JGTAI Journal of Global Trends in Artificial Intelligence

Volume 1, No. 1, June 2025-2026
Available online at: www.jgtai.com

A Review on Architectures and Techniques of Generative AI Frameworks for Intelligent Cloud Operations

Dr. Prashant Kumar Srivastava
PhD CSE
Associate Professor
Sanjeev Agrawal Global Educational (SAGE) University
SOCT
Bhopal
prashant.s@sageuniversity.edu.in

***Abstract*—Cloud computing has become one of the most important tools for making computing services scalable, flexible, and affordable. But, there are significant obstacles in the areas of operational automation, security, monitoring, and fault diagnosis due to the increasing complexity of cloud infrastructures, dynamic workloads, multi-cloud settings, and remote applications. Automated decision support, natural language interaction, and knowledge-driven automation are all made possible by new possibilities made possible by advances in Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs), which have created new paths for intelligent cloud operations. Here, takes a close look at the Generative AI frameworks' intelligent cloud operations and analyzes their structures and approaches in detail. Afterward, the fundamentals of cloud operations are covered, along with the current obstacles, Generative AI, core foundation models, and LLMs. Following that, it delves into a number of noteworthy generative architectures, including Transformers, GANs, VAEs, and Diffusion Models, along with well-known frameworks such as LangChain, LlamaIndex, AutoGen, CrewAI, and Semantic Kernel. Agentic AI, Retrieval-Augmented Generation (RAG), and Prompt Engineering are among the enabling techniques that are covered. Finally, the literature review highlights the present research trends, gaps, and potential future directions for autonomous, secure, and intelligent cloud administration.**

Keywords—Cloud Computing, Intelligent Cloud Operations, Generative Artificial Intelligence (GenAI), Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Agentic AI.

I. Introduction

Cloud computing is a shared computing environment that provides a variety of computing technologies and services through an on-demand model [1][2]. Cloud computing enables users to access and make use of shared computing resources—hardware, software, and networking—through the Internet. It provides infrastructure, platform, and application services to its customers. Cloud computing has changed the landscape of traditional IT by offering scalable software services and significant computing resources that are available at any time and anywhere based on the needs of educational institutions and other organizations [3][4][5]. In the cloud computing paradigm, the user is able to use services of high quality at a reduced price. Cloud computing has become a hot topic in both the academic and industrial community for various technical and economic benefits [6][7]. Industries can optimize workflows and boost decision-making processes with its capacity to consolidate operations and deliver on-demand resources. Cloud computing can also cut down costs and help organizations benefit from advanced technological capabilities while shifting away from extensive on-premises infrastructure [8]. But the complexity of cloud operations has escalated with the proliferation of cloud infrastructures, multi-cloud deployments, distributed applications, and dynamic workloads. As a result, intelligent cloud operations have become an effective method of automating infrastructure management, resource provisioning, system monitoring, incident response, security management, and performance optimization.

GenAI is the term given to the use of AI to produce new forms of content, including text, images, sounds or videos. Traditional or discriminative AI models try to learn a functional mapping for classification or regression. GenAI models, on the other hand, learn the underlying probability distribution of training data to create new content that is similar to the original data but also different [9][7][10]. The rise of GenAI has taken a major step beyond the capabilities of predictive AI, allowing for not just content automation and analytical decision-making, but intelligent content creation, reasoning, and problem-solving [11].

The next generation of AI is generative AI, powered by foundation models and LLMs, which can generate text, images, audio, 3D solutions and videos using natural language prompts [12]. The progress represents a major leap away from pattern recognition AI towards context-aware and human-like AI outputs. The combination of Generative AI and cloud operations can drive intelligent automation of tasks related to cloud management, such as log analysis, anomaly detection, infrastructure-as-code generation, cloud configuration management, troubleshooting, cloud incident summarization, and conversational support for cloud administrators [13]. While various studies have explored cloud computing and GenAI separately, studies that focus on the architecture, frameworks, enabling techniques, applications, and research challenges of Generative AI in intelligent cloud operations are sparse.

A. Structure of the paper

The following is the outline for the remainder of the paper. Section II introduces the fundamentals of cloud operations and discusses the most pressing concerns in today's cloud landscape. In Section III, learn about Generative AI, the core foundation models, and LLMs. In Section IV, the key architectures and leading Generative AI frameworks for intelligent cloud operations are discussed. The following sections elaborate on the key enabling techniques (Section V) and summarize the literature available (Section VI). Lastly, Section VII summarizes the paper and provides future research directions.

II. FUNDAMENTALS OF CLOUD OPERATIONS

Cloud Operations (CloudOps): Delivering, tuning, optimizing, and analyzing the performance of workloads and IT services deployed in cloud environments – from private cloud, multi-cloud, hybrid cloud, data centers and the edge. Like DevOps codifies operational processes for application development and delivery, CloudOps codifies operational processes using the cloud [14]. By leveraging automation, CloudOps teams can orchestrate complex workflows, ranging from provisioning infrastructure to deploying applications and monitoring performance and rapidly fixing issues [15]. Automating repetitive tasks minimizes human error, streamlines service delivery, and allows IT teams to prioritize strategic initiatives that fuel business value. CloudOps can provide a host of benefits to organizations, such as:

- **Accelerated automation:** Automation provisions infrastructure on the fly and increases reliability by enabling continuous monitoring that is used to automate remediation actions before the user is affected. This automation-first strategy allows CloudOps teams to concentrate on innovation and enhance deployment speed and efficiency.
- **Efficient operations:** Instant capacity increases, typically through a self-service portal, can satisfy growth or seasonal demands [16]. Organizations can be more flexible and meet demands as they come up by quickly setting up new virtual machines, physical servers, storage, and even network bandwidth.
- **Enhanced security:** Large hyperscale cloud providers take precautions against cybercrime and data exfiltration by implementing top-tier physical security measures and maintaining round-the-clock infrastructure monitoring.
- **Improved recovery:** A major motivator for moving to the cloud is the need to back up and restore data, which were the original use cases for the public cloud. Backup and replication methods can lower the RPO and RTO to almost nothing, which helps make sure that users and customers around the world can access the system at all times [17].
- **IT cost savings:** Cloud services remove huge capital expenditures, service contract purchases, and data center hardware upgrades to accommodate changing user demands because they are pay-as-you-go.
- **Anywhere access:** Cloud computing allows businesses to access their data and applications from virtually any device, anywhere in the world, as long as they have an internet connection.

A. Existing Challenges in Cloud Operations

Modern cloud environments are becoming increasingly complex due to dynamic workloads, distributed infrastructures, and multi-cloud deployments, creating several operational challenges.

- **Resource Management:** Efficient allocation of computing resources remains challenging due to dynamic and unpredictable workloads [18]. Poor resource provisioning can lead to increased operational costs, reduced resource utilization, and violations of SLAs.
- **Fault Detection and Incident Management:** Cloud infrastructures continuously generate large volumes of logs and monitoring data, making it difficult to detect failures and identify their root causes. Slow response to incidents can have a negative effect on system availability and user experience[19].
- **Security and Privacy:** Protecting sensitive information in shared cloud environments remains a major concern [20]. Organizations need to protect themselves from cyber-attacks, unauthorized access, data breaches and adhere to security and privacy regulations [21].
- **Multi-Cloud and Hybrid Cloud Management:** Operating across multiple cloud providers adds complexity of interoperability, resource orchestration, performance optimization, and vendor-specific configurations.
- **Automation and Operational Complexity:** Today, cloud-native applications built with containers, microservices and Kubernetes demand constant monitoring and management [22]. Traditional rule-based automation is often not enough and a need for intelligent automation solutions with AI arises[23].

III. CONCEPT OF GENERATIVE AI

The growth of GenAI has depended on mathematical progress, superior computing resources, and huge datasets. A statistical approach to probabilistic modeling was the ancestor of all generative models in artificial intelligence; it used hand-crafted probability distributions to synthesize data [24][25]. They highlight the need to take into account how extensively AI is replacing or complementing human efforts, how physical and digital spaces are merging, and how human capabilities are being drawn upon. GenAI is a collection of models that can discover a dataset's latent structure and use that knowledge to produce new data points that are statistically similar to existing ones. Despite writers' long-held belief that a model is an architecture plus a training phase, GenAI and AI in general lack a precise definition of a model [26].

A. Core Models of Generative AI

Modern GenAI is built upon large-scale pre-trained models that learn complex patterns from massive datasets consisting of text, images, code, audio, and videos [27]. Content generation, reasoning, summarization, translation, and decision support are just a few of the many generative activities made possible by these models. Generative models of today are more versatile than their task-specific predecessors thanks to advancements in prompting, fine-tuning, and retrieval-based techniques, which allow them to be used in more and more downstream applications. The most widely adopted core models of Generative AI include:

- **GPT (Generative Pre-trained Transformer):** The GPT models, which were created by OpenAI, are extensively utilized for reasoning, content production, code generation, conversational AI, and natural language understanding.
- **Llama (Large Language Model Meta AI):** Designed by Meta, Llama is an open-source language model that supports research and enterprise applications with minimal computational demands.
- **Gemini:** Gemini is an intelligent assistant and complicated reasoning candidate since it is a multimodal model (created by Google) that can comprehend and produce code, graphics, audio, video, and text.
- **Claude:** Developed by Anthropic, Claude emphasizes safe, reliable, and interpretable AI by incorporating constitutional AI principles, making it suitable for enterprise and cloud-based decision-support systems.
- **Mistral:** Developed by Mistral AI, Mistral models are lightweight and computationally efficient while delivering competitive performance in text generation, reasoning, and software development tasks [28].

The backbone of the modern ecosystem of Generative AI applications is these core models. They support intelligent cloud operations, such as automated log analysis, automated generation of infrastructure-as-code, cloud configuration management, anomaly detection, incident summarisation, conversational cloud assistants, and intelligent decision support.

B. Role of Large Language Models in Generative AI

LLMs are one of the most important developments in GenAI that help machines process and understand natural language, generate content, and reason about it [29]. Transformer models, used to build LLMs, are usually pre-trained on large text corpora using self-supervised learning; this allows them to grasp complex linguistic patterns and comprehend context. Text generation, summarization, translation, question answering, code generation, and conversational help are just a few of the NLP-related tasks they may accomplish.

The powerful reasoning and language understanding capabilities of LLMs have made them an essential component of intelligent cloud operations [30]. They help cloud administrators analyze logs, summarize incidents, generate infrastructure-as-code, configure cloud infrastructure, troubleshoot, and document automatically. Additionally, by combining with cloud monitoring tools and RAG, LLMs can deliver context-relevant and domain-specific answers, enhancing both operational efficiency and decision-making processes [31]. Despite these advantages, LLMs face challenges such as high computational requirements, training and inference costs, hallucinations, bias, and privacy concerns. Enhancing the trustworthiness, explainability and security of LLM-based systems thus remains a key research area for intelligent cloud-based systems.

IV. ARCHITECTURES OF GENERATIVE AI FRAMEWORKS FOR INTELLIGENT CLOUD OPERATIONS

In the context of cloud management and operations, GenAI frameworks facilitate intelligent reasoning, automation, decision-making, and content generation, optimizing cloud management, monitoring, security, and operational efficiency through powerful DL architectures.

A. Core Architectures of Generative AI

Machines can learn complicated data distributions and create realistic content with the help of various DL architectures used in generative AI. Generative tasks like text generation, picture synthesis, anomaly detection, and data representation are handled by architectures that use different learning mechanisms.

- **Transformer Architecture:** The self-attention processes of transformers allow modern generative models to quickly process sequential data and capture long-range connections without recurrence, making them the backbone of these models. The decoder-only configuration and the encoder-decoder configuration are two often used Transformer topologies for generative modeling [32]. There are primarily three types of large-scale pre-trained models now available: encoder-based, decoder-based, and encoder-decoder-based versions. Using the same self-attention method, Transformer resolves the issue of long-distance dependency. Sequence challenges are well-suited to encoder-decoders.
- **GANs:** A GAN is a type of unsupervised generative model that uses a generator and a discriminator neural network. In contrast to discriminators, generators try to create new data (false) that is nearly indistinguishable from real data. False data is produced by the generator network by feeding it noise. A sigmoid activation function and binary cross-entropy loss are used by the discriminator network to classify incoming data as real or fraudulent [33].
- **Variational Autoencoders (VAEs):** The three components of this type of artificial neural network are the encoder, the latent space (sometimes referred to as the "bottleneck layer"), and the decoder. Data input is mapped to latent space, and the process of producing the same data as input is subsequently taught. This is the definition of a probabilistic generative model called VAE, which is an enlarged version of AE [34]. The primary distinction between AE and VAE is that the former learns to compress the input and then decompress it so that it matches the input. In contrast, the VAE is a Bayesian model that, after learning the compressed AE representation, constructs parameters that characterize the probability distribution of the data.
- **Diffusion models:** Diffusion models work by slowly changing the observed data through a forward diffusion process, and then restoring the original data through a backward reverse process. In the forward process, the noise level is varied at each of the many steps of injection. The backward process, on the other hand, is a sequence of denoising operations designed to filter out the noise injected step by step [35].

B. Generative AI Frameworks

A framework is a set of tools to help build, launch, and run an application that utilizes an LLM. They streamline the integration of language models with external data sources, APIs, vector databases, and cloud platforms, and facilitate capabilities like workflow orchestration, retrieval, memory management, and multi-agent collaboration. These frameworks are essential for facilitating intelligent cloud operations, including automation, knowledge retrieval, and decision support.

- **LangChain:** The education sector and chatbots are two typical uses for LangChain, an open-source framework that facilitates the development of applications based on LLMs. Document loaders, embeddings, vector stores, and prompt templates are some of the framework's modular components that simplify the development of intricate workflows. By integrating RAG, LangChain can help LLMs access external knowledge bases, improving their accuracy in specific domains [36].
- **LlamaIndex:** LlamaIndex enables LLMs to efficiently access and retrieve information from structured and unstructured enterprise data. It allows for indexing with document vectors and retrieval, which makes it ideal for cloud documentation, knowledge management and troubleshooting.
- **AutoGen:** AutoGen is a multi-agent framework that enables multiple AI agents to collaborate on complex tasks. It supports automated workflow execution, cloud orchestration, and intelligent decision-making, reducing manual intervention in cloud operations.
- **CrewAI:** CrewAI coordinates multiple specialized AI agents to execute collaborative workflows. It is widely used for cloud monitoring, task automation, and operational management by assigning specific roles to different agents.
- **Semantic Kernel:** Semantic Kernel is a Microsoft framework that integrates LLMs with enterprise applications and cloud services. It supports workflow automation, plugin integration, and AI-powered cloud management, making it suitable for enterprise-scale intelligent cloud operations.

V. Techniques Used In Generative AI for Intelligent Cloud Operations

GenAI employs several advanced techniques to improve the accuracy, efficiency, and adaptability of intelligent cloud operations. These techniques enable LLMs to better understand user queries, retrieve domain-specific knowledge, and autonomously perform complex operational tasks.

A. *Prompt Engineering Techniques*

The field of software development and engineering has extensively embraced prompt engineering methodologies. To make generative models more effective in different kinds of development tasks, efficient prompts have been developed and used [37]. The two most common configurations are zero-shot and few-shot. In the zero-shot mode, an LLM completes a task based solely on the information it obtained during pre-training, without any training data, and with brief directions. However, in the few-shot or in-context option, the model is also given a small number of task-specific examples to help it understand the prompt better [38].

B. *Retrieval-Augmented Generation (RAG)*

The three main parts of a RAG system are indexing, retrieval, and creation. They optimize data embeddings during indexing to improve indexed content quality [39]. They learn dynamic embeddings for various kinds of content and fine-tune embedding models to compute the similarity between query content and text blocks during the retrieval phase, so they can find relevant content. During the generating phase, the user question and the retrieved content are combined to form the LLM's prompts. To overcome limits imposed by the context window, prompts may be compressed or reordered according to relevancy [40].

C. *Agentic AI Techniques*

Agentic AI expands upon multi-agent systems by giving software agents more complex cognitive models, allowing them to engage in deliberate, goal-directed behavior rather than just reacting to situations [41]. The development of agentic AI and the rise of LLM agents are both results of progress in the study of LLMs. Agents, a shorthand for LLM agents in this review, use LLMs to reason, make decisions, and carry out actions in agentic AI. These agents combine planning, memory management, and tool usage; they are built to execute multi-step operations, retrieve information, communicate with other APIs or knowledge bases, and collaborate. They find work in many different industries, such as software engineering, banking, and the scientific community [42].

VI. Literature Review

GenAI frameworks for intelligent cloud operations are examined in this review, along with their architectures and techniques.

Khara et al (2026) The objective is to enhance the caliber of strategic, operational, and tactical decision-making through the implementation of a data-driven approach to integrating GenAI with decision support systems within organizations. Organizational decision support systems rely on generative models, which are the subject of this study. Generative models are used to build scenarios, extract information, and make predictions [43].

Khan A (2026) explores RAG pipelines that integrate generative AI with "intelligent cloud augmentation systems" and "knowledge retrieval systems"; the principal emphasis is on the supplemental intelligent cloud operating systems and knowledge retrieval systems [44].

Gu et al (2025), suggest a smart monitoring system that combines DL with FL algorithms. Data from many sources, including video streams, employee whereabouts, and operational behaviors on the job site, is gathered in real-time by the system using edge terminals. This data is then processed by a cloud-based collaborative architecture [45].

Ratra et al (2025) offer a Hybrid Edge-to-Cloud Infrastructure powered by AI to address retail supply chain issues. Improved inventory demand forecasting and more efficient communication are two outcomes of this solution's focus on data visualization and streamlining. By providing immediate answers to existing inefficiencies, a platform that combines AI algorithms with a variety of models revolutionizes the supply chain [46].

Gao et al (2024) offers a ML algorithm-based system for network O&M that makes complete use of the cloud's resources while also making use of ML's analytical prowess to accomplish goals like automated processing of network status, fault prediction, and real-time monitoring [47].

Rani et al (2023) This paper offers a thorough analysis of generative AI models by comparing and contrasting various aspects such as input, output, development authority, frameworks, and tools. On top of that, covers the many domains where various generative AI models have found valuable applications [48].

Table I summarizes the existing literature on architectures and techniques of GenAI frameworks for intelligent cloud operations, highlighting the research focus, key findings, challenges, limitations, and future research directions.

TABLE I. LITERATURE OF REVIEW ON ARCHITECTURES AND TECHNIQUES OF GENERATIVE AI FRAMEWORK FOR INTELLIGENT CLOUD OPERATIONS

Reference	Study on	Key Findings	Challenges	Limitations	Future Work
Khara et al. 2026	Generative AI for Organizational Decision Support Systems	Integrated GenAI with decision support systems for strategic, operational and tactical decision-making; improved knowledge extraction, scenario generation and prediction.	Integration with enterprise workflows and reliable knowledge generation.	Limited validation across diverse organizations; dependent on high-quality data.	Develop domain-specific GenAI, improve explainability and validate at industrial scale.
Khan A., 2026	GenAI with Intelligent Cloud Augmentation and RAG	Integrated LLMs and RAG pipelines for intelligent cloud operations and knowledge retrieval.	Retrieval accuracy, hallucinations and secure cloud knowledge management.	Performance depends on retrieval quality and computational resources.	Optimize RAG efficiency and secure knowledge retrieval.
Gu et al., 2025	Cloud-Edge Monitoring using DL and FL	Cloud-edge collaborative monitoring using DL and FL for real-time multisource data analysis.	Edge heterogeneity, communication overhead and privacy.	High deployment complexity and edge computation cost.	Improve scalability and lightweight federated models.
Ratra et al., 2025	AI Hybrid Edge-to-Cloud Infrastructure	AI-enabled edge-cloud solution for retail supply chain, forecasting and communication.	Integration of heterogeneous infrastructure and latency.	Limited real-world scalability evaluation.	Enhance predictive analytics and resource allocation.
Gao et al., 2024	ML-based Intelligent Network Operation	ML-based cloud network O&M for monitoring, fault prediction and automation.	Accurate fault prediction in dynamic environments.	Limited validation on complex infrastructures.	Integrate advanced DL for autonomous network management.
Rani et al., 2023	Comparative Analysis of GenAI Models	Compared GenAI models, tools, frameworks and applications.	Rapid evolution and lack of standard evaluation.	Mostly comparative with limited implementation.	Develop standardized benchmarks and domain-specific optimization.

VII. CONCLUSION AND FUTURE WORK

GenAI is transforming intelligent cloud operations by enabling advanced automation, intelligent decision-making, and context-aware assistance across modern cloud environments. The rapid evolution of LLMs, foundation models, and GenAI frameworks has significantly improved capabilities such as cloud monitoring, log analysis, incident management, infrastructure automation, and resource optimization. The role of GenAI in resolving operational issues was covered in this paper, which also covered the basics of cloud operations. Along with popular frameworks like LangChain, LlamaIndex, AutoGen, CrewAI, and Semantic Kernel, it dived deeper into GenAI's fundamental architectures including Transformers, GANs, VAEs, and Diffusion Models. In addition, key techniques, including Prompt Engineering, RAG, and Agentic AI, were presented to highlight their contributions to intelligent cloud management. Despite the significant progress, challenges related to hallucinations, explainability, computational overhead, security, privacy, and the reliable deployment of GenAI in large-scale cloud environments remain unresolved. Future research should focus on developing trustworthy and explainable AI models, lightweight and energy-efficient LLMs, secure multi-agent cloud management systems, and domain-specific RAG frameworks. Furthermore, the integration of GenAI with AIOps, cloud-native technologies, edge computing, and autonomous orchestration platforms is expected to drive the next generation of intelligent, resilient, and self-managing cloud infrastructures.

REFERENCES

[1] S. Chatterjee, "A Data Governance Framework for Big Data Pipelines: Integrating Privacy, Security, and Quality in Multitenant Cloud Environments," *Tech. Int. J. Eng. Res.*, vol. 10, no. 5, 2023, doi: 10.56975/tijer.v10i5.158181.

[2] V. K. Sharma, "Cloud Computing IoT: 5G Focused IoT with Cloud Solutions," *Int. J. AI, Big Data, Comput. Manag. Stud.*, vol. 6, no. 3, 2025, doi: 10.63282/3050-9416.IJAIBDCMS-V6I3P103.

[3] S. K. Malaraju and S. K. Madishetty, "AI-Augmented Compiler Optimization for Energy-efficient Software Execution on Embedded Systems," in *2025 14th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, Uttar Pradesh, India: IEEE, 2025, pp. 1–6, November. doi: 10.1109/SMART66937.2025.11389313.

[4] K. Jangiti, "Design and Validation of a Machine Identity Governance Framework for AI Agents in Multi-Cloud Environments," in *SoutheastCon 2026*, IEEE, Feb. 2026, pp. 1–6. doi: 10.1109/SoutheastCon63549.2026.11476363.

[5] R. Palwe, "Three Layers of Trust in AI Interfaces: Interface, Behavior, and Organization," *Int. J. Sci. Res.*, vol. 15, no. 1, pp. 1152–1160, Jan. 2026, doi: 10.21275/SR26112072531.

[6] R. Azmeera and J. Hyatt, "Software Errors, Product Functionality, and Organizational Cascades: Evidence from Developer-Lived Experience," Mar. 2026. doi: 10.2139/ssrn.6924439.

[7] A. Joon, B. K. R. Janumpally, A. Gogineni, and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," in *2025 5th International Conference on Intelligent Technologies (CONIT)*, 2025, pp. 1–8. doi: 10.1109/CONIT65521.2025.11167760.

[8] E. Dritsas and M. Trigka, "A Survey on the Applications of Cloud Computing in the Industrial Internet of Things," *Big Data Cogn. Comput.*, vol. 9, no. 2, p. 44, Feb. 2025, doi: 10.3390/bdcc9020044.

[9] A. Alabdulatif, "A Novel Ensemble of Deep Learning Approach for Cybersecurity Intrusion Detection with Explainable Artificial Intelligence," *Appl. Sci.*, vol. 15, no. 14, p. 7984, Jul. 2025, doi: 10.3390/app15147984.

[10] S. R. Chanthati, "Implementing a Graph Neural Network (GNN) in Amazon Web Services (AWS) involves setting up infrastructure for data processing, training, and deployment," in *Proceedings of the 6th World Conference on Artificial Intelligence: Advances and Applications (WCAIAA 2025)*, May 2025. doi: 10.5281/zenodo.20313312.

[11] K. Patel, M. Shah, K. M. Qureshi, and M. R. N. Qureshi, "A

systematic review of generative AI: importance of industry and startup-centered perspectives, agentic AI, ethical considerations & challenges, and future directions," *Artif. Intell. Rev.*, vol. 59, no. 1, p. 7, 2025, doi: 10.1007/s10462-025-11435-z.

[12] J. B. Mehta, "AI-Driven Test Engineering For Cloud-Native Systems," *Int. J. Data Sci. IoT Manag. Syst.*, vol. 5, no. 1, Jan. 2026, doi: 10.64751/ijdim.2026.v5i1.297.

[13] R. Palwe and A. Kumar, "Redefining usability in the age of generative AI: Towards a new evaluation paradigm," *Int. J. Comput. Artif. Intell.*, vol. 6, no. 2, pp. 155–163, Jul. 2025, doi: 10.33545/27076571.2025.v6.i2b.193.

[14] K. Hamid, M. W. Iqbal, Q. Abbas, M. Arif, A. Brezulianu, and O. Geman, "Cloud Computing Network Empowered by Modern Topological Invariants," *Appl. Sci.*, vol. 13, no. 3, 2023, doi: 10.3390/app13031399.

[15] M. R. C. Mukkolakkal, "Deploy, Calibrate, Monitor, Heal -- No Human Required: An Autonomous AI SRE Agent for Elasticsearch," *arXiv.org*, Apr. 2026.

[16] E. Aviv, Itzhak and Gafni, Ruti and Sherman, Sofia and Aviv, Bertha and Sterkin, Asher and Bega, "Cloud infrastructure from Python code--breaking the barriers of cloud deployment," pp. 1--8, 2023.

[17] N. Kratzke, "A Brief History of Cloud Application Architectures," *Appl. Sci.*, vol. 8, no. 8, 2018, doi: 10.3390/app8081368.

[18] O. Islam, Rafia and Patamsetti, Vardhan and Gadhi, Aparna and Gondu, Ragha Madhavi and Bandaru, Chinna Manikanta and Kesani, Sai Chaitanya and Abiona, "The future of cloud computing: benefits and challenges," *Int. J. Commun. Netw. Syst. Sci.*, vol. 16, pp. 53--65, 2023.

[19] D. V K R K Koppireddy, D. GNR Parasad, D. Drijesh Patel, S. Sasidhar Reddy, and V. Kamaraj, Akshita Meher N, "AI Security Governance for Trustworthy Intelligent Systems: A Critical Analysis of Cyber Risks, Privacy Threats, and Responsible Deployment Frameworks," *Int. J. Res. Trends Innov.*, vol. 11, no. 7, 2026, doi: 10.56975/ijrti.v11i7.214029.

[20] S. Kumara and M. Shah, "Securing national connectivity infrastructure through identity resilience: implications for zero trust–aligned telecom security," *Futur. Technol.*, vol. 5, no. 6, pp. 276–286., 2026, doi: 10.55670/fpll.futech.5.3.24.

[21] R. K. Kanneganti, "Security and Compliance Issues in Cloud-Based Deployments of Content and Workflow Management Systems," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 2, no. 01, pp. 131–138, Aug. 2024, doi: 10.58812/esiscs.v2i01.1122.

[22] H. P. Cyril, N. D. Bhandarwar, S. Kumara, and S. Mathur, "Securing Containerized Applications Using Kubernetes-Based Orchestration in Software Development," in *2026 International Conference on Intelligent Computing, Networks, and Security (IC-ICNS)*, 2026, pp. 1–6. doi: 10.1109/IC-ICNS68863.2026.11537887.

[23] V. K. R. K. Koppireddy, "Passwordless Authentication: A Modern Approach to Secure Access," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 63, 2025.

[24] M. Trigka and E. Dritsas, "The Evolution of Generative AI: Trends and Applications," *IEEE Access*, vol. 13, pp. 98504–98529, 2025, doi: 10.1109/ACCESS.2025.3574660.

[25] M. N. Paralath, "Bridging Astronomical Imagery and Text with Advanced Multi-Modal AI," *2025 IEEE 7th Int. Conf. Comput. Commun. Autom.*, pp. 1–8, 2025.

[26] A. Bordas, P. Le Masson, M. Thomas, and B. Weil, "What is generative in generative artificial intelligence? A design-based perspective," *Res. Eng. Des.*, vol. 35, no. 4, pp. 427–443, 2024, doi: 10.1007/s00163-024-00441-x.

[27] M. Mohit, "The Rise of Generative AI: Evaluating Large Language Models for Code and Content Generation," *Int. J. Adv. Sci. Eng. Technol.*, vol. 10, no. 4, pp. 20643–20649, 2023.

[28] V. Chaturvedi, "Modern Software Development with Java , Spring Boot , and Python : A Survey of Frameworks and Best Practices," *ESP J. Eng. Technol. Adv.*, vol. 3, no. 4, pp. 188–197, 2023, doi: 10.56472/25832646/JETA-V3I8P121.

[29] S. Kakkar, "Prompt-Robotics: Leveraging Large Language Model (Llm) Agents To Enhance Robotic Process Automation (Rpa) Supervision And Efficiency," *Int. J. Financ. DATA Sci.*, vol. 3, no. 2, pp. 7–29, Sep. 2025, doi: 10.34218/IJFDS_03_02_002.

[30] S. R. Chanthati, "Integrating Large Language Models (LLMs) in Cognitive Medical Robotics," in *Proceedings of the International Conference on AI and Robotics (AIR 2025)*, May 2025. doi: 10.5281/zenodo.20316245.

[31] A. Naveed, Humza and Khan, Asad Ullah and Qiu, Shi and Saqib, Muhammad and Anwar, Saeed and Usman, Muhammad and Akhtar, Naveed and Barnes, Nick and Mian, "A comprehensive overview of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 16, pp. 1--72, 2025.

[32] J. Kosińska and M. Tobiasz, "Detection of Cluster Anomalies With ML Techniques," *IEEE Access*, vol. 10, pp. 110742–110753, 2022, doi: 10.1109/ACCESS.2022.3216080.

[33] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers," *IEEE Access*, vol. 12, pp. 69812–69837, 2024, doi: 10.1109/ACCESS.2024.3397775.

[34] K. A. Alaghbari, H. S. Lim, B. Jin, and Y. Shen, "Source Separation in Joint Communication and Radar Systems Based on Unsupervised Variational Autoencoder," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 56–70, 2024, doi: 10.1109/OJVT.2023.3335358.

[35] L. Lin, Z. Li, R. Li, X. Li, and J. Gao, "Diffusion Models for Time-Series Applications: A Survey," *Front. Inf. Technol. Electron. Eng.*, vol. 25, no. 1, pp. 19–41, 2024, doi: 10.1631/FITEE.2300310.

[36] D. Kim, S. Won Park, and J. Lee, "CQELedu: Design and Implementation of a LangChain and GPT-4o Mini-Based Web Application for Custom Question Generation and Error-Based Learning in Education," *IEEE Access*, vol. 13, pp. 208932–208947, 2025, doi: 10.1109/ACCESS.2025.3639087.

[37] A. F. Subahi, "Prompt Engineering for Advancing Software Engineering: Using Generative Language Models in the Development of Domain-Specific Languages," *IEEE Access*, vol. 14, pp. 22834–22850, 2026, doi: 10.1109/ACCESS.2026.3663331.

[38] S. Vatsal, H. Dubey, and A. Singh, "Multilingual Prompt Engineering in Large Language Models: A Survey Across NLP Tasks," *IEEE Access*, vol. 14, pp. 99057–99093, 2026, doi: 10.1109/ACCESS.2026.3702852.

[39] S. Devarakonda, R. Lingam, and V. Challa, "Confidence-Gated RAG for Adaptive Retrieval in Sequential Agents," in *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026, pp. 1–18. doi: 10.1109/TCASAI.2026.3653699.

[40] R. Hu, S. Liu, P. Qi, J. Liu, and F. Li, "ICCA-RAG: Intelligent Customs Clearance Assistant Using Retrieval-Augmented Generation (RAG)," *IEEE Access*, vol. 13, pp. 39711–39726, 2025, doi: 10.1109/ACCESS.2025.3544408.

[41] A. Hasan, E. Kandemir, D. Mordasov, and D. Trong Nguyen, "Agentic AI in Wind Energy Systems: Multi-Agent Architectures for Optimization and Resilience," *IEEE Access*, vol. 14, pp. 9935–9951, 2026, doi: 10.1109/ACCESS.2026.3654529.

[42] I. da Silva Portugal, P. Alencar, and D. Cowan, "Agentic Recommender Systems: A Systematic Literature Review," *IEEE Trans. Softw. Eng.*, vol. 52, no. 4, pp. 1234–1264, 2026, doi: 10.1109/TSE.2026.3657900.

[43] S V Khara, G. D. Tivari, S. Patel, J. Nathvani, R. Amrutiya, and D. Chavda, "Generative Artificial Intelligence for Intelligent Decision Support in Organizational Management: A Data-Driven Framework," in *2026 3rd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, 2026, pp. 1–5. doi: 10.1109/RMKMATE69073.2026.11518863.

[44] K. K. A, "Generative AI-Powered Rag Pipelines for Intelligent Cloud Operations and Knowledge Retrieval," in *2026 9th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2026, pp. 1388–1394. doi: 10.1109/ICICCS67901.2026.11502902.

[45] B. Gu, Y. Zhou, D. Ma, X. He, and L. Tian, "Performance Optimization Strategy for Power Operation Safety Intelligent Monitoring System under Cloud Edge Fusion Architecture," in *2025 2nd International Seminar on Artificial Intelligence, Computer Technology and Control Engineering (ACTCE)*, 2025, pp. 624–628. doi: 10.1109/ACTCE66599.2025.00123.

[46] K. K. Ratra and D. K. Seth, “AI-Driven Hybrid Edge-Cloud Architecture for Real-Time Big Data Analytics and Scalable Communication in Retail Supply Chains,” in *SoutheastCon 2025*, 2025, pp. 1023–1029. doi: 10.1109/SoutheastCon56624.2025.10971468.

[47] C. Gao, J. Zhang, D. Xuan, G. Deng, and W. Cai, “Network Operation and Maintenance Technology Based on Machine Learning Algorithms in Cloud Computing Environment,” in *2024 International Conference on Interactive Intelligent Systems and Techniques (IIST)*, IEEE, 2024. doi: 10.1109/IIST62526.2024.00058.

[48] G. Rani, J. Singh, and A. Khanna, “Comparative Analysis of Generative AI Models,” in *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, 2023, pp. 760–765. doi: 10.1109/ICAICCIT60255.2023.10465941.