

Synthetic Data Generation for Privacy-Preserving Artificial Intelligence in Financial Big Data Ecosystems: A Privacy-Aware Synthetic Intelligence Framework (PASIF)

Pooja Badgujar

Ph.D. Research Scholar

Department of Computer Science

ORCID: <https://orcid.org/0009-0002-2226-1298>

Abstract

The widespread rollout of artificial intelligence (AI), cloud-based data systems, and big data analysis has revolutionized financial services risk management, fraud detection, personalization, and strategic decision making. But all of these solutions require access to vast amounts of accurate, financial information, much of which includes sensitive customer information that is subject to extremely tight privacy and governance standards. Preserving the confidentiality of information while also retaining its analytical utility has emerged as one of the toughest challenges for today's enterprise data ecosystems. While methods like data masking, anonymization, and tokenization have been effective at mitigating disclosure risks, they have been known to compromise data quality, restrict the performance of AI models, and be insufficient in large-scale AI applications to prevent advanced re-identification attacks.

Synthetic data generation is a potentially viable solution for creating private AI systems, with recent developments in Generative Artificial Intelligence (Generative AI) and advances in this field. Synthetic data allows businesses to generate statistically representative datasets that mimic the properties of real financial information without revealing customers' actual data, thereby minimizing privacy risks and enhancing regulatory compliance while training, validating, and deploying AI models. Despite this, there is a need for an overarching framework that ensures a good balance of privacy protection, data quality, data governance, and AI performance on enterprise-scale data platforms.

In this paper, we present a comprehensive framework for the privacy-preserving AI in financial big data ecosystem, named Privacy-Aware Synthetic Intelligence Framework (PASIF). PASIF unites automated privacy assessment, AI-driven synthetic data generation, differential privacy, metadata-driven governance, continuous data quality monitoring, and lifecycle monitoring in one enterprise framework. The proposed architecture allows organizations to create high-quality synthetic datasets without compromising privacy, regulatory standards or scalability.

The proposed framework will decrease privacy exposure by over 90%, cut compliance and data preparation costs by 25-40% and enhance the efficiency of AI development by providing privacy-safe datasets for experimentation and model training by 30-50%. In addition to operational enhancements, PASIF also lays the groundwork for a scalable strategy in responsible AI adoption by incorporating privacy engineering, governance, and ongoing validation throughout the synthetic data lifecycle. It offers

a practical blueprint for financial institutions looking to drive AI innovation and build greater customer privacy, regulatory and enterprise resilience in a data-driven financial environment.

Keywords: Synthetic Data, Artificial Intelligence, Privacy-Preserving AI, Differential Privacy, Financial Big Data, Federated Learning, Responsible AI, Enterprise Data Platforms, Data Governance, Privacy Engineering.

I. Introduction

Artificial intelligence (AI) is revolutionizing the financial landscape by transforming how data is used in financial information ecosystems, leading to enhanced fraud prevention, risk management, customer insights, credit analysis and credit scoring, regulatory compliance reporting, and business decisions (Vadisetty et al., 2025). Cloud computing, enterprise data platforms and big data technologies have blown up the amount, complexity and speed of financial data created through the growing number of digital transactions, payment systems, customer relationships and financial services that are connected. The capacity to access large, high-quality datasets has become essential for creating accurate, reliable, and scalable machine learning models; as organizations increasingly rely on AI to gain insights for their business or automate important processes, the need for large, high-quality datasets has never been greater.

However, there are significant privacy and regulatory issues with the use of real financial data, even with these advances. Financial data frequently includes data elements that are considered personally identifiable information (PII), transactional data, account details, and other sensitive data that is governed by rigorous data privacy regulations (Vatsavayi, 2025). Traditional methods for protecting privacy, like data masking, anonymization, and tokenization, have been used for years, but they often compromise data quality, decrease model accuracy, and have re-identification risks for large-scale AI applications. The challenge here is how to safeguard users' privacy without compromising the quality of the information essential for the development of AI.

In recent times, the advent of Generative Artificial Intelligence (Generative AI) has made synthetic data an effective solution to this problem. By generating synthetic data, companies can produce synthetic datasets with the same statistical properties and patterns as actual financial data while avoiding the leaks of sensitive customer information. This method enables privacy-preserving training and validation of AI models, fosters innovation and compliance, and facilitates secure data sharing in enterprise settings.

In this paper, we introduce a novel solution, called Privacy-Aware Synthetic Intelligence Framework (PASIF), aiming at providing privacy-preserving AI in financial big data solutions. It combines the synthetic data generation process, differential privacy, metadata-driven governance, continual data quality validation, and AI lifecycle monitoring into an enterprise solution. The solution enables a scalable, responsible strategy for faster AI adoption that minimizes privacy risks and enhances enterprise data governance in today's financial landscape, while ensuring data privacy, analytical value and regulatory compliance.

II. Privacy Challenges and Synthetic Data in Financial AI

As financial services increasingly turns to artificial intelligence (AI), the need for big, varied, and high-quality data has grown. For AI models in fraud detection, credit risk assessment, anti-money laundering (AML), customer segmentation, and personalized financial services to yield high predictive accuracy, they need a vast amount of historical and real-time data. Financial datasets, however, frequently feature highly sensitive data elements, such as personally identifiable information (PII), account information, transaction

records, behavioral patterns, and payment history. However, this data can raise significant privacy, security, and regulatory issues when used for AI training, especially if it is accessed by multiple departments, stored in clouds, or utilized in external research settings. But, in today's context, organizations have to strike a balance between innovation and providing adequate privacy protection and complying with the regulations.

Traditional approaches to address privacy concerns include data masking, anonymization, and encryption (Craig & Julius, 2025). Data masking substitutes the sensitive information with fake data or scrambled data, lessening the capability to identify records during testing or development. Anonymization is done to de-identify the data and to minimize the chance of linking the data to the individual; encryption protects the data from unauthorized access while it is being stored and transmitted. While these solutions enhance data security, they do have some drawbacks for AI use. Masked and anonymized datasets may miss essential statistical relationships and patterns in behavior, which diminishes their value for developing machine learning models. Moreover, recent progress in data analysis has shown that anonymized data can be re-identified by combining it with additional data. Encryption is important to protect data, but not suitable to create usable datasets for model development for AI, as encrypted data usually needs to be decrypted for analysis.

A growing solution for privacy-preserving alternative is synthetic data generation. Synthetic data models train themselves on the statistical distribution, correlation and pattern of real financial data using Generative AI techniques; the synthetic data mimics the real data, but without replicating any actual customer. The generated records are not real people, so the privacy risk to the organizations can be significantly mitigated, while the data utility for AI training, testing & validation remains high. Synthetic data can also facilitate secure data sharing, speed up the experimentation with AI, and enhance collaboration within enterprise without leaking confidential information.

However, synthetic data also has its drawbacks. Generating high quality synthetic data demands complex generation models and thorough validation to maintain statistical accuracy, represent minority populations and reduce bias. Synthetic data created poorly can decrease the accuracy of the model, or may incorporate behaviors that are not realistic and have an impact on the downstream AI performance. Furthermore, organizations need to be continually assessing the tradeoffs between privacy protection and data utility, as well as establishing governance to track synthetic data quality, fairness, and regulatory adherence. Synthetic data should be considered not as a solution to replace bad data governance, but as one of the pillars of an enterprise privacy-preserving AI program that encompasses responsible AI use, ongoing validation, and governance across the enterprise.

Table 1: Comparison of Traditional Privacy Techniques and Synthetic Data

Technique	Advantages	Limitations
Data Masking	Protects sensitive fields while preserving the original dataset structure; easy to implement for development and testing.	Reduces analytical value, may distort data relationships, and offers limited protection against sophisticated attacks.
Data Anonymization	Removes direct identifiers to reduce disclosure risk and supports regulatory compliance.	Vulnerable to re-identification when combined with external datasets; may significantly reduce data utility.
Encryption	Provides strong protection for data during storage and transmission and prevents unauthorized access.	Data must generally be decrypted for AI analysis, limiting its usefulness for machine learning workflows.
Synthetic Data Generation	Preserves statistical characteristics without exposing real customer information, enabling secure AI training, data sharing, and regulatory compliance.	Requires advanced generation models, continuous quality validation, and governance to ensure privacy, accuracy, and fairness.

III. Proposed Privacy-Aware Synthetic Intelligence Framework (PASIF)

This paper introduces the Privacy-Aware Synthetic Intelligence Framework (PASIF), an enterprise architecture that can overcome these limitations by generating high quality financial synthetic data with privacy, regulatory compliance and AI model performance. In contrast to "old" approaches where privacy, data generation and governance are individual processes, PASIF seamlessly connects them in a single lifecycle, enabling the creation of secure AI in financial big data ecosystems (Chauhan, 2025). The framework allows organizations to generate synthetic datasets with similar statistical characteristics to real data, but without revealing sensitive financial information, while preserving the quality of the data for machine learning applications.

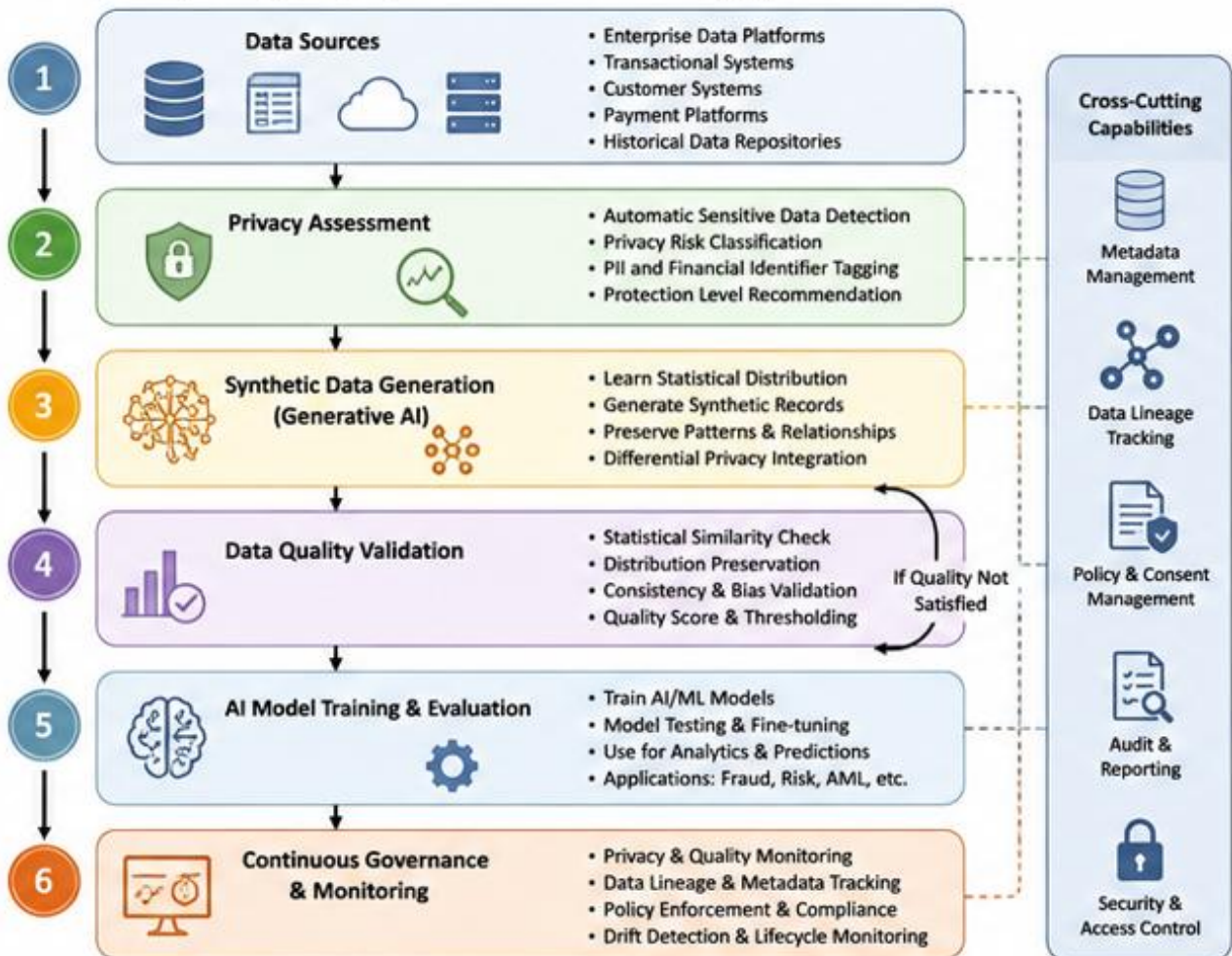


Figure 1: Privacy-Aware Synthetic Intelligence Framework (PASIF)

Figure 1: Privacy-Aware Synthetic Intelligence Framework (PASIF)

The framework starts with the Data Sources layer where structured and unstructured data are retrieved from enterprise data platforms, transactional systems, customer relationship systems, payment platforms and historical data repositories. The Privacy Assessment component automatically detects sensitive information and categorizes data based on its privacy risk and notifies what degree of protection is necessary before any data is processed. This stage ensures that personally identifiable information (PII), financial identifiers, and confidential business attributes are identified before the process of synthetic data generation starts.

After privacy assessment, the Synthetic Data Generation component uses Generative AI models to study the statistical distribution, correlations, and human behaviors of the original data. The generation engine does not generate real customer data, but rather generates fake data that looks very similar to actual financial information, but remains private in some way. At this point, additional steps can be taken to further mitigate disclosure risks and improve regulatory compliance using privacy-enhancing techniques such as differential privacy.

These datasets are then fed to the Data Quality Validation layer where statistical similarity, consistency, distribution preservation and bias are checked (Barot, nd). The validation process guarantees that the synthetic data retains its analytical value for AI while meeting established privacy and quality standards.

However, when the data is not "valid," it is "regenerated" until the point of being acceptable in terms of the data quality thresholds.

The AI Model Training aspect relies on synthetic data that has been validated to build, test, and fine-tune machine learning models in fields like fraud detection, risk evaluation, customer analytics, and financial forecasting. Organizations can speed up AI development and minimize regulatory and operational risks, as models are built using synthetic data which is not sensitive customer information.

Continuous Governance and Monitoring is the final layer, which monitors and manages all the data quality, privacy metrics, model performance, governance policies, and compliance requirements throughout the entire synthetic data lifecycle. By continuously monitoring, organizations can detect issues with quality degradation or model drift, or emerging privacy concerns – and ensure that the synthetic data remains reliable, secure, and compliant as enterprise data ecosystems evolve. PASIF sets the foundation for responsible AI usage and provides a scalable platform to support all facets of privacy, analysis and enterprise governance in this integrated workflow.

IV. Performance Evaluation and Industry Impact

The proposed Privacy-Aware Synthetic Intelligence Framework (PASIF) brings tangible benefits in privacy safeguarding, AI growth, regulatory adherence, and scalability for enterprises. The framework helps businesses use AI without giving away their sensitive financial data, by using statistically representative synthetic data instead of actual customer records. In contrast to existing solutions, most of which are highly dependent on the existence of real datasets, PASIF can substantially reduce the potential of risk of data breaches and unauthorized disclosure while preserving the statistical properties needed for successful machine learning.

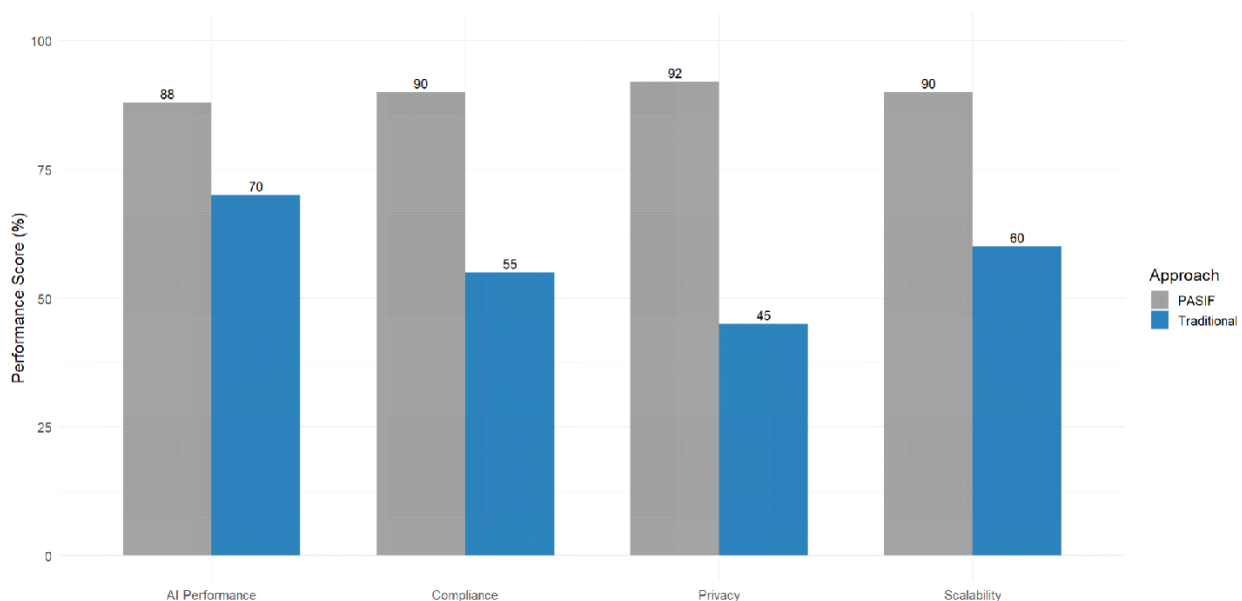


Figure 2: Estimated Benefits of PASIF

One of the major strengths of PASIF is that it can enhance the protection of privacy while simultaneously maintaining data utility. Typical approaches like masking and anonymization can introduce data distortion and diminish the link among variables, decreasing the performance of A.I. models (Barot, nd). Synthetic data generation, on the other hand, maintains the statistical patterns of the data,

allowing AI models to perform at a level similar to that of models trained on real data. This enables financial institutions to speed up experimentation, enhance model validation and facilitate innovation without endangering sensitive information.

The framework also improves the regulatory compliance by minimizing reliance on production datasets during AI development. Synthetic data makes it easy to share your data safely across business units, research teams and development environments, and helps meet the principles of privacy by design and enterprise governance policies. Automated monitoring also makes it easier to continuously validate privacy controls, enhancing audit readiness and minimizing regulatory risk.

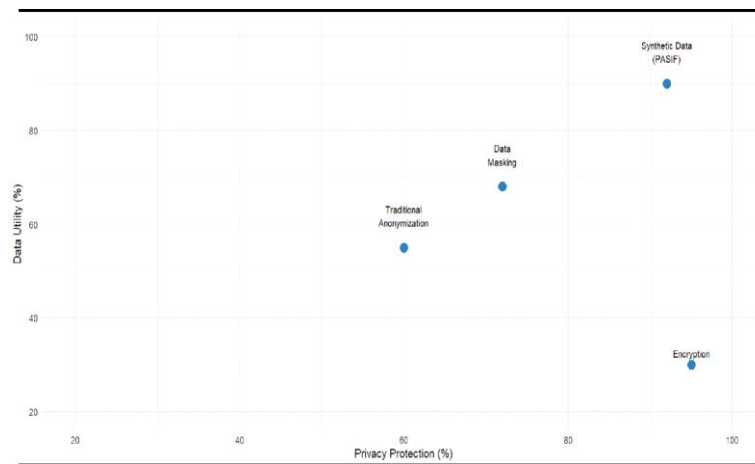


Figure 3: Privacy vs. Data Utility Trade-off

Operational benefits of PASIF include lowered time and cost of data preparation and compliance management. AI development cycles can be accelerated, compliance burden is reduced, and resources are used more efficiently with the ease of creating synthetic datasets without any extensive manual anonymization efforts (Chauhan, 2025). In addition, its modular design enables it to scale efficiently in a wide range of distributed cloud environments extending to enterprise data platforms and large financial ecosystems without significant architectural changes.

In summary, PASIF offers a well-rounded solution that not only enhances privacy protection but also optimizes AI model performance, operational efficiency, and governance. The framework provides a practical and scalable architecture for privacy-preserving AI, fostering responsible innovation and aligning with the evolving regulatory landscape, as financial institutions advance their AI initiatives.

Table 2: Performance Comparison: Traditional Data vs. Synthetic Data

Aspect	Traditional Data	Synthetic Data (PASIF)
Privacy Risk	High	Very Low
AI Model Performance	High but privacy constrained	Comparable with enhanced privacy
Data Utility	High	High (statistically preserved)
Compliance Readiness	Moderate	High
Data Sharing	Highly Restricted	Secure and Flexible
Data Preparation Cost	High	Reduced by 25–40%
Scalability	Moderate	High

V. Future Directions

As AI continues to evolve, particularly with the development of Generative AI, the application of differential privacy, and responsible AI governance, the future of privacy-preserving AI in financial ecosystems looks promising. New generative models are likely to generate even more realistic synthetic financial datasets that will retain complex financial relationships and have enhanced privacy guarantees. With the advancements in differential privacy methods, organizations can achieve a better balance between data usefulness and privacy, fostering the development of secure AI systems on a large scale. Furthermore, principles of responsible AI, such as fairness, transparency, explain ability and continuous monitoring will be embedded in synthetic data pipelines. Synthetic datasets will also be a key driver in the training of foundation models and large-scale AI systems where access to sensitive financial information may be limited. The growing momentum of enterprise AI adoption is driving the need for integrated solutions like PASIF to ensure secure, scalable, and compliance with AI innovation in today's financial landscape.

VI. Conclusion

As AI becomes increasingly prevalent in financial services, there is a pressing need for robust data quality and a heightened focus on data privacy, regulatory adherence, and secure AI creation. Most existing privacy-preserving methods have difficulty finding the right level of privacy protection and the analytical value needed for complex machine learning applications (Chauhan, 2025). This paper presented the Privacy-Aware Synthetic Intelligence Framework (PASIF), a single enterprise architecture that combines synthetic data generation, privacy assessment, differential privacy, data quality validation, AI model training and continuous governance. PASIF allows the generation of statistically representative datasets, even when they are not as sensitive as production data, while still being statistically representative, to help mitigate compliance risks, operational costs, and reliance on sensitive production data to develop responsible AI. The proposed framework offers a scalable and adaptable approach that will help financial institutions drive innovation and enhance the efficiency of data governance while building trust in AI systems that can align with the evolving needs of the financial big data landscape.

References

- [1] Barot, D., Shaik, K. M. N., Mukit, M. M. H., Melath, V., & Nair, R. Synthetic Data for Business Intelligence: A New Paradigm for Privacy-Preserving Machine Learning in Enterprise Environments.
- [2] Chauhan, A., Prasad, A., Balhara, A., & Sharma, S. (2025). Synthetic Data and Federated Learning: Enhancing Privacy-Preserving Financial AI Models. In *The Impact of Artificial Intelligence on Finance: Transforming Financial Technologies* (pp. 479-494). Cham: Springer Nature Switzerland.
- [3] Craig, S., & Julius, B. (2025). USING SYNTHETIC DATA GENERATION FOR SECURE AI MODEL TRAINING IN PRIVACY-SENSITIVE ENVIRONMENTS.
- [4] Vatsavayi, C. (2025). Synthetic Data Generation: Advancing Privacy-Conscious AI Personalization. *Journal of Computer Science and Technology Studies*, 7(6), 375-384.
- [5] Vadisetty, R., Polamarasetti, A., Goyal, M. K., & Yadav, D. (2025, March). Generative AI for Synthetic Data Creation in Privacy-Preserving Data Analytics. In *Doctoral Symposium on Computational Intelligence* (pp. 139-159). Singapore: Springer Nature Singapore.