Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 09| 2025

Research Article **Received:** 25-07-2025| **Accepted:** 15-08-2025 | **Published:** 12-09-2025

Enhancing Unstructured Document Text Extraction with LLMs in Cloud-Native Enterprise Solutions

Anupam Chansarkar

Amazon.com Services LLC, USA

Abstract: Text extraction from unstructured documents presents significant challenges for enterprises managing diverse content types across departments. Traditional methods struggle with variable formats, complex layouts, and contextual relationships within documents. The evolution from rigid rule-based systems to advanced language models marks a transformative shift in document processing capabilities. Cloud-native solutions leveraging Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) architectures offer superior performance by combining semantic understanding with precise context retrieval mechanisms. These architectures enable accurate information extraction from complex unstructured documents while maintaining enterprise-grade scalability. Implementation across domains like contract parsing and screenplay analysis demonstrates practical benefits in maintaining contextual awareness and semantic relationships that traditional methods cannot achieve. Ongoing challenges include multilingual support beyond major languages, handling degraded scanned documents, integration with existing systems, and optimizing performance in distributed environments.

Keywords: Unstructured Document Processing, Large Language Models, Retrieval-Augmented Generation, Vector Databases, Enterprise Document Management.

INTRODUCTION

Text extraction presents formidable challenges in enterprise environments where organizations must process diverse document types across departments. Most business documents lack standardized formats, making information extraction particularly difficult with conventional methods. The challenge intensifies with industry-specific terminology, varied document layouts, and multilingual content that requires contextual understanding rather than simple keyword matching. Enterprise document management systems must contend with this heterogeneity while maintaining processing speed and accuracy to deliver business value (Jaggavarapu, M. K. R. 2025).

The technological evolution of document processing reflects a journey from rigid rule-based approaches to sophisticated AI-driven solutions. Early systems employed template matching and regular expressions that required extensive manual configuration for each document type. As technology progressed, optical character recognition provided basic digitization capabilities but struggled with complex layouts. The field advanced through machine learning implementations that improved adaptability but still required significant feature engineering. Recent years have witnessed transformative change through deep learning models capable of understanding document context and adapting to variations without explicit programming. This progression marks a fundamental shift from brittle, rules-based systems toward intelligent, learning-based approaches (Aman, 2023).

Accurate text extraction serves as the foundation for critical business operations across enterprise workflows. Contract management depends on the precise extraction of terms and conditions to ensure compliance and optimize business relationships. Customer service effectiveness hinges on rapid access to relevant documentation during client interactions. Financial departments require accurate extraction of numerical data for accounting and planning. Legal teams rely on comprehensive document analysis for case preparation and compliance verification. The impact of improved extraction accuracy reverberates throughout organizations, affecting everything from strategic decision-making to operational efficiency (Jaggavarapu, M. K. R. 2025).

***Corresponding Author:** Anupam Chansarkar

DOI- https://doi.org/10.5281/zenodo.17105947

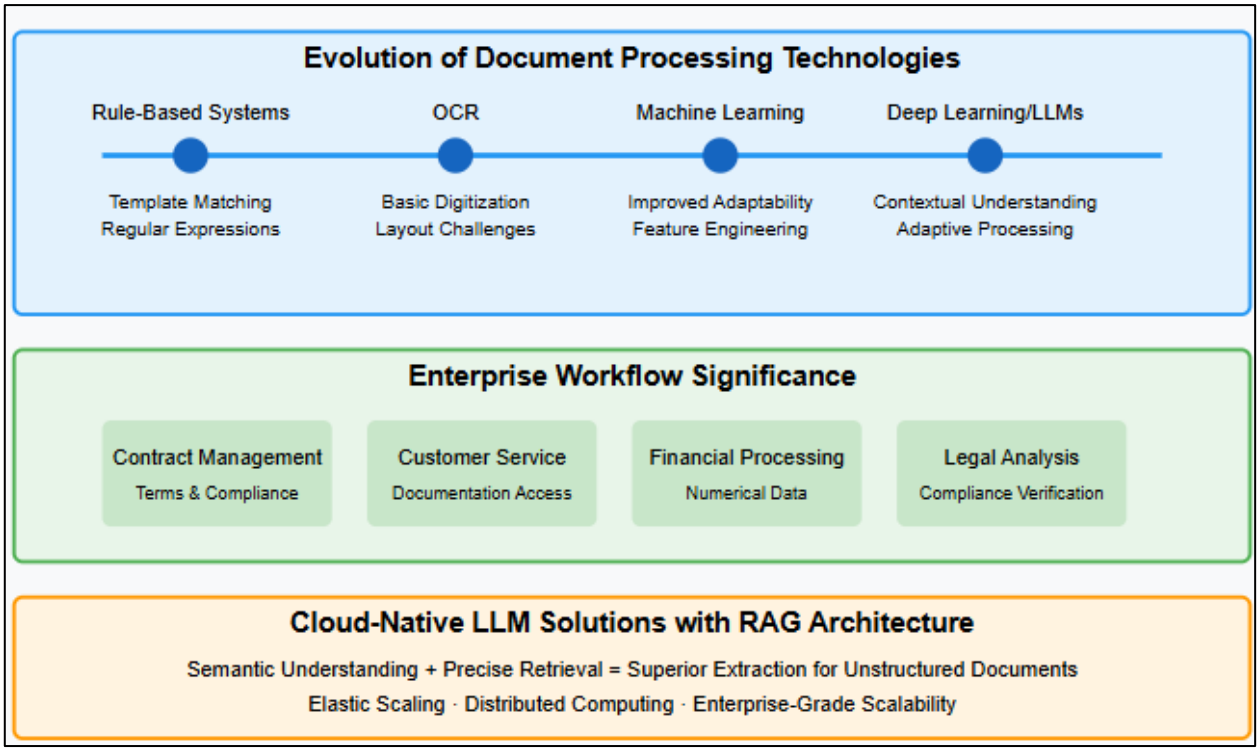

Fig 1: Evolution and Significance of Text Extraction in Enterprise Environments (Jaggavarapu, M. K. R. 2025; Aman, 2023).

Cloud-native architectures leveraging Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) represent the next frontier in document processing. These systems combine the semantic understanding capabilities of advanced language models with precise retrieval mechanisms that provide relevant context from document repositories. The integration enables accurate identification of key information even in highly unstructured content. When deployed in cloud environments, these solutions benefit from elastic scaling to accommodate processing demands and distributed computing for handling large document volumes. The architecture fundamentally addresses longstanding challenges in document extraction by merging the flexibility of modern language models with the precision of targeted information retrieval, all while maintaining enterprise-grade scalability (Aman, 2023).

CURRENT LANDSCAPE OF TEXT EXTRACTION TECHNOLOGIES

Traditional text extraction methods form the foundation of document processing systems, but face significant limitations with complex materials. Libraries like PyPDF2 and PDFMiner extract text directly from PDF object structures, performing adequately on digital-native documents with straightforward layouts. However, these approaches falter when confronted with multi-column formats, embedded graphics, or complex visual hierarchies. OCR technologies have evolved through neural network integration, improving baseline recognition for digitized documents, yet struggle with maintaining spatial relationships between textual elements. The fundamental limitation across traditional methods is their treatment of documents as collections of independent text fragments rather than semantically coherent information structures where meaning derives from both content and layout relationships (Mahadevkar, S. V. *et al.*, 2024).

Cloud provider solutions have transformed document processing through specialized AI services that combine OCR capabilities with advanced machine learning models. Major platforms offer document intelligence services with pre-built models for common document types while supporting custom training for organization-specific formats. These services implement multi-stage processing pipelines beginning with document normalization, continuing through text recognition, and culminating in semantic analysis to identify key fields and relationships. Cloud-based approaches integrate directly with complementary services, including storage, workflow automation, and analytics platforms,

Publisher: SARC Publisher

enabling comprehensive document processing solutions with reduced implementation complexity. The managed nature eliminates infrastructure concerns while providing scalability for variable workloads, though this convenience introduces usage-based pricing models requiring careful consideration for high-volume scenarios (Subhani, M. 2024).

Existing approaches demonstrate notable limitations when processing unstructured documents lacking consistent formatting. While current technologies perform adequately on forms and structured content, performance degrades substantially with documents like contracts, research papers, and technical manuals that combine narrative text with embedded tables and graphics. The challenge stems from the contextual nature of information, where meaning derives not only from text itself but from relationships to surrounding content. Most technologies struggle with maintaining hierarchical document structures, often flattening nested relationships essential for proper interpretation (Mahadevkar, S. V. *et al.*, 2024).

Performance characteristics vary significantly across extraction approaches, presenting tradeoffs in accuracy, speed, complexity, and cost. Open-source libraries offer flexibility but require substantial development resources for robust implementation. Self-hosted commercial solutions provide improved accuracy through specialized models but necessitate infrastructure management. Cloud-based services eliminate infrastructure concerns while offering state-of-the-art accuracy, though they introduce external dependencies and variable costs based on usage. Organizations increasingly implement hybrid approaches combining multiple technologies to address diverse document processing requirements across business functions (Subhani, M. 2024).

LLM-BASED TEXT EXTRACTION ARCHITECTURE FOR ENTERPRISE SCALE

Cloud-native Retrieval-Augmented Generation (RAG) systems form the foundation of modern enterprise text extraction solutions, addressing fundamental limitations in traditional approaches. The core architecture consists of interconnected components, including document ingestion pipelines, embedding services, vector storage systems, and orchestration layers. This approach enhances extraction capabilities by providing relevant contextual information during processing, enabling more accurate information identification even in complex unstructured documents. The RAG pattern demonstrates particular strength in maintaining semantic relationships across document boundaries, where conventional methods typically falter by treating sections as isolated units without preserving broader context (Sharma, C. 2025).

Vector database integration serves as the cornerstone for effective RAG implementations in document processing workflows. These specialized databases store and index high-dimensional vector representations of document fragments, enabling semantic search capabilities that transcend simple keyword matching. During extraction operations, relevant context is retrieved through similarity search rather than exact matching, allowing identification of conceptually related information even when terminology varies. Advanced implementations incorporate hybrid retrieval strategies combining dense vector search with sparse retrieval methods to maximize contextual relevance. The quality of retrieved context directly impacts extraction accuracy, as irrelevant information can actively degrade results by introducing competing noise (Sharma, C. 2025).

Distributed processing architectures enable enterprise-scale document handling through efficient resource allocation across available infrastructure. The processing model typically employs multi-stage pipelines where discrete operations occur in parallel across distributed nodes. Document ingestion, chunking, embedding generation, retrieval, and extraction operate as independent services communicating through message queues or APIs. This pattern enables both horizontal scaling for increased document volumes and specialized resource allocation based on workload characteristics. Edge computing approaches push initial document processing closer to data sources, reducing central infrastructure requirements while maintaining consistent extraction capabilities across geographically dispersed operations (Liang, W. *et al.*, 2025).

Publisher: SARC Publisher

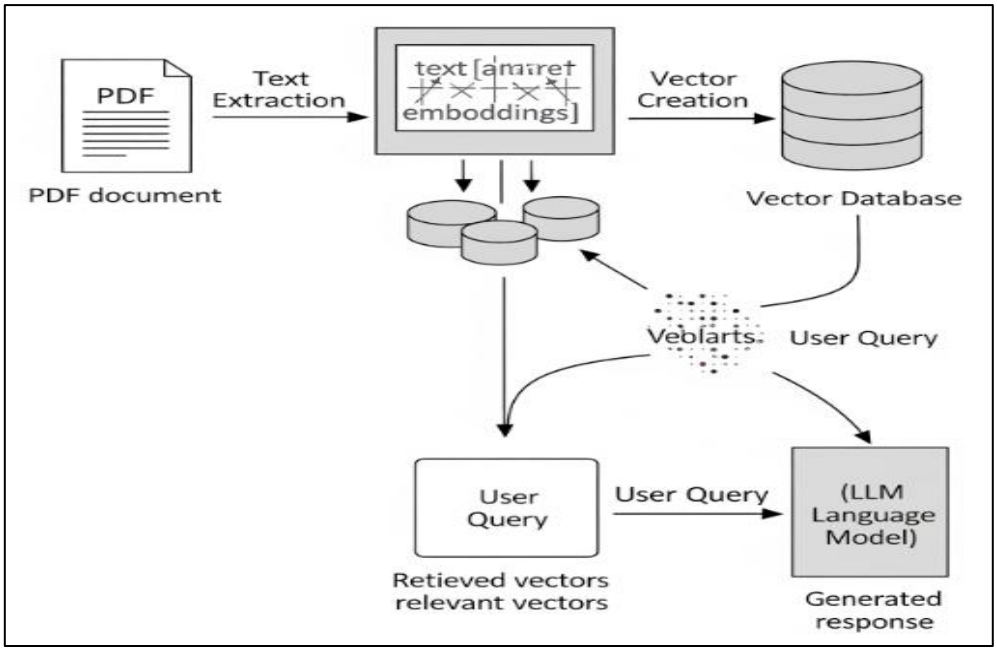

Fig 2: Retrieval-Augmented Generation (RAG) Architecture for PDF Document Processing (Sharma, C. 2025; Liang, W. *et al.*, 2025)

Enterprise deployment extends beyond technical architecture to encompass governance, security, and operational requirements. Successful implementations integrate with existing identity management systems for granular access control and implement differential security based on content sensitivity. Operational monitoring requires comprehensive observability with particular attention to drift detection in document characteristics and model performance. Integration with enterprise systems typically leverages event-driven architectures, maintaining loose coupling while ensuring consistent information flow. Resource management strategies must balance performance requirements against cost constraints, particularly for computationally intensive LLM operations requiring specialized hardware acceleration (Liang, W. *et al.*, 2025).

CASE STUDIES

Enterprise Applications

Unstructured contract parsing represents a significant application domain for advanced text extraction technologies in enterprise settings. Legal documents present unique challenges due to complex structure, specialized terminology, and business criticality. Successful implementations employ document segmentation that preserves logical structure rather than arbitrary chunking, maintaining relationships between clauses, sections, and definitions. Domain adaptation through legal-specific embeddings proves essential for capturing nuanced contractual language. The most notable advantage of RAG-enhanced systems is their ability to correctly interpret contextual references within contracts, such as definitions established in early sections that modify the interpretation of terms throughout the document. This contextual awareness significantly improves extraction accuracy for complex legal instruments with interdependent clauses and defined terms (Yang, T. 2025).

Screenplay analysis demonstrates the adaptability of extraction architectures to specialized document formats with unique structural requirements. Creative content evaluation demands understanding of narrative structure, character development, and stylistic elements beyond simple factual extraction. Implementation approaches employ specialized preprocessing to identify screenplay-specific elements, including scene headings, action descriptions, character dialogues, and transitions. These systems must maintain awareness of narrative progression throughout potentially hundreds of pages while identifying character relationships, plot developments, and tonal shifts. Beyond extraction, implementations often incorporate evaluation components that assess qualitative aspects, including pacing, dialogue quality, and commercial viability through comparison with historical examples (Yang, T. 2025).

Performance characteristics across document types reveal patterns informing implementation strategies for enterprise systems. Studies

Publisher: SARC Publisher

examining extraction quality across diverse categories demonstrate that performance variability correlates with specific document attributes rather than document type classifications. Documents with consistent structure and clear hierarchical organization typically yield higher extraction accuracy regardless of domain. Language complexity metrics, including technical terminology density and sentence complexity, show inverse correlations with extraction quality. These findings suggest optimization should address specific challenging document characteristics rather than building separate pipelines for different document types (Haefner, N. *et al.*, 2023).

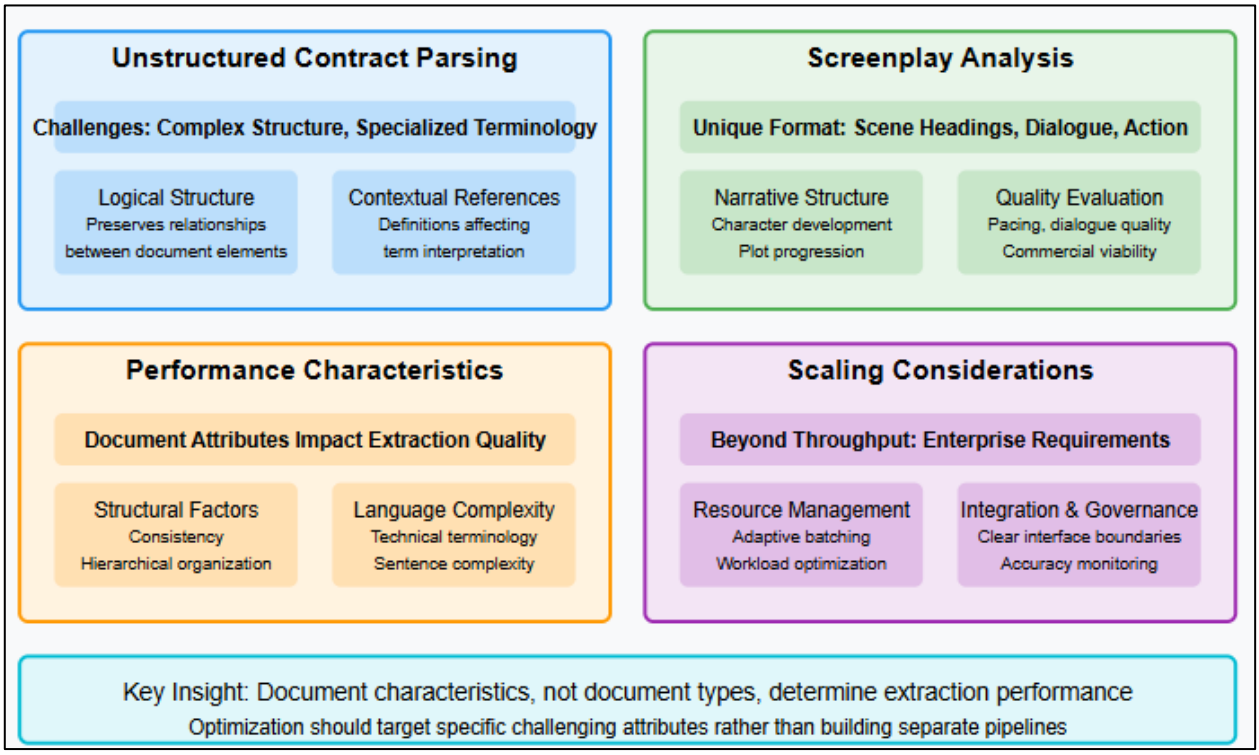

Fig 3: Case Studies: Enterprise Applications (Yang, T. 2025; Haefner, N. *et al.*, 2023)

Scaling considerations for enterprise workloads extend beyond throughput to encompass reliability, resource efficiency, and integration capabilities. Systems processing millions of documents annually must efficiently allocate computational resources across heterogeneous document types with varying processing requirements. Effective implementations employ adaptive batching strategies, optimizing resource utilization while maintaining acceptable latency. Integration with existing enterprise systems represents another critical consideration, with successful deployments implementing clear interface boundaries and robust error handling. Governance frameworks must address both accuracy monitoring and appropriate human oversight, particularly for high-consequence extraction tasks (Haefner, N. *et al.*, 2023).

CHALLENGES AND FUTURE DIRECTIONS

Handling multilingual documents beyond major European languages presents significant challenges for current document processing systems. Multilingual support has traditionally focused on high-resource languages with substantial training data available, creating performance disparities when processing content in less common languages. The challenges stem from several factors, including script complexity, morphological differences, and reading direction variations that impact extraction accuracy. Languages utilizing non-Latin scripts often require specialized tokenization approaches that respect character boundaries and contextual variations. Additionally, languages with complex morphological structures present difficulties for models trained primarily on English and European languages, requiring different segmentation strategies. Research into cross-lingual transfer learning shows promise for addressing these limitations, enabling knowledge transfer from high-resource to low-resource languages without requiring extensive language-specific training data. Future approaches focusing on language-agnostic

Publisher: SARC Publisher

document understanding models could potentially reduce the need for language-specific implementations while maintaining extraction quality across diverse linguistic content (Baviskar, D. *et al.*, 2021).

Strategies for improving scanned document extraction remain a persistent challenge for enterprise document processing systems. Digitized legacy documents often suffer from quality degradation, including low resolution, uneven illumination, geometric distortion, and noise artifacts that significantly impact extraction accuracy. Document preprocessing pipelines incorporating computer vision techniques show promise for addressing these limitations. Image enhancement through adaptive binarization improves contrast in low-quality regions while preserving detail in high-quality areas. Geometric correction algorithms address skew, rotation, and perspective distortion resulting from scanning processes. Specialized segmentation approaches for complex layouts enable more accurate structural understanding of tabular content, multi-column text, and embedded graphics. The integration of document reconstruction techniques shows particular promise, where fragmented or damaged content can be reconstructed through contextual understanding rather than relying solely on visible elements (Baviskar, D. *et al.*, 2021).

Integration with existing enterprise document management systems presents both technical and organizational challenges for advanced extraction implementations. Organizations typically maintain multiple document repositories across departments, each with distinct document structures, metadata schemas, and access patterns. The heterogeneity of these systems creates integration complexity, particularly when connecting modern extraction capabilities with legacy repositories. Successful integration approaches implement abstraction layers that standardize document access across repositories regardless of underlying technology. Event-driven architectures enable real-time processing of documents as they enter repositories, while batch processing approaches optimize throughput for historical document analysis. Metadata harmonization represents a particular challenge, requiring mappings between diverse classification schemes to maintain consistent document understanding (Kampana, V. & Daniel, J. 2024).

Table 1: Challenges and Future Directions in Document Processing (Baviskar, D. *et al.*, 2021; Kampana, V. & Daniel, J. 2024)

Focus Area	Challenges	Solutions
Multilingual Processing	Low support for non-European languages	Cross-lingual models, language-agnostic tools
Scanned Document Extraction	Low quality, distortions, complex layouts	Preprocessing, layout-aware segmentation
System Integration	Legacy systems, metadata mismatch	Abstraction layers, metadata harmonization
Performance Optimization	High cost, scaling issues	Pipelines, dynamic scaling, serverless model

Emerging techniques for performance optimization in distributed environments focus on maximizing throughput while managing computational costs. Document processing workloads benefit from pipeline architecture patterns that separate distinct processing stages, including document ingestion, preprocessing, embedding generation, and inference. This separation enables specialized resource allocation based on the computational profile of each stage. Dynamic scaling approaches adjust resource allocation based on current workload characteristics, expanding capacity during peak processing periods and contracting during lower demand. Processing optimization through batching strategies groups similar documents to maximize throughput, while specialized queuing mechanisms prioritize time-sensitive documents. The serverless computing paradigm shows particular promise for document processing workloads, enabling fine-grained resource allocation without maintaining idle capacity (Kampana, V. & Daniel, J. 2024).

CONCLUSION

LLM-based extraction with RAG architectures represents a paradigm shift in enterprise document processing by addressing fundamental constraints of previous methods. The combination of semantic understanding capabilities with context-aware retrieval mechanisms enables accurate extraction from even highly complex unstructured documents. Cloud-native implementations provide

Publisher: SARC Publisher

the necessary scalability and flexibility for enterprise workloads while maintaining integration with existing systems. Domain-specific applications demonstrate significant advantages in contextual understanding, particularly for documents with complex internal references and specialized terminology. Future advancements will likely focus on improving multilingual capabilities, enhancing performance on degraded documents, standardizing integration patterns, and optimizing computational efficiency. The long-term implications extend beyond technical improvements to fundamental changes in how organizations manage information assets, enabling more effective knowledge discovery and utilization across enterprise operations.

REFERENCES

1. Jaggavarapu, M. K. R. "Transforming Enterprise Document Management: AWS AI and RPA Integration." *International Journal of Computing and Engineering* 7.15 (2025): 50-69.
2. Aman, "History to Modern Era: The Evolution of Intelligent Document Processing." *AmyGB.ai*, (2023).
3. Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. "Exploring AI-driven approaches for unstructured document analysis and future horizons." *Journal of Big Data* 11.1 (2024): 92.
4. Subhani, M. "Impact of AI on Enterprise Cloud-Based Integrations and Automation." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 10 (2024): 1393-1401.
5. Sharma, C. "Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers." *arXiv preprint arXiv:2506.00054* (2025).
6. Liang, W. *et al.*, "Distributed Intelligence at the Edge." *ResearchGate*, (2025).
7. Yang, T. "LLM-Enhanced Data Management in Multi-Model Databases." (2025).
8. Haefner, N., Parida, V., Gassmann, O., & Wincent, J. "Implementing and scaling artificial intelligence: A review, framework, and research agenda." *Technological Forecasting and Social Change* 197 (2023): 122878.
9. Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions." *Ieee Access* 9 (2021): 72894-72936.
10. Kampana, V. & Daniel, J. "Scalable intelligent document processing using Amazon Bedrock." *AWS*, (2024).

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

Cite this article as:

Chansarkar, A. "Enhancing Unstructured Document Text Extraction with LLMs in Cloud-Native Enterprise Solutions." *Sarcouncil Journal of Engineering and Computer Sciences* 4.9 (2025): pp 251-257.

Publisher: SARC Publisher