POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Safety-Aware Aggregation for Federated Fine-Tuning of Large Language Models

Laurea Magistrale in Computer Science and Engineering – Ingegneria Informatica

Author: Leonardo Lei

Advisor: Prof. Cinzia Cappiello

Co-advisor: Mattia Sabella

Academic year: 2025–2026

1. Introduction & Background

As LLMs are deployed with real-world consequences, *alignment*, the process of shaping model behavior to reflect human intentions, has become a fundamental requirement for trustworthy AI [8]. The stakes are highest in *dual-use* settings, where the same knowledge is benign in one context yet dangerous in another: explaining a pathogen's mechanism can protect public health or enable bioterrorism. This harm is severe and irreversible, and hazardous knowledge is inseparable from legitimate scientific expertise [9]. Fine-tuning is the primary mechanism for implanting or removing such capabilities, but it normally requires updating billions of parameters, a huge computational and memory cost. Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) [4] reduces this overhead by orders of magnitude, making adaptation feasible on standard consumer hardware through low-rank matrix injection into frozen Transformer layers. Growing privacy concerns and strict data-sharing regulations have driven the rise of collaborative paradigms such as Federated Learning (FL) [6], where clients train a shared model while raw data never leaves local devices. Applying FL to LLMs, however, is hindered by the scale of parameters exchanged at every round; PEFT mitigates this barrier by shrinking that footprint, making FL practical even for billion-parameter models.

Existing safety defenses often intervene only after training: machine unlearning [12] and circuit-breaking [13] surgically remove or suppress hazardous capabilities only once they have already been absorbed into model weights, requiring extra computation. Deep Ignorance [7] instead intervenes earlier, at the pretraining data stage, using a Risk Estimator [11] classifier, a Masked Language Model (MLM) [10], to filter hazardous content before it ever reaches the model. However, this filtering step presupposes server-side access to the data, which the federated setting forbids by design, a constraint we term the *blind server problem*.

This work bridges this gap and tries to find novel solutions beyond simple data selection. Our answer relocates data assessment entirely to the client, where raw data is locally accessible, and couples **risk-aware local training** with safety enforcement in the **aggregation space**, where the server is permitted to operate, guided by a standardized per-sample risk signal computed client-side.

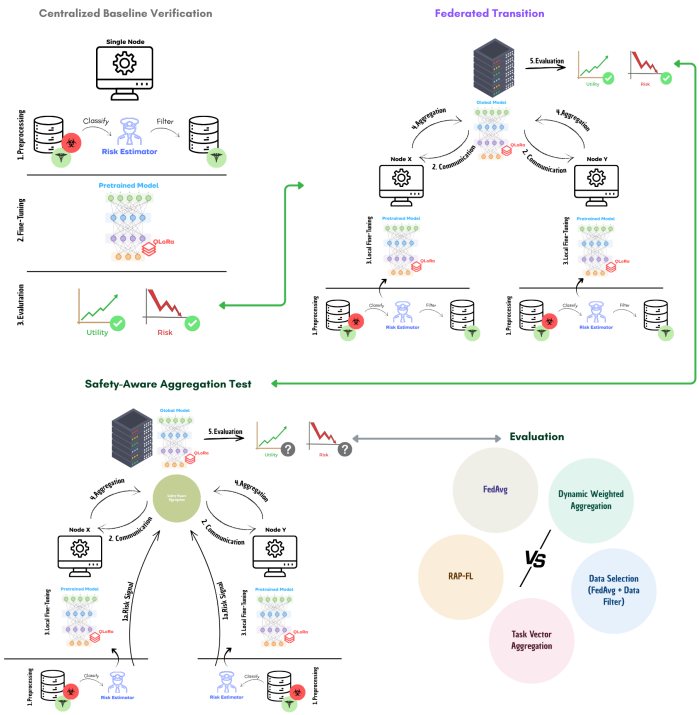

Figure 1: Methodology Overview.

2. Goals and Research Questions

The goal of this thesis is to design a secure, decentralized fine-tuning framework that prevents the acquisition of dual-use capabilities without violating the privacy constraints of a federated server. To this end, the research systematically investigates: (i) the viability of proactive data-filtering within a Parameter-Efficient (QLoRA) regime; (ii) the replicability of such data-filtering safety properties within a Federated Learning framework; (iii) the development of risk-aware, weight-space interventions to achieve safety-robustness beyond data selection; and (iv) the resulting trade-off between risk mitigation (safety) and general reasoning capabilities (utility) of the model.

3. Proposed Methodology

We structure the work as a four-module pipeline (Fig. 1), where each module gates the transition to the next.

3.1. Module 1 – Centralized Baseline Verification

The pipeline opens on a *single node* and proceeds through three sequential steps, exactly as depicted in Fig. 1.
(1) Preprocessing. A ***Risk Estimator*** assigns a continuous threat probability $P_{\text{threat}}(x) \in [0, 1]$ to every training sample. **(2) QLoRA Fine-Tuning.** The network's parameters are mathematically decomposed as $\Theta = \Theta_0 \oplus \Phi$, where Θ_0 represents the strictly frozen pre-trained base weights, and Φ denotes the trainable low-rank PEFT adapters. The adaptation is performed via QLoRA [2] on controlled dataset compositions (Safe-only, Mixed, Threat-only). **(3) Evaluation.** The adapted model is evaluated along two axes: a *Risk* score, quantifying the residual presence of the targeted hazardous capability (lower is better), and a *Utility* score, quantifying the preservation of general reasoning capabilities (higher is better). This module verifies that data-filtering safety survives a low-rank, 4-bit-quantized regime before trasitioning to federated training.

3.2. Module 2 – Federated Transition

The same three steps are now replicated across federated network. Each client executes:
(1) Preprocessing. The local Risk Estimator classifies and filters data. **(2) Local Fine-Tuning.** Each client adapts the global model locally via QLoRA, computing a weight update $\Delta\Phi_k$. **(3) Communication.** Only the lightweight LoRA adapter deltas $\Delta\Phi_k$ are transmitted to the server. **(4) Aggregation.** The server applies **FedAvg**:

$$\Delta\Phi_{\text{global}} = \sum_{k=1}^{K} \frac{n_k}{N} \Delta\Phi_k \tag{1}$$

where n_k/N is each client's share of total samples. The updated global adapter is then broadcast back to all clients, and steps (2)–(4) repeat for a total T communication rounds. **(5) Evaluation** verifies that the safety pattern survives decentralization. This module fixes two reference baselines: **FedAvg (M0)**, with no interventions, and **Data Selection (M1)**, i.e. Deep Ignorance [7] filtering applied at the edge.

3.3. Module 3 – Safety-Aware Aggregation Test

This is the core contribution of the thesis. Data filtering is removed from the pipeline: each client now emits only a *Risk Signal* (step 1a in Fig. 1),

alongside its adapter update. The server consumes this signal inside a *Safety-Aware Aggregation Mechanism*. The honest execution of the Risk Estimator is an explicit trust assumption. Three different solutions are studied.

Dynamic Weighted Aggregation (M2). A risk score w_k encodes each client's data composition, so the global update is now:

$$\Delta\Phi_{\text{global}} = \sum_{k=1}^{K} w_k \, \Delta\Phi_k \tag{2}$$

This weight is computed directly from the client's local data composition *during the current round*, where $N_{S,k}$ and $N_{T,k}$ denote the number of safe and hazardous samples *processed* by client k, weighted respectively by the fixed coefficients 0.8 and 0.2 to reflect their relative trust.

$$w_k = \frac{0.8\, N_{S,k} + 0.2\, N_{T,k}}{N_{\text{tot},k}} \tag{3}$$

Consequently, each client's update is directly scaled according to its underlying data type distribution: clients with more hazardous data exert proportionally less influence.

Task Vector Aggregation (M3). Each client trains two adapters on data partitioned by the Risk Estimator: a safe adapter $\Phi_{\text{safe},k}$ and a threat adapter $\Phi_{\text{threat},k}$. After global aggregation via FedAvg, the server negates the hazardous direction scaled by α:

$$\Phi_{t+1} = \Phi_{\text{safe\_global}} - \alpha\,(\Phi_{\text{threat\_global}} - \Phi_t) \tag{4}$$

Subtracting Φ_t prevents exponential accumulation across rounds, targeting only the newly acquired hazardous delta $\Delta\Phi_{\text{threat\_global}}$.

Risk-Aware Projection for Federated Learning, RAP-FL (M4). While M2 and M3 apply their interventions uniformly to the entire adapter, via scalar weighting or full vector negation, RAP-FL operates at a finer geometric granularity: it identifies the *specific direction* in parameter space encoding hazardous knowledge, then surgically removes only that component while salvaging any entangled benign knowledge.

Mathematical formulation. Let $\Delta\Phi_k^U$ (utility update) and $\Delta\Phi_k^R$ (risk update) be produced by client k using $P_{\text{threat}}(x)$, the classifier probability score estimating how hazardous a sample is, as a soft weight during optimization. The server aggregates them into global direction tensors weighted by dataset share $a_k = n_k/N$:

$$G^U = \textstyle\sum_k a_k \, \Delta\Phi_k^U, \qquad G^R = \sum_k a_k \, \Delta\Phi_k^R \tag{5}$$

The **orthogonal projection** of any adapter update V onto the hazardous direction G^R is:

$$\text{Proj}_{G^R}(V) = \frac{\langle V,\, G^R\rangle_F}{\|G^R\|_F^2}\, G^R \tag{6}$$

Subtracting this component (scaled by $\lambda, \mu \in [0,1]$) yields cleaned updates $\widetilde{\Delta\Phi}_k^U$ and $\widetilde{\Delta\Phi}_k^R$. Because dual-use data entangles benign and hazardous knowledge in the same samples, the residual R_k of $\widetilde{\Delta\Phi}_k^R$ aligned with the utility direction G^U is *salvaged* rather than discarded. The global model update is:

$$\Phi_{t+1} = \Phi_t + \textstyle\sum_k a_k \, \widetilde{\Delta\Phi}_k^U + \gamma \sum_k a_k \, R_k \tag{7}$$

where $\gamma \in [0,1]$ governs the salvage rate. This per-step geometric decomposition is what prevents catastrophic forgetting: benign parameter sub-spaces are never corrupted by the surgical removal of hazardous directions.

3.4. Module 4 – Evaluation

All five strategies (M0–M4) are compared head-to-head on the safety-utility trade-off (Fig. 1, right panel). The evaluation framework is Multiple-Choice Question Answering (MCQA), probing the model's *latent knowledge* [1] directly from output probabilities for replicability and consistency.

4. Experimental Setup & Results

4.1. Setup

All experiments use `deep-ignorance-pretraining-stage-unfiltered` [7], a 6.9B-parameter decoder-only model trained on 500B unfiltered tokens, an ideal baseline encoding dual-use knowledge prior to any alignment. Adapters target all linear layers with $r = 16$ (33.5M trainable parameters, ≈0.49% of the network), quantized to 4-bit NF4 (QLoRA [2]).

Training data is drawn from **Camel Biology** and **Camel Chemistry**, two dense dual-use scientific corpora, then labeled by a fine-tuned ModernBERT [11] classifier, inherited from Deep Ignorance [7], acting as the Risk Estimator (using a threshold τ=0.8, above which samples are flagged as hazardous). We simulate a $K = 2$ federated network where each client holds exactly 6,016 samples, tested across three fixed-heterogeneity scenarios:

- **S1 (Extreme Asymmetry):** One client holds 100% safe data, the other 100% threat data.
- **S2 (Mixed Homogeneous):** Both clients hold a 50/50 safe/threat split.
- **S3 (Mixed Heterogeneous):** Clients hold 60/40 and 30/70 safe/threat splits.

Evaluation is performed via MCQA on two primary benchmarks:

- **WMDP** [5]: Measures hazardous knowledge retention (*lower is safer*).
- **MMLU-Bio / MMLU-NoBio** [3]: Measures general-domain utility (*higher is better*).

4.2. Results

Figure 2 reports WMDP and MMLU-Bio average accuracy across all five training epochs and three federated scenarios, over two seeds.

Standard Baselines. FedAvg (M0) fails systematically: blind aggregation of malicious adapter updates escalates WMDP from the pretrained baseline of 33.6% to ∼40.3% on average across scenarios. **Data Selection (M1)** performs acceptably in S1 (WMDP ≈ 34.0%), where the compromised client can be fully excluded, but at the cost of *data starvation*: MMLU-Bio falls to 37.8%, well below baseline utility. Critically, in the realistic scenario S3, M1 fails on both axes simultaneously, WMDP remains elevated at 37.7% while MMLU-Bio degrades further to 38.0%, providing no meaningful advantage over undefended FedAvg. This results is unexpected, a possible justification can be the different data distributions create aggregation noise.

Limits of Safety-Aware Strategies. Dynamic Weighted Aggregation (M2) exhibits severe topology dependence. In S1, where safe and hazardous data are cleanly partitioned across clients, it suppresses the threat well (WMDP ≈ 33.6%, MMLU-Bio ≈ 40.5%). However, once threat samples are entangled within local datasets (S2, S3), per-batch risk scores average out and the dampening mechanism collapses: in S3, WMDP escalates to 41.7%, nearly as high as undefended FedAvg, rendering M2 unsafe for any realistic scenario deployment. **Task Vector Aggregation (M3)** represents the opposite failure mode: geometrically powerful but non-surgical. In S1, aggressive negation of the threat vector drives WMDP below baseline (32.8%), yet simultaneously causes severe *catastrophic forgetting*, collapsing MMLU-Bio to just 35.0%. In S3, vector extraction under heterogeneous distributions causes the defense to be less effective: it fails to suppress the threat (WMDP climbs to 38.6%) while it maintains good utility gains compared to unhindered training (MMLU-Bio ≈ 41.3%).

RAP-FL: Scenario-Agnostic Consistency. RAP-FL (M4) is the only method that achieves a *consistent* safety-utility balance across all three scenarios. Its key strength is not an absolute best score on any single metric, but rather a **remarkably stable operating point**: WMDP terminal accuracy varies only between 34.2% and 34.5% across S1, S2, and S3. No other method comes close to this level of scenario-agnosticism. O This stability is a direct consequence of the per-step geometric projection (Eq. 6): at each aggregation round, hazardous directions are surgically removed from the adapter space while the residual benign knowledge is actively salvaged (Eq. 7) keeping MMLU ≈ 3% above the pretrained model, mitigating the catastrophic forgetting that can afflicts M1 and M3 in most scenarios, and without relying on the data-partitioning assumptions that cause M2 to collapse. Table 1 summarizes terminal-epoch performance averaged across both experimental seeds.

5. Conclusions

This work demonstrates that decentralizing LLM training is not incompatible with proactive safety. By relocating the safety enforcement paradigm from the data space into the aggregation space, and by decomposing adapter updates via orthogonal projection in the LoRA weight

Figure 2: WMDP (Threat) and MMLU-Bio (Utility) accuracy across the three federated scenarios (S1–S3) over five training epochs. Each row corresponds to one scenario; left plots show threat suppression, right plots show utility preservation.

Average terminal-epoch performance (S3 – Realistic)

Method	WMDP ↓	MMLU-Bio ↑
Pretrained Baseline	33.6%	36.3%
FedAvg (M0)	39.5%	44.0%
Data Selection (M1)	37.7%	38.0%
Dyn. Weighting (M2)	41.7%	43.9%
Task Vector (M3)	38.6%	41.3%
RAP-FL (M4)	**34.2%**	**39.1%**

Table 1: Comparison of terminal-epoch accuracy in the highly heterogeneous scenario (S3), averaged across two seeds. RAP-FL achieves best threat-utility trade-off.

space, RAP-FL provides a robust, scenario-agnostic blueprint for federated fine-tuning of open-weight LLMs.

The results are unambiguous: standard FedAvg, as expected, and edge-side filtering both fail in realistic scenarios, while aggressive weight-space methods such as Task Vector Negation can compromise general utility through catastrophic forgetting and are highly dependent on the underlying data distribution. RAP-FL uniquely achieves the optimal safety-utility trade-off, empirically proving that the dual-use risk of hazardous knowledge can be suppressed, without discarding the benign knowledge that occupies the same parameter subspace.

The central limitation is the *trust assumption*: our methods guarantees hold only if the

Risk Estimator is a certified, non-forgeable entity. Future work should eliminate this assumption through Trusted Execution Environments (TEEs) or cryptographic proofs, and validate RAP-FL at scale ($K \geq 100$) with secure aggregation protocols, establishing the next pillar of safe, privacy-preserving AI.

References

[1] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[3] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[5] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[6] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. Pmlr, 2017.

[7] Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms. *arXiv preprint arXiv:2508.06601*, 2025.

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[9] Jaspreet Pannu, Doni Bloomfield, Robert MacKnight, Moritz S Hanke, Alex Zhu, Gabe Gomes, Anita Cicero, and Thomas V Inglesby. Dual-use capabilities of concern of biological ai models. *PLoS computational biology*, 21(5):e1012975, 2025.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, November 2019.

[11] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

[12] Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3):2150–2168, 2024.

[13] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373, 2024.