

ZT-RAG: A Zero-Trust Reference Architecture for Secure Cloud-Native Retrieval-Augmented Generation Systems

Atul Khanna

Enterprise Support Manager, AWS
Dallas, TX, USA
khanna_atul@outlook.com

Jay Nankani

Sr Software Developer, Redfin
Seattle, Washington, USA
jay.nankani1@gmail.com

Abhishek Shukla

Technical Account Manager, AWS
Brownsburg, Indiana
secondabhi@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a promising approach for enhancing the accuracy and contextual relevance of Large Language Model (LLM) applications by integrating external knowledge sources into the generation process. Despite its advantages, cloud-native RAG deployments introduce significant security, privacy, and governance challenges, including prompt injection attacks, retrieval poisoning, unauthorized knowledge access, insider threats, vector database compromise, and sensitive information leakage. Existing RAG architectures primarily focus on retrieval effectiveness and scalability while offering limited support for comprehensive security and trust management across the retrieval and generation lifecycle. This paper proposes ZT-RAG, a Zero-Trust Reference Architecture for Secure Cloud-Native Retrieval-Augmented Generation Systems. The proposed framework extends Zero-Trust principles beyond traditional network security by embedding continuous verification, identity-centric access control, policy-driven retrieval governance, trust-aware knowledge validation, secure inference, and output protection throughout the RAG pipeline. To strengthen retrieval security, ZT-RAG introduces a Retrieval Trust Score (RTS) framework that evaluates retrieved knowledge based on source credibility, integrity, freshness, and policy compliance before augmentation occurs. The architecture further defines trust boundaries across user interactions, retrieval services, knowledge repositories, language models, and response delivery channels. A threat-driven security analysis demonstrates how ZT-RAG mitigates major attack vectors affecting enterprise RAG deployments. A feature-based comparative analysis across ten security dimensions against four representative approaches - baseline RAG prompt-injection-focused defenses, PoisonedRAG countermeasures, and generic Secure-RAG frameworks, demonstrates that the proposed architecture provides stronger security, governance, accountability, and trust assurance than all compared approaches. ZT-RAG offers a practical foundation for developing trustworthy, secure, and compliant Retrieval-Augmented Generation systems in cloud environments.

Keywords

Zero Trust Architecture, Cloud Security, Large Language Models, Trustworthy AI, Access Control, AI Governance.

Biographies

Atul Khanna is the Enterprise Support Manager at Amazon Web Services (AWS), a leading global cloud computing platform. In his role, he leads high-performing teams of Technical Account Managers specializing in Data Analytics and Generative AI technologies, helping enterprise customers achieve their business targets by translating strategic business objectives into actionable work streams and scalable cloud solutions. Atul holds a CCIE certification in Data Center and Collaboration, reflecting his deep technical expertise across cloud infrastructure, networking, and enterprise systems. This multidisciplinary foundation allows him to bridge the gap between complex cloud technologies and real-world business outcomes for organizations at scale. Before joining AWS, Atul held progressive leadership roles at Cisco Systems and Twilio, where he led technical support teams, drove data center operations, managed strategic enterprise accounts, and contributed to the design of customer success frameworks across global organizations. As a technology leader, Atul is recognized for his expertise in cloud operations, C-suite engagement, cross-functional collaboration, and team development. His work focuses on driving operational excellence, accelerating cloud adoption, and enabling enterprise customers to realize the full value of modern cloud and AI-driven architectures. Atul is dedicated to advancing enterprise technology by championing methodologies that connect rigorous cloud engineering with meaningful business impact.

Jay Nankani is a seasoned Software Developer at Redfin and an active community researcher with more than five years of specialized experience in the technology sector. Driven by a passion for innovation, Jay has successfully engineered and delivered dozens of high-impact technical projects, consistently optimizing platform performance and enhancing user experiences. Beyond his core engineering responsibilities, Jay is deeply committed to advancing the frontiers of technology through academic and applied research. His scientific contributions include publishing multiple peer-reviewed research papers with a primary specialization in Artificial Intelligence. By exploring cutting-edge AI methodologies, his work aims to bridge the gap between complex algorithmic theory and practical, scalable industry applications. As a dedicated advocate for scientific progress and knowledge sharing, Jay frequently contributes to the broader technical community. He actively serves as a peer reviewer for prestigious conferences and journals, ensuring the integrity and quality of emerging research in his field. Through this blend of industry-leading software development, rigorous academic inquiry, and community mentorship, Jay continues to drive meaningful advancement in the ever-evolving landscape of artificial intelligence and software engineering.

Abhishek Shukla is a Senior Technical Account Manager at Amazon Web Services (AWS), specializing in cloud architecture optimization, storage, database modernization, and large-scale data migrations. With extensive expertise in enterprise cloud infrastructure, his work focuses on pioneering performance tuning methodologies, database engine transitions, and automated infrastructure management. As a thought leader in the technology sector, he actively contributes to the advancement of cloud computing paradigms, including the integration of agentic AI frameworks to enhance operational efficiency. His research and technical insights drive scalable, secure solutions for complex global enterprises. Beyond his corporate leadership, he is dedicated to publishing technical literature and mentoring organizations through critical digital transformations.