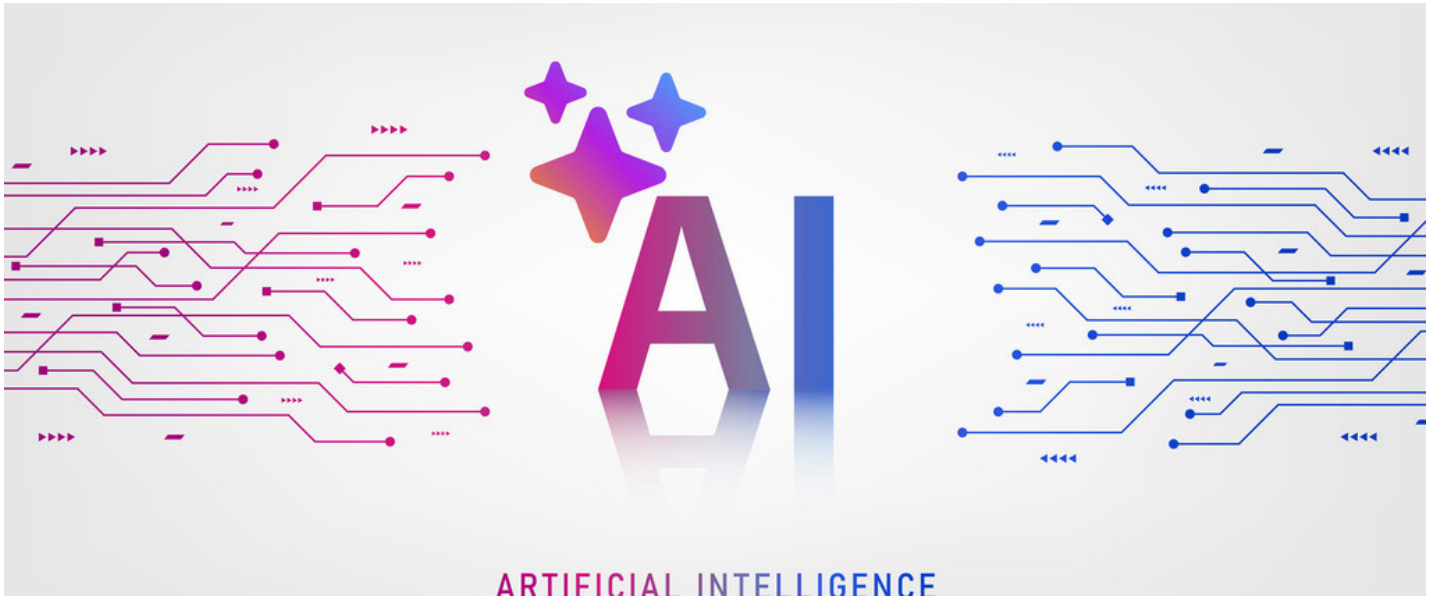
Link to article: https://discover.claromentis.com/knowledgebase/articles/1017/best-practice-recommendation-generative-ai-and-llm-implementation



Best Practice Recommendation Generative AI and LLM Implementation

Overview

Implementing Generative AI with Vector Search RAG (Retrieval Augmented Generation) within an organisation requires a robust approach to ensure data safety and proper permissions.

Here are the general approach & best practice recommendations.

1. Data Governance and Classification:

Comprehensive Data Inventory and Classification: Identify and classify all data sources that will be used for embedding and retrieval. Categorise data by sensitivity (e.g., public, internal, confidential, highly sensitive, PII, intellectual property) to inform access policies.

Data Minimisation: Only use the minimum amount of data necessary for training and generating outputs. Avoid ingesting or making available sensitive data that isn't absolutely essential for the RAG system's purpose.

Data Lineage and Provenance: Establish clear documentation of data origins, transformations, and usage within the RAG pipeline. This helps with auditing and understanding potential risks.

Data Anonymisation and Pseudonymization: For sensitive data, apply techniques like generalisation, suppression, or tokenisation before creating embeddings and storing them in the vector database. This reduces the risk of exposing identifiable information. Differential privacy can also be employed by adding noise to datasets.

2. Robust Access Control and Permission Management

Granular Role-Based Access Control (RBAC): Implement strict RBAC for both the Generative AI models and the vector database. Define roles with the principle of least privilege, ensuring users and AI agents only have access to the data necessary for their tasks.

- **Context-Aware Policies:** Go beyond static permissions by implementing policies that evaluate real-time factors like user location, device, and data sensitivity to dynamically adjust access permissions.
- **Relationship-Based Access Control (ReBAC):** For RAG systems, ReBAC is particularly useful. The system can retrieve only the documents the user is authorised to access by checking relationships between users and data.

Authorisation at the Vector Database Level: Authorise access to data before sending additional content as part of the prompt to the LLM. This can be implemented by:

- Creating separate vector databases for different departments or data sensitivity levels.
- Using metadata filtering within the vector database, where API calls include user or group membership information to filter results. This prevents prompt injection attacks from bypassing authorisation.

Secure API Endpoints: All APIs used for interacting with the Generative AI models and vector database must be secured with robust authentication and authorisation mechanisms (e.g., OAuth, API keys, mutual TLS).

Multi-Factor Authentication (MFA): Enforce MFA for all administrative access to the RAG infrastructure.

Human Oversight: Maintain human oversight mechanisms to review and validate generated outputs, especially for critical decisions or sensitive information.

Training and Awareness: Conduct regular training for all employees on the secure and ethical use of Generative AI and RAG systems, emphasising data handling procedures and incident reporting protocols.

3. Data Protection and Security Measures:

Encryption In-Transit and At-Rest: All data, including embeddings in the vector database, should be encrypted both at rest (storage) and in transit (network communication) using strong encryption standards (e.g., AES-256). Implement strong key management practices.

Secure Enclaves and Confidential Computing: Consider using confidential computing techniques, like secure enclaves, to protect data and models during processing, isolating sensitive computations.

Input Validation and Sanitisation: Implement stringent input validation and sanitisation techniques to prevent prompt injection attacks and ensure only clean, authorised data enters the system. This is crucial for preventing malicious inputs from compromising the AI system or causing data leakage.

Output Validation and Guardrails: Implement automated validation checks and guardrails to ensure outputs are accurate, relevant, appropriate, and do not inadvertently expose sensitive information. This can involve rule-based systems or custom-built validators that cross-reference generated content with trusted data sources.

Regular Security Assessments: Conduct frequent security assessments, vulnerability scanning, penetration testing, and red teaming exercises to identify and address vulnerabilities in the RAG pipeline.

Incident Response Plan: Develop a comprehensive AI incident response plan specifically for Generative AI and RAG systems to minimise damage and restore normal operations in case of a security breach or data leakage.

Vendor Due Diligence: Thoroughly vet any third-party vendors for Generative AI models, vector databases, or related services to ensure they meet your organisation's security and compliance standards.

4. Monitoring, Auditing, and Compliance:

Comprehensive Logging and Monitoring: Implement robust monitoring and logging mechanisms to track user inputs, queries, system behaviour, and generated outputs in real-time. Use anomaly detection algorithms to identify unusual patterns that may indicate malicious activity.

Audit Trails: Maintain detailed audit trails of all data access, model interactions, and configuration changes for compliance and forensic analysis.

Continuous Compliance Monitoring: Stay abreast of relevant data protection regulations (e.g., GDPR, CCPA, AI Act) and continuously monitor the RAG system's compliance with these regulations. Conduct regular compliance reviews.

Transparency and Explainability: Where possible, strive for transparency in how the RAG system processes and generates responses, especially concerning data retrieval. This aids in debugging and understanding potential biases or data exposures.

By implementing these best practices, organisations can significantly mitigate the risks associated with deploying Generative AI and Vector Search RAG, ensuring data privacy, security, and responsible AI usage.

Created on 2 June 2025 by Michael Christian