netidee

STIPENDIEN

Advancing Privacy in Federated Learning

Zwischenbericht | Call 20 | Stipendium ID 7832

Inhalt

1 Introduction

Artificial intelligence (AI) has become integral to virtually every industry, increasingly operating on sensitive, private, or legally protected data. Traditional AI systems require data to be collected centrally for model training — an approach that is often infeasible due to legal constraints, privacy regulations, or competitive concerns. Federated Learning (FL) addresses this by enabling collaborative model training without centralizing raw data: models are trained locally on each participant's device or server, and only model updates are shared, while the underlying data remains at its source[1].

Despite this privacy-preserving design, FL is not free of residual privacy risks. Shared model updates represent an abstraction of the original training data, and several classes of attacks have demonstrated that sensitive information can be reconstructed or inferred from them. These include membership inference attacks,[2] property inference attacks,[3] and model inversion attacks[4] [5]. Understanding and mitigating these risks is particularly critical in domains such as healthcare and finance, where FL is most promising yet the consequences of privacy breaches are most severe.

The present project, Advancing Privacy in Federated Learning, addresses this challenge through a three-phase research design. Phase 1 identifies and formalizes relevant privacy threats through expert interviews and threat modelling. Phase 2 empirically evaluates how these threats manifest under realistic FL conditions, including heterogeneous data distributions. Phase 3 investigates mitigation strategies — with a focus on Differential Privacy mechanisms — and analyses the resulting trade-offs between privacy and model utility. The ultimate deliverable is a comprehensive open-source privacy risk management framework for FL systems, grounded in both empirical evidence and practitioner input.

This report covers the project period from November 2025 to May 2026. It presents the progress achieved across all milestones, describes a minor but methodologically motivated update to the project timeline, and outlines the work planned for the remainder of the project. The report is structured as follows: Section 2 provides the status of each milestone; Section 3 summarizes the plan update.

---

[1] McMahan et al. (2017), "Communication-Efficient Learning of Deep Networks from Decentralized Data" — the original FL paper, Google. *AISTATS 2017.*

[2] Shokri et al. (2017), "Membership Inference Attacks Against Machine Learning Models" — *IEEE S&P 2017.* (the seminal MIA paper)

[3] Ganju et al. (2018), "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations" — *ACM CCS 2018.*

[4] Fredrikson et al. (2015), "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures" — *ACM CCS 2015.*

[5] Zhao et al. (2020), "iDLG: Improved Deep Leakage from Gradients" — covers gradient inversion in FL specifically.

2 Status

2.1 **Milestone 1** – Research Methodology

The main activities of Milestone 1 (November 2025 – February 2026) focused on refining research questions and establishing the methodological foundation. A systematic review of existing literature on privacy attacks in FL was conducted, with particular focus on Membership Inference Attacks (MIAs). Three research questions were formulated: RQ1 investigates which privacy risks and threat scenarios are relevant to domain stakeholders and how they can be systematically modelled; RQ2 examines how identified risks can be quantified under realistic FL conditions including non-IID data distributions; and RQ3 explores how risks can be mitigated and what privacy-utility trade-offs arise.
The research methodology was defined and is presented in Figure 1. The three-phase design covers risk identification (RQ1), empirical risk evaluation (RQ2), and risk mitigation (RQ3). The dissertation chapter on methodology was completed on schedule.

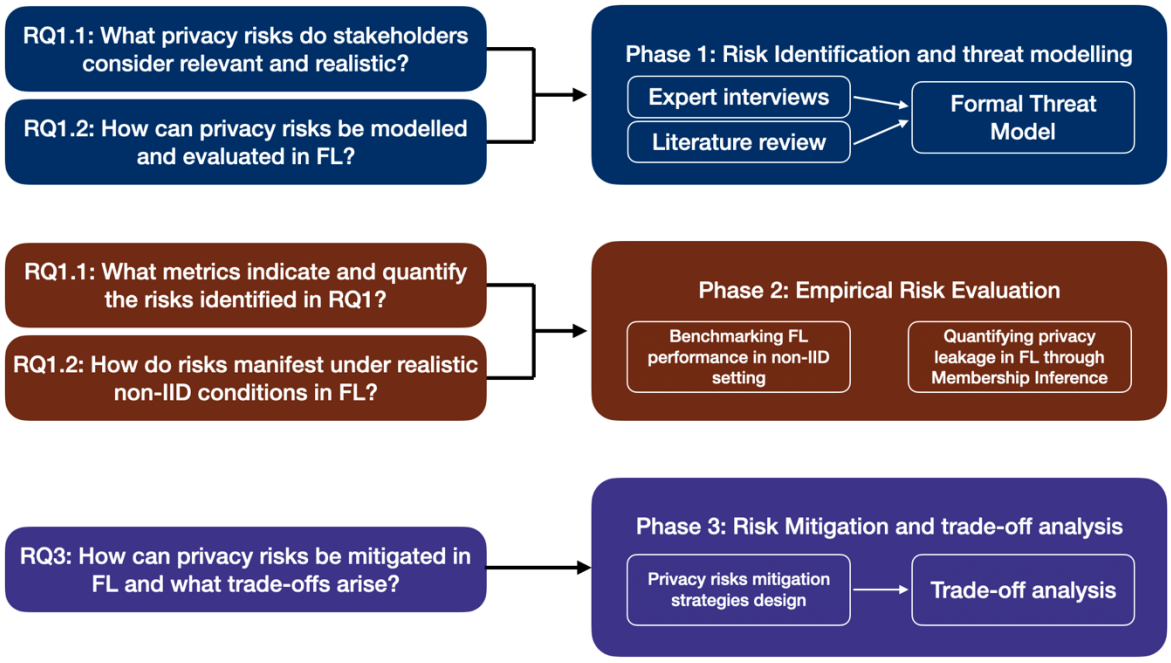

Figure 1: Research design and methodology

A key methodological decision was to ground the threat model empirically through expert interviews with domain stakeholders (RQ1.1). These interviews aim to provide a deeper understanding of privacy risks considered relevant in practice, as well as potential mitigation strategies from a practitioner perspective. While this represents an extension of the original plan, it strengthens the empirical foundation of the work and is expected to make the results more applicable to real-world FL deployments. The interview questionnaire has been finalized and the interviews, along with results analysis, are planned to take place over the next two months. The milestone was completed without significant deviations from the original plan.

2.2 Milestone 2 – Prototype Implementation

Milestone 2 (February 2026 – April 2026) covers the design and implementation of the proposed method and the development of a stable experimental prototype. FL training pipelines were implemented and tested under both IID (identically and independently distributed) and non-IID data settings to establish a comprehensive benchmark and baseline for subsequent privacy evaluations. Data heterogeneity across clients is a well-documented characteristic of real-world FL deployments, and non-IID settings are known to affect both model convergence and privacy vulnerability — making their inclusion in the benchmark essential for producing realistic and applicable results.

The prototype supports configurable federation parameters including the number of participating clients, the model aggregation strategy, and the number of local training rounds per communication cycle. This flexibility allows systematic variation of experimental conditions, which is necessary to isolate the effect of individual federation design choices on privacy risk. Baseline model performance metrics — including accuracy and loss convergence — have been recorded across multiple model architectures and dataset configurations, establishing the reference point against which privacy-utility trade-offs will be measured in the subsequent phases.

This infrastructure is directly essential for the empirical risk evaluation planned under RQ2, specifically for measuring Membership Inference Attack (MIA) success rates under varied and realistic FL conditions. The benchmark will enable a controlled comparison of privacy leakage and also impact of mitigations across different federation configurations, contributing to the privacy-utility trade-off evaluation in Phase 3. As a parallel research output during this period, a peer-reviewed journal paper was published[6].

2.3 Milestone 3 – Experimental Evaluation

Milestone 3 (April 2026 – May 2026) covers the execution of comprehensive experiments and evaluation of privacy and performance metrics. FL training experiments under IID and non-IID conditions are ongoing. A minor plan update applies: the full execution of privacy metric evaluations is partly dependent on the outcomes of the expert interviews (RQ1), which will identify the most relevant threat scenarios and metrics from a practitioner perspective. A portion of the experiments will therefore be completed after the interview analysis. This is a deliberate, methodologically justified adjustment.

[6] A. Pustozerova et al., “Lightweight Techniques for Federated Anomaly Detection in Log Data,” IEEE Transactions on Reliability, vol. 75, pp. 1641–1655, 2026, doi: 10.1109/TR.2026.3677674.

3 Summary of Plan Update

The project is progressing largely according to the original plan, with one structured update to the timeline. A subset of the Milestone 3 experiments — specifically the privacy metric evaluations — will extend beyond the original May 2026 timeline. This adjustment is methodologically justified: the expert interviews (RQ1) will determine which threat scenarios and metrics are most relevant for the empirical evaluation phase (RQ2), and running the full experiment suite before this input is available would risk producing results disconnected from realistic, practitioner-relevant threat scenarios.

To reflect the expanded scope of the interview-based work, a new Milestone 5 has been added to the project plan, covering the design, execution, and analysis of the expert interviews (June 2026 – July 2026). Consequently, Milestone 6 — covering the analysis and interpretation of experimental results — has been repositioned to follow the interview analysis, with the extended experiments expected to be completed around August 2026. All other project activities remain on schedule and within the original scope.