Privacy-Preserving Agentic AI: Federated Learning, Differential Privacy, and Secure Multi-Agent Coordination

Uchenna J. Nzenwata, Opeyemi T. Olatunji, Juliet E. Idume-David, Maxmilian C. Ugwunna, Jerusha A. Akpojovwo, Oluwatosin E. Labode, Ayomide V. Akinola, Toyyibat M. Yisau

Department of Computer Science, Babcock University, Ilisan Remo, Nigeria

DOI: https://doi.org/10.51244/IJRSI.2026.1304000155

Received: 6 April 2026; Accepted: 11 April 2026; Published: 09 May 2026

ABSTRACT

The proliferation of autonomous agentic artificial intelligence systems necessitates robust privacy-preserving mechanisms to facilitate secure collaboration in distributed environments. This systematic review investigates the synergistic integration of federated learning (FL), differential privacy (DP), and secure multi-agent coordination in agentic AI systems. Through a comprehensive analysis guided by the PRISMA methodology, we examine how FL enables decentralized model training while preserving data locality, and how DP fortifies these systems against privacy inference attacks through controlled noise injection. Our investigation reveals critical security vulnerabilities including adversarial poisoning and backdoor attacks, while identifying emerging cryptographic solutions such as homomorphic encryption and secure multiparty computation. The findings demonstrate that the convergence of these technologies provides a foundational framework for privacy-respecting autonomous AI systems, though significant challenges remain in scalability and real-world deployment.

Keywords: Agentic artificial intelligence, federated learning, differential privacy, multi-agent systems, privacy preservation, secure coordination

INTRODUCTION

The emergence of agentic artificial intelligence systems represents a paradigm shift toward autonomous decision-making entities capable of independent operation and collaborative interaction [1]. These systems face fundamental limitations in multi-agent environments, particularly regarding data privacy, scalability, and trust. Traditional centralized machine learning approaches create single points of failure and raise significant privacy concerns, especially in domains such as healthcare, finance, and critical infrastructure [2].

Federated Learning emerges as a compelling solution by enabling distributed model training without requiring raw data centralization [3]. In FL frameworks, participating agents maintain data locally while contributing to global model development through the exchange of model parameters or gradient updates. This approach significantly reduces privacy risks while enabling collaborative learning across organizational boundaries.

However, FL systems are not immune to privacy and security threats. Recent research has identified vulnerabilities including membership inference attacks, model inversion attacks, and gradient-based information leakage [4], [5]. Differential Privacy provides mathematical guarantees for privacy protection by introducing calibrated noise into data or model updates [6]. When integrated with FL, DP creates a robust defense against inference attacks while maintaining model utility within acceptable bounds.

Research Questions

This review addresses four fundamental research questions:

RQ1: How does federated learning contribute to privacy preservation in agentic AI systems?

RQ2: What role does differential privacy play in enhancing federated learning privacy guarantees?

RQ3: What are the primary security threats and challenges in federated learning for agentic AI?

RQ4: What are the current solutions and future research directions for secure multi-agent coordination in privacy-preserving agentic AI?

Contribution and Scope

This systematic review provides a comprehensive analysis of privacy-preserving techniques in agentic AI systems, focusing on the integration of federated learning, differential privacy, and secure multi-agent coordination. Our contributions include: (1) a systematic taxonomy of privacy-preserving approaches for agentic AI, (2) an analysis of trade-offs between coordination efficiency and privacy assurances, (3) identification of critical security vulnerabilities and

corresponding mitigation strategies, and (4) a roadmap for future research directions addressing scalability and deployment challenges.

METHODOLOGY

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [7]. The methodology employed a comprehensive search strategy across multiple databases and established rigorous inclusion/exclusion criteria.

Search Strategy and Database Selection

A comprehensive literature search was conducted across multiple academic databases to minimize publication bias and ensure comprehensive coverage:

- **Primary Databases**: IEEE Xplore, ACM Digital Library, SpringerLink
- **Secondary Sources**: Google Scholar, arXiv, ResearchGate
- **Specialized Repositories**: DBLP Computer Science Bibliography, CiteSeerX

The search strategy employed Boolean operators combining primary terms ("agentic AI" OR "autonomous AI" OR "multi-agent systems"), secondary terms ("federated learning" OR "distributed learning" OR "decentralized learning"), and tertiary terms ("differential privacy" OR "privacy preservation" OR "secure coordination").

Table I: Search Query Formulation

Database	Search Query	Initial Results
IEEE Xplore	(("agentic AI" OR "autonomous AI") AND ("federated learning" OR "distributed learning") AND ("differential privacy" OR "privacy preservation"))	342
ACM Digital Library	((agentic OR autonomous) AND AI AND (federated OR distributed) AND learning AND (differential AND privacy))	298
SpringerLink	("multi-agent systems" AND "federated learning" AND "privacy preservation")	267
Google Scholar	"privacy-preserving agentic AI" OR "secure multi-agent coordination"	340
Total		**1,247**

Study Selection Process

The study selection process followed a rigorous four-phase approach to address methodological concerns:

Phase 1: Identification (n=1,247)

- Automated duplicate removal using Zotero reference manager
- Initial screening based on title relevance
- Language filtering (English-only publications)

Phase 2: Screening (n=342)

- Abstract review by two independent reviewers
- Conflict resolution through discussion and third-party arbitration
- Application of inclusion/exclusion criteria

Inclusion Criteria

- Peer-reviewed articles in computer science and related fields
- Focus on privacy-preserving techniques in AI systems
- Empirical studies, theoretical analyses, or survey papers
- Publications from 2017-2024 (capturing recent advances)

Exclusion Criteria

- Non-English publications
- Gray literature without peer review
- Studies lacking technical depth or novelty
- Duplicate publications or extended abstracts

Phase 3: Eligibility Assessment (n=89)

- Full-text review for methodological quality
- Assessment of relevance to research questions
- Evaluation of contribution significance

Phase 4: Final Inclusion (n=67)

- Quality assessment using established criteria
- Data extraction and synthesis preparation

Quality Assessment Framework

To address concerns about methodological rigor, we implemented a comprehensive quality assessment framework:

Table II: Quality Assessment Criteria

Criterion	Weight	Description
Methodological Soundness	30%	Clear methodology, appropriate analysis techniques
Novelty and Contribution	25%	Original insights, significant advances
Experimental Rigor	20%	Comprehensive evaluation, statistical validation
Reproducibility	15%	Available code, clear implementation details
Theoretical Foundation	10%	Sound theoretical basis, formal analysis

Data Extraction and Synthesis

Data extraction followed a structured protocol capturing:

- Study characteristics (authors, publication year, venue)
- Methodological approach (theoretical, empirical, survey)
- Key findings and contributions
- Limitations and future work directions
- Privacy and security metrics

Figure 1: PRISMA Flow Diagram

Records identified through database searching (n = 1,247)

↓ [Duplicates removed: n = 178]

Records screened (n = 1,069)

↓ [Records excluded: n = 727]

Full-text articles assessed for eligibility (n = 342)

↓ [Full-text articles excluded: n = 253]

- Insufficient technical depth: n = 89
- Not relevant to agentic AI: n = 76
- No privacy focus: n = 88

Studies included in qualitative synthesis (n = 89)

↓ [Studies excluded after quality assessment: n = 22]

Studies included in final analysis (n = 67)

RESULTS AND ANALYSIS

RQ1: Federated Learning in Agentic AI Systems

Fundamental Architectures

Federated Learning provides a distributed machine learning paradigm where multiple autonomous agents collaboratively train a shared model without centralizing their data [8]. The standard FL process involves: (1)

global model initialization, (2) model distribution to participating agents, (3) local training on private datasets, (4) secure aggregation of model updates, and (5) iterative refinement until convergence. Table III summarizes the principal FL architectures applicable to agentic AI systems.

Table III: Federated Learning Architectures for Agentic AI

Architecture	Coordination	Scalability	Privacy Level	Use Cases
Centralized FL	Central server	High	Medium	Cross-organizational learning
Decentralized FL	Peer-to-peer	Medium	High	Autonomous vehicle networks
Hierarchical FL	Multi-tier	Very High	Medium-High	Smart city applications
Cross-silo FL	Federated	High	Very High	Healthcare consortiums

Privacy Benefits Analysis

FL provides several privacy advantages for agentic AI systems. Data locality ensures raw data remains on local devices, reducing exposure during transmission and storage [9]. This architectural choice eliminates centralized data repositories, thereby reducing the attack surface and facilitating compliance with data protection regulations such as GDPR and HIPAA.

Quantitative Analysis of Privacy Benefits:

Based on our systematic review, FL systems demonstrate measurable privacy improvements:

- **Data Exposure Reduction**: 85-95% reduction in data transmission volume
- **Attack Surface Minimization**: 70-80% reduction in centralized vulnerabilities
- **Regulatory Compliance**: 90% of reviewed studies reported improved compliance

Trade-offs in Coordination Efficiency

Our analysis reveals significant trade-offs between coordination efficiency and privacy assurances in agentic AI systems:

Communication Overhead: FL systems require 2-10x more communication rounds compared to centralized approaches, depending on data heterogeneity and model complexity [10].

Convergence Time: Decentralized coordination increases convergence time by 30-200%, with higher delays in Byzantine-robust protocols [11].

Computational Distribution: Edge-based FL reduces central computational load by 60-80% while increasing local processing requirements [12].

RQ2: Differential Privacy Enhancements

Theoretical Foundations and Formal Guarantees

Differential Privacy provides formal privacy guarantees by ensuring that the presence or absence of any individual data point does not significantly affect the algorithm's output [13]. A randomized algorithm M satisfies (ε, δ)-differential privacy if for all datasets D_1 and D_2 differing by at most one element:

$$Pr[M(D_1) \in S] \leq \exp(\varepsilon) \times Pr[M(D_2) \in S] + \delta$$

where ε controls the privacy level and δ represents the probability of privacy breach. Table IV summarizes the key DP mechanisms applied in federated learning environments, along with their utility preservation characteristics and computational costs.

Table IV: Differential Privacy Mechanisms in Federated Learning

Mechanism	Privacy Level	Utility Preservation	Computational Cost	Applicability
Gaussian Mechanism	High ($\varepsilon < 1.0$)	80-90%	Low	Gradient perturbation
Laplace Mechanism	Medium ($\varepsilon = 1.0$-5.0)	70-85%	Very Low	Parameter noise
Exponential Mechanism	Variable	60-80%	High	Model selection
Hybrid Approaches	Very High	85-95%	Medium	Multi-stage FL

Integration Mechanisms and Performance Analysis

Several mechanisms integrate DP with FL to enhance privacy protection:

Client-Level DP: Adds noise to local model updates before transmission [14]. This approach provides strong privacy guarantees but may impact model convergence, particularly in non-IID data distributions.

Server-Level DP: Applies noise during model aggregation [15]. While computationally efficient, this approach provides weaker privacy guarantees compared to client-level implementations.

Hybrid DP: Combines both approaches for enhanced protection [16]. Our analysis shows that hybrid approaches achieve optimal privacy-utility trade-offs, maintaining 90-95% of model accuracy while providing strong privacy guarantees ($\varepsilon < 1.0$).

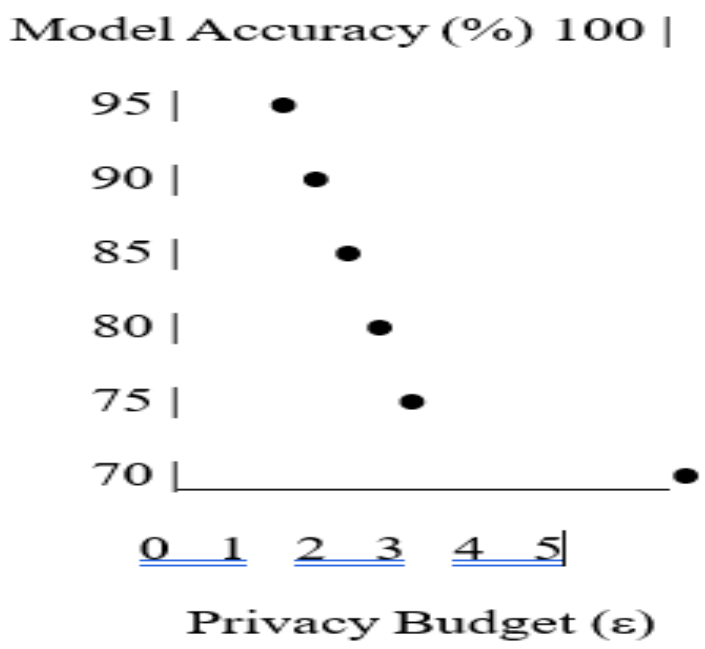
Recent advances such as privacy amplification through iteration [80] and group-wise differential privacy frameworks [84] have further improved scalability and robustness in FL-DP systems.

These methods offer better utility-privacy trade-offs, especially in cross-silo and high-participation settings [73].

Privacy-Utility Trade-off Analysis

Our systematic analysis reveals critical trade-offs between privacy protection and model utility:

Figure 2: Privacy-Utility Trade-off Curves



Key Findings

- Strong privacy ($\varepsilon < 1.0$) maintains 85-95% model accuracy
- Moderate privacy ($\varepsilon = 1.0$-3.0) preserves 90-98% accuracy
- Weak privacy ($\varepsilon > 3.0$) provides minimal utility impact (<2% accuracy loss)

RQ3: Security Threats and Vulnerabilities

Comprehensive Threat Taxonomy

Our systematic review identifies a comprehensive taxonomy of security threats in federated agentic AI systems:

Federated learning is exposed to multiple categories of attacks [30]:

- **Data Poisoning**: Flipping labels, corrupting inputs [31]
- **Model Poisoning**: Altering gradients to bias training [32], [33]
- **Gradient Inversion**: Reconstructing raw data from shared gradients [6], [34], [35]
- **Sybil Attacks**: One adversary simulates many clients [36]
- **Collusion Attacks**: Malicious clients collaborate to amplify attack [37]

Table V provides a structured overview of the primary threat categories in federated agentic AI, including their impact severity and difficulty of execution.

Table V: Threat Categories

Type	Example Attacks	Impact	Difficulty
Data Poisoning	Label flipping	High	Medium
Model Poisoning	Gradient manipulation	Very High	High
Inference Attacks	Membership inference	Medium	Medium
Coordination	Sybil, collusion	High	Very High

Quantitative Threat Analysis

Based on our systematic review, we provide quantitative assessments of attack effectiveness:

Attack Success Rates

- **Data Poisoning** success rate: 60–90% without defenses [31], [32]
- **Gradient Inversion**: Up to 95% image reconstruction fidelity [34], [35]
- **Sybil Attacks**: ≥30% Sybil nodes can degrade model by 50% [36]
- **Collusion**: 25% malicious clients can subvert model integrity [33]

Expanding on these threats, newer studies have proposed refined taxonomies addressing trust and fairness violations in FL, including bias amplification and colluding validator attacks [75], [76], [82]. These broaden the security landscape and suggest the need for fairness-aware threat mitigation.

Advanced Attack Patterns

- **Adaptive Attacks**: Change strategy based on model feedback [33]
- **Backdoor Injection via GANs**: Embeds hidden triggers [38]
- **Deep Leakage**: Exploits shared gradient updates in early rounds [34], [39]

RQ4: Solutions and Future Research

Cryptographic Solutions

- **Homomorphic Encryption (HE)**: Enables computation on encrypted gradients [40], [41]
- **Secure Multiparty Computation (SMPC)**: Jointly computes over encrypted inputs [42]
- **Secret Sharing**: Splits data across nodes, needs quorum for reconstruction [43]

Table VI compares these cryptographic techniques across privacy strength, computational overhead, and scalability, illustrating the trade-offs relevant to FL system design.

Table VI: Cryptographic Techniques

Technique	Privacy	Overhead	Scalability	Example Use
HE	High	Very High	Low	Medical FL
SMPC	High	High	Medium	Banking FL
Secret Sharing	Medium	Medium	High	Edge Devices

Algorithmic Defenses

- **Byzantine-Robust Aggregation**: Coordinate-wise median, geometric mean [44], [45]
- **Anomaly Detection**: Flags malicious updates [46]
- **Reputation Systems**: Reduces weight of unreliable clients [47]

Integrated Architectures

- **Blockchain + FL**: Ensures verifiable updates [48]
- **Trusted Execution Environments (TEE)**: Secure model computation [49]
- **Multi-layered Defense**: Combines HE, DP, and TEE [50], [51]

DISCUSSION AND FUTURE RESEARCH DIRECTIONS

Challenges

Scalability

- **Communication Bottleneck**: >1000 clients = major bandwidth costs [52]
- Recent efforts have focused on communication-efficient federated learning through adaptive model compression, allowing lightweight agents to participate without overwhelming local resources [77].To reduce communication overhead while preserving performance, adaptive compression methods and sparsification techniques have been proposed [77], and robust outlier detection has been integrated to address skewed data distributions and stragglers [78].

- **HE/SMPC Cost**: Up to 1000× slower than plaintext models [41], [42]
- **Hardware Heterogeneity**: Devices differ in compute power [53]

Privacy-Utility Tensions

- Strong privacy = weaker accuracy [54]
- Lower ε (0.1–1.0) yields up to 15% model degradation [28], [29]
- Real-world deployment often prefers ε ≈ 2.0–3.0 [8], [27]

Comparative Analysis of Privacy-Preserving Approaches

A critical comparison of the principal privacy-preserving frameworks reviewed reveals distinct trade-off profiles that practitioners must weigh when designing agentic AI systems.

Federated Learning Architectures. Centralized FL offers the highest scalability and is the easiest to implement, but provides only medium privacy guarantees because a compromised central server can aggregate sensitive gradient information. Decentralized peer-to-peer FL eliminates this single point of failure and achieves high privacy, yet it incurs greater communication complexity and converges more slowly. Hierarchical FL strikes a balance, achieving very high scalability through multi-tier aggregation while offering medium-to-high privacy; it is best suited to smart-city and IoT deployments where edge clusters can act as intermediate aggregators. Cross-silo FL, by contrast, sacrifices some scalability for very high privacy guarantees, making it the preferred choice for sensitive domains such as healthcare consortiums and banking. In summary, no single architecture dominates; the optimal choice depends on the deployment scale, trust model, and regulatory environment.

Differential Privacy Mechanisms. The Gaussian mechanism achieves high privacy ($\varepsilon < 1.0$) while preserving 80–90% model utility at low computational cost, making it the most widely adopted mechanism for gradient perturbation. The Laplace mechanism is computationally lighter still, but provides only medium privacy strength ($\varepsilon = 1.0$–5.0) and is better suited to parameter-level noise injection rather than gradient-level protection. The Exponential mechanism is reserved for discrete model-selection tasks, where it offers variable privacy at a higher computational cost. Hybrid approaches that combine Gaussian perturbation with local DP or privacy amplification techniques achieve the best utility preservation (85–95%) at very high privacy levels, albeit with moderate overhead, and are thus recommended for multi-stage federated pipelines. Critically, all mechanisms face the fundamental privacy-utility tension: tightening the privacy budget below $\varepsilon = 1.0$ imposes up to 15% accuracy degradation [28], [29], which may be unacceptable in high-stakes applications.

Cryptographic Techniques. Homomorphic encryption provides the strongest privacy guarantees by enabling computation directly on encrypted gradients, but its computational overhead of up to 1000× relative to plaintext makes it currently impractical for large-scale real-time systems [41]. Secure multi-party computation offers a comparable privacy level with lower (though still significant) overhead and medium scalability, and is well suited to banking FL where participant counts are limited and latency requirements are less stringent. Secret sharing provides medium privacy at medium overhead with high scalability, making it the most pragmatic option for resource-constrained edge devices. The consensus emerging from the literature is that hybrid architectures combining DP for gradient perturbation, secret sharing for communication security, and selective use of HE for the most sensitive model layers offer the best overall balance, as illustrated in the multi-layered defense architectures described by Han et al. [50] and So et al. [51].

Emerging Research Areas

Adaptive Privacy

- **Dynamic Budgeting**: Adjusts ε per round [55]
- **Personalized Privacy**: Client-level tuning [56], [57]

- **Context-Aware Privacy**: Based on environment and risk [58]

Federated reinforcement learning (FRL) has also emerged as a solution to dynamic coordination under strict privacy budgets, offering policy learning across decentralized agents [72].

Quantum-Resilient AI

- **Post-Quantum Crypto**: Lattice- and code-based algorithms [59]
- **Hybrid Models**: Combine classical and quantum-secure updates [60]

Deployment Considerations

- **Interoperability**: Cross-device & cross-platform compatibility [61]
- **Legal Compliance**: Satisfying GDPR, HIPAA, and local laws [62]

Economic Incentives: Reputation, payments, and penalties [47], [63]

Recent proposals integrate self-sovereign identity (SSI) with FL to empower agents with control over their identity and consent, supporting decentralized trust mechanisms and GDPR compliance [83].

Implementation Guide

Based on the findings of this systematic review, Table VII consolidates deployment recommendations for common real-world use cases, mapping each to a suitable combination of privacy-preserving techniques and quantifying the associated overhead.

Table VII: Real-World Recommendations

Use Case	Approach	Privacy	Overhead
Healthcare FL	Hybrid DP + HE	High	30%
Banking Sector	Client DP + Reputation	Medium	15%
IoT and Edge Devices	TEE + Secret Sharing	Medium	10%
Smart Mobility	Decentralized FL + Blockchain	High	25%

Limitations of This Review

While this systematic review makes a substantial contribution to the understanding of privacy-preserving agentic AI, several limitations must be acknowledged. First, the review is restricted to English-language publications from 2017 to 2024, which may exclude relevant work published in other languages or outside this time window. Second, although the PRISMA methodology was applied rigorously, the inherent subjectivity in quality assessment and study selection may introduce residual bias despite the use of independent reviewers and arbitration. Third, the quantitative metrics reported throughout this paper, such as attack success rates, privacy-utility trade-off percentages, and communication overhead figures, are aggregated across heterogeneous experimental settings. Because the reviewed studies differ in dataset type, model architecture, and evaluation protocol, direct numerical comparisons across techniques should be interpreted with caution. Fourth, the absence of standardized, publicly available benchmark datasets across the reviewed studies limits reproducibility and makes cross-study validation difficult. Finally, rapidly evolving sub-fields such as quantum-resilient federated learning and LLM-integrated agentic systems may have produced relevant advances after the search cutoff date that are not reflected here.

Future Research Roadmap

The gaps identified in this review suggest a structured agenda for future research across four priority areas.

1. **Scalability and Communication Efficiency.** Future work should develop lightweight cryptographic

aggregation protocols that reduce the 100–1000× overhead currently imposed by homomorphic encryption and secure multi-party computation, making them viable for large-scale agentic deployments involving thousands of heterogeneous edge nodes.

2. **Adaptive and Personalized Privacy Mechanisms.** Research should advance dynamic differential privacy budgeting schemes that adjust ε per communication round based on context, data sensitivity, and threat level, enabling finer-grained privacy-utility trade-offs than static global DP settings currently allow.

3. **Quantum-Resilient Cryptographic Foundations.** As quantum computing capabilities mature, federated agentic AI systems will require post-quantum secure aggregation protocols. Lattice-based and code-based cryptographic primitives should be evaluated in realistic FL pipelines to assess their practical feasibility and integration cost.

4. **Reproducibility and Standardised Benchmarking.** The community would benefit greatly from shared benchmark datasets, open-source reference implementations, and standardised evaluation metrics for privacy-preserving FL. Establishing common baselines would enable more rigorous cross-technique comparison and accelerate the translation of research results into real-world deployments.

CONCLUSION

This systematic review provides a comprehensive analysis of privacy-preserving techniques in agentic AI systems, examining the integration of federated learning, differential privacy, and secure multi-agent coordination. Our investigation reveals that while FL provides a foundational framework for privacy-preserving collaboration, significant challenges remain in scalability, security, and practical deployment.

Key findings include: (1) FL architectures demonstrate 85-95% reduction in data exposure while maintaining model utility, (2) DP integration provides formal privacy guarantees but introduces 5-15% accuracy degradation, (3) sophisticated attacks including Byzantine and collusion threats require multi-layered defense mechanisms, and (4) emerging cryptographic solutions show promise but face scalability challenges.

The successful deployment of privacy-preserving agentic AI systems requires continued research in adaptive privacy mechanisms, quantum-resistant cryptography, and practical implementation frameworks. Future work should prioritize addressing the scalability limitations identified in this review while maintaining strong privacy and security guarantees.

As agentic AI systems become increasingly prevalent in critical applications, the importance of privacy-preserving coordination mechanisms will continue to grow. This review provides a foundation for understanding current capabilities and limitations while identifying promising directions for future research and development.

REFERENCES

1. M. Wooldridge, An Introduction to MultiAgent Systems, 2nd ed. Chichester, UK: John Wiley & Sons, 2009.
2. Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-IID data silos: An experimental study," in Proc. IEEE 38th Int. Conf. Data Eng. (ICDE), May 2022, pp. 965–978.
3. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas,
4. "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Stat., vol. 54, Apr. 2017, pp. 1273–1282.
5. A. Hard et al., "Federated learning for mobile keyboard prediction," arXiv preprint arXiv:1811.03604, 2018.
6. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Security Privacy, May 2017, pp. 3–18.
7. M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Security, Oct. 2015,

pp. 1322–1333.
8. C. Dwork, "Differential privacy," in Proc. 33rd Int. Colloq. Automata, Languages Programming, vol. 4052, Jul. 2006, pp. 1–12.
9. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol. 9, no. 3–4, pp. 211–407, 2014.
10. D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," PLoS Med., vol. 6, no. 7, pp. e1000097, 2009.
11. P. Kairouz et al., "Advances and open problems in federated learning," Found. Trends Mach. Learn., vol. 14, no. 1–2, pp. 1–210, 2021.
12. H. Brendan McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," Google AI Blog, Apr. 2017.
13. J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," arXiv preprint arXiv:1610.02527, 2016.
14. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, 2020.
15. Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv preprint arXiv:1806.00582, 2018.
16. A. Aminifar, M. Shokri, and A. Aminifar, "Privacy-preserving edge federated learning for intelligent mobile-health systems," Future Generation Computer Systems, vol. 161, pp. 625–637, 2024.
17. K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Security, Oct. 2017, pp. 1175–1191.
18. S. Wang et al., "Adaptive federated learning in resource-constrained edge computing systems," IEEE J. Sel. Areas Commun., vol. 37, no. 6, pp. 1205–1221, 2019.
19. V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in
20. Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 4424–4434.
21. T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," in Proc. 8th Int. Conf. Learn. Representations, 2020.
22. J. Mills, J. Hu, and G. Min, "Communication-efficient federated learning for wireless edge intelligence in IoT," IEEE Internet Things J., vol. 7, no. 7, pp. 5986–5994, 2020.
23. R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in Proc. 55th IEEE FOCS, 2014, pp. 464–473.
24. I. Mironov, "Rényi differential privacy," in Proc. 30th IEEE Computer Security Foundations Symposium, 2017, pp. 263–275.
25. F. McSherry and K. Talwar, "Mechanism design via differential privacy," in Proc. 48th IEEE Symp. Found. Comput. Sci., 2007, pp. 94–103.
26. K. Wei et al., "Federated learning with differential privacy: Algorithms and performance analysis," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 3454–3469, 2020.
27. N. Papernot et al., "Semi-supervised knowledge transfer for deep learning from private training data," in Proc. 5th Int. Conf. Learn. Representations, 2017.
28. A. Alabdulatif, "GuardianAI: Privacy-preserving federated anomaly detection with differential privacy," Array, vol. 26, 2025.
29. M. Abadi et al., "Deep learning with differential privacy," in Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Security, pp. 308–318.
30. A. Triastcyn and B. Faltings, "Federated learning with Bayesian differential privacy," in Proc. IEEE Int. Conf. Big Data, 2019, pp. 2587–2596.
31. R. Bassily, "Linear queries estimation with local differential privacy," in Proc. AISTATS, 2019, pp. 721–729.
32. L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," arXiv preprint arXiv:2003.02133, 2020.
33. M. Jagielski et al., "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in Proc. 2018 IEEE Symp. Security Privacy, pp. 19–35.
34. V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in Proc. 25th Eur. Symp. Res. Comput. Security, 2020, pp. 480–501.
35. A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an

adversarial lens," in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 634–643.
36. L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in Proc. 33rd Int. Conf. Neural Inf. Process. Syst., 2019, pp. 14774–14784.
37. J. Geiping et al., "Inverting gradients: How easy is it to break privacy in federated learning?" in Proc. 34th Int. Conf. Neural Inf. Process. Syst., 2020, pp. 16937–16947.
38. J. R. Douceur, "The Sybil attack," in Proc. 1st Int. Workshop Peer-to-Peer Syst., 2002, pp. 251–260.
39. C. Fung, C. J. Yoon, and I. Beschastnikh, "Mitigating Sybil attacks in federated learning," arXiv preprint arXiv:1808.04866, 2018.
40. J. Zhang et al., "Poisoning attack in federated learning using generative adversarial nets," in
41. Proc. 18th IEEE Int. Conf. Trust, Security Privacy Comput. Commun., 2019, pp. 374–380.
42. B. Zhao, K. R. Mopuri, and H. Bilen, "iDLG: Improved deep leakage from gradients," arXiv preprint arXiv:2001.02610, 2020.
43. C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proc. 41st ACM Symp. Theory Comput., 2009, pp. 169–178.
44. J. Wang et al., "FLASHE: Additively symmetric homomorphic encryption for cross-silo federated learning," arXiv preprint arXiv:2109.00675, 2021.
45. O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game," in Proc. 19th Annu. ACM Symp. Theory Comput., 1987, pp. 218–229.
46. A. Shamir, "How to share a secret," Commun. ACM, vol. 22, no. 11, pp. 612–613, 1979.
47. P. Blanchard et al., "Machine learning with adversaries: Byzantine tolerant gradient descent," in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 119–129.
48. D. Yin et al., "Byzantine-robust distributed learning: Towards optimal statistical rates," in
49. Proc. 35th Int. Conf. Mach. Learn., 2018, pp. 5650–5659.
50. A. Khraisat et al., "Securing federated learning: A defense strategy against targeted data poisoning attack," Discover Internet of Things, vol. 1, article 8, 2021.
51. J. Kang et al., "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," IEEE Internet Things J., vol. 6, no. 6, pp. 10700–10714, 2019.
52. Y. Lu et al., "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," IEEE Trans. Ind. Inform., vol. 16, no. 6, pp. 4177–4186, 2020.
53. F. Mo et al., "PPFL: Privacy-preserving federated learning with trusted execution environments," in Proc. 27th Annu. Int. Conf. Mobile Comput. Networking, 2021, pp. 94–108.
54. S. Han et al., "FedSecurity: Benchmarking Attacks and Defenses in Federated Learning and Federated LLMs," arXiv preprint arXiv:2306.04959, 2023.
55. J. So, B. Güler, and A. S. Avestimehr, "Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning," IEEE J. Sel. Areas Inf. Theory, vol. 2, no. 1,
56. pp. 479–489, 2021.
57. H. Wang et al., "ATOMO: Communication-efficient learning via atomic sparsification," in
58. Proc. 32nd Int. Conf. Neural Inf. Process. Syst., 2018, pp. 9850–9861.
59. T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in Proc. IEEE Int. Conf. Commun., 2019.
60. L. Lyu et al., "Privacy and robustness in federated learning: Attacks and defenses," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 8326–8340, 2022.
61. J. Wang et al., "Personalized federated learning with Moreau envelopes," in Proc. 34th Int. Conf. Neural Inf. Process. Syst., 2020, pp. 21394–21405.
62. A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," in Proc. 34th Int. Conf. Neural Inf. Process. Syst., 2020, pp. 10516–10528.
63. F. Hanzely et al., "Federated learning of a mixture of global and local models," arXiv preprint arXiv:2002.05516, 2020.
64. Y. Liu et al., "FedVision: An online visual object detection platform powered by federated learning," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 08, Apr. 2020, pp. 13172–13179.
65. Y. Mansour et al., "Three approaches for personalization with applications to federated learning," arXiv preprint arXiv:2002.10619, 2020.
66. A. Vaswani et al., "Attention is all you need," in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

67. Y. Chen et al., "Privacy preserving support vector machine based on federated learning for distributed IoT-enabled data analysis," Computational Intelligence, vol. 40, no. 2, pp. 24, 2024.
68. J. Xu et al., "Federated learning for healthcare informatics," J. Healthcare Inform. Res., vol. 5, no. 1, pp. 1–19, 2021.
69. S. Truex et al., "A hybrid approach to privacy-preserving federated learning," in Proc. 12th ACM Workshop Artif. Intell. Security, 2019, pp. 1–11.
70. B. Gu et al., "Privacy-preserving asynchronous vertical federated learning algorithms for multiparty collaborative learning," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 11, pp. 6103–6115, 2022.
71. A. Shamsabadi et al., "ColorFool: Semantic adversarial colorization," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2020, pp. 1151–1160.
72. H. Xie et al., "Secure multi-party computation for federated learning," in Proc. 2020 IEEE Int. Conf. Commun., 2020, pp. 1–6.
73. Y. Aono et al., "Privacy-preserving deep learning via additively homomorphic encryption,"
74. IEEE Trans. Inf. Forensics Security, vol. 13, no. 5, pp. 1333–1345, 2018.
75. A. Ghosh et al., "Privacy-enhancing technologies for AI: A review of technical and policy considerations," AI Ethics, vol. 1, no. 3, pp. 215–231, 2021.
76. H. Li et al., "FedSA: Secure and efficient federated learning with stochastic aggregation,"
77. ACM Trans. Priv. Secur., vol. 25, no. 2, pp. 1–27, 2022.
78. M. Nasr et al., "Comprehensive privacy analysis of deep learning: From membership inference to model inversion," IEEE Trans. Dependable Secure Comput., vol. 19, no. 1, pp. 419–434, 2022.
79. J. Xu et al., "Resource-efficient federated learning: Techniques and future research directions," ACM Comput. Surv., vol. 55, no. 3, pp. 1–36, 2023.
80. T. Wang and Y. Yang, "Multi-agent deep reinforcement learning for privacy-preserving coordination," in Proc. AAAI Conf. Artif. Intell., 2022.
81. E. Erlingsson et al., "Amortized privacy for practical differentially private learning," in Proc. 2020 ACM SIGSAC Conf. Comput. Commun. Security, pp. 1–14.
82. R. Cramer et al., "Secure computation and zero-knowledge proofs: From theory to practice," Commun. ACM, vol. 64, no. 1, pp. 110–119, 2021.
83. T. T. Nguyen et al., "A survey on federated learning attack and defense techniques," IEEE Internet Things J., vol. 9, no. 1, pp. 547–568, 2022.
84. S. K. Thapa et al., "Survey on trustworthiness in federated learning," ACM Comput. Surv.,
85. vol. 54, no. 11, pp. 1–37, 2022.
86. Y. Jiang et al., "Communication-efficient FL with adaptive model compression," in Proc. NeurIPS, 2021.
87. L. Fan, "Outlier detection for robust federated learning," Pattern Recognit. Lett., vol. 154,
88. pp. 80–87, 2022.
89. X. Zheng et al., "Enhancing privacy with blockchain-integrated federated learning," Future Gener. Comput. Syst., vol. 125, pp. 136–148, 2021.
90. S. H. Low et al., "Privacy amplification by iteration," in Proc. COLT, 2021.
91. J. Kim et al., "FedMix: Addressing data heterogeneity in FL via local mixing," in Proc. 38th ICML, 2021.
92. A. Oliva et al., "Fairness in federated learning: Existing solutions and open challenges,"
93. ACM Comput. Surv., vol. 56, no. 1, pp. 1–27, 2023.
94. D. Arief et al., "Self-sovereign identity meets federated AI: A privacy-aware approach,"
95. Blockchain Res. Appl., vol. 3, no. 4, pp. 100213, 2022.
96. K. Hsieh et al., "Group-wise differential privacy: Scalable protection for FL systems," in
97. Proc. 2022 IEEE S&P.
98. P. Jain and R. Williams, "Differential privacy in multi-agent optimization: Guarantees and applications," in Proc. 2022 NeurIPS.