Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 07| 2025

Research Article **Received:** 26-05-2025 | **Accepted:** 16-06-2025 | **Published:** 02-07-2025

Secure by Design: Architecting Enterprise AI with Comprehensive Access Control across RAG, Fine-tuning, and Data Interaction Systems

***Junaid Syed*[1], *Zubair Atha*[2], *and Bushra Aijaz*[3]**
[1]*Georgia Institute of Technology, USA*
[2]*Columbia University, USA*
[3]*Georgia Institute of Technology, USA*

Abstract: This article presents a comprehensive framework for securing enterprise AI systems across their entire lifecycle, with a focus on Retrieval-Augmented Generation (RAG), fine-tuning, and data interaction interfaces. The architecture advocates for a "protection by design" philosophy that embeds robust authorization controls into every layer of the AI stack. Beginning with secure data preparation and storage, the framework addresses vector database vulnerabilities through the use of authorization metadata, section-level classification, and encrypted embeddings. It then explores access-controlled retrieval pipelines that maintain protection boundaries during inference through identity propagation, permission-based query augmentation, and multi-stage retrieval processes. For model adaptation, the article details strategies for safeguarding training data and hardening fine-tuning pipelines to prevent the extraction of confidential information. Finally, it examines protection mechanisms for user interaction interfaces, including natural language query fortification and comprehensive audit capabilities. Throughout, the article emphasizes that effective AI security requires not merely perimeter defenses but deeply integrated protection measures that preserve functionality while maintaining strict confidentiality boundaries.

Keywords: Enterprise AI Security, Access Control Architecture, Retrieval-Augmented Generation, Fine-Tuning Protection, Data Interaction Safeguards.

INTRODUCTION

The integration of advanced AI systems within enterprise environments represents a transformative shift in how organizations interact with their data assets. As Large Language Models (LLMs) and other AI technologies become increasingly sophisticated, their potential to unlock insights from vast, unstructured corporate datasets grows exponentially. However, this technological evolution introduces essential protection challenges that must be addressed to prevent unauthorized data access, ensure compliance with regulatory frameworks, and maintain the integrity of sensitive information [https://aws.amazon.com].

Enterprise data ecosystems typically contain a complex mixture of information with varying sensitivity levels, from publicly shareable content to highly restricted personal, financial, or strategic data. The democratization of access to this data through AI-powered interfaces creates a tension between accessibility and safeguarding that traditional authorization mechanisms are ill-equipped to resolve. When users can query data through natural language interfaces or leverage AI models trained on corporate data, conventional perimeter-based protection approaches prove insufficient. RAG systems particularly emphasize this challenge as they directly connect LLMs to organizational data stores [https://aws.amazon.com].

This article proposes a comprehensive protection architecture for enterprise AI systems that addresses these challenges through the lens of "protection by design." Rather than treating safeguarding as an afterthought, embedding comprehensive authorization frameworks into every layer of the AI stack creates resilient systems that maintain protection boundaries without compromising functionality.

1.1 The Enterprise AI Security Imperative

The imperative for enhanced safeguarding in enterprise AI systems stems from several factors. Data privacy regulations such as GDPR, CCPA, HIPAA, and industry-specific requirements mandate strict controls over how personal and confidential data is processed. Corporate datasets often contain proprietary information that requires defense against unauthorized extraction. Organizations must maintain auditability and accountability for all data access, particularly when automated systems are involved. The integration of external knowledge with AI reasoning capabilities further amplifies these concerns [https://aws.amazon.com].

1.2 The Security Challenges of Modern AI Architectures

Contemporary enterprise AI deployments typically leverage several interconnected approaches. Retrieval-Augmented Generation enhances LLM outputs by retrieving information from corporate

***Corresponding Author:** Junaid Syed

DOI- https://doi.org/10.5281/zenodo.15792315

knowledge bases, creating complex protection challenges at the retrieval layer [https://aws.amazon.com]. Fine-tuning adapts pre-trained models to specific organizational contexts, necessitating careful governance controls throughout the adaptation process. Data interaction systems provide interfaces that allow users to query, analyze, and visualize information through natural language. These hybrid approaches require layered defense strategies that address vulnerabilities at each stage of data processing and model interaction [Yuksel, I. *et al*., 2025].

The subsequent section examines the foundational layer where all data protection begins: preparation and storage architecture.

2. Secure Data Preparation and Storage Architecture

The foundation of secure enterprise AI systems begins with the data preparation and storage layer. This critical infrastructure must be designed to maintain protection controls and data governance principles throughout the lifecycle of information used by AI systems. Security vulnerabilities in vector databases and embedding systems have been identified as significant concerns in enterprise AI deployments, highlighting the need for thorough safeguarding strategies at this foundational level [Bar-El, T, 2025].

2.1 Secure Data Ingestion and Preprocessing

Before data enters the AI ecosystem, organizations must implement systems that capture and preserve authorization metadata. Document-level access control tags ensure each document ingested maintains metadata about who can access it, including role-based (RBAC) or attribute-based (ABAC) control information. Vector databases require particularly careful protection configuration as they represent a centralized repository of organizational knowledge that could be exploited if access controls are inadequate [Bar-El, T, 2025].

Section-level classification represents an advanced safeguarding approach for documents with varying sensitivity levels across sections. Preprocessing pipelines identify and tag sections according to their confidentiality requirements, allowing for more granular control than document-level permissions alone. Rigorous verification processes must be implemented for all vector store operations to prevent unauthorized access to sensitive embeddings [Bar-El, T, 2025].

Provenance tracking maintains clear records of data sources to ensure only authorized data enters the system. When preparing data for RAG systems, the chunking process must preserve protection contexts through protection-preserving chunk boundaries. Metadata inheritance ensures each chunk inherits the safeguarding properties of its source document or section, creating a continuous chain of defense controls throughout the data transformation process [Bar-El, T, 2025].

2.2 Secure Vector Stores and Embedding Management

To maintain strict separation between data with different confidentiality requirements, multi-tenant vector architectures implement separate vector spaces for different protection domains or business units. The privacy implications of vector embeddings extend beyond basic authorization, as these mathematical representations can potentially be reverse-engineered to reveal sensitive information about the original content [Koga, T. *et al*., 2024].

Logical partitioning uses sophisticated schemes within vector databases to enforce access boundaries even within shared infrastructure. Cross-domain query controls implement mechanisms to control when and how queries can span multiple protection domains, preventing unauthorized correlation of information across confidentiality boundaries [Koga, T. *et al*., 2024].

The embeddings themselves represent a potential vulnerability that must be addressed through encrypted embeddings. Applying encryption to vector embeddings prevents unauthorized analysis while maintaining retrieval functionality. Differential privacy in embeddings implements noise addition techniques to prevent the extraction of sensitive information, while embedding authorization controls regulate which systems and users can access the raw vector representations [Koga, T. *et al*., 2024].

With secure data foundations in place, the focus shifts to ensuring that retrieval processes, the dynamic bridge between stored data and AI models, maintain these protection boundaries during operation.

Publisher: SARC Publisher

Table 1: Enterprise AI Data Security Mechanisms [Bar-El, T, 2025; Koga, T. *et al*., 2024]

Security Component	Security Function
Document-level access controls	Maintain user permission metadata
Section-level classification	Enable granular security for varying sensitivity
Security-preserving chunks	Prevent leakage during RAG preparation
Multi-tenant vector spaces	Separate different security domains
Encrypted embeddings	Prevent unauthorized vector analysis

3. Access-Controlled Retrieval Pipelines

The retrieval mechanism in RAG systems represents a vital control point for enforcing protection policies. This section outlines architectural approaches to embedding authorization frameworks directly into the retrieval process. As illustrated in Fig. 1, protected retrieval pipelines implement multiple safeguarding layers throughout the data access flow. As LLM agents become increasingly autonomous in accessing and processing confidential information, the need for stringent safeguards integrated into the retrieval pipeline grows more essential [Bar-El, T, 2025].

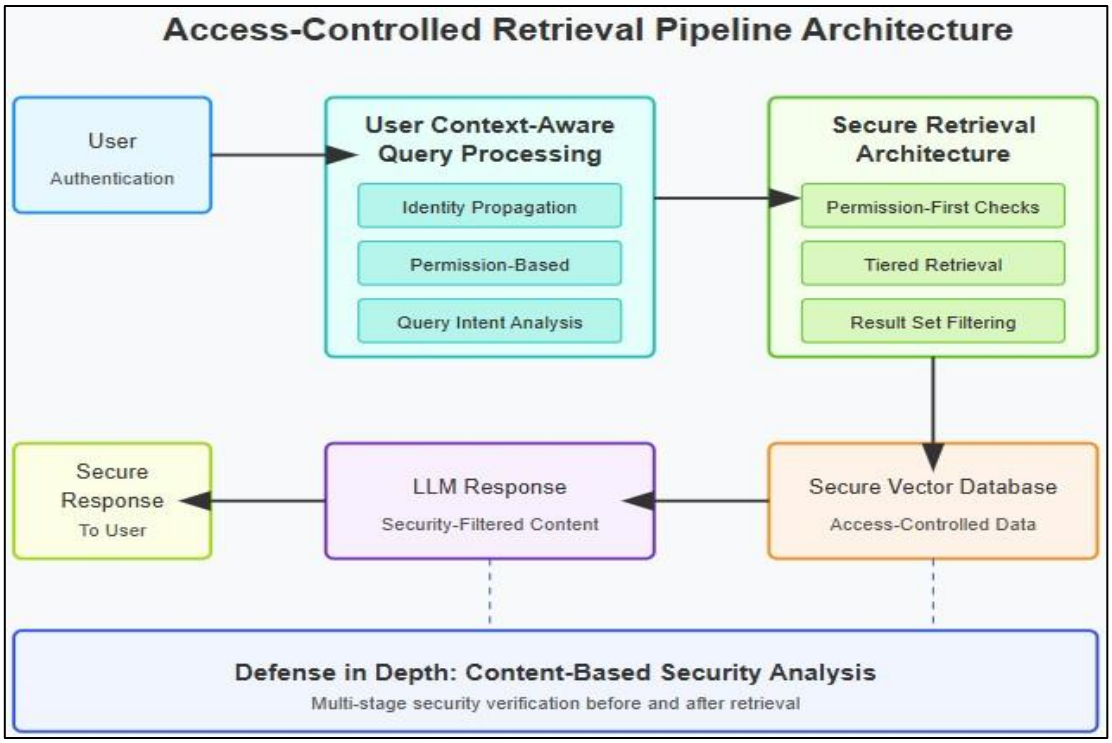

Fig 1: Security-First RAG Architecture: Multi-layered Access Controls in Retrieval Pipelines [Bar-El, T, 2025; Rahman, M. *et al*., 2024]

3.1 User Context-Aware Query Processing

Ensuring that user identity and permissions are tightly integrated with the query process creates the foundation for protected retrieval operations. As shown in the left section of Fig. 1, identity propagation establishes protected channels for passing authenticated user identity through all system components, maintaining a consistent authorization context from the initial user interface through to the retrieval engine. Token-based authentication implements verification tokens that encode user permissions relevant to data access, enabling stateless validation at each processing stage. LLM agents require particularly rigorous verification mechanisms as they operate with varying levels of autonomy while accessing sensitive enterprise data sources [Bar-El, T, 2025].

Continuous authentication implements methods to verify user identity throughout extended interaction sessions, addressing the protection challenges of long-running AI assistants. Modifying queries to incorporate authorization constraints before execution provides another essential protection layer. Permission-based query augmentation automatically enhances queries with access predicates based on user permissions, ensuring that retrieval operations respect

Publisher: SARC Publisher

authorization boundaries from the earliest processing stages. Access scope limiting adds constraints to vector similarity searches that restrict results to authorized data, while query intent analysis examines queries to identify potential attempts to circumvent governance controls [Bar-El, T, 2025].

3.2 Secure Retrieval Architectures

Implementing authorization verification before retrieval operations occur provides significant efficiency and protection benefits, as depicted in the central flow of Fig. 1. Permission-first architecture evaluates authorizations before executing expensive vector similarity searches, reducing computational waste and preventing potential information leakage through timing side channels. Governance-indexed retrieval maintains secondary indices of permission attributes for efficient filtering, enabling rapid authorization evaluation without compromising retrieval performance [Rahman, M. *et al*., 2024].

The tiered retrieval process implements a multi-stage approach where the initial stages focus on permission filtering before proceeding to more comprehensive semantic search. This architectural pattern enables more efficient processing by quickly eliminating unauthorized content from consideration [Rahman, M. *et al*., 2024].

Adding additional protection layers after initial retrieval provides defense-in-depth against potential safeguarding bypasses, as shown in the bottom section of Fig. 1. Result set filtering applies secondary authorization checks to candidate documents retrieved by vector search, while dynamic permission evaluation assesses complex access rules that cannot be efficiently encoded in the initial retrieval. Content-based analysis examines retrieved information for potential policy violations before passing content to LLMs for response generation, creating a comprehensive protection framework that safeguards sensitive information while maintaining system functionality [Rahman, M. *et al*., 2024].

While protected retrieval addresses data access during inference, organizations must also establish thorough governance controls during the model adaptation process itself, which introduces its unique challenges.

4. Protected Fine-Tuning and Model Management

Fine-tuning LLMs on enterprise data introduces unique safeguarding challenges that must be addressed through thorough controls throughout the model lifecycle. Recent research has demonstrated that fine-tuned models may inadvertently memorize confidential training data, creating potential vectors for information extraction through carefully constructed prompts or membership inference attacks [Mattern, J. *et al*., 2023].

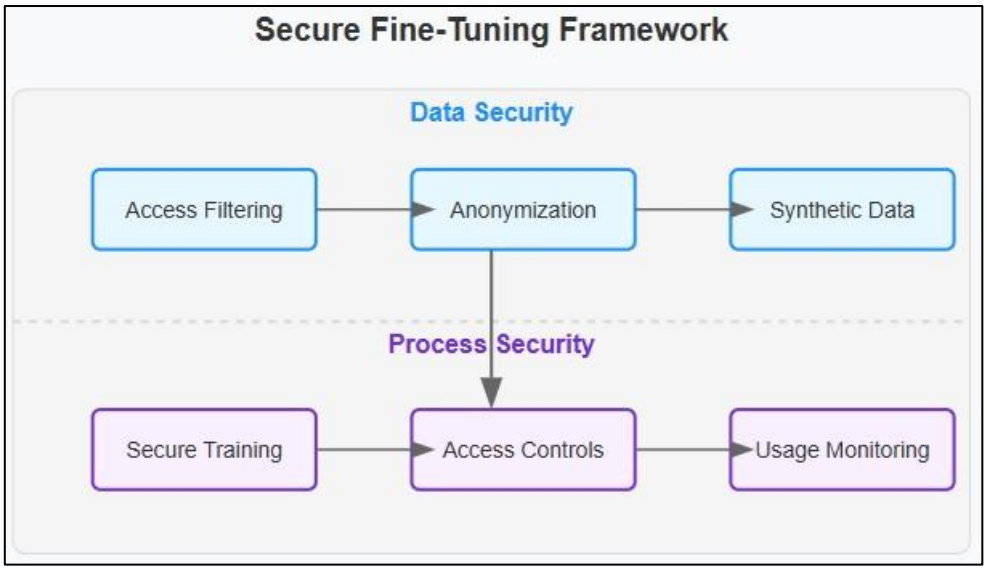

Fig. 2: Dual-Phase Enterprise LLM Governance: Data Protection and Process Controls [Mattern, J. *et al*., 2023; Parthasarathy, V.B. *et al*., 2024]

4.1 Protected Dataset Preparation for Fine-Tuning

As illustrated in the top section of Fig. 2, safeguarding the training data is the first essential phase in the fine-tuning protection framework.

Publisher: SARC Publisher

Access-based data filtering creates training datasets that only include data that the model's intended users will have permission to access. Studies examining attacks against language models have revealed that adversaries can extract training data through carefully crafted prompts, making rigorous data filtering vital for confidentiality [Mattern, J. *et al*., 2023].

The anonymization component implements techniques to remove sensitive information while preserving training utility. Research has shown that even when confidential information isn't directly provided to a model, it may still be vulnerable to extraction attacks through indirect inference [Mattern, J. *et al*., 2023].

The synthetic data approach depicted in the framework provides another important protection layer. Privacy-preserving synthetic data generates training examples that capture patterns without exposing actual confidential data. Differential privacy in training data prevents memorization of sensitive information, making extraction attacks significantly more difficult without compromising model utility [Mattern, J. *et al*., 2023].

4.2 Hardened Fine-Tuning Pipelines

As shown in the bottom section of Fig. 2, process protection forms the second phase of the framework. Isolated training environments implement physically or logically separated infrastructure for fine-tuning on highly confidential data, preventing potential exfiltration through network connections. Recent research has demonstrated the effectiveness of private multiparty computation techniques that enable collaborative fine-tuning while keeping sensitive data confidential [Parthasarathy, V.B. *et al*., 2024].

The authorization component implements strict governance around who can initiate and monitor fine-tuning processes. Studies examining hardened LLM adaptation have emphasized the importance of trusted execution environments for safeguarding sensitive operations [Parthasarathy, V.B. *et al*., 2024].

The usage oversight element extends protection beyond the training phase to ongoing model usage. Model permission systems determine which users can access which fine-tuned models, ensuring that models remain appropriately bounded in their deployment. Activity monitoring and quotas track model interactions and enforce appropriate limits, preventing potential abuse through excessive querying or pattern analysis [Parthasarathy, V.B. *et al*., 2024].

Having fortified both the data infrastructure and model lifecycle, the final essential area for protection is the interface where users actually interact with these AI systems.

5. Protected Data Interaction Systems

The interfaces through which users interact with AI-powered data systems represent the final vital component of the protection architecture. As enterprise AI systems increasingly provide natural language interfaces to confidential corporate data, safeguarding these interaction points becomes essential to maintaining data confidentiality throughout the AI lifecycle [Sanka, V, 2025].

5.1 Natural Language Query Protection

Analyzing and fortifying natural language queries provides the first line of defense in protected data interaction systems. Protection intent analysis classifies queries based on their confidentiality implications before processing, enabling early identification of potentially malicious requests. Comprehensive frameworks for natural language-driven business intelligence emphasize the importance of query intent classification as a foundational safeguarding measure [Sanka, V, 2025]. Permission-aware query routing directs queries to appropriate backend systems based on authorization requirements, ensuring that requests are handled by components with appropriate governance controls.

Query transformation rewrites queries to enforce access boundaries while preserving user intent, providing an additional layer of protection that modifies potentially problematic requests rather than simply blocking them. Research on conversational interfaces for enterprise data analytics has highlighted the importance of query transformation in maintaining both confidentiality and usability [Sanka, V, 2025].

Ensuring generated responses adhere to protection policies represents another essential aspect of interaction safeguarding. Content filtering pipelines implement multi-stage filtering for generated content based on confidentiality policies, preventing inadvertent disclosure of sensitive information. Attribution with contextual notices includes appropriate confidentiality context in responses to clarify information sources and restrictions. Adaptive response limitation dynamically adjusts the detail level of responses based on user permissions, ensuring that users

Publisher: SARC Publisher

receive only the level of information appropriate to their authorization clearance and role [Sanka, V, 2025].

5.2 Comprehensive Audit and Oversight

Creating thorough audit trails for all system interactions provides essential visibility into AI system usage and potential policy violations. Immutable audit logs implement tamper-resistant recording for all data access and model interactions, creating an authoritative record that can be used for compliance verification and incident investigation. Modern approaches to AI threat testing emphasize the importance of comprehensive logging to enable thorough risk analysis [Kumbhani, H, 2024].

Pattern analysis examines interaction behaviors to identify potential anomalies or misuse, enabling proactive detection of potential threats before they result in significant data breaches. Advanced testing methodologies leverage pattern analysis to identify subtle indicators of malicious activity [Kumbhani, H, 2024]. Real-time monitoring implements continuous oversight of AI system usage for policy violations, providing immediate visibility into potential incidents.

Anomaly detection develops models to identify unusual access patterns or potential breaches, leveraging machine learning techniques to identify subtle patterns that might indicate malicious activity. Automated response mechanisms create systems that can immediately react to detected threats, enabling rapid mitigation even before human analysts can intervene [Kumbhani, H, 2024].

By implementing comprehensive protection across data preparation, retrieval processes, model management, and user interfaces, organizations can confidently deploy enterprise AI systems that maintain data confidentiality while delivering valuable insights.

Table 2: Enterprise AI Interaction Security Controls [Sanka, V, 2025; Kumbhani, H, 2024]

Security Layer	Security Function
Query Intent Analysis	Classify queries based on security implications
Permission-Aware Routing	Direct queries to the appropriate backend systems
Query Transformation	Rewrite queries to enforce security boundaries
Content Filtering	Implement multi-stage filtering for generated content
Immutable Audit Logs	Create tamper-resistant records of all interactions

CONCLUSION

The architecture detailed in this article establishes a holistic protection framework that addresses the unique security challenges introduced by enterprise AI systems. By embedding authorization controls throughout the entire AI lifecycle, from initial data preparation through retrieval mechanisms, model adaptation, and user interfaces, organizations can realize the transformative potential of these technologies while maintaining strict governance over sensitive information. The multi-layered approach described ensures that protection measures operate in concert, with each layer reinforcing confidentiality boundaries established by others. This defense-in-depth strategy proves essential for mitigating the novel vulnerabilities introduced when granting AI systems access to organizational knowledge bases. As enterprise adoption of advanced AI technologies continues to accelerate, implementing such comprehensive protection architectures will be fundamental to balancing the competing imperatives of data accessibility and confidentiality. The framework presented provides a blueprint for organizations seeking to harness the insights of AI while maintaining rigorous safeguards over their most valuable information assets.

REFERENCES

1. AWS. “What is Retrieval-Augmented Generation (RAG)?" https://aws.amazon.com/what-is/retrieval-augmented-generation/
2. Yuksel, I. and Akhnoukh, K. “Tailoring foundation models for your business needs: A comprehensive guide to RAG, fine-tuning, and hybrid approaches.” *AWS* (2025). https://aws.amazon.com/blogs/machine-learning/tailoring-foundation-models-for-your-business-needs-a-comprehensive-guide-to-rag-fine-tuning-and-hybrid-approaches/
3. Bar-El, T. “Vector and Embedding Weaknesses in AI Systems.” *Mend.io*, (2025). https://www.mend.io/blog/vector-and-embedding-weaknesses-in-ai-systems/#:~:text=Implementing%20strict%20access%20controls%20is,write%20to%20the%20vector%20store.

Publisher: SARC Publisher

4. Koga, T., Wu, R. and Chaudhuri, K. "Privacy-Preserving Retrieval Augmented Generation with Differential Privacy." *arXiv preprint arXiv:2412.04697* (2024). https://arxiv.org/pdf/2412.04697
5. Edwards, J. "Understanding AI & LLM Agents: Architecture, Security, & Deployment." *Skyflow*, (2025). https://www.skyflow.com/post/understanding-llm-agents
6. Rahman, M., Piryani, K.O., Sanchez, A.M., Munikoti, S., De La Torre, L., Levin, M.S., Akbar, M., Hossain, M., Hasan, M. and Halappanavar, M. "Retrieval Augmented Generation for Robust Cyber Defense." *Pacific Northwest National Laboratory*, (2024). https://www.osti.gov/servlets/purl/2474934
7. Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M. and Berg-Kirkpatrick, T. "Membership inference attacks against language models via neighbourhood comparison." *arXiv preprint arXiv:2305.18462* (2023). https://arxiv.org/pdf/2305.18462
8. Parthasarathy, V.B., Zafar, A., Khan, A. and Shahid, A. "The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities." *arXiv preprint arXiv: 2408.13296* (2024). https://arxiv.org/html/2408.13296v1
9. Sanka, V. "Conversational AI for Enterprise Data Analytics and Governance: A Comprehensive Framework for Natural Language-Driven Business Intelligence." *International Journal of Scientific Research in Computer Science Engineering and Information Technology* 11.2 (2025): 3922-3928. https://www.researchgate.net/publication/392743081_Conversational_AI_for_Enterprise_Data_Analytics_and_Governance_A_Comprehensive_Framework_for_Natural_Language-Driven_Business_Intelligence
10. Kumbhani, H. "Transforming Security Testing With AI: Benefits and Challenges." *Cyber Defense Magazine*, (2024). https://www.cyberdefensemagazine.com/transforming-security-testing-with-ai-benefits-and-challenges/

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

Cite this article as:
Syed, J., Atha, Z. and Aijaz, B. "Secure by Design: Architecting Enterprise AI with Comprehensive Access Control across RAG, Fine-tuning, and Data Interaction Systems." *Sarcouncil Journal of Engineering and Computer Sciences* 4.7 (2025): pp 25-31.

Publisher: SARC Publisher