Confidencial

FEB 2024

Navigating the AI Gold Rush

Protecting sensitive data used for and by LLMs

Karim Eldefrawy, PhD
Co-Founder & CTO

How secure is your data amid the AI Gold Rush?

If you’re asking yourself that question, you’re not alone.

When training Large Language Models (LLMs), organizations face the dilemma of either (i) not using unstructured data with unprotected sensitive information, thereby impairing the effectiveness of the LLM by missing out on 80% of their training data; or (ii) using unstructured data to train LLMs, but (if not careful) potentially exposing unwanted sensitive information and violating multiple privacy laws in the process.

CTOs, CIOs, DPOs, CISOs in Fortune 5000 companies are trying to crack the code on how to securely leverage the enterprise data to rain new models and assist in their operations to boost productivity and automate tasks, while ensuring appropriate controls and protections of sensitive data in their enterprise stores (whether cloud-based or on-prem). We focus on LLM as a proxy for other machine learning (ML) techniques as they are the most commercially interesting at this moment, and a good representative of some of the most recent advances in ML in general.

There’s pressing need for a principled, well-thought out, and verifiably secure and compliant solution that is independent of the frameworks and platforms used for the training and execution of such LLM models, especially if they will be given access to unstructured content that may contain sensitive information, e.g., as part of Retrieval Augmented Generation (RAG) workflows.

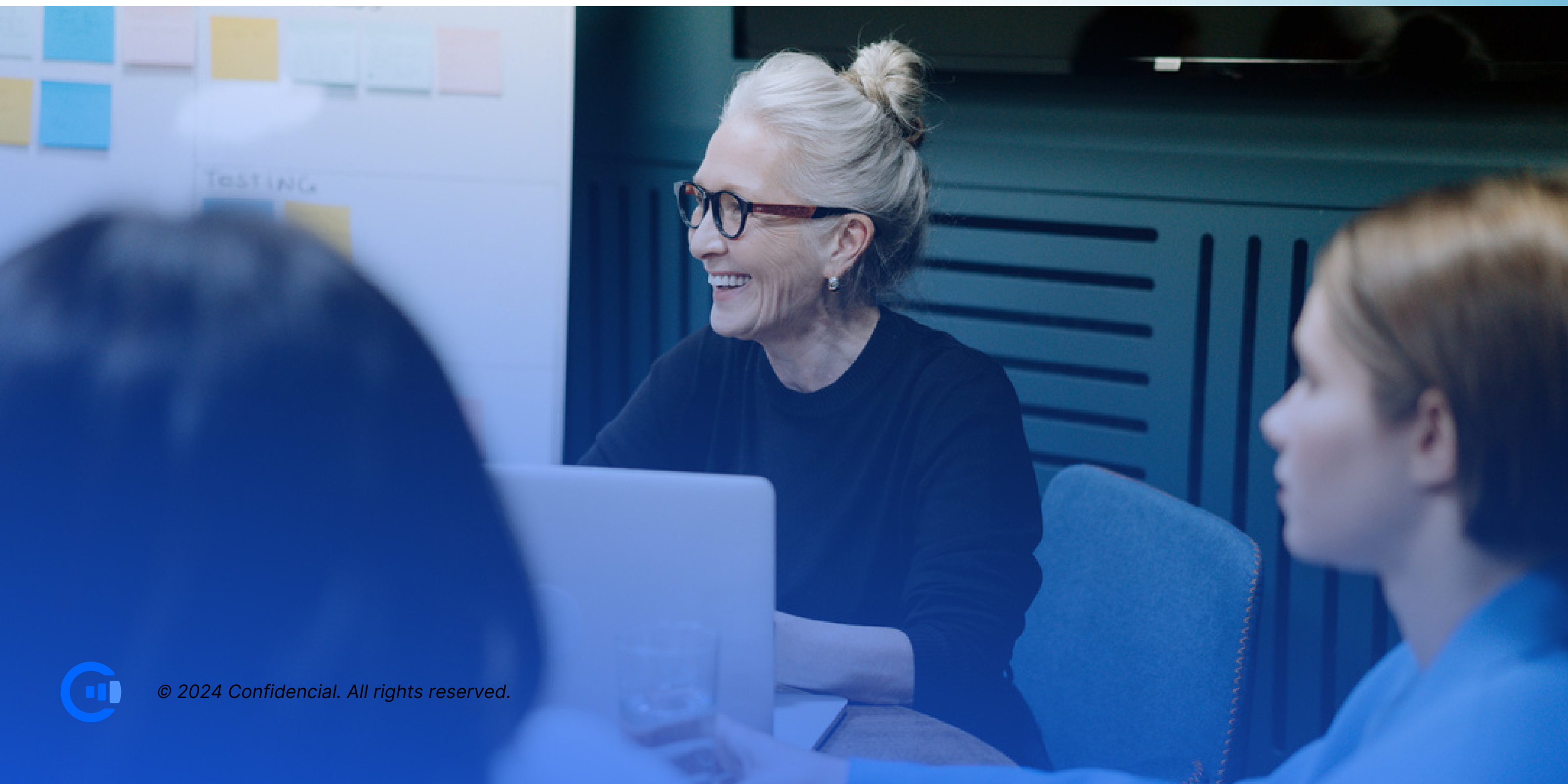
In this whitepaper, we’ll explore the advances in ML, with particular attention to LLMs and break down everything you need to know about how to optimize risk and value creation by protecting sensitive information in unstructured data when used in training and operations of LLMs and ML.

Who we are

Confidencial is built on the principle of data-centric security. We believe that true protection means ensuring data safety at every stage – from document creation to sharing, storage, and beyond. By making unstructured data security automatic and intuitive, we remove the complexities often associated with data protection, allowing our clients to focus on what they do best.

What we do

At Confidencial, your trust is our most valuable currency. We are committed to upholding the highest standards of data protection, ensuring that your information remains confidential and secure, always. Our flexible, adaptive approach ensures that your data remains secure against both current and emerging threats. Join us in our journey towards a more secure digital future.



Unignorable Advances Create Unignorable Challenges

> The Large Language Model (LLM) Market was valued at $10.5 Billion USD in 2022 and is anticipated to reach $40.8 Billion USD by 2029, witnessing a CAGR of 21.4% during the forecast period 2023-2029.
>
> - Valuates Reports

There's no doubt that LLMs are among the top technologies being explored by enterprises this year and into the future. The unique abilities of LLMs to understand, predict, and generate human language can't be understated. Contrary to traditional machine learning, LLMs can actually comprehend and interact with human language data with unprecedented sophistication, catalyzing innovation and leading to massive productivity gains.

> **However, in a recent McKinsey report, 79% have been exposed to AI, yet only 21% of organizations have established policies governing employees' use.**

This introduces unchartered risks to businesses including privacy concerns, which could arise in the event a user inputs information that later ends up in model outputs that make an individual(s) identifiable, as well as the leakage of proprietary or intellectual property. (McKinsley)

Understanding the LLM Lifecycle

To fully understand the risk and opportunity, you must understand the LLM lifecycle. Let’s dig in:

STEP 1

Data Source Integration

The process starts by linking the language model's training environment to a rich database of unstructured content. This extensive library, brimming with texts from diverse domains, is essential. It equips the model with a broad spectrum of linguistic structures and concepts, paving the way for nuanced language comprehension.

STEP 2

Data Assimilations

After establishing a connection, the next phase involves data assimilation. Here, the raw data from our extensive repository is systematically fed into the machine learning framework. Efficient data handling is crucial at this stage to accommodate the large-scale informational needs of language models.

STEP 3

Data Refinement

This step transforms raw data into a refined format suitable for model training. It entails purging superfluous or repetitive data and validating the remaining dataset against high-quality standards. Through iterative review, we identify the most instructive data, ensuring the model's training is both comprehensive and balanced.

STEP 4

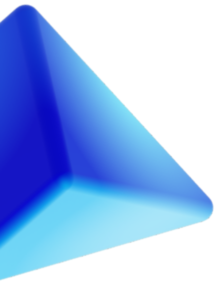
Model Training

With the data curated, we commence the training of the model. This stage is computationally demanding, as the model discerns and internalizes patterns within the data. This rigorous process is the cornerstone of the model's ability to later generate or comprehend text with precision.

STEP 5

Model Evaluation

The model's performance is then scrutinized through a series of evaluations and validations against a separate dataset. This phase is critical for gauging the model's proficiency in language generation and comprehension, highlighting both its strengths and areas needing improvement.



STEP 6

Model Packaging

Following successful training and validation, the model is prepped for deployment. This involves optimizing the model for real-world application efficiency and packaging it in a deployable format.

STEP 7

Model Deployment

This step integrates the packaged model into a live environment, where it begins to add real value. Deployment is executed with precision to ensure stability and consistent performance under operational conditions.

STEP 8

Evaluation & Improvement

Post-deployment, ongoing monitoring of the model's output is imperative. This ensures that the model meets performance benchmarks, adapts to any shifts in data trends or user interaction patterns, and continues to improve by learning from new information.

Figure 1: LLM Training, Operation, and Data Protection

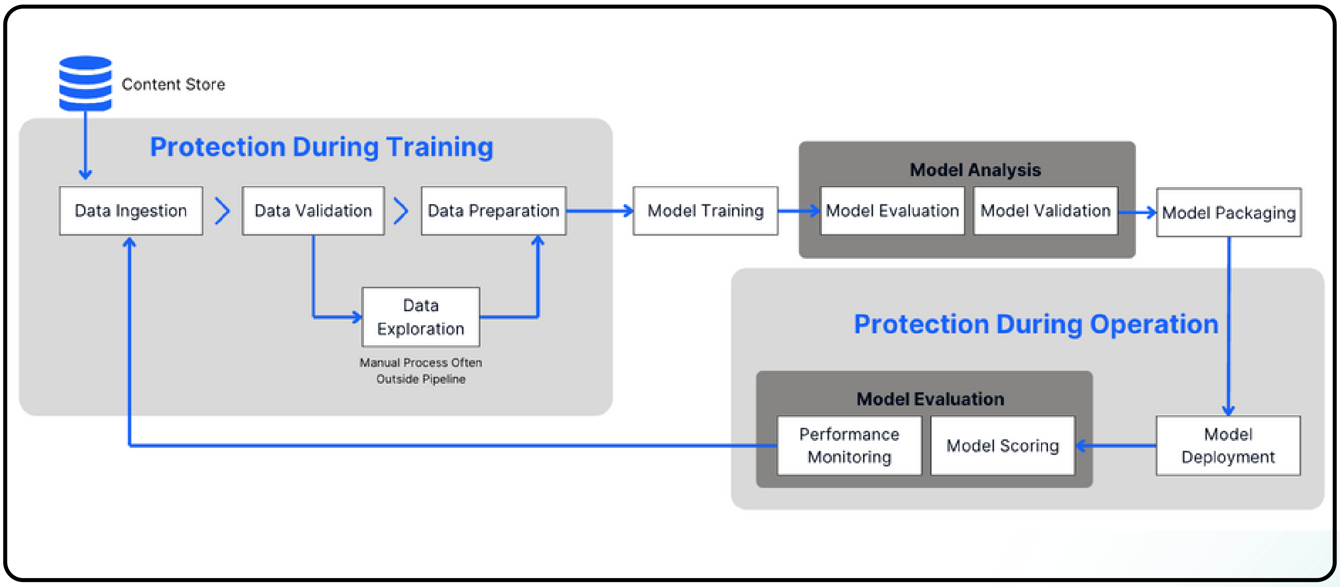


Figure 1 shows the typical steps and operations involved in LLM training and operation. When unstructured data is used, we must protect sensitive data found in unstructured content and documents.

Where is The Sensitive Data That Needs Protection?

In the construction and operation of LLMs, the use of unstructured data—spanning from documents and emails to instant messages—presents significant challenges and responsibilities. It's imperative that sensitive information, be it personally identifiable information (PII), protected health information (PHI), payment card information (PCI), intellectual property (IP), or trade secrets, is meticulously excluded in a demonstrable and verifiable manner.

This safeguarding is critical not only for adhering to stringent privacy laws and regulations but also for upholding the confidentiality and privacy mandates of enterprises.

Data Protection Requirements

To comply with privacy laws, ensure confidentiality, and meet data protection standards during LLM training, let's examine how the National Institute of Standards and Requirements (NIST) defines key protocols.

Data Protection and Compliance Mandates

Sensitive and private data must be demonstrably shielded throughout both the training and operational deployment of ML/LLMs, ensuring such data remains inaccessible.

Control mechanisms for this protection must reside with the data steward, aligning with frameworks like the General Data Protection Regulation (GDPR).

Data Training Protocols

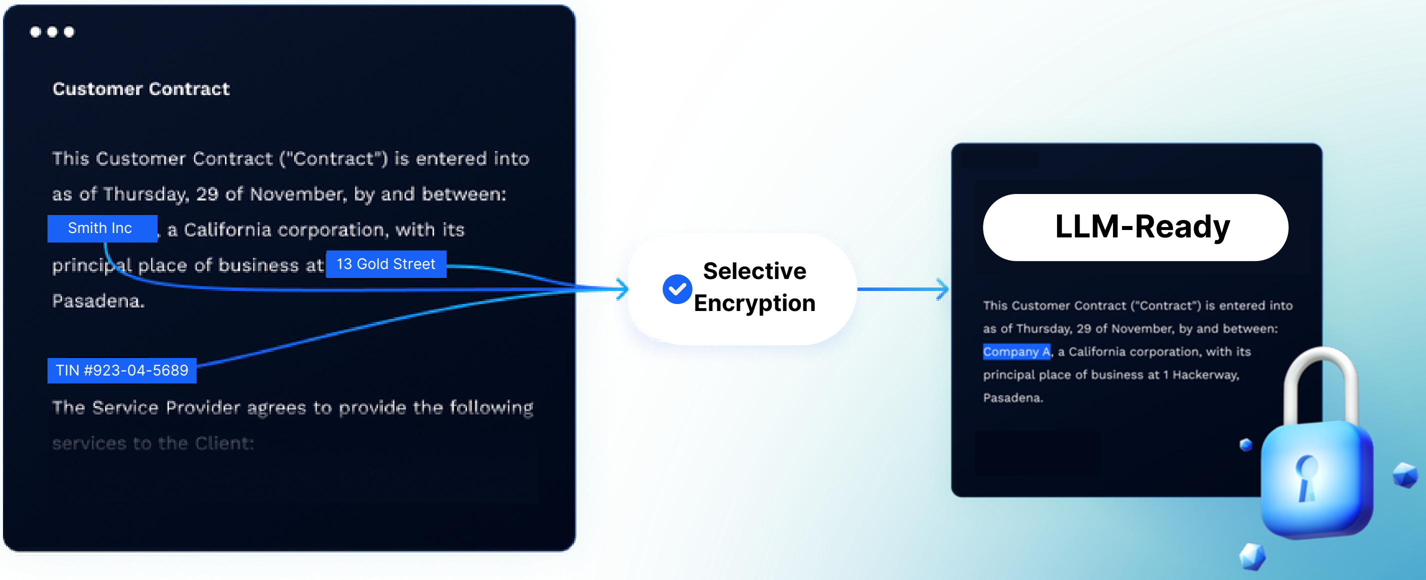
Data must remain accessible in a format that is intelligible to ML/LLM training algorithms, to facilitate effective learning.

Consistency in the representation of data must be maintained, meaning that data elements such as social security numbers should be uniformly obfuscated across all instances. This necessitates the use of techniques like tokenization and anonymization to disguise the actual values while preserving the data's structural integrity.

Using Selective Encryption to Protect Sensitive Information

A practical method we propose for safeguarding data involves the automated fine-grained use of encryption. This approach is secure when properly implemented and aligns with established NIST standards, making it straightforward to audit, automate, and adapt. Historically, encryption has effectively secured data when stored or sent across networks and, more recently, when used during processing.

Recent advances have led to the development of automated 'Selective Encryption,' allowing organizations and users to control and choose the extent of encryption down to specific words or paragraphs within documents and other unstructured data. Using encryption surgically—rather than bluntly encrypting entire files— to selectively shield sensitive or proprietary content within documents before being used in LLM training and operations, whether conducted on-premises or in the cloud, provides a definitive assurance that such data remains unused by LLMs during their training and operations.



Customer Contract

This Customer Contract ("Contract") is entered into as of Thursday, 29 of November, by and between: Smith Inc, a California corporation, with its principal place of business at 13 Gold Street Pasadena.

TIN #923-04-5689

The Service Provider agrees to provide the following services to the Client:

This Customer Contract ("Contract") is entered into as of Thursday, 29 of November, by and between: Company A, a California corporation, with its principal place of business at 1 Hackerway, Pasadena.

The Business Case for Selective Encryption

In addition to safely harnessing unstructured data for LLM use, there are many benefits to adopting Confidencial's selective encryption:

It's a flexible and configurable software solution, enabling businesses to fully leverage their data while demonstrating to regulators the steps taken to protect sensitive information.

Using AI to automatically scan and secure sensitive content as soon as it's created* eliminates the need for manual intervention by end-users, providing a streamlined, effective solution to close the security gaps left by current methods.

Encryption-based protection is compatible with various machine learning platforms and frameworks, offering a consistent and straightforward defense mechanism across different operational models.

It enables organizations to surgically encrypt sensitive data once and train multiple LLMs with varying levels of access to the sensitive data.

Selective encryption optimizes version control, allowing for a single version of a document that can be reused as is for multiple purposes without increasing storage or requiring a complex, costly detection and redaction pipeline.

Other Safeguarding Strategies

When considering the safeguarding of data during the ingestion, sanitization, and operation of LLM operations, the following alternative strategies are commonly evaluated by organizations, but their limitations emerge:

Internal Sanitation Measures

This requires a deep understanding of LLMs. It can be resource-intensive, necessitating ongoing updates and maintenance. An additional system may be needed for secure LLM access to documents during operation, especially if solutions like redaction, anonymization, and generalization (RAG) are applied.

Specialized Vendor Engagement

This option hinges on establishing trust with an external provider.
It may involve potential data exposure to the vendor, unless the solution is deployed on-premises. A separate mechanism might be needed for LLM document access in operational stages.

Built-in Sanitation

There's the risk of data exposure to the vendor without an on-premises arrangement. Such solutions are often specific to the chosen framework and may lack flexibility. Yet again, additional solutions may be required for document access during LLM operations.

Our Approach for Data Protection During Ingestion, Sanitization, and Operation of LLM

Security-aware and compliance-driven executives and teams require a singular, cost-effective solution that offers verifiable security, is auditable, and operates independently of vendors. Confidencial ensures that sensitive content within unstructured data is never compromised during the training or operational use of LLMs and other machine learning platforms.

Figure 2: How Confidencial Protects Sensitive Data

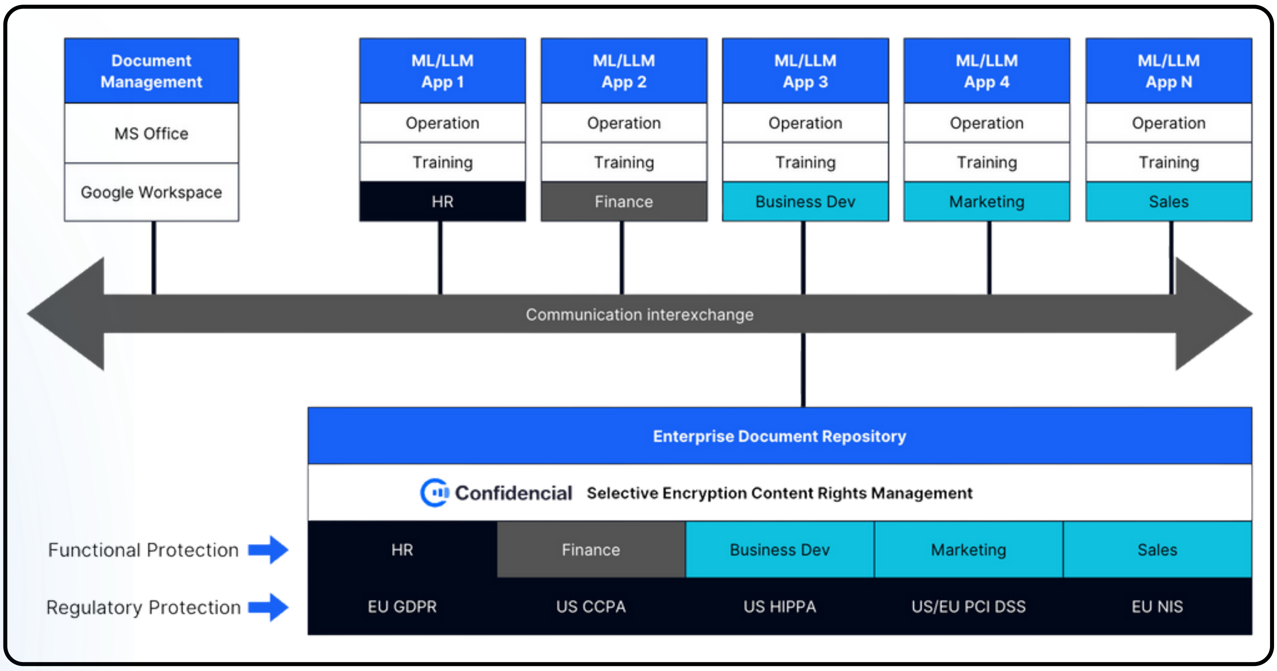


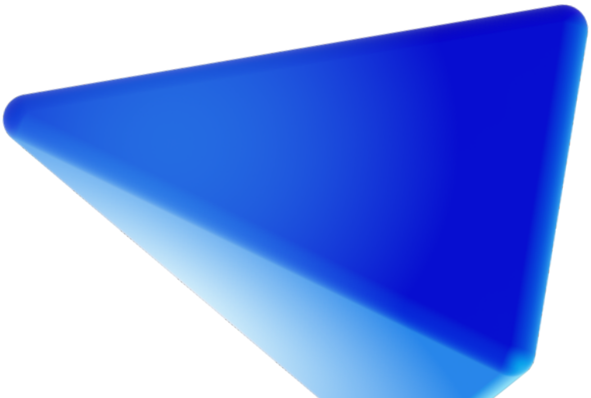
Figure 2: Confidencial maintains a singular version of each document while enabling differentiated access rights tailored to the specific needs of the end-user or group. As documents are retrieved from a centralized store for various LLM applications, the system's integrated in-document cryptographic access controls will automatically restrict usage, ensuring that each application can only interact with the portions of the documents for which it has been granted permission. This approach simplifies data management and enhances protection by embedding protection directly into the data container or documents.

The Role of Confidential Computing and Privacy-Enhancing Technologies

Privacy-enhancing technologies (PETs), such as confidential computing through Trusted Execution Environments (TEEs), Secure Multi-Party Computation (MPC), and Fully Homomorphic Encryption (FHE), are increasingly pivotal in the realm of ML and LLMs. These advanced cryptographic methods are setting the stage for a new era of secure data sharing, training, and operation of ML/LLMs.

Trusted Execution Environments offer a secure enclave within processors for code and data, maintaining integrity even if the broader system is compromised. For LLMs, this means that sensitive data and the model itself can be protected during training and use. In sectors where data sensitivity is critical, like healthcare and finance, TEEs offer a strong layer of security. They also help protect the LLM's intellectual property, keeping the model's parameters and design safe. However, while TEEs secure the execution, they do not address the risk of sensitive data being included in the training set inadvertently.

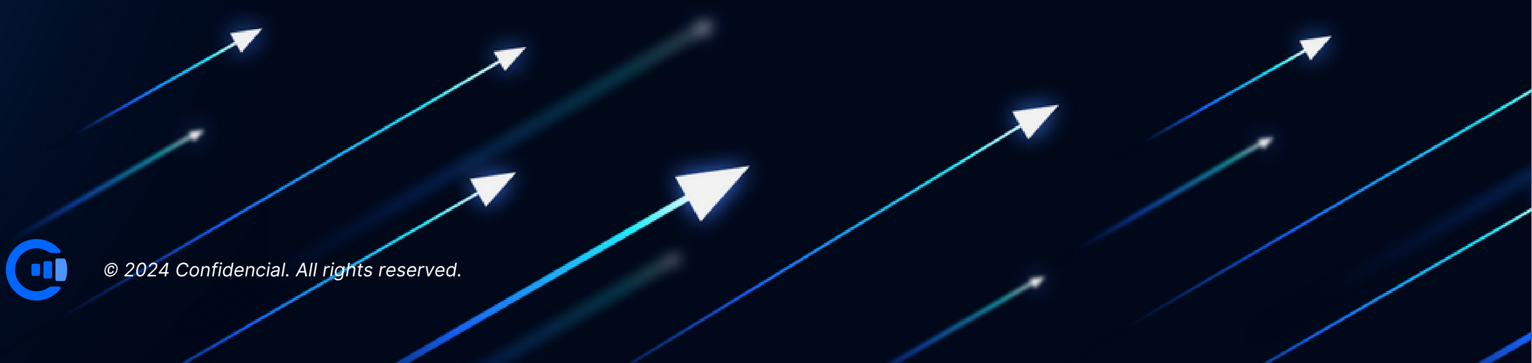
Secure Multi-Party Computation, on the other hand, allows multiple parties to compute a function over their inputs without revealing them to each other. This is incredibly useful when combining the scalability of public clouds with the need to keep enterprise data private. Like TEEs, though, MPC does not inherently filter out sensitive data from training sets, necessitating additional safeguards.



Fully Homomorphic Encryption allows computations on encrypted data, ideal for using untrusted cloud servers while keeping data private. Despite its strong security prospects, FHE is currently limited by its high computational overhead. Hardware accelerators on the horizon could reduce this overhead, potentially making FHE a more viable option for LLMs in the future.

Federated Learning and **Differential Privacy** bring decentralization and randomness to machine learning. Federated Learning trains models across multiple devices without centralizing data, which is ideal for maintaining privacy. Differential Privacy adds noise to data or algorithms to prevent attackers from identifying individual data points. Both are great for protecting user data but might not fully conceal sensitive patterns or information that an organization wants to keep hidden from the training software.

As we look ahead, it's also crucial to anticipate the shift to Post-Quantum Cryptography (PQC). The advent of quantum computing poses a significant threat to current encryption methods. Enterprises must prepare for this by auditing their cryptographic infrastructure and preparing for a smooth transition to PQC, as standards emerge. This forward-looking approach is essential for safeguarding data against future threats and maintaining trust in digital ecosystems as we move into the quantum computing era.



Challenges Facing Secure and Compliant Adoption of LLMs in the Enterprise

Integrating Large Language Models into the corporate landscape requires navigating a complex matrix of security, privacy, and compliance considerations. Here are the pivotal challenges organizations must address:

Data Privacy and Protection

Training LLMs with extensive datasets brings inherent risks of sensitive data exposure. Companies must balance the stringent requirements of data protection laws, like Europe's GDPR, with the efficacy of LLMs, ensuring privacy without compromising performance.

Security Vulnerabilities

LLMs, much like any advanced technology, are potential targets for cyber threats, including adversarial attacks aimed at manipulating model behavior. LLMs' intricate and often opaque nature amplifies the difficulty of safeguarding against such threats. To counteract these risks, businesses must deploy comprehensive security strategies, including real-time monitoring and enhanced training protocols.

Compliance and Ethical Use

The deployment of LLMs extends beyond adhering to privacy laws, demanding adherence to a wider spectrum of regulatory and ethical standards. This is crucial to prevent biases or discrimination in model outputs, necessitating vigilant oversight and continuous recalibration to align with ethical AI practices.

LLMs hold tremendous promise for enhancing enterprise capabilities, but unlocking their full potential demands a proactive stance on these challenges. Organizations must integrate best practices in data governance, ethical standards, and AI security to transform these challenges into opportunities.

That’s where we come in.

Confidencial's Strategic Edge

Selective Encryption

Confidencial automates selective, fine-grained encryption, ensuring data remains secure throughout LLM training and operations.

Verifiable Security

The platforms security is demonstrably robust, complete with compliance logs to back up its secure operations during LLM training.

Cost Efficiency

Confidencial minimizes expenses and simplifies the complexities traditionally associated with data preparation and sanitization.

Universal Integration

Our approach harmonizes with various LLM training frameworks and toolchains, enhancing existing data protection methods.

Future-Proofing

The platform is designed to seamlessly transition to NIST’s forthcoming Post-Quantum Cryptography (PQC) standards, keeping you ahead in data security for years to come.

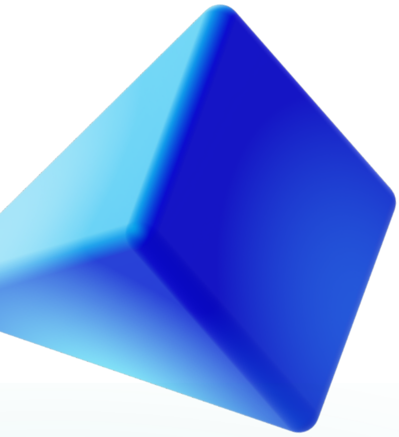
Inherent Document Defense

Confidencial fortifies documents and unstructured data from within, ensuring built-in protection.

In this rapidly evolving era of artificial intelligence, the integration of LLMs into enterprise systems epitomizes the cutting-edge of innovation. But it also brings to a critical need for robust protection mechanisms for sensitive and unstructured data to the forefront.

There is urgency when it comes to solutions that deliver verifiable security and compliance, to ensure that the adoption of LLMs and their potent capabilities does not come at the expense of data privacy and integrity. Enterprises need a cost-efficient, universally compatible, and inherently secure system capable of safeguarding the most confidential information.

Get in touch today to learn how we can support your AI innovation while maintaining the highest data security and compliance standards. We will ensure that your venture into AI is as secure as it is ambitious.



Get in touch.

🌐 confidencial.io

in linkedin.com/company/confidencial-inc/

✉ hello@confidencial.io