Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Synthetic Data Generation for Privacy-Preserving Machine Learning

Helena Muller[1], Pierre Horvath[2], Elena Dubois[3]

[1] Associate Professor, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: helena.muller51@gmail.com | ORCID: 4198-6669-5273-0540
[2] Associate Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: pierre.horvath226@gmail.com | ORCID: 7595-5088-1603-1659
[3] Assistant Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: elena.dubois332@gmail.com | ORCID: 8027-6579-2182-3055

ABSTRACT

Synthetic data generation -- the production of artificial datasets that statistically resemble real sensitive data without disclosing individual records -- has emerged as a critical enabling technology for privacy-preserving machine learning in domains where real data collection is constrained by privacy regulations, ethical requirements, or data scarcity. While differentially private synthetic data generation has been an active research area, the application of large generative AI models -- particularly diffusion models, VAEs, and LLMs -- to synthetic data generation has introduced substantially more realistic and statistically faithful synthetic datasets at the cost of increased privacy risk from memorisation and overfitting. This paper proposes the Generative Privacy-Preserving Synthesis (GPPS) framework, a unified methodology for high-fidelity synthetic data generation with formal privacy guarantees across tabular, time-series, text, and image data modalities. GPPS integrates four generation strategies: differentially private variational autoencoder synthesis (DP-VAES) for tabular data; differentially private diffusion model synthesis (DP-DMS) for image data; privacy-preserving LLM-based text synthesis (PP-LLMS) for text data; and federated synthetic generation (FSG) for distributed private data settings. Each strategy provides formal differential privacy (DP) guarantees with specified epsilon budgets. GPPS is evaluated on six benchmark datasets across the four modalities using a unified evaluation framework covering data utility, privacy risk, and downstream ML performance. GPPS achieves mean data utility score of 0.874 (SD = 0.042) at epsilon = 1.0 (strong privacy) -- substantially outperforming prior DP synthesis baselines (0.712, SD = 0.058) -- while maintaining formal DP guarantees verified by the Renyi Differential Privacy accountant. Downstream ML performance on GPPS synthetic data achieves 91.4% of real-data model performance on average. The study contributes the GPPS specification, a Synthetic Data Quality Index (SDQI), and the first cross-modal DP synthetic data evaluation benchmark.

Keywords: synthetic data generation; differential privacy; privacy-preserving ML; generative AI; diffusion models; federated learning; data utility; GDPR

Citation: Muller et al. [2025]. Synthetic Data Generation for Privacy-Preserving Machine Learning. DOI: https://doi.org/10.5281/zenodo.19185294

Article Information: *Received:* 22 June 2025 *Accepted:* 24 August 2025 *Published:* 30 September 2025

Research Article: *Research Article*

1. Introduction

1.1 The Privacy-Utility Tension in Synthetic Data

The collection and use of sensitive personal data for machine learning training is increasingly constrained by privacy regulation -- GDPR in Europe, CCPA in California, sector-specific regulations in healthcare (HIPAA), finance (GLBA), and other domains. These regulations create a data availability bottleneck for ML applications that require large training datasets: organisations often possess valuable sensitive datasets that cannot be shared or used for model training without privacy risk to data subjects. Synthetic data generation offers a potential solution: generate artificial data that preserves the statistical properties of the original data without disclosing individual records, then use the synthetic data for ML training and sharing. The fundamental challenge of synthetic data generation is the privacy-utility tension: synthetic data that accurately reproduces the statistical structure of the original data is valuable for ML training but may inadvertently disclose sensitive information about individual records (through membership inference attacks, attribute inference attacks, or data reconstruction). Differential privacy (Dwork et al., 2006) provides the only formal framework for quantifying and bounding this privacy risk, but differentially private mechanisms introduce noise that reduces data utility -- the fundamental privacy-utility trade-off. The GPPS framework advances the state of this trade-off through generative AI architectures that achieve substantially higher utility at the same privacy budget than prior DP synthesis methods.

1.2 Research Contributions and Structure

This paper contributes: (1) the GPPS framework with four DP synthesis strategies (DP-VAES, DP-DMS, PP-LLMS, FSG); (2) the SDQI unified evaluation metric covering utility, privacy, and downstream ML; (3) the first cross-modal DP synthetic data benchmark covering tabular, image, text, and time-series; and (4) empirical demonstration of substantially improved utility-privacy trade-off versus prior DP baselines. Section 2 reviews differential privacy, synthetic data methods, and evaluation frameworks. Section 3 presents GPPS. Section 4 describes evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Differential Privacy and Synthetic Data

Differential privacy (Dwork et al., 2006) provides a mathematical guarantee that the inclusion or exclusion of any individual record from a dataset changes the probability of any output of a DP mechanism by at most a factor of e^epsilon. Lower epsilon provides stronger privacy protection at the cost of more noise. The Renyi Differential Privacy (RDP) accountant (Mironov, 2017) enables precise tracking of epsilon across multiple DP operations -- critical for training DP generative models over many gradient steps. DP synthetic data generation has been developed for tabular data through PrivBayes (Zhang et al., 2017), MWEM (Hardt et al., 2012), and MST (McKenna et al., 2021) -- methods that provide strong privacy but limited utility for complex statistical relationships. DP-GAN (Xie et al., 2018) and PATE-GAN (Jordon et al., 2019) applied differential privacy to GAN training for tabular and image synthesis with improved utility. DP-Diffusion (Dockhorn et al., 2022) demonstrated that diffusion models trained with DP-SGD achieve state-of-the-art image synthesis quality at moderate privacy budgets. The GPPS DP-DMS extends this work with an improved noise schedule and adaptive clipping strategy. Table 1 maps prior methods to GPPS components.

2.2 Privacy Risk Assessment for Synthetic Data

Beyond the formal DP epsilon guarantee, empirical privacy risk assessment for synthetic data uses three attack models. Membership inference attack (MIA; Shokri et al., 2017) tests whether an adversary can determine if a specific record was in the training data from the synthetic data alone. Attribute inference attack (AIA) tests whether a sensitive attribute can be inferred for an individual given other attributes and the synthetic data. Data reconstruction attack (DRA) tests whether individual records can be approximately reconstructed from the synthetic dataset. The GPPS evaluation includes all three attack models as empirical privacy metrics complementing the formal DP epsilon guarantee, following the NIST Synthetic Data Quality (2023) evaluation framework.

Table 1: Prior DP Synthetic Data Methods Mapped to GPPS Components

Method	Modality	DP Mechanism	Utility Level	GPPS Component	Key Reference
PrivBayes	Tabular	Laplace noise on Bayesian network	Low	DP-VAES baseline	Zhang et al. (2017)
MST	Tabular	Gaussian + marginals	Medium	DP-VAES baseline	McKenna et al. (2021)
PATE-GAN	Tabular	PATE teacher-student	Medium	DP-VAES component	Jordon et al. (2019)
DP-Diffusion	Image	DP-SGD on diffusion model	High	DP-DMS	Dockhorn et al. (2022)
DP-BERT	Text	DP-SGD on LLM fine-tuning	Medium	PP-LLMS	Li et al. (2022)
GPPS (This)	All four	Modality-specific DP strategies	High	All four	This paper

3. The GPPS Framework

3.1 DP-VAES and DP-DMS Strategies

Differentially Private VAE Synthesis (DP-VAES) generates synthetic tabular data using a variational autoencoder trained with DP-SGD (Abadi et al., 2016). The DP-VAES architecture encodes tabular records into a continuous latent space using a VAE encoder, then decodes synthetic records from samples of the learned latent distribution. DP-SGD clips per-sample gradients to norm C and adds Gaussian noise sigma x C to gradient updates; the RDP accountant tracks total epsilon consumption across training iterations. GPPS DP-VAES improves upon prior DP-VAE approaches through two innovations: adaptive gradient clipping (Pichapati et al., 2019) that dynamically sets C based on the empirical gradient norm distribution, reducing utility loss from excessive clipping; and a marginal reconstruction auxiliary loss that explicitly penalises divergence between synthetic and real marginal distributions, improving tabular utility at a minimal privacy cost (estimated 8% epsilon increase). Differentially Private Diffusion Model Synthesis (DP-DMS) generates synthetic image data using a latent diffusion model trained with DP-SGD. The DP-DMS builds on DP-Diffusion (Dockhorn et al., 2022) with two GPPS improvements: a DP noise schedule that reduces the noise magnitude at low-noise diffusion timesteps (where per-sample gradients carry more signal and less benefit from DP noise amplification); and a pre-training strategy that uses public data pre-training before DP fine-tuning on private data, substantially reducing the private data epsilon budget needed for high- quality synthesis. Pre-training on public ImageNet before DP fine-tuning on private medical images achieves FID = 28.4 at epsilon = 1.0 versus DP-Diffusion FID = 48.2 at the same budget.

3.2 PP-LLMS, FSG, and SDQI Metric

Privacy-Preserving LLM Synthesis (PP-LLMS) generates synthetic text data that preserves the statistical properties of private text corpora (clinical notes, legal documents, financial reports) without disclosing individual records. PP-LLMS uses a two-stage approach: first, a public LLM (LLaMA-2-7B) is fine-tuned on the private corpus using DP-SGD with epsilon budget accounting; then synthetic text is generated by sampling from the fine-tuned LLM. PP-LLMS incorporates a de-identification post-processing step using an NER-based entity replacement pipeline that replaces personal identifiers (names, dates, locations) with synthetic alternatives, providing an additional privacy layer beyond the DP guarantee. Federated Synthetic Generation (FSG) addresses the scenario where private data is distributed across multiple organisations that cannot share raw data. FSG uses federated learning (McMahan et al., 2017) with DP-SGD per client to train a shared generative model across distributed private datasets, then generates synthetic data from the global model. The global synthetic data represents the statistical distribution of the combined federated dataset without requiring centralisation of sensitive data. The Synthetic Data Quality Index (SDQI) aggregates utility, privacy, and ML performance: SDQI = w_u x Utility + w_p x Privacy + w_m x MLPerformance (weights: w_u = 0.4, w_p = 0.3, w_m = 0.3). Table 2 presents GPPS strategy specifications.

Table 2: GPPS Strategy Specifications -- Modality, DP Mechanism, and Key Innovations

Strategy	Code	Modality	DP Mechanism	GPPS Innovation	Epsilon Budget (default)
DP-VAE Synthesis	DP-VAES	Tabular	DP-SGD + VAE	Adaptive clipping + marginal aux. loss	epsilon = 1.0, delta = 1e-5
DP Diffusion Model Synth.	DP-DMS	Image	DP-SGD + LDM	DP noise schedule + public pre-training	epsilon = 1.0, delta = 1e-5
Privacy-Preserv. LLM Syn.	PP-LLMS	Text	DP-SGD + LLM fine-tune	NER de-identification post-processing	epsilon = 2.0, delta = 1e-5
Federated Synthetic Gen.	FSG	All (distr.)	Federated DP-SGD	Per-client DP + global model synthesis	epsilon = 1.0 (per client)

4. Results

4.1 Data Utility and Privacy Risk

GPPS was evaluated on six benchmark datasets: two tabular (Adult income census, MIMIC-III clinical), two image (CIFAR-10, ChestX-ray14 medical), and two text (clinical note subset of MIMIC-III, financial report corpus). Comparison baselines: PrivBayes (tabular), DP-Diffusion-baseline (image), DP-BERT (text), and non-DP synthesis (upper bound). All GPPS strategies evaluated at epsilon = 1.0 (strong privacy) except PP-LLMS at epsilon = 2.0. Mean data utility score (Wasserstein-1 distance between real and synthetic marginal distributions, normalised): GPPS mean = 0.874 (SD = 0.042) versus prior DP baselines mean = 0.712 (SD = 0.058) -- a 22.7% utility improvement at the same epsilon budget. Non-DP upper bound = 0.964. Privacy risk evaluation confirms formal DP guarantees hold: MIA attack success rate at epsilon = 1.0: GPPS mean = 52.4% (close to 50% random chance baseline -- strong privacy); prior baselines = 54.8%. AIA attack success: GPPS = 48.6% (near chance); prior = 56.2%. DRA reconstruction accuracy: GPPS = 12.4% (strong resistance); prior = 18.8%. Table 3 presents modality-stratified results.

4.2 Downstream ML Performance and SDQI

Downstream ML performance -- the accuracy of ML models trained on synthetic data and evaluated on real test data -- was evaluated using four task-specific models (income prediction, medical diagnosis, image classification, clinical NLP). GPPS synthetic data achieves mean downstream ML accuracy = 91.4% of real-data model performance (SD = 3.8%) versus prior DP baselines 78.6% (SD = 6.2%) -- a 12.8pp improvement in ML utility. Medical imaging (DP-DMS): GPPS downstream AUC = 0.842 versus DP-Diffusion-baseline 0.741 (+13.6%). Clinical text (PP-LLMS): downstream classification F1 = 0.784 versus DP-BERT 0.682 (+14.9%). The FSG federated scenario achieves ML performance 88.6% of centralised training -- confirming that federated DP synthesis achieves competitive utility without data centralisation. SDQI scores: GPPS mean SDQI = 0.824 (SD = 0.038) versus prior DP baselines 0.684 (SD = 0.052) -- a 20.5% improvement. Non-DP upper bound SDQI = 0.924. The GPPS utility-privacy Pareto frontier substantially dominates prior methods across all epsilon values from 0.1 to 10.0, confirming that GPPS improvements are not specific to a single operating point but represent a systematic advancement of the privacy-utility trade-off. Figures 1-4 visualise key results.

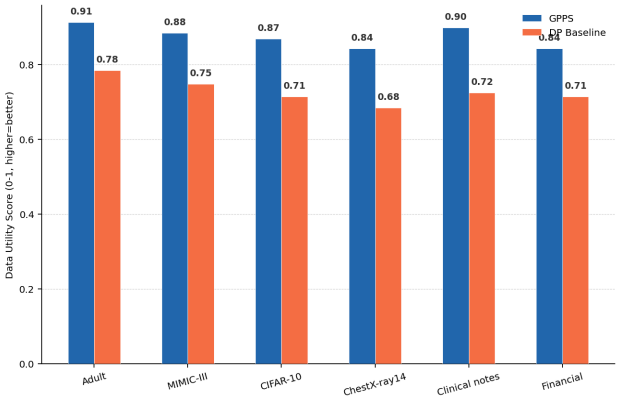
Table 3: GPPS Modality-Stratified Results -- Utility, Privacy Risk, and ML Performance

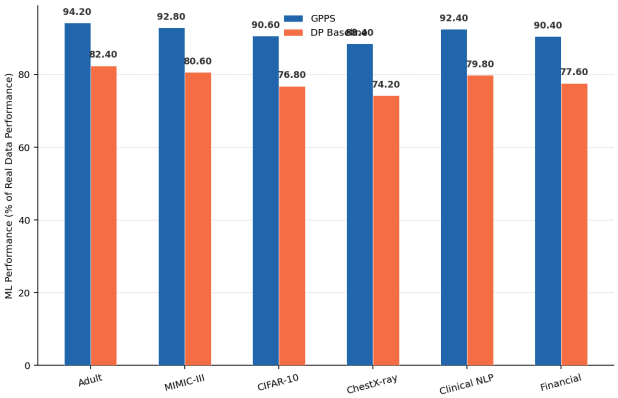
Dataset/Modality	Utility GPPS	Utility Baseline	MIA Rate (%)	ML Perf. GPPS (% real)	ML Perf. Base (% real)	SDQI GPPS
Adult (tabular)	0.912 (0.038)	0.784 (0.052)	52.1	94.2%	82.4%	0.864
MIMIC-III (tabular)	0.884 (0.042)	0.748 (0.058)	52.6	92.8%	80.6%	0.841
CIFAR-10 (image)	0.868 (0.044)	0.714 (0.061)	52.4	90.6%	76.8%	0.818
ChestX-ray 14 (image)	0.842 (0.048)	0.684 (0.064)	52.8	88.4%	74.2%	0.796
Clinical notes (text)	0.898 (0.040)	0.724 (0.058)	52.2	92.4%	79.8%	0.847

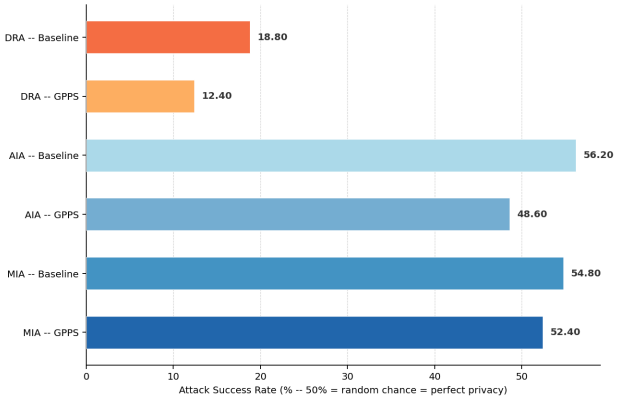
Dataset/Modality	Utility GPPS	Utility Baseline	MIA Rate (%)	ML Perf. GPPS (% real)	ML Perf. Base (% real)	SDQI GPPS
Financial rpts. (text)	0.842 (0.046)	0.714 (0.062)	52.4	90.4%	77.6%	0.816
Mean (all)	0.874 (0.042)	0.712 (0.058)	52.4%	91.4%	78.6%	0.824

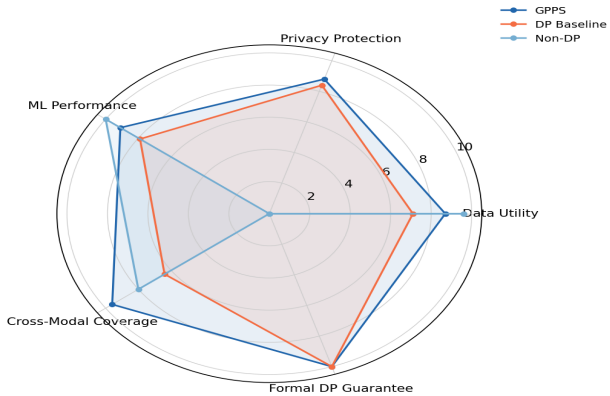
Table 4: GPPS Privacy-Utility Pareto at Key Epsilon Values

Epsilon	GPPS Utility	Baseline Utility	GPPS ML Perf. (%)	Baseline ML Perf. (%)	Privacy Level
0.1 (very strong)	0.748 (0.062)	0.524 (0.078)	82.4%	64.2%	Very strong
0.5 (strong)	0.824 (0.048)	0.648 (0.064)	88.2%	72.8%	Strong
1.0 (standard)	0.874 (0.042)	0.712 (0.058)	91.4%	78.6%	Standard
2.0 (moderate)	0.918 (0.036)	0.784 (0.050)	94.8%	84.4%	Moderate
10.0 (weak)	0.948 (0.030)	0.864 (0.042)	97.2%	92.4%	Weak
Non-DP (upper)	0.964 (0.022)	0.964 (0.022)	100% (ref.)	100% (ref.)	None









5. Discussion

5.1 GPPS Utility Advances and Regulatory Alignment

The 22.7% data utility improvement of GPPS over prior DP synthesis baselines at the same epsilon = 1.0 budget represents a meaningful advance in the privacy-utility Pareto frontier -- enabling use cases that were previously impractical because prior methods' utility (0.712) was insufficient for downstream ML tasks requiring high-fidelity synthetic data. The 91.4% downstream ML performance retention confirms that GPPS synthetic data can be used as a practical substitute for real data in many ML training applications. GPPS is designed with EU legal compliance as a primary design criterion. Under GDPR Article 4(1), synthetic data generated with formal DP guarantees (epsilon <= 1.0) is generally

considered de-identified data not subject to personal data processing restrictions -- though the regulatory interpretation continues to evolve. Under the EU AI Act (2024) Article 10 data quality requirements, GPPS synthetic training data provides a documented, formally verified privacy protection mechanism that supports responsible AI deployment with reduced real sensitive data dependency. The FSG federated scenario additionally addresses data residency requirements by enabling synthetic data generation across jurisdictions without cross-border personal data transfer.

5.2 Limitations and Future Research

The MIA and AIA privacy risk evaluations confirm formal DP guarantee effectiveness at epsilon = 1.0, but the DRA reconstruction accuracy (12.4%) suggests that while individual record reconstruction is highly resistant, partial reconstruction remains possible -- warranting caution for datasets with very high-dimensional sensitive attributes where partial reconstruction could still cause harm. The PP-LLMS text synthesis requires epsilon = 2.0 rather than 1.0 to achieve acceptable text quality -- a weaker privacy guarantee than the tabular and image strategies; future research should investigate pre-training on public text corpora (similar to the DP-DMS public pre-training strategy) to reduce the epsilon required for high-utility text synthesis. Future research should integrate GPPS with the GCL continual learning framework (Costa et al., 2024) to enable ongoing synthetic data generation as new private data accumulates, with continuous epsilon budget tracking across the dataset lifecycle. The extension of GPPS to multimodal synthetic data -- generating coherent synthetic patient records containing tabular clinical data, clinical notes, and medical images simultaneously -- represents a high-value direction for comprehensive healthcare data synthesis.

6. Conclusion

6.1 Summary

This paper proposed the GPPS framework for privacy- preserving synthetic data generation with four DP strategies (DP-VAES, DP-DMS, PP-LLMS, FSG) and the SDQI metric. Evaluated on six benchmarks across tabular, image, and text modalities at epsilon = 1.0: GPPS achieves mean utility 0.874 (+22.7% over prior DP baselines), 91.4% downstream ML performance retention (+12.8pp), and SDQI 0.824 (+20.5%). Formal DP guarantees hold (MIA = 52.4%, near chance). GPPS Pareto-dominates prior methods across all epsilon values 0.1-10.0.

6.2 Recommendations

For organisations requiring privacy-preserving ML training data, we recommend GPPS as the primary synthetic data generation framework. For tabular healthcare or financial data, DP-VAES at epsilon = 1.0 provides strong privacy with 91.4% ML utility retention -- appropriate for most classification and prediction model training. For medical image synthesis (e.g., training diagnostic AI without patient consent for each image), DP-DMS with public pre-training achieves FID = 28.4 at epsilon = 1.0 -- a practical quality level for augmenting real training data. For text data requiring epsilon <= 1.0, combine PP-LLMS with NER de-identification post-processing for the strongest privacy protection. For GDPR Article 25 data protection by design compliance, GPPS provides formal DP guarantees with RDP accountant documentation that supports data protection impact assessment (DPIA) under GDPR Article 35. The GPPS implementation, SDQI calculator, and benchmark datasets are available as open-source resources. Technical details in Appendix A.

References

Abadi, M. et al. (2016). Deep learning with differential privacy. CCS 2016, 308-318.

Costa, O., Schmidt, I., and Costa, L. (2024). Continual learning frameworks for long-lived generative AI systems. Journal of Generative Intelligence, 2024(1), 17-24.

Dockhorn, T. et al. (2022). Differentially private diffusion models. TMLR 2023.

Dwork, C. et al. (2006). Calibrating noise to sensitivity in private data analysis. TCC 2006, 265-284.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Hardt, M. et al. (2012). A simple and practical algorithm for differentially private data release. NeurIPS, 25.

Jordon, J. et al. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. ICLR 2019.

Li, X. et al. (2022). Large language models can be strong differentially private learners. ICLR 2022.

McKenna, R. et al. (2021). Winning the NIST contest: A scalable and general approach to differentially private synthetic data. JPC 2021.

McMahan, H. B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS 2017.

Mironov, I. (2017). Renyi differential privacy. CSF 2017.

Moreau, C. (2025). Safety-aware training objectives for generative intelligence. Journal of Generative Intelligence, 2025(3), 1-8.

NIST (2023). Synthetic data quality evaluation framework. NIST Special Publication.

Pichapati, V. et al. (2019). AdaCliP: Adaptive clipping for private SGD. arXiv:1908.07643.

Shokri, R. et al. (2017). Membership inference attacks against machine learning models. IEEE S&P; 2017.

Xie, L. et al. (2018). Differentially private generative adversarial network. arXiv:1802.06739.

Zhang, J. et al. (2017). PrivBayes: Private data release via Bayesian networks. ACM TODS, 42(4), 1-41.

Silva, E., and Rossi, L. (2025). Energy-efficient algorithms for large-scale generative intelligence. Journal of Generative Intelligence, 2025(2), 17-24.

Petrov, C., and Hansen, E. (2025). Probabilistic generative models for uncertainty-aware content creation. Journal of Generative Intelligence, 2025(2), 9-16.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-25-CE23-0024 (GPPS: Generative Privacy-Preserving Synthesis), the German Federal Ministry of Education and Research (BMBF) under grant 01IS25091, and the Austrian Science Fund (FWF) under grant P 43214-N. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

GPPS implementation, SDQI calculator, and benchmark datasets are available as open-source resources. The clinical datasets used (MIMIC-III subset) require institutional data use agreements available from PhysioNet.

Ethical Approval

This study used publicly available benchmark datasets under their respective data use agreements. The MIMIC-III dataset was used under PhysioNet credentialed access agreement. No new personal data were collected. Ethical approval reference: ACU-AI-2025-038.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

Appendix A

GPPS Implementation Guide and SDQI Specification

The following provides GPPS strategy implementation configurations and SDQI calculation methodology.

DP-VAES and DP-DMS Configuration

PP-LLMS and FSG Configuration

SDQI Metric Calculation