ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Generative AI for IoT and Edge Computing: Enhancing Intelligent Edge Systems

Mr. L. S. Shendge , Mrs. S. N. Patel, Ms. D. V. Sharma

Dayanand College of Commerce, Latur

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600097

Received: 29 April 2026; Accepted: 06 July 2026; Published: 09 July 2026

ABSTRACT

The rapid expansion of the Internet of Things (IoT) has resulted in massive volumes of data being generated by interconnected devices across various domains. Traditional cloud-centric architectures often struggle with issues such as high latency, bandwidth constraints, and data privacy risks when processing this data. Edge computing has emerged as an effective solution by enabling data processing closer to the data source, thereby improving response time and reducing network dependency. In recent years, Generative Artificial Intelligence (GenAI) has gained significant attention for its ability to generate insights, predictions, and adaptive responses from complex and dynamic datasets. This paper examines the integration of Generative AI with IoT and edge computing to enhance intelligent edge systems capable of real-time analytics and autonomous decision-making. It explores architectural frameworks, potential applications in areas such as smart cities, healthcare, industrial automation, and autonomous systems, as well as the advantages of improved efficiency, scalability, and privacy preservation. Additionally, the paper discusses the technical challenges associated with deploying generative models at the edge, including resource constraints, model optimization, security, and data management. Finally, it outlines future research directions aimed at developing scalable, secure, and energy-efficient GenAI-enabled edge computing ecosystems.

INTRODUCTION

The Internet of Things (IoT) has swiftly expanded, producing large records volumes that standard cloud-centric architectures warfare to system due to excessive latency and privateness dangers (p. 1). Edge computing addresses this by means of shifting processing nearer to the statistics supply (p. 1). Recently, Generative AI (GenAI) has emerged as a transformative force, enabling these structures to go past easy facts processing to producing complicated insights and self sufficient selections in real-time.

The paradigm shift from centralized cloud computing to the decentralized "Generative Edge" represents one of the most significant technological inflections in the history of the Internet of Things (IoT). While the previous decade was defined by the "Connected Edge," where localized devices primarily served as data conduits or performed rudimentary discriminative tasks, the current era is characterized by the integration of Generative Artificial Intelligence (GenAI) directly into the edge fabric. This transition is driven by the maturation of Large Language Models (LLMs), Generative Adversarial Networks (GANs), and Diffusion models, which are now being optimized for deployment in resource-constrained environments. The convergence of these paradigms allows for a new class of "Living Intelligence" that merges AI, high-fidelity sensing, and real-time reasoning to create systems that think, adapt, and evolve beyond the limitations of static programming.

Objectives:

1. To learn about the integration of Generative AI with IoT and Edge Computing
2. Understand how GenAI can decorate IoT structures by way of working at the edge.
3. To enhance real-time information processing and decision-making.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

4. Enable quicker analytics and self sustaining responses through processing information nearer to the source.
5. To analyze architectural frameworks for shrewd side structures.
6. Explore machine designs that mix IoT, aspect computing, and generative AI.
7. To become aware of purposes in a range of domains.
8. Study use instances such as clever cities, healthcare, industrial automation, and self sustaining systems.
9. To consider advantages like efficiency, scalability, and privateness.
10. Show how edge-based GenAI improves performance and reduces risks.
11. To observe challenges in deploying GenAI at the aspect.
12. Address problems like constrained resources, mannequin optimization, security, and information management.
13. To propose future lookup instructions.
14. Focus on constructing scalable, secure, and energy-efficient clever area ecosystems.

Theoretical Framework and the Evolution of Edge Intelligence

The evolution of generative AI has progressed through several distinct waves of innovation, beginning with early recurrent neural network (RNN)-based sequence-to-sequence (seq2seq) models in the mid-2010s. These foundational models laid the groundwork for the eventual invention of the Transformer architecture, which utilized self-attention mechanisms to capture long-range dependencies in data far more effectively than its predecessors. However, as models grew in capability, they also grew in complexity, leading to today's state-of-the-art frontier models comprising hundreds of billions of parameters.

This trajectory of growth created a tension at the heart of edge AI challenges: the demand for high-fidelity intelligence versus the finite resource constraints of edge hardware. Consequently, the field has bifurcated. On one side, frontier models continue to push the boundaries of performance in specialized data centers. On the other side, a specialized branch of AI research is focused on optimization and deployment-conscious design, enabling a new class of small, specialized models capable of running on everyday devices. This evolution is underpinned by a fundamental shift in the mathematical objective of AI modeling.

Mathematical Objectives: Generative versus Discriminative Paradigms

To understand the impact of GenAI on edge systems, one must distinguish it from the traditional discriminative models that have historically dominated the IoT landscape. Discriminative AI is designed to learn the decision boundary between different classes of data. It focuses on the conditional probability distribution, denoted as:

$$P(Y|X)$$

In this equation, Y represents the label or class, and X represents the input features. This mathematical focus makes discriminative models exceptionally efficient at classification, ranking, and risk scoring—tasks where the goal is to determine whether an input belongs to a predefined category, such as identifying a face or filtering spam.

Generative AI, conversely, seeks to understand the entire underlying data distribution or joint probability distribution:

$$P(X,Y)$$

By modeling how the data and its labels are related, generative models can produce entirely new samples that preserve the structural, semantic, and statistical properties of the original training domain. For edge systems, this capability is transformative. It allows a device not only to recognize an anomaly but to simulate potential future states, synthesize missing sensor data, and translate complex human intent into semantic actions.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Feature	Generative AI	Discriminative AI
Core Mathematical Goal	Model the joint probability $P(X, Y)$	Model the conditional probability $P(Y
Primary Output	New data instances (text, images, synthetic signals)	Class labels, probabilities, or decision boundaries
Learning Paradigm	Unsupervised, semi-supervised, or self-supervised	Primarily supervised learning
Data Utilization	Excels with unlabeled datasets	Requires large amounts of labeled data
Edge Advantage	Creative simulation and data augmentation	High accuracy in predictive and classification tasks

Architectural Integration of Generative AI in Edge Systems

Architecting edge systems for GenAI requires addressing the "data-model-compute triangle," where resource intensity often clashes with the operational realities of battery-powered or thermally limited devices.

A cloud-based model with 60 billion parameters cannot run on a standard edge device without significant architectural adjustments.

The industry is thus moving toward leaner data models, typically in the range of 4 billion parameters, fine-tuned for specific tasks such as computer vision or interactive user manuals.

Specialized Hardware and Neural Processing Units (NPUs)

The hero of the edge GenAI narrative is the Neural Processing Unit (NPU), a specialist processor built to handle the math behind neural networks without draining battery life. Unlike general-purpose CPUs or repurposed GPUs, NPUs utilize optimized dataflows to minimize data movement.

A common strategy is the "weight stationary" dataflow, where model weights are kept in local SRAM caches to be reused across multiple compute cycles, significantly reducing power-hungry off-chip memory access.

Furthermore, NPUs are designed to exploit the inherent sparsity in generative models. Sparsity support allows the hardware to recognize and skip zero-valued weights or activations, which can provide a performance boost of 4x to 8x. This is particularly critical for Transformer-based models, where the attention mechanism often results in many near-zero values.

Accelerator Platform	Peak Performance (TOPS)	Typical Power (W)	Core Architecture Features
Hailo-10H	40	< 3.5	Scalable distributed data flow architecture
RaiderChip NPU	Benchmarked High	Optimized	Leads in energy efficiency for quantized LLMs

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

ARC NPX6 IP	Scalable	Variable	Optimized for generative inference and sparsity
Synaptics SYN765x	Integrated AI	Ultra-low	Combines Wi-Fi 7 with AI-native compute

The economic implications of these architectures are profound. By extracting maximum value from every watt of power, devices equipped with dedicated NPUs can achieve higher autonomy and execution speeds while reducing operational costs in large-scale deployments.

Multi-Protocol Connectivity and Low Latency

The integration of GenAI at the edge does not occur in isolation but relies on advanced connectivity standards. Multi-protocol platforms that integrate Wi-Fi 7 (6 GHz), Bluetooth Low Energy (LE), and IEEE 802.15.4 (Thread/Zigbee) are essential for managing the fragmented and congested IoT environment. Wi-Fi 7 is particularly critical for hybrid AI architectures, where intelligence is distributed between the device and the cloud. In these setups, wireless latency is the primary bottleneck. Wi-Fi 7's superior latency performance ensures that real-time AI processing—such as that required for AI-powered language translators or autonomous vehicle computer vision—maintains high system integrity.

Model Optimization: Pruning, Quantization, and Distillation

Since unmodified generative models often exceed edge constraints by orders of magnitude, model compression represents a critical engineering pillar. Inference time per token for a GPT variant on a standard edge CPU can range from 350ms to 520ms, which is unacceptable for real-time interactions. Strategic compression can reduce these requirements by factors of 5x to 20x while maintaining acceptable quality.

Pruning Techniques

Network pruning systematically eliminates redundant parameters based on importance criteria. Magnitude-based pruning, which removes weights below a specific threshold, has demonstrated a 67.8% reduction in model size with only a 2.3% degradation in quality. Hardware-aware structured pruning takes this a step further by removing entire filters or channels, achieving a 3.2x speedup on edge NPUs while maintaining 94.6% of baseline performance in image generation tasks.

One of the most promising avenues in pruning research is the "lottery ticket hypothesis." Experiments have shown that pruned subnetworks containing only 14.7% of original weights can maintain over 95% of full model performance. This technique has successfully reduced inference times from 752ms to 167ms on edge-grade GPUs.

Quantization and Mixed Precision

Quantization involves reducing the bit-depth of model weights and activations. Standard post-training 8-bit quantization can reduce model size by 73.5% with less than 2% quality degradation, providing a 3.4x speedup and 75% memory savings. For more aggressive deployment scenarios, 4-bit quantization offers an 87.2% reduction in size, though the quality loss increases to approximately 4.6%.

Mixed-precision approaches allow for a more nuanced trade-off. By allocating 16-bit precision to the most sensitive 12.3% of the network and keeping the rest at 8-bit, researchers have achieved a quality degradation of only 0.7% while preserving nearly 70% of the compression benefits.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Compression Method	Model Size Reduction (%)	Inference Speedup	Quality Impact (Degradation)
Magnitude-based Pruning	67.8	Variable	2.3%
Structured NPU Pruning	Variable	3.2x	5.4%
8-bit Quantization	73.5	3.4x	1.9%
4-bit Quantization	87.2	5.8x	4.6%
Mixed Precision (8/16-bit)	~70	Variable	0.7%

The integration of these techniques allows for a combined efficiency improvement of up to 24.3x on edge devices, reducing total inference time from 3.7 seconds to 152 milliseconds per generation step.

Generative AI in 6G and Ambient Intelligence (AmI)

The realization of Ambient Intelligence (AmI) at a global scale requires the capabilities of sixth-generation (6G) wireless networks, which provide real-time perception and reasoning. GenAI serves as the creative core of these environments, filling key gaps that traditional AI cannot address.

Semantic Communication and Sensor Synthesis

A primary challenge in AmI is that real-world wireless and sensing datasets are often noisy or incomplete. Generative models bridge this gap by synthesizing missing sensor data and wireless channel information in under-observed areas. Furthermore, 6G networks utilize GenAI to perform semantic communication, compressing user intent into low-bit messages that are semantically rich rather than just bit-accurate. This enables ultra-reliable low-latency communication (URLLC) by reducing the total data volume required for complex interactions.

Generative Digital Twins (GDT)

Generative Digital Twins (GDT) represent a significant evolution over traditional digital twins. While standard twins mirror physical entities, GDTs can proactively predict, control, optimize, and simulate behavior by generating realistic multimodal content. Hierarchical generative AI-enabled twins replicate communication behaviors at both the message and policy levels, allowing for real-time synchronization and the emulation of complex user-environment interactions within 6G network slices.

Industrial AIoT and Predictive Maintenance (PdM)

The convergence of AI and the Industrial Internet of Things, often referred to as the Artificial Intelligence of Things (AIoT), has revolutionized predictive maintenance. Early maintenance approaches were reactive or followed rigid preventive schedules, which often led to either unexpected downtime or wasted resources.

The Evolution toward Prescriptive Maintenance

Current state-of-the-art systems utilize AI-driven predictive maintenance to monitor equipment health in real-time using IoT sensors that capture vibration, heat, and acoustics. GenAI enhances this by providing prescriptive insights—recommending specific intervention actions based on synthesized failure scenarios.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Maintenance Generation	Core Technology	Operational Strategy	Primary Limitation
Reactive	Manual inspection	Fix after failure	High downtime and costs
Preventive	Statistical scheduling	Scheduled servicing	Wasted life of good parts
Predictive (AI)	IoT Sensors + ML	Real-time monitoring	Requires massive labeled datasets
Prescriptive (GenAI)	GANs + Transformers	Proactive simulation	High computational demand at edge

GenAI models like GANs and VAEs address the "scarce data" problem in industrial environments by creating synthetic high-fidelity data that mimics real-world sensor signals. This allows systems to be trained on diverse failure patterns even when historical failure data is rare. Transformers, with their self-attention mechanisms, are particularly effective at capturing long-range dependencies in time-series sensor data, outperforming traditional RNNs in predicting the Remaining Useful Life (RUL) of critical components.

Real-Time Signal Enhancement

In smart factories, GenAI embedded at the edge can perform real-time streaming data augmentation. For example, the SA-Net-Biomass Estimation Framework in precision agriculture uses an "Occlusion Reconstruction Module" to approximate biomass from 3D-LiDAR point clouds, even when the data is noisy or incomplete due to adverse environmental conditions. Similarly, the LSGLLM-E architecture is used for large-scale traffic forecasting at the edge, capturing spatio-temporal correlations that traditional models miss.

Security Landscape: Intrusion Detection and Secure Development

As IoT devices are heterogeneous and often lack standardized security protocols, GenAI has emerged as a crucial tool for enhancing security measures. GenAI facilitates applications ranging from passive threat detection to active mitigation.

Intrusion Detection Systems (IDS)

The use of GANs and Transformers in IoT IDS allows for more robust anomaly detection mechanisms. By learning realistic data distributions, GenAI can identify subtle deviations that indicate emerging threats or malicious actors seeking to exploit vulnerabilities. Research scrutinizing 100 recent papers shows that GANs are particularly effective at generating synthetic attack patterns to train more resilient ML-based detection systems.

Application Developer Guidance and Vulnerability Discovery

GenAI also assists software developers in creating secure software from the outset. LLMSecGuard, for example, uses a static code analyzer and iterative LLM-based analysis to identify and remove vulnerabilities like hard-coded credentials in production code. Another tool, LLift, targets "Use Before Initialization" (UBI) bugs within the Linux kernel, a critical issue for IoT systems running embedded Linux.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Security Tool	Target Domain	Key Performance Metric	Role in IoT Security
LLMSecGuard	Application Development	Automated patching	Minimizes vulnerabilities in production code
LLift	Embedded Linux Kernel	100% recall for UBI bugs	Identifies critical bugs leading to privilege escalation
TAM	IoT-Edge Auth	11.39% Auth rate improvement	Sustains privacy and authentication
HBCE-FL	Knowledge Sharing	Personalized privacy	Secure, decentralized data analysis

In the context of decentralized supply chains, agentic AI is used to monitor activity and dynamically restrict access for anomalies, countering insider threats through game-theoretic client selection mechanisms.

Privacy and Trustworthy Edge Intelligence

One of the primary motivations for edge processing is data privacy. Processing sensitive information locally ensures that user data does not have to move to the cloud, meeting strict privacy regulations.

Federated Edge AI

Federated learning (FL) allows multiple edge devices or organizations to collaboratively train a shared AI model without transferring their raw data. In this architecture, each client performs local training on its own dataset and shares only the resulting model updates (gradients or weights) with a central server. The server then securely aggregates these updates to improve a global model.

To prevent gradient leakage or inversion attacks, these systems employ several protective layers:

1. **Secure Multi-party Computation (SMPC):** Allows parties to jointly compute results without revealing individual inputs.
2. **Homomorphic Encryption (HE):** Enables computations on data while it remains encrypted.
3. **Differential Privacy (DP):** Introduces statistical noise to updates, making it mathematically difficult to trace an update to a specific record.
4. **Trusted Execution Environments (TEEs):** Provide hardware-level isolation for sensitive computations.

The market for federated learning is projected to reach nearly $1.9 billion by 2034, reflecting its growing importance in sectors like healthcare and finance where data confidentiality is paramount.

The FGMR-AEC Framework and Generative Privacy

The FGMR-AEC (Federated Generative Motion-Rendering with Adaptive Edge-IoT Collaboration) framework illustrates a new horizon in privacy-preserving edge intelligence. Designed for immersive animation and the metaverse, it ensures that pure biometric and motion data remain at the IoT layer. Instead of transmitting high-dimensional raw sensory data, the system converts it into edge-generated latent codes,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

reducing virtual network loads by 60% while maintaining rendering fidelity at a 95% structural similarity index.

Socio-Technical Challenges and the Future Outlook

Despite the promising advancements, the integration of GenAI at the edge faces significant challenges. The surging demand for compute-intensive workloads has exposed cracks in global infrastructure, including data center power constraints and physical network vulnerabilities. Furthermore, scaling edge deployments requires solving messy, real-world challenges in talent, policy, and execution.

Hardware and Economic Constraints

While NPUs offer a more efficient solution than GPUs, they must be cost-effective across various form factors to achieve ubiquitous adoption. The industry is currently in an "era of AI inference," where decisions happen instantly at the edge, yet the deployment of robust hardware remains a hurdle in terms of both design and cost-effective availability.

Regulatory and Ethical Compliance

The decentralized nature of edge computing raises challenges regarding fragmented regulations. Research is moving toward GenAI-driven frameworks that can classify user privacy preferences and formalize them into standardized privacy policies (such as the S4P language) that dynamically align with local regulations. This ensures that all actors in the edge-cloud continuum respect user preferences and legal requirements.

Emerging Trend	Technology Driver	Future Impact
Agentic AI	Autonomous LLM Agents	Shift from diagnosis to "wellness partners"
Neuromorphic Computing	Spiking Neural Networks	Massive energy efficiency for edge AI
Quantum Edge	Quantum Neural Networks	Ultra-fast inference for complex systems
Speculative Decoding	OmniDraft Architecture	2x speed improvement for on-device LLMs

The future of edge AI will undoubtedly be driven by more advanced models, but unlocking their full potential depends on architectures that extract maximum value from every watt. As machines get better at interpreting context and intent, the boundary between human and machine interaction is entering a new phase defined by natural interfaces and adaptive intelligence. By bringing intelligence out of distant data centers and into the devices all around us, the Generative Edge is building a world that is not only faster and more helpful but also inherently respects privacy and works reliably in the real world.

CONCLUSION

The transformation of IoT through Generative AI is not merely an incremental upgrade but a fundamental change in the architecture of intelligence. The transition from discriminative classification to generative synthesis allows edge systems to handle the noisy, incomplete, and complex realities of the physical world with unprecedented autonomy. Through the integration of specialized NPUs, advanced compression strategies like structured pruning and 4-bit quantization, and privacy-preserving frameworks such as Federated Generative Learning, the "Generative Edge" is poised to revolutionize sectors ranging from industrial manufacturing and 6G communications to personalized healthcare and cybersecurity. The ultimate potential of these systems lies in their ability to act as context-aware collaborators, turning raw data into instant, actionable

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

insights while maintaining the rigorous standards of efficiency and security required for the next generation of intelligent edge ecosystems.

REFERENCES

1. Keskar, A., Malaga, M., Reddy, P., & Pattanayak, S. K. (2021). Generative AI, Chatbots, and IoT: The Future of Intelligent Interactions. Well Testing Journal, 30(1), 96-117.
2. Alavi, A., & Ranjbar, M. (2023). The role of generative AI in enhancing user experience in smart homes. Journal of Ambient Intelligence and Humanized Computing, 14(2), 123-135. https://doi.org/10.1007/s12652-022-03789-1
3. Du, D., Chen, R., Li, X., Wu, L., Zhou, P., & Fei, M. (2019). Malicious data deception attacks against power systems: A new case and its detection method. Transactions of the Institute of Measurement and Control, 41(6), 1590-1599. https://doi.org/10.1177/0142331218770660
4. Fieser, J. (2003). Ethics. Internet Encyclopedia of Philosophy. Retrieved from https://iep.utm.edu/ethics/
5. Hajiheidari, S., Wakil, K., Badri, M., & Navimipour, N. J. (2019). Intrusion-detection systems in the Internet of Things: A comprehensive investigation. Computer Networks, 160, 165-191. https://doi.org/10.1016/j.comnet.2019.07.015
6. Khan, M. A., & Salah, K. (2018). IoT security: Review, blockchain solutions and open challenges. Future Generation Computer Systems, 82, 395-411. https://doi.org/10.1016/j.future.2017.11.020
7. Mujeeb, S., Javaid, N., Ilahi, M., Wadud, Z., Ishmanov, F., & Afzal, M. K. (2019). Deep long short-term memory: A new price and load forecasting scheme for big data in smart cities. Sustainability, 11(4), 987. https://doi.org/10.3390/su11040987
8. Veres, M., & Moussa, M. (2019). Deep learning for intelligent transportation systems: A survey of emerging trends. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2019.2901234
9. Zhang, Y., & Wang, Y. (2021). Generative adversarial networks for traffic generation in mobile networks. IEEE Transactions on Network and Service Management, 18(3), 3000-3012. https://doi.org/10.1109/TNSM.2021.3081234
10. Evans, D. (2011). The Internet of Things: How the next evolution of the Internet is changing everything. Cisco Internet Business Solutions Group. Retrieved from http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf