# A Multi-Layer Architecture for Progressive Privacy Risk Mitigation in AI Applications

**Leonardo Bruscagini de Lima**[1], **Luca Misi Rocha**[1],
**Javier Martinez Silva**[1], **Alexandre Melo Braga**[1]

[1] CPQD, Rua Dr. Ricardo Benetton Martins, 1.000, Campinas - SP – Brazil

{leolima, lucamisi, javiers, ambraga}@cpqd.com.br

***Abstract.*** *AI applications process sensitive data and remain exposed to inference and reconstruction attacks. Privacy-Enhancing Technologies (PETs) are often deployed in isolation, allowing risks to accumulate across AI pipelines. This paper presents a Multi-Layer Architecture for Progressive Privacy Risk Mitigation that integrates semantic PII sanitization, structured anonymization, Differential Privacy, adversarial validation, and governance auditing. Experiments with Microsoft Presidio, ARX/PyCANON, and Opacus are combined with a reproducible pipeline that preserves layer interfaces and privacy metrics. Results show improved PT-BR PII detection, utility-aware anonymization, and lower inference and reconstruction exposure.*

## 1. Introduction

Modern AI applications continuously process sensitive information from users, sensors, APIs, enterprise systems, and distributed infrastructures. Foundation Models, Large Language Models (LLMs), and Edge AI increase privacy exposure because they operate on heterogeneous data and may retain latent information about training records [Bommasani et al. 2021]. Under centralized or Federated Learning settings, direct identifiers, quasi-identifiers, and latent sensitive attributes may be exploited through advanced inference attacks [Carlini et al. 2021]. Machine Learning models can also memorize training samples, enabling Membership Inference, Model Inversion, Attribute Inference, and Deep Leakage from Gradients (DLG) [Fredrikson et al. 2015, Shokri et al. 2017, Zhu et al. 2019].

Traditional mechanisms such as masking, tokenization, and identifier removal are insufficient against linkage and reconstruction attacks. Although Privacy-Enhancing Technologies (PETs) have evolved, semantic sanitization, structured anonymization, and Differential Privacy are usually evaluated as isolated controls. Previous studies on Differentially Private Statistics also show that privacy guarantees must be calibrated against utility and operational constraints [Paixão et al. 2025b]. Consequently, residual vulnerabilities may propagate between ingestion, anonymization, model training, and deployment.

This paper proposes a Multi-Layer Architecture for Progressive Privacy Risk Mitigation in AI Applications. The architecture coordinates semantic PII sanitization, structured anonymization, Differential Privacy, adversarial validation, and governance auditing within a unified Privacy-by-Design workflow. Its evaluation combines full-stack experiments using Microsoft Presidio, ARX/PyCANON, and Opacus with a lightweight executable notebook that preserves layer interfaces, parameter propagation, and privacy metrics.

The contribution is an operational and reproducible architecture that: (i) defines explicit input, transformation, output, and audit interfaces for five privacy layers; (ii) quantifies privacy–utility trade-offs using PT-BR text and real public census datasets; and (iii) validates progressive risk propagation through adversarial and governance indicators.

## 2. Related Work

The General Data Protection Regulation (GDPR) and the Lei Geral de Proteção de Dados (LGPD) accelerated the adoption of PETs, Privacy-by-Design methods, and privacy governance frameworks [Brazil 2018, Cavoukian 2011]. The NIST Artificial Intelligence Risk Management Framework (AI RMF) emphasizes continuous monitoring, accountability, and defense in depth throughout the AI lifecycle [National Institute of Standards and Technology 2023]. Architectures for privacy protection are also relevant to distributed and Federated Learning settings, where controls must cover data preparation, optimization, and model exchange [Kaissis et al. 2020].

Semantic privacy in unstructured data is commonly implemented through Named Entity Recognition (NER) [Ratinov and Roth 2009]. Microsoft Presidio is a PII analysis and anonymization tool [Microsoft 2023]; spaCy is an NLP library that provides language models and entity-recognition components [Explosion AI 2026]. PT-BR deployments require additional recognizers and checksum validation for identifiers such as CPF and CNPJ.

Structured privacy relies on k-anonymity [Sweeney 2002], $\ell$-diversity [Machanavajjhala et al. 2007], and t-closeness [Li et al. 2007]. These models are often implemented with multidimensional partitioning, including Mondrian [Lefevre et al. 2006]. Their protection can cause substantial information loss when quasi-identifier groups are sparse or highly dimensional [Aggarwal 2005].

In Machine Learning, Differential Privacy limits statistical inference through gradient clipping, calibrated noise, and privacy accounting [Dwork et al. 2006, Dwork and Roth 2014, Abadi et al. 2016]. Opacus operationalizes these mechanisms for PyTorch [Yousefpour et al. 2021]. Open-source DP studies further demonstrate the practical need for reproducible privacy–utility evaluation [Paixão et al. 2025b, Paixão et al. 2025a]. However, workflows coordinating semantic, tabular, model, adversarial, and governance controls remain comparatively underexplored.

## 3. Proposed Multi-Layer Privacy Architecture

The proposed architecture models privacy protection as a sequential composition of five complementary controls. Let $D_0$ denote the original data and let $L_i$ be the transformation implemented by layer $i$. Each layer is defined as

$$(D_i, M_i) = L_i(D_{i-1}; \theta_i), \quad i \in \{1, \ldots, 5\}, \tag{1}$$

where $\theta_i$ contains the layer configuration, $D_i$ is the protected artifact forwarded downstream, and $M_i$ contains privacy, utility, and traceability metadata. The final audit record is $A = \{M_1, \ldots, M_5\}$, enabling cross-layer verification rather than independent PET execution.

**Threat model and trust assumptions.** The architecture considers adversaries with auxiliary information for record linkage, black-box access to model confidence scores for Membership Inference, or access to shared gradients in distributed training. The attacker may correlate outputs produced at different stages, but cannot modify the trusted recognizer configuration, pseudonymization secrets, privacy accountant, or append-only audit store. The architecture therefore addresses data disclosure and inferential leakage, while compromise of trusted storage, malicious code execution, and denial-of-service attacks remain outside the present scope.

Figure 1 summarizes the data flow, mechanisms, and outputs. The vector diagram retains only the components needed to understand layer composition and the cross-cutting principles applied throughout the workflow.

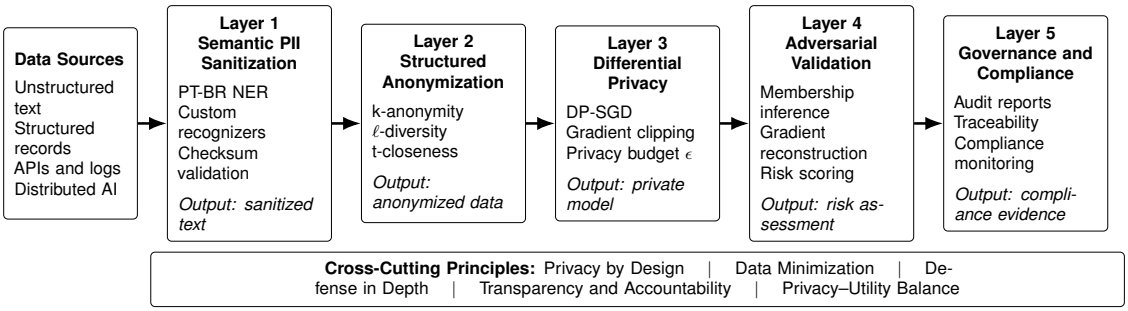


**Figure 1. Proposed multi-layer architecture and its cross-layer privacy controls.**

Layer 1 receives unstructured text and applies Presidio-compatible PT-BR NER, custom recognizers, regex validation, and deterministic replacement for CPF, e-mail, telephone, address, and personal-name entities. Layer 2 receives structured records and applies k-anonymity, $\ell$-diversity, and t-closeness using ARX/PyCANON-compatible hierarchies and metrics. Layer 3 receives anonymized training data and applies DP-SGD, per-sample clipping, Gaussian perturbation, and privacy accounting. Layer 4 consumes model outputs or gradients and estimates Membership Inference and reconstruction exposure. Layer 5 aggregates $M_i$ into audit records, parameter traces, risk indicators, and compliance evidence.

The layer contract enforces three invariants. First, raw inputs are not reintroduced after sanitization or anonymization. Second, each downstream decision must reference the metadata produced by its predecessor. Third, the governance layer must retain the parameterization and result of every applied control. These invariants support repeatable execution and prevent a local privacy metric from being interpreted as an end-to-end guarantee.

## 4. Experimental Evaluation

The evaluation uses two complementary modes. Full-stack experiments assess each technical layer with its native libraries, while a lightweight notebook validates end-to-end sequencing and audit propagation. Adult contains 30,162 census records and ACS PUMS 2022 contains a 120,000-record sample from California and Texas. These are real public census datasets. The semantic experiment uses a fixed-seed synthetic PT-BR corpus generated with Faker; representative entities include CPF, e-mail, telephone, postal-address,

and personal-name examples. Exact scripts, seeds, and configurations are available in the artifact repository.

**Evaluation questions and metrics.** The study addresses four questions: whether PT-BR customization reduces residual PII (**RQ1**); how stronger anonymity constraints affect utility across datasets of different sizes (**RQ2**); whether DP-SGD reduces inference and reconstruction exposure compared with heuristic pruning (**RQ3**); and whether parameters and evidence propagate across all five layers (**RQ4**). Layer 1 uses Precision, Recall, and F1-score. Layer 2 reports normalized Information Loss (IL), where larger values indicate greater utility degradation. Layer 3 uses predictive accuracy, MIA AUC, privacy budget $\epsilon$, and categorical DLG risk. Layers 4–5 record normalized risk and audit-generation status.

### 4.1. Layer 1: Semantic PII Sanitization

The first experiment compared the default spaCy PT-BR small model with a customized large-model configuration enhanced by recognizers and checksum validation. The fixed-seed corpus contains representative valid and invalid identifiers, allowing checksum-aware recognition to be distinguished from regex matching alone.

**Table 1. PII detection performance.**

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| spaCy PT-BR (sm) | 1.0000 | 0.9423 | 0.9703 |
| Custom PT-BR (lg) | 1.0000 | 0.9605 | 0.9799 |

Table 1 shows that the customized pipeline preserved Precision while improving Recall and F1-score. The result supports RQ1 by reducing the probability that undetected PII propagates to downstream layers, although it does not establish generalization beyond the tested PT-BR entity patterns.

### 4.2. Layer 2: Structured Data Anonymization

Adult uses age and education number as numerical quasi-identifiers and marital status, race, and sex as categorical quasi-identifiers; its sensitive target indicates annual income above USD 50,000. ACS PUMS uses a comparable quasi-identifier structure and *Class of Worker* as the sensitive attribute. Privacy constraints were strengthened through k-anonymity, $\ell$-diversity, and t-closeness.

**Table 2. Information loss across anonymization scenarios.**

| Dataset | Records | $k$ | $\ell$ | $t$ | IL |
|---|---|---|---|---|---|
| Adult | 30,162 | 5 | 2 | 0.40 | 0.1607 |
| Adult | 30,162 | 25 | 2 | 0.25 | 0.2564 |
| ACS PUMS | 120,000 | 25 | 6 | 0.20 | 0.0935 |
| ACS PUMS | 120,000 | 50 | 7 | 0.15 | 0.3516 |

Table 2 confirms the expected privacy–utility trade-off within each dataset. Adult loses more utility as constraints become stricter. ACS PUMS is more resilient under the moderate setting because its larger sample supports denser equivalence classes; the

aggressive ($k = 50, \ell = 7, t = 0.15$) setting nevertheless increases information loss substantially. Therefore, RQ2 is dataset-dependent rather than globally monotonic across datasets with different distributions.

### 4.3. Layer 3: Differential Privacy

The third experiment compared baseline training, parameter pruning, and DP-SGD under the same prediction task. Evaluation metrics were predictive accuracy, Membership Inference Attack (MIA) AUC, privacy budget $\epsilon$, and DLG reconstruction risk.

**Table 3. Differential Privacy protection against adversarial attacks.**

| Configuration | $\epsilon$ | Accuracy | MIA AUC | DLG Risk |
|---|---|---|---|---|
| Baseline model | – | 0.66 | 0.71 | Medium |
| Pruning defense | – | 0.65 | 0.71 | Medium |
| DP-SGD (Opacus) | 3.0 | 0.50 | 0.56 | Low |

As shown in Table 3, pruning did not reduce MIA exposure, whereas DP-SGD moved attack AUC toward random guessing and lowered reconstruction risk, with a measurable accuracy cost. This supports RQ3 and illustrates why model sparsification should not be treated as a formal privacy guarantee.

### 4.4. Layer 4: Adversarial Validation

Layer 4 evaluates whether previous controls remain effective against downstream inference. The notebook consumes MIA AUC and gradient-reconstruction indicators, normalizes them into risk levels, and records the evidence used for each classification. The validation separates the attack measurement from the mitigation layer: DP-SGD changes the model-training process, whereas Layer 4 verifies whether the resulting exposure falls within the configured risk policy. These proxy indicators validate orchestration and metric propagation; they do not replace a production-scale red-team environment.

### 4.5. Layer 5: Governance and End-to-End Validation

Layer 5 aggregates layer parameters, metrics, risk levels, and applied controls into a machine-readable audit report. Each record contains the execution identifier, dataset or artifact identifier, layer configuration, measured result, risk classification, and timestamp. This enables traceability from the raw-data transformation to the final privacy decision and provides evidence for operational monitoring.

**Table 4. End-to-end validation metrics.**

| Layer | Metric | Result |
|---|---|---|
| Semantic sanitization | Residual PII | 0 |
| Structured anonymization | Minimum $k$ | 5 |
| Differential Privacy | Simulated $\epsilon$ | 3.0 |
| Adversarial validation | Reconstruction risk | Low |
| Governance | Audit report | Generated |

Table 4 demonstrates that outputs and audit metadata were propagated across all five layers, supporting RQ4. The result validates the architecture's operational composition, while the isolated full-stack experiments provide the layer-specific evidence.

### 4.6. Cross-Layer Discussion and Limitations

The experiments show that no individual metric represents the privacy of the entire AI workflow. A high PII F1-score does not prevent linkage over structured attributes; k-anonymity does not constrain model memorization; and a finite privacy budget does not provide governance evidence by itself. The architecture addresses this gap by preserving the output and metadata of each layer for subsequent validation.

The current evidence must nevertheless be interpreted within its scope. Residual PII equal to zero refers to the fixed synthetic corpus, not every Portuguese-language domain. Information Loss depends on the selected hierarchies and dataset distributions. The adversarial layer uses reproducible indicators and categorical reconstruction risk rather than an exhaustive attack suite. Finally, an audit report provides traceability, but does not by itself demonstrate legal compliance. These limitations motivate evaluation with governed organizational datasets and production-scale red-team exercises.

## 5. Conclusion and Future Work

This paper presented a Multi-Layer Architecture for Progressive Privacy Risk Mitigation that coordinates semantic sanitization, structured anonymization, Differential Privacy, adversarial validation, and governance auditing. The experiments showed improved PT-BR PII detection, dataset-dependent anonymization trade-offs, and reduced inference and reconstruction exposure under DP-SGD. The end-to-end notebook further demonstrated metric propagation and auditable control orchestration.

Adult and ACS PUMS provide real public census evidence, whereas the PT-BR semantic corpus and some end-to-end adversarial indicators are synthetic or proxy-based. Future work will evaluate governed organizational datasets, production-scale red-team attacks, Federated Learning, Homomorphic Encryption, multimodal data, and transformer-based models.

## Artifact Availability

The implementation artifacts, custom recognizers, experimental configurations, and Jupyter notebooks are publicly available in the https://github.com/leobrusc/sbseg2026-progressive-privacy-risk-mitigation.

## Acknowledgment

This work is supported by the Ministry of Science, Technology and Innovation (MCTI), with resources from the National Fund for Scientific and Technological Development (FNDCT) administered by FINEP, within the INSPIRE projects – National Data Infrastructure, Core, and AI Solutions for Public Service (Contract 01.25.0728.00, Reference 2315/25), in partnership with the Ministry of Management and Innovation (MGI). The views expressed are solely those of the authors.

## Use of Generative AI

ChatGPT (OpenAI) was used exclusively to assist with LaTeX formatting, section organization, table and figure placement, and Overleaf editing. It was not used to generate datasets, source code, experiments, results, scientific claims, or validation procedures. The authors reviewed the complete manuscript and assume responsibility for its originality, correctness, and integrity.

## References

Abadi, M. et al. (2016). Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. *Proceedings of the VLDB Conference*, pages 901–909.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. Technical report, Stanford Center for Research on Foundation Models (CRFM).

Brazil (2018). Lei no 13.709, de 14 de agosto de 2018: Lei geral de proteção de dados pessoais (lgpd).

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security '21)*, pages 2633–2650. USENIX Association.

Cavoukian, A. (2011). Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284.

Dwork, C. and Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science.

Explosion AI (2026). spacy: Industrial-strength natural language processing in python.

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333.

Kaissis, G. A., Makowski, M. R., Rückert, D., and Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311.

Lefevre, K., DeWitt, D., and Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *International Conference on Data Engineering*.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *International Conference on Data Engineering*.

Machanavajjhala, A. et al. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1).

Microsoft (2023). Presidio: Pii data protection and anonymization sdk.

National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (ai rmf 1.0).

Paixão, A. C. P., Camargo, G. F. L., and Braga, A. M. (2025a). Testing open-source libraries for private counts and averages on energy metering time series. In *20th European Dependable Computing Conference*, pages 100–104.

Paixão, A. C. P. et al. (2025b). Understanding how to use open-source libraries for differentially private statistics on energy metering time series. In *Proceedings of the 10th International Conference on Internet of Things, Big Data and Security (IoTBDS)*, pages 289–296. SCITEPRESS.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Conference on Computational Natural Language Learning*.

Shokri, R. et al. (2017). Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

Yousefpour, A. et al. (2021). Opacus: User-friendly differential privacy library in pytorch.

Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems*.