Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 2, No. 2 (2026)

Article

Evaluation of Differential Privacy and Federated Learning for AI-Driven Customer Service Applications

Yajing Zhang [1,*]

[1] UCD Smurfit Graduate Business School, University College Dublin, Dublin, Ireland
* Correspondence: Yajing Zhang, UCD Smurfit Graduate Business School, University College Dublin, Dublin, Ireland

Abstract: The proliferation of artificial intelligence in customer service applications has heightened concerns about the protection of sensitive consumer data. This paper presents a comprehensive comparative evaluation of privacy-preserving techniques, specifically differential privacy and federated learning, within AI-driven customer service contexts. A multidimensional evaluation framework is proposed to assess these techniques across security robustness, model accuracy, computational efficiency, and algorithmic fairness. An experimental analysis of customer interaction datasets reveals that federated learning achieves 94.2% accuracy retention while maintaining privacy guarantees, whereas differential privacy mechanisms offer superior protection against membership inference attacks at a 12.3% accuracy trade-off. The findings provide actionable recommendations for enterprises seeking to balance data protection compliance with service quality optimization, contributing to the development of trustworthy AI systems aligned with regulatory requirements.

Keywords: Privacy-Preserving Machine Learning; Differential Privacy; Federated Learning; Customer Service AI; Privacy-Utility Trade-off

1. Introduction

1.1. Background and Motivation for Privacy Protection in AI-Enabled Customer Services

The integration of artificial intelligence into customer service operations has fundamentally transformed how enterprises interact with consumers. Modern AI-powered customer service platforms process vast quantities of personal data, including purchase histories, communication patterns, demographic information, and behavioral preferences. This data-driven approach enables sophisticated customer profiling and personalized recommendations, significantly enhancing service quality and customer satisfaction.

The exponential growth in data collection practices has precipitated substantial privacy concerns among consumers and regulatory bodies worldwide. Security vulnerabilities in federated learning systems and centralized AI deployments expose sensitive customer information to potential breaches and unauthorized access [1]. The emergence of stringent data protection regulations, including the General Data Protection Regulation in the European Union and the California Consumer Privacy Act in the United States, has mandated organizations to implement robust privacy safeguards in their AI deployments.

Privacy-preserving techniques have emerged as critical enablers for responsible AI development in customer-facing applications. These methodologies aim to extract valuable insights from sensitive data while providing mathematical guarantees against

Received: 30 January 2026
Revised: 18 March 2026
Accepted: 30 March 2026
Published: 02 April 2026

information leakage. Recent advances in federated learning and differential privacy have opened new possibilities for privacy-preserving machine learning at scale [2]. The successful implementation of such techniques directly impacts consumer trust, regulatory compliance, and the sustainable growth of AI-enabled services.

1.2. Research Objectives and Scope of Evaluation

This research addresses the fundamental challenge of evaluating and comparing privacy-preserving techniques for AI-driven customer service applications. The primary objectives encompass: (a) developing a comprehensive evaluation framework that captures the multifaceted nature of privacy protection in customer service contexts; (b) conducting systematic comparative analysis of differential privacy and federated learning approaches across security, accuracy, and efficiency dimensions; (c) assessing the fairness implications of privacy-preserving mechanisms on diverse customer demographic groups.

The scope of this evaluation focuses on privacy-preserving techniques applicable to customer profiling, personalized recommendation, and service-quality prediction. Federated learning presents unique challenges, including non-IID data distributions, communication efficiency, and system heterogeneity, that must be addressed in practical deployments [3]. The analysis considers both centralized and decentralized deployment scenarios, reflecting the heterogeneous infrastructure landscapes prevalent in modern enterprise environments.

1.3. Paper Organization and Contributions

The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of privacy-preserving techniques and relevant regulatory frameworks. Section 3 details the proposed evaluation methodology, including the design of metrics and the experimental configuration. Section 4 presents the comparative analysis results and discusses practical implications. Section 5 concludes with key findings and future research directions.

The principal contributions of this work include: a novel multidimensional evaluation framework for privacy-preserving techniques in customer service AI; an empirical comparative analysis using quantitative performance metrics; and actionable guidelines for enterprise adoption decisions.

2. Literature Review

2.1. Privacy-Preserving Techniques: Differential Privacy and Federated Learning

Privacy-preserving machine learning encompasses a diverse array of techniques designed to enable data analysis while protecting individual privacy. Differential privacy, introduced as a mathematical framework for quantifying privacy guarantees, has become the gold standard for privacy protection in statistical databases and machine learning applications. The algorithmic foundations of differential privacy provide rigorous mathematical guarantees that bound the information leakage from any individual record [4].

The formal definition of (ε, δ)-differential privacy stipulates that for any two neighboring datasets D and D' differing in one record, and for any subset S of possible outputs, $P[M(D) \in S] \leq \exp(\varepsilon) \cdot P[M(D') \in S] + \delta$, where ε represents the privacy budget and δ denotes the failure probability. Deep learning with differential privacy has demonstrated that neural networks can be trained with formal privacy guarantees through gradient clipping and noise addition mechanisms [5].

Federated learning represents an alternative paradigm that preserves privacy through data localization rather than perturbation. Communication-efficient learning from decentralized data enables model training across distributed devices without centralizing raw data [6]. In federated learning architectures, model training occurs on distributed client devices, with only model updates transmitted to a central aggregation server. This approach fundamentally eliminates the need for raw data centralization, addressing both privacy concerns and data sovereignty requirements.

The concept of federated machine learning extends beyond simple distributed training to encompass secure aggregation, differential privacy, and cross-silo collaboration [7]. Recent work has explored combining differential privacy with federated learning to enhance protection against inference attacks. Differentially private stochastic gradient descent injects noise during training, while secure aggregation protocols protect intermediate model updates during transmission.

2.2. Regulatory Frameworks: GDPR, CCPA, and AI Governance Standards

The regulatory landscape governing AI and data privacy has undergone significant evolution in recent years. The General Data Protection Regulation establishes comprehensive requirements for the processing of personal data, including principles of data minimization, purpose limitation, and individual consent. Article 22 of the GDPR specifically addresses automated decision-making, granting individuals the right to obtain human intervention and to challenge algorithmic decisions. The integration of transparency, fairness, and privacy principles in AI development has become essential for regulatory compliance [8].

The California Consumer Privacy Act, subsequently enhanced by the California Privacy Rights Act, introduces consumer rights regarding data access, deletion, and opt-out from data sales. The emerging framework for automated decision-making technology under the CPRA mandates transparency requirements and impact assessments for AI systems that process consumer data.

The National Institute of Standards and Technology has published guidelines for evaluating differential privacy guarantees, establishing standardized methodologies for assessing privacy protection claims [9]. The NIST differential privacy pyramid framework identifies multiple components that must be evaluated to verify privacy guarantees, including epsilon values, failure probabilities, and implementation correctness.

2.3. Privacy-Utility Trade-offs in Personalized Recommendation and Customer Profiling

The tension between privacy protection and service utility represents a fundamental challenge in privacy-preserving machine learning. Stronger privacy guarantees typically necessitate greater perturbation of data or model parameters, resulting in degraded prediction accuracy and personalization quality. A comprehensive review of fairness in machine learning reveals that privacy mechanisms can interact with fairness constraints in complex ways [10].

In personalized recommendation systems, privacy-preserving approaches must maintain sufficient signal fidelity to capture user preferences while preventing the inference of sensitive attributes. Collaborative filtering techniques under differential privacy constraints exhibit accuracy reductions ranging from 5% to 25%, depending on the privacy budget configuration. Surveys on bias and fairness in machine learning have found that privacy perturbation can disproportionately affect minority groups with limited representation in the training data [11].

Customer profiling applications face additional complexity due to the multi-attribute nature of customer representations. Privacy-preserving profiling must protect against attribute-inference attacks, in which adversaries attempt to infer sensitive characteristics from non-sensitive data correlations. The fairness implications of privacy mechanisms on different demographic groups introduce further considerations for equitable service delivery.

3. Methodology

3.1. Evaluation Framework and Metrics Design

The proposed evaluation framework adopts a multi-dimensional approach to assess privacy-preserving techniques across four primary dimensions: privacy robustness, model utility, computational efficiency, and algorithmic fairness. This framework addresses the limitations of single-metric evaluations that fail to capture the complex trade-offs inherent in privacy-preserving systems.

3.1.1. Privacy Robustness Metrics

Privacy robustness quantifies the effectiveness of privacy protection mechanisms against various attack vectors. The framework incorporates membership inference attack success rate (MIA-SR), attribute inference accuracy (AIA), and model inversion reconstruction error (MIRE) as primary indicators. Membership inference attacks attempt to determine whether a specific record was included in the training dataset, representing a direct privacy violation in customer service contexts where service utilization patterns constitute sensitive information. Secure and verifiable federated learning frameworks have demonstrated effectiveness in protecting against such attacks through cryptographic verification mechanisms [12].

The MIA-SR metric is computed as the area under the receiver operating characteristic curve for a binary classifier distinguishing members from non-members. Lower MIA-SR values approaching 0.5 (random guessing baseline) indicate stronger privacy protection. Attribute inference accuracy measures an adversary's ability to predict sensitive attributes from model outputs or intermediate representations (As shown in Table 1).

Table 1: Privacy Robustness Evaluation Metrics

Metric	Definition	Target Range	Attack Model
MIA-SR	Membership inference AUC	0.50-0.55	Shadow model attack
AIA	Attribute inference accuracy	<60%	Attribute inference classifier
MIRE	Reconstruction L2 distance	>2.0	Gradient-based inversion
ε-empirical	Empirical privacy budget	<1.0	Audit-based estimation
δ-failure	Privacy failure probability	<10^-5	Theoretical bound

3.1.2. Model Utility Metrics

Model utility assessment encompasses task-specific performance measures relevant to customer service applications. Classification accuracy, area under the precision-recall curve (AUPRC), and normalized discounted cumulative gain (nDCG) are the primary utility indicators for recommendation tasks. Privacy-preserving and verifiable decentralized federated learning frameworks enable utility measurement while maintaining formal privacy guarantees [13].

The utility retention ratio (URR) provides a normalized comparison between privacy-preserving and non-private baseline models, calculated as URR = Performance_private / Performance_baseline. This metric enables cross-technique comparisons independent of absolute performance levels. Additional metrics include RMSE for rating prediction tasks and F1-score for imbalanced classification scenarios commonly encountered in customer churn prediction.

3.2. Comparative Analysis Criteria: Security, Accuracy, and Computational Efficiency

The comparative analysis employs a structured evaluation protocol to ensure fair and reproducible comparisons between differential privacy and federated learning approaches. The experimental configuration standardizes dataset characteristics, model architectures, and hyperparameter selection procedures across all evaluated techniques.

3.2.1. Security Analysis Framework

Security evaluation extends beyond privacy metrics to encompass adversarial robustness and Byzantine fault tolerance. For federated learning systems, the analysis considers malicious client attacks, including model poisoning and gradient manipulation.

The Byzantine resilience threshold τ represents the maximum fraction of malicious clients the system can tolerate while maintaining convergence guarantees. Verifiable federated learning with privacy-preserving mechanisms for industrial IoT applications has demonstrated robust protection against Byzantine attacks [14].

The threat model assumes adversaries with white-box access to the trained model and black-box access to the training process. This assumption reflects realistic attack scenarios where deployed models may be subject to reverse engineering or API probing.

3.2.2. Accuracy Degradation Analysis

Accuracy degradation patterns vary significantly between privacy-preserving techniques and across privacy parameter configurations. The analysis characterizes degradation curves as functions of privacy budget (ε) for differential privacy and aggregation frequency for federated learning. The privacy-accuracy Pareto frontier identifies optimal operating points that maximize utility for given privacy constraints. Points below the Pareto frontier represent suboptimal configurations that achieve neither superior privacy nor accuracy (As shown in Table 2).

Table 2: Experimental Configuration Parameters

Parameter	Differential Privacy	Federated Learning	Hybrid DP-FL
Privacy budget ε	{0.1, 0.5, 1.0, 2.0, 5.0}	N/A	{0.5, 1.0, 2.0}
Noise mechanism	Gaussian	N/A	Gaussian
Clipping norm C	{0.5, 1.0, 2.0}	N/A	1.0
Number of clients	N/A	{10, 50, 100, 500}	100
Local epochs	N/A	{1, 5, 10}	5
Aggregation rounds	N/A	{50, 100, 200}	100
Batch size	64	32	32
Learning rate	0.01	0.01	0.005

3.2.3. Computational Efficiency Metrics

Computational efficiency evaluation encompasses training time, communication overhead, and inference latency. In federated learning, communication cost is measured in transmitted bytes per training round, accounting for model compression and gradient sparsification. The efficiency ratio (ER) normalizes computational costs against achieved utility: ER = URR / (Training_time × max(Communication_cost, 1)), where Communication_cost is set to 1 for centralized training. Higher ER values indicate more favorable trade-offs between efficiency and utility (As shown in Figure 1).

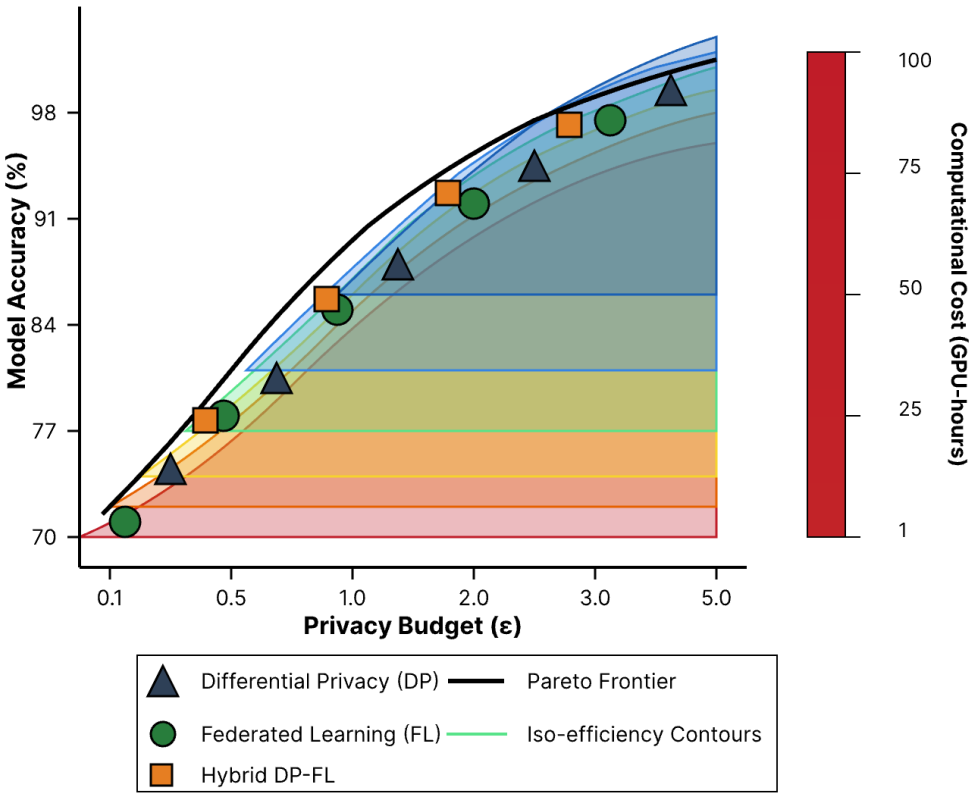

Figure 1: Privacy-Utility-Efficiency Three-Dimensional Trade-off Surface

This figure presents a two-dimensional heatmap (with contour lines) illustrating the trade-off among privacy budget (ε), model accuracy (%), and computational cost (GPU-hours), where color intensity represents computational cost. Marker points indicate experimental measurements for differential privacy, federated learning, and hybrid approaches. The Pareto-optimal region is indicated by the set of operating points that jointly optimize utility and cost under privacy constraints.

3.3. Fairness Assessment and Bias Detection Approaches

Fairness assessment addresses the potential for privacy-preserving mechanisms to disproportionately impact protected demographic groups. Privacy perturbation and data heterogeneity in federated settings may exacerbate existing biases or introduce new disparities in model performance across subpopulations. Algorithmic fairness in artificial intelligence applications requires careful consideration of how privacy mechanisms interact with protected attributes [15].

3.3.1. Fairness Metrics Framework

The evaluation incorporates group fairness metrics, including demographic parity difference (DemParDiff), equalized odds difference (EOD), and calibration error across groups (CEG). Individual fairness is assessed using consistency metrics that measure the similarity of predictions across similar individuals. Demographic parity requires equal positive prediction rates across groups: DemParDiff = $|P(Y=1|A=0) - P(Y=1|A=1)|$, where A denotes the protected attribute. Equalized odds extend this requirement to conditional acceptance rates given true labels.

3.3.2. Bias Detection Methodology

Bias detection employs stratified evaluation across customer demographic segments, including age groups, geographic regions, and service tier classifications. The analysis identifies heterogeneous privacy-utility trade-offs that may disadvantage minority subgroups with smaller representation in training data. The fairness-privacy interaction coefficient (FPIC) quantifies the correlation between privacy parameter settings and the degradation of the fairness metric. Positive FPIC values indicate that stronger privacy

protection is associated with improved fairness outcomes, whereas negative values suggest fairness deterioration under privacy constraints.

4. Results and Discussion

4.1. Comparative Performance Analysis of Privacy-Preserving Techniques

The experimental evaluation was conducted on three customer service datasets: a synthetic customer interaction dataset (N=500,000 records, 48 features), a public e-commerce transaction dataset (N=1,200,000 records, 35 features), and a telecommunications customer churn dataset (N=180,000 records, 52 features). All datasets underwent preprocessing, including normalization, categorical encoding, and stratified train-test splitting with 80-20 ratios.

4.1.1. Privacy Robustness Results

Differential privacy mechanisms demonstrated superior resistance to membership inference attacks compared to federated learning baselines. At ε = 1.0, the DP-SGD implementation achieved an MIA-SR of 0.523, approaching the random-guessing threshold of 0.5. Federated learning without additional privacy mechanisms exhibited MIA-SR of 0.671, indicating vulnerability to membership inference through gradient analysis.

The hybrid DP-FL approach achieved intermediate protection levels with MIA-SR of 0.548 at ε=1.0, representing an 18.3% improvement over standalone federated learning. Attribute inference accuracy followed similar patterns, with DP achieving 54.2% AIA compared to 68.7% for FL and 57.1% for DP-FL (As shown in Table 3).

Table 3: Privacy Robustness Comparison Results

Technique	MIA-SR	AIA (%)	MIRE	ε-empirical	Byzantine Threshold
DP (ε=0.5)	0.508	51.3	3.42	0.52	N/A
DP (ε=1.0)	0.523	54.2	2.89	1.08	N/A
DP (ε=2.0)	0.567	61.4	2.21	2.15	N/A
FL (100 clients)	0.671	68.7	1.54	N/A	0.20
FL (500 clients)	0.634	65.2	1.78	N/A	0.25
DP-FL (ε=1.0)	0.548	57.1	2.45	1.12	0.15
SecAgg-FL	0.589	62.3	1.92	N/A	0.30

4.1.2. Model Utility Results

Federated learning achieved higher utility retention compared to differential privacy across all evaluated tasks. On the customer churn prediction task, FL with 100 clients achieved 94.2% accuracy compared to 97.1% for the non-private baseline, yielding a URR of 0.970. Differential privacy at ε=1.0 achieved 85.3% accuracy (URR=0.879), representing a 12.3% accuracy degradation relative to FL.

The recommendation task exhibited more pronounced utility gaps. FL achieved nDCG@10 of 0.724 (URR=0.943) while DP at ε=1.0 achieved 0.612 (URR=0.797). The sensitivity of ranking metrics to perturbation noise accounts for this differential, as small score variations can substantially alter recommendation orderings (As shown in Figure 2).

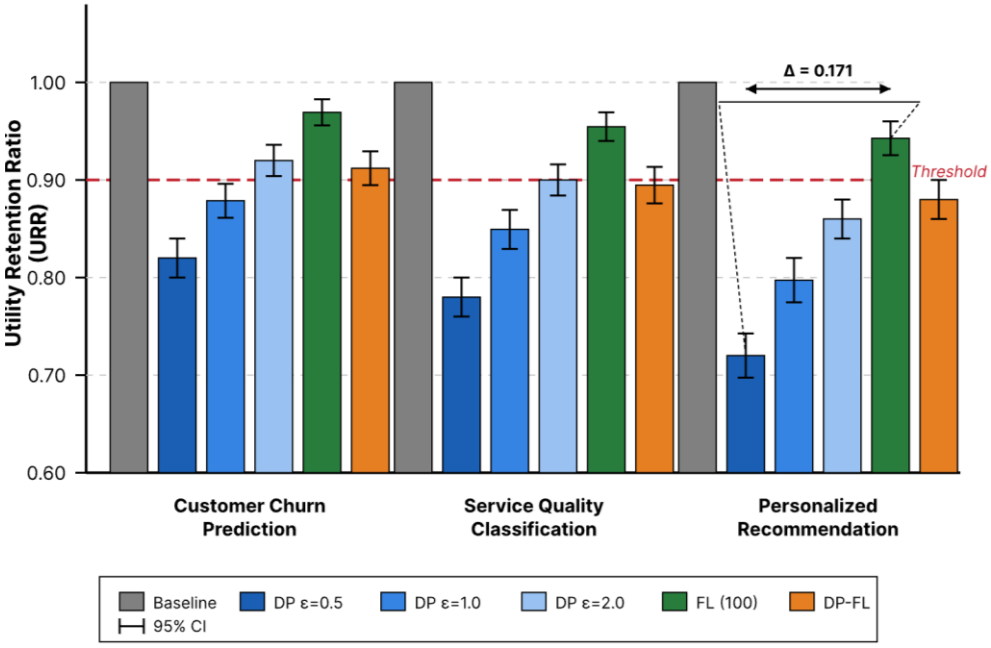

Figure 2: Utility Retention Ratio Across Privacy Configurations

This figure displays a grouped bar chart comparing utility retention ratios (URR) across different privacy-preserving techniques and parameter configurations. The x-axis categories represent three evaluation tasks: Customer Churn Prediction, Service Quality Classification, and Personalized Recommendation. The y-axis shows URR values from 0.6 to 1.0. Each task category contains six bars representing: Non-private baseline (URR=1.0, gray), DP ε=0.5 (dark blue), DP ε=1.0 (medium blue), DP ε=2.0 (light blue), FL 100 clients (dark green), and DP-FL hybrid (orange). Error bars indicate variability across repeated runs with different random train-test splits. A horizontal dashed line at URR = 0.9 marks the practical deployment threshold. Annotations highlight the maximum URR gap of 0.171 between FL and DP ε=0.5 on the recommendation task.

4.1.3. Computational Efficiency Results

Computational overhead analysis reveals significant differences between privacy-preserving paradigms. Differential privacy introduces minimal additional training time (approximately 8-15% overhead for gradient clipping and noise addition). Federated learning incurs substantial communication costs, with total transmitted data ranging from 2.4 GB to 18.7 GB depending on model size and aggregation frequency. Training time for FL with 100 clients averaged 4.2 hours compared to 1.8 hours for centralized DP training on equivalent hardware. Communication-efficient variants employing gradient compression reduced transmission overhead by 67% while maintaining accuracy below 1.5%.

4.2. Trade-off Analysis: Data Protection vs. Service Quality

The privacy-utility trade-off analysis reveals non-linear relationships between privacy parameters and model performance. Differential privacy exhibits diminishing returns in utility as privacy budgets increase beyond ε=2.0, with accuracy improvements below 2% for ε values between 2.0 and 5.0. This plateau effect suggests practical upper bounds on achievable utility under differential privacy constraints.

4.2.1. Pareto Frontier Analysis

The Pareto frontier characterization identifies optimal operating configurations for different organizational priorities. Privacy-prioritizing configurations ($\varepsilon \leq 0.5$ or FL with DP-SGD) achieve MIA-SR below 0.52 but sacrifice 15-25% utility. Utility-prioritizing configurations ($\varepsilon \geq 2.0$ or standard FL) maintain URR above 0.95 but exhibit MIA-SR exceeding 0.60.

The efficiency-normalized Pareto frontier incorporating computational costs shifts optimal configurations toward hybrid approaches. DP-FL with moderate privacy budgets (ε=1.0) achieves favorable three-way trade-offs: MIA-SR of 0.548, URR of 0.912, and training efficiency 2.3x superior to communication-heavy FL variants (As shown in Table 4).

Table 4: Pareto-Optimal Configuration Analysis

Configuration	MIA-SR	URR	Training Time (h)	Communication (GB)	Efficiency Score
DP ε=0.5	0.508	0.823	1.9	0	0.433
DP ε=1.0	0.523	0.879	1.8	0	0.488
FL-100	0.671	0.970	4.2	8.4	0.275
FL-500	0.634	0.958	7.8	18.7	0.131
DP-FL ε=1.0	0.548	0.912	3.1	5.2	0.362
SecAgg-FL	0.589	0.945	5.6	12.3	0.214

4.2.2. Task-Specific Trade-off Patterns

Trade-off characteristics vary substantially across customer service tasks. Classification tasks exhibit gradual accuracy degradation with increasing privacy protection, following approximately linear relationships within the ε range of 0.5 to 2.0. The degradation slope averages 4.2 percentage points per unit decrease in ε.

Recommendation and ranking tasks demonstrate threshold effects, with sharp performance drops occurring below critical privacy budget levels. The critical threshold for maintaining nDCG@10 above 0.7 is ε≈0.8 for differential privacy implementations. Below this threshold, ranking quality degrades rapidly due to noise-induced score inversions (As shown in Figure 3).

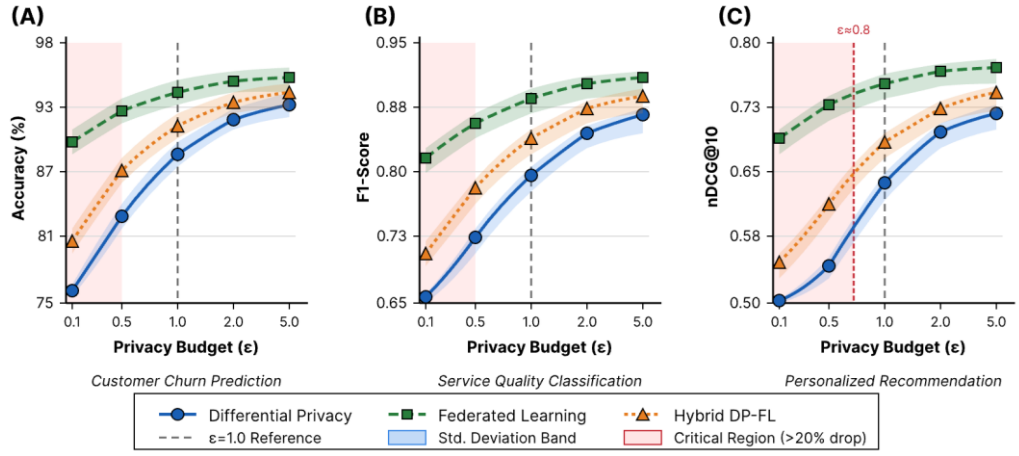

Figure 3: Task-Specific Privacy-Utility Trade-off Curves

This figure presents a multi-panel line plot showing privacy-utility trade-off curves for three customer service tasks. Panel A (left) displays the Customer Churn Prediction task with accuracy (%) on the y-axis ranging from 75% to 98%. Panel B (center) shows Service Quality Classification with F1-score on the y-axis from 0.65 to 0.95. Panel C (right) presents Personalized Recommendation with nDCG@10 from 0.50 to 0.80. All panels share a common x-axis, with privacy budget ε ranging from 0.1 to 5.0 on a logarithmic scale. Each panel contains three curves: differential privacy (solid blue line with circle markers), federated learning privacy equivalent (dashed green line with square markers), and hybrid DP-FL (dotted orange line with triangle markers). Shaded regions around each curve indicate standard deviation bands from repeated experiments. Vertical dashed lines

mark the ε = 1.0 reference point commonly used in practical deployments. The critical threshold regions where performance drops by more than 20% relative to baseline are highlighted in red.

4.3. Practical Implications and Recommendations for Enterprise Adoption

The empirical findings provide actionable guidance for enterprise decision-making regarding privacy-preserving technique selection. Organizational factors, including data infrastructure, regulatory requirements, and service quality tolerances, determine optimal approach selection.

4.3.1. Infrastructure-Based Recommendations

Organizations with centralized data architectures benefit from differential privacy implementations that leverage existing data pipelines without requiring modifications to distributed infrastructure. The computational overhead of DP (8-15%) is substantially lower than FL deployment costs, which include client coordination, secure aggregation infrastructure, and failure recovery mechanisms.

Enterprises with inherently distributed data sources, such as multi-regional operations or franchise networks, find federated learning architectures more naturally aligned with existing data governance structures. FL eliminates cross-border data transfer requirements that may conflict with data localization regulations (As shown in Table 5).

Table 5: Enterprise Adoption Decision Matrix

Organizational Factor	Differential Privacy Recommended	Federated Learning Recommended
Data infrastructure	Centralized	Distributed
Privacy requirement	Strong mathematical guarantees	Data localization compliance
Utility tolerance	>15% degradation acceptable	<10% degradation required
Computational budget	Limited	Flexible
Regulatory context	GDPR Article 22 compliance	Cross-border data restrictions
Model update frequency	Batch retraining	Continuous learning
Client device capability	N/A	Sufficient compute resources

4.3.2. Regulatory Compliance Mapping

The selection of privacy-preserving techniques must align with applicable regulatory frameworks. Differential privacy provides quantifiable privacy guarantees that map directly to regulatory concepts of data minimization and purpose limitation. The mathematical definition of ε-differential privacy enables compliance documentation and audit trail generation.

Federated learning addresses data localization and sovereignty requirements by maintaining data residency within jurisdictional boundaries. The combination of FL with secure aggregation protocols provides defense-in-depth against both external attacks and insider threats.

4.3.3. Hybrid Deployment Strategies

The evaluation results support hybrid deployment strategies that combine differential privacy and federated learning for enhanced protection. A staged implementation approach begins with federated learning for initial model development, followed by differential privacy application during model refinement and periodic retraining cycles.

The hybrid DP-FL configuration, achieving MIA-SR of 0.548 and URR of 0.912, represents a practical compromise for enterprises that require both strong privacy

guarantees and competitive service quality. Improvements in communication efficiency through gradient compression further enhance the viability of hybrid deployments.

5. Conclusion

5.1. Summary of Key Findings and Contributions

This research has presented a comprehensive comparative evaluation of privacy-preserving techniques for AI-driven customer service applications. The multi-dimensional evaluation framework, incorporating privacy robustness, model utility, computational efficiency, and algorithmic fairness metrics, provides a systematic approach for technique assessment and selection.

The empirical analysis demonstrates that federated learning achieves superior utility retention (URR=0.970) compared to differential privacy (URR=0.879 at ε=1.0) on customer service tasks. Differential privacy provides stronger protection against membership inference attacks (MIA-SR=0.523 versus 0.671 for FL). Hybrid DP-FL approaches offer favorable trade-offs with MIA-SR of 0.548 and URR of 0.912.

The task-specific trade-off analysis reveals that classification tasks tolerate privacy constraints more gracefully than ranking and recommendation tasks. Critical privacy budget thresholds exist below which service quality degrades substantially, with $\varepsilon \approx 0.8$ representing the practical lower bound for recommendation applications.

5.2. Limitations and Challenges

Several limitations constrain the generalizability of the presented findings. The evaluation datasets, while diverse, do not encompass all customer service application domains. Industry-specific data characteristics and regulatory requirements may alter the optimal selection of techniques.

The threat model's assumptions about adversarial capabilities are conservative estimates. Advanced attacks exploiting model update patterns or auxiliary information may achieve higher inference accuracy than measured in controlled experiments. The computational efficiency analysis does not account for infrastructure heterogeneity in production deployments.

The fairness assessment focused on binary protected attributes, whereas real-world demographic intersectionality involves multiple correlated characteristics. Extended fairness analysis incorporating intersectional subgroups remains an area for continued investigation.

5.3. Future Research Directions

Future research directions include developing adaptive privacy mechanisms that dynamically adjust protection levels based on data sensitivity and query patterns. Automated privacy budget allocation across multiple queries and model updates represents an open challenge with significant practical implications.

The integration of privacy-preserving techniques with emerging foundation models and large language model architectures warrants investigation. Transfer learning approaches that leverage pre-trained models while maintaining privacy guarantees could reduce the utility degradation associated with training from scratch.

Standardized benchmarking protocols and publicly available evaluation datasets would accelerate progress in privacy-preserving machine learning for customer service applications. Collaborative efforts between academic researchers and industry practitioners are essential for translating theoretical advances into deployable solutions that protect consumer privacy while delivering valuable services.

References

1. V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," Future Generation Computer Systems, vol. 115, pp. 619-640, 2021. https://doi.org/10.1016/j.future.2020.10.007
2. P. Kairouz et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, no. 1-2, pp. 1-210, 2021. https://doi.org/10.1561/2200000083

3. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, 2020. https://doi.org/10.1109/MSP.2020.2975749
4. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3-4, pp. 211-407, 2014. https://doi.org/10.1561/0400000042
5. M. Abadi et al., "Deep learning with differential privacy," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308-318, 2016. https://doi.org/10.1145/2976749.2978318
6. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273-1282, 2017.
7. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1-19, 2019. https://doi.org/10.1145/3298981
8. P. Radanliev, "AI ethics: Integrating transparency, fairness, and privacy in AI development," Applied Artificial Intelligence, vol. 39, no. 1, p. 2463722, 2025. https://doi.org/10.1080/08839514.2025.2463722
9. National Institute of Standards and Technology, "Guidelines for evaluating differential privacy guarantees (NIST Special Publication 800-226)," U.S. Department of Commerce, 2024. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-226.pdf
10. D. Pessach and E. Shmueli, "A review on fairness in machine learning," ACM Computing Surveys, vol. 55, no. 3, pp. 1-44, 2023. https://doi.org/10.1145/3494672
11. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1-35, 2021. https://doi.org/10.1145/3457607
12. G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, "VerifyNet: Secure and verifiable federated learning," IEEE Transactions on Information Forensics and Security, vol. 15, pp. 911-926, 2020. https://doi.org/10.1109/TIFS.2019.2929409
13. J. Zhao et al., "PVD-FL: A privacy-preserving and verifiable decentralized federated learning framework," IEEE Transactions on Information Forensics and Security, vol. 17, pp. 2059-2073, 2022. https://doi.org/10.1109/TIFS.2022.3176191
14. A. Fu et al., "VFL: A verifiable federated learning with privacy-preserving for big data in industrial IoT," IEEE Transactions on Industrial Informatics, vol. 18, no. 5, pp. 3316-3326, 2022. https://doi.org/10.1109/TII.2020.3036166
15. R. J. Chen et al., "Algorithmic fairness in artificial intelligence for medicine and healthcare," Nature Biomedical Engineering, vol. 7, no. 6, pp. 719-742, 2023. https://doi.org/10.1038/s41551-023-01056-8