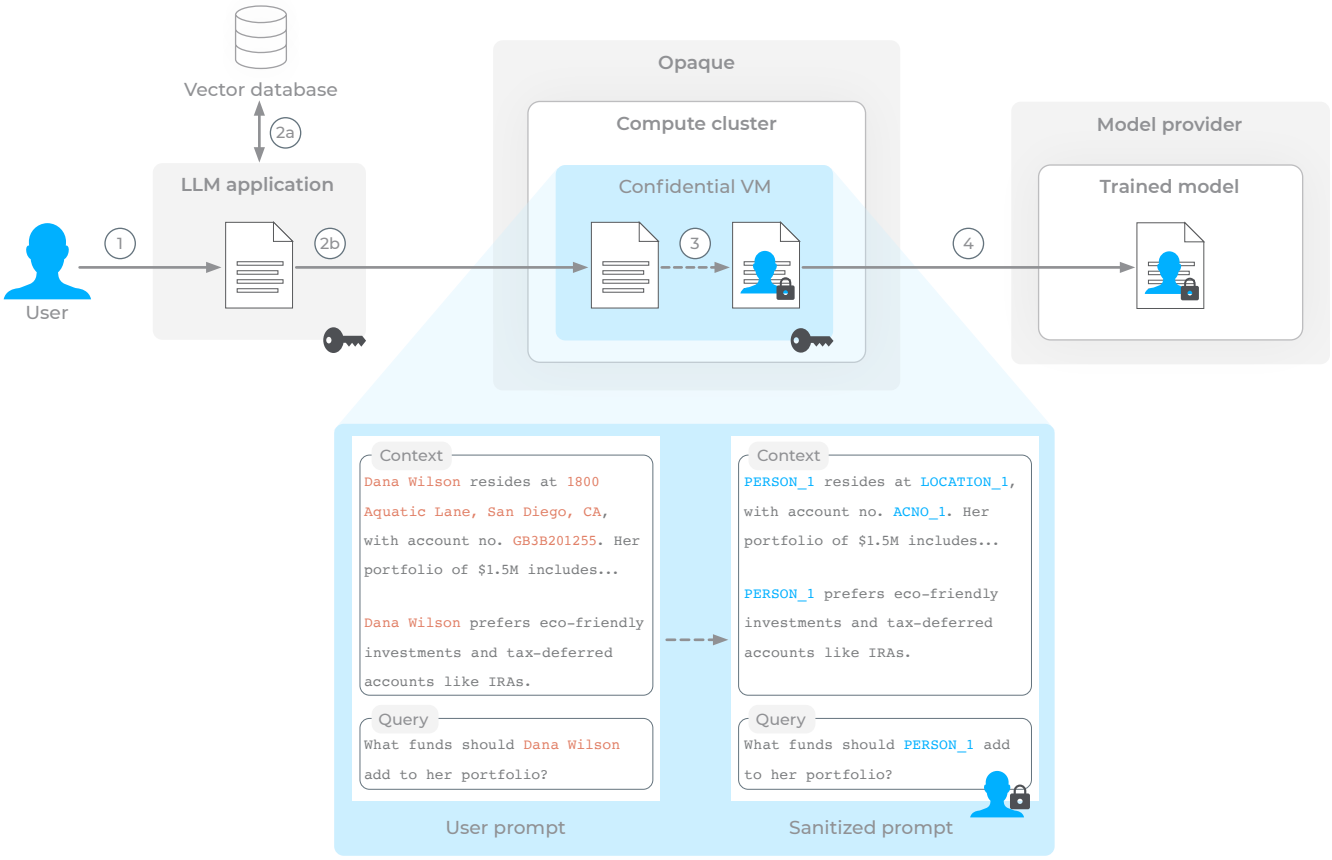


Figure 5: *To protect user data during LLM inference, sensitive information in prompts and queries are encrypted and redacted within an attested TEE. The user's sensitive data is never accessible to other entities, including Opaque and the model provider.*

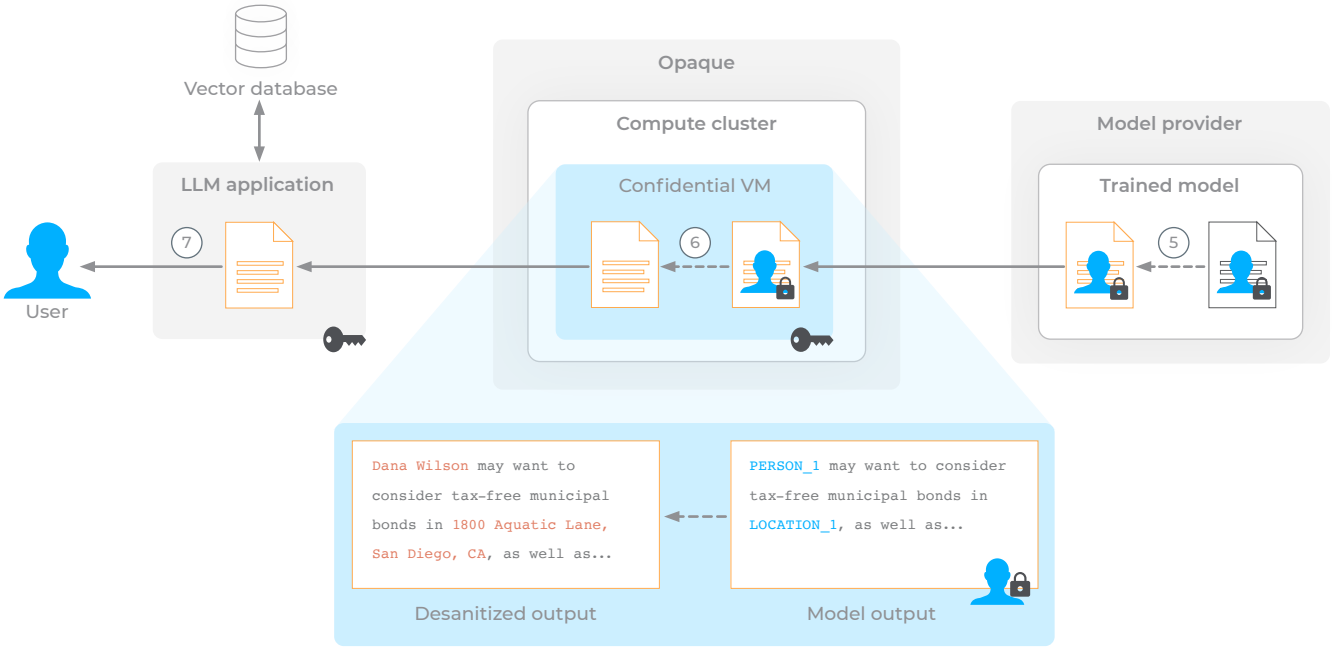


Figure 6: *Prior to returning the LLM-generated output to the user, the output passes through the TEE, in which de-identified data is restored to its original plaintext data.*

Comprehensive insights on prompts

LLM Gateway provides monitoring and reporting tools for enterprises to gain insights into their prompts and usage patterns.

Monitoring and reporting

Opaque's LLM Gateway provides monitoring and reporting capabilities to ensure the effective and responsible use of LLMs. Without insights into their prompts, enterprises may struggle to optimize LLM performance and resource usage. This could lead to higher costs and reduced effectiveness of their LLM implementations. In addition, a lack of insight into the types of information contained in prompts could potentially expose companies to a range of security threats, including adversarial attacks, privacy violations, and data leakage.

LLM Gateway features comprehensive monitoring and logging capabilities for businesses to gain insights into their prompts at an aggregate level. The tool provides advanced ML techniques to monitor prompts, log key usage metrics, and generate detailed reports and dashboards aimed at LLM optimization. These reporting capabilities are available to all Opaque users, requiring no additional engineering effort on their part.

For instance, businesses can pinpoint and address potential PII and proprietary data leaks based on reports that identify whether prompts contain sensitive information. This proactive measure enables them to reinforce their data protection strategies and optimize their LLM implementations. Based on their findings, businesses can apply custom filters to prompts, in addition to utilizing the out-of-the-box PII filtering and sanitization capabilities provided by LLM Gateway. The tool also provides insights into the dollars saved and the amount of tokens processed by LLMs as a result of applying prompt-level compression.

Data compression for efficiency

LLM Gateway incorporates data compression techniques to optimize performance and reduce costs for enterprises that need to apply analytics and machine learning to large amounts of data.

Compression

LLM Gateway's prompt compression capabilities enable enterprises to reduce the costs of their LLM usage, which are often driven by large token payloads in LLM API calls. The tool condenses prompts and contexts to make them more manageable for models to process. By compressing the input data, LLMs can handle more complex tasks with greater speed, reduced computational costs, and little performance loss, resulting in substantial cost savings for businesses.

Our prompt optimization tool compresses prompts to a fraction of their original size, reducing the number of tokens in LLM calls without losing the essence of the information conveyed. This compression not only preserves the semantic integrity of the prompts but also enhances efficiency by reducing the computational load on models. This means enterprises can achieve comparable results with significantly less resource expenditure.

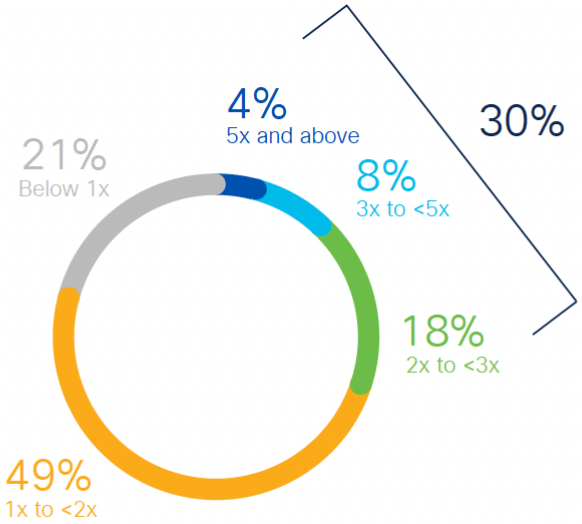
The benefits of prompt compression extend beyond just cost savings. Our compression tool accelerates inference, increasing the responsiveness of LLM applications. It also improves privacy by reducing the amount of sensitive information processed and exposed to models.

Trusted AI: The path forward

Comparing spending and benefits together, privacy remains a very attractive financial investment for most organizations.

The versatility and sophistication of generative AI and LLMs have positioned them as powerful tools across industries. The capabilities of AI and the adoption of LLM systems will only continue to increase, adding to the critical need for securing data and preserving privacy. The future of modern organizations that run on data is machine driven, and implementing a privacy-preserving AI solution is essential for all involved parties to trust that the privacy of their confidential data is protected.

The average organization reports getting privacy benefits of 1.6 times their investment.[1] In addition, as shown in Figure 7, 30% of organizations estimate returns at least two times, with some (12%) realizing returns upwards of three times their investment.



Source: Cisco 2024 Data Privacy Benchmark Study

Figure 7: *Estimated ROI ranges indicate that privacy is an attractive financial investment for most organizations.*[1]

Confidential computing provides organizations the ability to use AI and LLMs in a privacy-preserving manner. Opaque's trusted AI platform enables businesses to keep their data encrypted throughout its lifecycle, from model training and fine-tuning to inference. Data is kept confidential at rest, in transit, and in use, significantly reducing the likelihood of loss and guaranteeing the preservation of privacy.

Privacy for data and AI is the cornerstone of Opaque. Book a demo to discuss how we can help your organization adopt privacy-preserving technology.

References

[1] Cisco 2024 Data Privacy Benchmark Study. Privacy as an Enabler of Customer Trust. https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-privacy-benchmark-study-2024.pdf.

[2] Gurman, M. "Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak." Bloomberg. https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[3] OpenAI. "March 20 ChatGPT outage: Here's what happened." https://openai.com/blog/march-20-chatgpt-outage.

[4] Reuters. "OpenAI's ChatGPT breaches privacy rules, says Italian watchdog." https://www.reuters.com/technology/cybersecurity/italy-regulator-notifies-openai-privacy-breaches-chatgpt-2024-01-29.

[5] Artificial Intelligence Policy Institute (AIPI). "Poll Shows Overwhelming Concern About Risks From AI as New Institute Launches to Understand Public Opinion and Advocate for Responsible AI Policies." https://theaipi.org/poll-shows-overwhelming-concern-about-risks-from-ai-as-new-institute-launches-to-understand-public-opinion-and-advocate-for-responsible-ai-policies/.

[6] Gillespie, N., Lockey, S., Curtis, C., Pool, J., and Akbari, A. Trust in Artificial Intelligence: A Global Study. The University of Queensland and KPMG Australia. doi:10.14264/00d3c94. https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf.

[7] Code42. Annual Data Exposure Report 2023. https://www.code42.com/resources/reports/2023-data-exposure.

[8] Rocher, L., Hendrickx, J.M., and de Montjoye, YA. "Estimating the success of re-identifications in incomplete datasets using generative models." *Nature Communications* 10, no. 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3.

[9] Morris, J.X., Kuleshov, V., Shmatikov, V., and Rush, A.M. "Text Embeddings Reveal (Almost) As Much As Text." https://arxiv.org/pdf/2310.06816.pdf.

[10] Wu, T. "California Seeks to Be First to Regulate Business Use of AI." Bloomberg Law. https://news.bloomberglaw.com/in-house-counsel/california-seeks-to-be-first-to-regulate-business-use-of-ai.

[11] Innovation, Science and Economic Development Canada. "The Artificial Intelligence and Data Act (AIDA) – Companion document." https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

[12] European Parliament. "Texts adopted - Artificial Intelligence Act - Wednesday, 14 June 2023." https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

[13] Confidential Computing Consortium. "A Technical Analysis of Confidential Computing." https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf.

[14] Majmudar, J., Dupuy, C., Peris, C., Smaili, S., Gupta, R., and Zemel, R. "Differentially Private Decoding in Large Language Models." *ArXiv preprint arXiv:2205.13621* (2022). https://arxiv.org/pdf/2205.13621.pdf.