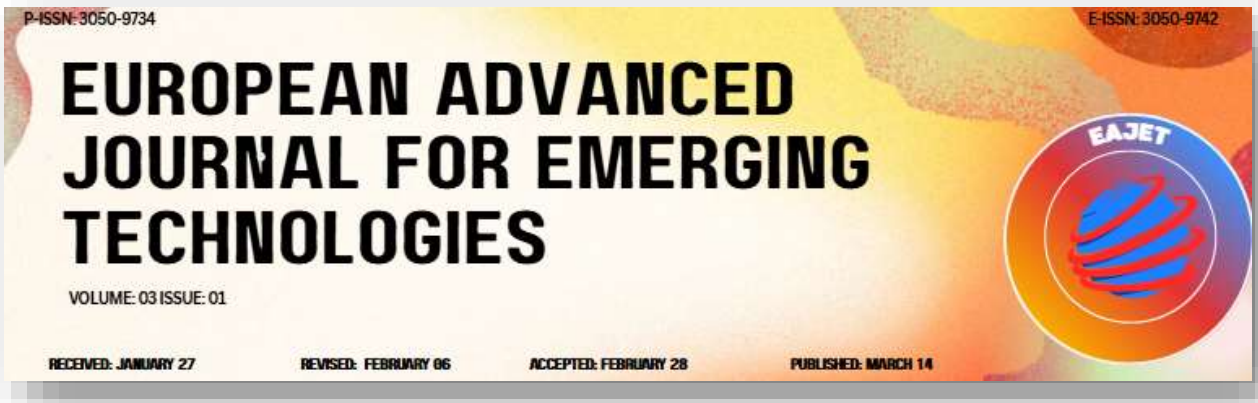


Secure Cloud-Native Architectures for Generative AI in Enterprise Workflows

Yannis Ioannidis
Asst Professor

Abstract

Generative artificial intelligence (GenAI) is a rapidly developing technology with potential for multiple applications, yet it is complex, resource-intensive, and prone to risks. Deployment of GenAI in enterprise workflow platforms requires approaches that enable secure operation while maintaining availability, reliability, and quality. A synthesis of cloud-native architectural patterns with contemporary risk frameworks provides insights into essential security aspects for GenAI. Findings indicate that careful consideration of the safeguards available for prompt injection mitigation, model inversion protection, and data privacy when developing GenAI within a service-mesh architecture can reduce the likelihood of future attack success or damage.

Cloud-native generative artificial intelligence (GenAI) deployment in enterprise workflow platforms is increasingly common, especially for support documentation creation. However, safeguarding the system against attacks that target the availability, reliability, or data privacy of the service remains challenging. Leveraging cloud-native patterns of scalability, resilience, composability, portability, and observability can guide security-enhancing measures. Mapping the security-by-design concept to GenAI services within a service-mesh architecture identifies a range of security controls rooted in established identity and access management principles, the concept of security through obscurity, the defense-in-depth principle, and the auditing of logs and monitoring alerts.

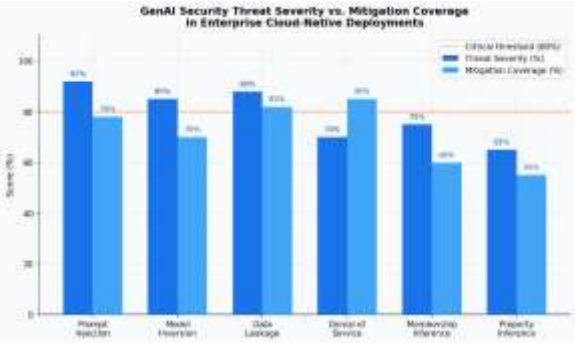
Keywords : Cloud-Native AI Infrastructure,Secure Generative AI Deployment,Enterprise Workflow Automation,Deep Learning Architecture Design,Kubernetes-Based AI Orchestration,AI Model Security and Governance,Scalable Generative AI Systems,MLOps for Enterprise AI Platforms,Zero-Trust AI Deployment Frameworks,Containerized Deep Learning Pipelines.

1. Introduction

Generative artificial intelligence (GenAI) enables strikingly realistic smart creators that can enrich enterprise workflows across engineering, research, marketing, sales, operations, and support. With GenAI adoption rapidly rising, deploying these technologies securely and responsibly has become paramount. However, enterprise deployment of these cloud-native workloads requires special attention to security—and especially privacy. Indeed, widespread, unregulated production usage of GenAI has raised issues including prompt injection, data leaks, hallucinations, model inversion, and biased, factually incorrect, and toxic outputs.

Represented as cloud-native architectures, the insights aid in in-depth scrutiny of the security posture surrounding a cloud-native GenAI capability, whether the platform is internal or external, self-managed or third-party owned. Examination of the use of these vast language models as part of the broader workflow in the cloud highlights the pros and cons of a microservice-based architecture compared to a more monolithic one. Security is best achieved when the flavour of internal GenAI deployment incorporates a security-by-design approach, whereby security considerations are mapped to the application using established cloud-native controls. In particular, the potential for model inversion must be addressed by including inspected parameterized copies of the training data, marking such content for deletion after model.

P-ISSN: 3050-9734
E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14



Grouped Bar — Security Threat Severity vs. Mitigation Coverage

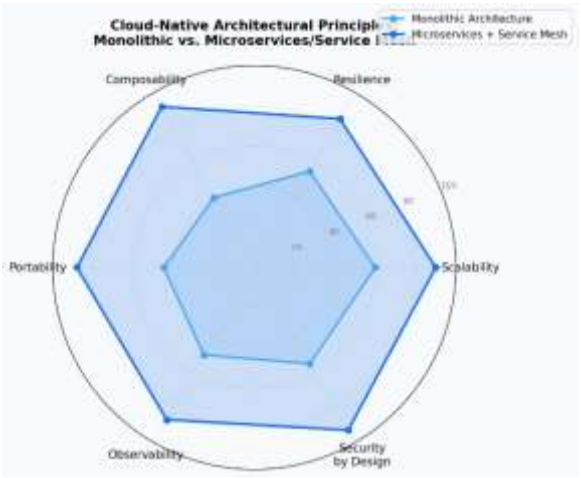


Flow Chart 1: Secure Cloud-Native GenAI Deployment

2. Background and Fundamentals

Generative artificial intelligence (GenAI) is an emerging subfield of deep neural networks capable of generating content from minimal input. A plethora of systems and applications based on GenAI have appeared in recent months, enabling users to specify large text and image-based requests and receive coherent, creative results. Yet, the risk and safety landscape for GenAI models remains poorly understood. Uncontrolled access may enable prompt injection, content generation of any kind, and, if the models are made available as a REST API, they may reveal sensitive training data via model inversion attacks. Without appropriate safeguards, GenAI deployments also risk being indiscriminately utilized for a variety of purposes—including illegal and immoral activities—leading to reputational, legal, and financial harm. The wished-for benefits may thus descend into a nightmare scenario.

The threats posed by GenAI can be approached via a combination of threat modeling, application of security mitigations during the development lifecycle, and integration with native cloud security controls such as identity and access management, micro-segmentation, managed detection and response, and logging and audit capabilities. The design and operation of GenAI in production environments thus requires careful adaptation of security best practices, safeguards based on the expected threat agents, and inclusion of security requirements during GenAI development and tuning.



Radar / Spider — Cloud-Native Principles: Monolithic vs. Microservices/Service Mesh

2.1. Safeguards Against Prompt Injection and Model Inversion

Prompt injection and model inversion rank among the most pressing GenAI security threats. Prompt injection risks arise from malicious input crafted to yield harmful model responses not originally intended by the user. Resilience hinges on deploying prompt-hardening procedures. These protective measures modify user-supplied prompts and associated metadata, focusing on application payloads, context data, system prompts, and model configuration.

P-ISSN: 3050-9734
E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

Denial of Service can cripple GenAI operation, spotlighting the criticality of data validation and sanitization. Foundations such as OpenAI's API and the Safety and Robustness Externalization Layer recommend standard countermeasures. Model inversion allows adversaries to elicit training corpus attributes by analyzing model output. Implementing prompt-hardening capabilities, rejecting toxic content, and augmenting training data with noise enhance GenAI security without overly burdening performance. The microservice exposé depicted in Figure [insert reference] confirms the potential for systematic defense-in-depth architecture.

Architectural Principle	Purpose in GenAI Deployment	Enterprise Benefit	Security Contribution
Scalability	Handle dynamic AI workloads	Supports enterprise growth	Prevents service overload
Resilience	Maintain uptime during failures	Improves reliability	Reduces attack impact
Composability	Modular AI component integration	Faster service updates	Enables isolated security controls
Portability	Deploy across multi-cloud platforms	Vendor flexibility	Secure workload migration
Observability	Monitor AI behavior and traffic	Operational transparency	Enables anomaly detection

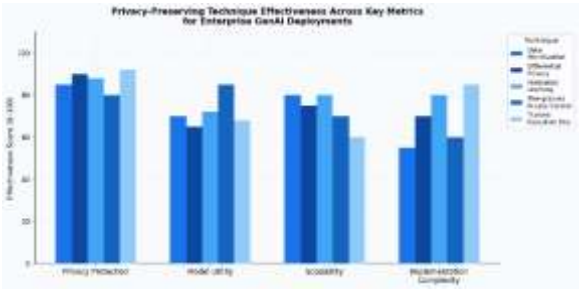
Table 1: Cloud-Native Architectural Principles for Enterprise GenAI



Flow Chart 2: Microservices and Service Mesh Architecture

3. Cloud-Native Architectural Principles for Generative AI

The development and deployment of Generative AI within enterprises are strongly aligned with major cloud-native architectural principles. Scalability and resilience are fundamental to addressing demand fluctuations, while the composability of Generative models naturally supports tiered service implementation. Sensitivity to costs and risks associated with overly expansive service deployments suggests that a monolithic approach is more suitable for initial usage patterns and overwhelming workloads, yet microservices serve as a pathway to future evolution. Composability also implies that key components, such as prompt-steering, prompt-hardening, content-filtering, and watermarking, are candidates for independent service development. Portability becomes prominent when these supporting elements are repurposed external to enterprise boundaries. Observability ensures that potential operational failures and malicious misuse or attack vectors are quickly detected and acted upon.



Multi-Series Bar — Privacy-Preserving Technique Effectiveness (Differential Privacy, Federated Learning, etc.)

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

3.1. Microservices and Service Mesh for GenAI Components

Generative AI capabilities map well to microservices and service mesh patterns, enabling a security architecture compliant with the defense-in-depth principle of cloud-native computing.

Built-from-scratch components and third-party providers can jointly fulfil enterprise requirements. For example, a content moderation microservice can add guardrails for image generation, and natural language understanding (NLU) can provide a query language compatible with a multimodal approach across text, image, and video creation services. A service mesh can enforce security, such as access control, traffic management, encryption using mutual Transport Layer Security (mTLS), and compliance with policy frameworks. Sidecar proxy instances fulfill these tasks without business logic hidden in the application code itself but instead in dedicated, configurable components of the service mesh. The control plane of the mesh manages a set of service definitions that, among other characteristics, specify the environment in which sidecars act.

Security Threat	Description	Potential Impact	Mitigation Technique
Prompt Injection	Malicious prompts manipulate model behavior	Harmful AI responses	Prompt hardening
Model Inversion	Extraction of training data from outputs	Sensitive data leakage	Differential privacy
Data Leakage	Exposure of confidential enterprise data	Compliance violations	Data minimization
Toxic Content Generation	AI produces unsafe outputs	Reputational damage	Content filtering
Denial of Service	Resource exhaustion attacks	Service disruption	Validation and sanitization

Table 2: Security Threats and Mitigation Techniques in Generative AI

Service-based architecture requires special attention to monitoring and observability because, as traffic patterns shift, new services come online or offline. Cloud environments typically provide dynamic dashboards, but service meshes go a step further by introducing service discovery, enabling automatic profiling of traffic flows using a classic data plane/control plane approach. Multi-cloud deployment of distinct services is often a business requirement, so assuring that packet content travelling across untrusted networks remains confidential requires coupling the service-mesh approach with secure policies of the external cloud providers.

Mathematical Formulas:

1. Deep Learning Loss Function

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

2. Secure Model Optimization

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

3. Generative AI Probability Distribution

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_{1:i-1})$$

4. Differential Privacy Mechanism

$$M(D) = f(D) + \mathcal{N}(0, \sigma^2)$$

5. Federated Learning Aggregation

$$W_{global} = \sum_{k=1}^{K} \frac{n_k}{n} W_k$$

6. Zero-Trust Access Validation

$$A(u,r) = \begin{cases} 1, & \text{if } Auth(u) \wedge Policy(r) \\ 0, & \text{otherwise} \end{cases}$$

7. Kubernetes Resource Scaling Equation

$$R_{scale} = \frac{C_{load}}{C_{node}}$$

8. Secure Service Mesh Communication

$$E_{msg} = Enc_{TLS}(Data, Key)$$

9. AI Inference Latency Model

$$T_{total} = T_{compute} + T_{network} + T_{storage}$$

10. Enterprise Workflow Automation Efficiency

$$E_w = \frac{T_{manual} - T_{AI}}{T_{manual}} \times 100$$

11. Model Inversion Risk Estimation

$$R_{inv} = P(D_{leak} \mid M_{output})$$

12. Cloud-Native Reliability Equation

$$R(t) = e^{-\lambda t}$$

13. Observability Metric Function

$$O_s = Logs + Metrics + Traces$$

14. Containerized AI Throughput

$$Throughput = \frac{Requests}{Second}$$

15. RBAC Authorization Model

$$Perm(u) = \bigcup_{r \in Roles(u)} Permissions(r)$$

16. Secure Data Minimization Function

$$D_{min} = D - D_{sensitive}$$

17. AI Governance Compliance Score

$$G_c = \frac{Controls_{implemented}}{Controls_{required}} \times 100$$

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

18. Service Mesh Traffic Routing

$$T_r = \sum_{i=1}^{n} w_i \, p_i$$

19. SIEM Threat Detection Probability

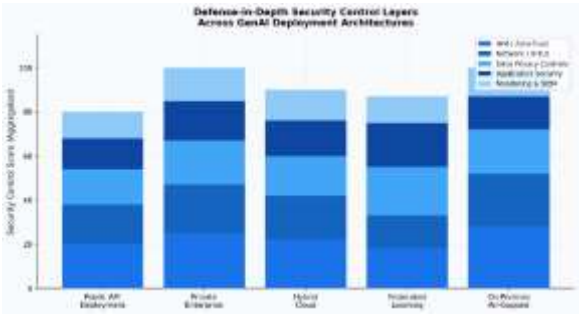
$$P(T_d) = 1 - \prod_{i=1}^{n} (1 - p_i)$$

20. Cloud-Native AI Availability

$$A = \frac{Uptime}{Uptime + Downtime}$$

4. Security by Design in Generative AI Deployments

Securing Generative AI frameworks in enterprise-grade systems requires an extended security-by-design process encompassing cloud-native controls, threat modeling, secure software development, and integration of security and functional requirements. Security challenges and mitigations should be documented in all project artifacts, forming the foundation of deployment-specific support documentation. Effective security requires depth across development, orchestration, and hosting. Brake glass methods for mitigation of operational failures are planed, supporting migration toward rich controls.



Stacked Bar — Defense-in-Depth Security Control Layers across Deployment Types



Flow Chart 3: Privacy-Preserving GenAI Workflow

4.1. Authentication, Authorization, and Least Privilege

Identity and access management (IAM) provides mechanisms for authentication and authorization of user requests and inherent observations. Zero-trust principles mitigate the inherent trustworthiness of entities within the information system perimeter by requiring rigorous validation of requests before granting permission to enterprise resources. Fine-grained access control further constrains resource accessibility, ensuring that entities can access only those resources crucial for task execution and to the least extent required. This principle is realized via role-based access control (RBAC) models that use a limited and manageable set of role definitions to enforce permissions or by attribute-based access control (ABAC) that evaluates a broader range of attributes across any resource or user. Auditing, logging, and monitoring requirements demand that requests, responses, and access patterns be logged for potential misuse detection, enable pattern monitoring for risk identification, guarantee system evolution traceability, and satisfy regulatory obligations.

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

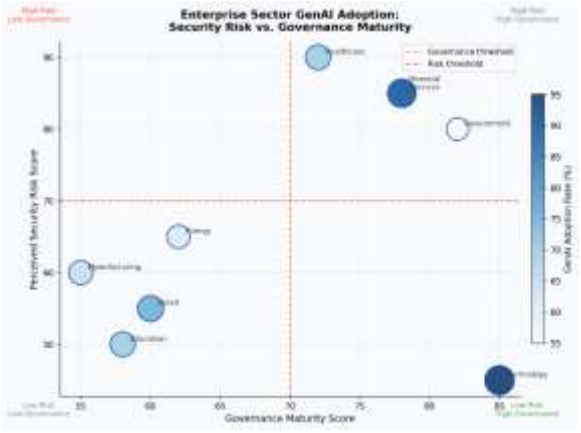
Component	Function	Cloud-Native Role	Security Capability
Sidecar Proxy	Handles service communication	Service mesh integration	mTLS encryption
Content Moderation Service	Filters unsafe AI outputs	AI governance	Toxicity prevention
Prompt Hardening Service	Sanitizes user prompts	Input protection	Injection defense
Traffic Manager	Controls service routing	Load balancing	Secure traffic policies
Control Plane	Manages service policies	Configuration management	Centralized governance

Table 3: Microservices and Service Mesh Components for GenAI

5. Privacy-Preserving Techniques for Enterprise GenAI

Generative AI has the potential to produce very sensitive data; thus it is important to implement protection of the data while still providing for its use in such capabilities. Establishing access restrictions and controls, as well as limiting the volume of data used by generative models, can reduce exposure. Processing private data in such a way that the data is not revealed outside of a trusted system also helps reduce the risk. Processing such sensitive data in a federated way, where models are trained at the location of the data but remain at the data owner's site, is becoming the preferred way of providing such services when using trusted data sets. Such an approach is also important for meeting regulatory compliance requirements concerning personally identifiable information.

Generating synthetic data in violation of privacy obligations can lead organizations to discrimination and other risks apart from financial penalties. Thus, these risks must be mitigated in the design and use of enterprise Generative AI services, through privacy-preserving techniques applied to the data, the models, and its access during inference. Such techniques include data minimization, access controls, and processing in trusted environments. Protection of sensitive data in the training process can also be accomplished through Differential Privacy and Federated Learning algorithms. In the first case, sensitive information is protected through the use of noise which acts like a barrier when the information contained in the model is attempted to be analysed or extracted. In the second case, models are trained at source locations by making use of the data but these models do not get out of their locations, hence the sensitive information remains within the trusted locations.



Bubble Scatter — Enterprise Sector Risk vs. Governance Maturity (bubble = adoption rate)

5.1. Differential Privacy and Federated Learning

Ensuring privacy during the deployment of GenAI technologies requires careful management of the data shared with the model and the predictions generated. Approaches such as data minimization, fine-grained access control, and

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

enabling the processing of Confidential Information in trusted environments help reduce privacy exposure but cannot eliminate the risk entirely. These techniques need to be complemented by safeguards applied to the model's training and inference processes. The main objective is to ensure that no sensitive information is leaked from either the training or the prediction phase. The joint modeling of training and inference risks offers a coherent framework for defining and investigating these measures.

Differential privacy provides a mathematically rigorous way of quantifying and controlling inference risks in supervised machine-learning workflows. DP is usually defined for a specific algorithm, describing a postprocessing requirement that must be satisfied for any two datasets differing in a single input. An attacker with knowledge of one of the datasets will have similar information about the outputs of the different algorithms, regardless of how much they know about the rest of the input data. A model with differential privacy guarantees is trained using data fulfilling a specific condition, and any attack relying on that condition will not yield a significant advantage. An alternative paradigm for risk mitigation is federated learning, where the model is hosted by an external party, and the data never leaves the data owners' environment. A trusted party provides an infrastructure for executing model training on behalf of data owners under a clear and controlled service level agreement.

Security Mechanism	Description	Enterprise Usage	Benefit
RBAC	Role-based permissions	Department-level access	Simplified access control
ABAC	Attribute-driven authorization	Dynamic policy enforcement	Fine-grained security
IAM	Identity management framework	User authentication	Centralized credential control
Zero-Trust Validation	Continuous verification	Secure AI service access	Minimizes insider threats
Audit Logging	Tracks user actions	Compliance monitoring	Incident traceability

Table 4: Authentication and Authorization Mechanisms



Flow Chart 4: Governance, Risk, and Compliance Flow

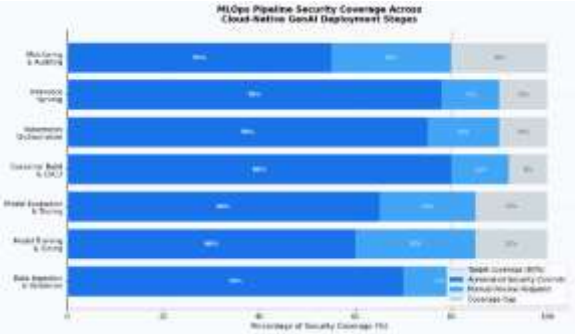
6. Governance, Risk, and Compliance in Cloud-Native GenAI

Establishing appropriate governance structures helps ensure that Generative AI systems meet stakeholder expectations for risk and impact through different decision-making processes aligned with the enterprise risk appetite. Governance requirements drive the design and implementation of risk assessment methodologies to evaluate the potential consequences of the use of such technologies, the breadth and detail of supporting policy frameworks, and the influence of applicable regulations and industry standards. In many enterprises, formal risk appetite statements indicate how willing the organization is to incur a specific type of risk, thus determining the level of necessary controls. The cloud-native environment provides a technological foundation that can assist in delivering a secure GenAI implementation. Integrating risk and compliance into the use of Generative AI enables security controls to be specified and implemented at a level informed by risk.

A comprehensive enterprise approach to governance, risk, and compliance across its cloud-native environment,

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

including operations for GenAI, thus requires the establishment of these additional tiers as part of compliance with ISO 27001 and ISO 27001:2013, together with supporting audit reporting aligned with ISAE 3000.



Horizontal Stacked Bar — MLOps Pipeline Security Coverage (automated / manual / gap)

6.1. Auditing, Logging, and Monitoring

Traceability is central to safeguarding GenAI. Center for Internet Security (CIS) recommendations for Detecting Indicators of Compromise (IOC) emphasize the central role of logs for forensics, detection, attribution, and remediation of breaches. Relevant logging requirements cover integrity, accessibility, availability, and expedition; tamper-evident guarantees preserve evidence when writing large volumes of data. Model Inversion and Data Leakage Attacks against Generative Neural Networks show that prompt injections can produce fine-tuning datasets that extract model internals; such entries must be detectable as they traverse entire workflows.

Tampering with log entries and alerts must be detectable after the fact. Auditing ensures that bypassing logging alerts for staging/tuning flows is detectable and attributes responsibility to user credentials. Anomaly detection detects IOCs and APTs. Tamper-proof anomaly logs are generated by a GenAI telemetry pipeline and linked to related incident response workflows. The logging model supports both machine analysis for IOCs and APTs and human investigation, where specific users have access twists for the least amount of time.

Tamper-proof logging and anomaly detects for Data Leakage Through Sensitive Parameters of Deep Generative Models enable cross-referencing with neural-network training and diagnostic tools, supporting feature-value-class association for stages of operations. Anomaly detection pipelines, integrated with Security Information and Event Management (SIEM) and Observability, correlate model-inversion vulnerabilities in Language Models with traffic patterns externalizing sensitive information via web-application firewall and proxy-logs. GenAI therefore complies with relevant security requirements and is integrated with cloud-native controls for auditing, anomaly detection, and forensics.

Technique	Core Function	Privacy Benefit	Enterprise Application
Differential Privacy	Adds noise to data	Prevents inference attacks	Secure model training
Federated Learning	Keeps data local	Maintains data sovereignty	Distributed AI collaboration
Data Minimization	Limits data exposure	Reduces leakage risk	Regulatory compliance
Trusted Environments	Secure processing zones	Protects confidential information	Enterprise AI hosting
Fine-Grained Access Control	Restricts sensitive access	Improves privacy governance	Secure AI inference

Table 5: Privacy-Preserving Techniques for Enterprise GenAI



7. Conclusion

In summary, secure adoption of Generative AI in Enterprise Workflow Platforms relies on the combination of applied suggested measures with the control and compliance capabilities embedded within Cloud-Native technology and principles. The proposed Patterns of Cloud-Native-Architecture-Aware-Governance cover the key aspects of security-by-design principles, Defence-in-Depth controls, least-privilege using Authentication and Authorisation methods, privacy-preserving techniques, Governance, Risk and Compliance measures, Auditing Logging and Monitoring together with a ready-to-use, practical example for risk-proof Decision Support System—Data understandability in understandable domain.

GenAI cloud-native architecture patterns and trade-offs decisions findings provide a foundation for advancing research in safely scale Generative AI across enterprise use cases to fulfil the high ambition set forth by the GenAI community of making it "the new electricity"—not only of the consumer world but of the enterprise world too.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2022). On the opportunities and risks of foundation models. arXiv.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2022). Language models are few-shot learners. Communications of the ACM, 65(4), 126–135.
3. Radha, S., Gottimukkala, V. R. R., Thottara, S., Vandhana, K., & J, Gokulraj. (2025). Adaptive Video Streaming Over 5G Networks Using Deep Reinforcement Learning with Closed-Loop Feedback Mechanism for Bitrate Control. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–6). IEEE. 2025 International Conference on Communication, Computer, and Information Technology (IC3IT). https://doi.org/10.1109/ic3it66137.2025.11341184
4. Chiang, W. L., Zheng, L., Sheng, Y., Li, T., Zhuang, Z., Wu, Y., ... Stoica, I. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality. arXiv.
5. Ding, Y., Qin, X., Wei, Z., & Choi, B. Y. (2023). Cloud-native computing: Architecture, security, and applications. IEEE Access, 11, 109215–109236.
6. Kummari, D. N., Burugulla, J. K. R., Malempati, M., Amistapuram, K., Garapati, R. S., & Nagabhyru, K. C. (2025, December). Enhancing Audit Compliance and Operational Efficiency in Manufacturing and Commercial Insurance Through Agentic AI and Data Engineering Frameworks. In 2025 IEEE International

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

Conference on Communication Networks and Computing (CNC) (pp. 714-720). IEEE.

7. Dong, X., Li, Y., Sun, H., Zhang, C., & Chen, J. (2024). Secure deployment architectures for generative AI services in cloud environments. Future Internet, 16(4), 112.
8. Guo, Y., Li, X., Wang, H., & Zhang, Y. (2024). Zero-trust security architecture for cloud-native AI platforms. Future Internet, 16(5), 154.
9. He, Y., Bian, S., Chen, L., Hui, Y., Lentz, M., Li, B., ... Zhuo, D. (2024). Computing in the era of large generative models: From cloud-native to AI-native. arXiv.
10. Kolla, T. (2025). Generative AI for Intelligent Medical Coding and Healthcare Analytics. International Journal of Advanced Research in Computer Science & Technology (IJARCST), 8(6), 13285-13299.
11. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D., ... Sayed, W. (2024). Mixtral of experts. arXiv.
12. Kandpal, N., Deng, H., Roberts, A., Wallace, E., & McCallum, A. (2023). Large language models struggle to learn long-tail knowledge. Proceedings of the 40th International Conference on Machine Learning.
13. Inala, R., Kaulwar, P. K., Nagabhyru, K. C., Adusupalli, B., & Arun Raj, S. R. (2025, October). Leveraging IEC 61850 for Interoperable and Resilient Smart Grid Communication Architecture. In International Conference on Microelectronics, Electromagnetics and Telecommunication (pp. 549-566). Cham: Springer Nature Switzerland.
14. Karpathy, A. (2023). Software 2.0 and the rise of AI-native application development. arXiv.
15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Riedel, S. (2022). Retrieval-augmented generation for knowledge-intensive NLP tasks. Journal of Machine Learning Research, 23, 1–48.
16. Amistapuram, K., Pandiri, L., Raju, V. R., Paleti, S., Singireddy, S., & Sheelam, G. K. (2025). AI-Based Cloud Infrastructure and MLOps Frameworks for Scalable Data Engineering Across Banking and Insurance. In 2025 IEEE International Conference on Communication Networks and Computing (CNC) (pp. 186–192). IEEE. 2025 IEEE International Conference on Communication Networks and Computing (CNC).

P-ISSN: 3050-9734
E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

https://doi.org/10.1109/cnc68716.2025.11484532

17. Li, X., Zhang, Y., Wang, Z., & Chen, H. (2023). Kubernetes-based orchestration for scalable AI microservices. IEEE Access, 11, 88431–88447.
18. Lu, Y., Bian, S., Chen, L., He, Y., Hui, Y., Li, B., ... Zhuo, D. (2024). AI-native computing: Challenges and opportunities for cloud infrastructure. arXiv.
19. OpenAI. (2023). GPT-4 technical report. arXiv.
20. Sudhakar, A. V. V., Inala, R., Verma, A. K., Nag, K., Pandey, V., & Anand, P. S. (2025, September). Hybrid Rule-Based and Machine Learning Framework for Embedding Anti-Discrimination Law in Automated Decision Systems. In 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT) (pp. 1-6). IEEE.
21. Patel, D., Raut, G., Cheetirala, S. N., Nadkarni, G. N., Freeman, R., Glicksberg, B. S., ... Klang, E. (2024). Cloud platforms for developing generative AI solutions: A scoping review of tools and services. arXiv.
22. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv.
23. Wang, Z., Chen, H., Li, X., & Zhang, Y. (2023). Secure microservice architectures for enterprise cloud applications. IEEE Access, 11, 97624–97641.
24. Loganathan, R. (2025). AGENTIC AI FRAMEWORKS FOR AUTONOMOUS RISK DETECTION AND COMPLIANCE REMEDIATION IN ENTERPRISE DATA CENTER OPERATIONS. Lex Localis-Journal of Local Self-Government, 23 (S6), 9672–9697.
25. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Zhou, D. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.
26. Zhang, C., Liu, J., Wang, H., & Sun, X. (2024). Secure cloud-native deployment framework for enterprise generative AI applications. Journal of Cloud Computing, 13(1), 58.
27. Nigam, N., Sireesha, B., Ediga, P., Segireddy, A. R., & Bokde, S. (2025, December). Comparative Evaluation of Cloud Security Algorithms Using Multiple Classifiers with an Optimized Intrusion Detection System. In 2025 IEEE 5th International Conference on ICT in

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

Business Industry & Government (ICTBIG) (pp. 1-6). IEEE.

28. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J. R. (2023). A survey of large language models. arXiv.
29. Deng, S., Zhao, H., Huang, B., Zhang, C., Chen, F., Deng, Y., Yin, J., Dustdar, S., & Zomaya, A. Y. (2024). Cloud-native computing: A survey from the perspective of services. Proceedings of the IEEE, 112(1), 12–46.
30. Marin, E., Perino, D., & Di Pietro, R. (2022). Serverless computing: A security perspective. Journal of Cloud Computing, 11, Article 69.
31. Amistapuram[1], K., Kolla, T., Bandi, V. D. V. K., Kolla[4], S. K., & Rani, P. S. Journal of Rare Cardiovascular Diseases.
32. OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774.
33. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.
34. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971.
35. Singh, H., Bose, D., Nagubandi, A. R., Prabhu, S., & Naik, S. G. Cryptocurrency Market Spillovers: Risk Contagion Across Global Financial Systems.
36. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., et al. (2023). A survey of large language models. arXiv.
37. Lu, Y., Bian, S., Chen, L., He, Y., Hui, Y., Lentz, M., Li, B., et al. (2024). Computing in the era of large generative models: From cloud-native to AI-native. arXiv:2401.12230.
38. Patel, D., Raut, G., Cheetirala, S. N., Nadkarni, G. N., Freeman, R., Glicksberg, B. S., Klang, E., & Timsina, P. (2024). Cloud platforms for developing generative AI solutions: A scoping review of tools and services. arXiv:2412.06044.
39. Rani, P. S., Kummari, D. N., Yellanki, S. K., Meda, R., Koppolu, H. K. R., & Inala, R. (2025, July). Blockchain and AI for Securing Electrical Infrastructure. In 2025 2nd International Conference on Computing and Data Science (ICCDS) (pp. 1-6). IEEE.
40. Ramesh, G., Pai, T. V., Birau, R., Poojary, K. K., Abhay, Shingad, A. R., Sowjanya, N., Popescu, V., Mitroi, A. T., Nioata Chireac, R. M., & Raj, K. M. K. (2025). A comprehensive review on scaling machine

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

learning workflows using cloud technologies and DevOps. IEEE Access.

41. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., et al. (2022). On the opportunities and risks of foundation models. arXiv.
42. Mangalampalli, B. M., Bandi, V. D. V. K., Kolla, S. K., & Kumar, M. V. K. (2025). Towards Self-Evolving Healthcare Intelligence: Integrating Advanced Learning Systems with Real-Time Clinical Data Pipelines. Cultura: International Journal of Philosophy of Culture and Axiology, 22(12s), 464-486.
43. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.
44. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D., et al. (2024). Mixtral of experts. arXiv.
45. Chiang, W.-L., Zheng, L., Sheng, Y., Li, T., Zhuang, Z., Wu, Y., et al. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality. arXiv.
46. Kumar, I., Nagabhyru, K. C., IG, N., MV, P., & KV, S. (2025, October). Adaptive Meta-Knowledge Transfer Network with Feature Hallucination and Attention for Low-Shot Object Detection in Aerial Images. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.
47. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., et al. (2022). Retrieval-augmented generation for knowledge-intensive NLP tasks. Journal of Machine Learning Research, 23, 1–48.
48. Kandpal, N., Deng, H., Roberts, A., Wallace, E., & McCallum, A. (2023). Large language models struggle to learn long-tail knowledge. In Proceedings of the 40th International Conference on Machine Learning.
49. Kolla, S. H., & Mangala, N. (2025). DESIGNING AUTONOMOUS LLM AGENT FRAMEWORKS USING GEN AI PIPELINES TO ENHANCE CUSTOMER SERVICE MANAGEMENT AND KNOWLEDGE WORKFLOWS. Lex Localis-Journal of Local Self-Government, 23, 9719-9733.
50. He, Y., Bian, S., Chen, L., Hui, Y., Lentz, M., Li, B., Liu, F., et al. (2024). AI-native computing for large generative models: Opportunities and challenges. arXiv.

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

51. Golda, A., Chamola, V., Hassija, V., & Sikdar, B. (2024). Security and privacy challenges in generative artificial intelligence: Current trends and future directions. IEEE Access.
52. Nabende, P., & Wanyama, T. (2008). An expert system for diagnosing heavy-duty diesel engine faults. In Advances in computer and information sciences and engineering (pp. 384-389). Dordrecht: Springer Netherlands.
53. Deng, S., Dustdar, S., Yin, J., & Zomaya, A. Y. (2024). Service-oriented cloud-native computing: Architecture, orchestration, and future research directions. Proceedings of the IEEE.
54. Marin, E., Perino, D., & Di Pietro, R. (2022). Security challenges and research opportunities in serverless cloud-native systems. Journal of Cloud Computing.
55. Mangalampalli, Bindu Madhavi, Sasi Kumar Kolla, Velangani Divya Vardhan Kumar Bandi, Uday Surendra Yandamuri, and PR Sudha Rani. "Designing intelligent healthcare ecosystems through adaptive data integration and autonomous learning systems." Vascular and Endovascular Review 8, no. 20s (2025): 330-347.
56. Andreoni, M., Lunardi, W. T., Lawton, G., & Thakkar, S. (2024). Enhancing autonomous system security and resilience with generative AI: A comprehensive survey. IEEE Access, 12, 109470–109493.
57. Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V., & Sikdar, B. (2024). Privacy and security concerns in generative AI: A comprehensive survey. IEEE Access, 12, 48126–48144.
58. Mangalampalli, B. M., & Kolla, S. K. (2025). Large Language Models for Automated Healthcare Data Dictionary Generation and Maintenance. Vascular and Endovascular Review, 8(20s), 363-375.
59. Deng, S., Zhao, H., Huang, B., Zhang, C., Chen, F., Deng, Y., Yin, J., Dustdar, S., & Zomaya, A. Y. (2024). Cloud-native computing: A survey from the perspective of services. Proceedings of the IEEE, 112(1), 12–46.
60. Syed, N. F., Shah, S. W., Shaghaghi, A., Anwar, A., Baig, Z., & Doss, R. (2022). Zero trust architecture (ZTA): A comprehensive survey. IEEE Access, 10, 57143–57179.
61. Ashokkumar, S., & Amistapuram, K. (2025, October). Attention-Guided Spatial Temporal Framework for Deepfake Detection on Social Video Platforms. In 2025 International Conference on

P-ISSN: 3050-9734 E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.

62. Poirrier, A., Cailleux, L., & Clausen, T. H. (2025). Is trust misplaced? A zero-trust survey. Proceedings of the IEEE, 113(1), 5–39.
63. Lu, Y., Bian, S., Chen, L., He, Y., Hui, Y., Lentz, M., Li, B., Liu, F., Li, J., Liu, Q., Liu, R., Liu, X., Ma, L., Rong, K., Wang, J., Wu, Y., Wu, Y., Zhang, H., Zhang, M., Zhang, Q., Zhou, T., & Zhuo, D. (2024). Computing in the era of large generative models: From cloud-native to AI-native. arXiv.
64. Reddy, V. A. R., & Kolla, S. K. (1984). Infrastructure-As-Code Practices For Regulated Healthcare Cloud Environments. Metallurgical and Materials Engineering, 30 (4), 1028–1042.
65. OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774.
66. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.
67. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D., et al. (2024). Mixtral of experts. arXiv:2401.04088.
68. Garapati, R. S., & Kanna, S. R. A Digital Twin-Enabled Predictive Maintenance Framework Leveraging Multi-Agent Reinforcement Learning and Industrial IoT Data.
69. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. arXiv.
70. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2022). On the opportunities and risks of foundation models. arXiv.
71. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.
72. Davuluri, P. N. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems.
73. Chiang, W.-L., Zheng, L., Sheng, Y., Li, T., Zhuang, Z., Wu, Y., et al. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality. arXiv.
74. Weinberg, A. I., & Cohen, K. (2024). Zero trust implementation in the emerging technologies era: Survey. arXiv.
75. Thijsman, J., Sebrechts, M., De Turck, F., & Volckaert, B. (2024). Trusting the

P-ISSN: 3050-9734
E-ISSN: 3050-9742
EUROPEAN ADVANCED JOURNAL FOR EMERGING TECHNOLOGIES
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 27 REVISED: FEBRUARY 06 ACCEPTED: FEBRUARY 28 PUBLISHED: MARCH 14

cloud-native edge: Remotely attested Kubernetes workers. arXiv.

76. LEBCIR, I., Shah, C. A., & Appa Rao Nagubandi, D. S. M. D. FinTech and Financial Inclusion: Empirical Evidence from Emerging Markets.
77. Nahar, N., Andersson, K., Schelen, O., & Saguna, S. (2024). A survey on zero trust architecture: Applications and challenges of 6G networks. IEEE Access.
78. Bandi, V. D. V. K. AI-Based Anomaly Detection Frameworks in Distributed Enterprise Data Systems.
79. Patel, D., Raut, G., Cheetirala, S. N., Nadkarni, G. N., Freeman, R., Glicksberg, B. S., Klang, E., & Timsina, P. (2024). Cloud platforms for developing generative AI solutions: A scoping review of tools and services. arXiv.
80. Marin, E., Perino, D., & Di Pietro, R. (2022). Serverless computing: A security perspective. Journal of Cloud Computing, 11, Article 69.
81. GARAPATI, R. S. SYNERGETIC INTELLIGENCE Converging AI, Cloud, IoT, and Smart Automation for Real-Time Futures. CANEDA GLOBAL JOURNAL GROUP.
82. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2022). Retrieval-augmented generation for knowledge-intensive NLP tasks. Journal of Machine Learning Research, 23, 1–48.
83. Kandpal, N., Deng, H., Roberts, A., Wallace, E., & McCallum, A. (2023). Large language models struggle to learn long-tail knowledge. In Proceedings of the 40th International Conference on Machine Learning.
84. Reddy, V. A. R. (2025). Journal of Rare Cardiovascular Diseases. Health, 5(3), 402-422.