Machine Intelligence & Responsible Systems
REVIEW | COMMISSIONED SYNTHESIS

Privacy-Preserving Adaptation of Large Language Models: An Evidence Framework

Kevin Adams - Corresponding author

Abstract

This review examines privacy-preserving adaptation of large language models. The organizing question is how adaptation utility should be balanced against memorization, inference attacks, and data governance. Ten related scholarly sources are synthesized through a decision-centered framework spanning problem definition, mechanism, measurement, evaluation, implementation, and governance. The review does not invent experiments, pooled estimates, or unreported quantitative results. It instead evaluates the strength and transferability of the available evidence, with particular attention to reporting a privacy mechanism without measuring the end-to-end exposure surface. The resulting framework links technical or empirical performance to explicit use conditions and identifies tests that should precede wider adoption in domain adaptation using sensitive organizational or personal data.

Keywords: large language models, differential privacy, federated learning, fine-tuning, governance

Synthesis lens	Principal question
Mechanism	the data path from collection through parameter updates, retrieval stores, logging, and model release
Evaluation	privacy accounting, attack-based testing, utility by subgroup, and repeated-composition effects
Primary risk	reporting a privacy mechanism without measuring the end-to-end exposure surface

1. Introduction

Research on privacy-preserving adaptation of large language models sits at the boundary between a promising component and a consequential decision. The core question is how adaptation utility should be balanced against memorization, inference attacks, and data governance. Answering it requires more than assembling favorable results. It requires a chain of evidence that makes the problem boundary, mechanism, measurement, comparator, uncertainty, and accountable decision visible to the reader.

This article presents a structured narrative review of ten related publications. Sources were selected for topical relevance, traceable publication metadata, and complementary coverage of technical, empirical, and governance questions. No meta-analytic pooling is attempted because the included studies differ in designs, populations, system boundaries, and outcomes. The purpose is decision-oriented synthesis: to identify what each source can support, where transfer remains uncertain, and which evidence would justify the next stage of use.

The central risk is reporting a privacy mechanism without measuring the end-to-end exposure surface. A top-line metric or broad policy label can appear decisive while concealing the conditions that produced it. Accordingly, the synthesis separates observation from interpretation and implementation from validation. This separation is especially important when the same system creates benefits for one user or institution while transferring cost, risk, or administrative burden to another.

2. Evidence Base and Problem Boundary

A defensible review begins by defining the unit of analysis. For the present topic, the relevant boundary includes inputs, institutional or technical context, the mechanism producing an output, the person or system acting on that output, and the downstream outcome. Evidence falls outside the boundary when it measures only an intermediate proxy yet is used to claim an end-to-end benefit.

Yin et al. (2021) addresses "A Comprehensive Survey of Privacy-preserving Federated Learning." In this synthesis, that work is used as a source on problem definition within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Liu et al. (2021) addresses "When Machine Learning Meets Privacy." In this synthesis, that work is used as a source on conceptual boundary within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Ye et al. (2023) addresses "Heterogeneous Federated Learning: State-of-the-art and Research Challenges." In this synthesis, that work is used as a source on measurement within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Together, these works provide complementary entry points, but their conclusions should not be flattened into a single ranking. Differences in data generation, setting, temporal horizon, and outcome definition may be substantively meaningful. A review should therefore preserve those differences and state where analogy, rather than direct replication, connects one domain to another.

3. Mechanism and System Design

The operative mechanism is the data path from collection through parameter updates, retrieval stores, logging, and model release. This formulation discourages component-only reasoning. A model, platform, policy, assay, or infrastructure service acquires practical meaning only when its interfaces and use conditions are specified. The evidence chain should describe what information is available, what transformation occurs, which assumptions make that transformation credible, and who is authorized to act.

Gosselin et al. (2022) addresses "Privacy and Security in Federated Learning: A Survey." In this synthesis, that work is used as a source on system mechanism within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Ferrag et al. (2024) addresses "Revolutionizing Cyber Threat Detection With Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices." In this synthesis, that work is used as a source on representation choice within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Wang et al. (2021) addresses "Fast-adapting and privacy-preserving federated recommender system." In this synthesis, that work is used as a source on failure analysis within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

These studies also show why modular claims require interface tests. A component may perform well while the complete pathway fails because inputs drift, users interpret outputs differently, recovery semantics are ambiguous, or institutional incentives change. Design documentation should therefore include not only the intended pathway but also foreseeable deviations, degraded modes, and conditions under which the system should stop or defer.

4. Evaluation and Uncertainty

Evaluation should center on privacy accounting, attack-based testing, utility by subgroup, and repeated-composition effects. Average performance remains useful, but it should be accompanied by distributions, error categories, relevant subgroups or operating regimes, and uncertainty at the point where a decision changes. Comparators must isolate the value of the proposed addition rather than compare a mature intervention with an intentionally weak baseline.

Hallaji et al. (2024) addresses "Decentralized Federated Learning: A Survey on Security and Privacy." In this synthesis, that work is used as a source on comparative evaluation within privacy-preserving adaptation of large language models. The connection is deliberately

bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Liu et al. (2020) addresses "Privacy and Security Issues in Deep Learning: A Survey." In this synthesis, that work is used as a source on uncertainty within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

External validation should introduce meaningful shifts in setting, time, population, workload, market structure, or environmental conditions. A nominally independent sample can still be too similar to test transfer. Conversely, a failed transfer is scientifically informative when the changed conditions are measured and the mechanism of failure is examined rather than hidden in an aggregate score.

5. Translation and Governance

Translation to domain adaptation using sensitive organizational or personal data depends on more than technical readiness. Decision rights, documentation, maintenance, monitoring, appeal, and recovery are part of the intervention. The governance requirement is purpose limitation, access controls, retention limits, and independent privacy review. Each control should be connected to a named failure mode and should identify an owner, evidence source, review interval, and escalation path.

Agrawal et al. (2022) addresses "Federated Learning for intrusion detection system: Concepts, challenges and future directions." In this synthesis, that work is used as a source on implementation within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Bouacida and Mohapatra (2021) addresses "Vulnerabilities in Federated Learning." In this synthesis, that work is used as a source on governance within privacy-preserving adaptation of large language models. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for domain adaptation using sensitive organizational or personal data. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

A credible implementation plan also defines sunset and revalidation conditions. New data, software updates, institutional restructuring, population change, or shifts in incentives can invalidate previously reasonable assumptions. Monitoring should therefore test the validity of the evidence chain, not merely whether a service remains online or a headline metric remains stable.

6. Research Agenda

Future studies should preregister or otherwise predefine the intended decision, principal outcomes, exclusions, and analysis plan when feasible. They should report missingness, sensitivity analyses, negative or null findings, and the cost of errors to differently situated stakeholders. Reproducible artifacts should include sufficient metadata to reconstruct data provenance, model or analytical versions, parameter choices, and the sequence from evidence to conclusion.

Cross-disciplinary work needs a shared object of explanation rather than a collection of adjacent vocabularies. For privacy-preserving adaptation of large language models, that shared object is the complete decision pathway. Technical, clinical, social, environmental, or economic expertise should each change the design or interpretation of the study. When evidence remains incomplete, the useful output is a bounded hypothesis, a transparent uncertainty statement, or a request for additional measurement.

7. Conclusion

The literature supports a disciplined approach to privacy-preserving adaptation of large language models: define the decision, preserve system boundaries, test consequential failure modes, connect uncertainty to action, and assign accountable control. The strongest conclusion is not that one method is universally superior. It is that credible use in domain adaptation using sensitive organizational or personal data requires an auditable link from source evidence to design choice, evaluation result, and governed decision.

References

Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Alazab, M., Bhattacharya, S., Maddikunta, P. K. R., & Gadekallu, T. R. (2022). Federated Learning for intrusion detection system: Concepts, challenges and future directions. Computer Communications, 195, 346-361. https://doi.org/10.1016/j.comcom.2022.09.012

Bouacida, N., & Mohapatra, P. (2021). Vulnerabilities in Federated Learning. IEEE Access, 9, 63229-63249. https://doi.org/10.1109/access.2021.3075203

Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro, L. C., Debbah, M., Lestable, T., & Thandi, N. S. (2024). Revolutionizing Cyber Threat Detection With Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices. IEEE Access, 12, 23733-23750. https://doi.org/10.1109/access.2024.3363469

Gosselin, R., Vieu, L., Loukil, F., & Benoit, A. (2022). Privacy and Security in Federated Learning: A Survey. Applied Sciences, 12(19), 9901. https://doi.org/10.3390/app12199901

Hallaji, E., Razavi-Far, R., Saif, M., Wang, B., & Yang, Q. (2024). Decentralized Federated Learning: A Survey on Security and Privacy. IEEE Transactions on Big Data, 10(2), 194-213. https://doi.org/10.1109/tbdata.2024.3362191

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When Machine Learning Meets Privacy. ACM Computing Surveys, 54(2), 1-36. https://doi.org/10.1145/3436755

Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., & Vasilakos, A. V. (2020). Privacy and Security Issues in Deep Learning: A Survey. IEEE Access, 9, 4566-4593. https://doi.org/10.1109/access.2020.3045078

Wang, Q., Yin, H., Chen, T., Yu, J., Zhou, A., & Zhang, X. (2021). Fast-adapting and privacy-preserving federated recommender system. The VLDB Journal, 31(5), 877-896. https://doi.org/10.1007/s00778-021-00700-6

Ye, M., Fang, X., Du, B., Yuen, P. C., & Tao, D. (2023). Heterogeneous Federated Learning: State-of-the-art and Research Challenges. ACM Computing Surveys, 56(3), 1-44. https://doi.org/10.1145/3625558

Yin, X., Zhu, Y., & Hu, J. (2021). A Comprehensive Survey of Privacy-preserving Federated Learning. ACM Computing Surveys, 54(6), 1-36. https://doi.org/10.1145/3460427