

Enterprise AI Security Architecture

A Comprehensive Enterprise Architecture Approach for Securing Generative AI, Agentic AI, Enterprise Data, Models, APIs, Identity, Cloud Platforms, and Responsible AI

Sanjeeve Kumar Gajadi*

*(HCL Technologies, Bengaluru, Karnataka, India)

Email: sanjeevekumar.g@hcltech.com

Abstract:

Artificial Intelligence is becoming a core enterprise capability across customer service, software engineering, finance, supply chain, analytics, knowledge management, cybersecurity, and autonomous operations. However, the rapid adoption of Generative AI, Large Language Models, Retrieval-Augmented Generation, AI copilots, and agentic AI introduces a new attack surface that extends beyond conventional application security. Enterprises must protect models, prompts, embeddings, vector stores, APIs, tools, identities, business data, AI agents, orchestration layers, and cloud platforms while maintaining privacy, compliance, explainability, availability, and business trust. This paper proposes an Enterprise AI Security Architecture that integrates Zero Trust, Identity and Access Management, DevSecOps, model security, data security, prompt protection, API security, agent governance, threat detection, secure AI operations, incident response, and Responsible AI into a unified enterprise architecture. The framework establishes security controls across the complete AI lifecycle, including design, data preparation, model onboarding, RAG, agent execution, deployment, monitoring, and retirement. The proposed architecture enables organizations to scale enterprise AI securely across hybrid and multi-cloud environments while reducing cyber risk, protecting sensitive information, supporting regulatory compliance, and strengthening operational resilience.

Keywords — Enterprise AI Security, Generative AI Security, Agentic AI, Zero Trust, SAP BTP, SAP AI Core, SAP AI Launchpad, SAP Joule, Identity and Access Management, Prompt Injection, RAG Security, API Security, Responsible AI, Cloud Security, AI Governance, Cybersecurity

I. INTRODUCTION

Artificial Intelligence is transforming enterprise operating models by embedding prediction, natural-language interaction, autonomous reasoning, and intelligent automation into business processes. Generative AI platforms and enterprise agents can summarize documents, generate software, retrieve business knowledge, execute workflows, invoke APIs, and recommend or perform actions. These capabilities create significant productivity and innovation opportunities, but they also create new security risks because AI systems are probabilistic, highly connected, data-intensive, and increasingly autonomous.

Traditional cybersecurity architectures focus on users, applications, networks, databases, and infrastructure. Enterprise AI expands this scope to include foundation models, model endpoints, prompts, system instructions, embeddings, vector databases, retrieval pipelines, agent memory, external tools, model routing, plugins, AI gateways, orchestration frameworks, and machine identities. Security must therefore be redesigned as an end-to-end architecture rather than added as a control after deployment.

II. BUSINESS PROBLEM

Many enterprises are adopting AI through isolated pilots, vendor copilots, public model APIs, custom RAG solutions, and low-code agent platforms. When these initiatives evolve without common security architecture, organizations may create uncontrolled model access, inconsistent data protection, shadow AI usage, excessive permissions, ungoverned plugins, weak prompt controls, and fragmented monitoring.

AI systems also face threats that differ from conventional application attacks. Prompt injection can manipulate model behavior; indirect prompt injection can enter through retrieved documents or websites; poisoned data can influence retrieval or training; model theft can expose intellectual property; over-permissioned agents can misuse enterprise tools; hallucinated outputs can cause unsafe decisions; and malicious users can exploit model interfaces for sensitive data extraction. The business therefore requires a security architecture that protects both the AI technology and the business processes that depend on it.

III. SECURITY OBJECTIVES

The Enterprise AI Security Architecture is designed around confidentiality, integrity, availability, privacy,

accountability, traceability, resilience, and business trust. Security controls should protect sensitive enterprise information, preserve model and prompt integrity, enforce least privilege, prevent unauthorized actions, detect abnormal behavior, provide audit evidence, and support rapid recovery from AI-specific incidents.

Security must also preserve appropriate business usability. Controls that completely block experimentation may drive users toward unsanctioned tools, while weak controls create unacceptable risk. The framework therefore applies risk-based governance and proportional controls based on data sensitivity, model capability, user population, autonomy, business impact, and regulatory significance.

IV. REFERENCE SECURITY ARCHITECTURE

The reference architecture contains several security layers. The Identity Layer authenticates users, developers, workloads, AI agents, service accounts, and machine identities. The Access and Policy Layer enforces authorization, least privilege, segregation of duties, and contextual access. The Data Security Layer protects structured data, documents, embeddings, vectors, logs, and knowledge repositories. The Model Security Layer governs model onboarding, model endpoints, versions, evaluation, and runtime access.

The Prompt and Interaction Security Layer protects system prompts, user prompts, retrieved context, content filters, and output validation. The Integration Security Layer protects APIs, tools, connectors, event channels, and external services. The Runtime Security Layer secures containers, Kubernetes, cloud workloads, secrets, networks, and execution environments. The Observability and Governance Layer provides audit logs, threat detection, policy enforcement, security analytics, incident management, and compliance reporting.

V. IDENTITY AND ACCESS MANAGEMENT

AI security begins with identity. Every human user, application, workload, and autonomous agent should have a verifiable identity. Shared credentials and generic technical users should be avoided because they weaken accountability and incident investigation. Enterprise identity services should support federation, Single Sign-On, OAuth 2.0, OpenID Connect, SAML, workload identity, machine identity, and privileged access management.

Authorization should follow the principle of least privilege. An AI assistant that only summarizes documents should not receive the same permissions as an agent that can create purchase orders or modify customer records. Access should be restricted by role, attribute, data classification, device posture, environment, business context, and risk. Sensitive AI actions should require step-up authentication, approval workflows, or dual control.

VI. DATA AND RAG SECURITY

Retrieval-Augmented Generation connects AI models to enterprise knowledge but also introduces security dependencies on source data, document permissions, metadata, vector stores, and retrieval logic. RAG security should enforce source-level authorization, document classification, data minimization, encryption, tenant isolation, and retrieval-time access checks. The model should receive only the context that the requesting user or agent is authorized to access.

Vector databases and embeddings must be treated as sensitive enterprise assets because embeddings may reveal semantic information even when the original document is not directly exposed. Indexes require encryption, access control, backup, retention, lifecycle management, and monitoring. Retrieval pipelines should validate document provenance and freshness to reduce the risk of poisoned or outdated information influencing model responses.

VII. PROMPT AND CONTENT SECURITY

Prompt security is a critical control domain for Generative AI. System prompts, policy prompts, user prompts, retrieved content, and tool instructions can all influence model behavior. Enterprises should protect system prompts from unauthorized disclosure or modification and validate user input for malicious instructions, sensitive data, disallowed content, and prompt injection patterns.

Indirect prompt injection is particularly important because malicious instructions may be embedded in documents, emails, websites, or data retrieved by the model. Defensive architecture should isolate untrusted content, distinguish data from instructions, apply allow-listed tool behavior, restrict privileged actions, and use policy enforcement outside the model. Output controls should inspect responses for confidential data, unsafe content, unsupported claims, and policy violations before results reach downstream systems.

VIII. MODEL SECURITY

Model security governs the provenance, approval, configuration, and runtime use of foundation models, fine-tuned models, embeddings models, and specialized AI services. Enterprises should maintain an approved model catalog containing provider, version, hosting region, licensing constraints, security posture, data handling terms, risk classification, and allowed use cases.

Model endpoints should be protected through authentication, network controls, throttling, logging, and usage limits. Model updates can change behavior, so version changes should trigger evaluation and approval before production use. Organizations should test models for jailbreak resistance, data leakage, hallucination, harmful content, bias, and security resilience. Where appropriate, sensitive workloads may require private endpoints, dedicated tenancy, or self-hosted models.

IX. AGENTIC AI SECURITY

AI agents increase security risk because they can plan tasks, select tools, call APIs, access data, and execute actions. Each production agent should have a distinct machine identity, explicit business owner, defined purpose, approved toolset, bounded permissions, risk tier, and auditable action history. The agent should never inherit broader permissions than required for its specific business function.

Tool access should be mediated through secure gateways and narrow APIs rather than direct administrative access. High-impact actions should use policy checks, human approval, transaction limits, or dual authorization. Multi-agent systems require delegation controls so one agent cannot transfer privileges beyond its own authority. Agent memory should be encrypted, access-controlled, retention-managed, and separated by business context.

X. API AND INTEGRATION SECURITY

Enterprise AI is highly dependent on APIs and integrations. API gateways should enforce authentication, authorization, rate limiting, schema validation, threat protection, token validation, logging, and usage analytics. AI tools and plugins should be registered, versioned, and reviewed before they become available to agents or copilots.

Integration with SAP S/4HANA, SAP SuccessFactors, SAP Ariba, SAP Customer Experience, SAP Business Data Cloud, SAP Datasphere, external SaaS applications, and public cloud services should follow secure integration patterns. Sensitive operations should use service-specific identities, least-privilege scopes, encrypted communication, and continuous monitoring. Event-driven integrations should validate message authenticity and prevent unauthorized event publication.

XI. CLOUD AND PLATFORM SECURITY

AI workloads frequently run across SAP BTP, hyperscaler services, Kubernetes, serverless platforms, and managed AI services. Cloud security should include secure landing zones, network segmentation, private connectivity, security groups, workload identity, container security, image scanning, Infrastructure as Code validation, secret management, vulnerability management, and Cloud Security Posture Management.

Platform engineering can provide secure-by-default templates for AI applications, RAG services, agents, and model gateways. Standardized deployment pipelines reduce misconfiguration and ensure security controls are applied consistently. Production AI services should be separated from development and experimentation environments to reduce exposure and control data movement.

XII. ZERO TRUST FOR ENTERPRISE AI

Zero Trust assumes that no user, workload, agent, model, API, or device is trusted solely because of network

location. Every access request should be authenticated, authorized, encrypted, and continuously evaluated. Zero Trust for AI extends this model to machine-to-machine communication, agent-to-tool calls, model endpoints, vector stores, knowledge repositories, and automated actions.

Continuous verification should consider identity, requested resource, data sensitivity, device posture, behavioral signals, workload status, and transaction risk. High-risk sessions or actions can trigger step-up controls, reduced privileges, human review, or immediate containment.

XIII. DEVSECOPS AND AI SUPPLY CHAIN

Enterprise AI security should be integrated into DevSecOps and MLOps pipelines. Source code scanning, dependency scanning, container image scanning, secret detection, Software Bill of Materials, Infrastructure as Code scanning, policy-as-code, model artifact validation, and automated security testing should be part of continuous integration and deployment.

AI supply-chain risk includes compromised libraries, malicious model artifacts, untrusted datasets, unsafe plugins, and third-party services. Organizations should verify the provenance and integrity of models, datasets, packages, containers, and external AI services. Production promotion should require signed artifacts, approved repositories, evaluation evidence, and change records.

XIV. OBSERVABILITY AND THREAT DETECTION

AI observability should capture user identity, agent identity, model version, prompt metadata, retrieved sources, tool calls, policy decisions, output classifications, latency, cost, exceptions, and security events. Sensitive prompts and outputs may require masking or tokenization, but the organization should preserve sufficient evidence for audit and incident response.

Threat detection should identify unusual prompt patterns, excessive model usage, unauthorized tool attempts, data exfiltration indicators, repeated policy violations, abnormal token consumption, suspicious retrieval behavior, credential misuse, and unexpected agent actions. Security analytics can correlate AI telemetry with identity, API, cloud, and SIEM data to detect complex attacks.

XV. INCIDENT RESPONSE

AI incidents require response playbooks that differ from conventional application incidents. Examples include prompt injection, model misuse, sensitive data disclosure, malicious agent actions, compromised tools, poisoned knowledge sources, model endpoint abuse, and unsafe automated decisions. Incident teams should be able to disable an agent, revoke credentials, block tools,

switch models, quarantine data sources, invalidate indexes, and roll back configurations.

Post-incident analysis should identify whether the failure originated in identity, data, prompts, model behavior, retrieval, tools, integrations, governance, or human oversight. Lessons learned should update policies, test suites, architecture standards, and security controls.

XVI. RESPONSIBLE AI AND COMPLIANCE

Security and Responsible AI are closely connected. A secure AI system can still create business harm if it produces unfair, misleading, unsafe, or unexplainable outcomes. The architecture should therefore integrate security with transparency, fairness, privacy, explainability, accountability, and human oversight.

Governance should align with enterprise risk management, information security management, AI management, privacy requirements, and sector-specific regulation. Evidence should include risk assessments, model evaluations, security reviews, approval records, access reviews, incident logs, and periodic recertification. High-impact AI use cases should receive stronger independent oversight.

XVII. IMPLEMENTATION METHODOLOGY

Implementation begins by establishing an enterprise AI security policy, security architecture principles, approved model catalog, data classification model, AI inventory, and risk-tiering method. Existing AI solutions should be assessed for identity, access, data exposure, model security, prompt security, agent authority, API controls, observability, and incident readiness.

Organizations should then create reusable secure architecture patterns for common use cases such as enterprise chat, RAG, document summarization, coding assistants, analytics copilots, and autonomous agents. Pilot projects should validate controls before broader rollout. Security automation, policy-as-code, centralized AI gateways, and continuous monitoring should be introduced progressively to reduce manual governance overhead.

XVIII. KEY PERFORMANCE AND RISK INDICATORS

Security KPIs can include percentage of AI workloads using approved models, percentage with centralized identity, security review completion, secure deployment lead time, remediation time, and coverage of AI telemetry. KRIs can include prompt injection detections, unauthorized tool attempts, sensitive-data exposure events, excessive privilege findings, unapproved model usage, policy violations, stale credentials, vulnerable AI components, and high-risk agents without current recertification.

Executive reporting should combine security posture with business adoption so leadership can understand whether AI growth is occurring within acceptable risk boundaries.

XIX. INDUSTRY USE CASES

In financial services, Enterprise AI Security Architecture protects customer data, fraud models, credit decision services, AI assistants, and regulated workflows. In healthcare, it protects clinical information, AI-assisted decision support, patient-facing assistants, and research data. In manufacturing, it secures AI-enabled maintenance, production analytics, autonomous operations, and industrial integrations.

In retail and consumer industries, the architecture protects customer profiles, recommendation services, pricing intelligence, and conversational commerce. In the public sector, it supports secure citizen services, document processing, knowledge assistants, and AI-assisted case management while maintaining transparency, privacy, and accountability.

XX. FUTURE TRENDS

Enterprise AI security will increasingly use automated policy enforcement, AI security gateways, confidential computing, private model hosting, continuous red teaming, autonomous threat detection, dynamic authorization, and agent behavior analytics. Security teams will need controls that operate at machine speed because autonomous agents may make large numbers of decisions and tool calls without direct human interaction.

Future architectures will also require stronger cryptographic provenance, content authenticity, model attestation, secure agent identity, and end-to-end traceability across multi-agent ecosystems. Security, AI governance, and enterprise architecture will converge into a unified discipline for governing intelligent digital systems.

XXI. CONCLUSION

Enterprise AI introduces a security landscape that extends beyond conventional application and cloud controls. Models, prompts, RAG pipelines, vector stores, agents, APIs, tools, identities, data, and orchestration layers must be protected as an integrated system. Security must therefore be designed into the AI lifecycle from architecture and data preparation through deployment, runtime operations, monitoring, and retirement.

The proposed Enterprise AI Security Architecture provides a comprehensive framework that combines Zero Trust, IAM, data protection, prompt security, model security, agent governance, API security, DevSecOps, observability, incident response, and Responsible AI. By applying these controls consistently, organizations can scale Generative AI and agentic AI while protecting sensitive information, reducing cyber risk, improving compliance, and maintaining enterprise trust.

REFERENCES

- [1] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.

- [2] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
- [3] ISO 31000:2018, Risk management — Guidelines.
- [4] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
- [5] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.
- [6] National Institute of Standards and Technology, SP 800-207: Zero Trust Architecture, 2020.
- [7] National Institute of Standards and Technology, Cybersecurity Framework (CSF 2.0), 2024.
- [8] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
- [9] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
- [10] Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing.
- [11] Cloud Native Computing Foundation, Kubernetes Security and Cloud Native Security guidance.
- [12] The Open Group, TOGAF Standard, 10th Edition, 2022.
- [13] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
- [14] SAP SE, SAP Business Technology Platform Documentation.
- [15] SAP SE, SAP AI Core and SAP AI Launchpad Documentation.
- [16] SAP SE, SAP Integration Suite Documentation.
- [17] SAP SE, SAP HANA Cloud Documentation.
- [18] SAP SE, SAP Cloud Identity Services Documentation.
- [19] SAP SE, SAP Cloud ALM Documentation.
- [20] SAP SE, SAP Business Data Cloud and SAP Datasphere Documentation.
- [21] SAP SE, SAP LeanIX Enterprise Architecture Documentation.
- [22] SAP SE, SAP Signavio Process Transformation Suite Documentation.
- [23] DAMA International, DAMA-DMBOK2: Data Management Body of Knowledge, 2017.
- [24] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.
- [25] OECD, Recommendation of the Council on Artificial Intelligence.

AUTHOR BIOGRAPHY

Sanjeeve Kumar Gajadi is an SAP Enterprise Architect and Lead Consultant with extensive experience in SAP BTP, enterprise architecture, data and analytics, integration, cloud architecture, security, governance, and AI-enabled transformation. His professional interests include Enterprise AI Security Architecture, Responsible AI, SAP BTP, Zero Trust, cloud security, enterprise data platforms, agentic AI, and digital transformation.

ACKNOWLEDGMENT

The author acknowledges the contributions of international standards bodies, cybersecurity communities, enterprise architecture practitioners, and SAP documentation teams whose published guidance has informed the security principles synthesized in this paper.