iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.77762

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Enterprise GenAI: LLM Deployment on AWS

Sufiyan Shaikh[1], Arbaaz Kazi[2], Aryan Kadam[3], Sonali Bansode[4], Prof. Nikhil .R. Zinzurke[5]
[1, 2, 3, 4]*Student, Department of Computer Engineering, KJCOEMR, Pune, India*
[5]*Professor, Department of Computer Engineering, KJCOEMR, Pune, India*

Abstract: ***Generative AI and Large Language Models (LLMs) have transitioned from experimental prototypes to critical enterprise assets, requiring robust, scalable, and secure deployment frameworks. This paper presents a comprehensive survey of LLM deployment strategies on Amazon Web Services (AWS), focusing on the shift from consumer-grade to enterprise-ready architectures. We analyze the AWS Generative AI stack, specifically comparing managed serverless approaches via Amazon Bedrock with customizable infrastructure through Amazon SageMaker. The survey highlights key architectural patterns, including Retrieval-Augmented Generation (RAG) for grounding models in proprietary data and multi-agent systems for complex task orchestration. Furthermore, we examine the critical role of LLMOps in managing the model lifecycle, ensuring security through Guardrails, and optimizing costs via quantization and provisioned throughput. By synthesizing real-world case studies and performance metrics, this paper provides a scalable roadmap for organizations to implement production-grade Generative AI solutions that maintain data sovereignty and operational excellence.***
Keywords: Generative AI, Large Language Model (LLMs), Medicine, AWS Cloud, Amazon Bedrock , Amazon SageMaker, LLMOps , RAG Architecture , Cybersecurity.

I. INTRODUCTION

Enterprise Generative Artificial Intelligence (GenAI) has emerged as one of the most transformative components of the modern digital ecosystem. It enables organizations to automate knowledge work, enhance decision-making, and improve operational efficiency at scale. However, the rapid adoption of Large Language Models (LLMs) in enterprise environments has introduced significant challenges related to security, scalability, governance, and cost management. The involvement of multiple data sources, business units, and deployment environments makes enterprise AI integration complex. Improper deployment of LLM systems can lead to risks such as data leakage, compliance violations, model hallucinations, and unreliable outputs. Apart from operational risks, poorly governed AI systems can also result in substantial financial losses and reputational damage. The main cause of these challenges is the absence of standardized, secure, and auditable frameworks for enterprise-grade LLM deployment. Conventional AI deployment methods, including standalone APIs and isolated cloud instances, provide limited control over data privacy, monitoring, and lifecycle management. These approaches also suffer from disadvantages such as vendor lock-in, limited customization, unpredictable inference costs, and lack of real-time performance optimization. This means that enterprises cannot fully leverage LLM capabilities in a controlled and scalable manner. Recently, cloud-native AI platforms have been identified as a promising approach to overcome these limitations. The scalability, elasticity, and managed service architecture of cloud platforms make them suitable for enterprise-level LLM deployment and governance. Various research studies have proposed cloud-based and hybrid solutions for securely deploying, fine-tuning, and monitoring LLMs in enterprise environments. This survey paper discusses and analyzes the existing work on LLM deployment frameworks with a focus on cloud-based implementations using Amazon Web Services. The aim of this survey paper is to present a clear understanding of the current state of advancements in enterprise LLM deployment on AWS

II. BACKGROUND AND BASIC CONCEPTS

To comprehend the difficulties involved in enterprise adoption of Generative Artificial Intelligence and the technological solutions developed to address them, it is important to discuss the primary challenges associated with deploying Large Language Models.'

A. Large Language Model (LLMs)

Large Language Models (LLMs) are advanced artificial intelligence systems trained on vast amounts of textual data to understand, generate, and analyze human language. These models are built using transformer-based architectures and are capable of performing tasks such as text generation, summarization, question answering, code generation, and conversational assistance.

The adoption of LLMs in enterprises presents significant opportunities for automation and knowledge enhancement. However, improper configuration or deployment of LLMs may result in inaccurate outputs, hallucinations, biased responses, or exposure of sensitive information. In large-scale environments, such risks may lead to operational inefficiencies, compliance violations, and reputational damage.

B. Enterprise AI Deployment Ecosystem

The enterprise AI deployment ecosystem consists of multiple components that work together to enable scalable and secure LLM integration. Typically, this ecosystem includes data sources, preprocessing pipelines, model hosting environments, application interfaces, monitoring tools, and end-users.

Although traditional AI deployment methods follow a structured workflow, they often lack centralized visibility, governance, and cost optimization mechanisms. Data silos, fragmented infrastructure, and limited observability create challenges in managing model performance and compliance.

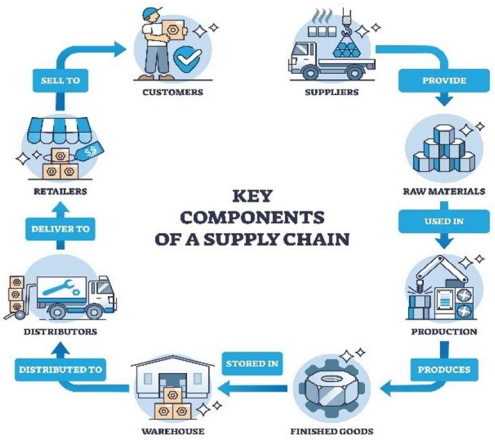


Fig. 1. Traditional Pharmaceutical Supply Chain

C. Cloud Computing Infrastructure (AWS) Architecture

Cloud computing platforms provide the foundational infrastructure required for enterprise-scale LLM deployment. Amazon Web Services (AWS) offers a distributed and elastic computing environment that supports model training, fine-tuning, and inference. Key characteristics of cloud platforms such as scalability, elasticity, reliability, and managed services make them suitable for hosting enterprise LLM workloads. Unlike traditional on-premise systems, cloud-native architectures reduce hardware dependency and operational overhead. These properties allow enterprises to deploy AI models in a flexible, secure, and cost-effective manner while maintaining performance and regulatory compliance.

D. Managed AI Services and Automation

Managed AI services are cloud-based solutions that simplify the deployment and management of LLMs in enterprise environments. These services provide pre-configured frameworks for model hosting, fine-tuning, monitoring, and scaling without requiring deep infrastructure management. Automation tools enable continuous integration, continuous deployment (CI/CD), logging, and performance tracking. By leveraging managed AI services, enterprises can implement reliable, transparent, and governance-driven LLM deployment strategies while minimizing operational complexity and infrastructure risks.

III. LITERATURE SURVEY

Enterprise adoption of Generative Artificial Intelligence has rapidly increased in recent years, particularly with the deployment of Large Language Models in cloud environments. However, secure, scalable, and cost-effective deployment of LLMs remains a major challenge for organizations.

The authors in [1] proposed a cloud-based framework that integrates managed LLM services with distributed computing resources to support enterprise workloads. Their solution improves scalability and deployment flexibility; however, it does not provide a detailed cost analysis

A hybrid deployment architecture combining managed foundation model services and custom model hosting was proposed in [2] to reduce vendor dependency and improve control over enterprise data. In this architecture, proprietary models are accessed through managed APIs, while domain-specific models are hosted separately.

The study introduced in [3] focuses on Retrieval-Augmented Generation (RAG) frameworks deployed on cloud infrastructure to enhance enterprise knowledge retrieval and contextual accuracy. The proposed solution improves response reliability and reduces hallucination; however, the system is not experimentally validated under high-concurrency enterprise workloads.

A similar concept is presented in [4], where cloud-based LLM deployment is introduced as a solution for enterprise automation, but performance optimization and real-time monitoring mechanisms are not deeply investigated.

In [5], the authors proposed an enterprise LLM management framework that integrates logging, access control, and model versioning features. The paper emphasizes governance, auditability, and responsible AI practices within enterprise environments. However, important aspects such as inference cost modeling, resource utilization efficiency, and privacy-preserving fine-tuning techniques are not explored in depth.

A broader perspective is provided in [6], which presents a systematic review of cloud-based AI deployment models across different industries. The authors highlight key challenges including scalability limitations, high operational costs, data privacy risks, and lack of interoperability between legacy enterprise systems and cloud-native AI services. From the literature reviewed, it is clear that the application of blockchain technology has been successful in improving the transparency and traceability of pharmaceutical supply chains. However, there are still challenges in scalability, cost optimization, privacy preservation, usability analysis, and real-time performance analysis, which require further research in this area.

In [8], the authors proposed a cloud-integrated LLM deployment framework enhanced with automated monitoring and feedback mechanisms to improve inference accuracy and system efficiency. The proposed method demonstrates improved resource allocation and better performance tracking in enterprise environments.

[9]. This study emphasizes the importance of secure data exchange, decentralized processing, and low-latency inference in enterprise ecosystems. It also highlights the potential of combining AI services with distributed systems to improve operational intelligence.

In [10], the authors specifically address secure LLM deployment using encrypted data pipelines and role-based access control mechanisms within cloud infrastructure. The system ensures controlled model interaction and reduces the risk of sensitive data exposure. The findings indicate that cloud-native security frameworks significantly enhance enterprise AI governance.

Similarly, [11] presents a solution that combines containerized LLM deployment with automated CI/CD pipelines and infrastructure-as-code practices. This enables enterprises to manage model updates, scaling policies, and compliance requirements efficiently. The study demonstrates that structured DevOps integration can substantially improve reliability and operational efficiency in enterprise GenAI systems.

IV. CHALLENGES AND OPEN ISSUES

Although the adaption of Generative Artificial Intelligence and Large Language Models on cloud platforms Such as AWS has the potential to transform enterprise operations there are several challenges that hinder large scale implementation. Based on the survey of existing literature and industry practices, the following key issues that have been identified in enterprise LLM Deployment .

A. Scalability and Throughput

Large Language Models require substantial computational resources for both training and inference. In enterprise environments where thousands or even millions of user queries may be processed daily, maintaining low latency and high throughput becomes a significant challenge. High concurrency workloads can strain GPU instances, leading to delayed responses and degraded user experience.

B. High Implementation Costs

Deploying LLMs in enterprise environments involves considerable costs related to compute infrastructure, storage, networking, and managed services. GPU-based instances, model fine-tuning processes, and continuous inference workloads significantly increase operational expenditure. Additionally, data transfer costs, monitoring services, and backup mechanisms contribute to the overall financial burden.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

C. Data Privacy vs. Transparency

Enterprise environments handle sensitive data including customer information, financial records, intellectual property, and confidential business documents. When deploying LLMs, particularly through managed APIs or shared cloud infrastructure, there is a risk of unintended data exposure, unauthorized access, or regulatory non-compliance.

V. COMPARATIVE ANALYSIS OF EXISTING APPROACHES

This section provides a comparative study of the most commonly used approaches for validating LLM based on certain parameters such as data storage, security, cost, speed, accessibility, and trust model. The comparison is made based on observations that have been reported in existing literature.

Table 1: Comparison of Existing Enterprise LLM Deployment Approaches

Feature	Managed Foundation Model Services (e.g., Bedrock)	Self-Hosted LLM on Cloud Infrastructure (EC2/SageMaker)	Hybrid Deployment Model
Data Storage	Managed cloud storage with integrated services	Enterprise-controlled storage (S3, databases, private VPC)	Managed APIs with private enterprise storage integration
Security	High (managed IAM, encryption, isolation controls)	High but depends on enterprise configuration	High (combined managed security + enterprise controls)
Cost	Moderate to high (usage-based pricing)	High (GPU instances, maintenance overhead)	Moderate (optimized workload distribution)
Speed	Moderate	High	Moderate to high
Accessibility	Automatic scaling managed by provider	Requires manual or configured autoscaling	Flexible scaling depending on workload type
Trust Model	Trust in cloud provider's managed services	Trust in enterprise-managed infrastructure	Shared trust between provider and enterprise governance

VI. RESEARCH GAPS IDENTIFIED

Based on the analysis of existing literature and comparative evaluation of current approaches, several research gaps have been identified in cloud-based Large Language Model Implementation on AWS:

1) Lack of cost-efficient architectures suitable for large-scale deployment
2) Limited focus standardized governance frameworks
3) Insufficient real-time performance evaluation in existing studies
4) Privacy and data residency concerns
5) Minimal integration with mobile platforms for end-user validation
6) - Absence of regulatory and government-level integration

These gaps indicate the need for further research toward scalable, secure, and LLM technologies in enterprise environments.

VII. CONCEPTUAL SOLUTION MOTIVATION (ENTERPRISEGENAI FRAMEWORK)

Based on the survey of existing enterprise LLM deployment approaches, it is evident that while cloud-based AI services have significantly improved scalability, accessibility, and automation, several limitations remain in areas such as cost optimization, governance enforcement, privacy preservation, and performance consistency. Managed foundation model services simplify integration but may introduce vendor dependency and limited customization. Fully self-hosted deployments offer greater control but require high infrastructure investment and specialized expertise. Additionally, many existing approaches lack standardized frameworks for real-time monitoring, compliance validation, and enterprise-wide governance.

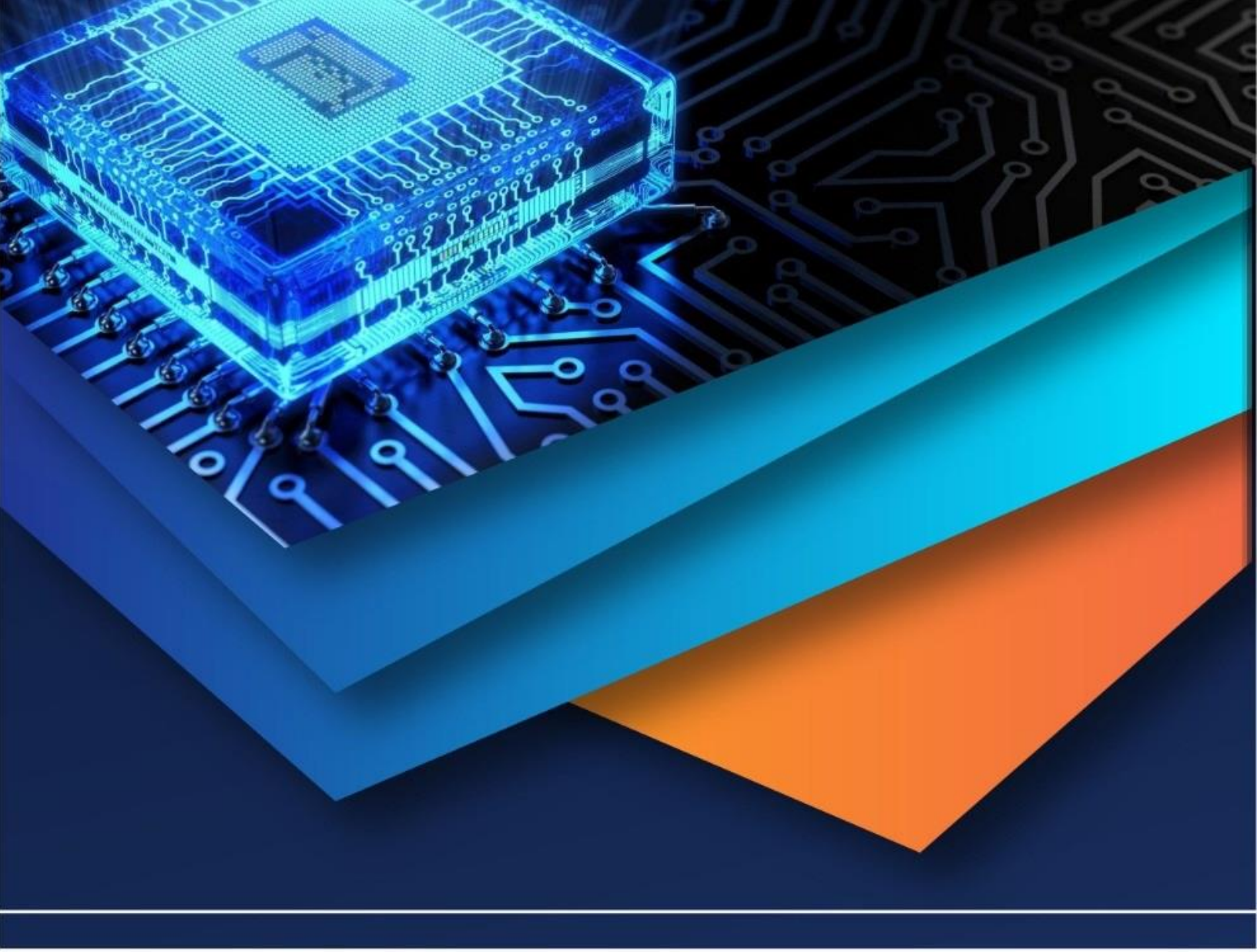
VIII. CONCLUSION

The rapid advancement of Generative Artificial Intelligence has transformed enterprise digital strategies, yet the deployment of Large Language Models at scale presents complex technical and operational challenges. This paper introduced the EnterpriseGenAI Framework, a cloud-based deployment approach for enterprise LLM implementation on AWS aimed at improving scalability, governance, cost efficiency, and security. The proposed framework adopts a hybrid architecture that combines managed foundation model services with enterprise-controlled infrastructure and monitoring mechanisms.

These advancements will contribute to building robust, transparent, and future-proof enterprise GenAI ecosystems powered by AWS.

REFERENCES

[1] T. Brown et al., “Language models are few-shot learners,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 1877–1901.
[2] A. Vaswani et al., “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[3] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” Journal of Machine Learning Research, vol. 23, no. 120, pp. 1–39, 2022.
[4] Amazon Web Services, “Amazon SageMaker: Developer guide,” AWS Documentation, 2023.
[5] Amazon Web Services, “Amazon Bedrock: User guide,” AWS Documentation, 2023.
[6] J. Dean et al., “Large scale distributed deep networks,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1223–1231.
[7] S. Rajbhandari et al., “ZeRO: Memory optimizations toward training trillion parameter models,” in Proc. International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2020, pp. 1–16.S. R. Patel, A. Verma, and P. K. Singh, “A blockchain-based drug supply management system using enhanced learning scheme,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 245–252, 2023.
[8] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 9459–9474.
[9] S. Bubeck et al., “Sparks of artificial general intelligence: Early experiments with GPT-4,” arXiv preprint arXiv:2303.12712, 2023.
[10] M. Mao et al., “Cost-efficient resource provisioning for cloud-based machine learning workloads,” IEEE Transactions on Cloud Computing, vol. 9, no. 2, pp. 456–469, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY