Project title	**ACHILLES Human-Centred Machine learning lighter, clearer, safer**		
Project acronym	ACHILLES		
GA number	101189689		
Project start date	01/11/2024	Duration	48 months

D2.3 - PRIVACY-PRESERVING LEARNING REPORT V1

Due date	30/04/2026	Delivery date	30/04/2026
Work package	WP2		
Responsible Author(s)	Sînică Alboaie (AXL)		
Contributor(s)	Sînică Alboaie (AXL), Maximilian Andreas Hoefler (FhHHI), Darlene MacDonald (CuomoIT), Filipe Soares (AICOS), Pablo Accuosto (ETICAS), Mariana Carvajal Sojo (ETICAS)		
Reviewer(s)	Rui Castro (FhAICOS), Pablo Accuosto (ETICAS), André Carreiro (FhAICOS)		
Version	V1		
Dissemination level	Public		

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

VERSION AND AMENDMENTS HISTORY

Version	Date (DD/MM/YYYY)	Created /Amended by	Changes
V0.1	09/04/2026	AXL	First draft
V0.2	15/04/2026	All task partners	Final review
V0.3	22/04/2026	All contributors	Review and fixes
V0.4	28/04/2026	All contributors and reviewers	Final draft
V1.0	29/04/2026	André Carreiro (FhAICOS)	Final revision and clean-up

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

TABLE OF CONTENTS

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

LIST OF TABLES

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

LIST OF ABBREVIATIONS

AEPD	Agencia Española de Protección de Datos (Spanish Data Protection Agency)
AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
APDP	Associação Protectora dos Diabéticos de Portugal
API(s)	Application Programming Interface(s)
AutoFedAvg	Automatic Federated Averaging
AWS	Amazon Web Services
COML	Collaborative Machine Learning without Centralized Training Data
CVM	Confidential Virtual Machine
DFL	Decentralised Federated Learning
DP	Differential Privacy – data privacy protection technique in machine learning
DP-FedAvg	Differentially Private Federated Averaging
DP-SGD	Differentially Private Stochastic Gradient Descent
DPS	Data Processing Space is a concept proposed during the initial research in Task 2.3. The name was subsequently changed to DPU (Data Processing Unit)
DPU	Data Processing Unit
DPU Agent	Data Processing Unit Agent– agent mediating access to private and protected resources
DSU	Data Sharing Unit
Dx.y	Deliverable x.y
EDPS	European Data Protection Supervisor

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

EC	European Commission
FATE	Federated AI Technology Enabler
FedAvg	Federated Averaging
FedBN	Federated Batch Normalisation
FedKT-CSD	One-shot differentially private federated learning method (FhHHI)
FedML	Federated Machine Learning
FedNova	Federated Normalised Averaging
FedProx	Federated Proximal
FedSGD	Federated Stochastic Gradient Descent
FedSyn-Refine	LLM-seeded federated learning with differentially private synthetic pretraining data (FhHHI)
FedXDS	Federated learning method by FhHHI for attribution-guided counteraction of data heterogeneity
FedYogi	Federated Yogi Optimiser
FL	Federated Learning – distributed learning without centralizing data
FL-LR	Federated Learning Log & Replay (INESC ID)
GDPR	General Data Protection Regulation
HDC	Hyperdimensional Computing – neuro-symbolic computing method using high-dimensional vectors
HE	Homomorphic Encryption – encryption that allows computation on encrypted data
HERA	Holistic Evaluation and Regulatory Adherence
HHI	(Fraunhofer) Heinrich Hertz Institute

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

HLEG	High-Level Expert Group
ICCV	International Conference on Computer Vision
ID	Identity Document
IDE	Integrated Development Environment – ACHILLES prototyping environment
IDRiD	Indian Diabetic Retinopathy Image Dataset
ISRUC	Research institute or organisation that contributes (pseudonymised) datasets for facial recognition
KeySSI	Key Self-Sovereign Identity – mechanism for identification and access control over digital artifacts
LLM	Large Language Model
MCP	Model Context Protocol– mechanism for exposing functionalities of Agents
ML	Machine Learning – statistical learning techniques
MLOps	Machine Learning Operations
MRP-VM	Neuro-symbolic execution model under development in T3.3, called “MRP Virtual Machine”
NIS2	Network and Information Security Directive 2
NVFLARE	NVIDIA Federated Learning Application Runtime Environment
OAuth	Standard protocol for authorisation and access control
OCR	Optical Character Recognition
OECD	Organisation for Economic Co-operation and Development
PA	Privacy-Aware – aware of data privacy / protection
PET	Privacy-Enhancing Technology
Ploinky	Container-oriented orchestration substrate for agents developed in T3.3

Funded by the European Union
This project has received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No 101189689.

QM	Quality Management
QMS	Quality Management System
RAG	Retrieval-Augmented Generation
R&D	Research & Development
REFUGE	Retinal Fundus Glaucoma Challenge
SA	Security-Aware – aware of security / secure execution
SCRIPTA	Structured Creative Writing Intelligent Platform for Textual Authoring
SERMAS	Servicio Madrileño de Salud
SMPC	Secure Multi-Party Computation
SOPs	Standard Operating Procedures
SSD	Semantic Specification Documents
SSO	Single Sign-On – unified authentication mechanism for distributed platforms
TCB	Trusted Computing Base
TEE(s)	Trusted Execution Environment(s)
UDC	University of A Coruña, Spain
UNESCO	United Nations Educational, Scientific and Cultural Organisation
VM	Virtual Machine
VSA	Vector Symbolic Architecture – neuro-symbolic architecture inspired by distributed vector representations
VSS	Visual Secret Sharing
WACV	Winter Conference on Applications of Computer Vision
WPx	Work Package x

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

1 EXECUTIVE SUMMARY

This deliverable presents the first project-wide baseline for privacy-preserving learning and confidentiality-aware AI support within ACHILLES. Its purpose is to consolidate the technical directions, methods, and partner assets that are currently most relevant to the project use cases and to the progressive integration of such capabilities into the ACHILLES platform environment, including ACHILLES IDE and the broader agentic execution architecture.

The report covers the main families of techniques that are most relevant at this stage of the project, including federated learning, differential privacy, secure computation, confidential computing, synthetic-data-related approaches, and supporting mechanisms for auditability and controlled execution. Rather than promoting a single privacy-preserving stack, the deliverable adopts a use-case-oriented perspective and identifies the combinations of methods that are most appropriate under different operational, regulatory, and architectural conditions.

A central contribution of this first version is to organise these technical directions in a form that is directly usable by the project. The deliverable relates privacy-preserving methods to the requirements emerging from the identity verification, healthcare, HERA, and SCRIPTA use cases; it documents the main tools and assets currently available across the consortium; and it clarifies how these capabilities can progressively become operational through integration into the ACHILLES environment. In this sense, the report serves both as a technical reference for ongoing development and as a baseline for the next phase of implementation and validation.

The document also introduces the Data Processing Unit (DPU) Agent as an operational integration pattern for capabilities that require stronger control conditions, such as protected data access, confidential processing, restricted retrieval, secure execution, or policy-aware exposure of specialised functions. In ACHILLES, privacy-preserving methods are not expected to operate in isolation; they must be deployed within larger workflows that also require governance, access control, auditability, and integration with platform components. The DPU Agent is therefore presented as a practical architectural mechanism that can support the deployment of privacy- and confidentiality-sensitive capabilities across different project settings.

Overall, this deliverable establishes a coherent first-phase basis for privacy-preserving work in ACHILLES. It identifies the methods and assets that are currently most relevant, connects them to the needs of the use cases and the platform architecture, and prepares the ground for more targeted integration, validation, and exploitation activities in the second part of the project.

2 INTRODUCTION

This document addresses privacy-preserving machine learning and related privacy- and confidentiality-aware technical mechanisms as enabling layers for trustworthy AI within ACHILLES. The relevance of this area is driven both by the broader regulatory context—particularly the increasing emphasis on

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

trustworthy and human-centric AI in Europe (EC, 2025; EC, 2026)—and by the practical need to support AI workflows over data that cannot be freely centralised, exposed, or shared (OECD, 2025).

Within ACHILLES, this topic is approached in an applied, use-case-driven manner. Task 2.3 is not intended as a standalone line of fundamental research. Its primary role is to support the needs emerging from the project's use cases and the surrounding technical work, in particular the multi-agent execution and orchestration directions described in (D3.3), as well as the integration and platform developments in WP6, including ACHILLES IDE (D3.3; D6.1). In this first phase, the work has therefore focused on privacy-preserving and confidentiality-related questions that are currently most relevant to these research and development lines. In this context, it is important to distinguish between privacy in the strict sense and confidentiality in a broader operational sense. Privacy is particularly relevant when AI systems handle personal data or other information subject to data protection requirements. Confidentiality, on the other hand, concerns protected enterprise data, trade secrets, restricted corpora, security-sensitive assets, and other non-public resources. While these dimensions must be distinguished from a legal and governance perspective, they often overlap significantly at the technical level. In practice, privacy-preserving AI cannot be implemented effectively without also considering secure execution, controlled access, limited disclosure, and the protection of confidential resources. For this reason, the scope of this deliverable extends beyond privacy in the narrow sense, while remaining fully aligned with the practical objectives of Task 2.3.

The ACHILLES proposal already recognises that no single technique is sufficient on its own. Federated learning helps keep data at the source, but model updates may still leak sensitive information. Differential privacy provides formal guarantees, though often at the cost of utility. Techniques such as synthetic data, anonymisation, and pseudonymisation can reduce the direct exposure of personal information, but their effectiveness depends on the context, data modality, and threat model. ACHILLES adopts a layered approach in which multiple techniques may need to be combined depending on use-case requirements, infrastructure maturity, and the acceptable balance between privacy, utility, efficiency, and operational complexity (Carreiro, 2024; EDPS-AEPD, 2025; OECD, 2025). Against this background, Task 2.3 focuses on the distributed and computational aspects of privacy-preserving and confidentiality-aware AI. Its scope includes not only the identification of relevant techniques, but also the study of how training, aggregation, inference, validation, and related computations can be organised when data must remain under local control and collaboration is required without unrestricted centralisation. This makes federated learning, secure computation, and related approaches particularly relevant, while also highlighting the need for reusable integration and execution patterns tailored to the ACHILLES environment. This deliverable addresses both privacy-preserving methods and the conditions under which such methods may later become operational within ACHILLES. Within this context, the DPU Agent is introduced. While the DPU Agent is not a privacy-preserving technique, it serves as a protected execution template associated with the Ploinky orchestration environment and related settings, such as the ACHILLES IDE. Its role is to support the controlled exposure of protected skills, access to private or confidential scopes, centralised secret management, and enhanced auditability and data locality (MCP, 2025; D6.1). In this sense, the DPU

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Agent reflects the integration needs emerging from D3.3 and WP6 rather than a standalone scientific objective.

This deliverable should therefore be read as an overview and positioning document for the privacy-preserving work currently relevant to ACHILLES. It reviews the technical directions that are most applicable at this stage, with particular attention to federated learning, secure computation, and related operational patterns, and relates them to the current architecture and implementation roadmap.

In this first phase, the main focus of Task 2.3 has been to identify the technical directions relevant to privacy and confidentiality most applicable to ACHILLES, to structure them in relation to the evolving architecture, and to prepare provisional integration patterns and requirements to guide subsequent implementation and validation. This first version of D2.3 provides:

i. a structured overview of privacy-preserving learning techniques relevant to ACHILLES;
ii. an initial mapping of these techniques to the four ACHILLES use cases;
iii. a catalogue of consortium assets relevant to federated learning, secure computation, confidential execution, privacy assessment, and controlled disclosure;
iv. an initial operational integration pattern, through DPU Agents, for exposing protected capabilities in ACHILLES environments; and
v. a baseline for the more implementation- and validation-oriented update in D2.4.

Together, these contributions support the role of Task 2.3 as a transversal task serving the project use cases, the D3.3 research direction, and the implementation paths being developed in WP6. The subsequent D2.4 report will be the appropriate point to report more mature evidence on selected integrations, use-case validation, and comparative performance, where implementation and pilot conditions have stabilised.

3 IMPORTANCE OF PRIVACY FOR AI

3.1 Privacy as a Core Ethical Principle for AI

The protection of privacy is widely recognised as a foundational ethical principle for the development and deployment of artificial intelligence. The right to privacy, together with the right to the protection of personal data, is enshrined in the Charter of Fundamental Rights of the European Union (Articles 7 and 8), the European Convention on Human Rights (Article 8), and international instruments such as the International Covenant on Civil and Political Rights. These rights provide essential prerequisites for the exercise of other fundamental freedoms, including freedom of expression, assembly, and religion.

In the context of artificial intelligence, the High-Level Expert Group on Artificial Intelligence (AI HLEG), established by the European Commission, identified privacy and data governance as one of seven key requirements for trustworthy AI. The AI HLEG's Ethics Guidelines for Trustworthy AI state that

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

AI systems must guarantee privacy and data protection throughout their entire lifecycle and that adequate data governance mechanisms must be implemented to ensure data quality and integrity. These guidelines, together with the associated Assessment List for Trustworthy Artificial Intelligence (ALTAI), have informed the EU AI Act and serve as a reference framework for the ACHILLES project.

Privacy is also recognised as a core principle in other major international policy frameworks. The OECD Principles on AI and the UNESCO Recommendation on the Ethics of Artificial Intelligence both emphasise the importance of privacy protection in AI systems. A widely cited review of over 80 AI ethics documents identified privacy as one of five common denominator principles across global AI governance frameworks (Jobin et al., 2019).

Within ACHILLES, these ethical commitments are operationalised through the project's alignment with Horizon Europe's ethical framework. As outlined in D4.4 – Ethics Guidelines v1, Horizon Europe requires that AI systems respect fundamental rights, including human dignity, privacy, and non-discrimination, and that projects integrate human oversight, fairness and bias mitigation, and transparency throughout AI development. ACHILLES applies the ALTAI framework and the Z-Inspection Process for Trustworthy AI (led by ARCADA) to ensure that these principles are systematically embedded across the project lifecycle.

3.2 Privacy Risks from Data Collection to Deployment

The development and deployment of AI systems create privacy risks at every stage, from data collection through to model deployment and beyond. These risks arise from the scale and nature of data processing involved in modern machine learning, which often requires large volumes of data - including personal or sensitive information - for training, testing, and deployment. The widespread use of AI has introduced new challenges in the collection, storage, and processing of personal data, often in complex and opaque ways, posing significant risks to privacy and data protection rights.

Privacy risks begin at the point of data collection and preparation. Personal data may be included in training datasets, sometimes without adequate consent or legal basis. Even when consent has been obtained for a specific purpose, the data may later be used in ways not originally intended, raising questions of purpose limitation and further processing under GDPR (Article 6(4)). The principle of data minimisation requires that only data strictly necessary for the intended purpose be collected; however, in practice, additional information is often gathered. Moreover, combining different and separately anonymised datasets can enable the re-identification of individuals, undermining the protections initially assumed. As highlighted in Navigating the Privacy Maze (Carreiro, 2024), three major challenges arise in the context of AI and data privacy: data persistence (data may exist long after the subjects who generated it), data repurposing (data may have value beyond its original scope), and data spillovers (data collection may impact unintended individuals who did not provide consent).

During model training and testing, privacy risks arise from several sources, including inadequate access controls, unintended data leakage during the training process, and the retention of personal data beyond the period necessary for the intended purpose. The machine learning process

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

itself can encode information about the training data into model parameters, creating potential avenues for subsequent extraction. As noted by the European Data Protection Supervisor (EDPS) (EDPS, 2025.), AI systems incorporating algorithms for continuous learning introduce additional privacy concerns, as operational data may be used for ongoing training without the same level of oversight applied during initial development.

Once a model is deployed, privacy risks manifest as data leakage through model outputs or inference attacks. Membership inference attacks can reveal whether specific individuals' data were included in the training set, while model inversion attacks can reconstruct aspects of the training data from the model's outputs. Even partial leakage may enable re-identification of individuals, depending on the context in which the AI system operates. Furthermore, using model outputs in ways not anticipated during training may violate the original purpose for which personal data was collected and consented, raising concerns under GDPR principles such as purpose limitation.

3.3 The Legal Framework for Privacy in AI

In the European Union, the protection of privacy in the development and deployment of AI systems is primarily operationalised through the General Data Protection Regulation (GDPR). The GDPR establishes a comprehensive framework of principles and obligations for the processing of personal data, including requirements for lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity, and confidentiality. A detailed analysis of these GDPR requirements and their application to the ACHILLES use cases is provided in the companion deliverable D2.1 (GDPR Compliance Report v1). The GDPR operates alongside other EU digital regulations that are relevant to the privacy aspects of AI development. The EU AI Act establishes risk-based obligations for AI systems, including specific requirements for high-risk applications, such as those involving biometric identification and health-related data, both of which are pertinent to ACHILLES use cases. The Data Act and the Data Governance Act define the conditions for data sharing and reuse, while the NIS2 Directive and the Cyber Resilience Act impose cybersecurity obligations that intersect with data protection requirements. References to legal and regulatory instruments in this deliverable are used to frame technical requirements and risk considerations. They do not replace the detailed legal analysis provided in the relevant D4.1 (Legal and Ethical Mapping) and D2.1 deliverables.

3.4 The Gap Between Ethics and Law

Despite the increasingly detailed legal framework, a gap persists between legal requirements and the ethical aspirations for AI, particularly as technology continues to evolve. The GDPR was originally designed for traditional data processing scenarios, and its application to modern machine learning raises questions that are not yet fully resolved: for example, the legal status of model updates exchanged in federated learning settings, or the extent to which differential privacy mechanisms can be considered sufficient to meet the GDPR's anonymisation requirements.

Emerging technologies introduce new challenges that existing legal frameworks are still adapting to. For example, the increasing use of AI agents acting on behalf of individuals raises complex

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

questions regarding consent management, purpose limitation, and data sharing; issues that go beyond what current regulations explicitly address. While a user may grant an AI agent broad access to their personal data, defining all circumstances in which the agent should restrict access to third parties remains a challenge that neither current ethical guidelines nor legal requirements fully resolve.

Within ACHILLES, this gap is recognised as an ongoing area of attention. The project's approach, combining technical privacy-preserving measures (Section 4) with legal compliance frameworks (D2.1) and ethical oversight (D4.4), aims to address both the letter and the spirit of privacy protection, acknowledging that responsible AI development requires going beyond minimum legal compliance.

4 LITERATURE REVIEW ON PRIVACY-PRESERVING TECHNIQUES

4.1 Overview and Taxonomy of Privacy-Preserving Techniques

The development and deployment of AI systems increasingly rely on large volumes of data, much of which may contain personal or sensitive information. As discussed earlier, this creates significant privacy risks across the AI lifecycle. In response, a variety of privacy-preserving techniques have emerged, each addressing different aspects of these challenges. These techniques can be broadly organised according to how and where data is processed. Following the taxonomy proposed by Carreiro (2024), four main approaches can be distinguished according to how the data is handled:

- **Data kept at source**: Techniques such as Federated Learning, in which models are trained locally and only model updates, not raw data, are shared between participants.
- **Data processed while encrypted**: Techniques such as Homomorphic Encryption and Secure Multi-Party Computation, which enable computations on data without revealing the underlying information.
- **Data partially changed**: techniques such as Differential Privacy, which adds calibrated noise to data or model outputs to prevent identification of individuals, as well as pseudonymisation and anonymisation methods that modify personally identifiable information.
- **Data completely replaced**: Techniques such as synthetic data generation, which create artificial datasets intended to reproduce relevant statistical or structural properties of the original data while reducing direct exposure of real records. Their privacy value depends on the generation process, similarity controls, leakage testing, and the applicable threat model.

These approaches are not mutually exclusive. The limitations of individual privacy-preserving methods make a multi-faceted strategy preferable, combining the strengths of different techniques while considering the required level of privacy in a specific context and the utility of the transformed data. This is the approach adopted within ACHILLES, where WP2 integrates GDPR compliance tools (Task 2.1), synthetic data generation (Task 2.2), and federated learning with secure computation strategies (Task 2.3).

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

The following sections focus on the privacy-preserving techniques most relevant to the ACHILLES project, in particular federated learning and secure computation. Synthetic data generation and anonymisation techniques are primarily addressed in the companion deliverable D2.1 (GDPR Compliance Report v1), and their relevance to privacy-preserving learning is also noted.

4.2 Federated Learning

4.2.1 Fundamentals

Federated Learning (FL) is a machine learning approach in which multiple data sources - such as devices or organisations - collaboratively train a shared model while keeping data decentralised. Rather than transmitting raw data to a central server, each participant processes its own data locally and shares only model updates, such as gradients or weights.

This approach is particularly relevant in scenarios where data sensitivity or regulatory requirements make data centralisation impractical – for example, in healthcare, where patient data is subject to strict legal frameworks, or in cross-organisational settings, where competitive or legal constraints limit data.

Federated Learning can be implemented with or without a central server. In the centralised variant, a server coordinates the process by distributing an initial model, collecting local model updates, aggregating them into a global model, and redistributing it. In the decentralised variant, known as Decentralised Federated Learning (DFL), participants exchange model parameters directly through peer-to-peer networks without a central coordinator (EDPS & AEPD, 2025).

4.2.2 Classification of Federated Learning Systems

Federated Learning (FL) systems can be classified along two dimensions (EDPS & AEPD, 2025; Ursache, 2025).

By data distribution:

- **Horizontal FL**: participants hold data with the same features but different samples - for example, hospitals in different regions each holding medical records with the same structure but for different patients.
- **Vertical FL**: participants hold data about the same entities but with different features - for example, a bank and a hospital holding different types of information about the same individuals.
- **Transfer FL**: a model is adapted to different contexts with different data distributions, not just different samples.
- **Hybrid FL**: a combination of the above approaches.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

By participant type:

- **Cross-device FL**: participants are individual users with personal devices (smartphones, wearables) in large numbers. Data is limited and dynamic. This suits scenarios like personalised model adjustment.
- **Cross-silo FL**: participants are organisations (hospitals, banks) in smaller numbers, each holding large datasets. Data is more stable and structured.

The distinction between cross-silo and cross-device settings has implications for privacy, communication efficiency, and security. Cross-device FL faces challenges with network stability and device security, while cross-silo FL can generally be managed through asynchronous learning and more robust infrastructure (EDPS & AEPD, 2025).

4.2.3 Privacy Benefits of Federated Learning

From a data protection perspective, Federated Learning offers several advantages over centralised approaches. By keeping personal data within local environments and sharing only model updates, FL is consistent with the GDPR principle of data minimisation. It can also reinforce accountability by allowing data controllers to maintain clearer oversight of processing activities under their responsibility. In scenarios involving special categories of data (such as medical records), FL reduces the risks associated with centralising sensitive information and can support more favourable outcomes in Data Protection Impact Assessments. FL also offers practical benefits for consent management, as data subjects retain better control over their personal data when it remains on their devices (EDPS & AEPD, 2025).

4.2.4 Privacy Risks and Limitations

Despite its advantages, FL does not eliminate privacy risks. Key concerns include (EDPS & AEPD, 2025):

- **Data leakage through model updates**: even without direct access to raw data, attackers may infer sensitive information by analysing the gradients or weights exchanged between participants. This enables membership inference attacks (determining whether specific data points were in the training set) and model inversion attacks (reconstructing aspects of the training data).
- **Security across the ecosystem**: security must be implemented across all participating devices and communication channels. An attacker compromising the weakest link can potentially compromise the whole system. Local models, trained on limited data, are particularly susceptible to revealing characteristics of their training data.
- **Data quality and bias**: in a federated setting, checking data quality is more difficult as data sources cannot be directly compared. Ensuring that the overall model is free from bias requires distributed quality management procedures and careful monitoring of statistical distributions across participants.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Integrity threats**: FL architectures are vulnerable to data poisoning (injecting false data into the training process) and model poisoning (modifying local model updates). Active defences that detect and eliminate poisoned models are currently the most promising mitigation approach.

It should not be assumed that model updates exchanged in Federated Learning constitute anonymous data; a careful technical and legal assessment must be conducted on a case-by-case basis to evaluate the risks associated with the shared information (EDPS & AEPD, 2025).

4.2.5 Federated Learning in the ACHILLES Context

Within ACHILLES, Federated Learning is being integrated into the project's development environment through two complementary approaches: the ACHILLES IDE, developed by Axiologic, and the COML platform, developed by FhAICOS. The ACHILLES IDE is an evolution of AssistOS, an open-source platform that provides collaborative workspaces for AI-assisted workflows. Unlike traditional development environments focused primarily on writing and debugging code, the ACHILLES IDE is organised around Specification Documents - interactive blueprints that guide the development process and integrate AI agents into the workflow (Ursache, 2025). To support FL, the IDE is being extended with new SSD types that structure and configure the distributed training process, covering model definition, training configuration, secure aggregation planning, and deployment strategy.

A key architectural concept underpinning FL in the ACHILLES IDE is the Data Processing Space (DPS) - a self-contained, Docker-based execution environment designed to enable secure, decentralised data processing (Ursache, 2025). DPS enables computation at the data source, avoiding the need to transfer sensitive data to centralised infrastructure and thereby supporting the fundamental federated learning principle of data locality. Access control is enforced through cryptographic identifiers - KeySSIs (OpenDSU, 2026), providing fine-grained and dynamic authorisation mechanisms, while inter-workspace communication is managed via secure APIs supporting model distribution, update collection, aggregation, and policy enforcement. The DPS concept was designed to support multiple deployment configurations, including cloud-based infrastructures, blockchain-integrated architectures, and environments compliant with European Data Spaces requirements, depending on the regulatory and data sovereignty constraints of each use case (Ursache, 2025). Following architectural extensions introduced in Task 3.3, the DPS concept has evolved and been redefined as the Data Processing Unit (DPU), reflecting its integration into the agent-based system architecture. In this context, a DPS is considered a specific instantiation of a DPU Agent within the broader system. The “DPS” terminology is being progressively phased out in favour of “DPU” to ensure conceptual and architectural consistency. However, the term is retained in this document for alignment with earlier work (Ursache, 2025), which has already been disseminated within the consortium.

In preparing the integration of FL capabilities, Ursache (2025) reviews several established open-source FL frameworks—including Flower, NVFLARE, PySyft, FedML, FATE, TensorFlow Federated, and Fed-BioMed—analysing their architectures, supported aggregation strategies, privacy-enhancing features (such as differential privacy, secure aggregation, and homomorphic encryption), and

Funded by the European Union
This project has received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No 101189689.

suitability for different deployment scenarios. These frameworks provide the building blocks that the ACHILLES IDE aims to integrate, making FL accessible to users with varying levels of technical expertise while ensuring compliance with data protection requirements.

Alongside the ACHILLES IDE, FhAICOS has developed COML (Collaborative Machine Learning without Centralised Training Data), a comprehensive FL platform built on top of Nvidia NVFlare. COML provides aggregation scheme selection and federated evaluation, with multiple aggregation strategies implemented and tested, a data analyser and partitioner, and MLOps components for job configuration, orchestration, and monitoring. The platform supports both edge and cloud deployments and is designed to be interoperable with the broader ACHILLES toolkit.

The specific tools, aggregation strategies, and partner contributions to the FL toolkit within ACHILLES are described in detail in Section 6. The roadmap for their development and integration is presented in Section 7.

4.3 Secure Computation Techniques

To address the residual privacy risks in federated learning, particularly data leakage through model updates, FL is often combined with additional security-enhancing techniques. These techniques can also be applied independently in scenarios requiring privacy-preserving computation (Carreiro, 2024; EDPS & AEPD, 2025)

4.3.1 Secure Multi-Party Computation (SMPC)

Secure Multi-Party Computation is a cryptographic protocol that enables multiple parties to jointly compute a function over their inputs while keeping those inputs private. No participant learns anything about the other parties' data beyond what can be inferred from the output. In the context of Federated Learning, SMPC can be used to aggregate local model updates without revealing individual contributions to the central server or other participants (Carreiro, 2024; EDPS & AEPD, 2025).

A common approach within SMPC is secret sharing, in which each participant splits their input into multiple shares distributed across the other parties. The computation is then performed on these shares in a way that produces the correct aggregated result without any single party being able to reconstruct another's original input. This makes SMPC particularly suitable for scenarios requiring multi-institutional collaboration - such as cross-organisational medical research or financial services - where data cannot be shared directly but joint analysis is needed (Carreiro, 2024). The main limitation of SMPC is its computational and communication overhead, which increases with the number of participants and the complexity of the function being computed, making it challenging to scale to large, federated settings (EDPS & AEPD, 2025).

4.3.2 Homomorphic Encryption

Homomorphic Encryption allows computations to be performed directly on encrypted data, producing an encrypted result that, when decrypted, matches the outcome of operations performed on

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

the raw data. This offers strong security guarantees, as data never needs to be decrypted during processing. However, this currently incurs significant computational overhead, limiting its practical applicability in some scenarios (Carreiro, 2024; Sereno, 2025). Within ACHILLES, FhAICOS has successfully integrated Homomorphic Encryption into its COML federated learning platform using the Nvidia Flare framework.

4.3.3 Trusted Execution Environments and Confidential Computing

Trusted Execution Environments (TEEs) use hardware-level isolation to create protected computing environments where data can be processed securely, offering a practical alternative to purely algorithmic approaches such as HE and SMPC. While TEEs rely on a larger Trusted Computing Base (TCB) than cryptographic techniques, they can support real-world workloads - including AI model training and inference - at performance levels that algorithmic approaches currently cannot achieve (Sereno, 2025). Confidential Virtual Machines (CVMs) extend this concept by protecting entire VM memory spaces from the hypervisor, enabling privacy-preserving computation in cloud environments without requiring application-level modifications (Misono, 2024; Sereno, 2025). A key trust mechanism is remote attestation, which allows data owners to cryptographically verify the state and integrity of the computing environment before provisioning sensitive data.

However, CVM deployments by major cloud providers present practical challenges. (Sereno, 2025) found that implementations across Google Cloud, Microsoft Azure, and AWS are inconsistent in their trust models and attestation mechanisms, with none fully leveraging the security capabilities offered by the underlying hardware. To address this, (Sereno, 2025) propose the Evident framework, which provides unified CVM lifecycle management, automating remote attestation and making the trust model transparent for each deployment. For federated learning scenarios, Evident supports attested workflows in which model owners and data owners can verify that only authorised components are running before sharing models or data. Within ACHILLES, INESC-ID is developing a prototype to integrate CVMs into the AssistOS platform, enabling privacy-preserving machine learning in cloud environments and supporting federated learning workflows. A platform-independent attestation mechanism is being developed to address the transparency challenges identified above, and an early version of a CVM-based plug-in has been built to support secure federated learning tasks across different AI models. The Evident framework, developed partly under the ACHILLES grant, represents the research output of this line of work.

4.4 Differential Privacy

Differential Privacy (DP) is a mathematical framework that protects the privacy of individuals within a dataset by introducing carefully calibrated random noise into data, queries, or model outputs (Dwork, 2006). The core guarantee is that the inclusion or exclusion of any single individual's data does not significantly affect the result of a computation, thereby limiting what can be inferred about any specific person. This guarantee is formally expressed through a "privacy budget" parameter (ε): lower values of ε provide stronger privacy protection but require more noise, while higher values allow greater accuracy

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

at the cost of weaker privacy guarantees. DP mechanisms can be applied at different stages - either locally (noise added before data leaves the owner's control) or globally (noise added to aggregated outputs) - and through different techniques such as Laplace noise addition, Gaussian mechanisms, and DP-SGD for deep learning (Alzoubi & Mishra, 2025).

The main challenge of DP is the inherent trade-off between privacy and utility: adding more noise provides stronger privacy guarantees but can reduce the accuracy and usefulness of the resulting model or analysis. Finding the optimal balance remains an active area of research, alongside concerns about computational overhead, the specialist expertise required to calibrate privacy budgets appropriately, and the fact that regulatory frameworks have not yet fully standardised how DP guarantees map to legal compliance requirements (Alzoubi & Mishra, 2025).

In the context of privacy-preserving machine learning, DP is most relevant as a complement to federated learning. By adding noise to local model updates before they are shared with the aggregation server, DP reduces the risk of membership inference or gradient-based reconstruction attacks (EDPS & AEPD, 2025; Ursache, 2025). Several approaches integrate DP into FL protocols: DP-FedAvg adds noise during the federated averaging process, while more recent methods explore using differentially private synthetic data to address data heterogeneity across participants (Ursache, 2025). This combination is important for compliance with data protection regulations such as the GDPR, as it limits the risk of re-identifying specific data points while still enabling meaningful model learning.

Within ACHILLES, differential privacy is being explored in combination with FL through several complementary approaches. FhHHI is investigating compressed model update techniques where quantisation noise inherently contributes to differential privacy guarantees. FhHHI has also developed approaches using differentially private synthetic data generation to address data heterogeneity in federated settings, including one-shot FL mechanisms that provide protection against membership inference attacks. This line of research has been disseminated through a publication at ICCV 2025 (Hoefler, 2025). Additionally, standard differential privacy–enabled federated learning baselines, such as DP-FedAvg, are considered as reference implementations within the ACHILLES project toolkit.

4.5 Synthetic Data Techniques

Synthetic data generation creates artificial datasets that mimic the statistical properties of real data, allowing model development and testing without exposing sensitive information (Carreiro, 2024; Facoco, 2025). While a detailed treatment of synthetic data techniques falls within the scope of D2.1 and especially the future D2.2, it is worth noting their relevance to privacy-preserving learning as well.

Within ACHILLES, synthetic data serves a dual purpose: as a privacy-preserving alternative to real data for model training (Task 2.2), and as a mechanism for addressing data heterogeneity in federated learning settings: for example, through differentially private synthetic data generation to improve convergence and personalisation across federated participants.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

4.6 Remaining Challenges and Open Questions

Despite significant progress, several challenges remain in the field of privacy-preserving machine learning. Perhaps the most fundamental is that no single technique addresses all privacy concerns. Each of the approaches described in this section has distinct strengths and limitations, and a multi-layered strategy combining several PETs is generally necessary - though this adds complexity to system design and deployment (Carreiro, 2024; EDPS & AEPD, 2025).

A recurring theme across techniques is the tension between privacy and utility. Differential privacy degrades model accuracy as noise increases (Alzoubi & Mishra, 2025), while cryptographic methods such as homomorphic encryption and secure multi-party computation impose computational overhead that limits their practical scalability (Sereno, 2025). Hardware-based approaches like CVMs offer better performance but introduce their own trust challenges, as current cloud provider implementations do not yet fully leverage the security capabilities of the underlying hardware (Sereno, 2025). Finding the right balance between privacy protection and data or model utility for each specific context remains an active area of research.

Distributed settings also raise challenges that centralised approaches do not face. In federated learning, data quality cannot be directly verified across participants, bias is harder to detect and mitigate, and the architecture is vulnerable to poisoning attacks that require continuous monitoring and active defences (EDPS & AEPD, 2025). At the same time, the threat landscape continues to evolve, and membership inference, model inversion, and gradient reconstruction attacks grow more sophisticated alongside the defences designed to counter them (Alzoubi & Mishra, 2025; EDPS & AEPD, 2025).

From a regulatory perspective, while techniques like FL align well with GDPR principles such as data minimisation, the legal status of model updates and resulting models with respect to personal data remains subject to case-by-case assessment. The regulatory landscape has not yet fully standardised how the formal guarantees offered by techniques like differential privacy map to legal compliance requirements (Alzoubi & Mishra, 2025; EDPS & AEPD, 2025).

ACHILLES addresses these challenges through its multi-faceted approach, combining GDPR compliance tools, synthetic data generation, and federated learning with secure computation - each reinforcing the others to provide layered privacy protection across the AI lifecycle. The specific tools and strategies being developed within the project are described in Sections 6 and 7.

Beyond the techniques described above, other privacy-preserving approaches are also being explored within ACHILLES. ISRUC is developing a framework based on Visual Secret Sharing (VSS) for privacy-preserving facial recognition in the context of the ID verification use case. This approach splits facial landmark data into encrypted shares distributed across multiple institutions, so that no single entity can reconstruct the original biometric data. These and other partner-specific tools are described in detail in Section 5.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

5 PRIVACY-PRESERVING REQUIREMENTS WITHIN ACHILLES USE CASES

This section summarises the privacy-preserving requirements identified for each of the four ACHILLES use cases, based on the use case definitions established in D7.1 (Use Cases and Evaluation Framework, M12) and the ongoing pipeline specification work in WP7 towards D7.2. For each use case, we describe the data handled, the main privacy concerns anticipated, and the privacy-preserving techniques considered most relevant at this stage. Where D7.1 prioritises "Privacy-preserving Learning" most strongly for specific use cases, D2.3 additionally considers confidentiality-aware execution and controlled access patterns that are relevant to agentic workflows, even when a use case does not yet require a full privacy-preserving-learning pipeline. These descriptions are necessarily preliminary, as pipeline definitions are being refined and category-specific test cases, including privacy and security, are currently being developed across all use cases. As these specifications mature, a more detailed account of how privacy-preserving tools are applied will be provided in D2.4.

5.1 Identity verification

The identity verification use case targets automated remote onboarding through two complementary processes: ID document analysis (classification, OCR, and fraud detection from smartphone-captured images) and biometric verification (face matching between a selfie/video and the document photo, including liveness detection). The data processed includes images of identity documents, facial images, and video selfies. IDnow uses proprietary production data collected under explicit user consent; ISRUC contributes a research face dataset with pseudonymised images and videos. Publicly available and synthetic datasets are also used. Both facial images and biometric templates qualify as special category data under Art. 9 GDPR.

The main privacy concerns are fourfold. First, production data from IDnow's industrial flow is inherently sensitive and subject to strict GDPR retention and purpose-limited constraints, requiring periodic dataset renewal and the deletion of expired samples. Second, the use case explicitly aims to increase adoption of synthetic data to reduce reliance on real personal data, which introduces the challenge of ensuring that synthetic samples are statistically representative without leaking information from the real data from which they are derived. Third, biometric data used for face verification and liveness detection carries elevated re-identification risk. Fourth, on-premises training requirements and the potential use of federated learning across IDnow and ISRUC raise questions about secure model exchange and data governance across organisations.

The most relevant privacy-preserving techniques for this use case are synthetic data generation, data minimisation strategies, and federated learning. The primary operational scenario focuses on progressively replacing real training data with synthetic samples while maintaining classification accuracy, assessed against fully annotated real data benchmarks. A second scenario addresses non-regression testing with GDPR-compliant data renewal cycles, ensuring that expired real data is replaced by newly collected samples whose distributional properties are validated against the original baseline. Federated learning is also relevant for enabling collaboration between IDnow and ISRUC

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

without direct data exchange. The respective data owners hold pseudonymisation keys. Differential privacy may be relevant in later stages to provide formal guarantees when sharing model updates across organisations.

5.2 Healthcare

The healthcare use case develops AI-based diagnostic support for three ophthalmological conditions (glaucoma, age-related macular degeneration, and diabetic retinopathy) from retinal fundus images. The data processed includes retinal images, clinical diagnoses, segmentation masks, and clinical metadata. Both publicly available datasets (e.g. REFUGE, DRISHTI-GS, IDRiD) and private clinical datasets from SERMAS/Hospital Clínico San Carlos (Spain) and APDP (Portugal) are used. All private data qualifies as special category data under Art. 9 GDPR (health data).

The main privacy concerns in this use case are threefold. First, retinal fundus images, while not directly identifying, can in principle be linked to individuals through matching against reference databases, making pseudonymisation rather than full anonymisation the realistic baseline. Second, the use case involves a multi-institutional setting where clinical data from different providers must be used for training and validation without centralising raw patient images. Third, potential memorisation effects in deep learning models could expose information about individual training samples.

The privacy-preserving techniques most relevant to this use case are federated learning and data pseudonymisation. Federated learning is the primary strategy being explored: it enables collaborative model training across institutions (e.g. UDC and FhAICOS each hosting local datasets) while keeping raw images at the source. A first privacy-specific test case has already been defined for this purpose, focusing on federated training of a glaucoma diagnosis model using the REFUGE and DRISHTI-GS datasets distributed across two sites, with centralised training as the performance baseline. Beyond federated learning, pseudonymisation of private datasets is enforced at the source by the clinical data providers, who exclusively hold the re-identification keys. Differential privacy and synthetic data generation are not currently planned for this use case, but may be considered in later iterations if the federated learning evaluation identifies residual privacy risks.

5.3 HERA

The HERA (Holistic Evaluation and Regulatory Adherence) use case applies AI to support regulatory compliance and quality management in the pharmaceutical sector. The system is built around Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) within a multi-agent architecture, providing context-aware document retrieval, automated compliance monitoring, document summarisation, and cross-document analysis. The data processed consists primarily of external regulations and standards, as well as internal company-specific policies, SOPs, and quality management data. This data is generally non-personal; personal data processing occurs only incidentally (e.g. names appearing in internal documents and records).

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Company-specific internal documents (policies, SOPs, audit records) may contain commercially sensitive or confidential information, and ingesting such data into LLM-based pipelines raises concerns about data leakage through model memorisation or inadvertent exposure in generated outputs. As the system matures towards deployment with pharmaceutical companies, multi-tenant scenarios may arise in which strict data segregation across different organisations must be ensured. As CuomoIT is based in Switzerland, a cross-border data transfer dimension must be considered, depending on the client's location, and may be covered by an EU adequacy decision.

The most relevant privacy-preserving techniques for this use case are confidential computing, access control, and data minimisation, rather than federated learning or synthetic data. Secure execution environments and tenant-level data isolation are important for preventing cross-contamination of organisations' confidential documents. Techniques to limit memorisation risks in LLMs (e.g. retrieval-based architectures that avoid fine-tuning on sensitive corpora, input filtering, and output monitoring) are also relevant. At this stage, no federated learning or differential privacy mechanisms are planned; however, these may become relevant if the system ingests company-specific information from legal entities located in different countries or if there is a movement toward collaborative model training across multiple pharmaceutical partners.

A summary of the Relevant Privacy-Preserving Techniques and the Privacy-Preserving Goal, taking into consideration the HERA QM/Partner AI System View, has been provided below. This overview of what must be considered at each stage of the “AI process” will facilitate the next step of mapping relevant ACHILLES Methods/Assets to support achieving the defined Privacy-Preserving Goals.

Table 1 - Privacy Protection Techniques and Objectives (HERA QM/AI View)

HERA QM/Partner AI System View	**Relevant Privacy-Preserving Techniques**	**Privacy-Preserving Goal**
Nature of the Data – External regulations and standards - Internal Policies, Standard Operating Procedures and Work Instructions	Data minimisation; removal/masking of incidental identifiers	Keep processing focused on largely non-personal. Regulatory and QMS content, and reduce the small amount of personal data and commercially sensitive/confidential content present in internal documents.
Ingestion of External Regulations and Internal company-specific documents into RAG/LLM pipelines	Access control on ingestion services; encrypted transfer and storage; restricted ingestion scopes	Protect commercially sensitive and confidential internal documents as they are brought into HERA, and prevent unnecessary expansion of

Funded by the European Union
This project has received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No 101189689.

		the in-scope Regulatory and QMS content
Use of RAG and LLMs for retrieval, summarisation and cross-document analysis	Retrieval-based architecture (avoid fine-tuning on sensitive content); input filtering; prompt-level minimisation	Limit memorisation risk by keeping sensitive content in retrievers rather than model weights, and ensure the LLM only receives the minimal context required per query.
Generation of outputs (Classification, Mapping, Reports Cross-document insights)	Output monitoring and redaction; template-based, privacy-aware reporting	Prevent inadvertent exposure of sensitive/confidential information or incidental personal data in generated text returned to users.
Multi-tenant Deployment across legal entities of a Pharmaceutical Company or Pharmaceutical companies	Tenant-level data isolation; strong role-based access control; organisation-scoped configuration and logging	Ensure strict segregation between different organisations' (or legal entities') content so documents, prompts and outputs cannot cross boundaries.
Cross-border processing (e.g. Swiss provider, EU clients) dependent upon location of pharmaceutical client and content	Data-location transparency; minimisation of cross-border transfers; use of adequacy-covered hosting and standard safeguards	Align with GDPR and data-transfer rules while operating LLM and RAG infrastructure in Switzerland under EU adequacy, limiting unnecessary data movement.
Execution of LLM-based multi-agent architecture	Confidential/secure execution environments; hardened runtime; strict endpoint authentication and authorisation	Prevent unauthorised access to prompts, retrieved context and intermediate representations during AI processing.
Future collaborative model training (consideration for the future)	(Consideration for the future) federated learning across legal entities of a global pharmaceutical company or	If the system later trains shared models across multiple legal entities, it can enable cross-partner learning without centralising all

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

	similar collaborative training approaches	sensitive QMS documents in one place.

5.4 SCRIPTA

SCRIPTA is relevant to Task 2.3 as a use case for controlled generation, policy-aware review, traceability, and governed AI-assisted workflows within the ACHILLES environment. Its current role is not to validate a full privacy-preserving learning pipeline, as in identity verification or healthcare scenarios, but to examine how protected knowledge, proprietary assets, constrained generation processes, and confidential editorial resources can be handled within agentic workflows.

In advanced creative and editorial settings, value depends not only on text generation, but also on provenance, editorial coherence, controlled originality, legal and ethical conformity, revision traceability, and evidence that specific constraints were respected during generation and review. From the perspective of Task 2.3, this makes SCRIPTA relevant to questions of protected knowledge bases, restricted retrieval, controlled access to private or proprietary assets, auditability of agentic interactions, and selective exposure of specialised functions to users or other system components.

These requirements do not necessarily imply the use of federated learning or formal differential privacy in the current phase. They do, however, motivate architectural patterns for controlled access, data minimisation, selective disclosure, protected execution boundaries, and confidentiality-aware management of resources in multi-step AI workflows. As SCRIPTA evolves, relevant techniques may include anonymisation, synthetic data for sensitive editorial materials, protected execution for specialised services, and, where later requirements justify it, differential privacy for selected release or training scenarios (Alzoubi & Mishra, 2025; Dwork, 2006).

In D2.3, SCRIPTA should therefore be understood as a use case that broadens the privacy discussion from personal-data-centric machine learning to the governance of protected assets and trustworthy AI-assisted services. Its practical contribution is to help identify requirements for ACHILLES IDE and related ACHILLES environments concerning controlled capability exposure, auditability, privacy-aware execution, and confidentiality-aware workflow design.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

6 PRIVACY-PRESERVING TOOLS WITHIN ACHILLES

6.1 Existing tools

The ACHILLES framework incorporates a suite of privacy-preserving tools contributed by the partners, collectively covering a broad spectrum of Federated Learning (FL) paradigms, privacy threat models, and deployment configurations.

Table 2 - Privacy-Preserving and Trusted Federated Learning Tools

Method / Asset Name	Short Description	Organisation
COML	Comprehensive FL platform (NVFlare-based) with aggregation scheme selection, differential privacy, homomorphic encryption, encrypted communications, GUI, REST APIs, and 7 FL strategies (FedAvg, FedProx, FedBN, FedSGD, FedYogi, FedNova, AutoFedAvg).	FhAICOS
Confidential VM Framework	Automates hardening, deployment, and remote attestation of Confidential VMs across cloud providers for secure and verifiable AI inference.	INESC ID
FedKT-CSD	One-shot differentially private federated learning for personalisation using tiny pretrained autoencoders. Provides strong membership inference defence	FhHHI
FedSyn-Refine	LLM-seeded federated learning with differentially private synthetic pretraining data generation. Targets text generation and classification tasks.	FhHHI
FedXDS	Attribution-guided data sharing between FL clients to counteract data heterogeneity. Defends against reconstruction attacks (ICCV 2025).	FhHHI
FL-LR	Log-and-replay auditing system for federated learning with defences against model poisoning; enables post-hoc investigation of client behaviour.	INESC ID
Privacy-preserving Sparse Collaborative Inference	Sparsification of intermediate activations for collaborative inference. Defends against reconstruction attacks (WACV 2026).	FhHHI
VeriProv	Provenance-based framework for verifiable, non-repudiable records across the ML lifecycle using trusted execution environments.	INESC ID

A structured view of the Privacy Protection-relevant methods and assets is available in the form of a FactSheet per relevant ACHILLES Asset, providing concise, standardised information on the underlying modules (e.g. TRL, roadmap, maturity), thereby making their contribution to the protection

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

of privacy transparent for the Use Case business, technical, and QA stakeholders. Please refer to the FactSheets in Appendix 1.

Contributions from Fraunhofer HHI

Fraunhofer HHI (Fraunhofer Heinrich-Hertz-Institut) contributes several peer-reviewed methods targeting distinct privacy threat models in distributed learning scenarios. The first, FedXDS: Leveraging Model Attribution Methods to Counteract Data Heterogeneity in Federated Learning (accepted at ICCV 2025), addresses the challenge of statistical data heterogeneity across clients in FL settings while simultaneously providing defences against reconstruction attacks - including the recovery of personal identities and scene-level information from model updates. The second, Leveraging Sparsity for Privacy in Collaborative Inference (WACV 2026), extends privacy protection to the collaborative inference regime, where partial model computations are offloaded across network nodes, and similarly defends against reconstruction of sensitive visual features.

In addition to these reconstruction-defence mechanisms, HHI contributes two one-shot FL approaches grounded in differentially private (DP) synthetic data generation. FedKT-CSD employs tiny pretrained autoencoders to guide one-shot differentially private federated learning, enabling personalisation while maintaining strong formal privacy guarantees against membership inference attacks. Complementing this, FedSyn-Refine leverages large language models seeded within an FL pipeline to generate differentially private synthetic pretraining data, targeting text generation and classification tasks. Both methods offer strong formal privacy guarantees under the differential privacy framework. HHI further contributes Diffusion-guided FL, which employs diffusion models to generate synthetic data that bridges statistical heterogeneity across clients and accelerates convergence. Finally, to ensure methodological rigour and reproducibility, well-established benchmark algorithms are also incorporated as reference baselines, enabling systematic comparison across the privacy-utility trade-off spectrum.

Contributions from Fraunhofer AICOS

Fraunhofer AICOS contributes the Collaborative Machine Learning without Centralised Training Data (COML) platform, a comprehensive and production-ready infrastructure for federated learning, aggregation strategies selection, and federated evaluation of AI models. COML is built on NVFlare, a domain-agnostic and extensible SDK that enables distributed machine learning across geographically and organisationally separated sites without centralised data sharing. The platform integrates a range of security and privacy-enhancing technologies (PETs), including differential privacy, homomorphic encryption, mutual authentication, and end-to-end encrypted communications.

From an MLOps perspective, COML provides a full-featured operational stack that includes: an administrative web API and graphical user interface for job configuration, orchestration, and lifecycle management; a data analyser and partitioner for both real and synthetic data quality assessment; and monitoring and tracking services for visualising training progress and cross-model comparison.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Deployment is supported via both direct and Docker-based installation, making COML suitable for edge and cloud environments alike, covering the entire pipeline from development to production.

COML v3 currently implements seven FL aggregation strategies: AutoFedAvg (Automatic Federated Averaging), FedAvg (Federated Averaging), FedProx (Federated Proximal), FedBN (Federated Batch Normalisation), FedSGD (Federated Stochastic Gradient Descent), FedYogi (Federated Yogi Optimiser), and FedNova (Federated Normalised Averaging). This diversity of strategies enables flexible adaptation to heterogeneous data distributions, varying client participation patterns, and asynchronous communication constraints commonly encountered in real-world federated deployments.

Contributions from INESC ID

INESC ID provides a set of complementary tools that address security, auditability, and trusted execution in federated and cloud-based ML pipelines. VeriProv is a provenance-based framework that captures verifiable and non-repudiable records across all phases of the ML lifecycle, binding each participant to the actions they perform. By leveraging trusted execution environments (TEEs), VeriProv prevents privileged cloud actors from tampering with data or execution environments, providing end-to-end auditability for distributed ML workflows.

Addressing the specific challenges of federated learning integrity, FL-LR (Federated Learning Log & Replay) enables auditing of FL systems that deploy defences against model poisoning. FL-LR allows auditors to re-execute and analyse FL training runs, supporting post-hoc investigation of client behaviour and verification that the produced model has not been compromised. Finally, the Confidential VM Framework provides a unified, auditable approach to managing secure Confidential Virtual Machines (CVMs), automating image hardening, deployment, and remote attestation across cloud providers. This framework supports verifiable trust for third-party ML workloads - including confidential AI inference - ensuring that deployment in public cloud environments remains compatible with the privacy and security requirements of sensitive applications.

6.2 DPU Agents

The privacy-preserving techniques relevant to ACHILLES differ not only in their assumptions and threat models, but also in the operational conditions required for their deployment. Federated learning keeps data at the source; differential privacy reduces inferential leakage; secure multi-party computation and homomorphic encryption protect aspects of distributed computation; confidential computing protects execution environments; and synthetic data can reduce direct reliance on personal information. In practical settings, these techniques usually require orchestration, policy-aware exposure, monitoring, auditability, and disciplined integration into larger workflows (Carreiro, 2024; EDPS & AEPD, 2025; Sereno, 2025).

Within ACHILLES, this integration problem is particularly relevant because Task 2.3 is not centred on a single privacy-preserving method. Different techniques may become relevant depending on the

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

use case, the participating partners, the maturity of the infrastructure, and the deployment model adopted. For this reason, the DPU Agent is introduced here as an architectural pattern for integrating privacy- and confidentiality-sensitive capabilities into the broader ACHILLES environment, especially in connection with the orchestration and platform work described in D3.3 and D6.1 (D3.3; D6.1).

In this deliverable, the DPU Agent should be understood as a protected-execution template associated with Ploinky and related ACHILLES environments, including ACHILLES IDE. Its role is to host selected capabilities that require stronger control conditions than those of ordinary agents, such as restricted retrieval, protected preprocessing, confidential inference, privacy-sensitive validation, controlled access to non-public corpora, or adapters for privacy-enhancing technologies. The DPU Agent is therefore not presented as a privacy-preserving technique. It is the execution and integration layer through which such techniques and capabilities can be packaged, exposed, and governed more coherently inside a modular agentic environment.

This distinction matters for the report's logic. The preceding sections discuss privacy-preserving methods themselves: what they protect, under which assumptions, and with which limitations. The DPU Agent belongs to a different layer. It concerns how selected protected capabilities may be operationalised inside ACHILLES in a reusable, policy-aware, and auditable manner. Its relevance in D2.3 follows directly from the need to connect technical methods to real workflows, especially where non-public resources, protected connectors, or constrained execution conditions are involved.

The proposed pattern is aligned with Ploinky, understood here as a container-oriented orchestration substrate for agent systems. In such an environment, a protected unit may expose only internal APIs, only MCP-compatible tools, or a combination of both, depending on the deployment profile and the sensitivity of the functionality involved. This flexibility is important because privacy-sensitive operations may need different integration modes across use cases and deployment settings (MCP, 2025). A DPU Agent may therefore encapsulate protected state, scoped references to external private systems, deployment-specific policy settings, restricted indexes, connector-specific credentials, and selected skill packs. Its architectural value lies in providing a disciplined boundary for these elements, rather than dispersing privacy-sensitive logic across multiple ordinary agents or services. In this sense, the DPU pattern is best understood as a reusable packaging and governance model for protected capabilities.

This also clarifies how the DPU concept relates to the rest of the deliverable. The privacy-preserving methods reviewed earlier can, where relevant, be mapped to operational capabilities that may later be hosted or mediated through such a protected runtime. Federated learning may require functions for local training orchestration, secure aggregation coordination, federated evaluation, or update handling (EDPS & AEPD, 2025; Ursache, 2025). Differential privacy may require release-control functions, privacy-budget management, or DP-aware training support (Alzoubi & Mishra, 2025; Dwork, 2006). Confidential computing, secure computation, and synthetic-data workflows may likewise require specialised adapters, restricted execution paths, or more controlled backend bindings

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

(Carreiro, 2024; Hoefler, 2025; Misono, 2024; Sereno, 2025). These examples indicate how the architectural pattern can be used; they do not imply uniform maturity across all such capabilities within ACHILLES.

The DPU Agent is already being approached through early iterations in connection with the project's agentic architecture and platform work. At the current stage, these iterations establish the direction and the general execution model, while the surrounding implementation, deployment profiles, policy layer, and capability coverage are still maturing. In this deliverable, the role of the DPU concept is therefore to position a coherent operational model for privacy- and confidentiality-sensitive functionality inside ACHILLES and to connect that model to the use-case and platform needs emerging in the first phase of the project.

6.3 Mapping Tools to Use Cases

The analysed use cases demonstrate that privacy requirements in real-world AI deployments are typically not fully defined at the initial stages, but instead evolve progressively as pipelines mature, data flows are refined, and threat models are systematically clarified. The ACHILLES privacy framework addresses this by organising its tools around a structured mapping between requirements and capabilities. Each deployment scenario is characterised along three axes: the protection level needed (user-level participation privacy, sample-level membership privacy, or reconstruction defence), the threat model assumed (honest-but-curious server, malicious participants, or external adversaries), and the deployment regime (federated learning, collaborative inference, or centralised pipelines with sensitive corpora). Standardised asset fact sheets are used to map the portfolio against these dimensions, providing a clear breakdown of capabilities, limitations, and formal guarantees for each tool. Critically, the framework also encompasses privacy auditing tools that quantify residual leakage empirically — gradient reconstruction tests, membership inference benchmarks, and activation inversion evaluations. This dual emphasis on protection and measurement reflects a core conviction: privacy in AI systems cannot be assured by mechanisms alone but requires continuous empirical verification that deployed defences hold under realistic attack scenarios.

This structure serves two purposes. First, it lets use case owners match known requirements to available tools. Second, it provides a principled response to requirements that are not yet defined. As pipelines evolve and new risks surface — through auditing, through the addition of data partners, or through the transition from research to production — the mapping allows the consortium to identify gaps and compose defences systematically rather than ad hoc. The toolkit thereby functions as a living framework rather than a static catalogue. To support this practically, tools are gathered through a shared repository with standardised metadata, and well-established baselines (e.g. DP-FedAvg, standard secure aggregation) are included as reference points for systematic comparison across the privacy-utility spectrum.

Mapping this to individual use cases reveals distinct requirements and cross-cutting opportunities. Identity verification, with its biometric special-category data and cross-organisational

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

training, requires the broadest combination: synthetic data generation to replace real personal data (Diffusion-guided FL, FedKT-CSD), federated learning with formal DP for secure model exchange, and reconstruction defence for facial images (FedXDS). Healthcare's multi-institutional retinal image analysis maps to federated learning with sample-level privacy, where COML provides the production-ready backbone and HHI's one-shot methods offer communication-efficient alternatives for clinical sites. HERA's LLM-based compliance pipelines face memorisation and cross-tenant leakage rather than gradient attacks, making INESC ID's confidential computing stack and VeriProv the most relevant tools, complemented by retrieval-augmented architectures as an architectural defence against memorisation. SCRIPTA, as an exploratory environment for protected creative workflows, does not yet have fully specified requirements but is positioned as a natural testbed for differential privacy, controlled anonymisation, and the broader framework for matching emerging needs to available defences as the project progresses.

7 ROADMAP FOR TOOLS

7.1 Tool Integration Roadmap

The tools, methods, and architectural patterns described in this deliverable form the current technical basis for the privacy-preserving and confidentiality-aware work relevant to ACHILLES. In the second part of the project, these assets are expected to be improved, extended, and operationalised to the extent that they are integrated into ACHILLES IDE and thereby made available to the use cases. Their evolution is therefore primarily driven by platform integration and concrete requirements emerging from the use cases, rather than by a separate, standalone roadmap internal to Task 2.3.

For this reason, ACHILLES does not define a rigid per-tool development schedule at this stage within T2.3. Some tools may remain primarily partner assets, reference implementations, or research outputs, while others may be selected for deeper integration, interface refinement, validation, and operational exposure inside ACHILLES IDE. In practical terms, candidate tools for deeper integration will be prioritised according to:

i. direct relevance to at least one WP7 use case;
ii. maturity and maintainability of the implementation;
iii. clarity of privacy/security guarantees and assumptions;
iv. compatibility with ACHILLES IDE, Ploinky, or DPU-based execution patterns;
v. availability of suitable evaluation data or test scenarios; and
vi. expected contribution to the D2.4 validation narrative.

7.2 DPU Agents Evolution

Alongside the individual tools described in this report, ACHILLES is also developing a more general integration pattern for protected execution through the DPU Agent. The DPU Agent is introduced

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

as an architectural mechanism through which privacy- and confidentiality-sensitive capabilities can be packaged, exposed, and governed within the broader ACHILLES agentic environment (D3.3; D6.1).

This direction already has early iterations and is being matured in connection with the orchestration and platform work of the project. Its purpose is to provide a more controlled execution boundary for selected capabilities that require stronger policy, audit, and access conditions than ordinary agents, especially where protected data, restricted connectors, confidential state, or controlled output release are involved. In the next phase of the project, the DPU Agent is expected to evolve in step with ACHILLES IDE and with the use cases that require this type of protected operational model. Its maturation, therefore, follows the same general logic as the rest of the tool landscape: progress depends on concrete integration needs, implementation maturity, and the practical value of making protected capabilities reusable inside ACHILLES workflows.

7.3 Emerging Directions

The roadmap outlined above also leaves room for directions that are not yet separate implementation lines but may become relevant as the ACHILLES architecture evolves. In particular, the work reported in D3.3 suggests that some future developments may connect privacy-preserving questions not only to distributed machine learning in the narrow sense, but also to agent orchestration, constrained execution, and emerging neuro-symbolic execution models. In such settings, part of what is treated today as learning may increasingly involve the synthesis, adaptation, and controlled composition of executable structures, which shifts part of the privacy and confidentiality problem toward execution-time protection, controlled disclosure, protected intermediate representations, and governed runtime boundaries (D3.3).

This perspective is one reason why ACHILLES has begun to consider privacy-related questions in connection with VSA/HDC-inspired approaches, hybrid neuro-symbolic directions, and the evolving MRP-VM execution model (D3.3). At the current stage, these directions remain exploratory and do not define a separate committed roadmap within T2.3. Their relevance lies in indicating where additional integration pressure may appear later, depending on platform evolution and use-case needs. The overall project position remains clear: the tools already identified in this deliverable provide the basis for future work, while the depth and sequencing of their evolution will be determined by ACHILLES IDE integration, use-case uptake, and the operational value of the resulting capabilities.

8 CONCLUSION

This deliverable establishes the first consolidated baseline for privacy-preserving learning and confidentiality-aware AI support in ACHILLES. The work carried out so far shows that the project does not require a single uniform privacy-preserving solution, but a structured set of methods, tools, and integration patterns that can be selected and combined according to the characteristics of each use case, the sensitivity of the data involved, and the maturity of the surrounding technical environment. From this perspective, the main result of the present report is threefold. First, it identifies and structures

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

the privacy-preserving techniques most relevant to ACHILLES, with particular attention to federated learning, differential privacy, secure computation, confidential computing, and synthetic-data-related approaches. Second, it maps these techniques to the requirements emerging across the project use cases, recognising that privacy, confidentiality, and controlled execution play different roles in identity verification, healthcare, HERA, and SCRIPTA. Third, it documents the main tools and assets currently available in the consortium and positions their progressive integration into the ACHILLES platform environment.

An important practical conclusion is that the operational value of privacy-preserving methods in ACHILLES depends not only on the underlying techniques, but also on the conditions under which they are deployed. Access control, protected execution, auditability, controlled capability exposure, and integration with broader AI workflows are necessary for these methods to become usable in realistic settings. This is why the deliverable also introduces the DPU Agent as an architectural pattern that can support the governed deployment of privacy- and confidentiality-sensitive capabilities within ACHILLES IDE and related ACHILLES environments.

The next phase of the project will build on this baseline through deeper integration of selected tools, stronger alignment with the evolving use-case pipelines, and more focused validation under realistic operational conditions. The overall direction is clear: ACHILLES will advance those privacy-preserving and confidentiality-aware capabilities that provide concrete value for the project's workflows, strengthen trustworthy AI deployment, and support transfer toward operational applications and partner exploitation paths.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

9 REFERENCES

Alzoubi, Y. I., & Mishra, A. (2025). Differential privacy and artificial intelligence: potentials, challenges, and future avenues. https://doi.org/10.1186/s13635-025-00203-9

Carreiro, A. (2024). Navigating the Privacy Maze in Artificial Intelligence. Center for Responsible AI, Fraunhofer AICOS. https://centerforresponsible.ai/navigating-the-privacy-maze-in-artificial-intelligence/

D2.1 Internal project reference. Deliverable D2.1.

D3.3. Internal project reference. Deliverable D3.3.

D6.1. Internal project reference. Deliverable D6.1.

DSSC-2025. Data Spaces Support Centre. DSSC Blueprint (2025). https://blueprint.dssc.eu/

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. https://link.springer.com/chapter/10.1007/11681878_14

EC-2025. European Commission. AI Act | Shaping Europe's digital future (2025). https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

EC-2026. European Commission. Navigating the AI Act (2026). https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act

EDPS-AEPD-2025. European Data Protection Supervisor, Agencia Española de Protección de Datos. TechDispatch #1/2025: Federated Learning (2025). https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-06-10-techdispatch-12025-federated-learning_en

EDPS-2025. European Data Protection Supervisor. *Generative AI and the EUDPR: Orientations for ensuring data protection compliance when using Generative AI systems (Version 2)* (28 October 2025). https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf

Facoco, I et al. (2025) "Adapting Stable Diffusion Models for Domain-Specific Medical Imaging: A Case Study in Synthetic Retinal Fundus Image Generation", accepted in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD). https://zenodo.org/records/17207258

Hardt, D. (2012). The OAuth 2.0 Authorization Framework. https://www.rfc-editor.org/rfc/rfc6749

Hoefler, M. A., Mueller, K., & Samek, W. (2025). FedXDS: Leveraging Model Attribution Methods to Counteract Data Heterogeneity in Federated Learning.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

https://openaccess.thecvf.com/content/ICCV2025/papers/Hoefler_FedXDS_Leveraging_Model_Attribution_Methods_to_counteract_Data_Heterogeneity_in_ICCV_2025_paper.pdf

Jobin, A., Ienca, M., Vayena, E. The global landscape of AI ethics guidelines (2019). https://www.researchgate.net/publication/335579286_The_global_landscape_of_AI_ethics_guidelines

Hu, V., Ferraiolo, D., Kuhn, R., et al. (2014). Guide to Attribute Based Access Control (ABAC) Definition and Considerations. https://doi.org/10.6028/NIST.SP.800-162

Jones, M., & Hardt, D. (2012). The OAuth 2.0 Authorization Framework: Bearer Token Usage. https://www.rfc-editor.org/rfc/rfc6750

MCP-2025. Model Context Protocol. Specification and documentation (2025). https://modelcontextprotocol.io/specification/2025-11-25

Misono, M., et al. (2024). Confidential VMs Explained: An Empirical Analysis of AMD SEV-SNP and Intel TDX. https://dse.in.tum.de/wp-content/uploads/2024/11/sigmetrics25summer-CVM-Explained.pdf

OECD-2025. OECD. Sharing Trustworthy AI Models with Privacy-Enhancing Technologies. 2025. https://doi.org/10.1787/a266160b-en

OpenDSU-About. OpenDSU. About OpenDSU. Accessed 2026. https://www.opendsu.org/pages/about.html

Richer, J. (2015). OAuth 2.0 Token Introspection. https://www.rfc-editor.org/rfc/rfc7662

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero Trust Architecture. https://csrc.nist.gov/pubs/sp/800/207/final

Sereno, J., Castro, D., Santos, N., & Rodrigues, L. (2025). Secure Lifecycle Management of Confidential Virtual Machines in Public Clouds. https://ieeexplore.ieee.org/document/11261550

Ursache, C. (2025). Integrating Federated Learning in ACHILLES IDE. Internal ACHILLES project document. https://www.researchgate.net/publication/403758655_Integrating_Federated_Learning_in_Achilles_IDE

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

APPENDIX 1 – ASSET FACT SHEETS

The asset descriptions in this public deliverable are intentionally limited to non-confidential information. Detailed implementation artefacts, deployment configurations, security-sensitive parameters, and partner-confidential technical documentation remain governed by the project's internal access-control and confidentiality procedures.

Asset Fact Sheet: COML - Collaborative Machine Learning without Centralized Training Data

Version: 0.4
Date: 2026-04-06
Author: João Gonçalves, Filipe Soares (Fraunhofer Portugal AICOS)
Category: Federated Learning / Machine Learning / AI Infrastructure

---

1. What It Does (Executive Summary)

Collaborative Machine Learning without Centralized Training Data (COML) v3 is a comprehensive platform for federated learning, aggregation scheme selection and federated evaluation. It is powered by NVFlare, a domain-agnostic and extensible SDK, that enables distributed machine learning across multiple sites while keeping data locality. Over the security and privacy-enhancing technologies available as differential privacy, homomorphic encryption, authentication, and encrypted communications, it provides several components prepared by Fraunhofer AICOS with docker containers and REST API: Admin Web API and graphical user interface for job configuration, orchestration and management, multiple aggregation schemes implemented and tested, a data analyser and partitioner based on real or synthetic data quality, along with tracking services for monitoring training progress and comparison of machine learning models. The platform supports various deployment ML models for federated learning including direct and docker-based installations (edge and cloud), making it suitable for both development and production environments.

---

2. Why It Matters (Business Impact)

Federated learning addresses critical challenges in distributed machine learning by enabling collaborative model training without centralizing sensitive data. This approach offers significant business value:

Key Benefits:

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Privacy and data sovereignty preservation:** Organizations can collaborate on machine learning models without sharing raw data, addressing regulatory concerns and maintaining competitive advantage.
- **Robustness through more diversity:** Privacy and data sovereignty can limit collaboration between institutions with real-world data, which reduces the data diversity needed for robust AI models development that can tackle some of the screening challenges.
- **Scalability:** The system supports multiple client sites, enabling large-scale distributed learning across geographically dispersed locations. It is also based on a modular architecture that supports the extension to new ML models, metrics, and customizable FL aggregation strategies.
- **Regulatory Compliance:** By keeping data localized, the platform helps organizations meet GDPR and other data protection regulations.
- **Cost Efficiency:** Reduces the need for data consolidation, lowering infrastructure and data transfer costs.
- **Flexible Deployment:** Supports both cloud and on-premise deployments, accommodating various organizational requirements.

---

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

[List relevant regulations and standards this asset helps address]

Applicable Regulations:

☑ Digital Governance Act
☐ Digital Markets Act
☐ Digital Services Act
☑ EU AI Act
☐ EU Cyber Resilience Act
☐ EU Cyber Security Act
☑ GDPR
☐ NIS2 Directive
☑ Other: HIPAA (Health Insurance Portability and Accountability Act)

Sector Specific Regulations:

☐ GxP
☑ HIPAA
☑ Medical Device Regulation
☐ Other: [Specify]

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Compliance Support:

The platform's federated learning architecture inherently supports regulatory compliance related with privacy of the data by enabling collaborative model training without data centralization. This approach helps organizations meet data privacy requirements while still benefiting from distributed machine learning.

On the other hand, the EU AI act is a regulatory driver for better design and consideration for system data sovereignty (how and when to use data) and localization, data privacy and ownership, and particularly for high-risk AI systems. It is also a catalyst for the wider adoption of federated learning operations (FLOps). For example, using federated learning and the traceability provided by COML could act as evidence that an organisation has met its data governance requirements under Article 10 of the EU AI Act.

3.2 Guidelines and Standards Relevance

☑ ISO 13485
☐ ISO 42001
☐ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)
☐ ISO/IEC 23894 (AI Risk Management – conceptual alignment)
☐ ISO/IEC 27001 (Information Security)
☐ MPEG Feature Coding for Machines (FCM)
☐ NIST Risk Management Framework for AI
☐ OECD AI Principles
☐ Other: [Specify]

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☑ Yes ☐ No
Sample-level	Prevents identification of individual training examples	☑ Yes ☐ No

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☑ Yes ☐ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No
Malicious server	Server deviates from protocol, sends crafted queries	[] Yes ☑ No
External adversary	Attacker with black-box or white-box model access	☑ Yes ☐ No

Privacy Mechanism:

- ☑ Differential Privacy — Central DP
- ☑ Differential Privacy — Local DP
- ☑ Secure Aggregation
- ☐ Protected Activations / Learned Obfuscation
- ☑ Cryptographic (MPC / HE / TEE)
- ☑ DP Synthetic Data
- ☐ None (utility-focused baseline)
- ☐ Other: [Specify]

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☐ Yes ☑ No	[Link / Metric]
Gradient Reconstruction Attack	☐ Yes ☑ No	[Link / Metric]

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Activation Inversion Attack	☐ Yes ☑ No	[Link / Metric]
Formal ε Verification	☐ Yes ☑ No	[ε value]

Notes:
[Any additional context on privacy limitations, assumptions, or recommended configurations]

---

4. How It Works (Technical Approach)

COML uses NVFlare as its core framework, with an Admin API serving as the bridge between users and NVFlare services. The system architecture includes:

Core Technology:

- NVFlare (NVIDIA Federated Learning framework)
- FastAPI for the Admin Web API
- Docker for containerization
- MLFlow for experiment tracking
- MinIO artifact Store to save models and other artifacts

Key Components:

- Admin Web API: Manages and orchestrates jobs and system operations
- NVFlare Server: Coordinates the federated learning process
- NVFlare Clients: Run on individual sites to perform local training
- Tracking Services: Provide monitoring and visualization of training progress through MLFlow

COML supports various job templates for different machine learning scenarios, enabling flexible deployment of asynchronous federated learning workflows. The platform provides 7 federated learning strategies: - AutoFedAvg (Automatic Federated Averaging) - FedAvg (Federated Averaging) - FedProx (Federated Proximal) - FedBN (Federated Batch Normalization) - FedSGD (Federated Stochastic Gradient Descent) - FedYogi (Federated Yogi optimizer) - FedNova (Federated Normalized Averaging)

These strategies are available through pre-configured job templates that can be customized for specific ML models and datasets.

In summary, federated aggregation strategies that rely solely on straightforward model-averaging perform well when the local datasets are roughly independent and identically distributed, offering a low-complexity solution for simple federations. More sophisticated schemes introduce additional regularization or normalization to accommodate skewed data distributions, client heterogeneity or frequent partial participation, trading higher computational overhead for greater

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

adaptability. These FL aggregation schemes can be selected based on whether preserving convergence speed outweighs the need to handle distribution drift or varying system conditions.

---

5. Deployment Model

[Describe how the asset can be deployed]

Available Options:

- ☑ Cloud (SaaS)
- ☑ On-Premise / Bare metal
- ☑ Container (Docker/Kubernetes)
- ☑ API Service
- ☐ Library/SDK
- ☑ Framework
- ☐ Guideline / Checklist

Recommended:

Docker container deployment for production environments in edge and cloud, or multi-cloud services. Full local deployment (client nodes plus server) for development and testing.

---

6. Effort to Implement

Overall Complexity: ☐ Low ☑ Medium ☐ High
Time Estimate: 2-4 weeks

Breakdown:

- **Preparation & Requirements:** 1-2 days
- **Setup & Configuration:** 2-4 days
- **Integration:** 3-5 days
- **Testing & Validation:** 2-3 days
- **Training:** 2-3 days

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- Python 3.8+
- Git

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- Make
- Docker (for containerized deployment)
- NVFlare 2.4.1+

System Requirements:

- Linux-based operating system (ARM or x86)
- Minimum 32 GB RAM for the Federated Learning server.
- Sufficient CPU/GPU resources based on workload
- Network connectivity between sites

Dependencies:

- nvflare >= 2.4.1
- fastapi >= 0.104.1
- mlflow-skinny >= 2.8.1
- uvicorn >= 0.24.0
- pydantic >= 2.8.2

---

8. License, Cost & Classification

Software Type:

☑ Open-Source Software
☑ Commercial Software
☐ Cloud Service
☐ Library/SDK
☑ Framework
☐ Standalone Tool

License Type:

☐ Open Source ([License name, e.g., MIT, Apache 2.0, GPL])
☑ Proprietary
☐ Freemium
☐ Enterprise License

Cost Structure:

- To be defined as it uses some open source, and commercial software by Fraunhofer which is initially available via Evaluation License Agreement (ELA).
- Current running cost of the cloud is: US$500 (five hundred dollars) per month.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

• Local client nodes may use their current computing power (CPU only or CPU+GPU), or request help for the specification of expansion on-premisses or Cloud GPUs.

Commercial Restrictions:

• Subject to contractual terms. Check with project maintainers for license agreement or commercial use restrictions.

---

9. Maturity Level

Technical Readiness Level (TRL): 6-7

Stability:

☐ Experimental
☐ Proof of Concept (PoC)
☑ Stable with active development
☐ Stable with maintenance mode
☐ Legacy/End-of-life

Version History:

• Current Version: 3.0.0
• First Release: 2021 (internal R&D) and 2024 (external deployments).
• Release Frequency: Iterative, pilot-driven with ongoing development.

Adoption:

• **Number of users/installations:** 5 external installations of local clients, currently limited to Fraunhofer AICOS users and international partners who collaborate in ongoing projects copromoted by Fraunhofer. New installations under onboarding process: 2.
• **Notable Users:** Fraunhofer AICOS research projects and Hospital partners. All users working in the healthcare application field.
• **Relevant Artifacts:** Code repository, documentation, job templates for various Machine Learning scenarios.

---

10. Vendor/Community Support

Support Channels:

☑ Official Documentation
☐ Community Forum
☐ GitHub Issues

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Commercial Support (SLA available)
☑ Email Support
☐ Slack/Discord Community
☑ Private JIRA Issues

Documentation Quality: ☐ Basic ☑ Good ☐ Excellent

Community Activity:

- GitHub Stars: N/A (private repository)
- Contributors: Fraunhofer AICOS team
- Last Update: 2026-01-21

Training Resources:

☐ Official Training Available
☑ Tutorials/Guides
☑ Video Content (promotional only)
☐ Certification Programs

Contact for Questions: <filipe.soares@aicos.fraunhofer.pt>

Related Assets: NVFlare documentation, FastAPI documentation

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Asset Fact Sheet: Evident

Version: 0.1
Date: 2026-03-26
Author: João Sereno, Daniel Castro, Nuno Santos, Luís Rodrigues
Category: Confidential Computing / Cloud Security / Privacy Enhancement Technologies

---

1. What It Does (Executive Summary)

Evident is a framework for secure lifecycle management of Confidential Virtual Machines (CVMs) in public clouds. It automates remote attestation, enhances transparency of cloud trust models, and enables verifiable deployment and interaction with virtual machines executing confidential workloads (e.g., AI inference or federated learning).

---

2. Why It Matters (Business Impact)

Cloud providers offer Confidential Virtual Machines based on hardware technologies such as AMD SEV-SNP and Intel TDX, but their trust models are opaque, inconsistent, and difficult to verify. Organizations handling sensitive workloads (e.g., AI models, healthcare data, financial analytics, or regulated enterprise systems) lack transparency regarding what is protected and what must still be trusted in the Cloud Service Provider (CSP).

Evident bridges this gap by providing automated, reproducible, and transparent attestation workflows across major cloud platforms. It reduces reliance on implicit trust in CSPs and replaces it with verification of firmware, boot components, and runtime state. This significantly strengthens assurance for confidential AI deployments, federated learning scenarios, and third-party workload execution.

By simplifying secure CVM deployment and interaction, Evident lowers the operational barrier to adopting confidential computing. It enables secure collaboration between model owners and data owners without exposing sensitive assets. This improves regulatory posture, reduces risk of data leakage, and strengthens trust in outsourced cloud computation.

Key Benefits:

- Transparent Trust Verification Provides automated remote attestation and explicit verification of CVM integrity, reducing blind trust in cloud providers.
- Secure Third-Party Workload Provisioning Enables confidential AI inference and federated learning workflows where model and data owners independently verify execution integrity.
- Cloud-Agnostic Security Layer Abstracts inconsistencies across AWS, Azure, and GCP, offering a unified and reproducible lifecycle management framework.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

[List relevant regulations and standards this asset helps address]

Applicable Regulations:

- ☐ Digital Governance Act
- ☐ Digital Markets Act
- ☐ Digital Services Act
- ☑ EU AI Act
- ☑ EU Cyber Resilience Act
- ☑ EU Cyber Security Act
- ☑ GDPR
- ☑ NIS2 Directive
- ☐ Other: [Specify]

Sector Specific Regulations:

- ☐ GxP
- ☐ HIPAA
- ☐ Medical Device Regulation
- ☐ Other: [Specify]

Compliance Support:

Evident strengthens compliance by ensuring data-in-use protection through hardware-backed isolation and verifiable execution environments. It supports GDPR principles such as data protection by design and by default, minimizing unauthorized access risks. For the EU AI Act, Evident enhances traceability, integrity, and accountability of AI systems deployed in cloud infrastructures. It also contributes to auditability and risk mitigation under broader cybersecurity frameworks.

3.2 Guidelines and Standards Relevance

- ☐ ISO 13485
- ☑ ISO 42001
- ☐ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)
- ☑ ISO/IEC 23894 (AI Risk Management – conceptual alignment)
- ☑ ISO/IEC 27001 (Information Security)
- ☐ MPEG Feature Coding for Machines (FCM)

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☑ NIST Risk Management Framework for AI
☐ OECD AI Principles
☐ Other: [Specify]

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☐ Yes ☑ No
Sample-level	Prevents identification of individual training examples	☐ Yes ☑ No
Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☐ Yes ☑ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No
Malicious server	Server deviates from protocol, sends crafted queries	☐ Yes ☑ No
External adversary	Attacker with black-box or white-box model access	☐ Yes ☑ No

Privacy Mechanism:

☐ Differential Privacy — Central DP
☐ Differential Privacy — Local DP
☐ Secure Aggregation

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Protected Activations / Learned Obfuscation
☑ Cryptographic (MPC / HE / TEE)
☐ DP Synthetic Data
☐ None (utility-focused baseline)
☐ Other: [Specify]

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☐ Yes ☑ No	[Link / Metric]
Gradient Reconstruction Attack	☐ Yes ☑ No	[Link / Metric]
Activation Inversion Attack	☐ Yes ☑ No	[Link / Metric]
Formal ε Verification	☐ Yes ☑ No	[ε value]

Notes:

When using Evident, trust in the underlying hardware is required. It is assumed that the hardware vendor does not introduce backdoors or critical vulnerabilities in the Trusted Execution Environment (e.g., AMD SEV-SNP).

Confidential Virtual Machines protect against unauthorized access from cloud providers by enforcing hardware-backed isolation, assuming correct implementation of the underlying technology. However, misconfigurations or deviations from the expected protocols may occur in practice.

Evident mitigates this risk by enabling remote attestation, allowing users to verify that the CVM is correctly instantiated and configured. This ensures that only trusted software components are executed and that security guarantees advertised by the cloud provider can be independently validated.

---

4. How It Works (Technical Approach)

Evident builds on hardware-based Trusted Execution Environments, specifically Confidential Virtual Machines enabled by AMD SEV-SNP and similar technologies. It automates remote attestation to verify the integrity of VM firmware, boot components, and runtime configuration before sensitive workloads are provisioned.

The framework consists of two main components: Evident-server (deployed inside the CVM) and Evident-client (used by remote parties). Evident-server bootstraps applications inside the CVM and

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

generates cryptographic evidence of the system's integrity. Evident-client verifies attestation reports, validates measurements, and establishes secure communication channels only after successful verification.

The system makes trust assumptions explicit and transparent, mitigating inconsistencies across cloud providers. It supports secure provisioning workflows where model owners and data owners independently verify that only trusted components execute within the CVM.

Core Technology:

• Trusted Execution Environments (AMD SEV-SNP)
• Remote Attestation (cryptographic verification)

Key Components:

• Evident-server (in-VM attestation and bootstrap service)
• Evident-client (remote verification and secure interaction interface)

---

5. Deployment Model

Evident is deployed in conjunction with Confidential Virtual Machines in public or private cloud environments.

Available Options:

- ☑ Cloud (SaaS)
- ☑ On-Premise / Bare metal
- ☐ Container (Docker/Kubernetes)
- ☐ API Service
- ☐ Library/SDK
- ☐ Framework
- ☐ Guideline / Checklist

Recommended: Deploy in Google Cloud, Azure or AWS, choose AMD-SEV servers.

After the attestation step is completed, any software can be run within the Confidential Virtual Machine.

---

6. Effort to Implement

Overall Complexity: ☐ Low ☑ Medium ☐ High

Time Estimate: 4--8 weeks (depending on integration depth)

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Breakdown:

- **Preparation & Requirements:** 1--2 weeks
- **Setup & Configuration:** 1 week
- **Integration:** 1--2 weeks
- **Testing & Validation:** 1--2 weeks
- **Training:** 3--5 days

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- Access to CVM-enabled cloud instances (e.g., AMD SEV-SNP capable VMs)
- Underlying CPU hardware supporting SEV-SNP or equivalent
- Understanding of remote attestation workflows

System Requirements:

- Cloud environment supporting Confidential VMs
- Secure key management infrastructure

Dependencies:

- Cloud CLI tools and attestation services (AWS, Azure, or GCP)
- [NIX](https://nixos.org/) for building the VM images
- The server is coded in [Rust](https://rust-lang.org/), the client in [Go](https://go.dev/)

---

8. License, Cost & Classification

Software Type:

- ☑ Open-Source Software
- ☐ Commercial Software
- ☐ Cloud Service
- ☐ Library/SDK
- ☐ Platform Component
- ☐ Standalone Tool

License Type:

- ☑ Open Source (Apache 2.0)
- ☐ Proprietary

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Freemium
☐ Enterprise License

Cost Structure:

• Free (research prototype; infrastructure costs depend on cloud provider)
• Estimated Cost: Cloud CVM instance costs only

Commercial Restrictions:

• None

---

9. Maturity Level

Technical Readiness Level (TRL): 3 (proof-of-concept validation)

Stability:

☐ Experimental
☑ Proof of Concept (PoC)
☐ Stable with active development
☐ Stable with maintenance mode
☐ Legacy/End-of-life

Version History:

• Current Version: 0.1
• First Release: 2026-03-04
• Release Frequency: N/A

Adoption:

• Research deployments
• **Notable Users:** INESC-ID researchers
• **Relevant Artifacts:** [Gitlab repository] (https://gitlab.com/dpss-inesc-id/achilles-cvm)

---

10. Vendor/Community Support

Support Channels:

☑ Official Documentation
☐ Community Forum
☐ GitHub Issues

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Commercial Support (SLA available)
☐ Email Support
☐ Slack/Discord Community

Documentation Quality: ☐ Basic ☐ Good ☐ Excellent

Community Activity:

- GitHub Stars: N/A
- Contributors: 1 (INESC-ID)
- Last Update: March 2026

Training Resources:

☐ Official Training Available
☑ Tutorials/Guides
☐ Video Content
☐ Certification Programs

Contact for Questions: João Sereno (joaohsereno@tecnico.ulisboa.pt), Daniel Castro (daniel.castro@tecnico.ulisboa.pt)

Related Assets: N/A

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Asset Fact Sheet: FedKT-CSD - Collaborative Synthetic Data

Version: 0.1 (Under Review)
Date: 2026-01-20
Author: FhHHI
Category: Federated Learning / Synthetic Data / Privacy

---

1. What It Does (Executive Summary)

FedKT-CSD is a personalized Federated Learning (pFL) framework that enables the generation of high-quality synthetic training data in a single communication round ("One-Shot"). Instead of sharing model updates, clients optimize low-dimensional latent vectors using a frozen, public Variational Autoencoder (VAE) and share differentially private class-wise statistics. The server generates a synthetic dataset to train a global feature extractor, which clients then download and personalize.

---

2. Why It Matters (Business Impact)

Traditional FL requires continuous, high-bandwidth communication and heavy on-device computation, making it impractical for many real-world deployments. FedKT-CSD drastically reduces the overhead for clients and networks while maintaining formal privacy guarantees. The one-shot communication design eliminates the need for hundreds of synchronization rounds, cutting infrastructure costs and reducing the attack surface for communication-channel adversaries. This is especially relevant in cross-device settings such as mobile health apps or IoT sensor networks where battery life and bandwidth are constrained. The framework's use of a public VAE means clients never need to train or transmit generative model weights, keeping computational demands minimal on edge hardware. By providing formal Differential Privacy guarantees on the shared statistics, FedKT-CSD offers a clear compliance story for regulated industries. Its demonstrated scalability to 2000+ clients makes it suitable for large-scale production deployments.

Key Benefits:

- **Communication Efficiency:** "One-Shot" method—requires only a single round of communication (uploading small statistics), unlike the hundreds of rounds in standard FL. This dramatically reduces infrastructure and bandwidth costs.
- **Low Computational Footprint:** Clients do not train generative models; they only perform lightweight gradient optimization on latent vectors. Suitable for mobile/edge devices with limited compute and battery.
- **Formal Privacy:** Provides record-level Differential Privacy (DP) guarantees via the Gaussian mechanism, offering a clear mathematical privacy bound for regulatory compliance.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Extreme Heterogeneity Handling:** Proven robustness to extreme label skew and domain shifts where standard FL fails, validated across diverse benchmark scenarios.

---

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

Applicable Regulations:

- ☐ Digital Governance Act
- ☐ Digital Markets Act
- ☐ Digital Services Act
- ☑ EU AI Act
- ☐ EU Cyber Resilience Act
- ☐ EU Cyber Security Act
- ☑ GDPR
- ☐ NIS2 Directive
- ☐ Other: [Specify]

Sector Specific Regulations:

- ☐ GxP
- ☐ HIPAA
- ☐ Medical Device Regulation
- ☐ Other: [Specify]

Compliance Support:

FedKT-CSD ensures that no raw data and no model parameters derived directly from raw data are ever shared. Only aggregate, noisy statistics (means and covariances) are transmitted. The generation of synthetic data occurs on the server side using public priors, ensuring a clean separation between private client data and the global model training process. This supports GDPR principles of Anonymization and Data Minimization, as well as EU AI Act requirements for Transparency and Data Governance.

3.2 Guidelines and Standards Relevance

- ☐ ISO 13485
- ☐ ISO 42001
- ☐ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)
- ☐ ISO/IEC 23894 (AI Risk Management – conceptual alignment)

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ ISO/IEC 27001 (Information Security)
☐ MPEG Feature Coding for Machines (FCM)
☐ NIST Risk Management Framework for AI
☐ OECD AI Principles
☐ Other: [Specify]

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☐ Yes ☑ No
Sample-level	Prevents identification of individual training examples	☑ Yes ☐ No
Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☑ Yes ☐ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No
Malicious server	Server deviates from protocol, sends crafted queries	☐ Yes ☑ No
External adversary	Attacker with black-box or white-box model access	☐ Yes ☑ No

Privacy Mechanism:

☑ Differential Privacy — Central DP

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Differential Privacy — Local DP
☐ Secure Aggregation
☐ Protected Activations / Learned Obfuscation
☐ Cryptographic (MPC / HE / TEE)
☑ DP Synthetic Data
☐ None (utility-focused baseline)
☐ Other: [Specify]

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☐ Yes ☑ No	—
Gradient Reconstruction Attack	☐ Yes ☑ No	—
Activation Inversion Attack	☐ Yes ☑ No	—
Formal ε Verification	☑ Yes ☐ No	ε values reported in paper

Notes:

Privacy is provided via the Gaussian mechanism applied to aggregated class-wise sums and covariances after per-client clipping. The formal ε guarantee is record-level. No adversarial privacy auditing (MIA, reconstruction) has been performed yet.

---

4. How It Works (Technical Approach)

FedKT-CSD operates in three main stages. First, a frozen, publicly available Variational Autoencoder (VAE) is distributed to all clients. Each client then optimizes low-dimensional latent vectors to reconstruct its local private data through the VAE decoder, without training any model weights. Clients clip their optimized latent vectors and compute class-wise sufficient statistics (means and covariances) which are sent to the server. The server applies calibrated Gaussian noise to these aggregated statistics to satisfy Differential Privacy. Using the noisy statistics, the server samples synthetic latent vectors from Gaussian distributions and decodes them through the VAE to produce a full synthetic dataset. A global feature extractor is then trained on this synthetic dataset. Finally, clients

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

download the global feature extractor and train a lightweight local classification head on their private data for personalization. The entire client-to-server communication happens in a single round.

Core Technology:

• Public Pretrained VAE: Uses a frozen encoder/decoder (e.g., from ImageNet) distributed to all clients.
• Latent Space Optimization: Clients find latent vectors that reconstruct their private data.
• Differential Privacy: Clients clip vectors; Server adds noise to aggregated class-wise sums and covariances.

Key Components:

• Client Side: Optimization of latent codes z (not model weights).
• Server Side: Generation of synthetic dataset X* via sampling from DP-protected Gaussian distributions.
• Personalization: Clients receive the global feature extractor and train a lightweight local head.

---

5. Deployment Model

Available Options:

☐ Cloud (SaaS)
☑ On-Premise / Bare metal
☑ Container (Docker/Kubernetes)
☐ API Service
☑ Library/SDK
☐ Framework
☐ Guideline / Checklist

Recommended: Ideal for cross-silo scenarios with bandwidth constraints or cross-device scenarios (mobile phones) where client battery/compute is limited.

---

6. Effort to Implement

Overall Complexity: ☑ Low ☐ Medium ☐ High
Time Estimate: 1-2 weeks

Breakdown:

• **Preparation & Requirements:** Selection of public VAE

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Setup & Configuration:** 2-3 days
- **Integration:** Replacing standard `fit()` loop with latent optimization loop
- **Testing & Validation:** 1 week
- **Training:** Tuning synthetic data budget

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- A publicly available VAE relevant to the data modality (e.g., images).
- PyTorch / TensorFlow.

System Requirements:

- Low requirement for Client (CPU or weak GPU).
- High requirement for Server (GPU for feature extractor training).

Dependencies:

- Standard Deep Learning libraries.
- No heavy dependency on specific FL frameworks (agnostic).

---

8. License, Cost & Classification

Software Type:

☑ Open-Source Software
☐ Commercial Software
☐ Cloud Service
☐ Library/SDK
☐ Platform Component
☐ Standalone Tool

License Type:

☑ Open Source (Pending publication)
☐ Proprietary
☐ Freemium
☐ Enterprise License

Cost Structure:

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

• Low: Significant reduction in communication costs (uplink/downlink) compared to FedAvg.
• Compute: Server bears the load of training the feature extractor; Clients have minimal load.
• Estimated Cost: Minimal operational cost

Commercial Restrictions:

• None anticipated (Research Code).

---

9. Maturity Level

Technical Readiness Level (TRL): 3-4 (Research Prototype)

Stability:

☑ Experimental
☐ Proof of Concept (PoC)
☐ Stable with active development
☐ Stable with maintenance mode
☐ Legacy/End-of-life

Version History:

• Current Version: 0.1
• First Release: 2026 (Pending)
• Release Frequency: N/A

Adoption:

• Demonstrated on CIFAR-10, CIFAR-100, Tiny-ImageNet, PACS, DomainNet.
• Scalability tested up to 2000+ clients.
• **Notable Users:** Fraunhofer HHI
• **Relevant Artifacts:** Paper under review

---

10. Vendor/Community Support

Support Channels:

☐ Official Documentation
☐ Community Forum
☐ GitHub Issues
☐ Commercial Support (SLA available)
☑ Email Support

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Slack/Discord Community

Documentation Quality: ☐ Basic ☑ Good ☐ Excellent

Community Activity:

- GitHub Stars: N/A
- Contributors: N/A
- Last Update: N/A

Training Resources:

☐ Official Training Available
☐ Tutorials/Guides
☐ Video Content
☐ Certification Programs

Contact for Questions: maximilian.andreas.hoefler@hhi.fraunhofer.de

Related Assets: [Links to related fact sheets]

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Asset Fact Sheet: FedSyn-Refine - Federated Synthetic Data Generation via LLM Refinement

Version: 0.1 (Under Review)
Date: 2026-01-20
Author: FhHHI
Category: Federated Learning / Synthetic Data / NLP / Privacy

---

1. What It Does (Executive Summary)

FedSyn-Refine is a one-shot federated learning framework that generates differentially private synthetic text data from aggregated client signals. Clients compute and transmit class-conditional mean embeddings of their local data in a single round under DP constraints. The server extracts semantically meaningful key phrases from the noisy aggregates and refines them into full text samples via LLM prompting with class-conditioning and style guidance, producing a centralized synthetic training corpus without sharing raw data.

---

2. Why It Matters (Business Impact)

Deploying federated learning at scale is hindered by statistical heterogeneity, multi-round communication overhead, high on-device computational demands, and the utility loss from iterative differential privacy noise injection. Existing LLM-based synthetic data approaches still rely on multi-round protocols or require clients to perform expensive tasks like on-device LLM inference. FedSyn-Refine eliminates these bottlenecks by collapsing training into a single communication round where clients perform only a lightweight embedding forward pass. All heavy computation—including LLM inference and model training—is offloaded to the server, reducing the client's role to a single efficient operation completing in under 3 seconds on commodity CPU hardware. The framework scales favorably with client participation: as more clients contribute, sensitivity remains fixed while aggregate signal strength grows, reducing DP noise variance without weakening privacy. On text classification benchmarks, FedSyn-Refine matches or exceeds the accuracy of state-of-the-art multi-round FL methods while demonstrating substantially improved fairness across clients.

Key Benefits:

- **One-Shot Communication:** Requires only a single communication round per client (6 kB upload), eliminating multi-round synchronization overhead and making it suitable for large-scale asynchronous deployments under severe resource constraints.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Minimal Client Computation:** Clients perform only a single forward pass through a lightweight sentence encoder (all-MiniLM-L6-v2), completing in under 3 seconds on standard CPU hardware with 46 MB peak memory—no GPU required.
- **Heterogeneity Invariance:** Aggregating class-conditional statistics is mathematically equivalent to centralized computation, eliminating optimization-induced degradation from non-IID data. Performance varies only 2-3pp across heterogeneity levels.
- **Improved Fairness:** Achieves substantially lower cross-client variance ($\sigma = 2.10\%$) and higher minimum client accuracy (70.92%) compared to standard pFL baselines, many of which exhibit 0% minimum accuracy on some clients.

---

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

Applicable Regulations:

☐ Digital Governance Act
☐ Digital Markets Act
☐ Digital Services Act
☑ EU AI Act
☐ EU Cyber Resilience Act
☐ EU Cyber Security Act
☑ GDPR
☐ NIS2 Directive
☐ Other: [Specify]

Sector Specific Regulations:

☐ GxP
☑ HIPAA
☐ Medical Device Regulation
☐ Other: [Specify]

Compliance Support:

FedSyn-Refine provides document-level (ε, δ)-Differential Privacy guarantees. No raw data leaves client devices—only clipped, class-conditional mean embeddings are transmitted via secure aggregation. The server applies calibrated Gaussian noise, and all downstream processing (seed extraction, LLM generation) inherits the DP guarantee via post-processing. This supports GDPR principles of data minimization and privacy by design, and EU AI Act requirements for data governance

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

and transparency in high-risk AI systems. The framework has been validated on medical text data (PubMed abstracts), demonstrating applicability to healthcare domains.

3.2 Guidelines and Standards Relevance

☐ ISO 13485
☐ ISO 42001
☐ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)
☐ ISO/IEC 23894 (AI Risk Management – conceptual alignment)
☐ ISO/IEC 27001 (Information Security)
☐ MPEG Feature Coding for Machines (FCM)
☐ NIST Risk Management Framework for AI
☐ OECD AI Principles
☐ Other: [Specify]

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☐ Yes ☑ No
Sample-level	Prevents identification of individual training examples	☑ Yes ☐ No
Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☑ Yes ☐ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Malicious server	Server deviates from protocol, sends crafted queries	☐ Yes ☑ No
External adversary	Attacker with black-box or white-box model access	☐ Yes ☑ No

Privacy Mechanism:

- ☑ Differential Privacy — Central DP
- ☐ Differential Privacy — Local DP
- ☑ Secure Aggregation
- ☐ Protected Activations / Learned Obfuscation
- ☐ Cryptographic (MPC / HE / TEE)
- ☑ DP Synthetic Data
- ☐ None (utility-focused baseline)
- ☐ Other: [Specify]

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☐ Yes ☑ No	—
Gradient Reconstruction Attack	☐ Yes ☑ No	—
Activation Inversion Attack	☐ Yes ☑ No	—
Formal ε Verification	☑ Yes ☐ No	$\varepsilon = 6.0, \delta = 1\times10^{-3}$ (document-level DP)

4. How It Works (Technical Approach)

FedSyn-Refine operates in three stages. In the first stage, each client embeds its local text documents using a lightweight pretrained sentence encoder (all-MiniLM-L6-v2), applies per-document ℓ_2 clipping, and computes per-class sums and counts of the clipped embeddings. These additive

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

statistics are transmitted once via secure aggregation. In the second stage, the server receives only the global sums, applies calibrated Gaussian noise to satisfy (ε, δ)-DP at the document level, and computes noisy class-conditional mean embeddings. In the third stage, the server transforms each noisy mean vector into semantically rich seed phrases through a pipeline of Top-K token selection (centred cosine similarity against a public vocabulary), compositional n-gram phrase generation, stochastic hill-climbing search for semantically coherent phrases, and diversity-aware MMR reranking. These seed phrases are used to prompt a server-side LLM (Mistral-7B) with class-conditioning and style guidance to generate a high-quality synthetic text corpus. A central model (TinyBERT) is then trained on this corpus and distributed to clients for optional local personalization.

Core Technology:

• Sentence Embeddings: all-MiniLM-L6-v2 for lightweight document encoding (384-dimensional).
• Differential Privacy: Gaussian mechanism with document-level (ε, δ)-DP via clipped class-conditional sums.
• LLM Synthesis: Mistral-7B-Instruct for contextual text generation from DP-protected seed phrases.
• Secure Aggregation: Compatible with SecAgg protocols due to additive statistics.

Key Components:

• Client Side: Single forward pass-through sentence encoder, per-class sum/count computation (CPU-only, <3s).
• Server Side: DP noise injection, lexical seed extraction (Top-K + n-grams + hill-climbing + MMR), LLM-based synthetic data generation.
• Personalization: Optional two-stage pFL where clients freeze the pretrained encoder and fine-tune a local classification head.

---

5. Deployment Model

Available Options:

☐ Cloud (SaaS)
☑ On-Premise / Bare metal
☑ Container (Docker/Kubernetes)
☐ API Service
☑ Library/SDK
☐ Framework
☐ Guideline / Checklist

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Recommended: Cross-device deployment where lightweight clients (CPU-only, mobile/edge) transmit embeddings to a GPU-equipped server that handles LLM synthesis and model training. Ideal for large-scale asynchronous deployments (tested with 1000+ clients).

---

6. Effort to Implement

Overall Complexity: ☐ Low ☑ Medium ☐ High
Time Estimate: 2-3 weeks

Breakdown:

- **Preparation & Requirements:** Selection of sentence encoder, LLM, and public vocabulary (2-3 days)
- **Setup & Configuration:** Server-side LLM deployment and DP parameter calibration (1 week)
- **Integration:** Client embedding pipeline and secure aggregation setup (3-5 days)
- **Testing & Validation:** End-to-end pipeline validation and synthetic data quality assessment (1 week)
- **Training:** Server-side LLM generation (~20 min on 2×RTX 5090) + downstream model training (~10 min)

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- Server with GPU capability for LLM inference and model training.
- Pretrained sentence encoder (all-MiniLM-L6-v2 or equivalent).
- Access to an LLM for server-side generation (Mistral-7B-Instruct or equivalent).
- Public vocabulary with word embeddings (e.g., GloVe, ~20K words).

System Requirements:

- Client: CPU-only, 46 MB peak memory, Python environment.
- Server: GPU (2× RTX 5090 or equivalent), 12 GB+ memory for LLM inference.

Dependencies:

- PyTorch
- HuggingFace Transformers / Sentence-Transformers
- LLM weights (Mistral-7B-Instruct-v3.0 or equivalent)

---

8. License, Cost & Classification

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Software Type:

- ☑ Open-Source Software
- ☐ Commercial Software
- ☐ Cloud Service
- ☐ Library/SDK
- ☐ Platform Component
- ☐ Standalone Tool

License Type:

- ☑ Open Source (Code available on GitHub)
- ☐ Proprietary
- ☐ Freemium
- ☐ Enterprise License

Cost Structure:

• Free (Open-Source research code).
• Client-side cost: Negligible (CPU-only, <3s per client).
• Server-side cost: One-time ~30 min GPU compute for synthesis + training; reusable synthetic dataset.
• Estimated Cost: Free software; server GPU hardware required

Commercial Restrictions:

• Subject to the license of the LLM used for generation (e.g., Mistral model license).

---

9. Maturity Level

Technical Readiness Level (TRL): 3-4 (Research Prototype)

Stability:

- ☑ Experimental
- ☐ Proof of Concept (PoC)
- ☐ Stable with active development
- ☐ Stable with maintenance mode
- ☐ Legacy/End-of-life

Version History:

• Current Version: 0.1
• First Release: 2026 (Pending, MLSys submission)

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

• Release Frequency: N/A

Adoption:

• **Validated on:** AG News (4-class), DBpedia14 (14-class), Amazon Reviews (5-class), PubMed (5-class medical abstracts), SST-2.
• Scalability tested with 1000 clients, fragmentation tested up to 3000 clients.
• **Notable Users:** Fraunhofer HHI
• **Relevant Artifacts:** Code at https://github.com/an7123/FedSyn-Refine

---

10. Vendor/Community Support

Support Channels:

☐ Official Documentation
☐ Community Forum
☑ GitHub Issues
☐ Commercial Support (SLA available)
☑ Email Support
☐ Slack/Discord Community

Documentation Quality: ☐ Basic ☑ Good ☐ Excellent

Community Activity:

• GitHub Stars: N/A
• Contributors: Fraunhofer HHI Team
• Last Update: N/A

Training Resources:

☐ Official Training Available
☐ Tutorials/Guides
☐ Video Content
☐ Certification Programs

Contact for Questions: maximilian.andreas.hoefler@hhi.fraunhofer.de

Related Assets: [Links to related fact sheets]

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Asset Fact Sheet: FedXDS - XAI-Guided Data Sharing

Version: 1.0
Date: 2026-01-20
Author: FhHHI
Category: Privacy-Preserving Machine Learning / Explainable AI (XAI)

---

1. What It Does (Executive Summary)

FedXDS is a privacy-preserving framework designed to mitigate performance degradation in Federated Learning (FL) caused by heterogeneous (non-IID) data. It utilizes Explainable AI (XAI) attribution methods to identify and selectively share only the most task-relevant features from client data. These features are masked and perturbed with noise to satisfy Metric Differential Privacy, allowing the central server to utilize global data knowledge without accessing raw private data.

---

2. Why It Matters (Business Impact)

Data heterogeneity is a primary barrier to deploying FL in real-world scenarios like healthcare and manufacturing, where data distributions vary significantly between sites. FedXDS addresses this by enabling a compliant form of data sharing that transmits only the information strictly necessary for model improvement. Unlike standard FL that shares model gradients (which can leak information), FedXDS shares sparsified, noisy feature representations that are provably protected under Metric Differential Privacy. This makes it particularly attractive for regulated industries where both model performance and data protection are non-negotiable requirements. The XAI-driven approach also provides inherent interpretability, as stakeholders can inspect which features are deemed relevant and shared. By reducing the dimensionality of shared data, the framework also lowers communication overhead, making it practical for bandwidth-constrained deployments. The combination of formal privacy guarantees and demonstrated robustness against inference attacks provides a strong compliance narrative for cross-organizational collaborations.

Key Benefits:

- **Improved Model Accuracy:** Significantly outperforms standard FL methods (e.g., FedAvg, FedProx) in highly heterogeneous settings by aligning client distributions. Validated across multiple benchmark datasets including CIFAR-10/100, Tiny-ImageNet, CelebA, and FEMNIST.
- **Enhanced Privacy-Utility Trade-off:** Uses XAI to sparsify data before adding noise, allowing for stronger privacy guarantees (Differential Privacy) while retaining higher utility compared to protecting full raw images.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Defence Against Attacks:** The attribution-based masking provides robust protection against Membership Inference Attacks (MIA) and Feature Inversion attacks, as demonstrated in empirical evaluations.
- **Communication Efficiency:** Shares only sparse, masked representations, reducing bandwidth usage compared to raw data sharing approaches.

---

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

Applicable Regulations:

☐ Digital Governance Act
☐ Digital Markets Act
☐ Digital Services Act
☑ EU AI Act
☐ EU Cyber Resilience Act
☐ EU Cyber Security Act
☑ GDPR
☐ NIS2 Directive
☐ Other: [Specify]

Sector Specific Regulations:

☐ GxP
☐ HIPAA
☐ Medical Device Regulation
☐ Other: [Specify]

Compliance Support:

FedXDS supports the GDPR principle of Data Minimization by using XAI to strip away irrelevant information before transmission. It incorporates Metric Differential Privacy, providing a formal mathematical guarantee of privacy to support compliance with data protection standards for collaborative AI. For the EU AI Act, it addresses High-Risk AI System requirements around Robustness and Accuracy by improving FL performance in heterogeneous settings.

3.2 Guidelines and Standards Relevance

☐ ISO 13485
☐ ISO 42001
☐ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ ISO/IEC 23894 (AI Risk Management – conceptual alignment)
☑ ISO/IEC 27001 (Information Security)
☐ MPEG Feature Coding for Machines (FCM)
☐ NIST Risk Management Framework for AI
☐ OECD AI Principles
☐ Other: [Specify]

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☐ Yes ☑ No
Sample-level	Prevents identification of individual training examples	☑ Yes ☐ No
Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☐ Yes ☑ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No
Malicious server	Server deviates from protocol, sends crafted queries	☐ Yes ☑ No
External adversary	Attacker with black-box or white-box model access	☑ Yes ☐ No

Privacy Mechanism:

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Differential Privacy — Central DP
☐ Differential Privacy — Local DP
☐ Secure Aggregation
☐ Protected Activations / Learned Obfuscation
☐ Cryptographic (MPC / HE / TEE)
☑ DP Synthetic Data
☐ None (utility-focused baseline)
☑ Other: Metric Differential Privacy with XAI-guided sparsification

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☑ Yes ☐ No	Demonstrated robust protection in paper
Gradient Reconstruction Attack	☐ Yes ☑ No	—
Activation Inversion Attack	☑ Yes ☐ No	Feature Inversion attack evaluated in paper
Formal ε Verification	☑ Yes ☐ No	Metric DP ε values reported

Notes:
The attribution-based masking acts as a first line of defence by removing irrelevant features before noise is applied. This two-stage approach (sparsify then perturb) achieves a better privacy-utility trade-off than applying noise to full representations.

---

4. How It Works (Technical Approach)

The method operates in a two-stage process involving local attribution computation and central aggregation. During a warmup phase, the global model is trained using standard FL to produce a reasonable feature extractor. Each client then computes Layer-wise Relevance Propagation (LRP) attribution maps on its local data using the current global model. These attribution maps identify which pixels or features are most relevant to the classification task. A binary mask is created by thresholding the relevance scores, retaining only the top-k most informative features. The masked data is then

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

perturbed with calibrated Gaussian noise to satisfy Metric Differential Privacy guarantees. Clients upload these sparse, noisy representations to the server, which assembles a global auxiliary dataset. The server uses this dataset alongside the standard FL aggregation to train the global model using a hybrid objective that balances local data performance with knowledge from the shared global dataset. The sparsification step is crucial: by reducing dimensionality before adding noise, the sensitivity of the mechanism is lowered, allowing tighter privacy budgets with less utility loss.

Core Technology:

• Explainable AI (XAI): Uses propagation-based attribution methods (specifically Layer-wise Relevance Propagation - LRP) to generate relevance maps.
• Metric Differential Privacy: Applies Gaussian noise calibrated to the sensitivity of the masked features.

Key Components:

• Attribution-Guided Masking: Creates a binary mask to retain only pixels with high relevance scores.
• Privacy Mechanism: Adds noise only to the masked features, reducing the dimensionality and sensitivity of the shared data.
• Hybrid Training: Clients train using a composite objective function balancing local data performance with knowledge distilled from the shared global dataset.

---

5. Deployment Model

Available Options:

- ☑ Cloud (SaaS)
- ☑ On-Premise / Bare metal
- ☑ Container (Docker/Kubernetes)
- ☐ API Service
- ☑ Library/SDK
- ☑ Framework
- ☐ Guideline / Checklist

Recommended: Containerized deployment where the FedXDS client runs on local edge devices (or hospital servers) and connects to a central aggregation server.

---

6. Effort to Implement

Overall Complexity: ☐ Low ☑ Medium ☐ High
Time Estimate: 2-3 weeks (assuming existing FL infrastructure)

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Breakdown:

- **Preparation & Requirements:** 3-5 days
- **Setup & Configuration:** 1 week
- **Integration:** 3-5 days (integration with FL loop)
- **Testing & Validation:** 1 week
- **Training:** N/A (Part of the training process)

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- Existing Federated Learning setup (server/client architecture).
- Pre-training/Warmup phase capability (to generate initial attribution maps).

System Requirements:

- Python 3.8+
- PyTorch

Dependencies:

- Captum (or similar XAI library for attribution)
- Opacus (optional, for advanced DP accounting)

---

8. License, Cost & Classification

Software Type:

☑ Open-Source Software
☐ Commercial Software
☐ Cloud Service
☐ Library/SDK
☐ Platform Component
☐ Standalone Tool

License Type:

☑ Open Source (Apache 2.0 / MIT - check repository)
☐ Proprietary
☐ Freemium
☐ Enterprise License

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Cost Structure:

• Free (Open-Source implementation).
• Computational cost: Moderate overhead for computing attribution maps (single backward pass).
• Estimated Cost: Free

Commercial Restrictions:

• None (Standard Open-Source terms apply).

---

9. Maturity Level

Technical Readiness Level (TRL): 4-5 (Validated in lab environment)

Stability:

☐ Experimental
☑ Proof of Concept (PoC)
☐ Stable with active development
☐ Stable with maintenance mode
☐ Legacy/End-of-life

Version History:

• Current Version: 1.0
• First Release: 2024
• Release Frequency: Research Code

Adoption:

• **Validated on benchmarks:** CIFAR-10/100, Tiny-ImageNet, CelebA, FEMNIST.
• **Notable Users:** Fraunhofer HHI
• **Relevant Artifacts:** Code repository at `github.com/MaxH1996/FedXDS`

---

10. Vendor/Community Support

Support Channels:

☐ Official Documentation
☐ Community Forum
☑ GitHub Issues
☐ Commercial Support (SLA available)
☑ Email Support

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Slack/Discord Community

Documentation Quality: ☐ Basic ☑ Good ☐ Excellent

Community Activity:

- GitHub Stars: [Check Repo]
- Contributors: [Check Repo]
- Last Update: [Check Repo]

Training Resources:

☐ Official Training Available
☐ Tutorials/Guides
☐ Video Content
☐ Certification Programs

Contact for Questions: maximilian.andreas.hoefler@hhi.fraunhofer.de

Related Assets: [Links to related fact sheets]

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Asset Fact Sheet: FL-LR

Version: 0.1
Date: 2026-03-25
Author: Mafalda Fernandes, Daniel Castro, Nuno Santos, Luís Rodrigues
Category: Federated Learning / AI Auditability / Provenance / Trustworthy AI

1. What It Does (Executive Summary)

FL-LR is an auditability framework for federated learning (FL) systems, enabling post hoc analysis of training processes through verifiable logging and deterministic replay.

It records cryptographically verifiable metadata during training and provides an auditor-side replay engine that reconstructs and analyses the behaviour of aggregation algorithms. This allows auditors to investigate decisions such as client exclusion, detect poisoning attacks, and assess the correctness of defence mechanisms.

2. Why It Matters (Business Impact)

Federated learning is increasingly used in high-stakes domains such as healthcare and finance, where incorrect decisions or undetected attacks can have severe consequences. However, current FL systems suffer from lack of transparency, especially in server-side aggregation and defence mechanisms.

This creates critical risks: incorrect exclusion of benign participants; undetected model poisoning attacks; inability to justify decisions to regulators or stakeholders.

FL-LR addresses these issues by transforming FL from a "black box" into an auditable and explainable process.

By enabling replay-based forensic analysis, organizations can: - validate whether defences behaved correctly - identify malicious or faulty participants - justify decisions for compliance and accountability

Key Benefits:

- Post-Mortem Auditability of FL Systems Enables detailed forensic analysis after training without impacting runtime performance.
- Transparent and Explainable Aggregation Decisions Provides visibility into why clients were accepted or rejected by defence mechanisms.
- Forensic Replay and Investigation Capabilities Allows auditors to reconstruct training rounds and test alternative hypotheses or defences.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

Applicable Regulations:

- ☐ Digital Governance Act
- ☐ Digital Markets Act
- ☐ Digital Services Act
- ☑ EU AI Act
- ☑ EU Cyber Resilience Act
- ☑ EU Cyber Security Act
- ☑ GDPR
- ☑ NIS2 Directive
- ☑ Other: Digital Operational Resilience Act (DORA)

Sector Specific Regulations:

- ☐ GxP
- ☐ HIPAA
- ☐ Medical Device Regulation
- ☐ Other: [Specify]

Compliance Support:

FL-LR supports compliance by providing auditability, traceability, and explainability of federated learning processes.

For the EU AI Act, it enables logging, transparency, and post-hoc explainability of automated decision systems. Under GDPR, it contributes to accountability and integrity, ensuring that processing decisions can be reconstructed and justified.

In NIS2 and DORA contexts, FL-LR strengthens incident investigation and operational resilience, enabling forensic analysis of anomalous or compromised training processes.

3.2 Guidelines and Standards Relevance

- ☐ ISO 13485
- ☑ ISO 42001
- ☐ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)
- ☑ ISO/IEC 23894 (AI Risk Management – conceptual alignment)
- ☐ ISO/IEC 27001 (Information Security)
- ☐ MPEG Feature Coding for Machines (FCM)
- ☑ NIST Risk Management Framework for AI
- ☐ OECD AI Principles

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☑ Other: Data provenance and auditability best practices

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☐ Yes ☑ No
Sample-level	Prevents identification of individual training examples	☐ Yes ☑ No
Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☐ Yes ☑ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No
Malicious server	Server deviates from protocol, sends crafted queries	☐ Yes ☑ No
External adversary	Attacker with black-box or white-box model access	☐ Yes ☑ No

Privacy Mechanism:

☐ Differential Privacy — Central DP
☐ Differential Privacy — Local DP
☐ Secure Aggregation
☐ Protected Activations / Learned Obfuscation
☑ Cryptographic (MPC / HE / TEE)

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ DP Synthetic Data
☐ None (utility-focused baseline)
☐ Other: [Specify]

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☐ Yes ☑ No	[Link / Metric]
Gradient Reconstruction Attack	☐ Yes ☑ No	[Link / Metric]
Activation Inversion Attack	☐ Yes ☑ No	[Link / Metric]
Formal ε Verification	☐ Yes ☑ No	[ε value]

Notes:

FL-LR is not a privacy-preserving mechanism but an auditability and observability tool for federated learning systems.
It follows a minimal disclosure principle, where sensitive client updates are only accessed on demand and strictly when required for an audit. During training, only metadata (e.g., hashes, metrics) is recorded.

Privacy guarantees depend on the underlying FL system and any applied techniques such as secure aggregation or differential privacy. FL-LR complements these by enabling accountability without continuous exposure of sensitive data.

---

4. How It Works (Technical Approach)

FL-LR operates in two main phases: an online logging phase during training and an offline audit phase for replay and analysis.

During the online logging phase, FL-LR integrates directly into the federated learning server pipeline, where it passively records relevant information about each training round. This includes client identifiers, cryptographic hashes of model updates, intermediate defence metrics such as scores, rankings, and norms, as well as the final aggregation decisions. All collected information is structured into audit log entries and stored in an append-only hash chain, ensuring tamper-evident integrity and preserving the chronological order of events.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

In the offline audit phase, an external auditor retrieves the committed logs and, when necessary, additional evidence such as the original client updates. Using a Directed Acyclic Graph (DAG)-based Replay Engine, the auditor can reconstruct the execution of specific training rounds in a deterministic manner. This replay capability allows auditors not only to validate the correctness of past decisions but also to extend the analysis by introducing alternative diagnostic modules, enabling deeper forensic investigations into the behaviour of the federated learning process.

Core Technology:

- Federated Learning frameworks (e.g., ByzFL)
- Cryptographic hash chains
- Trusted Execution Environments (e.g., Confidential Virtual Machines)
- DAG-based replay engines

Key Components:

- Server-side accountability layer (logging hooks)
- Audit log (append-only, tamper-evident)
- Auditor toolkit (Log Manager, Evidence Collector)
- Replay Engine (modular DAG execution)

---

5. Deployment Model

FL-LR is deployed as an extension to federated learning systems.

Available Options:

- ☑ Cloud (SaaS)
- ☐ On-Premise / Bare metal
- ☐ Container (Docker/Kubernetes)
- ☐ API Service
- ☐ Library/SDK
- ☑ Framework
- ☐ Guideline / Checklist

Recommended: Deploy alongside FL frameworks (e.g., ByzFL) with CVM-backed infrastructure for secure logging.

---

6. Effort to Implement

Overall Complexity: ☐ Low ☐ Medium ☑ High

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Time Estimate: 6--10 weeks (depending on integration depth)

Breakdown:

- **Preparation & Requirements:** 2 weeks
- **Setup & Configuration:** 1--2 week
- **Integration:** 2--3 weeks
- **Testing & Validation:** 2--3 weeks
- **Training:** 1 week

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- Federated learning environment (cross-silo preferred)
- Understanding of aggregation defences (e.g., MultiKrum)
- Access to CVM-enabled infrastructure (optional but recommended)

System Requirements:

- FL framework (e.g., ByzFL)
- Storage for audit logs and optional evidence
- Auditor-side compute environment

Dependencies:

- Cryptographic hash functions
- DAG execution framework
- Secure storage for audit logs and client updates

---

8. License, Cost & Classification

Software Type:

- ☑ Open-Source Software
- ☐ Commercial Software
- ☐ Cloud Service
- ☐ Library/SDK
- ☑ Platform Component
- ☐ Standalone Tool

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

License Type:

- ☑ Open Source (Apache 2.0)
- ☐ Proprietary
- ☐ Freemium
- ☐ Enterprise License

Cost Structure:

- Free (research prototype; infrastructure costs depend on cloud provider)
- Estimated Cost: Cloud CVM instance costs only

Commercial Restrictions:

- None

---

9. Maturity Level

Technical Readiness Level (TRL): N/A

Stability:

- ☑ Experimental
- ☐ Proof of Concept (PoC)
- ☐ Stable with active development
- ☐ Stable with maintenance mode
- ☐ Legacy/End-of-life

Version History:

- Current Version: N/A
- First Release: N/A
- Release Frequency: N/A

Adoption:

- N/A
- **Notable Users:** N/A
- **Relevant Artifacts:** N/A

---

10. Vendor/Community Support

Support Channels:

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Official Documentation
☐ Community Forum
☐ GitHub Issues
☐ Commercial Support (SLA available)
☐ Email Support
☐ Slack/Discord Community

Documentation Quality: ☑ Basic ☐ Good ☐ Excellent

Community Activity:

- GitHub Stars: N/A
- Contributors: 1 (INESC-ID)
- Last Update: March 2026

Training Resources:

☐ Official Training Available
☐ Tutorials/Guides
☐ Video Content
☐ Certification Programs

Contact for Questions: Mafalda Fernandes (mafalda.m.fernandes@tecnico.ulisboa.pt), Daniel Castro (daniel.castro@tecnico.ulisboa.pt)

Related Assets: N/A

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Asset Fact Sheet: Privacy-Preserving Sparse Collaborative Inference

Version: 1.0
Date: 2026-01-20
Author: FhHHI
Category: Edge AI / Collaborative Inference / Privacy

---

1. What It Does (Executive Summary)

This asset is a framework for privacy-preserving Collaborative Inference (CI) that leverages activation sparsity as a dual-purpose mechanism for both communication efficiency and formal privacy guarantees. A lightweight Sparse Autoencoder (SAE) learns a sparse representation of intermediate activations, which is then protected by a novel two-channel noise mechanism: Gumbel noise obfuscates which features are active (index channel), while Gaussian noise protects their numerical values (value channel). The framework provides tunable, information-theoretic privacy budgets with provable lower bounds on adversarial reconstruction error.

---

2. Why It Matters (Business Impact)

Collaborative inference splits deep neural networks between resource-constrained edge devices and powerful servers but transmitting intermediate activations creates both a communication bottleneck and a severe privacy risk from exposing data to potentially untrusted servers. Existing solutions address these challenges separately: compression methods lack privacy guarantees, while privacy defences ignore communication costs or impose heavy computational overhead. This framework is the first to systematically unify sparsity with formal privacy guarantees, treating sparsity as a dual-purpose tool. The SAE learns to concentrate task-relevant information into a small fraction of features (as low as 3%), enabling extreme compression while creating the structural foundation for rigorous privacy accounting. The two-channel noise design provides independent, tunable privacy controls for the index pattern and activation values, allowing practitioners to precisely calibrate the privacy-utility trade-off. Evaluations on CIFAR-10, Tiny-ImageNet, and the privacy-sensitive FaceScrub dataset demonstrate state-of-the-art privacy-utility trade-offs, sustaining high accuracy at up to 97% sparsity while offering superior resilience against strong GAN-based model inversion attacks.

Key Benefits:

- **Dual-Purpose Sparsity:** Achieves up to 97% sparsity, simultaneously reducing communication costs (compressed to <1% of original size via NNCodec) and providing the structural foundation for formal privacy guarantees.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Formal Privacy Guarantees:** Provides rigorous information-theoretic bounds linking noise parameters to a quantifiable privacy budget in bits, with a provable lower bound on any adversary's reconstruction error (Corollary 4.5).
- **State-of-the-Art Privacy-Utility Trade-off:** Achieves the highest reconstruction MSE (best privacy) across all benchmarks while maintaining competitive task accuracy compared to existing defences including PATROL, ARL, and Noise-ARL.
- **Lightweight Client-Side Mechanism:** Requires only a single forward pass through the SAE encoder on the edge device, making it practical for real-world deployment on resource-constrained hardware.

---

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

Applicable Regulations:

☐ Digital Governance Act
☐ Digital Markets Act
☐ Digital Services Act
☑ EU AI Act
☐ EU Cyber Resilience Act
☐ EU Cyber Security Act
☑ GDPR
☐ NIS2 Directive
☐ Other: [Specify]

Sector Specific Regulations:

☐ GxP
☐ HIPAA
☐ Medical Device Regulation
☐ Other: [Specify]

Compliance Support:

The framework supports GDPR principles of Data Minimization by transmitting only a sparse, noise-protected subset of intermediate activations (3-15% of features). The formal information-theoretic privacy bounds provide a quantifiable privacy budget in bits, offering a clear compliance narrative for data protection requirements. The provable lower bound on reconstruction error gives a mathematical guarantee that adversaries cannot recover input data beyond a certifiable threshold.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

3.2 Guidelines and Standards Relevance

☐ ISO 13485
☐ ISO 42001
☑ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)
☐ ISO/IEC 23894 (AI Risk Management – conceptual alignment)
☑ ISO/IEC 27001 (Information Security)
☑ MPEG Feature Coding for Machines (FCM)
☐ NIST Risk Management Framework for AI
☐ OECD AI Principles
☐ Other: [Specify]

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☐ Yes ☑ No
Sample-level	Prevents identification of individual training examples	☐ Yes ☑ No
Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☑ Yes ☐ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No
Malicious server	Server deviates from protocol, sends crafted queries	☐ Yes ☑ No

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

External adversary	Attacker with black-box or white-box model access	☑ Yes ☐ No

Privacy Mechanism:

☐ Differential Privacy — Central DP
☐ Differential Privacy — Local DP
☐ Secure Aggregation
☑ Protected Activations / Learned Obfuscation
☐ Cryptographic (MPC / HE / TEE)
☐ DP Synthetic Data
☐ None (utility-focused baseline)
☑ Other: Information-theoretic two-channel noise (Gumbel index obfuscation + Gaussian value obfuscation)

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☐ Yes ☑ No	Not in scope
Gradient Reconstruction Attack	☑ Yes ☐ No	—
Activation Inversion Attack	☑ Yes ☐ No	GAN-based and decoder-based inversion attacks evaluated; state-of-the-art resilience demonstrated
Formal ε Verification	☑ Yes ☐ No	

4. How It Works (Technical Approach)

The framework operates in two phases. In the training phase, a neural network is split between client and server, and a Sparse Autoencoder (SAE) is inserted at the split point. The SAE encoder projects the dense intermediate features into a higher-dimensional overcomplete basis (4× expansion),

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

and a hard Top-K operator enforces a strict sparsity budget (typically 3% non-zero features). The SAE is trained with a dual objective: the main task loss for end-to-end utility, and a local reconstruction loss to ensure the sparse code faithfully represents the original features. During inference, the client runs the initial network layers and the SAE encoder. Before transmission, two independent noise mechanisms are applied: Gumbel noise is injected into the pre-activation scores before Top-K selection to randomize which features are selected (index channel obfuscation), and Gaussian noise is added to the selected non-zero activation values (value channel obfuscation). The noise scales are calibrated from formal information-theoretic bounds (Propositions 4.2 and 4.3) to meet target privacy budgets in bits. The protected sparse activations are compressed using NNCodec (ISO/IEC standard) and transmitted to the server, which decompresses and runs the remaining network layers. Theorem 4.4 bounds the total information leakage, and Corollary 4.5 translates this into a provable lower bound on any adversary's reconstruction error.

Core Technology:

• Sparse Autoencoder (SAE): Learned overcomplete basis with hard Top-K sparsity for deterministic payload size.
• Two-Channel Noise Mechanism: Gumbel noise for index channel obfuscation, Gaussian noise for value channel obfuscation.
• Information-Theoretic Privacy: Formal mutual information bounds with provable reconstruction error lower bounds.
• NNCodec: ISO/IEC standard neural network compression for sparse activation tensors.

Key Components:

• Client Side: Feature extraction (initial network layers) → SAE encoder → Gumbel-perturbed Top-K selection → Gaussian noise injection → NNCodec compression → transmission.
• Server Side: NNCodec decompression → SAE decoder → remaining network layers → prediction.
• Privacy Calibration: Adaptive per-batch noise calibration from empirical activation variances and target bit budgets.

---

5. Deployment Model

Available Options:

☐ Cloud (SaaS)
☑ On-Premise / Bare metal
☑ Container (Docker/Kubernetes)
☐ API Service
☑ Library/SDK
☐ Framework

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

☐ Guideline / Checklist

Recommended: Edge-cloud split architecture where the client runs on resource-constrained devices (NVIDIA Jetson, Raspberry Pi, mobile) and communicates with a central GPU inference server. The SAE encoder adds only a single forward pass to client-side computation.

---

6. Effort to Implement

Overall Complexity: ☐ Low ☑ Medium ☐ High
Time Estimate: 2-3 weeks

Breakdown:

- **Preparation & Requirements:** Selecting split point in neural network, SAE architecture sizing (2-3 days)
- **Setup & Configuration:** SAE training with dual-objective loss, privacy parameter calibration (1 week)
- **Integration:** NNCodec compression pipeline, noise injection mechanisms (3-5 days)
- **Testing & Validation:** Privacy-utility evaluation, inversion attack benchmarking (1 week)
- **Training:** SAE fine-tuning ~40 epochs with cosine annealing

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- A pretrained deep learning model (CNN or Transformer) suitable for splitting.
- Network connectivity between client and server.

System Requirements:

- Client: Edge device capable of running partial neural network + SAE encoder forward pass.
- Server: GPU for remaining network layers and SAE decoder.

Dependencies:

- PyTorch
- NNCodec (ISO/IEC standard implementation)
- AdamW optimizer with cosine annealing scheduler

---

8. License, Cost & Classification

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Software Type:

- ☑ Open-Source Software
- ☐ Commercial Software
- ☐ Cloud Service
- ☐ Library/SDK
- ☐ Platform Component
- ☐ Standalone Tool

License Type:

- ☑ Open Source (Code available on GitHub)
- ☐ Proprietary
- ☐ Freemium
- ☐ Enterprise License

Cost Structure:

• Free (Open-Source research code).
• Operational Savings: Drastic reduction in communication costs (<1% of original activation size after compression).
• Client overhead: Minimal (single SAE encoder forward pass, e.g., 2.57s inference for sHidden=4 on FaceScrub).
• Estimated Cost: Free software; edge and server hardware required

Commercial Restrictions:

• NNCodec is ISO/IEC standards-based; check specific licensing terms for commercial deployment.

---

9. Maturity Level

Technical Readiness Level (TRL): 4 (Validated in lab on standard benchmarks)

Stability:

- ☐ Experimental
- ☑ Proof of Concept (PoC)
- ☐ Stable with active development
- ☐ Stable with maintenance mode
- ☐ Legacy/End-of-life

Version History:

• Current Version: 1.0

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

• First Release: 2026 (WACV)
• Release Frequency: Research Code

Adoption:

• **Validated on:** CIFAR-10, Tiny-ImageNet, FaceScrub.
• Outperforms state-of-the-art baselines (PATROL, ARL, Noise-ARL, Top-K, DistCorr, Dropout, Bottleneck, Laplace Noise).
• **Notable Users:** Fraunhofer HHI
• **Relevant Artifacts:** Code at https://github.com/an7123/privacy_ci

---

10. Vendor/Community Support

Support Channels:

☐ Official Documentation
☐ Community Forum
☑ GitHub Issues
☐ Commercial Support (SLA available)
☑ Email Support
☐ Slack/Discord Community

Documentation Quality: ☐ Basic ☑ Good ☐ Excellent

Community Activity:

• GitHub Stars: [Check Repo]
• Contributors: Fraunhofer HHI Team
• Last Update: [Check Repo]
• Builds on the broader MPEG/JPEG AI and NNCodec standardization community.

Training Resources:

☐ Official Training Available
☐ Tutorials/Guides
☐ Video Content
☐ Certification Programs

Contact for Questions: maximilian.andreas.hoefler@hhi.fraunhofer.de

Related Assets: Relevance-Guided Collaborative Inference (related CI framework using LRP-based static masking + NNCodec)

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Asset Fact Sheet: VeriProv

Version: 0.1
Date: 2026-03-25
Author: Daniel Ferreira, Daniel Castro, Nuno Santos, Luís Rodrigues
Category: AI Governance / Provenance / Confidential Computing / Cloud Security

---

1. What It Does (Executive Summary)

VeriProv is a framework for capturing, enforcing, and auditing verifiable and non-repudiable provenance across the full lifecycle of machine learning (ML) pipelines executed in cloud environments. It combines Trusted Execution Environments (TEEs) and cryptographically verifiable provenance records to ensure that every ML phase (data processing, training, evaluation, inference) is executed as expected, bound to a responsible actor, and recorded in a tamper-evident provenance graph.

---

2. Why It Matters (Business Impact)

Cloud-based AI systems introduce significant trust and accountability gaps. Organizations often lack visibility into how models were trained, what data was used, and who performed each step. This creates risks in regulatory compliance (e.g., AI Act, GDPR); auditability and accountability; insider threats and malicious data/model manipulation.

VeriProv addresses these challenges by enabling end-to-end verifiable ML pipelines with strong guarantees on execution integrity and actor accountability.

By binding each ML phase to a cryptographic identity and storing attestable records, VeriProv ensures that actions cannot be denied (non-repudiation) and that workflows can be independently audited.

Key Benefits:

- End-to-End Verifiable Provenance Captures cryptographically verifiable records for all ML lifecycle phases, enabling full auditability and traceability of data, models, and computations.
- Non-Repudiable Accountability Binds each pipeline step to a signed actor identity, ensuring that participants cannot deny their actions and enabling strong accountability.
- Policy-Enforced ML Execution Enforces user-defined policies (e.g., fairness, accuracy thresholds) at runtime inside trusted environments, preventing non-compliant executions.

---

3. Regulatory, Guidelines, Standards and Privacy Relevance

3.1 Regulatory Compliance

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Applicable Regulations:

- ☐ Digital Governance Act
- ☐ Digital Markets Act
- ☐ Digital Services Act
- ☑ EU AI Act
- ☑ EU Cyber Resilience Act
- ☑ EU Cyber Security Act
- ☑ GDPR
- ☑ NIS2 Directive
- ☑ Other: Digital Operational Resilience Act (DORA)

Sector Specific Regulations:

- ☐ GxP
- ☐ HIPAA
- ☐ Medical Device Regulation
- ☐ Other: [Specify]

Compliance Support:

VeriProv strengthens compliance by enabling auditability, traceability, and accountability of AI systems. It directly supports requirements in the EU AI Act related to logging, transparency, and risk management by maintaining verifiable records of all ML lifecycle phases.

For GDPR, it contributes to accountability and integrity principles, ensuring that data processing activities are traceable and verifiable. Under NIS2 and DORA, it enhances operational resilience and incident investigation capabilities by providing tamper-evident execution logs.

3.2 Guidelines and Standards Relevance

- ☐ ISO 13485
- ☑ ISO 42001
- ☐ ISO/IEC 15938-17 (Neural Network Coding / NNCodec)
- ☑ ISO/IEC 23894 (AI Risk Management – conceptual alignment)
- ☐ ISO/IEC 27001 (Information Security)
- ☐ MPEG Feature Coding for Machines (FCM)
- ☑ NIST Risk Management Framework for AI
- ☐ OECD AI Principles
- ☑ Other: Data provenance and auditability best practices

3.3 Privacy Protection (For FL / Collaborative Inference Tools) [x] Applicable [ ] Not Applicable

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Privacy Objective:

Level	Definition	Provided?
User-level	Prevents inference about which users participated in training	☐ Yes ☑ No
Sample-level	Prevents identification of individual training examples	☐ Yes ☑ No
Reconstruction	Prevents recovery of raw inputs from e.g. gradients/activations	☐ Yes ☑ No

Threat Model:

Threat Model	Description	Addressed?
Honest-but-curious server	Server follows protocol but inspects received data	☑ Yes ☐ No
Malicious server	Server deviates from protocol, sends crafted queries	☑ Yes ☐ No
External adversary	Attacker with black-box or white-box model access	☐ Yes ☑ No

Privacy Mechanism:

- ☐ Differential Privacy — Central DP
- ☐ Differential Privacy — Local DP
- ☐ Secure Aggregation
- ☐ Protected Activations / Learned Obfuscation
- ☑ Cryptographic (MPC / HE / TEE)
- ☐ DP Synthetic Data
- ☐ None (utility-focused baseline)
- ☐ Other: [Specify]

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Privacy Auditing:

Audit Type	Performed?	Result / Reference
Membership Inference Attack (MIA)	☐ Yes ☑ No	[Link / Metric]
Gradient Reconstruction Attack	☐ Yes ☑ No	[Link / Metric]
Activation Inversion Attack	☐ Yes ☑ No	[Link / Metric]
Formal ε Verification	☐ Yes ☑ No	[ε value]

Notes:

VeriProv does not provide intrinsic privacy-preserving mechanisms such as differential privacy. Instead, it focuses on accountability and auditability, ensuring that all data processing steps are recorded and verifiable.

Sensitive data is protected during execution through Trusted Execution Environments (TEEs), while provenance records are encrypted and accessible only to authorized auditors. Privacy guarantees depend on the security of the underlying hardware and correct system configuration.

For stronger privacy guarantees (e.g., resistance to inference attacks), VeriProv should be combined with complementary techniques such as differential privacy or secure multi-party computation.

---

4. How It Works (Technical Approach)

VeriProv executes each machine learning phase inside TEEs, ensuring isolated and verifiable execution.

For every phase (data processing, training, evaluation, inference), the system generates an attestable provenance record containing hashes of inputs and outputs; code identifiers; actor identity (via digital signatures); attestation proof of execution environment.

These records are stored in an encrypted provenance graph, where nodes represent computations and edges capture dependencies between phases.

A policy engine enforces user-defined constraints (e.g., accuracy thresholds, dataset properties) before execution. Policies are evaluated within the TEE using the provenance graph, ensuring they cannot be bypassed.

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Core Technology:

- Trusted Execution Environments (e.g., Confidential Virtual Machines)
- Remote Attestation
- Cryptographic Signatures (non-repudiation)
- Provenance Graph Databases (e.g., RDF, property graphs)

Key Components:

- CVM-based execution environment
- Record Generation module
- Provenance Store (encrypted graph)
- Policy Engine (runtime enforcement)

---

5. Deployment Model

VeriProv is deployed on cloud infrastructures supporting Confidential Virtual Machines.

Available Options:

- ☑ Cloud (SaaS)
- ☐ On-Premise / Bare metal
- ☐ Container (Docker/Kubernetes)
- ☐ API Service
- ☐ Library/SDK
- ☐ Framework
- ☐ Guideline / Checklist

Recommended: Deploy in Google Cloud, Azure or AWS, choose AMD-SEV servers.

After the attestation step is completed, any software can be run within the Confidential Virtual Machine.

---

6. Effort to Implement

Overall Complexity: ☐ Low ☐ Medium ☑ High
Time Estimate: 6--10 weeks (depending on integration depth)

Breakdown:

- **Preparation & Requirements:** 2 weeks
- **Setup & Configuration:** 1--2 week
- **Integration:** 2--3 weeks

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

- **Testing & Validation:** 2--3 weeks
- **Training:** 1 week

---

7. Prerequisites & Dependencies

Technical Prerequisites:

- Access to CVM-enabled cloud infrastructure
- Public Key Infrastructure (PKI) for identity binding
- Understanding of provenance systems and attestation

System Requirements:

- Confidential VM support
- Secure key management for actors and auditors
- Graph database (e.g., RDF or Neo4j)

Dependencies:

- Cloud attestation services
- Cryptographic libraries for signing and verification (in development)
- Provenance storage/query systems (in development)

---

8. License, Cost & Classification

Software Type:

☑ Open-Source Software
☐ Commercial Software
☐ Cloud Service
☐ Library/SDK
☐ Platform Component
☐ Standalone Tool

License Type:

☑ Open Source (Apache 2.0)
☐ Proprietary
☐ Freemium
☐ Enterprise License

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Cost Structure:

- Free (research prototype; infrastructure costs depend on cloud provider)
- Estimated Cost: Cloud CVM instance costs only

Commercial Restrictions:

- None

---

9. Maturity Level

Technical Readiness Level (TRL): N/A

Stability:

- ☑ Experimental
- ☐ Proof of Concept (PoC)
- ☐ Stable with active development
- ☐ Stable with maintenance mode
- ☐ Legacy/End-of-life

Version History:

- Current Version: N/A
- First Release: N/A
- Release Frequency: N/A

Adoption:

- N/A
- **Notable Users:** N/A
- **Relevant Artifacts:** N/A

---

10. Vendor/Community Support

Support Channels:

- ☐ Official Documentation
- ☐ Community Forum
- ☐ GitHub Issues
- ☐ Commercial Support (SLA available)
- ☐ Email Support
- ☐ Slack/Discord Community

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.

Documentation Quality: ☑ Basic ☐ Good ☐ Excellent

Community Activity:

- GitHub Stars: N/A
- Contributors: 1 (INESC-ID)
- Last Update: March 2026

Training Resources:

☐ Official Training Available
☐ Tutorials/Guides
☐ Video Content
☐ Certification Programs

Contact for Questions: Daniel Ferreira (unidsferreira2003@gmail.com), Daniel Castro (daniel.castro@tecnico.ulisboa.pt)

Related Assets: N/A

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101189689.