Privacy-Preserving Federated Large Language Models: Techniques and Trade-offs

Runhua Xu[†], Guoan Wan[†] and James Joshi[‡]

[†] School of Computer Science and Engineering, Beihang University, Beijing, China
[‡] School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA

{runhua,gawan}@buaa.edu.cn,jjoshi@pitt.edu

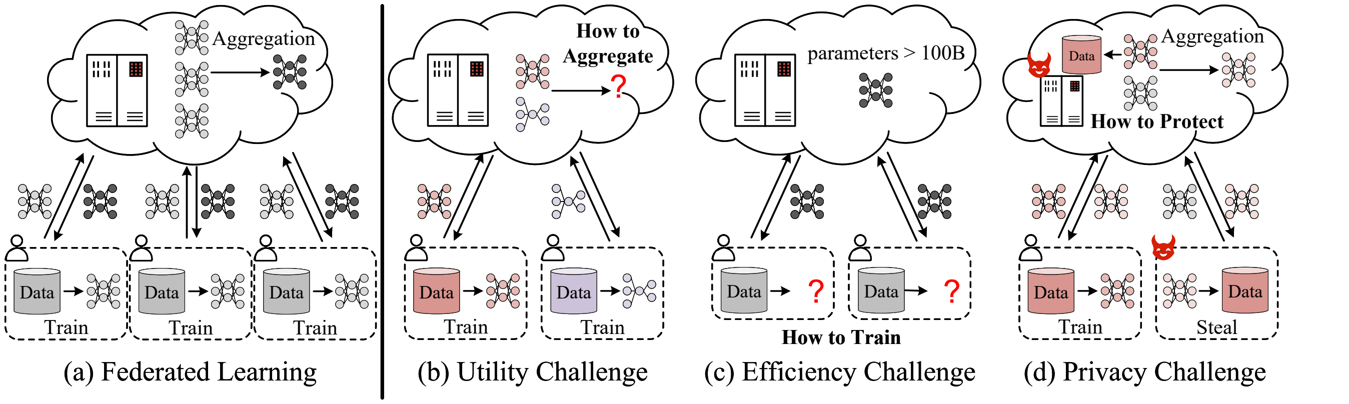
Abstract

The remarkable capabilities of Large Language Models (LLMs) rely on massive, diverse datasets. This dependence creates a fundamental tension in privacy-sensitive domains such as healthcare and finance, where data is siloed and tightly regulated. Federated Learning (FL) offers a privacy-by-design approach that enables collaborative fine-tuning of foundation models on decentralized data. However, combining FL with LLMs—forming Federated LLMs (FedLLMs)—introduces a critical utility–efficiency–privacy trilemma. This study systematically analyzes this trilemma by outlining three core challenges: (1) maintaining model utility amid statistical and system heterogeneity; (2) ensuring efficiency by alleviating severe communication and computation bottlenecks, even with Parameter-Efficient Fine-Tuning (PEFT); and (3) safeguarding privacy against powerful attacks. We formalize these interrelated challenges, examine their trade-offs, review existing defense mechanisms and optimization strategies, and conclude by outlining key open issues and future research directions.

1 Introduction

The remarkable capabilities of large language models (LLMs) have transformed natural language processing, yet their efficacy fundamentally depends on access to massive, diverse training datasets. This dependence creates a critical tension in high-impact domains, such as healthcare, finance, and enterprises, where the most valuable data remain highly privacy-sensitive and organizationally siloed. Simultaneously, emerging and fast-evolving regulatory frameworks, including the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the European Union Artificial Intelligence Act (EU AI Act), impose stringent data protection and privacy requirements, making traditional centralized training increasingly untenable.

In response, federated learning (FL) addresses the challenges faced by centralized training by enabling collaborative training without a need to bring all raw data in one place, allowing multiple parties to jointly train a shared global model while keeping data localized [10]. The central principle of FL is data minimization. Under this principle, only model updates traverse the network, thereby establishing a privacy-by-design architecture better aligned with regulatory data protection and privacy requirements. The convergence of FL with LLMs represents the emerging frontier of privacy-preserving artificial intelligence (AI). Given the prohibitive cost of training billion-parameter models from scratch, we focus on federated fine-tuning of pre-trained foundation models that use distributed private datasets. Large-scale FL deployments, including Google's Gboard with differentially private learning [15] and the SWIFT consortium for cross-border fraud detection [7], indicate operational feasibility at scale. As shown in Figure 1, the current design space of federated LLMs still faces three key challenges: utility, efficiency, and privacy.



(a) Federated Learning (b) Utility Challenge (c) Efficiency Challenge (d) Privacy Challenge

Figure 1: Key Challenges in Privacy-Preserving Federated Large Language Models

Utility Challenge. Federated LLM deployments span cross-device and cross-silo contexts, with pronounced data and model heterogeneity. FedLLM-Bench shows that client sampling strongly affects convergence, with importance-weighted sampling requiring about 2.3× fewer rounds than random sampling [5]. Under a pre-trained global model, utility should be evaluated along two dimensions: aggregation utility, ensuring heterogeneity-aware participation and update fusion, and continual learning utility, which focuses on improving the global model without catastrophic forgetting amid shifting client distributions. Routine data operations such as profiling, schema harmonization, quality control, and maintaining consistent tokenization, vocabulary, and encoding across nodes remain essential for preserving semantic coherence.

Efficiency Challenge. Billion-parameter LLMs induce extreme communication pressure. For example, transmitting full GPT-3 gradients (175B parameters [14]) at FP32 is roughly 700 GB per client per round, rendering naive federated training infeasible. Parameter-Efficient Fine-Tuning (PEFT) provides a practical route: DoRA [2] and VeRA [3] approach full-tuning performance with approximately 100× fewer trainable parameters, and prompt engineering can further reduce on-device computation and communication. Yet these choices introduce new concerns, including guarantees on convergence under non-independent and non-identically distributed (non-IID) client updates and the way PEFT modules interact with privacy amplification and compression.

Privacy Challenge. Although FL keeps raw data local, model updates can leak sensitive information. The DAGER attack [4] demonstrates exact gradient inversion for LLMs, reconstructing training sequences with ROUGE > 0.99 for batches up to 128 tokens. Such near-perfect reconstruction necessitates stronger protection; however, differential privacy may reduce accuracy due to noise injection, whereas cryptographic methods often impose substantial computational overhead. In multi-jurisdictional deployments, designers must also reconcile heterogeneous regulations while maintaining data sovereignty, which further constrains feasible mechanisms and the design of secure aggregation and auditing pipelines.

This comprehensive study provides a systematic analysis of techniques to navigate these challenges. We first delineate three core difficulties in federated LLMs, as reflected in Figure 1(b)–(d): *Aggregation Utility in Heterogeneous Settings*; *Federated Client-Efficient Training of LLMs*, and *Privacy Attacks and Defense Mechanisms*. Next, we formalize the *utility–efficiency–privacy trilemma* that is related to these challenges, specifying threat models, deployment assumptions (cross-device and cross-silo), and evaluation metrics that will guide the subsequent discussion. Finally, we discuss open challenges and future directions.

Table 1: List of abbreviations for partial terms.

Abbreviation	Full Term
CCPA	California Consumer Privacy Act
CKKS	Cheon–Kim–Kim–Song (approximate homomorphic encryption scheme)
DP	Differential Privacy
DP-SGD	Differentially Private Stochastic Gradient Descent
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
LoRA	Low-Rank Adaptation
PEFT	Parameter-Efficient Fine-Tuning
PIPL	Personal Information Protection Law (China)
RAG	Retrieval-Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SGX	(Intel) Software Guard Extensions
SWIFT	Society for Worldwide Interbank Financial Telecommunication

2 Aggregation Utility in Heterogeneous Settings

In federated LLMs, aggregation utility reflects accuracy, stability, and fairness across clients, depending on how heterogeneity is represented and reconciled [5]. In cross-device deployments, strict limits on memory, computation, and bandwidth lead to model heterogeneity[1, 28]. In cross-silo collaborations, differing objectives, label spaces, and domains create data heterogeneity[16, 18].

2.1 Cross-Device Model Heterogeneity and Utility-Oriented Aggregation

Cross-device FL operate under uneven device memory, compute throughput, and network quality. Clients may load different subsets of the model, employ mixed precision, or attach lightweight adaptation modules, which breaks the assumption that every participant optimizes an identical parameterization. Benchmark and toolkit studies in federated LLM fine-tuning show that naive aggregation under such heterogeneity degrades convergence and yields uneven user quality [5]. Open implementations also document mixed local configurations that complicate server-side fusion [1].

The first family of approaches includes those that align parameters before averaging so that the server only fuses structurally compatible updates. When a device trains only a subset of layers, selective or masked aggregation updates those layers while leaving absent components unchanged [6]. When clients expose compact modules such as low-rank adapters, server-side merging that preserves the intended subspace geometry improves stability compared to element-wise averaging [16]. These choices reduce undesirable interference between clients that optimize different slices of the model thereby protecting aggregation utility.

The second family of approaches includes those that are architecture agnostic and rely on distillation to bridge model mismatches. Each client, regardless of backbone or head configuration, produces predictions on a proxy or public corpus. The server distills these signals into a shared student that captures improvements discovered on heterogeneous devices [29]. In cross-device deployments, this reduces payload size, removes strict layer alignment, and remains robust under device churn. Toolkits that support both parameter aggregation and prediction distillation demonstrate that these paths can be combined within one workflow [6].

Personalization improves perceived utility without fragmenting the deployable model. A shared representation is maintained at the server, while devices attach small local heads or adapters that encode preferences and context [27]. Evidence from federated LLM toolchains shows that this separation reduces cross-device update conflicts and improves user-level metrics across wide variation in hardware and availability [1].

In short, structure-aware averaging, architecture-agnostic distillation, and lightweight personalization form a complementary toolkit for cross-device FedLLMs. These methods enhance average accuracy, stabilize convergence, and reduce performance degradation on under-resourced clients by aligning aggregation with each device's training capabilities.

2.2 Cross-Silo Data Heterogeneity and Utility-Oriented Aggregation

Cross-silo FL approaches connect institutions or organizations that pursue different objectives and curate domain-specific datasets. Partnering institutions may use labels following distinct taxonomies, optimize different task mixes, or operate under diverse risk considerations and compliance requirements. If the server aggregates updates without accounting for these differences, the global model can drift toward the majority objective and underperform on minority tasks. Evidence from federated LLM benchmarks shows that utility is sensitive to participation policies and weighting under such heterogeneity, and that naive averaging widens performance gaps across silos [5].

Data-aware aggregation aligns each silo's contribution with relevance to the shared objective. Importance-weighted participation reduces the rounds needed to reach target quality when some participants provide more informative gradients for the current global state [5]. In multilingual settings, clustering institutions by linguistic proximity and aggregating within clusters before global fusion preserves domain signal and lifts perplexity and accuracy, especially for low-resource groups [34]. In multimodal collaborations, separating client-agnostic knowledge from client-specific signals and distilling them into a shared student can help stabiliz training when distributions differ substantially [35].

Personalization further raises perceived utility while maintaining a single deployable backbone. A practical design keeps a shared representation at the server and allows silos to attach small local heads or adapters that encode institutional constraints and preferences [27]. Recent work for federated LLM tooling have shown that this separation reduces inter-silo conflict during aggregation and improves user-level metrics because global updates carry a common structure while local components handle idiosyncrasies [6].

Optimization and sampling policies are key issues when participation in FL training is intermittent and objectives differ. Proximal regularization reduces client drift under objective mismatch [23]. Control variates reduce variance in aggregated updates and enhance stability under uneven connectivity [24]. Adaptive server optimizers help maintain steady progress with non-IID updates [25]. Normalizing local steps improves fairness when silos contribute different amounts of work per round [26]. Priority-based client-selection accelerates convergence to target quality and complements the aggregation strategies above [5].

Overall, cross-silo aggregation that accounts for data heterogeneity, combined with personalization and principled optimization and sampling, expands the utility frontier. The global model benefits from shared structure while each institution retains control over local adaptations that reflect task priorities and regulatory context.

2.3 Utility-Oriented Continual Global Learning under Joint Model and Data Heterogeneity

Federated LLMs begin from a strong pre-trained global model rather than random initialization. This starting point creates a utility risk: if collaborative updates erode base capabilities, usefulness declines for all participants. Measurable catastrophic forgetting appears during continual adaptation and manifests as losses on previously mastered domains after task- or domain-specific tuning [43]. Privacy-compatible rehearsal methods further confirm the phenomenon and propose mitigations [44]. Therefore, continual global learning is essential to preserve and expand utility as new sites, skills, and distributions enter the federation.

Joint heterogeneity arises because devices expose different trainable slices while institutions contribute data with divergent label spaces and objectives. Layered aggregation improves utility when it respects both dimensions. Within groups that share similar adapter layouts, structure-preserving fusion of low-rank updates avoids distortions caused by element-wise averaging [16]. Across groups that differ by language or domain, cluster-wise aggregation retains local signal before global fusion and benefits minority clients [34]. Benchmarks that vary participation and task composition report high sensitivity of utility to these choices [5].

Continual learning mechanisms should be integrated with aggregation strategies to prevent regression of the shared model. Weight consolidation protects parameters important for earlier skills and reduces interference during new updates [30]. Federated transfer decomposes knowledge into global and task-specific components so that the server accumulates stable competence while clients keep sparse local parameters [33]. Tooling that supports selective synchronization and modular adapters enables lifecycle operations such as freezing, merging, or retiring client-side modules without destabilizing the shared backbone [6].

When original data cannot be retained, synthetic rehearsal generated by the model can maintain earlier capabilities with low leakage. Yet, it requires careful filtering and scheduling to avoid drift [44]. The broader literature warns that uncontrolled reliance on model-generated data may degrade generality, which motivates audits and curriculum policies for any replay pipeline [45]. In a federated context, these safeguards should be coupled with per-skill evaluation so that global updates are admitted only when they do not reduce established competencies [5].

Based on the above discussion, we can see that a utility-oriented approach can align aggregation with model slices and data groups, protect previously acquired skills through consolidation or parameter isolation, and maintain them with privacy-compatible rehearsal and module lifecycle management.

3 Federated Client-Efficient Training of LLMs

LLM capability rises predictably with increases in parameters, data, and compute, as captured by empirical scaling laws. Compute-optimal analyses further show that sufficiently trained smaller models can outperform under-trained larger models, although the capability still lies at the scael of billion parameter. In a federated setting, transmitting and updating all weights at these scales is impractical and increases bandwidth and memory costs, which degrades training efficiency [14]. In this section, we therefore focuse on parameter-efficient adaptation as the organizing principle. Clients learn compact adjustments on top of frozen backbones and keep base weights in a compressed form so that the device and network constraints are met while task quality is maintained [12]. In the remainder of this section, we examine federated specializations of these adaptations, introduce randomized or compressed update schemes that further reduce traffic, and outline alternatives that avoid weight updates through retrieval, prompting, and agent coordination [9].

3.1 Federated Specializations of PEFT: Methods and Evidence

A growing body of work tailors PEFT to the specifics of FL optimization, with the goal of reducing communication and memory footprints while keeping round time stable under data and model variation and under intermittent participation. These specializations differ in how they partition knowledge between global and client-specific components and in how they regularize the server-side fusion of client updates. The resulting designs are attractive for production because they align with governance requirements that discourage wholesale model sharing.

FedTune provides a systematic comparison of prompt-, adapter-, and bias-level tuning for pre-trained Transformers in FL settings[22]. It reports fast convergence with substantially reduced communication and, in many cases, performance surpassing local-only baselines, indicating that cross-client collaboration remains effective even when only a tiny subset of parameters is updated.

For multilingual FLs, FedLFC freezes the backbone and trains Low-Rank Adaptation (LoRA) modules per language family, then aggregates at the family level to reflect linguistic proximity [34]. This design improves perplexity and accuracy—especially for low-resource languages—while keeping the adapter budget small and respecting language-specific heterogeneity. Family-wise structuring also reduces negative interference during aggregation and facilitates controlled personalization at deployment time without retraining the backbone.

In multi-modal heterogeneous scenarios, FedDAT introduces a dual-adapter teacher with mutual knowledge distillation; here, a global adapter captures client-agnostic information, while a local adapter encodes client-specific signals [35]. Separating these roles stabilizes aggregation and improves results on vision–language benchmarks compared with centralized PEFT-to-FL pipelines, suggesting that adapter topology is as important as adapter size. The formulation naturally extends to other cross-modality or cross-domain FL setting where shared and private factors must be disentangled.

Beyond module choice, aggregating PEFT parameters requires care to avoid structural drift. FLoRA addresses inconsistencies in naive averaging of LoRA factors by preserving their intended low-rank structure during fusion, yielding consistent quality gains across client–server compositions [16]. At the system layer, SLORA exploits structured sharing across layers to further reduce trainable budgets without sacrificing accuracy, which is useful when client memory ceilings bind participation. Together, these results indicate that PEFT is not only communication-efficient but also amenable to FL-aware rules that directly prioritize global utility under heterogeneous clients.

3.2 Randomization and Compression for Communication Efficiency

Randomized and compressed updates offer an orthogonal means to reduce communication overhead, and they can be layered over adapter- or LoRA-based clients without changing learning objectives. The central idea is to approximate dense updates with sketches or quantized representations that preserve informative directions for aggregation, while variance control and bias correction restore convergence guarantees. This family of methods is particularly an attractive option when bandwidth fluctuates or when client links are asymmetric across rounds.

Ferret, proposed in [36], performs a first-order full-parameter tuning with low-dimensional projections and shared randomness to reconstruct updates at the server. By decoupling local optimization from transmitted dimensionality, Ferret combines the benefits of full parameter training with communication comparable to compressed methods, and it exhibits favorable convergence relative to zeroth-order alternatives. The projection–reconstruction pipeline also interacts well with straggler mitigation because it decimates payloads without altering local objectives, thereby keeping device-side software simple.

Classical quantization and error-feedback mechanisms further lower bandwidth and correct compression bias. QSGD quantizes gradients under variance control [37]. EF SGD feeds back the compression error to recover the trajectory of the uncompressed method [38]. In heterogeneous deployments, combining PEFT with randomized projections yields additive gains. Clients keep tiny trainable modules and transmit sketched deltas when links are constrained, while the server aggregates in a manner consistent with reliability and fairness policies.

3.3 Alternatives to Weight Updates: RAG, Prompts, and Agents

Some alternative approaches to improving FL efficiency is to avoid gradient updates altogether. These approaches leverage local data and global coordination while keeping base model weights fixed [9, 12], which aligns with privacy and governance constraints in both cross-device and cross-silo settings [11, 19]. They are particularly attractive when regulatory or operational constraints limit the sharing of model parameters, even in modified form.

Retrieval-Augmented Generation (RAG) maintains private corpora locally and coordinates retrieval over distributed indices, after which a frozen LLM conditions on retrieved evidence for generation [39]. Synchronization shifts from heavyweight deltas to lightweight retrieval statistics and index metadata, reducing update traffic and simplifying audits because content remains under local control. In practice, FL can align retrieval strategies with local curation standards, seamlessly integrating with evaluation protocols that prioritize accuracy and source attribution.

Prompt-based collaboration includes exchanging soft prompts or prefixes rather than gradients, yielding huge communication savings and natural personalization under tenant or domain structure [40, 41]. Prompts can be shared, clustered, or composed to reflect organizational boundaries without modifying backbone weights, and they can be rotated or gated to manage risk. Because prompts are small, they are amenable to secure aggregation and differential privacy, which further broadens the set of compliant deployment regimes.

Agent orchestration frameworks coordinate tool use, retrieval, and prompting policies across sites, leveraging local data without weight updates and facilitating policy-compliant workflows [42]. By turning adaptation into planning and tool selection rather than gradient descent, agents avoid heavy synchronization while still exploiting situational context and institutional knowledge. In hybrid pipelines, agents can call RAG for evidence, choose prompts for control, and fall back to PEFT only when sustained drift necessitates weight changes.

Overall, these alternatives coexist with parameter-efficient adaptation and randomized compression to enlarge the feasible region of FedLLM design. By reducing or eliminating gradient traffic, they broaden participation, lower costs, and sustain utility under heterogeneous constraints, while leaving room for targeted weight updates when enduring domain shifts demand persistent changes.

4 Privacy Attacks and Defense Mechanisms

Understanding adversarial capabilities against FedLLM systems is crucial for designing effective defenses and ensuring realistic privacy guarantees. Recent studies have revealed advanced attack vectors that can extract sensitive information from model updates in FL process, highlighting the need for thorough analysis of attack methods, success conditions, and layered defense strategies.

4.1 Gradient Inversion Attacks: From Theory to Practice

Gradient inversion attacks pose one of the most serious threat to privacy in federated learning, as they can reconstruct the entire training dataset from observed gradients. These attacks rely on the fact that gradients encode rich information about the data that produced them. For a neural network with parameters θ and loss function ℓ, the gradient $g = \nabla_\theta \ell(\theta; x)$ is a deterministic function of both the parameters and the input data x. Under sufficient constraints, this relationship can be inverted to recover x from g.

Traditional gradient inversion methods frame reconstruction as an optimization problem: find an input $\hat{x}$ that minimizes $\|\nabla_\theta \ell(\theta; \hat{x}) - g\|^2$. Early works have demonstrated successful reconstructions for simple networks and small batches but have struggled with the high-dimensional, discrete nature of language data. The non-convex optimization landscape and the discrete token space of language models pose fundamental challenges, limiting reconstruction accuracy to semantic similarity rather than exact recovery.

The DAGER (Differentially Private Aggregated Gradient Extraction with Restoration) method [4] marks a major breakthrough in gradient inversion. DAGER achieves exact reconstruction of training sequences—recovering text with ROUGE-1 and ROUGE-2 scores above 0.99—for batch sizes up to 128 tokens. The implications for federated learning are severe. An honest-but-curious server observing gradient updates can reconstruct entire training texts, including medical records, financial documents, or private communications. The attack remains effective even when gradients are aggregated over multiple training steps, provided the batch size stays within feasible limits. Moreover, DAGER shows resilience to common defenses: gradient clipping only slightly reduces reconstruction quality, while compression techniques such as top-k sparsification offer limited protection unless applied aggressively [4].

4.2 Membership and Property Inference in FL Settings

While gradient inversion attacks aim to reconstruct specific training examples, membership and property inference attacks target different forms of privacy leakage that can be equally harmful in practice.

Membership inference attacks determine whether particular data points were used during training without necessarily reconstructing their content. In FL settings, these attacks exploit the distributed nature of learning to achieve higher success rates than those on centralized models. FedMIA [8] leverages the "all for one" principle inherent in federated learning: each client's update encodes information about all its local training samples simultaneously. By analyzing how model updates affect predictions on candidate data points across multiple rounds, FedMIA achieves 63–68% attack success rates compared with 45–52% for similar centralized attacks.

The attack methodology integrates multiple signals to improve inference accuracy. Temporal analysis tracks how prediction confidence on target examples evolves, with members typically showing steadily increasing confidence. Update correlation analysis measures alignment between model updates and gradients computed on target examples, with stronger correlations indicating membership. Influence estimation assesses how removing hypothetical examples would alter model updates, based on the insight that true members measurably influence parameter changes. Combined through ensemble methods, these signals enable robust membership detection even when individual indicators are noisy.

Property inference attacks extract statistical characteristics of training datasets rather than details about specific samples. In FL settings, such attacks can expose sensitive information about participating institutions/clients. A hospital's model updates might reveal unusual disease prevalence patterns, demographic distributions, or treatment protocols that constitute valuable competitive intelligence; financial institutions' updates could disclose customer segment traits, risk profiles, or business strategies embedded in their data.

FL setting paradoxically makes property inference both easier and harder than centralized training does. The isolation of client updates provides clearer signals about individual dataset properties since they are not diluted by mixing with others' data. However, limited visibility—observing only periodic model updates instead of continuous dynamics—reduces available information for adversaries. Recent studies show that sophisticated attackers can overcome this constraint by correlating observations across rounds and exploiting the temporal consistency of dataset properties[8, 15].

4.3 Multi-Layered Defense Strategies

Defending against this range of attacks requires comprehensive strategies that integrate multiple protection mechanisms, each addressing distinct threat vectors while collectively providing defense-in-depth.

Differential privacy offers the strongest theoretical safeguard against inference attacks by ensuring that model updates reveal only limited information about the training data. DP-SGD's noise injection fundamentally restricts what adversaries can learn, with formal guarantees that hold regardless of their capabilities or auxiliary knowledge. Empirical results show that applying differential privacy with $\epsilon = 1.0$ reduces DAGER's reconstruction quality from ROUGE > 0.99 to ROUGE < 0.3, effectively preventing meaningful text recovery. Similarly, differential privacy provides provable bounds on membership inference advantage, limiting adversarial success to near-random levels under reasonable privacy parameters.

However, differential privacy alone is insufficient. It does not protect against Byzantine attacks where malicious clients craft updates exploiting the noise distribution. Moreover, strong differential privacy often incurs unacceptable utility loss for some applications, making complementary defenses necessary to achieve practical protection with lower performance costs.

Secure aggregation protocols defend against server-side threats by ensuring servers see only aggregated updates rather than individual contributions. Cryptographic secure aggregation [11] uses secret sharing or homomorphic encryption to compute aggregates without exposing individual model updates. This approach strongly protects against honest-but-curious servers attempting gradient inversion or membership inference on single clients. However, its benefits diminish when few clients are aggregated since limited aggregation yields minimal privacy amplification. Additionally, secure aggregation cannot prevent malicious clients from analyzing the global model to infer information about others.

The most effective defense strategy combines multiple mechanisms targeting different threats. A robust configuration might use LoRA for parameter efficiency (reducing attack surface), DP-SGD with a moderate privacy budget for inference resistance, secure aggregation to hide individual updates from the server, robust aggregation to filter Byzantine inputs, and gradient compression to limit information leakage. While no single method provides complete protection against all attacks, this layered approach substantially increases the attack difficulty while preserving practical utility and efficiency.

5 The Privacy-Utility-Efficiency Trilemma

The techniques surveyed in previous sections do not operate independently but rather interact within a complex optimization landscape characterized by fundamental trade-offs that constrain achievable system configurations. Understanding these trade-offs through the lens of a privacy-utility-efficiency trilemma provides essential guidance for practical privacy-preserving FedLLM system design and helps explain why no single solution dominates across all deployment scenarios.

5.1 Mathematical Formalization and Constraint Analysis

The privacy–utility–efficiency trilemma can be formally defined by three interdependent metrics that together determine the feasible operating region of PP-FedLLM systems. The privacy level $\mathcal{P}$ measures protection against information leakage, quantified by the differential privacy parameter ϵ (where smaller values indicate stronger privacy) or by cryptographic security parameters in encryption-based methods. Model utility $\mathcal{U}$ reflects the performance of the trained model, typically evaluated using task-specific metrics such as accuracy, F1 score, or perplexity on held-out test sets. System efficiency $\mathcal{E}$ represents computational and communication costs, measured by training time, number of communication rounds, bandwidth usage, or total computational operations.

The trilemma manifests as fundamental constraints on the achievable region in $(\mathcal{P},\mathcal{U},\mathcal{E})$ space. These constraints arise from information-theoretic limits, computational complexity barriers, and statistical requirements that cannot be overcome through engineering alone.

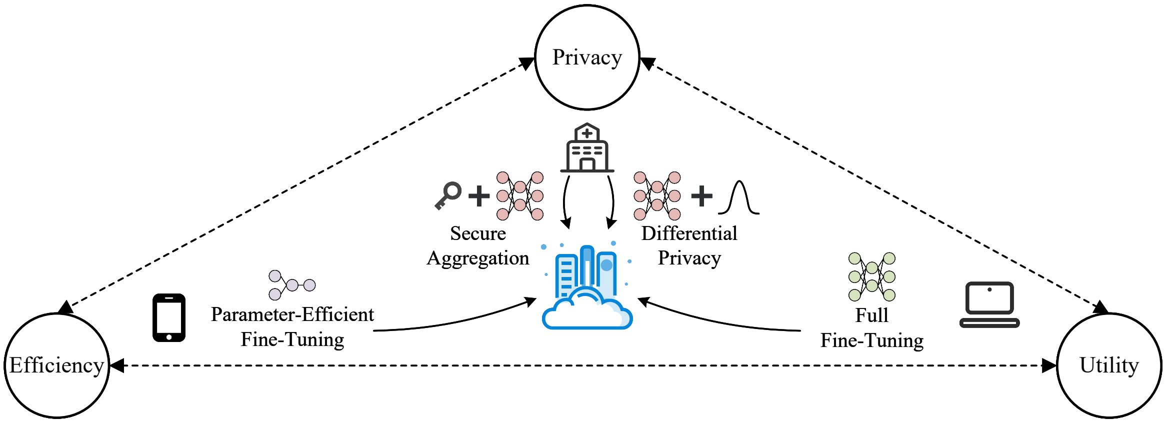


Figure 2: The illustration of privacy-utility-efficiency trilemma in privacy-preserving federated LLM.

The privacy-utility trade-off reflects an information-theoretic reality: protecting privacy requires limiting information flow about training data, but model learning fundamentally requires extracting information from that same data. Differential privacy makes this trade-off explicit through noise injection that provides privacy at the cost of accuracy. The privacy–efficiency trade-off arises from the computational complexity of privacy-preserving mechanisms. For example, homomorphic encryption perfectly preserves utility—the encrypted computation yields the same results as plaintext computation—but incurs a 50–300% computational overhead and a 70–200% increase in communication due to larger ciphertexts [46]. Cryptographic theory establishes fundamental lower bounds indicating that these overheads cannot be fully eliminated, only reduced through improved implementations [47].

The utility–efficiency trade-off is most evident in the choice between full fine-tuning and parameter-efficient methods. Full fine-tuning offers optimal task adaptation but requires updating and transmitting all model parameters, whereas PEFT methods such as LoRA achieve 95–98% of full fine-tuning performance while using 100–1000 times fewer parameters.

5.2 Empirical Characterization Through Systematic Evaluation

Recent systematic evaluations provide quantitative characterization of these trade-offs in practical systems. The FedLLM-Bench evaluation [5] tests configurations across the trilemma space, revealing consistent patterns that guides system design. Their experiments with LoRA (rank 8) on a 7B parameter model demonstrate that this configuration achieves 96.3% of full fine-tuning performance while reducing communication cost by 267 times and computation cost by 4.7 times. This highly favorable utility-efficiency trade-off that is primarily the reason for LoRA's widespread adoption. Adding differential privacy with $\epsilon = 4.0$ to this LoRA-based system reduces accuracy by an additional 4.2% while providing formal privacy guarantees, illustrating how privacy protection compounds with efficiency optimizations to impact utility. Applying selective homomorphic encryption to the most sensitive 10% of LoRA parameters adds 38% communication overhead and 42% computation time while maintaining perfect utility for those protected parameters.

These empirical observations show that the trade-offs are neither linear nor independent. Privacy mechanisms interact with efficiency optimizations in complex ways: the utility loss from differential privacy is amplified when paired with aggressive parameter reduction, as the smaller parameter space offers less redundancy to absorb noise. Conversely, cryptographic methods that preserve utility become more practical when combined with PEFT, since the reduced parameter count keeps encryption overhead manageable. Understanding these interactions is essential for effectively navigating the trilemma.

5.3 Hybrid Solutions: Synergistic Combinations for Practical Deployment

The most successful practical deployments employ hybrid solutions that combine multiple techniques synergistically to achieve favorable positions within the trilemma space. These combinations leverage the strengths of different approaches while mitigating their individual weaknesses through careful system design.

PEFT serves as the foundational enabler for most hybrid solutions, dramatically improving the efficiency baseline creating headroom for adding privacy protections. The combination of LoRA with differential privacy exemplifies this synergy. LoRA's parameter reduction from 7 billion to 21 million parameters (for a 7B model with rank 8) not only reduces communication cost by 333 times but also fundamentally changes how employing differential privacy affects the model. The DP noise added to protect privacy has smaller impact on a lower-dimensional parameter space, as the relative signal-to-noise ratio improves. Empirical studies show that LoRA with $\epsilon = 1.0$ differential privacy achieves accuracy within 5% of non-private full fine-tuning, whereas applying the same privacy budget to full fine-tuning degrades accuracy by 12-15%. This dramatic difference demonstrates how efficiency optimizations can indirectly improve the privacy-utility trade-off.

5.4 Application-Driven Trade-off Navigation

Optimal navigation of the trilemma depends heavily on application-specific requirements, threat models, and deployment constraints. Different domains emphasize different aspects of the trilemma, resulting in distinct solution strategies.

Healthcare applications operating under strict regulations such as HIPAA typically prioritize privacy above all else. A multi-hospital consortium analyzing electronic health records may accept significant utility loss to ensure strong privacy guarantees. Efficiency takes a back seat to meeting privacy requirements, though PEFT remains crucial for practical deployment.

Financial fraud detection systems face a different set of constraints, in which both false negatives (missed fraud) and false positives (legitimate transactions flagged as fraudulent) incur substantial costs. These systems prioritize utility while maintaining sufficient privacy protection against realistic threats. The Swift consortium's use of cryptographic safeguards for highly sensitive features combined with moderate differential privacy for others reflects this balance [7].

Enterprise knowledge management systems encounter yet another trade-off landscape. When deploying federated RAG across corporate departments, the main concern often lies in preserving departmental autonomy and intellectual property rather than individual privacy. Such systems may rely on trusted execution environments (TEE) that offer strong practical protection with minimal overhead, accepting the trust assumptions inherent in hardware-based security.

6 Open Challenges and Future Directions

Despite remarkable progress in making FedLLMs practical, fundamental challenges remain that will determine whether federated learning achieves its potential as a transformative paradigm for privacy-preserving AI or remains limited to specialized applications.

6.1 Foundational Challenges Across Three Dimensions

Deploying FedLLMs involves three interrelated challenges that define the technical landscape of this survey: maintaining utility amid statistical and system heterogeneity, ensuring efficiency under tight communication and computation constraints, and preserving privacy against advanced inference and reconstruction attacks.

6.1.1 Utility: Statistical and System Heterogeneity

Real-world federated networks exhibit significant statistical heterogeneity that violates the Independent and Identically Distributed (IID) assumption of classical learning [18, 21]. In healthcare, institutions serving different populations adopt distinct treatment patterns and documentation practices. In finance, regional regulations and local payment behaviors lead to divergent transaction distributions. This non-IID structure drives client gradients in conflicting directions, slowing convergence and causing uneven model quality across cohorts. FedLLM Bench reports that vanilla FedAvg may require two to three times more rounds to reach a target accuracy under realistic heterogeneity than under IID conditions, and that performance disparities across client groups persist even after convergence [5]. More research is needed to develop heterogeneity-aware aggregation and fairness-aware objectives for FedLLMs, and to standardize non-IID benchmarks with group-level reporting and convergence criteria.

This necessitates more targeted research into advanced heterogeneity-aware aggregation techniques to combat gradient divergence and stabilize convergence. Future work should also design robust fairness-oriented objectives—such as optimizing worst-case client performance or deploying personalized models—to address disparities in model quality. Moreover, there is an urgent need to standardize non-IID benchmarks that realistically capture data skew and to establish group-level metrics for properly evaluating algorithmic fairness and robustness.

System heterogeneity amplifies these effects [28]. Clients vary in memory, compute power, network bandwidth, and availability. Cross-device deployments must accommodate everything from high-end smartphones with neural processing units to entry-level devices with limited memory. Network quality ranges from gigabit fiber to intermittent 3G connections, leading to significant differences in latency and reliability. Synchronous aggregation that waits for all selected clients can be delayed by stragglers, extending round time. Asynchronous aggregation reduces waiting but introduces staleness when slower clients train on outdated parameters. Both approaches require designs that stabilize optimization under delayed or partially aligned updates [20]. More research is needed on staleness-robust asynchronous methods with convergence guarantees under bounded delay, and on adaptive client selection/model sizing that respects device and network constraints.

Future research should develop communication-efficient, staleness-robust asynchronous methods that offer formal convergence guarantees under bounded update delays. This includes exploring adaptive client selection strategies that prioritize available clients without introducing bias and dynamically adjusting model components or quantization to accommodate heterogeneous device constraints. Such system-aware techniques are crucial for stabilizing optimization and mitigating the effects of stragglers and partial client participation.

6.1.2 Efficiency: Communication and Computational Bottlenecks

At modern LLM scales, communication and memory dominate the costs in federated settings. Transmitting full updates for GPT-3 with 175 billion parameters at FP32 requires about 700 GB per client per round, and even FP16 still needs around 350 GB—far beyond typical network capacities [14]. Computation is similarly limiting: fine-tuning LLaMA2 70B demands roughly 512 GB of GPU memory for standard training with the Adam optimizer, accounting for weights, gradients, optimizer states, and activations [13]. Most edge participants cannot host such models, motivating approaches that reduce bytes per round and device memory while maintaining stable training.

Extending from adaptation to foundation model pretraining further amplifies these constraints. Photon demonstrates feasibility for 7B-parameter models across sixteen high-performance computing centers, but scaling to 100B and beyond requires orders of magnitude more tokens and wall-clock time. Autoregressive dependencies restrict parallelism, while heterogeneous pretraining corpora increase optimization drift among participants. Promising directions include hierarchical federation with intra-cluster optimization, bounded staleness asynchrony to handle stragglers, curriculum-style token scheduling by domain or quality, and phase-aware privacy budgets during early training. These remain open challenges that need to be addressed to ensure that federated pretraining at a foundation model is practical for deployment at scale.

6.1.3 Privacy: Threats and Adaptive Adversaries

Keeping data local does not prevent information leakage from model updates. Membership inference attacks, which determine whether a record was used in training, become stronger because each client update reflects many local examples. FedMIA reports success rates of 63–68% under realistic conditions [8]. Gradient inversion attacks pose an even greater threat. DAGER reconstructs LLM training sequences with ROUGE scores above 0.99 for batches of up to 128 tokens by searching the discrete token space [4]. These findings show that data localization alone cannot ensure privacy.

Noise-based defenses such as Differential Privacy (DP) are not a cure-all and can introduce new attack surfaces. Adversaries may shift from direct gradient inspection to exploiting side-channel vulnerabilities—like timing variations or resource usage patterns—or use active probing to maximize information leakage within a fixed privacy budget. Advanced threat models also include colluding clients and Sybil attacks, which can evade basic anomaly detection and amplify targeted or poisoning attacks. Effective mitigation therefore requires a defense-in-depth strategy that combines training-time privacy accounting with strong system-level hardening. This holistic approach should establish a hardware-rooted chain of trust to ensure platform integrity through remote attestation. Key defensive measures should be applied throughout the pipeline: traffic shaping and batched reporting to mask timing side channels; robust aggregation protocols with gradient clipping and cross-client correlation checks; and identity management via public key infrastructure (PKI) to enforce rate limits and defend against Sybils. These controls should be unified under a secure aggregation framework that enables auditable, per-round privacy accounting.

This highlights several critical research needs: (i) developing end-to-end privacy guarantees that cover the full pipeline—from data collection and training to aggregation and inference logging—is essential to meet regulatory compliance with auditable accounting; (ii) integrated objectives that jointly optimize privacy with security and fairness are required, particularly when facing heterogeneous clients and adversaries (iii) new machine unlearning protocols suitable for federated and DP-constrained settings are necessary to address the "right-to-be-forgotten" and facilitate data-correction requests without costly full retraining [48].

6.2 Data Engineering Challenges in Federated LLM

While most existing research on federated LLMs focuses on training algorithms and privacy mechanisms, the data engineering pipeline poses equally critical yet underexplored challenges [17, 18]. Unlike centralized machine learning, where data scientists can directly inspect, clean, and preprocess data, federated settings impose fundamental constraints that complicate traditional data engineering practices.

6.2.1 Data Quality Assessment Without Centralization

In traditional centralized ML pipelines, data quality assessment is performed through exploratory data analysis, profiling, and visualization of the entire dataset. However, in privacy-preserving FL, the orchestrating server cannot directly access raw data, making it impossible to assess data quality through conventional means. This creates several critical challenges:

Distributed Data Profiling. Without centralized access, evaluating data quality metrics—such as missing value rates, class distributions, feature correlations, and outliers—requires privacy-preserving distributed algorithms [17]. Recent research has proposed differentially private data profiling methods, but the added noise can obscure real data quality issues. For example, a hospital with systematically miscoded diagnoses might go unnoticed if privacy noise hides the anomalous patterns.

Heterogeneous Data Schemas. Real-world federated deployments often involve participants with heterogeneous data schemas, especially in healthcare, where institutions use different Electronic Health Record (EHR) systems. Even when training on the same task (e.g., clinical note classification), hospitals may differ in field names, coding systems (ICD-9 vs. ICD-10), and data granularity. Automated schema matching and harmonization in federated settings remain open challenges, as current solutions either demand extensive manual alignment or cause information loss.

Data Drift Detection. In production federated systems, data distributions at client nodes can shift over time due to changing user behavior, seasonal effects, or systematic changes in data collection practices. Detecting such concept drift without centralizing data requires new distributed monitoring methods. For LLMs in particular, vocabulary drift—such as the emergence of new terms and evolving language usage—introduces additional challenges that existing federated learning frameworks do not adequately address.

6.2.2 Data Governance and Compliance

The decentralized nature of federated learning introduces complex data governance challenges that go beyond technical privacy mechanisms [19]. Each participating organization must maintain sovereignty over its data while contributing to a collaborative model, requiring new governance frameworks that balance autonomy with coordination.

Multi-Jurisdictional Compliance. Federated LLM deployments across multiple countries must navigate a complex landscape of data protection laws, including the GDPR (Europe), CCPA (California), PIPL (China), and HIPAA (healthcare). Each jurisdiction imposes distinct requirements for data localization, consent management, and breach notification. For instance, the GDPR's right to erasure ("right to be forgotten") poses technical challenges in federated settings: if a user requests deletion of their data, how can we ensure their contribution is removed from a model trained across hundreds of devices? Current federated unlearning methods remain immature and computationally costly.

Data Lineage and Provenance Tracking. In regulated industries, maintaining detailed records of data lineage—tracking how data moves through ML pipelines and influences model predictions—is often a compliance requirement. In FL settings, this task becomes far more complex: the global model results from aggregated updates across multiple sources, each with its own preprocessing pipeline and data quality controls. Blockchain-based methods have been proposed to establish immutable audit trails for FL training, but they introduce additional computational overhead and privacy risks, as even encrypted metadata can reveal participation patterns.

Dynamic Participant Management. Deployable FL systems must manage participants joining, leaving, or being excluded for poor data quality or malicious behavior. This requires governance mechanisms to (1) assess new participants' data quality and security practices, (2) fairly attribute credit for model improvements among contributors, and (3) manage intellectual property rights in the jointly trained model. Multidisciplinary approaches to address these challenges. Existing frameworks offer limited support for these governance processes, especially when determining fair compensation in cases where participants contribute varying amounts or qualities of data.

6.2.3 Real-Time Federated Learning and Streaming Data

Most existing federated LLM studies assume static datasets and batch training. However, many real-world applications—such as mobile keyboard prediction, content recommendation, and real-time fraud detection—demand continuous learning from streaming data [20, 21]. This creates several data engineering challenges:

Online Data Preprocessing. Traditional ML pipelines perform extensive data preprocessing (tokenization, normalization, feature engineering) as a separate batch step before training. In FL settings that need to use streaming data, preprocessing must occur online at each client [6], requiring careful coordination to maintain consistency. For LLMs, this involves keeping tokenizer vocabularies synchronized as new terms appear, handling out-of-vocabulary words, and determining when to update preprocessing pipelines without disrupting existing models. Further research is needed to develop lightweight on-device tokenization algorithms and decentralized vocabulary synchronization protocols that are communication-efficient and resilient to network dropouts.

Temporal Data Alignment. In cross-device FL (e.g., training across millions of mobile phones), devices may go offline for long periods, causing temporal misalignment where some clients train on outdated data while others use fresh data [20]. For time-sensitive tasks such as news classification or trend detection, this temporal skew can severely degrade model performance. Designing aggregation algorithms that appropriately weight contributions by data freshness remains an open challenge [21]. This requires research into novel staleness-aware aggregation functions that explicitly model temporal dependencies and can dynamically discount or re-weight client updates based on their data timestamps.

Incremental Model Updates. Streaming data demands incremental model updates instead of full retraining. For LLMs, this is especially difficult due to catastrophic forgetting—the tendency of neural networks to lose previously learned knowledge when exposed to new data [43]. Federated continual learning must balance plasticity (learning new information) and stability (preserving prior knowledge) across distributed nodes while maintaining privacy guarantees. A key research and development challenge is to develop federated continual learning (FCL) strategies—such as parameter isolation or rehearsal-based methods—that prevent catastrophic forgetting while preserving privacy and minimizing communication overhead.

6.3 Future Directions

Federated Data Marketplaces. One promising research direction is to design innovative economic mechanisms and technical infrastructure for federated data marketplaces to enable future data economy where participants can discover collaboration opportunities, negotiate data-sharing terms, and receive fair compensation for their contributions. This involves addressing technical challenges (e.g., privacy-preserving dataset search, contribution valuation) and creating governance structures that incentivize high-quality and trustworthy participation.

Adaptive Preprocessing Pipelines. Complementary to this, developing adaptive preprocessing techniques that automatically adjust to heterogeneous data distributions and evolving characteristics in FL settings is crucial. This includes automated feature engineering, dynamic tokenizer updates for LLMs, and context-aware normalization strategies that respect local data properties while maintaining global consistency.

Privacy-Preserving Data Quality Tools. To support both the economic models of data marketplaces and the technical demands of adaptive pipelines, there is a significant need to develop privacy-preserving versions of standard data engineering tools (profilers, validators, schema matchers) with formal privacy guarantees. This research must extend beyond purely technical solutions—such as differentially private exploratory data analysis or secure multi-party computation for joint schema inference—to also address

critical human factors. This includes designing usable interfaces and auditable governance workflows that build trust and incentivize high-quality, trustworthy participation from data providers.
Standardization and Interoperability. Significant effort is also needed to establish standardized interfaces and protocols for federated data engineering, similar to how the OMOP Common Data Model standardizes clinical data. This includes standardized APIs for data quality reporting, common metadata schemas for describing federated datasets, and interoperable governance frameworks that span multiple organizations and jurisdictions.

Addressing these challenges represents critical research opportunities that could greatly accelerate the real-world deployment of privacy-preserving federated LLMs.

7 Conclusion

The convergence of LLMs and FL provides a crucial pathway to unlock the value of decentralized data in privacy-sensitive domains such as healthcare and finance. This paper presents a systematic analysis of the fundamental utility–efficiency–privacy trilemma inherent in designing Federated LLM (FedLLM) systems. We outline key challenges, including managing statistical heterogeneity, mitigating communication bottlenecks, and defending against advanced privacy attacks such as gradient inversion. By analyzing these trade-offs, reviewing existing techniques, and exploring practical applications, we highlight that no single solution suffices. Instead, effective and resilient FedLLM deployment requires a comprehensive and integrated approach. Addressing open research directions—such as developing heterogeneity-aware aggregation methods, robust privacy accounting frameworks, and standardized data engineering practices—is essential for the continued trustworthy and effective advancement of this transformative technology.

References

[1] Ye, R., Wang, W., Chai, J., Li, D., Li, Z., Xu, Y., Du, Y., Wang, Y. & Chen, S. OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning. *Proceedings Of The 30th ACM SIGKDD Conference On Knowledge Discovery And Data Mining.* pp. 6137-6147 (2024)

[2] Liu, S., Wang, C., Yin, H., Molchanov, P., Wang, Y., Cheng, K. & Chen, M. DoRA: Weight-Decomposed Low-Rank Adaptation. *Proceedings Of The 41st International Conference On Machine Learning.* pp. 32100-32121 (2024)

[3] Kopiczko, D., Blankevoort, T. & Asano, Y. VeRA: Vector-based Random Matrix Adaptation. *The Twelfth International Conference On Learning Representations.* pp. 1-14 (2024)

[4] Petrov, I., Dimitrov, D., Baader, M., Müller, M. & Vechev, M. DAGER: Exact Gradient Inversion for Large Language Models. *Advances In Neural Information Processing Systems.* (2024)

[5] Ye, R., Ge, R., Zhu, X., Chai, J., Du, Y., Liu, Y., Wang, Y. & Chen, S. FedLLM-Bench: Realistic Benchmarks for Federated Learning of Large Language Models. *Advances In Neural Information Processing Systems.* (2024)

[6] Kuang, W., Qian, B., Li, Z., Chen, D., Gao, D., Pan, X., Xie, Y., Li, Y., Ding, B. & Zhou, J. FederatedScope-LLM: A Comprehensive Package for Fine-tuning Large Language Models in Federated Learning. *Proceedings Of The 30th ACM SIGKDD Conference On Knowledge Discovery And Data Mining.* pp. 5260-5271 (2024)

[7] Dave, V. & Santhanagopalan, A. Google Cloud and Swift pioneer advanced AI and federated learning tech to help combat payments fraud. (Google Cloud Blog,2024,12), Accessed October 12, 2025

[8] Zhu, G., Li, D., Gu, H., Yao, Y., Fan, L. & Han, Y. FedMIA: An Effective Membership Inference Attack Exploiting "All for One" Principle in Federated Learning. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition.* pp. 20643-20653 (2025)

[9] Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances In Neural Information Processing Systems.* (2023)

[10] McMahan, H., Moore, E., Ramage, D., Hampson, S. & Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings Of The 20th International Conference On Artificial Intelligence And Statistics.* pp. 1273-1282 (2017)

[11] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H., Patel, S., Ramage, D., Segal, A. & Seth, K. Practical Secure Aggregation for Privacy-Preserving Machine Learning. *Proceedings Of The 2017 ACM SIGSAC Conference On Computer And Communications Security.* pp. 1175-1191 (2017)

[12] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *The Tenth International Conference On Learning Representations.* pp. 1-14 (2022)

[13] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S. & Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv Preprint ArXiv:2307.09288.* (2023)

[14] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. Language Models are Few-Shot Learners. *Advances In Neural Information Processing Systems.* pp. 1877-1901 (2020)

[15] Xu, Z., Zhang, Y., Andrew, G., Choquette-Choo, C., Kairouz, P., McMahan, H., Rosenstock, J. & Zhang, Y. Federated Learning of Gboard Language Models with Differential Privacy. *Proceedings Of The 61st Annual Meeting Of The Association For Computational Linguistics (Volume 5: Industry Track).* pp. 629-639 (2023)

[16] Wang, Z., Shen, Z., He, Y., Sun, G., Wang, H., Lyu, L. & Li, A. FLoRA: Federated Fine-Tuning Large Language Models with Heterogeneous Low-Rank Adaptations. *Advances In Neural Information Processing Systems.* (2024)

[17] Gonzalez Zelaya, C. Towards Explaining the Effects of Data Preprocessing on Machine Learning. *Proceedings Of The 2019 IEEE 35th International Conference On Data Engineering.* pp. 2086-2090 (2019)

[18] Li, Q., Diao, Y., Chen, Q. & He, B. Federated Learning on Non-IID Data Silos: An Experimental Study. *Proceedings Of The 2022 IEEE 38th International Conference On Data Engineering.* pp. 965-978 (2022)

[19] Habu, J., Dhabariya, A., Pal, B. & Abubakar, F. Decentralized Data Governance and Regulatory Compliance in Federated Learning and Edge Computing for Healthcare. *Research Square.* (2025)

[20] Chen, Z., Liao, W., Hua, K., Lu, C. & Yu, W. Towards asynchronous federated learning for heterogeneous edge-powered internet of things. *Digital Communications And Networks.* pp. 317-326 (2021)

[21] Lu, Z., Pan, H., Dai, Y., Si, X. & Zhang, Y. Federated Learning With Non-IID Data: A Survey. *IEEE Internet Of Things Journal.* pp. 19188-19209 (2024)

[22] Chen, J., Xu, W., Guo, S., Wang, J., Zhang, J. & Wang, H. FedTune: A Deep Dive into Efficient Federated Fine-Tuning with Pre-trained Transformers. *ArXiv Preprint ArXiv:2211.08025.* (2022)

[23] Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *Proceedings Of The 2020 Conference On Machine Learning And Systems.* (2020)

[24] Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. *Proceedings Of The 37th International Conference On Machine Learning.* (2020)

[25] Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; McMahan, H.B. Adaptive Federated Optimization. *arXiv preprint arXiv:2003.00295.* (2020)

[26] Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling The Objective Inconsistency Problem In Heterogeneous Federated Optimization. *Advances In Neural Information Processing Systems.* (2020)

[27] Arivazhagan, M. G.; Aggarwal, V.; Singh, A. K.; Choudhary, S. Federated Learning with Personalization Layers. *ArXiv Preprint ArXiv:1912.00818.* (2019).

[28] Diao, E.; Ding, J.; Tarokh, V. HeteroFL: Computation And Communication Efficient Federated Learning For Heterogeneous Clients. *International Conference On Learning Representations.* (2021)

[29] Li, D.; Wang, J. FedMD: Heterogenous Federated Learning Via Model Distillation. *Proceedings Of The 33rd Conference On Neural Information Processing Systems Workshops.* (2019)

[30] Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; et al. Overcoming Catastrophic Forgetting In Neural Networks. *Proceedings Of The National Academy Of Sciences.* (2017)

[31] Li, Z.; Hoiem, D. Learning Without Forgetting. *European Conference On Computer Vision.* pp. 614–629 (2016)

[32] Farajtabar, M.; Azizan, N.; Mott, A.; Li, A. Orthogonal Gradient Descent For Continual Learning. *Proceedings Of The 23rd International Conference On Artificial Intelligence And Statistics.* (2020)

[33] Yoon, J.; Jeong, W.; Ju, J.; Hwang, S.J.; Yang, E. Federated Continual Learning With Weighted Inter-Client Transfer. *Proceedings Of The 38th International Conference On Machine Learning.* (2021)

[34] Guo, Zhihan, et al. Fedlfc: Towards efficient federated multilingual modeling with lora-based language family clustering. *Findings of the Association for Computational Linguistics: NAACL 2024.* 2024

[35] Chen, H.; Zhang, Y.; Krompass, D.; Gu, J.; Tresp, V. FedDAT: An Approach for Foundation Model Finetuning in Multi-Modal Heterogeneous Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence.* (2024)

[36] Shu, Y.; Hu, W.; Ng, S. K.; Low, B. K. H.; Yu, F. R. Ferret: Federated Full-Parameter Tuning at Scale for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning.* (2025)

[37] Alistarh, D.; Grubic, D.; Li, J.; Tomioka, R.; Vojnovic, M. QSGD: Communication-Efficient SGD Via Gradient Quantization And Encoding. *Advances In Neural Information Processing Systems.* pp. 1709–1720 (2017)

[38] Karimireddy, S.P.; Rebjock, Q.; Stich, S.; Jaggi, M. Error Feedback Fixes SignSGD And Other Gradient Compression Schemes. *Proceedings Of The 36th International Conference On Machine Learning.* pp. 3252–3261 (2019)

[39] Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation For Knowledge-Intensive NLP. *Advances In Neural Information Processing Systems.* (2020)

[40] Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* pp. 3045–3059 (2021)

[41] Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts For Generation. *Proceedings Of The 59th Annual Meeting Of The Association For Computational Linguistics.* pp. 4582–4597 (2021).

[42] Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y.ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the 11th International Conference on Learning Representations.* (2023)

[43] Shi, H.; Xu, Z.; Wang, H.; Qin, W.; Wang, W.; Wang, Y.; Wang, Z.; Ebrahimi, S.; Wang, H. Continual Learning of Large Language Models: A Comprehensive Survey.*ACM Computing Surveys.* (2025).

[44] Huang, J.; Cui, L.; Wang, A.; Yang, C.; Liao, X.; Song, L.; Yao, J.; Su, J. Mitigating Catastrophic Forgetting in Large Language Models with Self-Synthesized Rehearsal. *Proceedings Of The 62nd Annual Meeting Of The Association For Computational Linguistics.* pp. 1450-1466 (2024)

[45] Shumailov, I.; Shumailov, R.; Papernot, N.; et al. AI models collapse when trained on recursively generated data. *Nature.* 630, pp. 971–978 (2024)

[46] Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., Liu, Y. BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning. *Proceedings of the 2020 USENIX Annual Technical Conference.* pp. 493–506 (2020)

[47] Chen, W.-N., Özgür, A., Cormode, G., Bharadwaj, A. The Communication Cost of Security and Privacy in Federated Frequency Estimation. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics.* pp. 4247–4274 (2023)

[48] Tang, L.; Joshi, J. Towards Privacy-Preserving and Secure Machine Unlearning: Taxonomy, Challenges and Research Directions. *Proceedings of the IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications.* pp. 280–291 (2024)