An Enterprise’s Guide to AI and LLM Data Protection

PROTOPIA®

About Protopia AI

Protopia AI is a leader in privacy preserving data protection technologies for AI. Protopia enables AI algorithms and software platforms to operate without the need to access plain-text information. Collaborating with enterprises, cloud service providers, and GenAI/LLM providers, Protopia AI ensures that organizations can retain ownership and confidentiality of their data while leveraging AI/ML solutions.

The analysis and insights presented in this piece are based on expert interviews, thorough research, and thoughtful examination. Despite the marketing team's rigorous approach, we acknowledge that it may not fully encompass every aspect or nuance of the topics. We are committed to providing accurate, comprehensive content, and we welcome any feedback or additional information. If you notice any inconsistencies or have supplementary information, please reach out to us at info@protopia.ai

Introduction

DATA CONFIDENTIALITY IS A TOP CONCERN IN THE ADOPTION OF GENERATIVE AI

The Generative AI market, including tools like ChatGPT, has taken off. Yet, **71% of Senior IT leaders hesitate to adopt Generative AI due to security and privacy risks**. AI's reliance on vast amounts of data amplifies the risk of data exposure and leakage.[1]

71%

OF IT LEADERS ARE WARY OF ADOPTING GENAI

AN APPROACHABLE GUIDE TO SIMPLIFY THE MULTITUDE OF SOLUTIONS

Many savvy CIOs and data leaders have found themselves disoriented amidst several AI privacy solutions, often conflating distinct methods and underestimating resource commitments.

Tech leaders seek clarity on topics from basic encryption to advanced security measures such as Synthetic Data, Confidential Computing, and Randomized Re-Representations, which this guide will delve into.

Ensuring the protection and accessibility of data is vital to harnessing the full potential of AI. This guide offers insights into various data protection solutions, gauges their efficacy, and sheds light on both practical and conceptual approaches that practitioners explore.

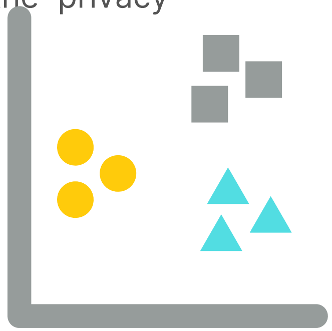
[1] Salesforce, 2023: IT Leaders Call Generative AI a 'Game Changer' but Seek Progress on Ethics and Trust:https://www.salesforce.com/news/stories/generative-ai-research/

Comparing the Solutions

KEY CONSIDERATIONS

This guide offers a simplified comparison of various AI data protection solutions, spotlighting their benefits, drawbacks, and real-life applications. Key considerations include:

- **Training:** The protected data can be used by the AI system for training the model.
- **Inferencing:** The AI system's ability to protect data during ML inferencing/deployment.
- **Inferencing with sensitive information:** The AI system's ability to safely use sensitive data during ML inferencing/deployment.
- **Risk of Exposure**: An assessment of potential threats and vulnerabilities that could lead to data breaches or misuse in AI systems, especially during compute.
- **Data Types/Modalities:** They types of data types that can be protected with an approach to privacy in AI.
- **Compute, Feasibility and Cost Investment:** The rate or speed the solution processes for it to be viable for ML tasks along with the practicality and cost-effectiveness of the privacy solution including the ease of integration and scalability.



Disclaimer: *The analysis presented in this piece is based on expert interviews, thorough research, and thoughtful examination of the solutions discussed. We acknowledge that it may not fully encompass every aspect or nuance of the topics at hand. We welcome any feedback or additional information.*

Guide Outline

Click into each chapter to jump to section

Masking and Tokenization: *The "Hide & Seek" of Data Security*

This guide groups **data masking and tokenization** together due to their notable similarities and their shared status as traditional data protection options.

Both are centered on changing sensitive data, such as phone or customer ID numbers, into non-sensitive "tokens" that can also maintain the field type and length of the data being hidden. A secure, centralized system is typically needed to generate and manage tokens and their associated data.

EXAMPLE

SOCIAL SECURITY NUMBER

INPUT 012-44-1870
TOKENIZATION: 555-65-6644

Data masking is often used to obscure sections of sensitive data with fictitious data, while tokenization is used when it is necessary to revert to the original data.

SIMPLE

- Tokenization and data masking techniques add a level of data privacy and protection while preserving some data utility.
- Tokenization can maintain the type and length of data, allowing it to be processed in systems without altering their operations.
- Integrates easily with minor modifications to existing systems.

INSUFFICIENT

- Tokenization can often be reverse-engineered by analyzing patterns in the non-tokenized components of the data record and context (see next page).
- The centralized system managing tokens becomes a single data exposure and risk point. If compromised, it exposes entire datasets with sensitive information.
- Maintenance can be resource-intensive, which may be pricey for modern high-scale and dynamic AI data.

THE NETFLIX CHALLENGE

DATA MASKING SHRUGGED

Netflix publicly challenged individuals to enhance their movie recommender engine, sharing 10 million anonymized movie rankings from half a million customers.

However, two University of Texas-Austin researchers successfully de-anonymized some users by correlating Netflix's data with public IMDb data. This demonstrated the weakness of Netflix's anonymization method, even with randomized timestamps or unique IDs.

This incident underscores that even slightly identifiable information can jeopardize a database's anonymity, highlighting the limitations of traditional tokenizing/masking data against advanced data mining and AI techniques.

Masking and tokenization fall short in safeguarding high-stakes data sets and are best suited for projects with low risk from potential data leaks.

2. S Lohr, 2012. Netflix Cancels Contest After Concerns Are Raised About Privacy. Wired. Retrieved from https://www.wired.com/2010/03/netflix-cancels-contest/

Homomorphic Encryption: *Computing on Encrypted Data without Decrypting It*

Homomorphic Encryption (HE) allows computation to be performed on data while it remains encrypted. This could provide greater privacy and security for sensitive data since the data never needs to be revealed in plain-text form.

ALWAYS ON

- HE operates on encrypted data without needing to decrypt it, and thereby preserves privacy as sensitive data is never exposed, even during processing.
- HE enables secure outsourcing of data storage and computation. Companies can use third-party cloud providers without risking their sensitive data.
- Since the data is always encrypted, it is easier to meet requirements of regulations, such as GDPR.

FEASABILITY

- Prohibitively high hardware costs, excess storage needs, and speed limitations prevent real-world use.
- Not practical for most AI tasks. Data scientists often use deep neural networks that are limited or unable to be implemented in a homomorphically encrypted fashion.

Think of HE as using an armored vehicle to transport your cash when you want just to make a safe wire transfer.

There are no public real-world examples of HE in enterprise environments to date. The technology is still evolving. Google and IBM have run internal exercises using the solution and offer a few open source developer tools.

[1] https://developers.googleblog.com/2023/08/expanding-our-fully-homomorphic-encryption-offering.html. [2] https://research.ibm.com/topics/fully-homomorphic-encryption

Confidential Computing: *The Data Processing Safehouse*

Confidential Compute or Trusted Execution Environments (TEEs) aim to protect data while it is being used by isolating computations to hardware-based TEEs.[2] Trusted execution environments only enable authorized software processes to access and decode the encryptions.

Leading technology companies, such as Google, Nvidia, Intel, Meta, and Microsoft, are part of the Confidential Computing Consortium, advocating for hardware-based TEEs.

DYNAMIC

- TEEs limit exposure of decrypted data and reduces attack surface.
- TEEs aims to protect sensitive information throughout its entire lifecycle – in transit, at rest, and in use.

At the time of writing, no major confidential computing solutions have been deployed for ML. The technology shows great promise, and will see key developments in coming years.

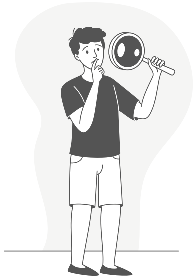
OBSTACLES

- Enterprises must incur additional costs to execute their AI services on platforms with specialized hardware support for confidential computing.
- Reported risks - In May 2021, an attack corrupted data from TEEs that rely on Intel SGX technology.[3]
- Key management in TEEs can result in slower processing than plain text, resulting in performance overhead.

[3] Y. Metha, 2023..How can CIOs protect PII for a new class of data consumers? CIO. Retrieved from https://www.cio.com

Differential Privacy: *When Your Data Plays It "Within Range"*

Differential Privacy is a framework for sharing information about a dataset's patterns while withholding specific individual data. Differential privacy provides margins on how much a single data record from a training dataset contributes to a machine-learning model.



Think of Differential Privacy as a means by which to use a dataset of household incomes to train a model so that the model can be used to understand roughly how much money a typical resident of a particular neighborhood makes, without exposing any one person's income from the training dataset.

APPROXIMATE

- Differential Privacy provides strong mathematical guarantees for privacy of individual data records in training datasets.
- Differential Privacy is applicable to various analytics and ML training tasks to gain general insights and trends.

INCOMPLETE

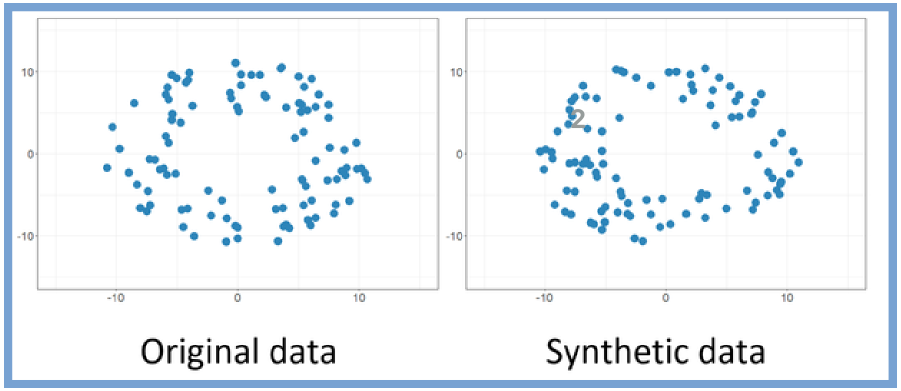
- Differential Privacy requires a data scientist to access and handle large volumes of plain-text data.
- Differential Privacy does not address the ML inference stage in any capacity.

Linkedin uses differential privacy internally to help marketers get information about users reading their content. This gives them a broad sense of audience engagement.[4]

[4] 4.Transcend. 2022, industry Perspective: How LinkedIn Developed a Differentially Private Data Pipeline: https://transcend.io/blog

Synthetic Data: *The Stand-in*

Synthetic data is artificially generated data that is used to mimic real-world data. Synthetic data is often used for testing and training machine learning models.



In this example, synthetic data retains the structure of the original data but is not the same.

Source: Data in Government, GOV.UK (2022)

FORTIFIED

- Just like a movie set is created to look like a real location, but with actors and props that are not real, synthetic data is created to look like real data.
- Synthetic data is easy to generate and train models. Since the data is made up, there is minimal risk of data breach or misuse.
- Synthetic data is useful for filling in missing data or augmenting data.

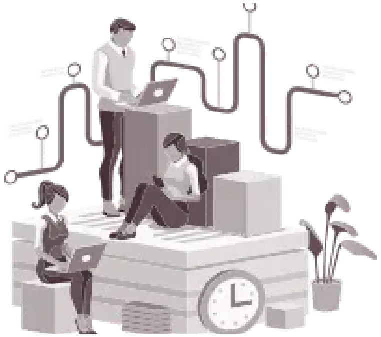
FAKE IT TILL YOU MAKE IT

- Synthetic data can lead to larger than acceptable accuracy losses by missing outliers and edge cases.
- A trained model still needs to use real data to perform inferencing and prediction for the real show; a key component missing in synthetic data.

Synthetic data was used for **Alexa's** new language versions to fill gaps in data for NLP models with sample utterances. [5]

[5] Amazon Science., May 2023 .Tools for generating synthetic data helped bootstrap Alexa's new language releases.

Secure Multi-Party Computing: *The Multi-Player Puzzle*



Secure Multi-Party Computation (SMPC) is a cryptographic protocol that allows multiple parties to compute a function over their inputs while keeping those inputs private.

The computation is distributed among the parties, and no single party can access the complete dataset.

MULTI-BENEFITS

- SMPC offers the dual advantage of data privacy and utility-data remains encrypted yet usable, reducing data breach risks.
- Efficient compared to protocols like fully homomorphic encryption, SMPC is less resource-intensive than HE because the computations are performed on smaller, partitioned data sets, reducing the computational load.

TOO MANY COOKS

- SMPC requires communication among all parties, escalating overhead and costs, and slows down processes for large datasets necessary for many ML applications.
- SMPC might be susceptible to colluding parties scheming to discover the private inputs of the other participants, undermining the system.
- Implementing SMPC, particularly with numerous parties and large datasets, is complex, introducing management challenges and error potential.

Two US county depts have used SMPC to analyze datasets on pay equality with the assistance of Boston University.[6]

[6] 6.United Nations Statistics Division.. 2023. Welcome to the United Nations Statistics Division Wiki. Retrieved from https://unstats.un.org/wiki/pages/viewpage.action?pageId=150012020

Federated Learning: *Collaborate Without Sharing Data*

Federated learning is a decentralized machine learning solution that allows multiple parties to train a shared model on their local datasets without revealing the underlying data to each other.

POWER IN NUMBERS

- Federated learning allows data to remain on the device where it was created, protecting it from exposure during transmission or storage.
- Federated learning can allow for more data to be used in training a model, potentially improving the model's performance.

PITFALLS

- Federated learning fails when one party or device does not do its part or shuts down.
- Federated learning does not address the ML inference stage, which still exposes data to the ML model during deployment.
- An attacker could use observations on an ML model's parameters to infer private information about the training data from other parties.[3]

Federated learning has shown utility in examining COVID-19 patient data. Researchers were able to train models with 20 international datasets while maintaining user anonymity.[7]

[7] Blankenberg & Schonberger,.2021. Machine learning for clinical neuroscientists: an introductory tutorial. NeuroImage: Clinical, 32, 102784.

Data Clean Rooms: *The Data Chaperone*

A data clean room is a secure sandbox where teams can share and analyze sensitive data without worrying about privacy breaches (in principle). Think of it as a data playground that is fenced off for safety and monitoring.

Parties who are exchanging data must be present simultaneously, put data into the clean room, do their analyses, get the necessary insights, and then leave. They do not get to take any unprotected personal data out of the data room, just the aggregated output.

Google's Ads Data Hub is an instance of data clean room. It allows advertisers to study their campaign data and user behaviors, but only in aggregated form. This respects user privacy and helps advertisers to enhance their campaigns.[8]

WHAT IS NEAT

- Through (pseudo-)anonymization, personal data is rendered safe for combining datasets and analysis in data clean rooms.
- Data clean rooms foster a secure environment for collaborative data sharing and exploration.

DIFFICULTIES

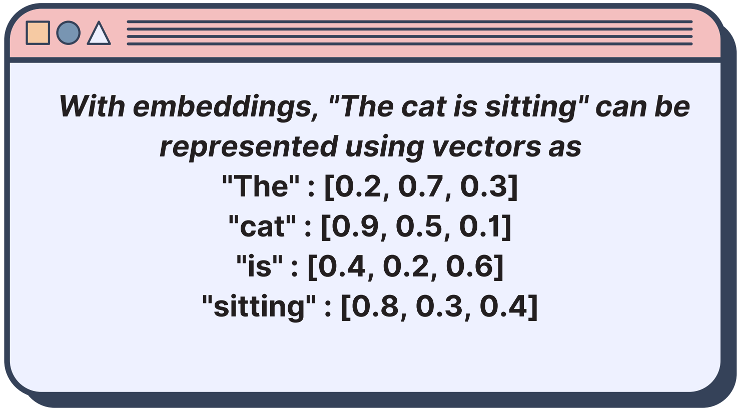
- Data clean rooms requires relevant parties to be present at the same time.
- Third-party reliance requires trust in an environment without universally accepted implementation guidelines. The black-box nature of clean rooms can make using the data for ML/DL tough.
- Data clean rooms require extensive technical know-how and are expensive to set up, maintain, and operate.

[8] G. Sloane. .2022. Why Google is building clean rooms for audience targeting? AdAge: https://shorturl.at/aceX5

Embeddings: *The Features of Language Processing*

Embeddings in language processing are numerical representations of words or phrases that capture their meaning and context. Embeddings help Natural Language Processing (NLP) models efficiently process and understand natural language. They provide a level of abstraction but are not meant to, and do not, inherently provide privacy or security.

Embeddings are meant to enable efficient Natural Language Processing and are not a security measure.



With embeddings, "The cat is sitting" can be represented using vectors as

"The" : [0.2, 0.7, 0.3]
"cat" : [0.9, 0.5, 0.1]
"is" : [0.4, 0.2, 0.6]
"sitting" : [0.8, 0.3, 0.4]

YES...

- Embeddings are how NLP models understand underlying natural language, and are only a numerical representation for the purpose of efficient computation.

BUT...

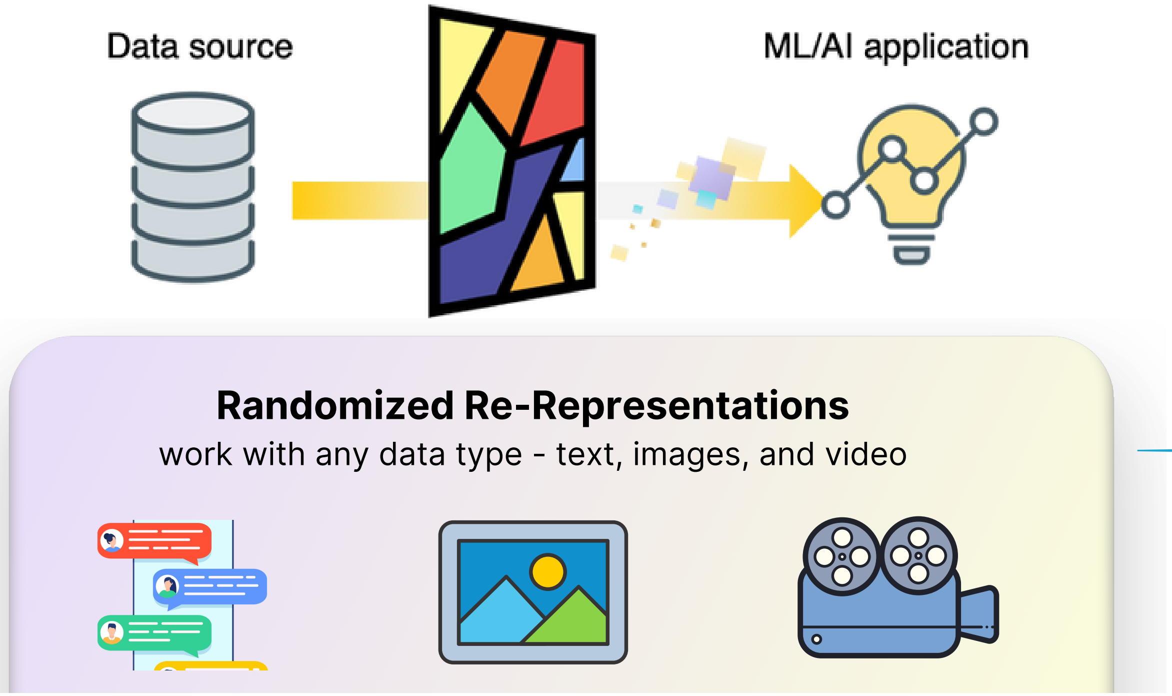
- Embeddings are not a security or privacy measure. There is a 1-to-1 correspondence with the tokens they represent.
- Improper use, sharing, or leakage of embeddings may expose sensitive information.
- Attackers can reconstruct confidential data from the embeddings themselves.

Stained Glass Transform™ *Privacy Uncompromised*

Stained Glass Transform™ (SGT) is a new category of AI privacy-enhancing technology that enables enterprises to retain data ownership while using cutting-edge ML solutions.

Stained Glass holistically transforms data to a **Randomized Re-Representation** through a stochastic data transformation which is learned based on Protopia AI's patented technology. These representations are irreversible to the original data but are understandable by the target AI model. Randomized Re-Representations can be used by AI tasks with high accuracy without exposing the original representations in unprotected form.

Protopia AI Stained Glass Transform™



Randomized Re-Representations

work with any data type - text, images, and video

HOLISTIC

- Stained Glass Transform is a lightweight and non-intrusive solution that retains data ownership.
- SGT enables using AI models to gain insight from more real-world enterprise data without compromising confidentiality.
- SGT protects data during ML training and inferencing, and LLM fine-tuning/querying.

LIMITATIONS

- Stained Glass Transform is intended for data that is going to be used for AI, not any arbitrary target algorithm.
- Fine-tuning the levels of added stochasticity/entropy may be needed to obtain desired accuracy with transformed data.

Existing Stained Glass Use Case

FINE-TUNING DATA FOR NETWORK LOG ANALYSIS

A customer used Stained Glass to transform their data and fine-tune an open-source LLM to analyze network logs.

Despite the transformed data being unrecognizable, the fine-tuned model was effective with almost identical accuracy to that of a model fine-tuned with unprotected plain-text logs

PROTOPIA

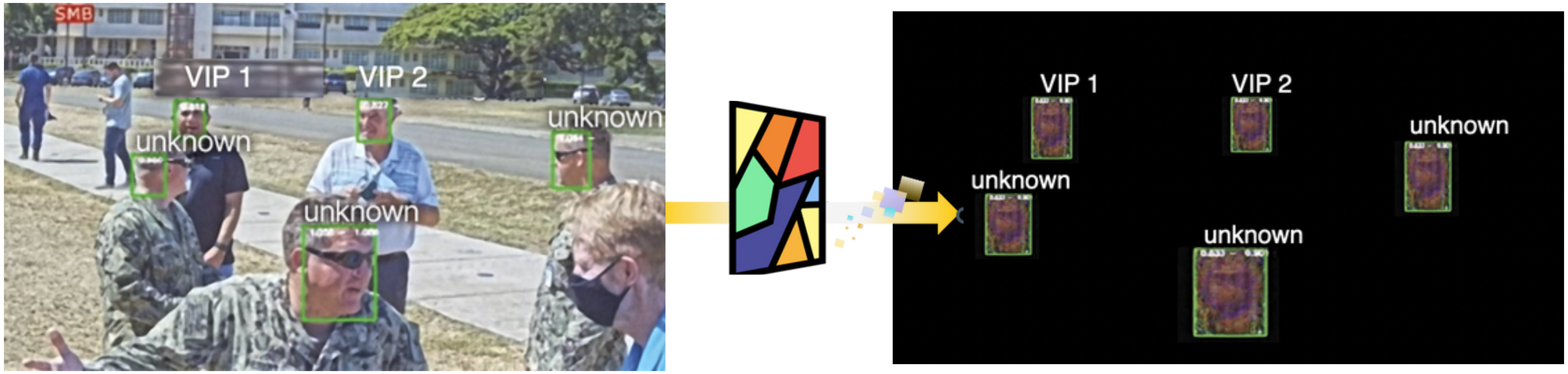
Unprotected Text Data	Reconstructed from Stained Glass Embeddings
15, 1, 72, Europe \ Berlin \ Sales, Gerrard S, 25. 33. 8. 119, kenta. s, FileCopy, File copy Source,\\ 192. 168. 102. 241 \ PUBLIC \ Confidential \ customer information 2. XLS,,, confidential folder—,,,	*Front Solitary Inheritance, 73 Catch Remember Shanghai Ndlauuri Muroran Axillary Nakamura Trump Card Shimbashi Rado for short Military history Sonouchi Masuda Body Opportunity Chasing ft No. Ritsu Animal Royon 2 Maneki ioni*
12, 1, 47, India Headquarters \ Delhi Office \ Support Department, Rajesh N, rajesh. n, lsass. exe, C :\ Windows \ System 32 \ lsass. exe, 9465 E 465 E 65 EB 4303 BF 87483 B 9621 D 402 E 848 A 50 E 6 D	*op Arrogance Dust Coordinates Load Mr. and Mrs. Husband Princess Happ Prefectural Road y Sou Appearance match Wo GP ě Dinner Mistake Momomori Minato : Group Capture Physics Chief Indispensable One Side After*

Stained Glass Transform™ in Facial Recognition

Use Case Definition

Protopia AI implemented Stained Glass Transform™ to demonstrate the breakthrough capability of performing real-time face recognition of VIPs for the US Navy without exposing unprotected images. This capability was showcased at the annual 2022 Trident Warrior exercise.

Solution for the US Navy: Transform data to match intent of model consuming it with Stained Glass



Live detection of VIPs exposes base personnel and assets

SGT only passes necessary information to model to identify VIPs

Solution

SGT transformed subjects' faces into **Randomized Re-Representations** that were only readable by the facial recognition model with near-perfect accuracy to identify VIPs. The transformation ran at the same frame as the real-time recognition without interference. Individual faces were not identifiable. The use case showcased Stained Glass Transforms's ability to protect AI deployment successfully.

Comparative Summary

	Masking / Tokenization	Homomorphic Encryption	Confidential Computing	Differential Privacy	Synthetic Data	Secure Multi-Party Computation	Federated Learning	Data Clean Rooms	***Stained Glass Transform***
Support for Training	Yes	Yes, but very slow	Requires specialized hardware	Yes	Yes	Yes, but very slow	Yes	Yes, but limited	Yes
Support for Inferencing	Yes	Yes, but prohibitively slow	Requires specialized hardware	Not Applicable	Not Applicable	Slow	Not Applicable	Not Applicable	Yes
Inferencing with sensitive information	No	Yes	Yes	No	Not Applicable	Yes	No	No	Yes
Data Exposure during Compute	Yes, partially due to context	No	Theoretically no, but breaches have occurred	Yes	No	No	Partially	Yes	No
Data Types / Modality	All	All	All	All	All	All	All	All	All
Compute	Low	Requires specialized hardware	Requires specialized hardware	Normal	Normal	High communication overhead	Normal	Normal	Low

Comparisons are simplified interpretations to be used as a quick guide. For any given use-case, full information along with potentially other factors could influence selection.

Final Takeaways

This guide provides a comparative summary of ten AI privacy solutions into distilled key points. Much of the technology is still developing, so this guide's analysis cannot encapsulate every nuance. Nonetheless, the clear imperative emerges: ***proactive steps are essential for safely navigating the rise of generative AI and enable the widespread use of LLMs.*** Finding the right solution must be feasible and provide the protection needed. There is no one-size-fits-all solution.

- Employing old-school methods like data masking and tokenization can still leak the context of the data record beyond just PII (or other specific parts of the data record) that can be masked/tokenized. Be aware that this can undo the impact of the performed masking or tokenization. Additionally, you give up the ability to actually use the masked/tokenized information which may be necessary to your use-case.
- Modern innovations like homomorphic encryption and confidential computing hold promise but require specialized hardware and have limitations that have prevented them from being broadly usable, especially for more sophisticated machine learning such as Deep Learning, Generative AI, and LLMs.
- Consider synthetic data as a training boot camp - it's great for augmenting data sets and patching up their weaknesses. However, when it comes to making predictions about your business' data after the model has been trained, you need to use your real data securely.
- Stained Glass Transform allows enterprises to protect their data and preserve privacy without adding resource intensive computation or requiring specialized hardware. It can be used for any modality of data and at any part of the AI cycle: inferencing, training, and fine tuning.

Disclaimer

The analysis and insights presented in this document are based on expert interviews, research literature, and thoughtful examination of the solutions discussed. Despite the marketing team's rigorous approach, we acknowledge that it may not fully encompass every aspect or nuance of the topics. We are committed to providing accurate, comprehensive content, and we welcome any feedback or additional information. If you notice any inconsistencies or have supplementary information, please reach out to us at info@protopia.ai