Sarcouncil Journal of Multidisciplinary

ISSN(Online): 2945-3445

Volume- 05| Issue- 11| 2025

Review Article **Received:** 20-10-2025 | **Accepted:** 14-11-2025 | **Published:** 27-11-2025

LLM on Private Data Using a Combination of LLM, Fine-tuning, Orchestration, and Tools

Siddhant Sonkar

University of California, Irvine, USA

Abstract: The use of Large Language Models on private organizational data marks an important advancement in enterprise artificial intelligence and is focused on the central challenge of using large language processing capabilities while keeping data safe and staying compliant. This in-depth review examines the multi-dimensional technical landscape for the enterprise use of foundation models, particularly architecture adaptations, privacy-preserving solutions, fine-tuning methods, and orchestration strategies. Foundation model architectures have evolved to safely accommodate private data use cases, primarily because of improvements to the attention mechanism, normalization methods, and context management properties. Retrieval-Augmented Generation architectures allow organizations to produce model outputs that are grounded in private/organizational knowledge bases while avoiding encoding private information in the model's parameters. Specific fine-tuning methods significantly lower compute requirements for downstream applications compared to retraining the model entirely. Agent-based orchestration frameworks create pathways for non-linear text generation, along with augmentation with tools, multi-step processes, and coordination with work streams. Privacy-preserving methods also represent a highly important consideration and lead to the need for differentially private model training methods, federated learning architectures, and hardware-based trusted execution environments. Access control, data use tracing, and full governance frameworks provide oversight into whether the model use is appropriate or lawful. Deployment strategies must consider competing priorities such as performance, reliability, cost, and security, whether the deployment is cloud-based, on-premises, or hybrid.

Keywords: Large Language Models, Private Data Processing, Fine-tuning Methodologies, Retrieval-Augmented Generation, Privacy-Preserving Machine Learning.

INTRODUCTION

A new deployment method of Large Language Models on private data organizations represents a new generation of enterprise artificial intelligence that utilizes complex model architectures while upholding a variety of data governance requirements. Implementations typically rely on pre-trained models from publicly available training corpora; however, the enterprise environment requires a model to be able to work with proprietary data without sacrificing confidentiality, compliance, or competitive advantage. The integration of fine-tuning methodologies, orchestration frameworks, and specialized tooling has emerged as the predominant approach for adapting foundation models to private data contexts without compromising sensitive information. Organizations across healthcare, telecommunications, financial services, and e-commerce sectors are increasingly adopting these technologies to leverage extensive proprietary knowledge bases while maintaining strict data governance protocols that comply with regulations, including GDPR, HIPAA, and industry-specific compliance frameworks (Abadi, M. *et al.,* 2016).

Contemporary architectures demonstrate remarkable capabilities across diverse natural language processing tasks, achieving performance metrics that approach or exceed human-level competency in specific domains. Advanced models exhibit strong accuracy rates on standardized reasoning benchmarks such as the Massive Multitask Language Understanding evaluation, while domain-specific implementations have demonstrated substantial accuracy improvements when fine-tuned on proprietary datasets containing labeled examples. Recent implementations in healthcare technology have shown enhanced diagnostic suggestion accuracy when trained on institutional clinical databases, compared to baseline performance using general-purpose models. In telecommunications network optimization contexts, fine-tuned implementations have achieved superior fault prediction accuracy, representing considerable improvement over traditional machine learning approaches. Yet, implementing these models with confidential enterprise data presents several challenges, both technical and cultural, that relate to implementation, data security, compliance, and scale.

The technical landscape for private data implementations encompasses several complementary approaches that address distinct use case requirements and constraint profiles. Retrieval-Augmented Generation architectures

***Corresponding Author:** Siddhant Sonkar

DOI- https://doi.org/10.5281/zenodo.17741274

enable models to access private knowledge bases without encoding sensitive information directly into model parameters, achieving efficient context retrieval latencies for enterprise-scale document repositories containing extensive indexed passages. Vector embedding models transform organizational documents into high-dimensional representations, enabling semantic similarity search across massive corpora with rapid query response times. Fine-tuning strategies adapt pre-trained foundation models to specific organizational vocabularies and workflows through continued training on curated private datasets, with training durations manageable on modern GPU infrastructure consisting of multiple accelerators per training cluster (Parthasarathy, V. B. *et al.*, 2024).

Privacy-preserving techniques have become essential components, with implementations employing differential privacy mechanisms that introduce calibrated noise into training processes while achieving appropriate epsilon values depending on privacy-utility trade-off requirements. These implementations maintain acceptable model accuracy degradation compared to non-private baselines, ensuring practical utility while providing mathematical privacy guarantees for sensitive organizational data.

LLM ARCHITECTURES AND PRIVATE DATA INTEGRATION STRATEGIES

Foundation Model Architectures for Enterprise Deployment

Foundation models designed for enterprise private data applications represent significant evolutions from their public-domain predecessors, incorporating architectural modifications that enhance security, controllability, and domain adaptability. Transformer-based architectures remain the dominant paradigm, with contemporary enterprise-focused implementations employing multiple attention layers and substantial embedding dimensions depending on intended application scope and computational budget constraints. These models achieve competitive perplexity scores on domain-specific evaluation benchmarks, indicating strong language modeling capabilities within specialized vocabulary contexts that include medical terminology, legal nomenclature, financial instruments, and technical specifications. The parameter count for enterprise-deployed models varies considerably, with organizations increasingly adopting mid-scale models that balance capability requirements against inference cost structures and latency constraints (Huang, S. *et al.*, 2024).

Architectural innovations specifically targeting private data scenarios include enhanced attention mechanisms with selective information flow controls, enabling fine-grained management of which knowledge sources influence generation processes. Multi-head attention configurations typically employ numerous parallel attention heads, with each head specializing in distinct syntactic, semantic, or contextual pattern recognition tasks. Layer normalization techniques have evolved beyond standard approaches, with recent implementations incorporating adaptive normalization strategies that adjust based on input data distributions, reducing domain shift effects when models encounter private data with statistical properties divergent from public training corpora. Position encoding mechanisms have been augmented with relative position representations capable of handling extended document sequences, addressing enterprise requirements for processing lengthy contracts, technical documentation, and comprehensive reports within single inference passes.

Retrieval-Augmented Generation for Private Knowledge Bases

Retrieval-Augmented Generation architectures provide mechanisms for grounding responses in private organizational knowledge without requiring comprehensive fine-tuning of model parameters, offering advantages in terms of knowledge update agility, interpretability, and reduced training overhead. Contemporary implementations segment private document corpora into manageable chunks, balancing retrieval granularity against context window utilization efficiency. Vector embedding models transform these chunks into high-dimensional representations, enabling semantic similarity computations that identify relevant knowledge sources for specific queries. Advanced embedding architectures achieve strong cosine similarity scores for semantically related content, with competitive retrieval precision rates for top candidates on enterprise evaluation benchmarks (Maity, S., & Saikia, M. J. 2025).

The indexing infrastructure for enterprise-scale systems involves sophisticated vector databases capable of managing substantial embedded document segments while supporting concurrent query loads. Hybrid retrieval strategies combine vector-based semantic search with traditional keyword-based retrieval mechanisms, addressing

Publisher: SARC Publisher

limitations in pure neural approaches for queries involving specific terminology, identifiers, or rare entities. These hybrid systems demonstrate precision improvements compared to vector-only approaches on enterprise benchmarks involving technical documentation and specialized domain knowledge.

Prompt Engineering and Context Management

Effective prompt engineering represents a critical determinant of private data performance, with prompt structure, content organization, and instruction clarity significantly influencing response quality, hallucination rates, and task completion accuracy. Systematic prompt engineering methodologies have demonstrated substantial performance improvements on complex enterprise tasks compared to naive query formulations. Context window management poses significant challenges for enterprise applications requiring the synthesis of information across extensive document collections, with contemporary models supporting varying context window sizes depending on architecture and optimization priorities.

Table 1: Foundation Model Architecture Components for Enterprise Private Data Applications (Huang, S. *et al.*, 2024; Maity, S., & Saikia, M. J. 2025)

Architecture Component	Key Characteristics	Enterprise Applications
Multi-head Attention Mechanisms	Parallel attention heads specializing in syntactic, semantic, and contextual pattern recognition	Medical terminology processing, legal document analysis, and financial instrument classification
Adaptive Layer Normalization	Dynamic adjustment based on input data distributions to reduce domain shift effects	Cross-domain knowledge transfer, specialized vocabulary handling, proprietary data processing
Extended Position Encoding	Relative position representations for lengthy document sequences	Contract analysis, technical documentation processing, and comprehensive report generation
Vector Embedding Models	High-dimensional representations enabling semantic similarity computations	Enterprise knowledge base search, document retrieval, contextual information identification

FINE-TUNING METHODOLOGIES FOR DOMAIN ADAPTATION

Full Model Fine-tuning Approaches

Full model fine-tuning involves continued training of all model parameters on domain-specific private datasets, enabling comprehensive adaptation to specialized vocabularies, reasoning patterns, and knowledge requirements characteristic of enterprise contexts. This approach typically requires substantial computational resources, with training costs varying significantly depending on model size and dataset magnitude. Training datasets for effective enterprise fine-tuning generally contain substantial collections of high-quality examples spanning representative task distributions, with data curation and annotation efforts frequently representing the majority of total project costs and timelines. The training process typically involves multiple epochs over the complete dataset, with learning rates carefully scheduled to balance adaptation speed against catastrophic forgetting risks. Enterprise implementations in healthcare domains have demonstrated significant improvements in clinical terminology recognition accuracy, while financial services applications have achieved enhanced performance in regulatory compliance document analysis and risk assessment tasks (Oyamada, R. S. *et al.*, 2025).

Hardware requirements for full model fine-tuning scale substantially with model size, with smaller models typically requiring multiple GPUs with substantial memory per device, while larger models may require extensive GPU arrays utilizing tensor parallelism and pipeline parallelism strategies to distribute computation and memory loads. Gradient accumulation techniques enable fine-tuning on hardware configurations with limited memory capacity by computing gradients across multiple micro-batches before applying parameter updates, though this approach increases training time proportionally. Mixed-precision training utilizing reduced floating point representations reduces memory requirements substantially compared to full-precision training while maintaining model quality within acceptable tolerances, though numerical stability considerations require careful hyperparameter tuning.

Publisher: SARC Publisher

Catastrophic forgetting represents a significant challenge in full model fine-tuning, where adaptation to private domain data degrades model performance on general capabilities and knowledge domains. Regularization techniques, including elastic weight consolidation and knowledge distillation, help mitigate these effects by constraining parameter updates that would significantly alter important general capabilities. Evaluation protocols assess both in-domain performance on private data tasks and out-of-domain retention of general capabilities, with acceptable implementations typically demonstrating minimal degradation in general benchmarks while achieving substantial improvements on domain-specific metrics.

Parameter-Efficient Fine-tuning Techniques

Parameter-efficient fine-tuning methods address the computational and resource challenges of full model fine-tuning by adapting only small subsets of model parameters or introducing lightweight auxiliary components while keeping foundation model parameters frozen. Low-Rank Adaptation represents the most widely adopted approach, introducing trainable rank decomposition matrices into attention layers that enable adaptation using a minimal fraction of the parameter count required for full fine-tuning. Implementations typically achieve the majority of performance gains observed with full fine-tuning while reducing training costs by substantial factors and enabling multiple task-specific adaptations to coexist through adapter module swapping. Enterprise deployments have successfully utilized these techniques for domain adaptation in telecommunications network management, customer service automation, and technical documentation generation with significantly reduced infrastructure requirements (Software Mind, 2025).

Prefix tuning approaches prepend learnable continuous vectors to input sequences at each transformer layer, providing task-specific context that modulates model behavior without modifying foundation parameters. Adapter modules insert small feedforward networks between transformer layers, enabling task-specific adaptation while preserving foundation capabilities. These adapter architectures demonstrate particularly strong performance on tasks requiring integration of private domain knowledge while maintaining general reasoning capabilities.

Instruction Tuning and Alignment for Enterprise Context

Instruction tuning optimizes models to follow explicit directives and complete specified tasks based on natural language instructions, representing a critical adaptation layer for enterprise deployments where consistent adherence to organizational policies, formatting standards, and interaction protocols proves essential. Instruction tuning datasets for enterprise contexts typically contain substantial collections of carefully curated instruction-response pairs spanning anticipated task distributions, business processes, and interaction scenarios. Reinforcement learning from human feedback provides mechanisms for aligning model behavior with organizational values, preferences, and quality standards through iterative refinement based on human evaluations of model outputs.

Table 2: Fine-tuning Methodologies and Adaptation Strategies for Domain Specialization (Oyamada, R. S. *et al.*, 2025; Software Mind, 2025)

Fine-tuning Method	Technical Approach	Computational Efficiency
Full Model Fine-tuning	Continued training of all model parameters on domain-specific datasets	Requires substantial GPU arrays with tensor and pipeline parallelism for large models
Low-Rank Adaptation	Trainable rank decomposition matrices in attention layers with frozen foundation parameters	Minimal parameter fraction enables multiple task-specific adaptations through adapter swapping
Prefix Tuning	Learnable continuous vectors prepended to input sequences at each transformer layer	Task-specific context modulation without foundation parameter modification
Instruction Tuning	Optimization for the following explicit directives with curated instruction-response pairs	Critical adaptation layer for organizational policy adherence and formatting standards

ORCHESTRATION FRAMEWORKS AND TOOL INTEGRATION

Agent Architectures and Multi-Step Reasoning

Agent-based orchestration architectures enable models to decompose complex enterprise tasks into sequences of reasoning steps, tool invocations,

Publisher: SARC Publisher

and information gathering operations, substantially extending capabilities beyond single-pass text generation. Contemporary agent frameworks implement perceive-reason-act cycles where models analyze situations, formulate plans, execute actions through tool interfaces, and refine strategies based on observed outcomes. These iterative reasoning processes typically involve multiple discrete steps for moderately complex enterprise tasks, with execution times varying depending on computational complexity and external tool latency profiles. Advanced frameworks represent widely adopted agent architectures, achieving substantial success rates on complex multi-step reasoning benchmarks compared to direct prompting approaches. Enterprise deployments in supply chain optimization, financial analysis, and customer support automation have demonstrated substantial improvements in task completion rates and accuracy through iterative reasoning patterns that decompose complex queries into manageable subtasks (Sonkar, S. 2025).

Planning and decomposition strategies enable agents to break down high-level objectives into manageable subtasks amenable to individual tool invocations or focused reasoning operations. Hierarchical task networks provide formal frameworks for representing task decomposition structures, with typical enterprise workflows decomposing into multiple hierarchical levels of abstraction. Chain-of-thought prompting techniques encourage models to articulate intermediate reasoning steps explicitly, improving accuracy on complex reasoning tasks compared to direct answer generation. Self-consistency approaches generate multiple independent reasoning traces and aggregate results through voting or confidence-weighted combinations, reducing error rates on tasks involving numerical reasoning or multi-hop inference.

Tool-augmented agents extend language model capabilities by providing programmatic interfaces to databases, search engines, calculators, code interpreters, and specialized domain APIs. Typical enterprise agent implementations incorporate numerous distinct tool interfaces, each with formal specifications defining input parameters, output formats, and operational semantics. Function calling mechanisms enable models to generate structured tool invocation requests that orchestration frameworks parse, validate, execute, and integrate results back into agent reasoning contexts (Sonkar, S. 2025).

Workflow Orchestration and Pipeline Management

Enterprise applications often need to orchestrate many stages of processing between the parsing of documents, information extraction, reasoning operations, validation, and output generation. Thus, they need a workflow orchestration capability. Directed acyclic graphs provide a formal way to describe processing pipelines. Typically, enterprise applications adopt even simple workflows composed of several nodes representing processing steps, with directed edges encoding data dependencies and flow control. Parallel execution strategies take advantage of the functional independence of branches of the workflow to minimize end-to-end latency compared to sequences of processing activity, although resource contention and coordination cost must be managed carefully.

Error handling and reliability engineering become especially important for production enterprise deployments, as processing failures cascade through multi-stage work processes and produce disruptions to business operations. Human-in-the-loop workflows add a human verification and confirmation step at crucial decision points to navigate a course between the efficiency of automation and the need to monitor oversights when important decisions are being rendered in enterprise applications.

Tool Ecosystems and Integration Patterns

Comprehensive tool ecosystems for enterprise agents encompass diverse capabilities, including data access, computation, communication, and domain-specific operations. Database integration tools enable structured querying of relational, document, and graph databases containing private organizational data. Security and governance aspects for integrating tools include authentication, authorization, input validation, output sanitization, and audit logging.

Table 3: Agent Orchestration and Workflow Management Capabilities (Sonkar, S. 2025; Sonkar, S. 2025)

Orchestration Component	Functional Capabilities	Enterprise Implementation
Multi-Step Reasoning Agents	Perceive-reason-act cycles with iterative task decomposition and refinement	Supply chain optimization, financial analysis, and customer support automation

Publisher: SARC Publisher

Chain-of-Thought Prompting	Explicit articulation of intermediate reasoning steps for complex tasks	Multi-source information integration, decision audit trails, and validation processes
Tool-Augmented Agents	Programmatic interfaces to databases, calculators, code interpreters, and domain APIs	Structured data querying, quantitative analysis, specialized domain operations
Workflow Orchestration	Directed acyclic graph representations with parallel execution strategies	Document parsing, information extraction, validation checks, output formatting

SECURITY, PRIVACY, AND IMPLEMENTATION ASPECTS

Privacy-Preserving Techniques and Data Protection

Privacy preservation is an important consideration for organizations deploying models on sensitive private data, and technical approaches are employed singly or in combination to reduce the risk of privacy loss through training, inference, or operating lifecycles. Differential privacy protections add calibrated noise to model training procedures or query responses, and mathematical guarantees exist that ensure specific data points cannot be rebuilt or inferred based on model behaviors. Privacy budget parameters modulate the trade-off between privacy and model utility, and typical enterprise use-case implementations generally conform to various epsilon values that depend on the sensitivity of underlying data to regulatory requirements. It is possible that differential privacy implementations reduce model accuracy compared to baseline non-private implementations, depending on privacy budget, and model complexity parameters considered in specific use contexts require clear evaluation using privacy-utility trade-offs. More strict privacy budgets are often deployed for healthcare organizations processing patient records, and within financial institutions handling transaction process data, as part of regulatory privacy frameworks for sensitive information processing (Siddhant, S. 2025).

Federated learning architectures enable collaborative model training across distributed private datasets without requiring data centralization, addressing data sovereignty concerns and reducing risks associated with consolidated data breaches (Mintah, 2022). Federated learning protocols coordinate iterative training rounds where participating organizations compute local model updates on private data and share only aggregated gradient information with central coordination servers. Secure aggregation protocols employ cryptographic techniques, including secret sharing and homomorphic encryption, to prevent exposure of individual organization updates during aggregation processes. Federated learning systems have been deployed in multi-institution translational research collaborations, cross-border financial networks, and healthcare consortia to utilize collaborative knowledge and adhere to data locality requirements (Rai, 2023).

Secure enclaves and trusted execution environments provide hardware-based isolation of sensitive computations that can prevent unauthorized access to both data and models, even from system administrators and operating system components with privileged access. Technologies enable the creation of protected memory regions where code executes with cryptographic attestation of integrity and confidentiality. Financial services organizations processing payment card information and government agencies handling classified data increasingly leverage trusted execution environments to maintain air-gapped security guarantees while enabling advanced analytics capabilities (Sonkar, S. 2025).

Access Control and Governance Frameworks

Comprehensive access control frameworks for enterprise implementations account for authentication, authorization, audit logging, and policy enforcement methods to ensure access to model capabilities and underlying private data can be appropriately controlled. Role-based access control models assign permissions based on an organization's set of roles, wherein implementations naturally divide roles into user communities, administrative functions, and system integrations. For role-based access control models, attribute-based access control augments and extends the role-based model by applying more granular conditional policies to evaluate user attributes or characteristics, resource characteristics or context, environmental context, and request characteristics (Boadi-Mensah, 2023). Data lineage tracking and provenance management capabilities help organizations understand how and where private data flows through processing pipelines, to satisfy compliance requirements, and

Publisher: SARC Publisher

to gauge impact when data needs to be deleted or access changes.

Deployment Options and Operational Considerations

Production deployment options for enterprise contexts will weigh competing priorities such as performance, reliability, cost, and security across varying operational contexts. Deployments in the cloud leverage managed infrastructure services with the potential for elastic handling capacity, managed operations, and geographic distribution. Conversely, on-premises deployments provide maximum control over data management and security but typically involve significant upfront investment in hardware and the expertise to administer. Also consider hybrid architectures that can combine cloud and on-premises elements that process less-sensitive data in a cloud environment but maintain strict on-premises controls for very sensitive data (Fernandes, 2024).

Table 4: Privacy-Preserving Techniques and Security Frameworks for Sensitive Data Processing (Siddhant, S. 2025; Sonkar, S. 2025)

Security Mechanism	Protection Approach	Regulatory Compliance
Differential Privacy	Calibrated noise injection with mathematical guarantees against data inference	Healthcare patient records, financial transaction data, GDPR, and HIPAA compliance
Federated Learning	Collaborative training across distributed datasets without data centralization	Multi-institutional research, cross-border financial networks, and healthcare consortia
Trusted Execution Environments	Hardware-based isolation with cryptographic attestation of integrity	Payment card processing, classified government data, and air-gapped security requirements
Attribute-Based Access Control	Fine-grained conditional policies evaluating user attributes and contextual factors	Dynamic authorization decisions, role-based permissions, and compliance enforcement

CONCLUSION

Combining Large Language Models with private organizational data introduces an impactful ability for organizations to leverage advanced natural language processing with a commitment to governing sensitive data, privacy, and security. The technical solutions explored throughout this review demonstrate the maturity of methodologies available for adapting foundation models to proprietary contexts through diverse complementary techniques. Architectures for retrieval-augmented generation offer knowledge-grounded mechanisms that provide information currency, and this is done without necessitating model retraining, and parameter-efficient fine-tuning methods allow for domain adaptations with much lower compute costs than model retraining. The emergence of advanced orchestration frameworks and agent architectures enables models to do so much more than produce text—they are now capable of complex multi-step reasoning, integration with tools, and coordination of workflows, which are necessary to meet the reality of enterprise needs. The article on maintaining privacy noted thus far, including differential privacy, federated learning, and trusted execution environments, provides alternative frameworks and mathematical assurances for safeguarding sensitive data throughout the lifecycle of a model. Complete accountability, governance frameworks, and deployment solutions allow organizations to make trade-offs between performance, costs, security, and compliance across various operational contexts. Overall, with the continued maturation of these technologies, the usability of advanced language models can potentially be improved in evermore challenging and regulated environments and facilitate greater accessibility to advanced AI while also adhering to trust and security concerns in an enterprise setting. Ultimately, viable implementations will need to apply logical technical, operational, and organizational considerations for the greatest probability of success that is fit for purpose, aligned to meeting unique outcomes of the business, as well as regulatory frameworks.

REFERENCES

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. "Deep learning with differential privacy." *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. (2016).
2. Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. "The ultimate guide to fine-tuning

Publisher: SARC Publisher

llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities." *arXiv preprint arXiv:2408.13296* (2024).
3. Huang, S., Yang, K., Qi, S., & Wang, R. "When large language model meets optimization." *Swarm and Evolutionary Computation* 90 (2024): 101663.
4. Maity, S., & Saikia, M. J. "Large Language Models in Healthcare and Medical Applications: A Review." *Bioengineering* 12.6 (2025): 631.
5. Oyamada, R. S., Peeperkorn, J., De Weerdt, J., & De Smedt, J. "Domain Adaptation of LLMs for Process Data." *arXiv preprint arXiv:2509.03161* (2025).
6. Software Mind, "Parameter-efficient fine-tuning (PEFT): benefits and techniques," (2025).
7. Sonkar, S. "Fine-tuning AI Models for code generation: Advances and applications," *World Journal of Advanced Research and Reviews*, (2025).
8. Sonkar, S. "Reducing Seller Friction through Generative AI: An Analysis of E-commerce Innovation." *IJSAT-International Journal on Science and Technology* 16.1 (2025).
9. Siddhant, S. "Recent Innovations in AI Privacy: Protecting Data in the Age of Machine Learning." *International Journal* 11.2 (2025): 611-619.
10. Sonkar, S. "Boosting Developer Efficiency with AI: A Technical Review." *Journal Of Engineering And Computer Sciences* 4.9 (2025): 500-508.
11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.
12. Yang, Q., Liu, Y., Chen, T., & Tong, Y. "Federated machine learning: Concept and applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019): 1-19.
13. Mintah, P. A. "Asset-liability management practices and risk mitigation in banking systems." *Journal of Computational Analysis and Applications* 30 (2022): 835–850.
14. Rai, C. "Developing nutrient-enriched, health-forward bakery products: A case study on ube-infused and fermented-garlic sourdoughs." Evolutionary Studies in Imaginative Culture (2023): 97–103.
15. Boadi-Mensah, J. "The environmental impacts of poor waste management: A call for sustainable action." Sarcouncil Journal of Applied Sciences 3 (2023): 1–9.
16. Fernandes, N. "Counsellor decision-making in sex therapy within low-resource and multicultural contexts." *Evolutionary Studies in Imaginative Culture* 8 (2024): 1680–1686.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Sonkar, S. "LLM on Private Data Using a Combination of LLM, Fine-tuning, Orchestration, and Tools." *Sarcouncil Journal of Multidisciplinary 5.11* (2025): pp 118-125.

Publisher: SARC Publisher