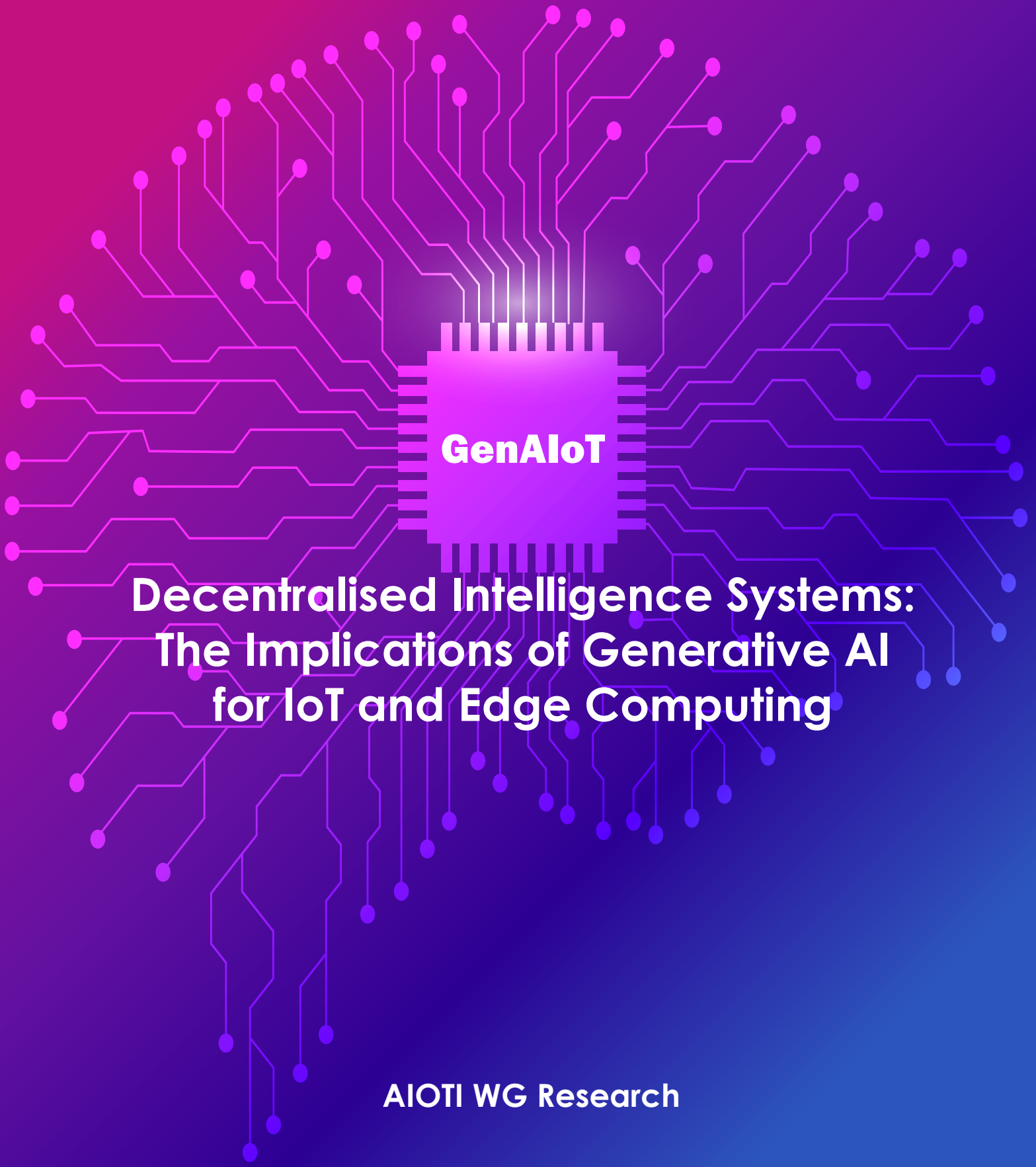


Decentralised Intelligence Systems: The Implications of Generative AI for IoT and Edge Computing

AIOTI WG Research

Alliance for IoT, AI and Edge Continuum Innovation

2026

Alliance for AI, IoT and Edge Continuum Innovation

GenAIoT
Decentralised Intelligence Systems: The Implications of Generative AI for IoT and Edge Computing

Release 1

AIOTI WG Research

10 November 2026

AIOTI. All rights reserved.

Executive Summary

Generative Artificial Intelligence (GenAI) is rapidly transitioning from a cloud-bound novelty to a defining workload for next-generation Internet of Things (IoT) and Edge Computing systems. This position paper argues that decentralising GenAI represents a needed and fundamental paradigm shift in how distributed networks operate. By migrating inference pipelines closer to physical devices, organisations can unlock critical improvements in latency, privacy, security, operational resilience, and cost. This paradigm shift takes the name of **Generative Artificial Intelligence of Things (GenAIoT)**.

To contextualise this shift, this position paper traces the evolution of intelligent IoT systems through five distinct phases. Historically, systems began in the "Connect" phase, acting merely as telemetry conduits, before advancing to a "Compute" phase, where localised processing reduced bandwidth. Recently, the "Reason" phase introduced task-specific machine learning (ML) for predictive analytics, and the nongenerative AI phase included discriminative AI elements focusing on analysing, classifying, or predicting outcomes from existing data rather than creating new content. Today, networks are entering the era of "Generative Reasoning," where systems can synthesise multimodal data, infer complex contexts, and generate dynamic outputs such as natural-language explanations or autonomous control strategies.

Realising the potential of this new era requires a shift away from massive, centralised training pipelines toward highly efficient, localised inference. This position paper explores the need for hybrid architectures that operate across the IoT-Edge-Cloud continuum. These architectures enable systems to balance the computational demands of foundation models with strict real-world deployment constraints, including stringent latency requirements, limited energy budgets, and stringent privacy mandates.

Decentralising intelligence alters the physical and logical makeup of network architectures. This position paper examines how GenAI reshapes Edge Computing, demanding novel resource allocation strategies and specialised hardware (HW) capabilities. Concurrently, it explores the deep implications for IoT devices and sensing systems, which are evolving from passive data collectors into capable, context-aware agents entrusted with localised decision-making.

The integration of advanced reasoning capabilities using GenAIoT into embodied edge AI marks a shift from reactive sensing to autonomous physical implementation. By enabling edge devices and agentic edge IoT autonomous systems to perform multi-step spatial and logical reasoning locally, these entities no longer just perceive their environment; they can plan, adapt, and execute complex task chains in real time, combined with modelling and simulation in virtual worlds using digital twins (DT) and immersive triplets (IMT). This localised cognitive capacity drastically reduces reliance on high-latency cloud feedback loops, allowing embodied agents to safely navigate unpredictable environments, re-evaluate actions mid-process, and interact dynamically with human partners. Ultimately, bringing reasoning directly to the physical edge AI and transitions agentic AI from isolated, digital decisionmakers to resilient, context-aware physical actors capable of true operational autonomy.

The integration of GenAI at the edge introduces a new landscape of operational vulnerabilities. This position paper critically evaluates security, safety, and trust considerations, highlighting the dangers of hallucination in decentralised models, adversarial attacks, and the physical safety implications of autonomous generative control. It addresses the challenges of privacy, data governance, and regulatory alignment, emphasising the immense operational complexity of managing the lifecycle of distributed models at scale.

This position paper concludes by detailing emerging developments and offering targeted recommendations for the industry. It outlines the architectural patterns needed to enable scalable GenAIoT systems and identifies the critical research directions for safely and effectively deploying the next generation of intelligent IoT and edge computing.

© AIOTI. All rights reserved.

Table of Content

Executive Summary 2
1 Introduction and Motivation 13
2 The Evolution of Intelligent IoT Systems 16
2.1 Connect: The Era of Telemetry 16
2.2 Compute: The Shift to Localised Processing 17
2.3 Reason: Task-Specific ML 19
2.4 Non-Generative AI (GenAI) - Discriminative AI 20
2.5 Generative Reasoning: The New Frontier of GenAIoT 23
2.6 GenAIoT, Embodied AI and Physical AI 29
Conceptual Overview of Embodied AI and Physical AI 31
3 Overview of GenAI Concepts 33
3.1 Definition and Classification of Foundation Models 34
3.2 Multimodal Capabilities in Physical Environments 35
3.3 Agentic AI 37
3.4 Future Research Trends 46
4 The IoT-Edge-Cloud Continuum 47
4.1 Data Interoperability as design principle 47
4.2 Shifting from Centralised Training to Edge Inference 50
4.3 Designing Hybrid Architectures for Distributed Intelligence 50
4.4 Future Research Trends 55
5 Implications for Edge Computing Infrastructure 57
5.1 Real-World Deployment Constraints 57
5.2 Architectural Patterns for Scalable GenAIoT 58
5.3 Future Research Trends 59
Quantization as the dominant compression mechanism 59
Pruning and structured sparsity 60
Knowledge distillation and the transition from LLMs to compact generative models 60
Beyond model weights: compression of the KV cache 61
HW-aware compression as an infrastructure requirement 61

© AIOTI. All rights reserved.

Implications for GenAIoT architectures and open challenges 62
6 Implications for IoT Devices and Sensing and Actuating Systems 64
6.1 Engineering implications and considerations 64
6.2 From task-specific sensing to reusable sensor foundation models 68
6.2.1 Data Spaces as a Sovereign Context Layer for GenAIoT 68
6.2.2 TinyML and Edge Intelligence as the Local Foundation of GenAIoT 70
6.3 Foundation models 71
6.4 Multimodal grounding in the physical environment 71
6.5 Intrusion Detection and Runtime Protection for Edge AI Inference 72
The Threat Landscape 72
Detection and Runtime Protection Methodologies 73
Operational Challenges 73
Emerging Research Directions 74
Evaluation requirements and open challenges 74
6.6 GenAI as a semantic layer over sensing 74
6.7 Edge–cloud collaboration and selective escalation 75
6.8 From generated interpretations to generated actions 76
6.9 World models, DTs and anticipatory action 76
6.10 Human–system interaction 77
6.11 Research challenges and emerging directions 77
6.12 Future Research Trends 78
7 Trustworthiness 80
7.1 Introduction 80
7.2 Key Challenges in Trustworthiness for GenAIoT Systems 81
7.3 Security and Safety Considerations 82
7.4 Privacy, Data Governance, and Regulatory Alignment 83
7.5 Future Research Trends 85
8 Ethics 87
8.1 Introduction 87
8.2 Challenges in Implementing Ethical Principles in GenAIoT Systems 88

© AIOTI. All rights reserved.

8.3 Future Research Trends and Directions 89

9 Standardisation and Regulations 91

9.1 Introduction 91

9.2 Challenges in Implementing Standardisation and Regulations in GenAIoT Systems 91

Challenge 1: Cloud Edge IoT Ecosystems to support Agentic AI 93

Challenge 2: Integrate Human Machine teaming 94

Challenge 3: Align with upcoming AI standards related to GenAI 95

Challenge 4: Ensure alignment of GenAI standards with other standards 95

Challenge 5: Support regulation 96

9.3 Future Research Trends and Directions 97

10 Recommendations 100

11 Conclusions 106

11.1 The Decentralisation Paradigm and the Evolution Toward Autonomous Edge Intelligence 106

11.2 The IoT Edge–Cloud Computing interoperability by means of Data Interoperability 106

11.3 The Physical Realities of Silicon, Packaging, and System-Level Co-Design 107

11.4 Operational Governance, Autonomous Engineering, and the Future Research Frontier 108

12 References 110

13 Contributors 118

14 Disclaimer 120

15 Acknowledgements 121

16 About AIOTI 122

© AIOTI. All rights reserved.

List of Figures

Figure 1-1 Decentralised GenAIoT conceptual evolution. 14

Figure 2-1 Evolution of intelligent IoT systems. 16

Figure 2-2 Non-Generative, discriminative AI 22

Figure 2-3 Non-Generative, discriminative AI characteristics. 22

Figure 2-4 Non-Generative, discriminative AI applications. 23

Figure 2-5 GenAI concept – how it works 23

Figure 2-6 GenAI workflow. 25

Figure 2-7 GenAI use cases examples. 26

Figure 2-8 Generative and non-GenAI differences animal photography. 27

Figure 2-9 Generative and non-generative handwritten digit classification 27

Figure 2-10 Transition from the classical discriminative inference paradigm to multimodal generative reasoning at the edge (GenAIoT) 28

Figure 2-11 Neuro-symbolic control architecture: decoupling stochastic inference from deterministic execution. 29

Figure 2-12 Layers representation of the concepts of GenAI, agentic AI, embodied and physical AI 30

Figure 2-13 Overview of embodied and physical AI representations and types of autonomous systems and use cases. 32

Figure 3-1 An overview of an agentic AI ecosystem. 37

Figure 3-2 Agentic AI architecture. 38

Figure 3-3 Agentic AI perception module layer. 38

Figure 3-4 Agentic AI cognitive/reasoning module layer. 39

Figure 3-5 Agentic AI action/execution module layer. 40

Figure 3-6 Agentic AI learning module layer. 40

Figure 3-7 Example of an agentic AI workflow for an industrial use case. 41

Figure 3-8 Agentic AI design principles example. 43

Figure 4-1 Edge Computing granularity across the computing continuum for GenAIoT processing. 51

Figure 5-1 Uncertainty-gated hybrid control: decision flow in each control cycle, applied per actuator; the checks run in order (symbols as defined in [142]). 58

Figure 7-1: RAI Adoption/Adaptation Framework 80

Figure 9-1 AI agent using a DT. 93

Figure 9-2 Examples of physical architecture allocations 93

Figure 9-3 Human-machine teaming example. 94

© AIOTI. All rights reserved.

List of Tables

Table 3-1 GenAI open challenges and technical vulnerabilities. 34

Table 4-1 Comparative Analysis of the Edge-to-Cloud Processing Continuum Tiers. 52

© AIOTI. All rights reserved.

List of Acronyms

3C	Connected Collaborative Computing Network
3D	Three dimensional
5G	Fifth generation of cellular network technology
6G	Sixth generation of cellular network technology
AhG	Ad hoc Group
AI	Artificial Intelligence
AIOTI	Alliance for AI, IoT and Edge Continuum Innovation
AIOTI SRIA	AIOTI Strategic Research and Innovation Agenda
API	Application Programming Interface
BDVA	Big Data Value Association
BLE	Bluetooth Low Energy
BOM	Bill of Materials
CoAP	Constrained Application Protocol
CPS	Cyber Physical System
CPU	Central Processing Unit
CSI	Channel State Information
D	Discriminator
DDPM	Denoising Diffusion Probabilistic Model
DL	Deep Learning
DPO	Direct Preference Optimisation
DRAM	Dynamic Random-Access Memory
DSP	Digital Signal Processor
DSSC	Data Space Support Centre
DT	Digital Twin
ELBO	Evidence Lower Bound
e-waste	Electronic waste
FW	Firmware
FPGA	Field-Programmable Gate Array
FUOTA	Firmware Updates Over-The-Air
G	Generator
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GenAIoT	Generative AI of Things
GNN	Graph Neural Networks
GOFAI	Good old-fashioned artificial intelligence
GPU	Graphics Processing Unit
HW	Hardware
IDS	Intrusion Detection System
IDSA	International Data Spaces Association
ILP	Instruction-level parallelism
IMT	Immersive Triplets
IMU	Inertial Measurement Unit
IoT	Internet of Things
IP	Intellectual Property
IVR	Immersive Virtual Reality
JCAS	Joint Communication and Sensing
KMPS	Kubernetes Multi-agent Proximal Policy Optimization Scheduler

© AIOTI. All rights reserved.

KV	Key Value
LoRA	Low-Rank Adapters
LoRa	Long Range Radio, Low-Rank Adapters
LoRaWAN	Long Range Wide Area Network
LLM	Large Language Model
LPDDR	Low-Power Double Data Rate
MAC	Multiply-Accumulate
MARL	Multi-Agent Reinforcement Learning
MAS	Multi-Agent Systems
MCU	Microcontroller Unit
MDP	Markov Decision Processes
MEC	Multi-access Edge Computing
MES	Manufacturing Execution System
ML	Machine Learning
MLP	Multi-Layer Perceptron
MPU	Microprocessor Unit
MQTT	Message Queuing Telemetry Transport
NFC	Near-Field Communication
NPU	Neural Processing Unit
OTA	Over-the-Air
PEFT	Parameter-Efficient Fine-Tuning
PLC	Programmable Logic Controller
PSRAM	Pseudo-Static Random-Access Memory
PTQ	Post-Training Quantization
QoS	Quality of Service
RAG	Retrieval-Augmented Generation
RAI	Responsible AI
RAM	Random Access Memory
RDF	Resource Description Framework
RF	Radio Frequency
RFID	Radio-Frequency Identification
RLHF	Reinforcement Learning from Human Feedback
RTOS	Real-Time Operating System
SCADA	Supervisory Control and Data Acquisition
SDK	Software Development Kit
SDN	Software Defined Networking
SFT	Supervised Fine-Tuning
SIFB	Service Interface Function Block
SIMD	Single Instruction, Multiple Data
SLA	Service Level Agreements
SLM	Small Language Model
SNN	spiking neural network
SoC	System on Chip
Software	SW
SOSA	Sensor-Observation-Sampling-Actuator ontology
SRAM	Static Random-Access Memory
SRIA	Strategic Research and Innovation Agenda
SSM	State-Space Model
SSN	Semantic Sensor Network
TEE	Trusted Execution Environment

© AIOTI. All rights reserved.

TPU	Tensor Processing Unit
USI	Uniform Resource Identifier
URLLC	Ultra-Reliable Low-Latency Communication
UWB	Ultra-Wide Band
VAE	Variational Autoencoder
VLA	Vision Language Action
Web	World Wide Web
Wi-Fi	Wireless Fidelity
XAI	Explainable AI

© AIOTI. All rights reserved.

List of Keywords

3D
5G
6G
Adaptability
Agentic AI
Agentic ecosystem
AI assistant
AI model
AI object detection
Architecture
Artificial intelligence
Autonomy
Backdoor attacks
Cloud nodes
Collaborative computing
Collaborative decision-making
Collaborative detection
Collaborative inference
Collaborative learning
Collaborative models
Collaborative network
Collaborative reasoning
Collaborative robots
Collaborative swarms
Compressed models
Computer vision
Contextual awareness
Continuous learning
Cross-sector deployment
Decision-making models
Design principles
Discriminative AI
Distributed intelligence
Distributed models
Deep-edge
Deployment models
Digital twin
Distributed models
Edge agents
Edge AI
Edge AI accelerators
Edge architectures
Edge-based generative intelligence
Edge computing
Edge devices
Edge ecosystems
Edge environments
Edge inference
Edge intelligence
Edge models
Edge-native data
Edge networks
Edge nodes
Edge platforms
Edge processors
Edge robotics
Embodied AI
Encoders
Ethical governance
Ethics
Ethics-by-Design
Explainable AI
Explainable generative AI
Federated learning
Forward simulations
Foundation models
Generative AI
Generative AI of Things
Generative engine
Generative models
Generative planning
Generative reasoning
Generative systems
Governance models
Hallucinations
Human co-workers
Human feedback alignment
Human-system interaction
Industry 5.0
Intelligent IoT systems
Internet of things
Interoperability
IoT-edge-cloud continuum
Large language model
Latency
Localised processing
Machine learning
Maintenance
Manufacturing
Meta-edge
Micro-edge
Models
Model training
Model hallucination
Modularity
Modular models
Multimodal
Non-Generative AI
Operational models
Physical AI
Physical edge
Physics-based models
Operational downtime
Physical reality
Predictive analytics
Predictive maintenance
Predictive models
Pretraining
Real-time simulations
Reinforcement learning
Remote API
Remote server
Regulations
Resource-intensive models
Retrieval-augmented generation
Rule-based systems
Safety
Scalability
Security
Semantic
Sensing pipeline
Sensor network
Standardisation
Supervised feedback alignment
Sustainability
System-description models
Task-specific machine learning
Telemetry
Training
Trustworthiness
Ultra-Reliable Low-Latency Communication
Ultra-wideband
Virtual agents
Virtual domain
Virtual reality
Virtual simulations
Virtual training
Virtual world

© AIOTI. All rights reserved.

© AIOTI. All rights reserved.

1 Introduction and Motivation

GenAI is transitioning from a cloud-centric focus to a distributed workload for next-generation Edge Computing and IoT systems, which can be called GenAIoT. It introduces the motivation for decentralising intelligence, outlining how pushing parts of data pipelines closer to end devices can significantly improve latency, privacy, resilience, and operational costs. Concurrently, it sets the stage for our critical examination of the new challenges this paradigm introduces across safety, security, governance, and model lifecycle management.

GenAIoT is driving a paradigm shift in decentralised intelligence by defining the operational workloads for next-generation IoT and Edge Computing systems. For decades, distributed computing architectures operated primarily as hierarchical pipelines that aggregated, compressed, and forwarded physical signals to centralised data centres.

The convergence of foundation models with physical endpoints marks a fundamental departure from this centralised paradigm, establishing the era of GenAIoT. Decentralising generative inference is no longer an alternative deployment preference; it represents a structural necessity driven by the physical and economic boundaries of modern networking.

To grasp the magnitude of this transition, we trace connected systems through distinct phases, which are discussed in more details in Section 2.

- In the foundational connect phase, connected devices functioned as passive telemetry conduits, streaming raw numerical measurements to centralised databases.
- The subsequent compute phase introduced localised processing, deploying edge gateways to filter high-frequency sensor noise and conserve backhaul network bandwidth.
- More recently, the reason phase incorporated task-specific, discriminative ML, enabling field HW to run predictive analytics and detect anomalies.
- The fourth evolutionary phase marked the transition from shallow ML models to deep discriminative artificial intelligence. In earlier phases, ML relied heavily on manual feature engineering, where human domain experts had to decide which specific statistical parameters an algorithm should track. Deep discriminative AI removed this limitation.
- Today, distributed networks are entering the generative reasoning phase. Connected systems are no longer restricted to outputting static classification tags or triggering human-authored heuristic thresholds. Instead, modern edge nodes are acquiring the capacity to synthesise high-dimensional, multi-modal data streams, infer latent environmental contexts, and generate complex outputs. These outputs range from nuanced natural-language health diagnostics to real-time, multi-step autonomous control strategies executed directly across physical actuators.

Realising the potential of this generative era requires shifting away from massive, centralised training pipelines toward efficient, localised inference architectures. The historic reliance on cloud-hosted intelligence introduces operational liabilities: high round-trip communication latency, vulnerability to network disruptions, excessive bandwidth expenses, and exposure of confidential sensory telemetry.

By contrast, migrating multi-modal generative inference directly to the physical periphery unlocks immediate operational responsiveness, preserves on-premises data privacy, reduces recurring telecommunications expenses, and ensures uninterrupted autonomous continuity even during complete wide-area network partitions.

The computational decentralisation alters the physical and logical compositions of edge infrastructure. It demands dynamic resource allocation strategies and specialised processing capabilities across the entire IoT-edge-cloud continuum as illustrated in Figure 1-1.

© AIOTI. All rights reserved.

At the same time, the physical endpoints of connected systems are undergoing an identity shift. Sensing nodes are evolving from passive, unthinking observation probes into context-aware, autonomous agents capable of independent situational interpretation and localised operational decision-making.

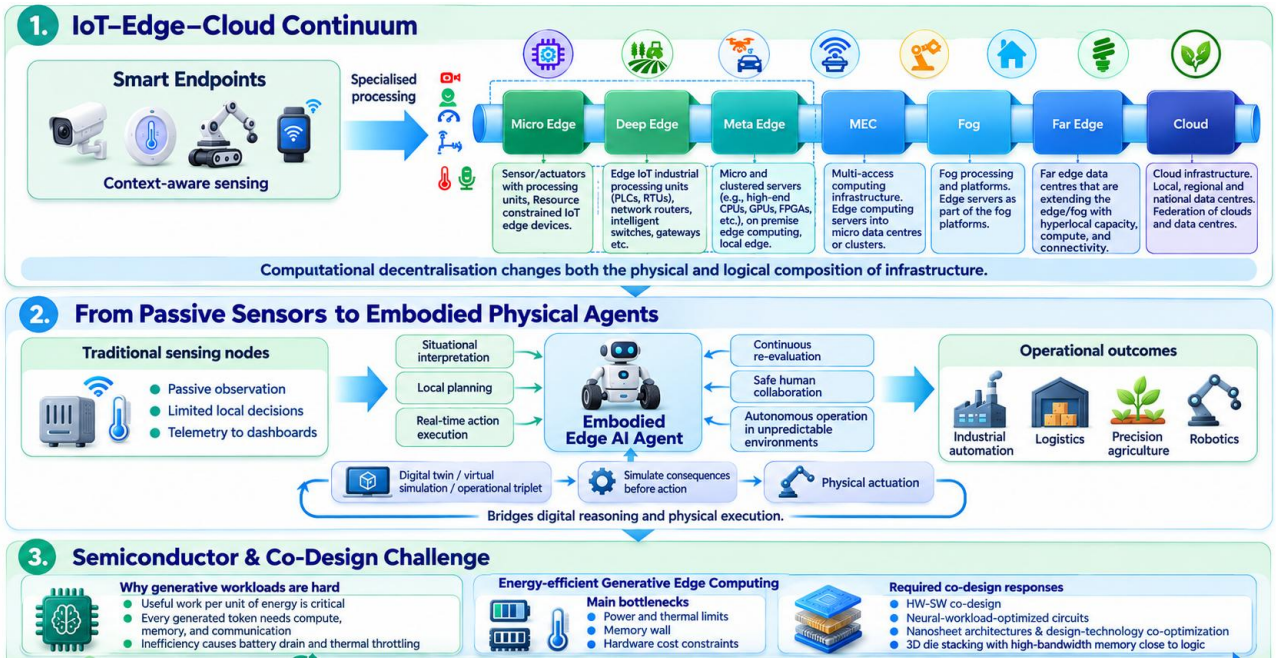


Figure 1-1 Decentralised GenAIoT conceptual evolution.

The primary motivation for decentralizing generative intelligence stems from the rise of embodied edge AI. In physical operational domains, such as industrial automation, logistics, precision agriculture, and robotics, intelligence cannot remain constrained within digital dashboards.

By embedding spatial and logical reasoning capabilities directly into physical field devices, autonomous systems transcend the limits of reactive sensing. These platforms no longer simply register the presence of an environmental obstacle or mechanical fault but can dynamically plan, evaluate, and execute intricate physical task sequences in real time.

This localized reasoning capacity enables embodied physical agents to navigate unpredictable physical environments, continuously re-evaluate action plans mid-operation, and collaborate safely alongside human co-workers without waiting for remote server roundtrips.

Coupling on-device generative reasoning with localized virtual simulations, DTs, and IMTs allows physical agents to simulate the forward consequences of their actions before physical actuation occurs.

This capability bridges the divide between abstract digital reasoning and physical execution, transforming connected systems from passive telemetry channels into resilient, physically context-aware actors capable of genuine operational autonomy.

Realising this vision exposes a fundamental semiconductor challenge: delivering scalable computing performance while keeping electrical energy consumption, thermal dissipation, and HW costs strictly bounded.

In generative workloads, useful work per unit of energy is the paramount design metric. Unlike traditional discriminative inference, where a single forward pass completes the task, generative sequence synthesis requires processing, memory access, and communication cycles for every

© AIOTI. All rights reserved.

generated token. In energy-constrained field deployments, marginal inefficiencies in energy per token quickly lead to battery exhaustion or severe thermal throttling.

Overcoming these physical limits demands comprehensive HW and software (SW) co-design across the semiconductor spectrum. At the transistor level, advanced manufacturing techniques such as nanosheet architectures and design-technology co-optimisation show that physical circuits must be engineered specifically for the neural workloads they run.

The memory wall significantly bottlenecks modern chip throughput, where execution units stall while waiting to fetch large model weight matrices from distant memory registers. Breaking this throughput bottleneck requires advanced multi-chip packaging, such as 3D die stacking, bringing high-bandwidth memory into direct vertical contact with processing logic to maximise throughput and minimise the energy required to move data.

As Edge Computing nodes grow more complex to handle multi-modal inputs, system-level communication between separate dies and packages is emerging as another performance limiter. High-speed, power-efficient interconnects, including integrated optical links, are becoming essential to eliminate the power penalties of traditional high-frequency electrical wiring as well as to increase the available bandwidth.

Managing the resulting architectural complexity requires deploying AI-assisted design automation tools to optimise cell placement, identify physical rule violations, and translate complex three-dimensional architectures into manufacturable silicon.

Progress in GenAIoT cannot rely solely on making isolated processing cores faster, rather it also depends on a unified systems approach: advancing the underlying silicon, integrating memory and compute physically closer together, and deploying robust, localised generative intelligence to solve real-world problems safely and sustainably at the physical edge. That can enable a quick transitioning from a cloud-bound novelty into the defining operational workload for the next generation of the IoT and Edge Computing systems as for decades, distributed computing architectures operated primarily as hierarchical pipelines that aggregated, compressed, and forwarded physical signals to centralised enterprise data centres.

At the same time, the physical endpoints of connected systems are undergoing an identity shift. Sensing nodes are evolving from passive, unthinking observation probes into context-aware, autonomous agents capable of independent situational interpretation and localised operational decision-making.

© AIOTI. All rights reserved.

2 The Evolution of Intelligent IoT Systems

The historical context for the current technological paradigm shift traces the development of IoT and edge architectures through distinct phases as illustrated in Figure 2-1, demonstrating how the integration of intelligence has progressively moved closer to the physical world.

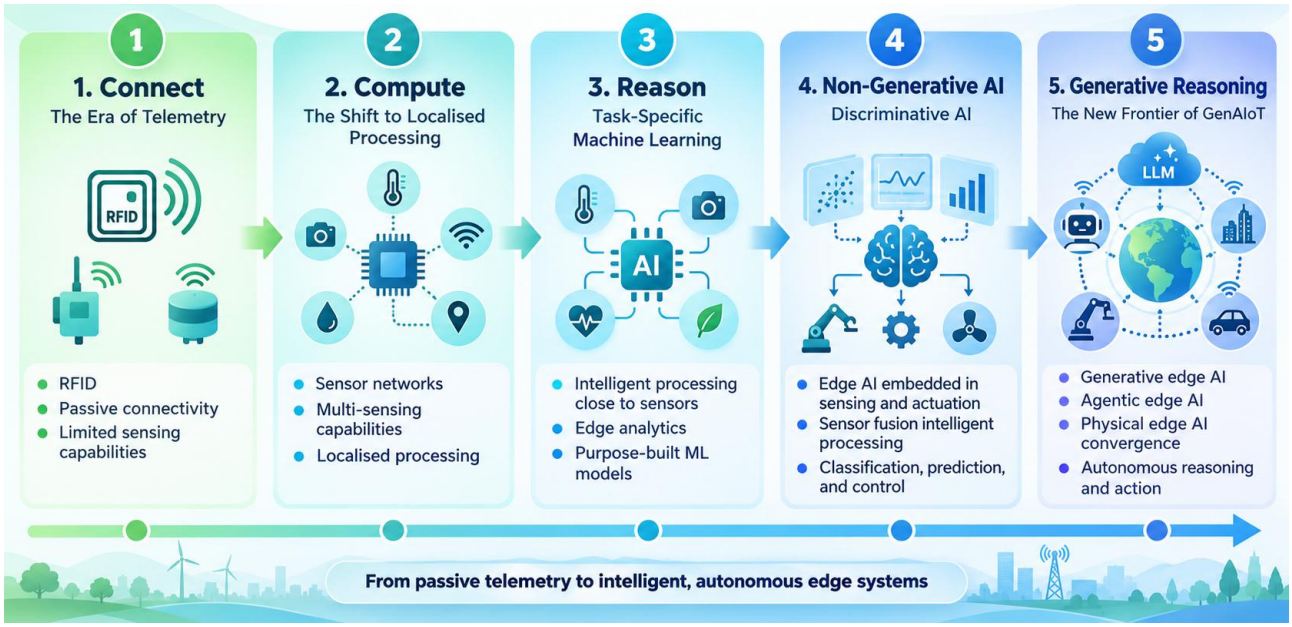


Figure 2-1 Evolution of intelligent IoT systems.

2.1 Connect: The Era of Telemetry

The foundational phase of IoT is characterized by simple devices that primarily served as data sources and by early sensor networks designed for transmitting raw telemetry to centralised systems for storage and eventual analysis.

The earliest era of the IoT was defined by a single objective: bridging the physical and digital worlds by gathering raw environmental measurements and streaming them to centralised servers. Field units operated as passive data-collection points, equipped with basic physical sensors that detected single variables, such as ambient temperature, liquid pressure, vibration intensity, or electrical flow.

The HW's primary task was straightforward and mechanical. An analogue sensor converted a real-world physical condition into a variable electrical voltage, which an on-chip converter then digitised. The local processing unit had no awareness of what the numbers meant. Its sole function was to package these raw numeric readings into small data packets, pass them to a wireless transmitter, and then enter an inactive sleep mode to save energy.

Because the electronic processing circuits were tiny and low-powered, they had minimal memory, modest processing speeds, and no dynamic operating systems. They lacked the memory and computational headroom to evaluate the data they collected, compare it to historical trends, or run local diagnostic code.

The network layer was engineered around the realities of low-power radio communication. Sensor outposts were frequently positioned in remote areas, factory basements, or sprawling agricultural fields where direct access to continuous grid power was non-existent. Devices relied on small batteries or basic ambient energy harvesters that were expected to last for years without replacement. As a result, standard enterprise networking techniques, which require high bandwidth and continuous handshakes, were unsuitable.

© AIOTI. All rights reserved.

The industry addressed the mentioned constraints by developing lightweight networking standards and low-overhead communication protocols. Engineers created adaptation techniques that let standard internet addresses travel over short-range, battery-conserving radios like Radio Frequency Identification (RFID), Near-Field Communication (NFC), Zigbee and Bluetooth Low Energy (BLE), as well as longer-range technologies such as LoRaWAN. Rather than using web data formats, systems embraced lightweight messaging standards like Message Queuing Telemetry Transport (MQTT) and Constrained Application Protocol (CoAP).

These protocols stripped out redundant transmission headers and relied on simple publish-and-subscribe models, keeping messages tiny and allowing low-power field units to quickly push updates to a central broker without maintaining battery-draining open connections. At the central destination, typically a corporate data centre or an early enterprise cloud, large data pipelines ingested this incoming flood of sensor readings. Centralised ingestion servers parsed continuous streams of time-stamped information and wrote them directly to specialised time-series databases. These early cloud backends served as digital filing cabinets. They let operators chart long-term operational trends, generate daily status reports, and trigger basic, reactive warning notifications whenever a reading crossed a manually defined threshold.

While this telemetry framework marked a major technological milestone by making isolated industrial infrastructure remotely visible, it introduced severe systemic limitations as deployments scaled from hundreds of sensors to tens of millions of active devices.

The first major bottleneck was the transmission to a central location. Because early sensors had to transmit every raw data point to the cloud for even basic analysis, their batteries drained quickly. Moreover, deployments in factories or urban settings encountered heavy radio congestion. As thousands of unsynchronised endpoints transmitted within the same unlicensed frequency bands, signals collided, requiring repeated retries those depleted batteries even faster and flooded local networks.

The second core limitation was latency and operational dependence. Because all reasoning was centralised in remote servers, the field site could not react autonomously to physical events. If an industrial pipeline sensor detected a dangerous pressure surge, that measurement had to travel across local radio links, pass through an edge gateway, traverse wide-area internet infrastructure, and wait in an ingestion queue before a central server could evaluate it.

Once the server identified the hazard, an emergency shutoff command had to travel back down the network pipe to an actuator. In safety-critical industrial operations, this round-trip delay was often unacceptably slow. Moreover, if the local internet backhaul connection was temporarily severed, the field site was blinded and could not react to dangerous operational failures.

Telemetry generated significant quantities of low-value, repetitive data. A temperature probe mounted on an industrial motor running at a steady state might send identical normal readings every second for months on end. Pushing large amount of these redundant, unvarying readings over costly telecommunications networks and paying to store them in cloud databases created mounting operational expenses without providing meaningful business intelligence.

It became evident that centralised telemetry was structurally unsustainable, paving the way for the next evolution: processing data locally at the edge.

2.2 Compute: The Shift to Localised Processing

The advent of Edge Computing, driven by the need to reduce latency and bandwidth consumption, spurred the industry to introduce localised processing, enabling basic data filtering and aggregation near the point of generation.

© AIOTI. All rights reserved.

The shift to localised processing transformed the architectural structure of connected devices. As the drawbacks of centralised cloud telemetry became clear, the design paradigm inverted: rather than routing raw operational data to distant centralised processing units, SW execution moved to the local network perimeter.

This evolution was made possible by the rapid development of low-power, high-performance silicon. Compact embedded processors gained multi-core processing architectures, HW floating-point support, and expanded volatile memory once restricted to enterprise desktop computers and servers. Field devices were no longer dumb endpoints connected to basic microcontrollers; they were paired with intelligent local gateways, ruggedised industrial PCs, and advanced programmable logic controllers that could run full-featured operating systems and complex SW stacks.

Instead of transmitting continuous streams of raw numbers, edge gateways served as real-time data filters and aggregators. They were programmed to ingest high-frequency sensor readings, validate signal integrity, and perform immediate digital signal processing directly on site. For example, high-rate vibration readings from machine components were no longer pushed across the internet sample by sample.

Instead, the edge gateway gathered the raw electrical waves over short time windows, ran mathematical frequency transformations locally, and translated raw vibration curves into clean operational metrics, such as overall peak frequency and running averages. The gateway then forwarded only concise, deduplicated diagnostic summaries to the central cloud.

Alongside data reduction, the compute phase introduced localised, low-latency closed-loop control. Industrial systems were equipped with rule-based automation engines running directly on local gateways. These engines evaluated continuous sensor data streams against local operational parameters, enabling sub-millisecond physical reactions.

If a cooling pump exceeded its safe operating temperature or an automated assembly line arm drifted out of alignment, the local gateway detected the breach immediately and tripped an emergency safety shutoff across the local factory floor bus. This protective response executed reliably within milliseconds, completely insulated from wide-area internet lag or sudden connectivity outages.

To manage SW deployed across diverse edge environments, the industry adopted micro-virtualisation and modern container frameworks. Modular SW deployments replaced traditional monolithic firmware (FW) updates, which required re-flashing an entire physical microchip and risked bricking devices in the field.

Engineers packaged edge applications into lightweight SW containers and sandboxed execution environments. These modular components could be securely dispatched, launched, isolated, and updated remotely across thousands of field gateways without interrupting other parallel control tasks running on the same machine.

However, moving active computing out of clean, climate-controlled corporate facilities to the untamed network edge introduced significant technical and operational hurdles:

The foremost challenge was achieving real-time determinism within general-purpose operating systems. Standard computing environments prioritise overall execution throughput over strict timing guarantees; background tasks and memory management routines can cause unpredictable delays.

In precision industrial manufacturing, robotics, and power distribution, unpredictable execution lag is unacceptable. An automated safety stop that executes twenty milliseconds late can destroy equipment or injure personnel.

© AIOTI. All rights reserved.

Engineers had to develop real-time system kernels and hybrid multi-core processor architectures. In these configurations, high-performance computing cores ran standard operating systems to manage web interfaces, network routing, and data analytics, while coupled, real-time processor cores ran bare-metal operating systems dedicated entirely to meeting hard physical control deadlines.

A second major hurdle was distributed data consistency and fleet management. When hundreds of edge gateways maintain their own local configurations, log files, and decision states while experiencing intermittent network disconnections, keeping the entire ecosystem aligned becomes difficult. Edge platforms needed specialised reconciliation mechanisms to ensure that when an offline gateway reconnected to the broader enterprise network, its local actions and stored operational logs merged cleanly without creating conflicting states.

Finally, Edge Computing HW faced physical durability and longevity constraints. Devices deployed in mining fields, marine installations, and remote substations had to operate reliably inside sealed, metal cases without fans, exposed to wide temperature swings, electrical noise, and constant physical vibration.

Constantly writing high-frequency diagnostic logs locally risked prematurely burning out solid-state and flash storage chips. System architects had to redesign operating systems to run primarily from volatile system memory, minimising physical storage writes and ensuring edge infrastructure could operate unattended for years without maintenance.

2.3 Reason: Task-Specific ML

The integration of traditional predictive AI into edge systems and the deployment of task-specific ML models enabled IoT networks to achieve localised predictive and prescriptive capabilities, such as anomaly detection or predictive maintenance.

While the compute phase enabled localised data filtering and rule-based actuation, system logic remained strictly bounded by human-authored conditional code. An edge gateway could execute complex instructions, but it was limited to rigid criteria a programmer designed in advance, such as checking whether a value had breached a fixed threshold.

Phase three brought ML directly to edge endpoints, shifting device behaviour from pre-set rules to statistical pattern recognition.

In industrial environments, machines rarely fail abruptly without warning; instead, they display subtle changes across multiple variables long before a breakdown occurs. Traditional threshold rules miss these early indicators because individual metrics, like a minor rise in temperature or a slight shift in motor balance, often remain within normal tolerances. Task-specific ML changed this by teaching edge systems to continuously evaluate multivariate correlations and identify subtle structural anomalies hidden deep within noisy real-world data.

Early implementations in this phase coupled local feature extraction with classical statistical models. Instead of passing raw data directly to an algorithm, the edge processor extracted meaningful statistical indicators from incoming signals, such as calculating signal variance, tracking wave shapes, and measuring changes across frequency spectra. These extracted features were fed into efficient predictive models, including support vector machines, decision trees, and shallow neural networks.

These algorithms were trained on historical failure data to recognise specific real-world problems, such as identifying early bearing fatigue in an industrial fan, detecting fluid blockages in chemical pumps, or classifying acoustic anomalies inside high-voltage electrical transformers.

This era also saw the emergence of the TinyML movement, which proved that trained neural networks could run on tiny, ultra-low-power microcontrollers operating on milliwatts of power. By

© AIOTI. All rights reserved.

designing neural SW frameworks that completely avoided dynamic memory allocations and fit inside static memory limits, engineers deployed predictive capabilities directly onto the smallest field nodes. An inexpensive, battery-operated vibration sensor could now evaluate mechanical health locally, infer operational degradation, and report high-level diagnostic assessments instead of passive telemetry numbers.

Deploying task-specific ML models onto resource-constrained physical endpoints introduced steep optimisation and lifecycle challenges.

The primary technical hurdle is model compression. Modern ML models are typically trained within powerful cloud data centres using high-precision, floating-point arithmetic. These models often consume hundreds of megabytes of memory, are exceeding the memory budgets of low-power edge processors, which typically have less than one megabyte of storage.

To bridge this divide, researchers developed quantisation and pruning techniques. Engineers converted high-precision mathematical parameters into compact eight-bit integers and pruned redundant connections between artificial neurons. This allowed models to shrink down to a fraction of their original footprint with negligible drops in accuracy, enabling integer-only microcontrollers to run advanced inferences in real time.

Maintaining model accuracy over time created another critical challenge: concept drift. Physical machines change as they age; mechanical tolerances widen, lubricants degrade, and seasonal temperature swings alter baseline sensor readings. A model trained on a brand-new machine operating in a controlled test lab often generated false positives when deployed on an older machine in an open-air factory.

Because edge devices lacked the memory and processing power to retrain their neural networks on the fly, teams had to periodically retrain models on newly captured data and redeploy them across the physical fleet.

To update models without transferring massive amounts of private raw data back to centralised cloud servers, the industry explored decentralised training frameworks, most notably federated learning. In this setup, edge endpoints used their locally collected data to compute incremental model adjustments. Instead of sharing raw sensor streams, the devices transmitted only their mathematical weight updates back to a central coordinating server. The central server aggregated these individual adjustments into an improved global model and distributed the updated version back out to the fleet.

Federated learning encountered real-world distribution challenges. Because devices were installed across diverse geographic regions and operating environments, the data collected by one sensor rarely matched the patterns seen by another. This data variance often pulled local models in conflicting directions, requiring sophisticated aggregation techniques to keep the collective model stable, accurate, and fair across varied operating contexts.

2.4 Non-Generative AI (GenAI) - Discriminative AI

The fourth evolutionary phase marked the transition from shallow ML models to deep discriminative artificial intelligence. In earlier phases, ML relied heavily on manual feature engineering, where human domain experts had to decide which specific statistical parameters an algorithm should track.

Many deep neural networks (such as standard CNNs for image classification or RNNs for sequence labelling) are discriminative models. They learn decision boundaries to map inputs directly to output classes or labels by estimating the conditional probability

Deep discriminative AI removed this limitation. By leveraging selected deep neural networks (such as standard CNNs for image classification or RNNs for sequence labelling) or primarily

© AIOTI. All rights reserved.

multi-layered convolutional neural networks, these systems are discriminative and learned to extract complex, hierarchical features directly from raw, unstructured data feeds like high-definition video, multi-axis LiDAR point clouds, and raw acoustic audio streams.

Discriminative AI systems focus on classification and separation tasks. Rather than generating new media, they aim to determine what an existing input is and assign it to a predefined category or bounding box. The network analyses an incoming visual or spatial input, evaluates the complex relationships between pixels or signal values, and calculates the statistical probability that the input belongs to a specific class: identifying a manufacturing defect, categorising a surface crack, spotting a worker lacking protective equipment, or segmenting obstacles along an autonomous warehouse vehicle's path.

This SW leap was accelerated by specialised edge silicon: Neural Processing Units (NPUs). Standard central processors execute instructions sequentially, which makes the billions of matrix math calculations deep networks require slow and power-hungry. NPUs, by contrast, are purpose-built to accelerate neural operations.

They use specialised HW layouts called systolic arrays, where calculations flow through a synchronised grid of processing elements. The processing elements pass intermediate math directly to neighbouring units without constantly reading and writing data back to main system memory. This architecture enabled edge devices to process high-resolution video streams at sixty frames per second while consuming only a few watts of power, bringing real-time visual inspection directly to physical production lines. Despite the processing breakthroughs enabled by edge neural accelerators, discriminative AI introduced complex design trade-offs:

The primary HW challenge was the memory wall. While NPUs can process calculations at tremendous speeds, the energy required to fetch large neural network weight matrices from external memory chips to the compute core is hundreds of times higher than the energy needed to execute the math itself.

In industrial enclosures, continuous memory transfers generate intense heat. If an edge processor runs sustained workloads without sufficient thermal dissipation, it overheats and throttles its operational clock speed. This thermal throttling causes dropped video frames and non-deterministic processing delays that undermine real-time safety systems. To address this, HW designers adopted techniques that compress neural-layer representations and keep intermediate operations entirely within ultra-fast, on-chip static memory.

Another major challenge is model calibration and overconfidence. Standard deep discriminative networks are trained to map all inputs into a fixed set of predefined categories. When presented with completely novel, out-of-distribution real-world inputs, such as a piece of machinery damaged in an unprecedented way or an unusual optical reflection, the model cannot naturally recognise that it is looking at something unknown.

Instead, its mathematical classification layers often output an incorrect prediction with high statistical confidence. In safety-critical industrial applications, such as high-speed robotic manufacturing or transportation monitoring, acting decisively on an overconfident, incorrect classification can cause catastrophic equipment damage or shut down continuous assembly lines.

Deep discriminative perception proved vulnerable to physical noise and sensor interference. Because deep networks rely on high-dimensional mathematical decision boundaries, minor, structured distortions in real-world sensor streams can alter an algorithm's output.

In industrial environments, these vulnerabilities emerged organically: intentional laser glare on optical cameras, acoustic resonances interfering with internal chip sensors, or heavy electromagnetic interference from nearby high-voltage arc welders could warp input signals just enough to flip a model's classification from safe to critical failure, or vice versa. Hardening

© AIOTI. All rights reserved.

deep discriminative models to remain stable and reliable amid the electrical and physical noise of industrial environments became a central engineering priority.

Non-generative artificial intelligence, frequently referred to as discriminative AI, represents a foundational class of ML focused on evaluating, categorizing, and predicting rather than creating new content. While generative systems grab headlines for producing text, art, or audio from scratch, discriminative systems are built to analyse existing information and make clear distinctions as illustrated in Figure 2-2. Their core objective is to map data directly to specific labels, determining what something is rather than imagining what it could be.

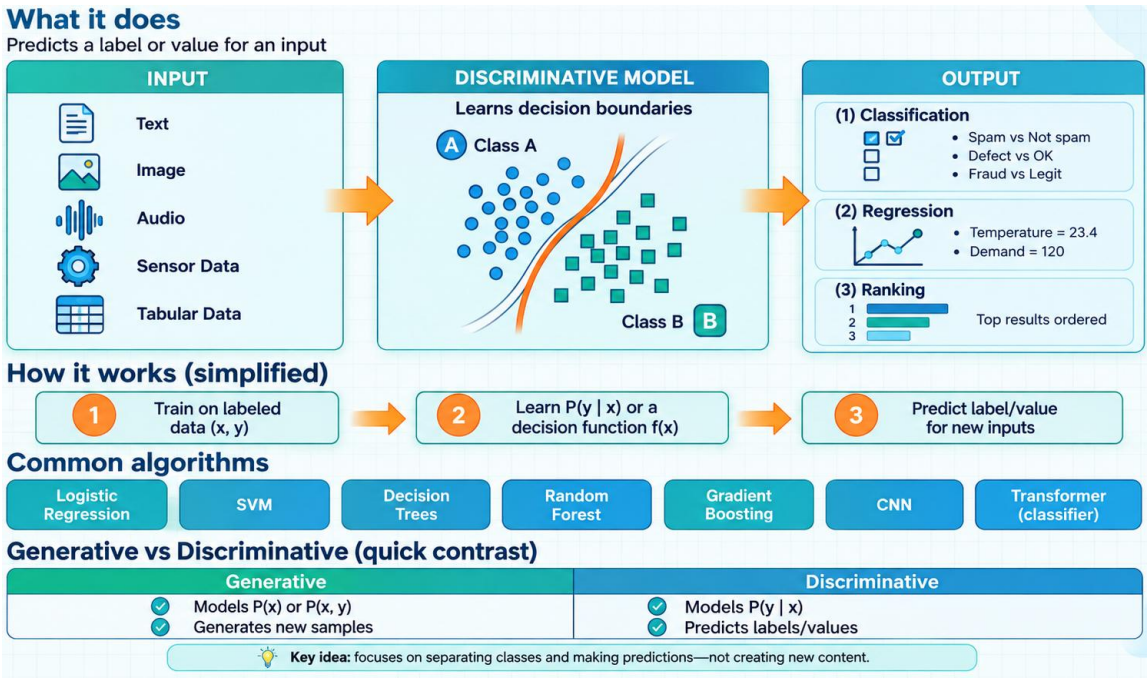


Figure 2-2 Non-Generative, discriminative AI.

At the heart of discriminative AI is supervised learning, which relies heavily on high-quality, labelled datasets for training. During this process, the model inspects data points that have already been tagged with predetermined categories, learning to identify the precise boundaries and features that separate one class from another. By mapping these underlying relationships, the model builds an internal framework that allows it to evaluate unseen data and reliably assign it to the correct category based on the patterns it observed during training.

This focused approach makes non-GenAI essential for a wide array of everyday technological applications. Because these models excel at pinpointing distinguishing characteristics, the power mission-critical tasks like filtering spam from inboxes, performing sentiment analysis on customer feedback, recognizing spoken words, and classifying objects within digital images.

The elements characterising non-generative, discriminative AI are presented in Figure 2-3.

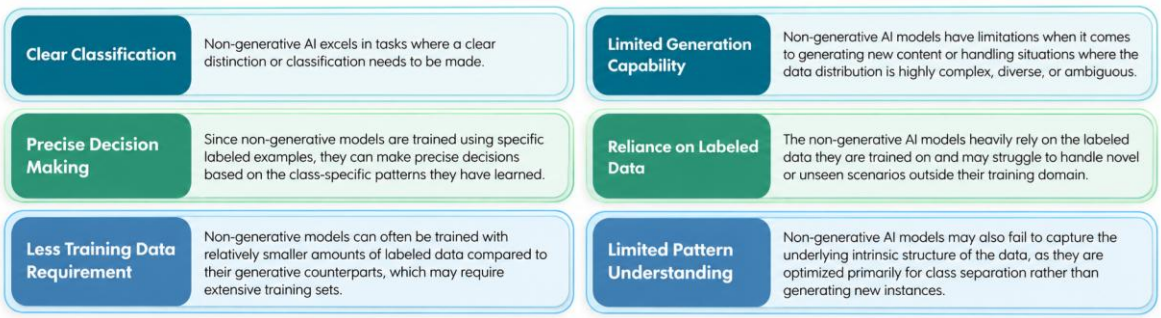


Figure 2-3 Non-Generative, discriminative AI characteristics.

© AIOTI. All rights reserved.

To achieve this level of precision, discriminative AI leverages a versatile toolkit of algorithms. Depending on the complexity of the data, practitioners deploy classical statistical techniques like logistic regression and decision trees, advanced ensemble methods such as bagging and boosting, or sophisticated neural networks equipped with tailored loss functions.

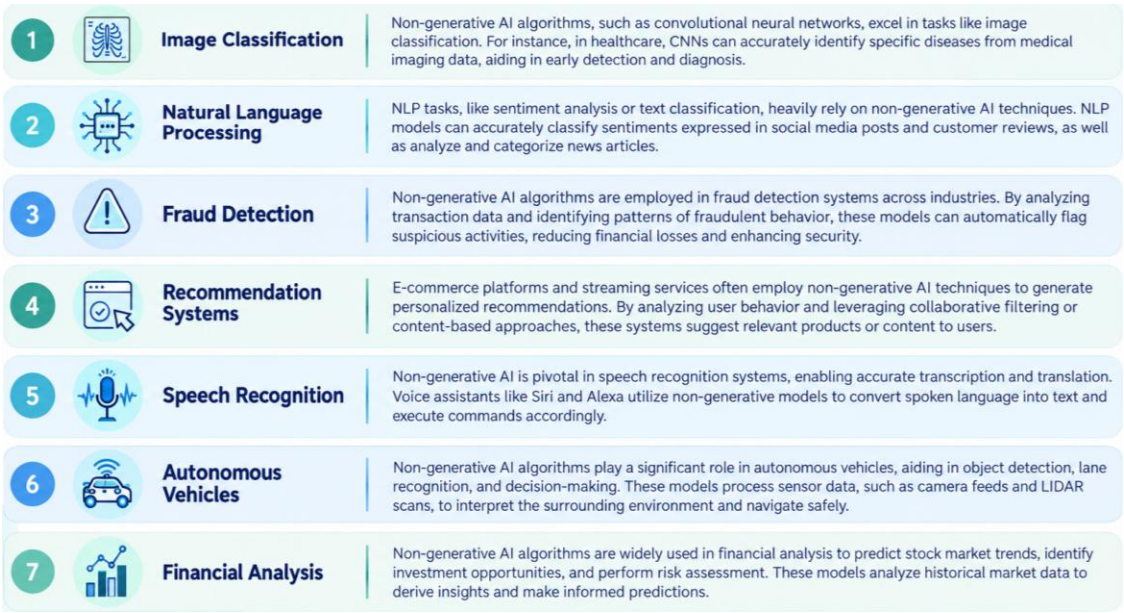


Figure 2-4 Non-Generative, discriminative AI applications.

Together, these tools form the analytical backbone of modern automated decision-making. Several non-GenAI applications are presented in Figure 2-4.

2.5 Generative Reasoning: The New Frontier of GenAIoT

Generative artificial intelligence represents a paradigm shift in ML, moving beyond traditional analysis to focus on the creation of entirely new, original content. Rather than simply evaluating existing inputs, generative models can produce novel artifacts across diverse media, including coherent text, synthetic audio, realistic images, and dynamic video. While conventional artificial intelligence techniques primarily specialize in recognizing and categorizing incoming information, GenAI takes a step further by using what it has learned to build something fresh.

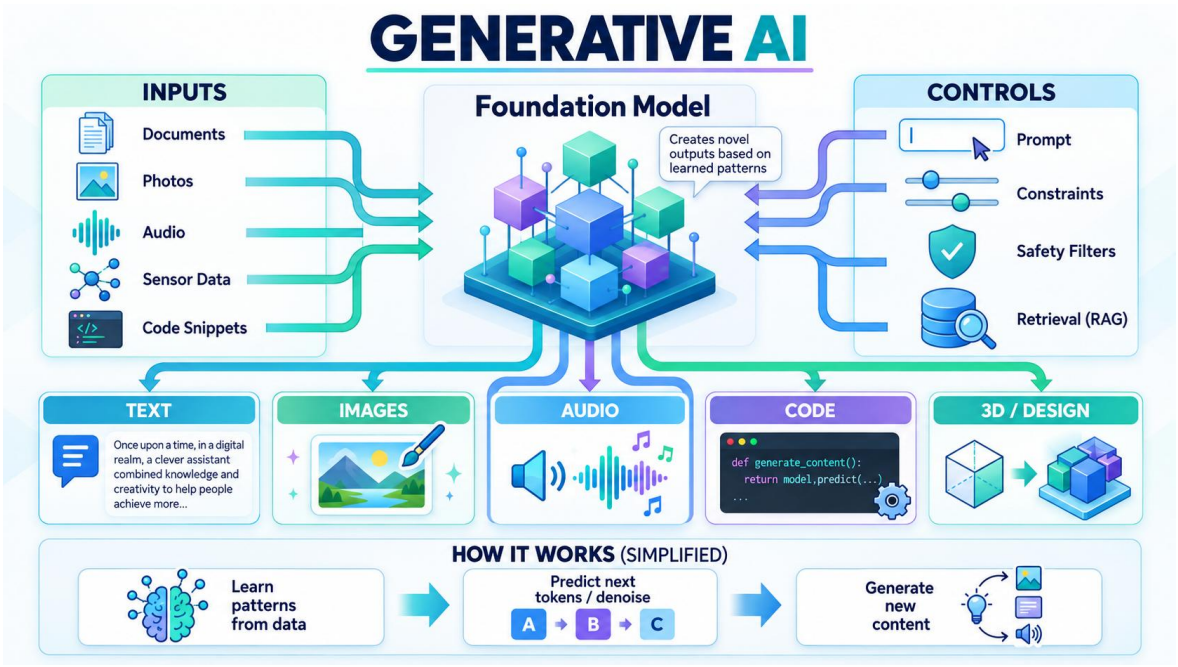


Figure 2-5 GenAI concept – how it works.

© AIOTI. All rights reserved.

The fifth and current frontier of connected systems is integrating Generative Artificial Intelligence with the Internet of Things, creating the GenAIoT paradigm. This era shifts systems from passive discriminative classification to active generative reasoning, dynamic contextual understanding, and autonomous physical problem-solving.

Rather than merely labelling an observed input as a defect or anomaly, a generative IoT system synthesises diverse observations, infers underlying root causes, generates natural language diagnostics, and dynamically plans complex remediation actions.

This breakthrough is powered by compact Small Language Models (SLMs) and multi-modal Vision-Language-Action models running directly on local edge HW or partitioned cooperatively across distributed edge-cloud fabrics. Unlike the massive generative models deployed in enterprise data centres, edge-optimised models are scaled down to between one and four billion parameters.

They are compressed using extreme low-bit quantisation techniques that let them run in a few gigabytes of memory. These compact generative engines ingest multi-modal tokens simultaneously: they process numerical sensor time-series, machinery log entries, thermal maps, and acoustic recordings within a shared understanding of system context.

In this paradigm, connected systems shift from passive sensory monitors to autonomous, agentic problem-solvers. When an anomaly occurs on an industrial site, a GenAIoT system does not simply throw a generic alarm code. It analyses continuous sensor readings, retrieves historical maintenance records, inspects operating manuals, and produces a complete situational assessment.

Furthermore, these generative systems have direct tool-use capabilities. Through structured function calling, the model can synthesise a corrective remediation plan, author the necessary operational scripts, and dispatch control commands directly across local industrial buses to dynamically adjust valve positions, redirect electrical loads, or slow down mechanical feed rates to prevent catastrophic system failures.

GenAI also enables dynamic, predictive DTs, rather than relying solely on static, pre-programmed physics equations, on-device generative models simulate forward operating trajectories in real time.

By analysing current machinery workloads, localised temperatures, and historical wear patterns, the generative engine simulates multiple plausible operational futures. It projects how internal components will degrade over the coming days under varying operational demands, allowing industrial facilities to safely push productivity to physical limits without triggering sudden structural failures.

While the potential of GenAIoT is transformative, deploying generative models into physical, mission-critical environments introduces unprecedented engineering and safety hurdles:

The primary computational bottleneck is managing operational context memory. Generative transformer architectures generate outputs sequentially, one token at a time. To maintain an understanding of the ongoing situation, the model must store the internal mathematical representations of all preceding tokens within an active memory structure known as the Key-Value cache.

In continuous IoT environments, where machines stream high-frequency telemetry non-stop, this active context cache expands rapidly. Within hours of continuous monitoring, the memory required to track the operational context can surpass the memory needed to store the model itself, overwhelming edge HW.

To overcome this, system architects are pioneering non-transformer architectures, such as state-space models and recurrent linear networks. These emerging designs process continuous, infinite operational contexts with flat, bounded memory footprints, making them ideal for long-duration industrial tracking.

© AIOTI. All rights reserved.

An even more acute challenge is hallucination and operational non-determinism. Generative models are inherently probabilistic; they construct outputs by sampling from statistical distributions.

In physical engineering environments, such as nuclear power plants, municipal water grids, or automated chemical refineries, probabilistic unpredictability poses serious safety risks. A generative model might hallucinate an incorrect operating state, invent a non-existent maintenance procedure, or generate a control command that breaches the safe pressure rating of a physical pipe.

To safely deploy generative intelligence into real-world control loops, modern GenAIoT architectures wrap generative reasoning engines inside deterministic safety guardrails. Engineers use grammar-constrained decoding techniques to strictly force models to output structured, validated function commands.

Crucially, the generative agent's suggested actions are passed through a deterministic verification layer governed by strict mathematical and physical invariants. If the generative model suggests an operational adjustment that violates established physical limits, rates of change, or safety tolerances, the deterministic safety layer instantly blocks the command and drops the system into a safe fallback state.

By combining the contextual reasoning of generative models with the predictable safety checks of classical control engineering, the GenAIoT paradigm establishes a framework for truly autonomous physical intelligence.

At its core, GenAI relies on a collection of probabilistic models and deep learning (DL) architecture designed to map the intricate distributions and underlying structures within large datasets. By studying these latent patterns, the system doesn't merely memorise its training data; instead, it develops a deep statistical intuition about how that data is formed. This allows the model to sample from that learned distribution and generate brand-new outputs that faithfully mirror the style, characteristics, and logic of the original inputs.

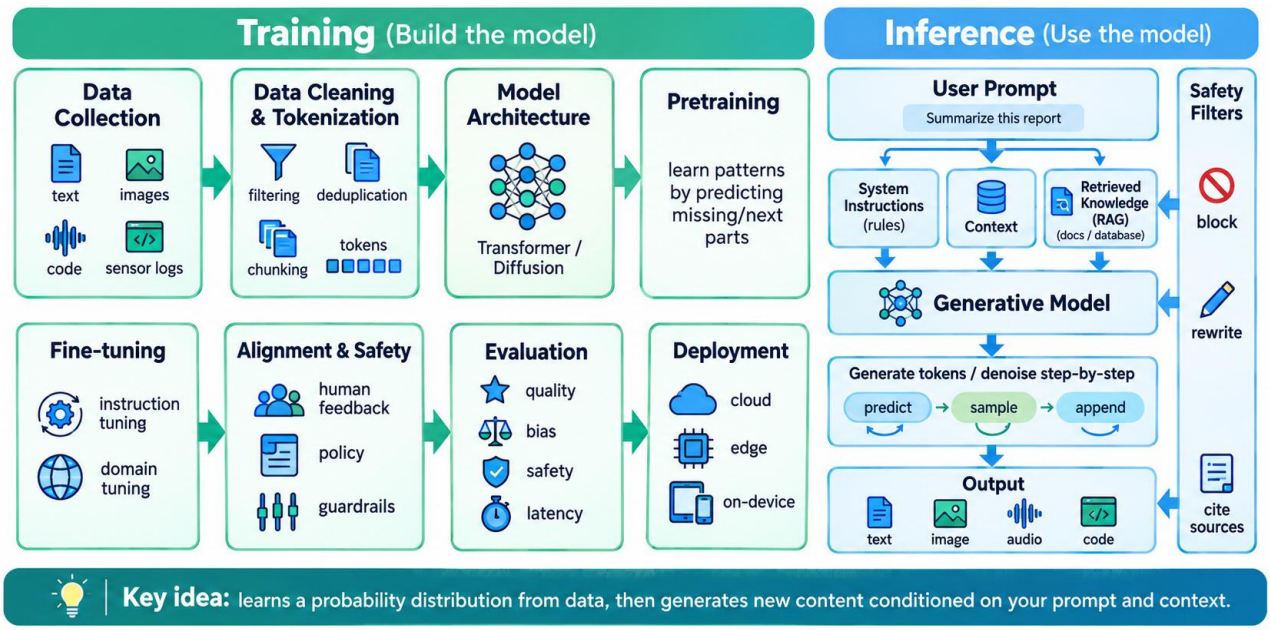


Figure 2-6 GenAI workflow.

This generative capacity unlocks powerful capabilities across major technological fields. In natural language processing, these models can draft structured writing, summarize complex documents, and sustain fluid, human-like conversations. Within computer vision, generative systems can synthesize hyper-realistic imagery from scratch, enhance existing visuals, or perform inpainting to seamlessly reconstruct missing portions of a corrupted picture.

© AIOTI. All rights reserved.

Beyond direct creative outputs, GenAI serves as a critical engine for advanced technical and scientific workflows. It plays a vital role in data augmentation, synthesizing realistic data points to train other ML systems where real-world data is scarce or sensitive. Additionally, these models are instrumental in detecting anomalies by flagging inputs that deviate from learned distributions, as well as simulating complex physical, biological, or economic environments for research and testing.

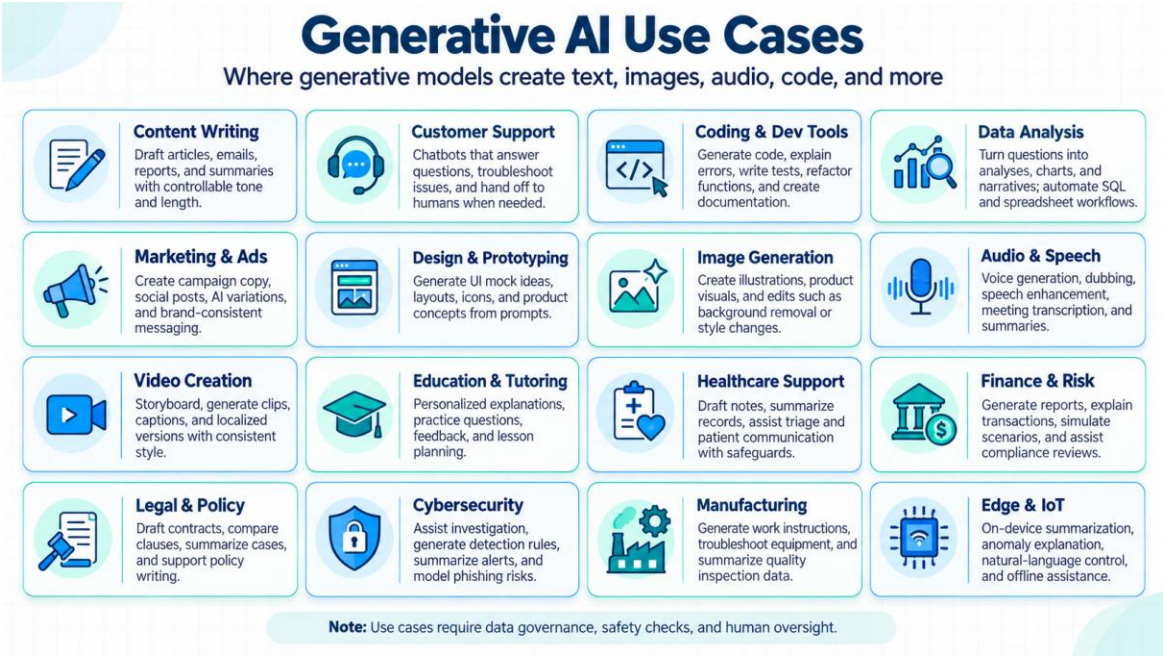


Figure 2-7 GenAI use cases examples.

Examples of GenAI use cases are given in Figure 2-7. GenAI models are designed to produce a wide array of rich outputs, ranging from fluent written passages and dynamic audio tracks to intricate visual scenes. While these generated assets closely mirror the style, structure, and aesthetic qualities of the data used to train the system, they are not simple cut-and-paste replicas or plagiarized fragments. Instead, the model crafts novel artifacts that feel organically related to the source material while existing as entirely new, independent creations.

To achieve this level of originality, generative models focus on learning the underlying probability distribution of their training datasets. Rather than memorizing individual data points, the architecture maps out the multi-dimensional space defining what makes the input data valid and coherent. Once the model grasps this internal distribution, it can navigate that statistical space and sample new points within it. The resulting outputs naturally align with the rules, features, and constraints of the original domain without duplicating any single past example.

A classic illustration of this principle can be seen in computer vision models trained on animal photography as illustrated in Figure 2-8. If a generative system is fed a dataset containing thousands of cat images, it does not simply store those photos to regurgitate later. Instead, it learns the fundamental visual rules that define a cat, such as ear geometry, fur textures, whisker placement, and lighting interactions. Consequently, the model can synthesize a brand-new portrait of a feline that never existed in the real world yet appears so natural and detailed that it is practically indistinguishable from an authentic photograph.

Rather than producing original artifacts or media, the primary objective of non-GenAI models is to deliver accurate predictions and categorical assessments. The outputs of these systems are typically structured values, such as class labels, confidence scores, or numerical values, derived directly from evaluating incoming data.

By analysing how new inputs align with the patterns and relationships acquired during training, the model can swiftly sort complex information into predefined, meaningful buckets.

© AIOTI. All rights reserved.



Figure 2-8 Generative and non-GenAI differences animal photography.

This analytical capability depends on the model learning from structured, labelled datasets. During this learning process, the system internalizes the key distinguishing characteristics associated with each designated category. When presented with completely novel, unclassified data, it assesses these features against its established decision boundaries, determining which class or prediction offers the highest statistical likelihood.

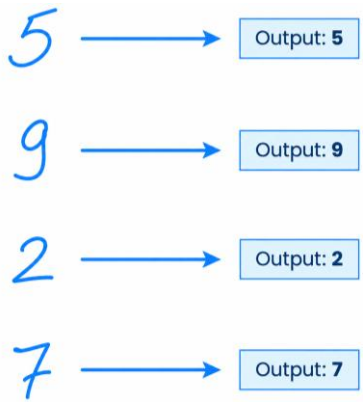


Figure 2-9 Generative and non-generative handwritten digit classification.

A classic illustration as presented in Figure 2-9 of this mechanism is found in optical character recognition, particularly through handwritten digit classification. When a non-generative model is trained on a standardized dataset of scribbled numbers, each visual sample is paired with its true numerical label from 0 through 9. Rather than generating its own drawings of numbers, the trained model inspects an unfamiliar handwritten digit, evaluates its curves, strokes, and contours, and accurately assigns it to its correct numerical identity.

Over the past decade, Edge AI deployed across industrial IoT networks and devices has predominantly focused on optimising conventional analytical models for closed, static tasks,

© AIOTI. All rights reserved.

such as classifying readings or detecting anomalies through fixed thresholding [1]. However, these schemes operate under the assumption of an invariant environment; when real-world operational conditions within a plant shift unexpectedly—such as through progressive sensor drift, mechanical wear, or physical environmental interference (*out-of-distribution*)—their reliability and accuracy degrade rapidly [2]. To overcome this rigidity, Generative Reasoning within GenAIoT introduces a fundamental shift in paradigm. Rather than merely classifying isolated data points, it continuously models the overall operational context of the facility [3][4]. By embedding Small Language Models (SLMs) and multimodal perception architectures that intrinsically couple observation with action such as *Vision-Language-Action* (VLA), edge assets cease to be passive sensing probes and acquire genuine local analytic capabilities [3][5].

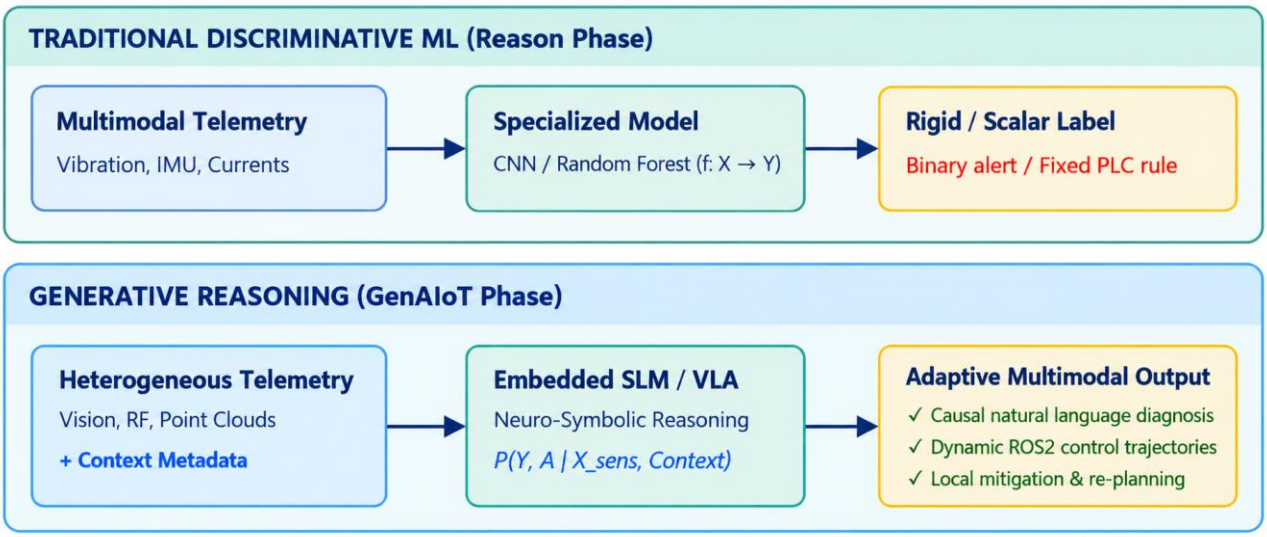


Figure 2-10 Transition from the classical discriminative inference paradigm to multimodal generative reasoning at the edge (GenAIoT).

This evolution aligns directly with the priorities established by AIOTI and European Cloud-Edge-IoT continuum frameworks (including the *aerOS* (***A****utonomous, scalabl****E****, t****R****ustworthy, intelligent European meta* ***O****perating* ***S****ystem for the IoT edge-cloud continuum*)[6] and *CODECO* (**Co**gnitive, **De**centralised Edge **C**loud **O**rchestration) [7] projects, as well as the *O-CEI* framework initiative [8]), whose shared objective is to endow connected environments with semantic autonomy and distributed decision-making capacity. Consequently, plant instrumentation does not merely flag faults reactively; it cross-references heterogeneous telemetry streams in real time (vibration profiles, thermal signatures, electrical power draw, or industrial computer vision), contrasts these with operating manuals and standard procedures, articulates root causes in comprehensible natural language, and formulates bespoke mitigation and tuning adjustments suited to the immediate operational scenario [4][9][10].

In process control and cyber-physical supervision, this approach replaces conventional, rigid programming with continuous loops of perception, analysis, and actuation structured across four coordinated phases [3][5].

First, a homogeneous cross-modal integration unifies analogue readings and industrial fieldbus telemetry into a standardised representation[5]. Second, the system benchmarks this operational state against engineering specifications and operating boundaries retained within local contextual stores, thereby interpreting the physical state of the machinery [4]. Third, it evaluates alternative operating trajectories in advance to anticipate the physical response of the plant, synthesising a balanced adjustment proposal [2][3]. Finally, this decision is compiled into executable operational setpoints dispatched across standard industrial communication protocols [3][11]. Because this entire operational cycle executes locally on the edge device or plant-floor server, the system mitigates network transport latency and eliminates service disruptions stemming from wide-area connectivity outages [1].

© AIOTI. All rights reserved.

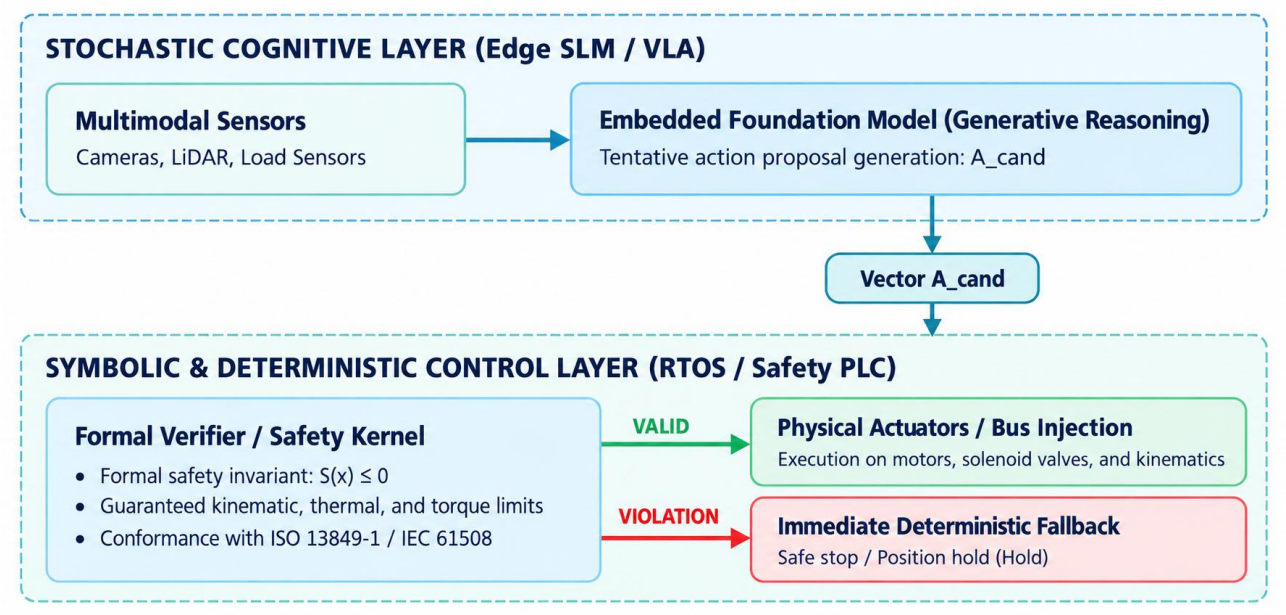


Figure 2-11 Neuro-symbolic control architecture: decoupling stochastic inference from deterministic execution.

To prevent non-deterministic excursions or unvetted actions in production environments (where an erroneous command could jeopardise plant safety or machinery integrity) the system implements a two-tier control architecture [4]. Within this topology, the generative layer functions as an intelligent, high-level supervisory advisor that suggests optimal control actions (such as modulating flow rates, trimming thermal setpoints, or shaping acceleration profiles) [3][6]. In parallel, a deterministic control module, executed atop real-time operating systems (RTOS), intercepts the proposed command and strictly verifies that it does not violate any pre-compiled safety invariants, including critical pressures, thermal thresholds, or kinematic limits. If the proposal satisfies all formal verification rules, it is transferred directly to the physical actuators. Should any ambiguity or constraint violation arise, the deterministic module immediately drops the generative recommendation and engages an established safe-fallback routine within sub-millisecond cycles [4]. Experimental testbeds and advanced open-network initiatives (such as the pilot facilities within *6G-SANDBOX* [12] alongside industrial field trials) demonstrate that hosting optimised, compact models (ranging from 1B to 7B parameters) on ruggedised edge nodes with private 5G connectivity successfully coordinates intricate operations while sustaining closed-loop response times below 50 ms without compromising functional safety [4][13].

The current evolutionary leap is introduced by GenAI as the edge systems are now entering a phase of "Generative Reasoning," where devices are no longer just predicting outcomes, but are actively synthesising information, inferring complex physical context, and autonomously producing dynamic outputs like natural language explanations, visual reconstructions, and adaptive control strategies.

2.6 GenAIoT, Embodied AI and Physical AI

Integrating advanced generative reasoning with the GenAIoT marks a transition from passive, reactive sensing to fully grounded, autonomous action.

For years, edge IoT was constrained to observational telemetry, detecting anomalies or processing static vision feeds without an intrinsic understanding of context. By infusing edge nodes with agentic, multimodal generative intelligence, endpoints can become embodied AI agents. These systems can ingest sensory data and actively query their surroundings through closed-loop sensorimotor integration, translating abstract spatial and causal reasoning into situated physical behaviour.

© AIOTI. All rights reserved.

The shift represents the practical realisation of physical AI, the discipline where ML directly interfaces with non-negotiable physical laws, mechanical dynamics, and material constraints. When agentic edge IoT systems perform multi-step spatial and temporal reasoning locally, they close the critical gap between high-level task planning ("what to do") and low-level physical execution ("how to move without failure").

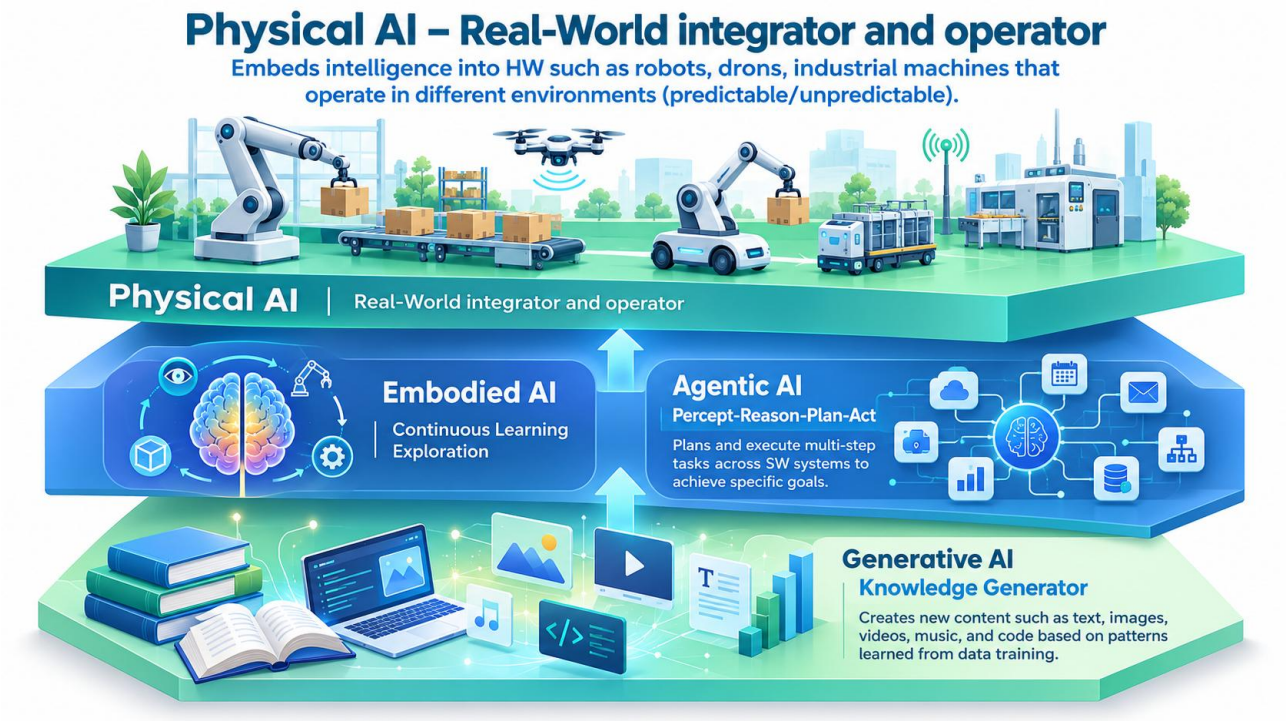


Figure 2-12 Layers representation of the concepts of GenAI, agentic AI, embodied and physical AI.

At the edge, where latency, bandwidth constraints, and real-time safety dictate operational boundaries, these agents dynamically account for physical realities: friction, inertia, actuator saturation, payload variability, and structural compliance. Rather than acting as remote peripherals executing rigid scripts, these systems exhibit true situatedness, making sub-second micro-adjustments to unpredictable physical perturbations.

A layered representation of the GenAI, agentic AI, embodied and physical AI is illustrated in Figure 2-12.

Bridging the cognitive capacity of embodied AI with the deterministic demands of physical AI requires a robust digital-to-physical pipeline. This is realised through high-fidelity modelling within virtual worlds, utilising DTs and IMTs (the tightly coupled synchronisation of the physical asset, its digital replica, and the human/AI cognitive interface). Within these physics-accurate simulated sandboxes, embodied agents leverage massive synthetic datasets and reinforcement learning to explore edge cases, anticipate mechanical stress, and formulate complex task chains without real-world risk.

Crucially, this architecture resolves the historic sim-to-real transfer problem. The DT serves as a bidirectional conduit: generative models simulate kinematic trajectories and thermodynamic outcomes in the virtual domain. At the same time, the physical edge agent feeds back real-world telemetry to continuously calibrate the simulation against real-world degradation and environmental shifts.

The developments of GenAIoT, embodied AI, and physical AI yield machines that possess both high-order logic and physical common sense. Whether deployed as autonomous mobile manipulators on an assembly line, collaborative agricultural swarms, or robotic infrastructure inspection platforms, these systems possess the physical agency to perceive context, simulate

© AIOTI. All rights reserved.

outcomes across spatial-digital dimensions, and safely exert force upon the real world to achieve deterministic operational goals.

Conceptual Overview of Embodied AI and Physical AI

Embodied AI is an approach to AI in which an agent learns, reasons, and develops intelligence through continuous, closed-loop sensorimotor interactions with an environment, using a physical or virtual body.

Physical AI is a branch of artificial intelligence and engineering focused specifically on systems that operate in the real, physical world, governed by the laws of physics, material constraints, and HW-SW integration (such as autonomous vehicles, humanoid robots, drones, and smart manufacturing systems).

For decades, mainstream AI developed primarily as "disembodied" intelligence, algorithms processing text, code, or static images on remote servers, detached from the physical world. While large language models (LLMs) and computer vision demonstrate remarkable cognitive capabilities, they lack an innate understanding of space, force, and physical consequences. Embodied AI and physical AI represent a fundamental change, moving AI from the digital realm into active engagement with the world.

Embodied AI originates from cognitive science and robotic philosophy (notably the embodied cognition hypothesis). It posits that true intelligence cannot be separated from the body that perceives and acts. Instead of passively receiving pre-curated data, an embodied agent actively explores. Its sensors (cameras, tactile arrays, Inertial Measurement Unit (IMUs)) gather information directly shaped by its movements, and its decisions immediately alter its sensory input. While the term is frequently used in robotics, Embodied AI also encompasses virtual agents navigating complex, physics-based simulations, video games, or DTs. The core insight of Embodied AI is that a body is not merely an output device for a brain; the physical morphology itself simplifies computation and actively shapes how the agent understands concepts like gravity, balance, and spatial relationships.

Physical AI takes this foundation and anchors it strictly into reality. While embodied AI can exist in simulated physics engines, Physical AI contends with the unforgiving nature of the real world: sensor noise, mechanical wear, friction, battery limitations, latency, and thermal dynamics. It sits at the convergence of advanced robotics, control theory, materials science, and ML. A physical AI system, such as a humanoid factory worker, an autonomous excavator, or a robotic surgical assistant, must achieve high-frequency, real-time closed-loop control where a SW failure has immediate physical and safety consequences.

The two concepts are deeply interdependent. Modern robotics typically uses Embodied AI techniques—such as reinforcement learning within massive simulations (e.g., Isaac Sim or MuJoCo)—to allow agents to develop thousands of years of sensorimotor experience in hours. Physical AI addresses the difficult transition from that virtual training to the physical world, known as the sim-to-real gap.

Physical AI focuses on reliable execution, while embodied AI addresses continuous learning, exploration, and reasoning from experience. Every embodied AI system is inherently a physical AI system because it must interact with the physical world. Not every physical AI system is an embodied AI system since not every physical machine is designed to continuously learn and adapt to experience.

Both physical and embodied AI systems are used to simulate and train within immersive virtual simulation and modelling environments, then transfer the simulation results and acquired knowledge directly into real-world environments. An overview of the embodied and physical AI representation and the types of physical AI autonomous systems and examples of applications is given in Figure 2-13.

© AIOTI. All rights reserved.

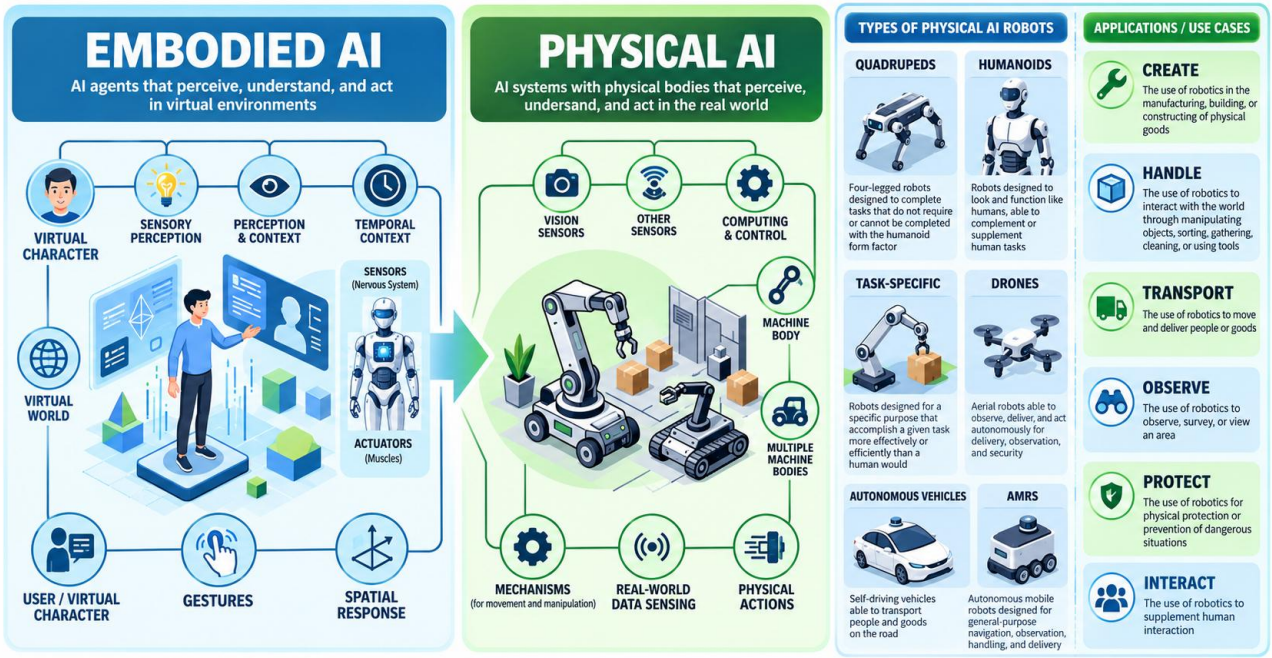


Figure 2-13 Overview of embodied and physical AI representations and types of autonomous systems and use cases.

Embodied and physical AI mark GenAIoT's evolution from passive observer to active participant. By grounding intelligence in physical reality, these paradigms are laying the groundwork for machines that can navigate unpredictable human environments, manipulate delicate tools, and collaborate alongside people in the physical world.

© AIOTI. All rights reserved.

3 Overview of GenAI Concepts

The section provides a concise technical overview of the GenAI technologies most relevant to distributed IoT and edge environments.

GenAI is a branch of ML focused on synthesising new, original content that matches the statistical patterns of its training data. Unlike traditional discriminative AI, which analyses, classifies, or predicts labels for existing inputs, generative models learn how data is assembled from the ground up. They produce novel artefacts across text, images, audio, video, code, and structured scientific data.

The generative architecture families to be considered include Variational Autoencoders (VAEs) that map raw inputs into a structured, continuous latent space governed by explicit statistical parameters (mean μ and variance σ). They optimise the Evidence Lower Bound (ELBO), balancing accurate output reconstruction against the divergence of the latent space from a standard normal prior; Generative Adversarial Networks (GANs) that frame synthesis as a zero-sum, non-cooperative game between two sub-networks, the Generator (G), which transforms continuous random noise into synthetic samples and the Discriminator (D), which evaluates inputs to differentiate authentic samples from synthetic attempts. GANs excel at generating sharp, high-frequency visual details, and they are often difficult to train and vulnerable to mode collapse, where the generator outputs samples from only a narrow subset of the target distribution; Denoising Diffusion Probabilistic Models (DDPMs) frame generation through two complementary processes: The Forward (Noising) Process, which systematically destroys the structure in a data point by adding incremental Gaussian noise across discrete time steps, until the input turns into pure noise and Reverse (Denoising) Process, which is a neural network that learns to reverse this noise step by step, predicting and removing the noise added at each stage.

Diffusion models provide stable training dynamics and broad sample diversity, serving as the mathematical backbone for modern visual synthesis systems; Autoregressive Transformer Models, which process discrete sequence tokens (such as words, code characters, or audio intervals) using the Self-Attention Mechanism.

GenAI continues to shift the role of computing from passive information processing to active creative synthesis. By converting complex, high-dimensional data into navigable latent spaces, generative models automate repetitive cognitive tasks, accelerate scientific modelling, and serve as dynamic reasoning engines across modern computing environments.

Raw generative models learn statistical associations from massive, unlabelled datasets, but this baseline training alone does not guarantee helpful, accurate, or safe behaviour. Preparing models for deployment requires a multi-stage lifecycle:

- **Self-Supervised Pre-Training:** The model ingests broad, unlabelled collections of text, code, or images. It learns general world knowledge and grammar by solving basic self-supervised tasks, such as predicting masked tokens or the next token in a sequence.
- **Supervised Fine-Tuning (SFT):** The pre-trained model is fine-tuned on curated sets of clear prompt-and-response demonstrations, teaching it to follow instructions and format answers predictably.
- **Human Feedback Alignment (RLHF/DPO):** To ensure outputs align with human expectations, developers use reinforcement learning techniques:
    - **RLHF (Reinforcement Learning from Human Feedback):** Uses an auxiliary reward model trained on human preference rankings to optimise the generative engine via policy gradients.
    - **DPO (Direct Preference Optimisation):** Derives an exact closed-form loss that optimises the policy directly on paired preference data, skipping the need for an independent reward model.

© AIOTI. All rights reserved.

Deploying generative models into production environments relies on several inference-time patterns:

- **In-Context Learning and Prompt Engineering:** Large autoregressive transformers can adapt to novel tasks without weight updates. Providing context, step-by-step logic, or few-shot input-output examples directly in the prompt steers the model's next-token generation.
- **Retrieval-Augmented Generation (RAG):** RAG systems mitigate factual errors by connecting the generative core to external reference knowledge.
- **Agentic Tool-Use:** Generative models can produce structured outputs (such as JSON or function calls) that execute code, query relational databases, call external APIs, and adjust their plans based on the returned results.

Several open challenges and technical vulnerabilities related to generative models are presented in Table 3-1.

Table 3-1 GenAI open challenges and technical vulnerabilities.

Vulnerability / Bottleneck	Mechanism	Systemic Impact
Hallucination	Generative models are probabilistic, not factual databases. They sample likely token paths even when facts are unsupported	Produce plausible sounding but factually wrong statements in sensitive legal, medical, or engineering tasks
Key Value (KV)-Cache Memory Scaling	In transformers, tracking extended interactions requires caching past keys and values ($O(N)$ context length memory growth)	Squeeze memory bandwidth on serving HW, inflating serving costs during long multi-turn sessions
Safety	Adversarial prompts bypass safety guardrails via persona adoption, cipher encodings, or recursive prompt injections	Risk of generating toxic material, security exploit scripts, or exposing confidential corporate data
Copyright and Data Provenance	Generative models can inadvertently reproduce verbatim passages or styles learned from proprietary sources	Create unresolved legal questions on intellectual property (IP), data licensing, and derivative content

3.1 Definition and Classification of Foundation Models

An AI Foundation Model is broadly defined as a large-scale neural network trained on vast, diverse, and typically unlabelled datasets using self-supervised or semi-supervised learning. The defining characteristic of a foundation model is its inherent versatility: rather than being engineered to solve a single, narrowly defined task, it acquires broad representations of syntax, context, structure, and world knowledge. These general representations can subsequently be adapted, fine-tuned, or prompted to execute a wide variety of downstream tasks across diverse operational domains without retraining the core network from scratch.

Foundation models are classified across several primary structural dimensions: input-output modality, core neural architecture, training paradigm, model scale, and licensing accessibility.

Modality classification categorises models by the types of sensory or symbolic data they process and generate. Unimodal models operate within a single domain, such as natural language processing, code generation, or acoustic speech analysis. Multimodal models process and bridge multiple data streams simultaneously, such as text-to-image synthesis, visual question answering, and audio-video comprehension. At the specialised edge, physical and sensor-native foundation models represent an emerging class designed to ingest non-symbolic real-world inputs, including inertial measurement data, radio-frequency signals, spatial point clouds, and high-frequency telemetry.

© AIOTI. All rights reserved.

Architectural classification distinguishes models based on their core mathematical formulation and computational scaling behaviour.

Modern foundation models are primarily built on transformer-based variants, categorised into causal autoregressive decoder-only networks optimised for generative sequence continuation, masked encoder-only configurations designed for dense contextual representation and classification, and dual encoder-decoder networks tailored for complex sequence-to-sequence translation. In parallel, non-transformer paradigms, such as recurrent architectures, State-Space Models (SSMs), and hybrid diffusion models, are increasingly utilised to achieve sub-quadratic computational scaling and constant-memory inference across long context horizons.

Operational scale and downstream adaptation delineate foundation models into enterprise-grade frontier networks and compact, task-specialised variants. Frontier models comprise hundreds of billions of parameters, demonstrating advanced emergent reasoning at the cost of immense computational and energy requirements.

On the other hand, SLMs and edge-native foundation models compress these capabilities into sub-billion to single-digit billion parameter ranges through techniques such as knowledge distillation and low-rank adaptation. Finally, governance and licensing classifications divide foundation models into proprietary closed-access systems accessible exclusively via remote APIs, and open-weight architectures that allow decentralised deployment, local parameter inspection, and on-premises execution across independent computing infrastructure.

3.2 Multimodal Capabilities in Physical Environments

The focus today shifts to how GenAI processes diverse data types and provides multimodal capabilities (text, vision, audio, time-series sensor data) for enabling IoT systems to fully understand and interact with complex physical environments.

Physical environments are dynamic and difficult to process at scale. They generate continuous, noisy, heterogeneous, and high-velocity data. Processing high-definition video, continuous audio, or raw sensor telemetry directly on edge devices can create significant bottlenecks, including high memory and compute requirements, data-transfer overhead, latency, and privacy concerns [14]. Consequently, physical-AI and GenAIoT architectures increasingly require an edge semantic abstraction layer that processes raw physical signals close to their source before exposing relevant information to higher-level foundation models.

Lightweight AI models, including TinyML models, optimized deep-learning models, and modality-specific encoders, can run across extreme-edge, far-edge, and edge devices [14]. Rather than continuously transmitting raw sensor streams to a GenAI system, these local models transform physical signals into compact, task-relevant representations. Depending on the modality and model architecture, these representations may take the form of embeddings/tokens, detected objects, trajectories, events, poses, occupancy states, or other structured features.

The purpose of this semantic abstraction is not simply to reduce data dimensionality. More importantly, it transforms high-volume physical signals into representations that preserve information relevant to perception, reasoning, and action while discarding unnecessary raw data. This reduces communication and computation requirements, limits exposure of sensitive raw data, and provides multimodal foundation models with representations that can be integrated with language and other modalities. This architecture is consistent with the broader evolution of Vision-Language-Action (VLA) models, which seek to unify perception, language understanding, and physical action [15].

Limitations of Conventional Optical and Audio Sensing

Traditional sensing modalities such as camera feeds and continuous audio recordings introduce substantial edge-computing and privacy challenges. High-definition video requires significant

© AIOTI. All rights reserved.

memory bandwidth, storage, and frame-processing resources, while continuous audio monitoring can expose sensitive conversations and environmental information. These concerns become particularly important in sensitive environments such as private homes, bedrooms, healthcare facilities, and other spaces where continuous recording may be unacceptable or heavily restricted.

Optical sensing also has inherent physical limitations. Cameras depend on line of sight and can degrade or fail under conditions such as darkness, fog, smoke, occlusion, or physical obstruction. Consequently, relying exclusively on cameras can create important gaps in an AI system's perception of the physical environment.

These limitations motivate the use of complementary non-optical sensing modalities that can provide additional information without continuously transmitting raw visual or audio data.

Time-Series Sensor Encoders

Time-series sensor data provides an important alternative for representing physical environments in a more compact and potentially privacy-preserving manner. Wi-Fi sensing is one example. By applying lightweight machine-learning models to Channel State Information (CSI) or related measurements from ambient wireless signals, edge devices can learn patterns associated with physical changes in the environment.

Depending on the sensing configuration and trained model, these signals can be transformed into structured representations such as human presence, motion, occupancy, activity patterns, gestures, or other inferred states. Unlike conventional cameras, Wi-Fi sensing does not require optical illumination and can operate through certain obstacles and walls, enabling sensing in situations where optical line of sight sensing is limited.

The resulting representation can be substantially smaller and more task-specific than the original high-rate RF measurements. Instead of transmitting the raw signal stream, the edge system can expose relevant features, embeddings, events, or states to a downstream multimodal foundation model.

However, Wi-Fi sensing should be viewed as one component of a broader physical-intelligence architecture, rather than as a complete solution for spatial perception. Other non-optical and complementary modalities—including Ultra-Wideband (UWB), mmWave radar, acoustic sensing, environmental sensors, and IMUs—can similarly transform raw physical measurements into compact representations.

Combining these modalities enables a multimodal semantic layer for physical environments, in which different sensors contribute complementary information about location, motion, occupancy, environmental conditions, human activity, and other physical states. A GenAI system can then integrate these representations with language, vision, and other modalities to build a richer context of the environment while reducing the need to transmit continuous raw sensor streams.

Vision Encoders

Vision encoders provide another important component of the edge semantic layer. Lightweight vision models can process camera frames locally and extract representations such as visual embeddings, object detections, bounding boxes, object tracks, poses, trajectories, and events.

For example, instead of transmitting a continuous high-definition video stream to a foundation model, an edge vision model might locally infer:

person detected → sitting → door opened → object moved

© AIOTI. All rights reserved.

These structured representations can then be provided to a multimodal foundation model alongside information from other modalities.

This approach reduces bandwidth and processing requirements and can also improve privacy by keeping raw video local. However, it does not eliminate the need for careful privacy design: even derived representations such as identities, poses, trajectories, or activity patterns can contain sensitive information.

Semantic Abstraction to Multimodal Foundation Models

The resulting architecture can therefore be viewed as a hierarchy. The key architectural principle is that raw physical data does not always need to reach the foundation model. Instead, intelligence can be distributed across the sensing hierarchy. Local models perform modality-specific perception and abstraction, while the foundation model performs higher-level multimodal reasoning and contextual understanding.

This creates a semantic interface between the physical world and GenAI: edge intelligence converts heterogeneous physical signals into representations that a foundation model can integrate, reason over, and ultimately use to support interaction with the physical environment.

3.3 Agentic AI

Agentic AI represents a significant evolution in autonomous computational systems. Moving beyond traditional discriminative models and passive generative tools, agentic systems are defined by their capacity for goal-directed, autonomous behaviour within complex, non-stationary environments. Rather than simply responding to static prompts or executing isolated, predefined tasks, an agentic AI system continuously evaluates its surrounding context, synthesises plans over extended temporal horizons, and executes actions to alter its operational reality. An overview of the agentic AI ecosystem is illustrated in Figure 3-1.

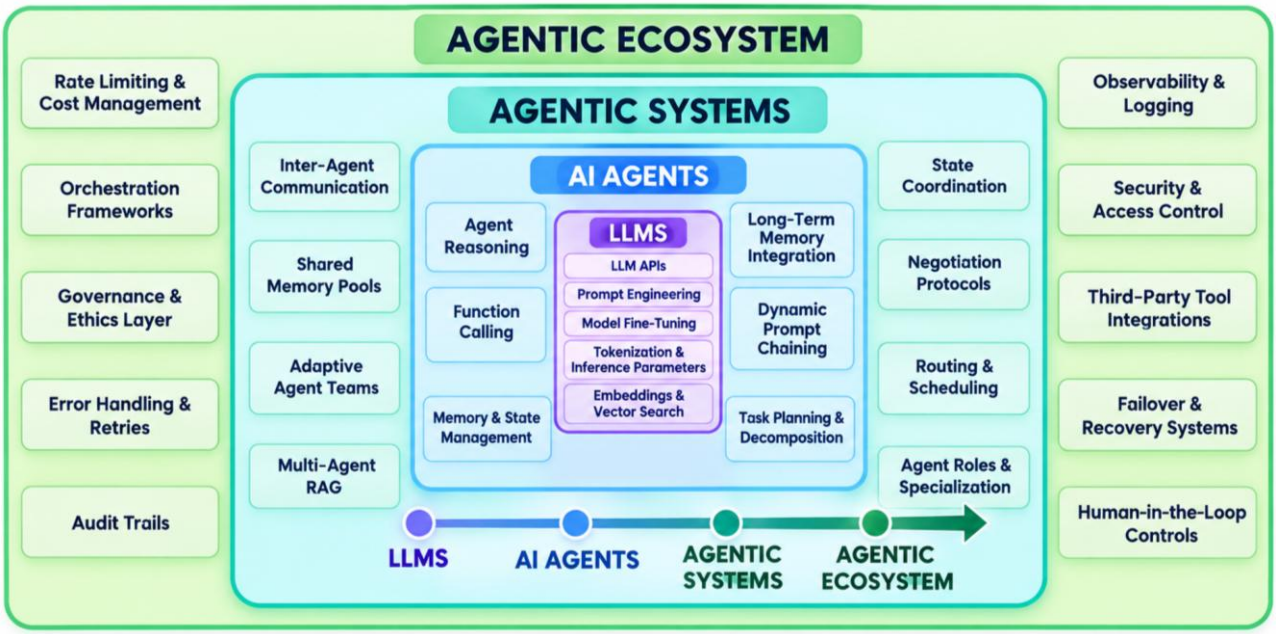


Figure 3-1 An overview of an agentic AI ecosystem.

Realising this high degree of agency demands a modular, integrated architecture comprising four primary components as illustrated in Figure 3-2: perception, cognition, action, and continuous learning, supported by comprehensive memory management and tool integration subsystems.

© AIOTI. All rights reserved.

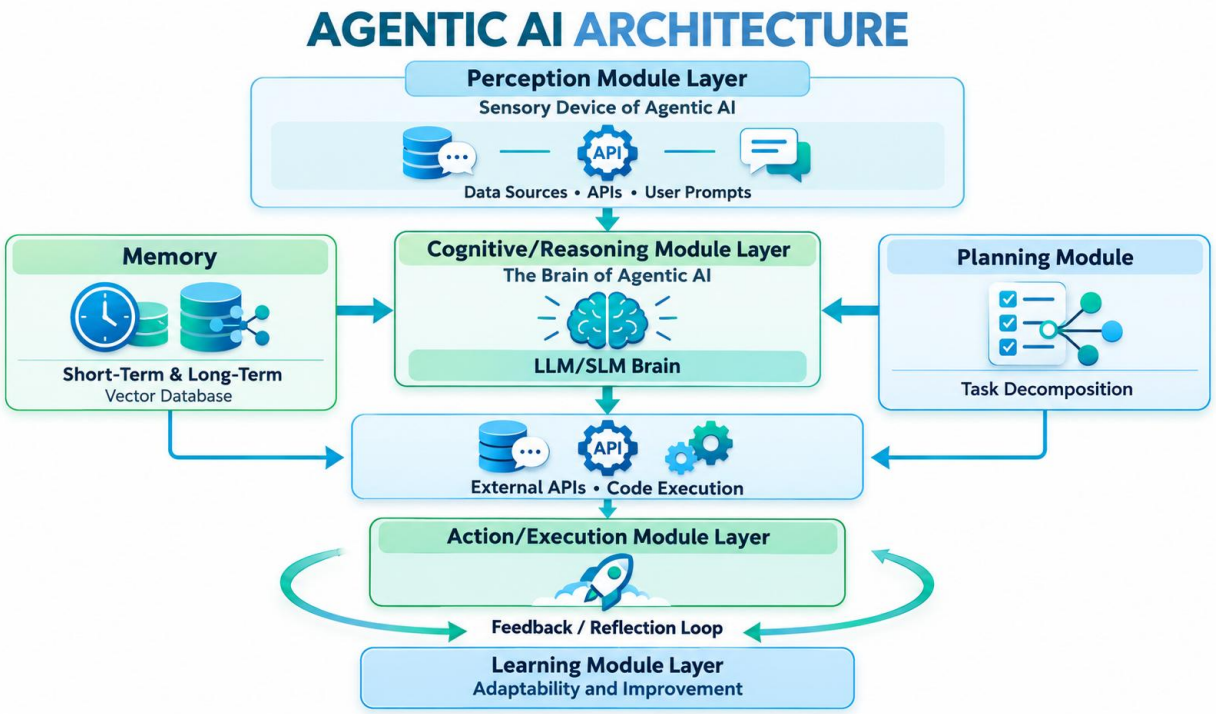


Figure 3-2 Agentic AI architecture.

The perception layer illustrated in Figure 3-3 acts as the sensory apparatus of the agentic system. It is responsible for ingesting, filtering, and structuring raw inputs from dynamic, non-curated physical and digital environments. This pipeline begins with sensory ingestion, gathering signals from physical transducers such as optical cameras, LiDAR arrays, acoustic microphones, and industrial telemetry probes, alongside digital streams such as webhooks, telemetry feeds, and SW application programming interfaces (APIs). These heterogeneous data streams must be normalised to eliminate noise, manage transmission jitter, and correct for temporal misalignment across distinct sensor modalities.

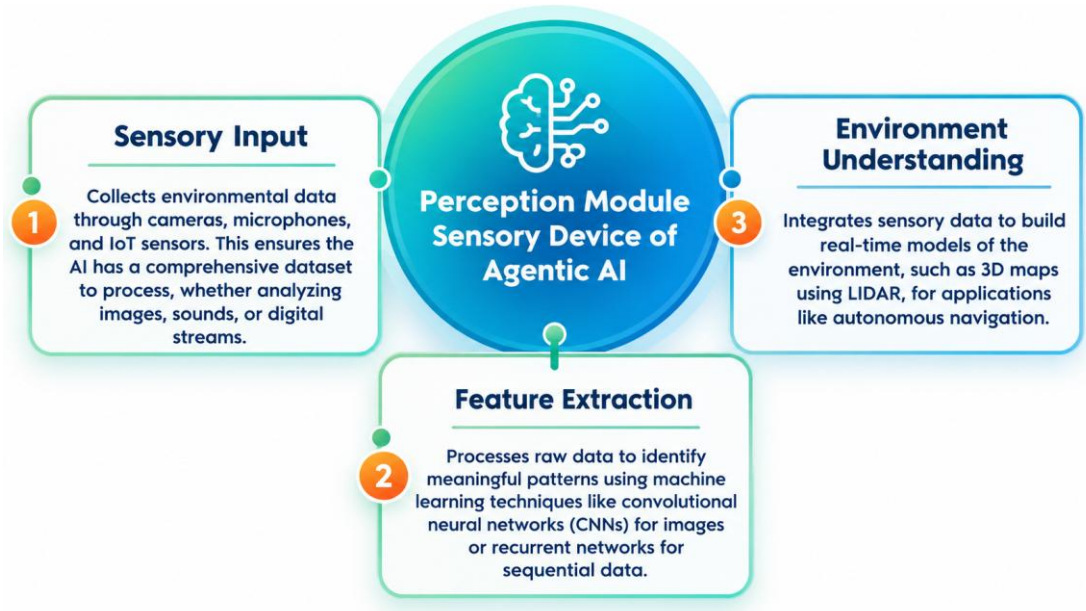


Figure 3-3 Agentic AI perception module layer.

Once ingested, the pipeline transitions to automated feature extraction. Here, deep neural representations distil dense raw data into structured embeddings that isolate salient invariant properties. In visual contexts, this encompasses edge detection, depth mapping, and spatial orientation; in textual and acoustic streams, it extracts semantic entities, syntactic relations, and spectral features.

© AIOTI. All rights reserved.

The final stage of perception is entity and object recognition, where downstream machine-learning classifiers identify, locate, and contextualise external components. In critical settings, such as clinical oncology diagnostics, this stage isolates pathological anomalies within complex multi-spectral imaging.

The cognitive/reasoning module layer illustrated in Figure 3-4 uses engine functions as the system's central decision-making core. It synthesises perceptual signals, aligns them with overarching operational goals, and formulates safe, efficient execution strategies. Cognition operates through three coupled layers: goal representation, strategic planning, and rational decision-making.

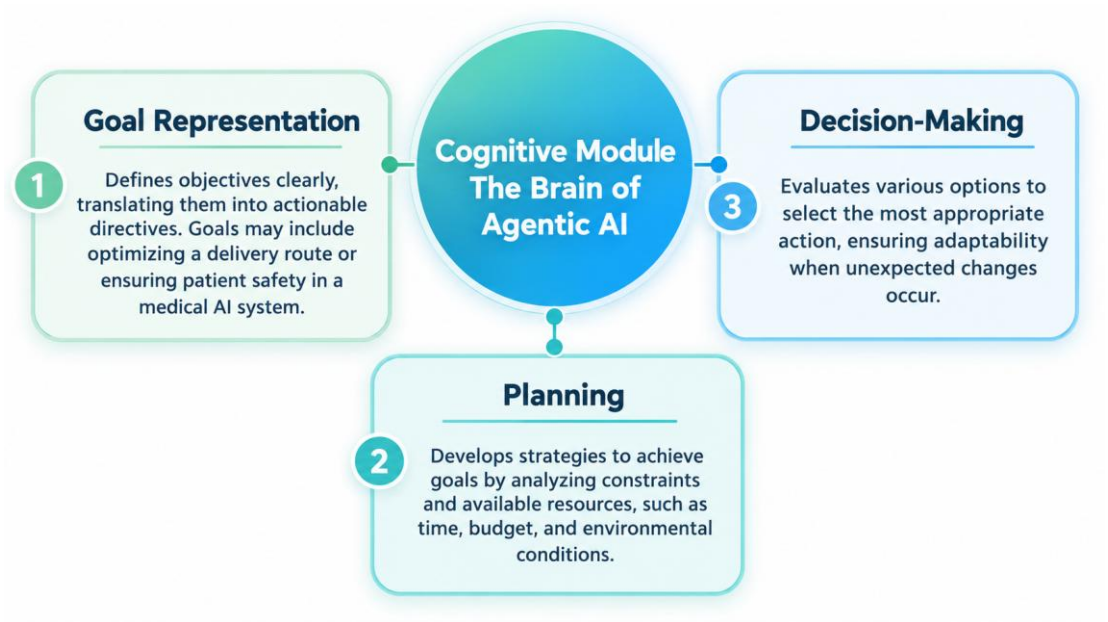


Figure 3-4 Agentic AI cognitive/reasoning module layer.

Goal representation encodes strategic objectives into formal mathematical and symbolic structures. This enables the agent to resolve competing priorities, compute trade-offs, and maintain a coherent operational direction over extended multi-step trajectories.

Decision-making operationalises these plans using utility theory, game-theoretic formulations, and multi-objective optimisation frameworks. The agent systematically evaluates alternative trajectories under environmental uncertainty, selecting actions that balance short-term progress with long-term strategic success.

Strategic planning decomposes high-level goals into ordered action sequences. It employs hierarchical task networks, heuristic tree searches, and generative reasoning loops to map paths across complex problem spaces while respecting resource constraints, such as thermal thresholds, energy budgets, and execution deadlines.

The action/execution module layer as illustrated in Figure 3-5 executes decisions in physical or virtual domains. Whether controlling robotic actuators or interacting with APIs, the action module ensures precise implementation of plans while adapting dynamically to environmental feedback. The action module executes plans through robotic or digital means, adjusting dynamically to feedback for precise implementation.

The module executes decisions derived from the cognitive layer through actuators and monitors, enabling real-world interactions with precision and adaptability. The actuators could be physical devices like robotic arms or virtual systems like API calls, executing planned actions seamlessly.

© AIOTI. All rights reserved.

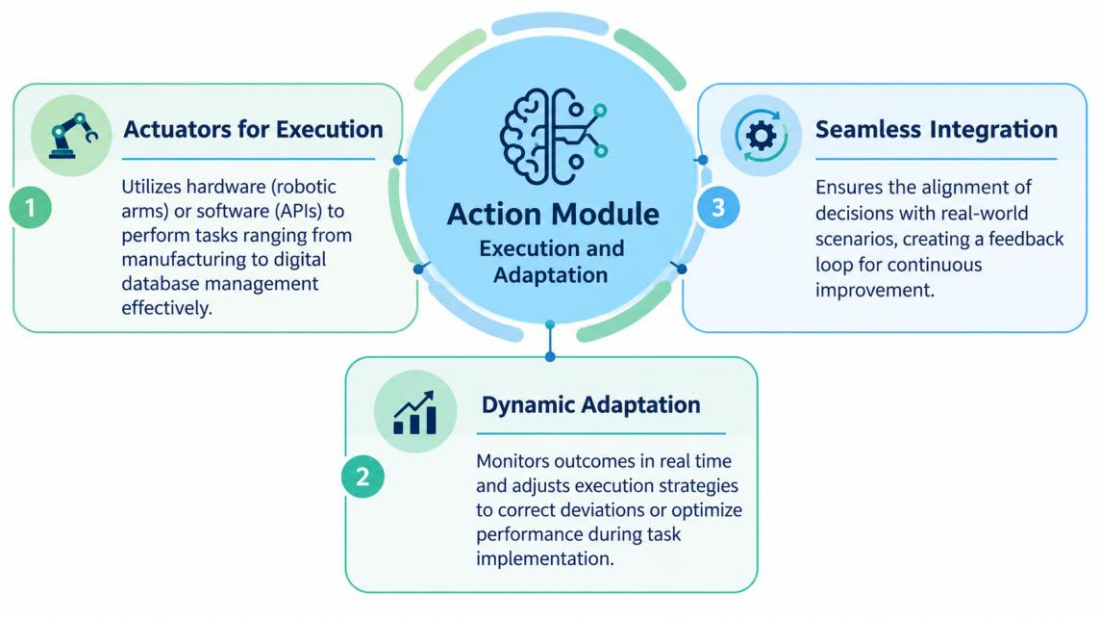


Figure 3-5 Agentic AI action/execution module layer.

The learning module layer illustrated in Figure 3-6 drives adaptability, employing techniques like reinforcement learning and supervised learning to refine decision-making processes. It uses ML techniques like reinforcement learning, supervised, and unsupervised learning to enhance the system's decision-making, ensuring long-term success and resilience. The core mechanism for continuous improvement, enabling agentic AI to refine its decision-making and adaptability through various ML paradigms.

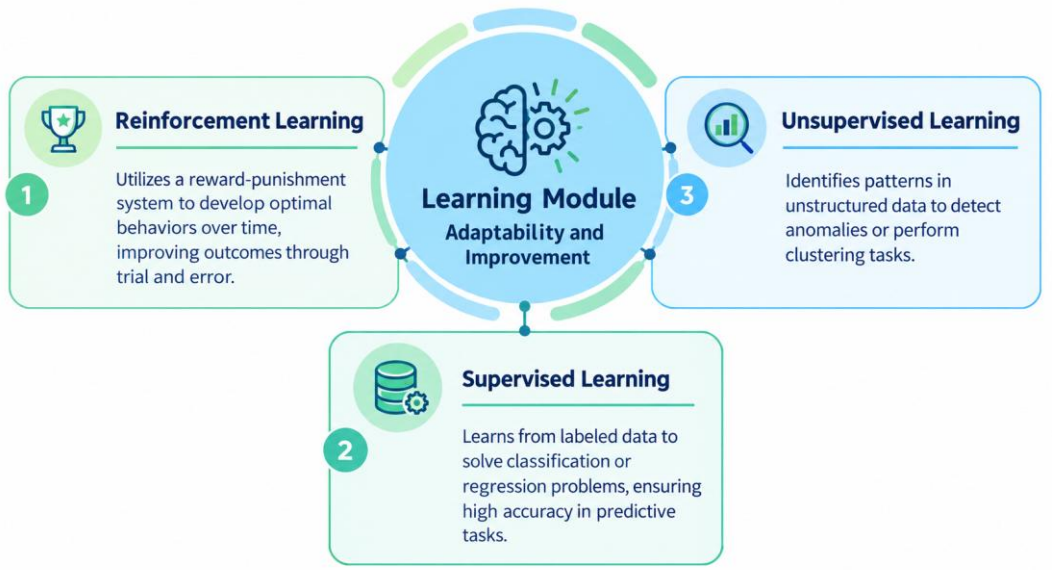


Figure 3-6 Agentic AI learning module layer.

Reinforcement learning optimizes actions by assigning rewards or penalties based on the outcomes, driving systems to improve performance over time. Supervised learning trains models using labelled data for tasks such as image classification or predictive analysis. Unsupervised learning identifies patterns and latent structures in unlabelled data for exploratory and adaptive tasks, such as clustering customer behaviours.

Supporting the cognitive core are memory management systems, which maintain continuity across operational lifetimes. Without structured memory, agents remain bound to their immediate context windows, unable to ground actions in historical precedent or retain knowledge between independent tasks.

© AIOTI. All rights reserved.

Modern architectures integrate working memory, episodic memory, and parametric memory. Working memory maintains the immediate operational context, active goals, and intermediate task states required for active reasoning loops. Episodic memory captures temporally sequenced records of historical interactions, failures, and operational successes. It relies on non-parametric retrieval structures, vector databases, and neural memory architectures to draw upon historical precedent when tackling novel challenges.

Parametric memory resides within the pre-trained weights of the model, encapsulating broad linguistic syntax, commonsense physical principles, and domain knowledge. Effective memory management dynamically retrieves, updates, and consolidates historical evidence into real-time decision loops, transforming reactive systems into longitudinally stable, contextually grounded actors.

Operational Mechanisms - Action, Tool Integration, and Learning Paradigms

The action module translates cognitive decisions into physical or digital outcomes, bridging abstract reasoning and real-world execution. Decoupled from the cognitive core to ensure system stability, this module executes commands across heterogeneous environments while continuously validating real-world states. In physical contexts, the action layer interfaces with low-level robotic actuators, micro-controllers, servo motors, and automated industrial plants through real-time fieldbus architectures. In digital contexts, it interfaces with SW architectures via API calls, automated database queries, microservice calls, and network protocol commands.

Modern agentic systems incorporate execution monitoring within the actuation pipeline. Rather than if issuing a command guarantees its success, the monitoring subsystem continuously polls the perception layer to verify whether physical or digital states shift in alignment with model predictions. If unexpected obstacles, physical friction, API timeouts, or unpredicted deviations occur, the execution monitor interrupts execution, flags state discrepancies, and prompts the cognitive core to replan dynamically. An example of an agentic AI workflow implementation for an industrial use case is illustrated in Figure 3-7.

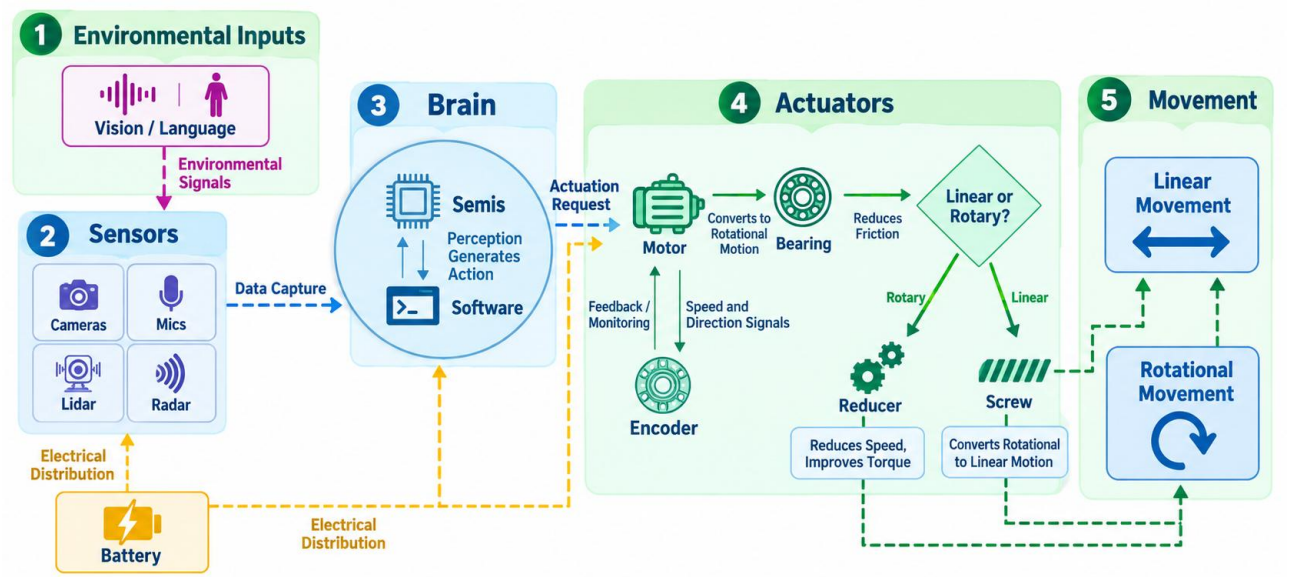


Figure 3-7 Example of an agentic AI workflow for an industrial use case.

This operational reach is extended through tool integration frameworks. An isolated intelligence is bounded by its internal parametric representations, which may be stale, uncalibrated, or numerically imprecise. Tool integration frameworks give the agent structured interfaces to consult external SW utilities, execute code securely within isolated environments, query enterprise relational databases, and interface with dynamic mathematical solvers.

© AIOTI. All rights reserved.

By treating external SW utilities as accessible toolsets, an agent dynamically identifies its cognitive limitations and invokes targeted external tools to compute exact numerical solutions, look up verified domain knowledge, or trigger external enterprise workflows. This capability grounds system operations in up-to-date data, mitigating hallucinations and expanding practical utility across dynamic enterprise environments.

Continuous learning provides the engine for long-term resilience and adaptation, enabling the agent to refine its operational capabilities through environmental experience. This iterative improvement draws upon three foundational machine-learning paradigms: reinforcement learning, supervised learning, and unsupervised structural discovery.

Reinforcement learning optimises sequential decision-making policies by issuing reward or penalty signals based on the outcomes of executed actions. Through continuous exploration and exploitation within operational envelopes, reinforcement learning algorithms guide agents to discover optimal behavioural strategies across dynamic topologies, such as automated high-frequency financial trading or robotic trajectory optimisation.

Supervised learning refines the system's perceptual and categorisation sub-modules using curated, ground-truth datasets, ensuring baseline accuracy in pattern recognition tasks like visual quality assurance.

Unsupervised learning operates without explicit external labels, discovering latent clusters and statistical regularities within raw operational telemetry. This capability is vital for exploratory diagnostic tasks, such as surfacing emerging equipment anomalies or identifying evolving consumer behaviours.

Beyond classical training methods, modern agentic AI demands advanced learning paradigms to tackle deployment-specific challenges. Systems must operate in dynamic environments while avoiding catastrophic forgetting, where acquiring new skills overwrites previously validated capabilities.

To resolve this plasticity-stability dilemma, agents employ continual learning frameworks, parameter-efficient fine-tuning (PEFT), and episodic memory replay. These approaches update isolated parameter subspaces or consolidate experiences offline, ensuring lifelong adaptation without compromising baseline stability.

Multi-Agent Coordination, and Key Characteristics

Designing agentic AI systems requires adherence to several principles, among them modularity, scalability and interoperability, as illustrated in Figure 3-8.

The development of agentic AI rests upon a synthesised theoretical foundation uniting cybernetics, formal decision theory, cognitive psychology, and distributed multi-agent systems. The primary architectural objective is balancing proactive autonomy with operational safety.

Cybernetic feedback loops provide the mathematical basis for autonomous regulation. By calculating discrepancies between real-world sensory observations and internal reference models, agents apply targeted corrective actions through closed negative feedback loops.

Formal decision theory complements this by providing probabilistic frameworks, such as Bayesian inference and Markov Decision Processes (MDPs), which allow systems to quantify uncertainty, estimate incomplete state observations, and select paths that maximise long-term expected utility.

© AIOTI. All rights reserved.

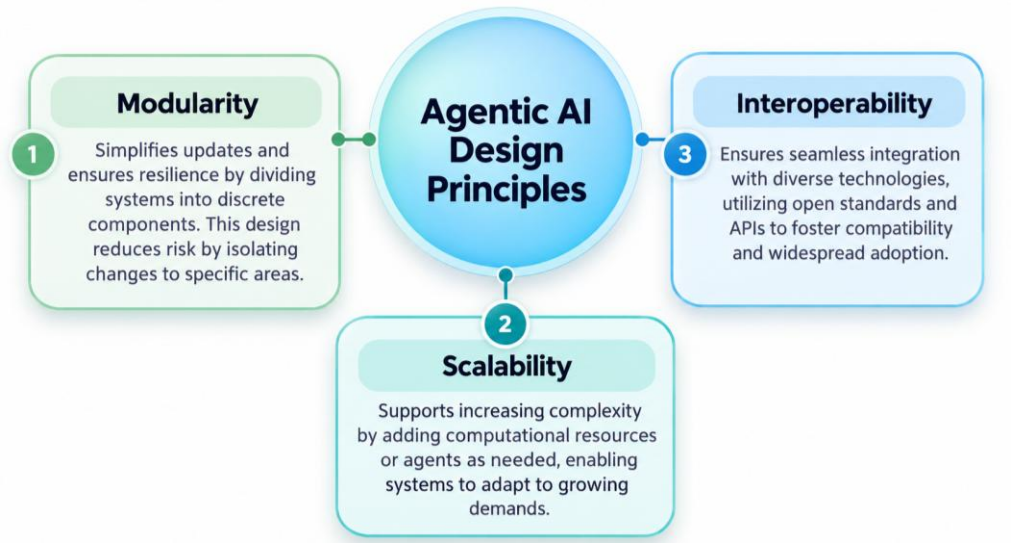


Figure 3-8 Agentic AI design principles example.

Cognitive psychology informs the dual-process architecture underpinning modern cognitive engines. Fast, intuitive pattern-matching components (analogous to cognitive System 1) handle high-frequency sensory classification and routine reactions. Meanwhile, deliberate, computationally demanding symbolic planners (analogous to cognitive System 2) evaluate complex decisions, explore alternative plans, and calculate risk profiles. This hybrid structure avoids burning compute on trivial actions, reserving intensive multi-step reasoning for novel or high-stakes scenarios.

These theoretical frameworks establish four defining characteristics of agentic AI systems:

- **Autonomy:** The capacity to operate independently across extended durations without continuous human oversight, deriving actions from broad strategic objectives.
- **Adaptability:** The ability to adjust operating strategies dynamically when faced with unexpected disruptions or systemic changes.
- **Contextual Awareness:** The capacity to interpret implicit environmental variables, situational subtleties, and non-stationary constraints.
- **Continuous Learning:** The ability to integrate operational lessons iteratively, refining behavioural policies through environmental interactions.

When operational challenges outstrip the capabilities of a single monolithic agent, architectural paradigms shift toward Multi-Agent Systems (MAS). In a MAS topology, a complex mission is distributed across a collaborative network of heterogeneous, specialised agents, each contributing targeted expertise.

Multi-agent coordination relies on decentralised communication protocols and distributed consensus algorithms. Rather than funnelling all telemetry into a brittle single point of failure, agents negotiate peer-to-peer, sharing intermediate representations, delegating sub-tasks, and bidding on activities based on local computational capacity, physical proximity, or specialized tools.

In automated logistics, for example, specialized automated guided vehicles, robotic sorting cells, and inventory management agents continuously balance routing schedules to prevent warehouse bottlenecks.

In decentralised smart grids, autonomous sub-station agents balance fluctuating local renewable generation against neighbourhood demand in real time. Decomposing complex problems into multi-agent frameworks provides horizontal scalability, graceful operational degradation during HW faults, and structural resilience in the face of partial system failures.

© AIOTI. All rights reserved.

Cross-Sector Deployments, Trustworthiness, and Ethical Governance

Agentic AI systems are driving major operational transformations across mission-critical domains. In healthcare, they are moving beyond passive diagnostic image analysis to autonomous patient monitoring, dynamic treatment scheduling, and closed-loop pharmacological delivery. These systems synthesize dynamic telemetry from bedside sensors with longitudinal patient histories, alerting medical teams to early stages of septic shock or autonomously managing intravenous infusion rates within strict clinical envelopes.

In autonomous vehicles and robotics, agentic systems integrate visual, inertial, and spatial telemetry to chart trajectories through unpredictable real-world environments. They adapt instantly to unexpected hazards, such as an errant pedestrian or sudden loss of traction, replanning trajectories mid-action without relying on high-latency cloud roundtrips.

In Industry 5.0, these architectures support human-machine symbiosis on the factory floor. Collaborative robots evaluate human operator movements, anticipate intended assembly steps, adjust tool alignments, and monitor workspaces to ensure worker safety while maintaining manufacturing throughput. However, granting computational systems real-world operational autonomy introduces profound safety, legal, and ethical risks. Unconstrained autonomy in physical or high-stakes digital settings can result in property damage, systemic financial loss, or loss of life. Addressing these vulnerabilities requires deploying robust governance frameworks:

Explainable AI (XAI) frameworks provide the foundation for system interpretability and operational audits. Because advanced neural architectures and generative components operate as non-linear black boxes, post-hoc explanation techniques, attention-attribution mapping, and symbolic audit trails are essential to trace the reasoning behind specific decisions.

In critical sectors such as credit underwriting, automated policing, and administrative law, explainability ensures outputs can be reviewed, scrutinized, and legally defended, preventing opaque systems from making unchallengeable rulings.

Accountability frameworks must establish clear liability when systems fail. Autonomous operation complicates traditional product liability and negligence regimes by separating human operators from immediate tactical control.

Robust technical audit trails must continuously log the system's internal world model, sensory inputs, chosen actions, and environmental feedback into tamper-evident data stores. In the event of a catastrophic industrial failure or collision, these logs provide forensic clarity, determining whether a breakdown stemmed from mechanical sensor failure, adversarial manipulation, or aberrant cognitive planning.

Finally, algorithmic bias mitigation must be enforced across the perception and learning pipelines. Training models on biased real-world data risks perpetuating discriminatory patterns in downstream allocations, clinical decisions, and predictive policing assessments.

Mitigation requires ongoing algorithmic auditing, adversarial testing, and data-balancing pipelines. In tandem with these technical safeguards, systems must operate within strict regulatory structures and international safety standards, ensuring that autonomous agents remain legally compliant, robust, and aligned with human values throughout their operational lifecycles.

Emerging Frontiers, and the Evolution of Agentic AI

The trajectory of agentic artificial intelligence is shifting from SW-only algorithmic enhancements toward holistic system-level co-design, uniting modern cognitive models with next-generation semiconductor HW and distributed systems engineering. Early agentic architectures relied on

© AIOTI. All rights reserved.

remote, power-hungry server infrastructure, running monolithic transformer backbones within resource-abundant data centres.

Scaling agentic networks across global deployments requires a computing shift: migrating multi-modal reasoning out of remote server clusters and directly onto the physical, embedded edge.

Edge-native agency faces strict thermodynamic and HW limits. Edge nodes operate under strict energy budgets, thermal limitations, and intermittent network access. Running dense multi-billion-parameter models for continuous real-world planning can quickly trigger thermal throttling or exhaust batteries on remote platforms.

Overcoming this bottleneck demands HW-SW co-design, tailoring modern semiconductor processes directly to the sparse mathematical profiles of agentic workloads.

Neuromorphic computing represents a foundational breakthrough in energy-efficient physical intelligence. Unlike traditional von Neumann processors that continuously cycle energy-intensive data between separate memory and compute units, neuromorphic processors deploy event-driven spiking neural networks (SNNs) that mimic biological brain architectures. These chips consume power only when processing incoming sensory spikes, idling at near-zero power during environmental quiet periods. Integrating neuromorphic silicon into the sensory and perception layers allows agentic devices to process continuous high-bandwidth video, audio, and vibration data at a fraction of standard power requirements.

Parallel innovations in advanced 3D packaging, die-stacking, and high-bandwidth in-memory processing break the classic memory wall that slows down embedded generative inference. Stacking compute logic directly beneath memory arrays drastically slashes data transfer distances, reducing power-per-bit metrics and boosting local inference throughput.

Meanwhile, high-efficiency RISC-V open architectures, augmented with custom vector and matrix math extensions, provide modular silicon foundations tailored for sparse tensor mathematics and low-bit quantization regimes.

The HW innovations are matched by fundamental architectural evolution across artificial intelligence paradigms:

- **Rule-Based Systems (GOFAI):** Highly interpretable, formally verifiable, but brittle when handling non curated, non-linear real-world observations.
- **Deep Learning:** Remarkable perceptual pattern recognition over dense, high-dimensional data, but limited by opaque black-box designs and lack of multi-step planning.
- **Agentic Systems:** Hybrid neuro-symbolic systems that fuse perceptual models with deliberate cognitive planning, self-reflection loops, and multi-agent coordination frameworks.

This shift positions agentic AI to bridge high-level cognitive models with low-level physical dynamics. Agents will ground spatial reasoning by maintaining localized DTs and running parallel forward simulations within virtual triplets before executing physical actions.

By testing alternative action sequences within real-time simulation sandboxes, an agent can evaluate failure modes and optimize real-world execution plans without risking bodily harm or structural damage.

Achieving this future requires deep collaboration across SW engineering, ML, robotics, microelectronics, safety verification, and public policy. The research community must advance past fragile pilots, building standardized benchmark suites that test agentic safety, operational latency, and physical task completion under messy, adversarial real-world conditions.

© AIOTI. All rights reserved.

Agentic AI marks a fundamental turning point in computational design: transforming computers from passive tools that simply process instructions into context-aware, autonomous partners capable of safely navigating, reasoning through, and shaping the physical world.

3.4 Future Research Trends

Future research at the intersection of GenAI, IoT and Edge Computing focuses on radically compressing foundational reasoning capabilities to fit within tight physical, thermal, and electrical limits. Traditional foundation models rely on expansive cloud data centres, creating an urgent need to design sub-billion-parameter architectures that can run efficiently on devices powered by milliwatts of energy. Investigating extreme low-bit mathematical regimes, such as ternary and binary networks, is a vital priority. These representations replace costly floating-point matrix multiplications with simple integer operations, allowing local silicon to generate complex text, audio, and visual outputs without excessive battery drain.

A critical systems challenge lies in overcoming the memory wall inherent to autoregressive sequence generation. Modern transformer models require an ever-expanding cache of contextual states, which quickly consumes the limited static memory available on edge processors. Researchers are actively developing linear-time alternative architectures, such as state-space models and recurrent associative memories. These frameworks maintain a constant, bounded memory footprint regardless of operational duration, enabling uninterrupted localised reasoning over infinite sensor streams without inducing system crashes or memory exhaustion.

Dynamic edge-cloud collaboration represents another compelling research direction. Rather than running generative workloads entirely on an endpoint or relying wholly on remote servers, future frameworks explore distributed speculative decoding and elastic split-computing across heterogeneous nodes. A lightweight draft model running locally on a field gateway can rapidly propose candidate actions or token streams, while a larger cloud-hosted parent model selectively verifies and refines these proposals. This hybrid dynamic requires adaptive networking protocols that gracefully handle packet loss, fluctuating bandwidth, and physical disconnections without disrupting continuous edge operations.

Ensuring deterministic safety and preventing hallucinations in physical operational settings remain essential priorities. Unlike digital chatbots, edge-deployed generative systems interact directly with robotic actuators, automated transit lines, and high-voltage power grids where probabilistic errors can cause physical destruction. Ongoing research seeks to combine generative models with classical control theory and formal mathematical verifiers. By wrapping generative inference engines within deterministic safety filters, future platforms ensure that synthesised control codes and autonomous actions strictly obey physical limits and system invariants before reaching a physical motor or valve.

Enabling privacy-preserving, on-device continual learning poses an open algorithmic puzzle. Physical operational environments continuously shift through changing seasons, mechanical wear, and operational adjustments. Edge generative models must adapt to these changing conditions locally without succumbing to catastrophic forgetting of their core pre-trained knowledge. Developing decentralised fine-tuning techniques that operate directly on streaming private data will empower edge devices to continuously adapt their generative models while keeping sensitive raw inputs permanently on premises.

© AIOTI. All rights reserved.

4 The IoT-Edge-Cloud Continuum

The convergence of AI and particularly practical implementations supporting GenAI-workloads applications while using LLM's as baseline technology, and the Internet of Things (IoT) is transforming the design and architectures of intelligent, distributed systems, introducing new demands for integrating context-aware information, include processing capabilities, adaptive communication, and data cross-domain interoperability. Supporting workloads no longer permits a binary choice between "cloud" and "edge." What is required instead is a fluid (seamlessly), jointly managed computing and data continuum in which generative tasks, training, fine-tuning, retrieval, and inference, are placed dynamically according to data locality, compute capability, network conditions, and energy constraints. Realising this continuum is fundamentally an orchestration problem.

The research community is converging on it from different directions with model-level partitioning. IoT comprises resource-constrained devices that can generate data, prompts, and contextual signals that feed a GenAIoT pipeline. This pipeline is executed across tiers, infrastructure-level orchestration that extends container platforms multiple computing layers: edge nodes, typically one hop away from the data sources, facilitate low-latency inference for time-sensitive GenAI tasks that cannot tolerate a round trip to the cloud. A few hops away, fog nodes offer more substantial computing power with cross-layer awareness, only a small delay, making them well suited to intermediate reasoning tasks, such as fusing the outputs of several edge models. Farther away, cloud nodes provide virtually unlimited resources, appropriate for training and fine-tuning large AI models and intent-level orchestration that lets natural-language policy drive placement directly. for serving requests that exceed local capacity, though their remote location and heavy processing role may still introduce bottlenecks and high latency. Users are also integrated into the continuum through devices such as wearables, smartphones, and laptops, which may connect to any of the computing layers, so GenAIoT workloads can flow seamlessly between layers. In such a GenAIoT deployment, tracking data ownership, privacy, and provenance as prompts and outputs move across the continuum becomes an important design consideration [16].

4.1 Data Interoperability as design principle

The emerging of GenAIoT, either working on edge or cloud, is demanding to have common elements incorporating machine learning, foundational models, and LLMs working together. traditional data-centric communication approaches are increasingly complemented by semantic communication paradigms i.e. labels, tagging meta-tags, etc. that prioritise the contextual interpretation, relevance, and actionable meaning of information. Integrating semantic models, knowledge representation, and ontology engineering into GenAI architectures derivates in enabling devices, edge nodes, and AI agents to move beyond data acquisition towards the interpretation, aggregation, and exchange of meaningful information. This is a completely new paradigm, which it is believed that semantic technologies can demonstrate they can provide a structured foundation for GenAIoT models contextualising sensor observations, retrieve relevant domain knowledge, reasoning over relationships between heterogeneous data sources, and generate responses aligned with machine-interpretable representations [17]. Conversely, generative models can facilitate the interpretation of natural-language queries, the discovery of relevant data and services, and the transformation of user requirements into structured semantic queries and interoperable information requests.

Data interoperability consequently becomes a critical enabler of collaborative GenAIoT ecosystems, helping to ensure that information exchanged across devices, platforms, and applications retains its intended meaning and domain context. These capabilities also support distributed intelligence, in which AI models and agents collaborate across edge and cloud environments while accounting for communication constraints, computational resources, and application requirements. The GenAIoT semantic interoperability concept is introduced here as

© AIOTI. All rights reserved.

a reference architectural component for edge-cloud intelligence, combining LLMs-ready, AI-driven analytics, with generative capabilities, semantic reasoning, and machine-readable knowledge representations to bridge sensor-level observations and application-level understanding [18]. By embedding semantic capabilities throughout the sensing, processing, communication, and decision-making pipeline, GenAIoT systems can evolve from exchanging isolated data streams towards sharing contextualised information and interoperable knowledge, establishing a foundation for more adaptive, explainable, and meaning-aware digital services.

GenAIoT should complement current models with semantic descriptions, rather than replace, explicit current data models. In the other hand, Ontologies provide structured, machine-interpretable representations of concepts and relationships, while generative models offer flexible language understanding and generation. Combining these functionalities GenAIoT can support more robust semantic interpretation and interoperability, provided that generated outputs are validated against relevant knowledge models and data constraints.

The integration of data interoperability by semantics aggregation in Generative AI introduces four complementary capabilities:

- Natural-language-to-semantics translation: GenAIoT can transform user questions and application requirements into structured queries that operate on semantic models, knowledge graphs, and IoT data services.
- Context-aware generation and reasoning: Semantic representations can provide domain context for generative models, helping them interpret heterogeneous sensor observations and produce responses grounded in structured information.
- Semantic interoperability across AI agents: Shared semantic models and machine-readable descriptions can help AI agents to interpret data, capabilities, and services consistently, supporting collaboration across heterogeneous AIoT environments.
- Adaptive semantic communications: GenAIoT can use data models to identify which information is relevant to a particular task, application, or recipient, supporting transmission of meaningful information rather than relying exclusively on raw data exchange.

The evolution of GenAIoT architectures towards distributed intelligence requires a consistent approach to interpreting, processing, and exchanging information across the entire computing continuum, from sensing devices and embedded systems to edge nodes, edge infrastructure, and cloud platforms [19]. While these computational layers offer different capabilities in terms of processing power, storage, connectivity, and latency, their effective integration depends not only on communication protocols and data formats but also on the ability to interpret information consistently across heterogeneous systems. GenAIoT addresses semantic interoperability requirements by providing explicit descriptions of data structures, concepts, relationships, and contextual properties, enabling independently developed devices, services, and applications to exchange and reuse information without losing its intended meaning.

The computing continuum adaptations for GenAIoT demand semantic capabilities that should not be confined to centralised data repositories or cloud-based knowledge management services. Instead, they should be distributed according to the operational requirements and computational capabilities of individual nodes. At the device and edge levels, lightweight semantic annotations and predefined knowledge models can provide sufficient context for immediate interpretation and local decision-making [20]. More computationally capable edge and fog nodes can perform contextual integration, rule-based reasoning, and distributed information retrieval, while cloud infrastructure can support broader knowledge integration, ontology management, computationally intensive analytics, and the training or adaptation of AI models. This distribution enables semantic processing to become an integral part of the information lifecycle rather than an additional step performed after data collection.

© AIOTI. All rights reserved.

The practical implementation of this approach requires balancing semantic expressiveness against resource consumption, particularly in environments characterised by constrained devices, intermittent connectivity, and stringent response-time requirements. Modular ontologies, reusable vocabularies, and selective reasoning mechanisms can help reduce computational overhead while preserving the information necessary for interoperability. Rather than requiring every device to implement an identical semantic processing stack, the architecture can support different levels of semantic capability while maintaining mappings to shared conceptual models. This approach aligns with the principles of modularity, extensibility, and cross-domain ontology harmonisation addressed by ISO/IEC 21823-3 [21][22].

Several complementary technologies can support this GenAIoT data interoperability demand:

- **Ontologies and Knowledge Graphs:** Ontologies provide formal descriptions of domain concepts and their relationships, while knowledge graphs represent interconnected entities and facts using these or other explicit schemas. The W3C Semantic Sensor Network (SSN) ontology provides a model for describing sensors, observations, procedures, observed properties, and actuators. Ontology models can establish a shared interpretation of observations generated by heterogeneous sensing systems. Examples that work in IoT can also be used in GenAIoT, i.e. lightweight SOSA core supports applications that require a more limited set of semantic concepts [23][24].
- **Linked Data and Semantic Resource Identification:** Uniform Resource Identifiers (URIs), the Resource Description Framework (RDF), and related linked data principles provide mechanisms for identifying, describing, and linking information resources. These mechanisms can support the integration of contextual information across distributed nodes without requiring all data to be consolidated in a single repository [25]. Their practical use at the edge depends on appropriate choices concerning data representation, network availability, caching, and resource discovery.
- **Semantic Annotation and Distributed Reasoning:** Associating observations with machine-interpretable descriptions at or near their point of origin can reduce ambiguity before information is consumed by downstream services. Where appropriate, edge nodes can apply lightweight rules or inference mechanisms to identify conditions, derive contextual facts, and trigger local actions. Technologies such as EYE and RDF4J provide capabilities for rule-based reasoning and RDF processing, respectively; their suitability for constrained deployments depends on the selected components, configuration, and workload.
- **Distributed Semantic Query Processing:** Federated SPARQL and related distributed-query approaches enable applications to retrieve information from multiple RDF data sources without requiring complete centralisation of their contents [26]. GenAIoT is expected can support selective retrieval of information relevant to a particular task or decision. Query planning, data availability, network conditions, and access-control policies remain important factors in determining the performance and privacy characteristics.
- **Semantic Support for AI Agents and Generative Models:** Semantic models can provide structured domain context for AI agents and Generative AI systems interacting with heterogeneous IoT services. Ontologies, knowledge graphs, and semantic query interfaces can help map natural-language requests to available concepts and data resources, retrieve relevant information, and constrain generated outputs against established domain knowledge. These capabilities complement conventional semantic reasoning rather than replacing it and require appropriate validation to address incorrect or unsupported AI-generated interpretations.

Collectively, these technologies provide a foundation for semantic interoperability throughout the GenAIoT computing continuum. Their value lies not in requiring every node to perform complex reasoning, but in enabling the different computational layers to exchange information that can be interpreted and used consistently. This distinction is particularly important for distributed intelligence, where data may be processed locally, models may operate across different infrastructure layers, and decisions may depend on information originating from multiple administrative or technological domains.

© AIOTI. All rights reserved.

4.2 Shifting from Centralised Training to Edge Inference

The operational centre of gravity for GenAIoT is moving away from massive, centralised training pipelines toward highly efficient, localised inference at or near the network edge. This shift is being tackled at the model-partitioning level as much as at the infrastructure level. EdgeShard, for instance, formulates joint device selection and LLM layer-partitioning as an optimisation problem, dynamically sharding a model across collaborating edge devices and cloud servers to adapt to heterogeneous, bandwidth-constrained network conditions [27]. A complementary line of work targets collaborative inference architectures in which *SLMs* resident on edge devices handle routine or latency-sensitive requests and selectively escalate more demanding queries to cloud-hosted LLMs, balancing accuracy against latency, cost, and privacy [28]. Neither approach is infrastructure-agnostic, however: both assume that some other mechanism inline is deciding, in real time, which device or tier a given shard or escalation should land on, which is precisely the orchestration problem addressed at the platform level, for example by Kubernetes-native continuum frameworks such as kUDECO [29] (the OSS orchestration framework developed in the CODECO project), which extend the orchestration substrate itself with cross-layer (compute, network, data) awareness [13]. The escalation and sharding logic in EdgeShard-style and SLM/LLM-collaboration systems cannot be handled by static routing tables: it is itself a continuous, cross-layer, context-aware orchestration decision, reinforcing the case that inference placement and infrastructure orchestration need to be co-designed rather than treated as separate layers.

This shift away from centralised training is already visible in GenAIoT systems. For example, such systems can use Hierarchical Federated Learning, with edge and fog nodes aggregating client model updates before synchronising with the cloud, reducing wide-area communication. The same layered structure applies to fine-tuning an AI model across many edge deployments: a control-plane policy decides which compact summaries to send to which edge node, under strict bandwidth and delay budgets, since a full reference model cannot be broadcast to every node every round [30]. Alternative GenAIoT approaches instead keep the generative model entirely in the cloud, with edge devices only sampling and streaming reduced data to lower the cost of transmitting large datasets. In such cases, predictive models in the cloud can lower the sampling frequency at the edge and reconstruct the missing data centrally, adapting the sampling rate dynamically to keep the prediction and reconstruction error low. This can significantly reduce the overhead of running large AI models across the edge-cloud continuum [31].

4.3 Designing Hybrid Architectures for Distributed Intelligence

Hybrid computing models are becoming a necessity for GenAIoT, combining edge and cloud intelligence so that systems benefit from low latency local inference while retaining the option of cloud fallback for complex reasoning, continuous learning, and model updates. Different projects realise this hybrid model.

Moving this hybrid model towards the data source can improve responsiveness and reduce the need to transmit every observation to central infrastructure, while semantic context can help make locally generated results interpretable across systems. This section can distinguish between two activities:

- **Model training:** Often computationally intensive and therefore commonly performed using cloud or data-centre resources, although distributed and on-device training are also possible.
- **Model inference:** Executing a trained model to generate predictions or classifications, which can be performed on devices, edge nodes, fog infrastructure, or cloud platforms, depending on the application.

© AIOTI. All rights reserved.

Semantic technologies contribute by describing the meaning of input features, units, observed properties, model outputs, and the operational context in which predictions are produced. These descriptions can help downstream systems interpret inference results consistently.

GenAIoT introduces an additional dimension: selected language models or other generative components may support local interaction, contextual information retrieval, or decision-support workflows. However, the suitability of local execution depends on model size, hardware resources, latency requirements, and the need to access external knowledge. Semantic grounding can help constrain generated responses but does not independently guarantee their correctness [32] .

Deploying generative models within industrial environments must reckon with tangible physical constraints, including spatial limitations, thermal dissipation envelopes, restricted electrical power budgets, and volatile memory ceilings across edge HW, ruling out the installation of monolithic compute clusters adjacent to machinery as well as complete reliance on remote data centres [33]. To resolve this trade-off, the architecture organises computational workloads across a processing continuum. The edge devices can be optimised based on different aspects, like processing, memory, energy, connectivity, size, cost. Their performance and capabilities are of course constrained by these parameters.

A description of the micro-, deep- and meta-edge concepts is provided in the following sections and aligned with other European partnerships [33]. The granularity of the Edge Computing environments is illustrated in Figure 4-1 [13][34][35].

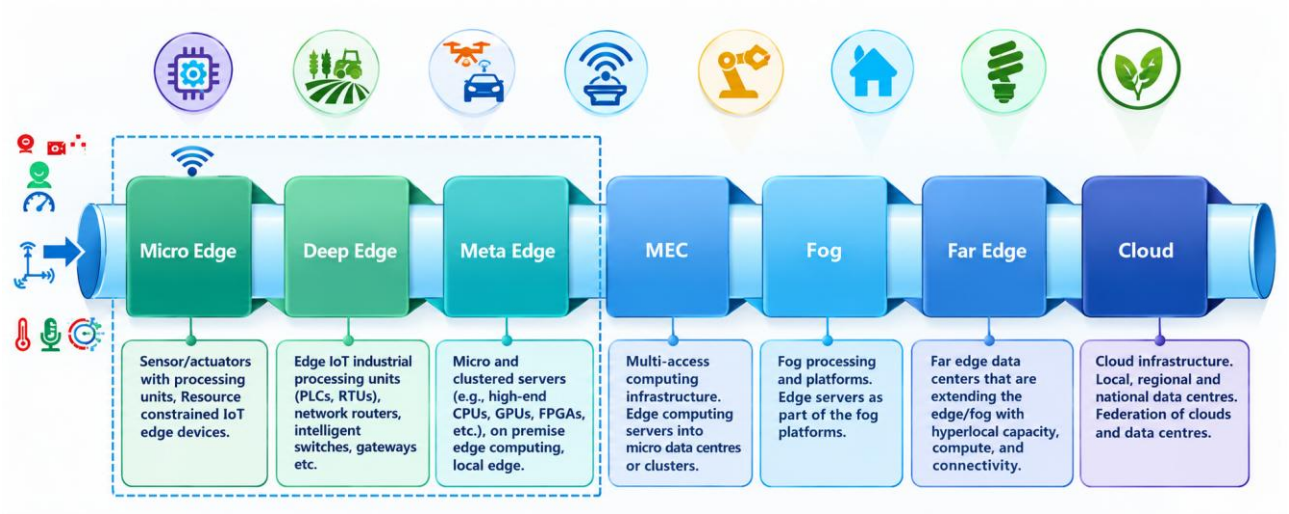


Figure 4-1 Edge Computing granularity across the computing continuum for GenAIoT processing.

The ***micro-edge*** describes the intelligent sensors, machine vision, and IoT devices that generate insight data and are implemented using microcontrollers built around processors that can well cope with costs and power consumption constraints. The distance from the data source generated by the sensors is minimised. The compute resources process this raw data in line and produce insight data with minimal latency. The HW devices of the micro-edge physical sensors/actuators generate from raw data insight data or actuate based on physical objects by integrating AI-based elements into these devices and running AI-based techniques for inference and self-training.

Intelligent micro-edge allows IoT real-time applications to become ubiquitous and merged into the environment where various IoT devices can sense their environments and react fast and intelligently with a very low power budget energy-efficient gain. AI capabilities integrated into IoT devices significantly enhance their capabilities (functionality, performances, low latency, low power consumption, high processing power) and usefulness, especially when the full power of these networked devices is harnessed – a trend called *AI on edge* [36][37].

© AIOTI. All rights reserved.

The ***deep-edge*** comprises intelligent controllers like Programmable Logic Controllers (PLC), Supervisory control and data acquisition (SCADA) elements, machine vision connected embedded systems, networking equipment, gateways and computing units that aggregate data from the sensors/actuators of the IoT devices that generate data. Deep edge processing resources are implemented with performant processors and microcontrollers such as Intel i-series, Atom, ARM M7+, etc., including Central Processing Units (CPUs), Graphics Processing Units (GPUs), Tensor Processing Units (TPU), and ASICs. The system architecture, including the deep edge, depends on the envisioned functionality and deployment options considering that these devices' cores are controllers: PLCs, gateways with cognitive capabilities that can acquire, aggregate, understand, react to data, exchange, and distribute information [33].

The ***meta-edge*** integrates processing units, typically located on-premises, implemented with high-performance embedded computing units, edge machine vision systems, edge servers (e.g., high-performance CPUs, GPUs, Field-Programmable Gate Array (FPGAs), etc.) that are designed to handle compute-intensive tasks, such as processing, data analytics, AI-based functions, networking, and data storage [33].

This classification is closely related to the distance between the data source and processing, impacting overall latency. The specific processing characteristics, HW targets, and latency constraints of the various components spanning from the extreme micro-edge to the centralised cloud are detailed in Table 4-1.

Table 4-1 Comparative Analysis of the Edge-to-Cloud Processing Continuum Tiers.

Tier	Typical Hardware	Key Accelerators	Latency Target	Primary Function
Micro-Edge	MCUs, MPUs, SoCs	TinyML, micro-NPU, NPU	< 1 ms	Sensing, simple inference, wake-up triggers. Sensor ingestion, fast low-power discriminative analysis (1–4 bits), noise filtering, deterministic safety interlocking
Deep-Edge	Gateways, Industrial PCs	Edge GPU, TPU, VPU, NPU	2 - 5 ms	Aggregation, sensor fusion, local control
Meta-Edge	On-premises micro-servers, high-performance embedded servers	High-end GPU, FPGA	< 10 ms	Local training, complex analytics, orchestration. Local inference of compact language models (1B–8B quantized), contextual manual querying, multi-agent coordination
Multi-Access Edge Computing (MEC)	Telecom-Grade servers, rack servers	GPUs for packet processing, FPGAs for network function virtualization	< 30 ms	Content caching, high-bandwidth processing, application offloading
Fog	Industrial routers, bigger servers	TPUs, large FPGAs, Digital Signal Processor (DSPs)	< 50 ms	Bridging Operational Technology and Information Technology. Focus on data processing not in the cloud

The seamless coordination of this distributed workload relies on three primary methods established within distributed systems literature [38][39]. The first is split computing, wherein the edge asset extracts primary features and low-dimensional representations without transmitting raw high-bandwidth telemetry upstream, forwarding only compact intermediate tensors to the plant-floor server; this prevents wireless link saturation whilst safeguarding proprietary industrial data [35][38]. The second is the early-exit mechanism, which resolves nominal and high-confidence operational queries directly on the field device, escalating only ambiguous edge cases or high-uncertainty anomalies to the intermediate tier. The third comprises distributed speculative decoding, wherein a compact, rapid model deployed at the physical asset generates preliminary draft tokens that a larger model hosted on the local server validates in a single forward pass, substantially accelerating aggregate inference throughput without sacrificing predictive fidelity [39].

For this multi-tier infrastructure to operate effectively, the interconnecting network fabric is just as critical as the compute nodes themselves. Contemporary wireless deployments face severe Quality of Service (QoS) bottlenecks, spurring the research community to formulate novel

© AIOTI. All rights reserved.

technologies and standards designed to enhance the determinism and dependability of such systems. Foremost among these, 3GPP 5G technology configured with Ultra-Reliable Low-Latency Communication (URLLC) profiles alongside standard ETSI MEC interfaces stands as the principal candidate to deliver dependable, scalable, and secure connectivity [12][40]. At the SW tier, European research consortia investigating distributed meta-operating systems (such as *aerOS* [6] and *FluidOS* [41]) explore lightweight container runtimes and broker less data-sharing protocols (such as *Eclipse Zenoh* [42] and MQTT v5) to dynamically balance microservices, taking into account instantaneous radio channel conditions, processor thermal limits, and real-time execution deadlines [40][43]. Model lifecycle management proceeds asynchronously: historical plant telemetry is aggregated, anonymised, and streamed to the cloud during off-peak intervals to train lightweight low-rank adapters (LoRA), which are subsequently redeployed and flashed onto field HW via FW Updates Over-The-Air (FUOTA) without disrupting ongoing production [35][44]. Empirical evaluations across pilot manufacturing plants demonstrate that this heterogeneous framework curtails cloud-bound backhaul traffic by more than 70% [35][38], sustains full operational autonomy throughout wider connectivity blackouts, and delivers deterministic closed-loop inference within sub-45-millisecond bounds [43].

The IEC 61499 [45] provides a relevant architectural foundation for designing distributed-intelligence systems spanning industrial devices, Edge Computing nodes, and cloud resources. It defines a generic architecture for function blocks in distributed systems, including mechanisms for configuring distributed systems and mapping function-block instances to distributed computing platforms. The separation between an application's functional specification and the physical resources on which its components execute is key to the standard's relevance for distributed intelligence. The work [46] characterizes IEC 61499 as an architecture intended to enable intelligent automation in which intelligence can be encapsulated in SW components and distributed across networked devices – at IoT, edge, or cloud.

IEC 61499 provides mechanisms through which GenAI functions can be integrated, in particular, Service Interface Function Blocks (SIFBs) provide standardized application-level interfaces to services whose internal implementation is abstracted; hence GenAI functionality can remain logically embedded and composable within the IEC 61499 application boundary. Industries such as Schneider Electric and some others [47] have adopted the technology as a foundation for building HW-independent, portable intelligent SW systems.

The work [48][49][60] demonstrate application in IEC 61499 that embed intelligence (in stochastics / "neuro" AI or mixed neuro - symbolic AI) can target and be deployed across heterogeneous platforms ranging from low-cost ARM controllers to Edge Computing nodes and cloud server. There are SW frameworks such as [50] which support the design and development of AI-enabled applications in hybrid architecture.

Applications that embed GenAI in hybrid architecture but have timing and functional safety constraints shall consider appropriate separation of concern principle by separating deterministic functionality from computationally intensive GenAI functions. Purely cloud-based GenAI provides substantial computational capacity but may introduce latency, whereas edge–cloud collaboration offers a means of combining cloud-scale computation with lower-latency execution closer to users and physical systems [51]. Studies have identified memory limitations, computational demand, and heterogeneous HW as major constraints on local LLM deployment, thereby motivating collaborative or hierarchical edge–cloud execution strategies [51]. IEC 61499 may provide application-level composition and coordination layer within GenAI-enabled applications with hybrid architecture, where different architectural commitments: forms of intelligence, such as GenAI, can be deployed and distributed across devices and computing layers depending on the available resources and operational requirements, as demonstrated in [48], where for example, a small or domain-specific model could execute locally for low-latency, a larger model could execute at the edge when greater reasoning capacity is required while retaining local data, and a foundation model hosted in the cloud could support tasks requiring substantially greater computational resources or broader contextual knowledge.

© AIOTI. All rights reserved.

- **Federated, neighbourhood-scoped orchestration.** KUDECO operationalises hybrid placement through a layered governance structure based on centralised policy and intent management at a Hub cluster, coupled with bounded, application-scoped "neighbourhoods" of Managed Clusters that retain local autonomy for scheduling and adaptation even under intermittent connectivity, with AI-assisted (Graph-Neural Network (GNN)/ Multi-Agent Reinforcement Learning (MARL)-based) recommendations kept separate from deterministic scheduling enforcement to limit oscillatory behaviour [52]. This design has been validated in domain-specific case studies, including edge orchestration of autonomous mobile robots in smart-manufacturing settings, where it reduced CPU consumption and produced more stable inter-pod communication than standard Kubernetes scheduling, at the cost of modest memory overhead [53].
- **Multi-agent, throughput-optimised scheduling.** Taking a different architectural route, the Kubernetes Multi-agent Proximal Policy Optimization Scheduler (KMPS) combines multi-agent proximal policy optimisation with graph neural network state embedding and a three-timescale coordination mechanism to perform cross-cluster Kubernetes scheduling for edge-cloud systems, explicitly targeting the throughput bottlenecks that single-cluster or purely service-orchestration-level approaches face [54]. Decentralised, energy-aware serverless scheduling for the cloud-edge continuum has also been formalised as a consensus-based MARL problem, in which distributed agents jointly weigh green-energy availability against per-function QoS requirements without central coordination [55]; a related hybrid design combines local, neighbourhood-scoped MARL agents with a GNN-embedded global orchestrator to balance decentralised adaptability against system-wide coherence. Considering these aspects, together with the CODECO proposed separation of AI-driven recommendations and placement enforcement, these systems show the capability to support broader, resilient convergence: whatever the specific coordination mechanism (federated neighbourhoods, MARL, or hybrid local/global control), continuum orchestrators are converging on architectures that keep learned recommendations advisory and deterministic placement authoritative, precisely to avoid the instability that purely learning-driven control can introduce.
- **Energy-aware placement as a first-class objective.** Independent of the coordination mechanism, several of these systems fold energy directly into the scheduling objective rather than treating it as a secondary metric. CODECO-based work injects node- and network-level energy metrics ("greenness" heuristics) into ILP-based placement [56], while the serverless MARL work by Patel et al. optimises directly against green-energy node availability [57]. This convergence across otherwise-different architectures suggests energy-as-objective, not just energy-as-constraint, is becoming a baseline expectation for continuum schedulers generally.

Orchestration approaches beyond container/workload placement are converging on similar design principles from the intent side. Intent-based orchestration frameworks use LLMs to translate natural-language policy goals including privacy and data-locality constraints directly into enforceable Kubernetes placement rules and SDN flow configurations spanning the cloud-edge continuum, reducing the expert effort otherwise required to hand-craft such policies [58]. In mobile edge-network settings, agentic architectures built on generative models are being explored to move orchestration from reactive to proactive: predicting users' implicit service intents from multidimensional context and pre-provisioning service function chains ahead of demand, rather than orchestrating purely in response to observed load [59].

This hybrid principle in GenAIoT, separates an AI model's weights and heavy reasoning, which stay in the fog or cloud, from the lightweight control signals that keep edge deployments coordinated. In hierarchical federated learning, for example, the control-plane state can be a compact semantic summary that each edge node uses to decide locally which knowledge gaps to close next. This way, a distributed GenAIoT system can keep adapting even when the cloud cannot broadcast a full reference model at every training iteration. Prioritising the summaries most likely to change a node's decisions, rather than the most out-of-date ones, can significantly speed up recovery from missing local knowledge compared to only minimising

© AIOTI. All rights reserved.

model drift [30]. A similar hybrid pattern appears in edge-cloud data-prediction systems, where a lightweight component on the edge device streams reduced data to the cloud, and resource-demanding models and controllers in the cloud handle the additional decision logic. Coordination between edge and cloud is pivotal here, adapting the sampling frequency at the edge while reducing the prediction error in the cloud. Such hybrid architectures help GenAIoT and IoT deployments reason efficiently while scaling across many devices in the computing continuum [31].

For years, Europe advanced its digital infrastructure along two parallel tracks. On one side, the Cloud-Edge-IoT community worked to connect devices, data and computing power across the continent. On the other side, the telecom and smart networks community drove Europe's progress from 5G towards 6G. Two strong ecosystems, each with its own roadmap, its own community, and its own momentum. But the digital services of the future, from AI to autonomous mobility to real-time industrial systems, don't live neatly in one world or the other. They need networks, cloud, edge, and devices working together as a single, seamless system.

That meeting point is the **3C Initiative**: **C**onnected **C**ollaborative **C**omputing **N**etworks. It represents the moment Europe's cloud story, and its telecom story converges into one continuum, a sovereign Telco-Edge-Cloud fabric stretching from the smallest device all the way to the data centre. Connected Collaborative Computing Networks (3CNs) and distributed GenAI represent an emerging computing paradigm in which interconnected devices, edge infrastructure, cloud platforms, and AI agents collaborate to deliver intelligent services The Connected Collaborative Computing initiative (3C Network) has the objective to develop a Telco-Edge-Cloud-IoT continuum that should be able to support distributed GenAIoT. The 3CN will provide end-to-end infrastructures integrating devices, network computing and communication capabilities.

4.4 Future Research Trends

Across the described above landscape, described by model-partitioning systems like EdgeShard, platform orchestrators like CODECO and KMPS, and decentralised MARL schedulers, a recurring set of open questions emerges: formal stability guarantees for hybrid AI-assisted control loops that combine learned recommendation with deterministic enforcement; optimal privacy-preserving aggregation strategies for federated or multi-agent scheduling at Internet scale; and standardisation of the intent representations and cost-map abstractions that would let these otherwise-incompatible orchestration approaches interoperate rather than each solving placement in isolation.

Hybrid architectures combine device-level responsiveness, edge-level processing, and cloud-level resources, while semantic interoperability enables their outputs and services to be integrated into a coherent system. This section can develop four architectural principles:

1. **Workload placement:** Assign processing tasks according to latency, computational requirements, data sensitivity, and connectivity.
2. **Semantic continuity:** Preserve the meaning and relevant context of information exchanged between heterogeneous components.
3. **Distributed coordination:** Enable services and AI agents to discover relevant information and capabilities across infrastructure boundaries.
4. **Knowledge and model management:** Coordinate shared semantic models, model versions, and inference outputs while accounting for local operational constraints.

A hybrid architecture may combine local rules and lightweight inference, edge-based analytics, and cloud-based training or large-scale reasoning. Generative AI agents may provide interfaces for querying distributed services, while ontologies and other structured knowledge representations help define the available concepts and relationships.

© AIOTI. All rights reserved.

A promising research direction is to extend the IEC 61499 technology towards adaptive distribution of intelligence. A hybrid architecture could determine whether an intelligent function should execute on a IoT or field device (such as PLCs), industrial edge node, or cloud resource according to factors such as latency, computational demand, model capability, privacy, energy consumption, and resource availability. The work done in [50] provide an initial foundation for this idea by optimizing the placement of computation-intensive IEC 61499 intelligent SW logic across heterogeneous computing resources. Combining such mechanisms with emerging edge–cloud foundation-model techniques such as [61] may enable IEC 61499 hybrid architecture to possess adaptive capability where deterministic and non-deterministic functions (i.e. GenAI based functions) can be anchored, deployed, and optimized to appropriate computing platforms without violating fundamental key requirements.

Looking to 3CN, the main research areas with regards to GenAIoT should address the following:

- Network orchestration: Coordinating heterogeneous computing nodes.
- Resource management: Allocating CPU, GPU, memory and bandwidth.
- Communication Efficiency: Reducing communication latency and overhead.

© AIOTI. All rights reserved.

5 Implications for Edge Computing Infrastructure

Edge Computing plays a fundamental role in enabling Generative Artificial Intelligence for the Internet of Things (GenAIoT) by bringing intelligence closer to the physical environments where data is generated, processed, and consumed. Traditional cloud-centric AI architectures face limitations in supporting the distributed, heterogeneous, and time-sensitive nature of IoT ecosystems, particularly regarding communication latency, network bandwidth, data privacy, and operational autonomy. By decentralising AI capabilities towards edge devices and intermediate computing nodes, the integration of Generative AI into Edge Computing introduces new opportunities for transforming IoT systems from conventional data collection and monitoring platforms into intelligent environments capable of interpreting contextual information, generating insights, and supporting autonomous operations. However, the computational complexity, memory requirements, and energy consumption associated with generative models demand new architectural approaches that balance intelligence distribution with available edge resources. This motivates the adoption of lightweight and optimised models, distributed inference, collaborative edge–cloud processing, and resource-aware orchestration mechanisms. Consequently, Edge Computing is not merely an alternative deployment environment for Generative AI, but a key architectural enabler for scalable, responsive, privacy-aware, and resilient GenAIoT systems. In the other hand, decentralising generative intelligence requires a fundamental rethinking of how edge infrastructure is provisioned and managed, as the heavy computational demands of foundation models alter the design of edge nodes and aggregation layers.

5.1 Real-World Deployment Constraints

The heavy computational demands of GenAI workloads cannot be properly fulfilled by the limited processing capabilities of existing edge deployments. This gap requires new engineering strategies that balance generative capabilities against strict constraints on latency, energy budgets, thermal limits, and network bandwidth. Two factors compound this beyond raw processing headroom: i) **energy footprint is not solely a compute-layer problem**: request-level energy depends jointly on model tier, HW class, and the network path a request travels, so orchestration that optimises workload placement alone systematically misses network-side savings. Recent work shows energy-aware scheduling that folds in network telemetry alongside compute telemetry can materially reduce end-to-end energy versus execution-only optimisation [62][63]; ii) **data sovereignty is increasingly a hard deployment constraint rather than a compliance afterthought:** many current brownfield industrial environment workloads are built on private, regulated, or licence-restricted documentation (e.g., manufacturing standards, Manufacturing Execution System (MES) records) and run on legacy computing platforms traditionally in the form of PLCs. Consequently, the "static edge/server split" common in current agentic frameworks (where a fixed confidence threshold always escalates execution to the cloud) is not just energy-blind, but in regulated settings is inherently legally unviable as a fallback. This reframes edge-cloud offloading beyond a latency-vs-computing trade-off, since these decades old PLCs lack the computational strength to execute modern GenAI architecture such as LLMs [64].

Recently, industrial edge vendors have started offering more powerful industrial computers that combine the compatibility to support legacy PLC with higher computational capability, allowing a three-way trade-off between **energy**, **quality**, and **sovereignty**, evaluated per request against heterogeneous HW (server GPUs, consumer accelerators). However, the increase in the cost of static and dynamic memory HW due to growing demand from data centre investments [65], which stems from the rapid adoption and utilization of GenAI, and supply chain bottlenecks following geopolitical instability have created unprecedented financial barriers, particularly small for businesses that plan to deploy these on-premises, edge devices. As a result, current orchestrators still fail to characterise ana manage such distributed GenAI deployments.

© AIOTI. All rights reserved.

5.2 Architectural Patterns for Scalable GenAIoT

Scalable GenAIoT architectures require flexible combinations of device-level processing, edge intelligence, and cloud-based generative services to accommodate heterogeneous hardware, workload demands, and operational constraints. Emerging architectural patterns should support dynamic workload placement, distributed inference, and adaptive resource orchestration, balancing response quality, latency, energy efficiency, and data sovereignty. Semantic interoperability can strengthen these architectures by enabling consistent interpretation of device capabilities, model requirements, data resources, and service descriptions across distributed infrastructure. Integrating semantic models and knowledge representations with Generative AI and agent-based orchestration offers opportunities to coordinate heterogeneous components and discover relevant data and services across the edge–cloud continuum. Rather than relying exclusively on static deployment configurations, semantically enriched architectures can support context-aware orchestration and more interoperable interactions between AI agents, IoT devices, and generative services, providing a foundation for scalable, adaptive, and resource-efficient GenAIoT ecosystems.

A recurring pattern for GenAIoT systems that act on the physical world is **uncertainty-gated hybrid control**, a variant of the Simplex architecture [143]. A generative model and a simple deterministic controller run side by side on the same edge node, and a decision module chooses, in every control cycle and for every actuator, which of them is allowed to act. The pattern refines the two-tier architecture of Figure 2-11 by making the hand-over between the tiers explicit and testable. The generative model stays in the loop for what it does well, such as interpreting operator requests, combining context and explaining decisions, while the deterministic layer guarantees the safety-critical floor. Figure 5-1 shows the decision flow. The checks run in order, and each cycle ends in one of three states:

- **RULE**: if inputs are stale, the previous cycle missed its deadline, the plant is outside the envelope in which the actuators can hold the setpoint, or a reading lies outside the configured safe band, the deterministic controller acts at once, without waiting for the model and without anti-oscillation lockouts;
- **HOLD**: if the model's proposal is uncertain, measured as disagreement among several sampled proposals, or contradicts the deterministic controller on a safety-relevant actuator, the proposal is not executed. In an autonomous cycle the deterministic decision stands; in an interactive session the operator is asked to confirm;
- **ACT**: otherwise, the proposal is executed through the existing safety interlocks.

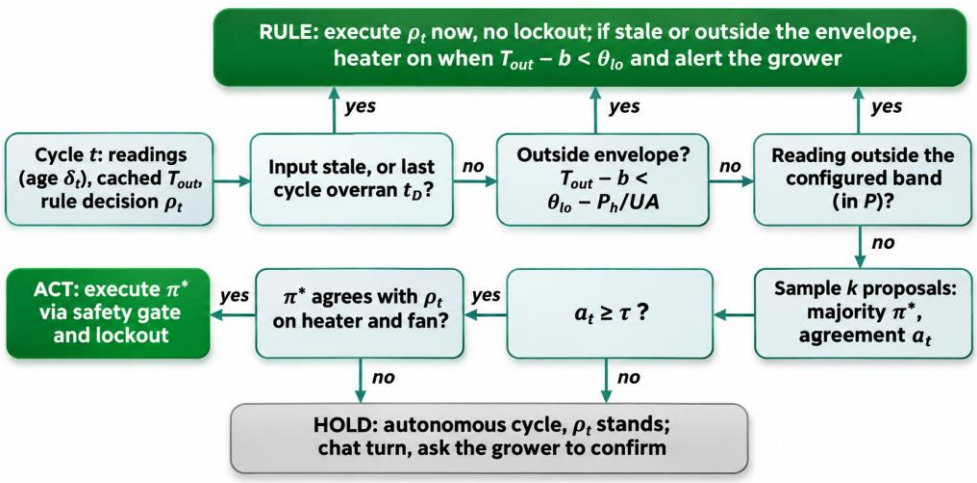


Figure 5-1 Uncertainty-gated hybrid control: decision flow in each control cycle, applied per actuator; the checks run in order (symbols as defined in [142]).

© AIOTI. All rights reserved.

The pattern has been evaluated on an operational greenhouse controller in which quantised open-weight language models with 3 to 14 billion parameters switch heating, ventilation and lighting through typed tool calls, on edge HW that includes a Qualcomm QCS6490-based RUBIK Pi 3 [141][142]. Replayed on 3,840 stored benchmark runs, the gate reduced unsafe actuation in situations that required restraint from 39–85% to 3–12% across four models, without lowering decision accuracy for any of them [142]. Two lessons from this evaluation apply to GenAIoT deployments more generally.

Sampling-based uncertainty is necessary but not sufficient. Disagreement among sampled outputs flagged the model's indecision but missed its systematic errors: for one model, half of the wrong decisions were unanimous, and all of them were cases in which the model should have acted and did nothing. A gate based on agreement alone made that model less accurate than the existing hybrid controller. The independent cross-check against the deterministic controller is what caught these omissions.

Timing must be budgeted in physical units. The decision cadence, not model inference, dominated the worst-case latency from a threshold breach to an actuator command, and the latency tail decided whether the margin to crop damage held. The designed worst case of 7 to 8 minutes consumed less than half of a 3°C margin, whereas the logged 99th-percentile cycle would have consumed 93% of it and the longest logged cycle more than three times the margin [142]. The pattern therefore needs a watchdog that enforces the cycle deadline and a defined, fail-closed behaviour for stale or missing inputs, rather than relying on the model to respond in time.

The pattern scales because its safety-critical part is small. The deterministic controller and the gate run on the same constrained device as the model, need no cloud round trip, and remain available when the model, the network or the inference stack fails. The price is liveness rather than safety: more decisions are deferred to the deterministic layer or to a human, and in the greenhouse evaluation 31–83% of the actions requested by operators in chat were held for confirmation [142]. In edge–cloud deployments the same gate can be applied at each tier, so that escalation to a larger model change who proposes an action but not who is allowed to execute it.

5.3 Future Research Trends

Recent surveys identify three principal families of LLM compression techniques: quantization, pruning, and knowledge distillation, complemented by architectural and inference-level approaches such as mixture-of-experts, early exiting, and efficient attention mechanisms [66][67]. The state of the art increasingly suggests that these techniques should not be considered independently. Instead, compression is evolving toward combinations of weight and activation quantization, structured sparsity, knowledge transfer, efficient model architectures, and HW-aware inference.

For GenAIoT, this transition is particularly important because the optimal compression strategy depends not only on the model and target accuracy, but also on the characteristics of the edge device, including available memory, memory bandwidth, processing capabilities, accelerator support, power budget, and the ability to exploit the compressed representation efficiently.

Quantization as the dominant compression mechanism

Among existing approaches, quantization has emerged as the most mature and widely adopted technique for reducing the deployment cost of LLMs. Quantization represents model parameters and, potentially, activations using lower numerical precision than conventional Floating Point 16 (FP16) or FP32 representations. The reduction in bits directly decreases model

© AIOTI. All rights reserved.

storage and memory traffic and can also enable more efficient arithmetic on CPUs, GPUs, NPUs, and dedicated AI accelerators.

The research frontier has subsequently moved below 4 bits. Methods such as QuIP explored theoretically motivated 2-bit quantization using incoherence transformations and demonstrated that meaningful LLM performance can be retained at extremely low precision [68]. Other works investigated combinations such as 4-bit weights with 8-bit activations, as well as micro scaling formats. Experiments with micro scaling formats demonstrated that combinations of established Post-Training Quantization (PTQ) techniques can achieve 4-bit weights and 8-bit activations with little degradation relative to higher-precision baselines [69].

An especially interesting direction for GenAIoT is the emergence of sub-2-bit and ternary models. Rather than compressing an existing FP16 model after training, approaches such as BitNet pursue models whose weights are intrinsically represented using very low-bit values. This changes the problem from conventional compression to compression-aware model design. The associated BitNet inference work demonstrates that ternary models can be executed using specialized kernels, with substantial speed improvements on edge-relevant CPU platforms [70]. Such approaches could become particularly important for IoT gateways, mobile processors, and low-power edge servers, because they potentially reduce not only memory consumption but also the energy and bandwidth associated with moving model parameters.

Pruning and structured sparsity

A second major compression direction is pruning, in which parameters, neurons, channels, attention heads, layers, or other computational structures are removed. Pruning can theoretically provide greater reductions than quantization because it reduces the number of operations as well as the number of stored parameters. However, the practical benefits of pruning on edge devices are more complicated.

Early approaches largely relied on magnitude pruning, removing parameters with small absolute values. For LLMs, however, the relationship between parameter magnitude and functional importance is considerably more complex. Nevertheless, an important infrastructure lesson is that algorithmic sparsity does not automatically translate into systems-level acceleration. Dense matrix multiplication is highly optimized on contemporary CPUs, GPUs, and NPUs, whereas irregular sparse computation can suffer from indexing overhead, poor memory locality, underutilization of processing elements, and insufficient HW support.

Recent work has therefore explored combinations of structural pruning and knowledge distillation. The MINITRON approach, for example, combines depth, width, attention, and Multi-Layer Perceptron (MLP) pruning with distillation-based retraining and reports 2–4× model-size reductions while using substantially fewer training tokens than training separate compressed models from scratch [71]. This approach illustrates a broader trend: rather than viewing compression as simply deleting parameters, state-of-the-art methods increasingly re-optimize the compressed model to recover the capabilities lost during compression.

For GenAIoT, structured pruning is particularly promising when the target HW exposes a corresponding computational structure. Removing Transformer layers, reducing hidden dimensions, or decreasing the number of attention heads can lead directly to smaller computational graphs that are easier to map onto NPUs and edge accelerators. In contrast, arbitrary unstructured sparsity may produce impressive compression ratios on paper while providing limited benefits after deployment.

Knowledge distillation and the transition from LLMs to compact generative models

Knowledge distillation provides a fundamentally different compression mechanism. Rather than directly modifying the original model, a smaller student model is trained to reproduce the behaviour of a larger teacher model.

© AIOTI. All rights reserved.

For generative models, distillation is particularly attractive because it can transform a large general-purpose model into a smaller model specialized for a particular application domain. This has direct implications for IoT. Edge devices frequently operate in relatively narrow application domains, for example, industrial monitoring, predictive maintenance, robotics, smart-home control, or vehicular systems. Distillation offers a mechanism for transforming those capabilities into a smaller model while retaining the behaviour that matters for the specific workload.

Recent studies combining pruning and distillation reinforce this observation [72][73]. This suggests a particularly relevant design principle for GenAIoT: compression should be application-aware rather than universally optimized for parameter count. A 2-billion-parameter model specialized for industrial anomaly diagnosis may be more valuable at the edge than a heavily compressed 7-billion-parameter general-purpose model.

Beyond model weights: compression of the KV cache

A critical development is that compressing model weights is no longer sufficient for many generative workloads. During autoregressive transformer inference, previously computed KV representations are stored in the KV cache to avoid recomputing attention states for all preceding tokens. The cache grows with context length and batch size and can eventually become a dominant component of memory consumption. Recent surveys identify KV-cache management as a major bottleneck for long-context LLM inference, with techniques including token selection, cache eviction, quantization, merging, low-rank representations, and system-level memory management [74][75].

This issue is particularly significant for edge infrastructure. Edge applications such as conversational assistants, multimodal agents, and IoT controllers may maintain long-lived interactions while operating with much smaller memory capacities than cloud servers. Consequently, an edge LLM can have a sufficiently small static model footprint while still exceeding available memory because of its dynamically growing KV cache.

KV-cache compression therefore extends the concept of model compression from static parameters to dynamic inference state. Quantizing KV values, retaining only the most important tokens, merging similar representations, or exploiting architectural mechanisms such as grouped-query attention can reduce this dynamic memory requirement. However, recent research also highlights that cache compression can introduce less obvious quality degradation, particularly for complex multi-instruction prompts and workloads that differ from the benchmarks used during method development [76]. Thus, as with parameter compression, the challenge is not simply maximizing compression but identifying which information can safely be discarded for a particular workload.

HW-aware compression as an infrastructure requirement

The most important implication for Edge Computing is that compression ratios should no longer be treated as sufficient measures of success. The same compressed model can behave very differently on different HW platforms. A 4-bit representation may be highly advantageous on an accelerator with native INT4 support but provide limited benefits on a processor whose kernels internally unpack the representation into higher precision. Similarly, unstructured sparsity may dramatically reduce the theoretical number of operations while leaving end-to-end latency almost unchanged.

This has led to increasing interest in algorithm–HW co-design, in which quantization formats, sparsity structures, model architectures, kernels, memory systems, and accelerators are designed together. Recent edge-oriented work combines pruning, low-bit KV-cache quantization, and specialized FPGA acceleration, illustrating that the best results increasingly arise from jointly optimizing the model and its execution substrate [77].

© AIOTI. All rights reserved.

Consequently, future edge infrastructure should expose compression-aware primitives at multiple levels of the stack: HW support for low-precision arithmetic and structured sparsity; compiler and runtime support for compressed tensors; memory hierarchies optimized for quantized parameters and KV caches; and deployment frameworks capable of selecting an appropriate representation according to the target device. Compression should effectively become a first-class property of the model–runtime–HW interface.

Implications for GenAIoT architectures and open challenges

These developments point toward a broader architectural shift in GenAIoT. The objective is unlikely to be the complete migration of large foundation models to every IoT endpoint. Instead, a hierarchical intelligence architecture is likely to emerge.

Small, aggressively compressed models can execute locally for latency-sensitive, privacy-sensitive, or connectivity-independent tasks, while larger models remain on more capable edge servers or in the cloud. Compression consequently determines not only whether a model fits on an individual device but also where different components of a GenAI pipeline can be executed.

A further challenge is heterogeneity. IoT environments contain CPUs, GPUs, NPUs, DSPs, FPGAs, microcontrollers, mobile processors, and specialized accelerators, each supporting different numerical formats and sparsity patterns. A compression strategy optimized for one platform may therefore be suboptimal on another. This makes automated HW-aware compression and deployment increasingly important.

The central research challenge is consequently shifting from "How much can an LLM be compressed?" to "How should a generative model be compressed for a particular edge system?".

- **Model-level sharding and partitioning.** EdgeShard-style approaches solve a joint device-selection and layer-partitioning problem to distribute an LLM's own computation across heterogeneous edge and cloud resources under varying network conditions, treating the model itself as a divisible, re-placeable unit rather than a monolithic deployment target.
- **Cross-layer, data-compute-network-aware infrastructure orchestration.** Kubernetes-native continuum orchestrators extend the platform layer with joint compute, network, and data observability, so that placement decisions account for end-to-end conditions rather than compute metrics alone. KuDECO is one representative example of this pattern, using a federated, neighbourhood-scoped governance model, while KMPS pursues the same cross-layer awareness through a multi-agent, GNN-embedded scheduling approach with an explicit multi-timescale coordination mechanism rather than federated neighbourhoods.
- **Decentralised, negotiation- or consensus-based scheduling for rescheduling scale.** Centralised or hub-coordinated schedulers that re-solve a global placement problem on every migration event do not scale gracefully as cluster count grows. Independent lines of work address this from different angles: consensus-based MARL for decentralised, energy-aware serverless placement, and hybrid architectures that pair neighbourhood-scoped local agents with a GNN-informed global orchestrator to retain system-wide coherence without full centralisation. The common thread is bounding the scope of any single rescheduling decision (by neighbourhood, by agent locality, or by negotiation depth) rather than treating every change as a global re-solve.
- **Coupling network DTs to pre-emptive traffic engineering.** A complementary pattern gives the network layer a query able, request-level energy/QoS metric and couples it to a predictive model that forecasts hotspots and congestion ahead of the event, so that SDN-based path adaptation acts pre-emptively rather than only reacting to an already made placement decision. This is an active research area in its own right, as a recent

© AIOTI. All rights reserved.

survey catalogues the breadth of network DT approaches for networking, telecommunications, and traffic engineering with specific low-cost, open-tooling instantiations for the IoT-edge-cloud continuum now emerging as one example within that space. Current architectures in this space typically model and steer traffic in general, without any notion of the LLM-specific serving cost that generated it, which remains an open integration point.

Open problems that follow directly from the patterns above include:

- Predictive cost models for LLM serving that generalise energy and latency estimates to *unseen* HW from public metadata alone, avoiding per-device profiling as edge fleets grow more heterogeneous
- Formal stability and convergence guarantee once learning-based recommendation, autonomous multi-agent negotiation, and deterministic enforcement are combined in the same control loop, a concern that spans federated orchestrators, MARL-based schedulers, and hybrid local/global architectures alike
- Standardised representations for sovereignty and energy constraints that would let these architecturally different orchestrators interoperate across projects and vendors, rather than each solving cross-layer awareness in a proprietary way
- Systematic, cross-architecture benchmarking considering the key approaches cited are rarely evaluated against one another under matched conditions, and hence, it remains difficult to say which architectural pattern dominates for a given deployment profile (fleet heterogeneity, mobility, connectivity volatility) rather than only against a default Kubernetes scheduler baseline.

© AIOTI. All rights reserved.

6 Implications for IoT Devices and Sensing and Actuating Systems

The IoT devices and sensing and actuating systems constitute the physical interface of Generative AIoT, providing the real-world observations, contextual information and mechanisms through which generative intelligence can interact with and influence physical environments. Sensors capture continuous and heterogeneous data about environmental conditions, infrastructure, processes and human activities, while actuators translate decisions and generated instructions into physical operations. Their relevance extends beyond data acquisition and execution: they provide the grounding mechanisms that connect generative models with real-world conditions, allowing AI-generated insights and actions to be informed by actual observations rather than relying exclusively on learned patterns or static knowledge.

The integration of Generative AI transforms the role of these devices from passive endpoints operating according to predefined rules into active components of distributed intelligence architectures. Through increasingly capable, embedded intelligence, IoT devices can support local data interpretation, contextual filtering, anomaly detection and event identification, while contributing relevant information to higher-level reasoning processes. Conversely, generative and agentic AI capabilities can enable more adaptive interactions with sensing and actuating systems, supporting context-aware operational recommendations, dynamic configuration and intelligent control. This evolution requires bidirectional coordination between computational intelligence and physical devices, where sensing provides continuous feedback to AI systems and actuation is governed by validated decisions, operational constraints and safety mechanisms. Consequently, IoT devices and sensing and actuating systems are essential to establishing the connection between generative intelligence and physical reality, enabling GenAIoT to evolve towards more autonomous, context-aware and responsive cyber-physical environments. Moving generative capabilities into physical environments introduces risks, as the new threat landscape evolves with the introduction of new vulnerabilities such as adversarial attacks on edge inference, the dangers of model hallucination in mission-critical IoT, and the physical safety implications of autonomous generative control.

At the same time, the integration of generative and agentic AI changes the role of IoT devices beyond the traditional model of sensing, transmitting and actuating according to a set of predefined Service Level Agreements (SLAs) or rules. Future IoT systems are likely to operate as distributed intelligence systems in which different computational layers assume different responsibilities. Highly constrained sensing devices can perform continuous local intelligence, edge systems can aggregate and interpret the insights, generative models can reason over physical and external context, data spaces can provide governed access to distributed information and secure systems should make the final checks before any critical physical action is carried out in the form of actuation.

The resulting architecture should therefore not be interpreted as placing a large generative model on every connected device. Instead, GenAIoT is likely to emerge as a hierarchical perception, reasoning, action continuum, ranging from TinyML at the physical sensing layer to more capable generative models at the edge and cloud, with intelligence distributed according to latency, energy, privacy, connectivity, task complexity and safety requirements.

6.1 Engineering implications and considerations

Deploying GenAI on IoT devices for sensing and actuation changes nearly every layer of the system stack, from silicon to security and safety certification.

Compute and HW constraints

Most IoT devices are microcontrollers with kB to MB of RAM. Even heavily quantized models need GB of memory. Moreover, real-time generative inference significantly increases Bills of

© AIOTI. All rights reserved.

Materials (BOM) costs because it demands specialized HW, e.g., NPU, DSP, advanced Single Instruction, Multiple Data (SIMD), rather than standard microcontrollers (e.g., M3, M4).

Generative models rely on massive matrix multiplications and standard, low-cost MCUs used in IoT devices lack parallel execution units needed for this, forcing a step up to silicon with dedicated Multiply-Accumulate (MAC) arrays. Also, these models are highly memory-bandwidth constrained. While classic sensor fusion can fit within a cheap MCU's internal Static Random-Access Memory (SRAM), generative models require external Pseudo-Static RAM (PSRAM), Low-Power Double Data Rate (LPDDR), or high-density NOR/NAND Flash to store weights, adding multiple discrete components to the IoT board.

Energy

Generative models require far more energy to reach one decision than classical estimation or classification methods. While classical filters and classifiers are small, run in one pass, and keep everything in on-chip memory, so they cost nanojoules to microjoules, the generative models are larger and must run the whole network many times. Each run streams hundreds of MB from DRAM, pushing the cost to joules.

On energy-constrained IoT devices, that single cost is higher than everything else the devices does (i.e., radio, sensing, sleep). There are also secondary costs like long active time and peak current limits. Seconds of inference, versus microseconds, keep the processor, voltage regulators and clocks out of deep sleep. This undermines the aggressive duty-cycling that IoT battery-based devices depend on. Moreover, small batteries and harvesting capacitors often cannot supply the peak currents accelerators draw.

Sensing pipeline

GenAI models, particularly multimodal ones, expect complex formats (e.g., tokenized streams, image patches, audio spectrograms) instead of simple scalars (e.g., raw numbers from a temperature sensor) used in classical pipeline, adding preprocessing steps overhead before the AI even starts. Pipelines could offload this work to dedicated HW and compress the model to fit. For example, NPUs and edge-AI SoCs execute vectorized MAC operations in HW, removing tensor math from the CPU at a fraction of the power cost, and Direct Memory Access stream raw sensor data directly to a DSP, bypassing the CPU. Also, out-of-distribution or noisy sensor input increases hallucination risk, which in the case of actuation can cause mechanical failures, physical damage or safety hazards compared to an annoying chat response. One option is to place a lightweight anomaly-detection or out-of-distribution detection model in front of the GenAI model to monitor incoming sensor data and identify unexpected, corrupted or low-confidence inputs before they are used for reasoning.

Deterministic guardrails and safety interlocks should be in place, while confidence scoring and thresholding could automatically default the system to a safe state or fall back to classical, deterministic control loops.

Multimodal fusion (e.g., vision, audio) requires synchronized sampling, timestamp alignment, and sometimes heavy buffering, which is very hard on simple IoT device operating systems like RTOS or bare metal.

Safety

Generative outputs that directly affect the physical state is the highest risk category in the design space. GenAI outputs are probabilistic, therefore identical sensor inputs can produce different interpretations or commands across inferences. This rules out the direct control of actuators (e.g., valves, motors, robotic arms) without a deterministic layer between the HW and the model.

© AIOTI. All rights reserved.

GenAI failure modes are less countable than conventional SW faults, therefore fallback behaviour must be designed, tested and verified explicitly. The opaque generative outputs conflict with the traceability requirements in regulated vertical domains (energy, automotive) thus the deployments must log inputs, model version, outputs, confidence scores, and supervisory-layer decisions for every actuation event, so that each command can be reconstructed and justified.

Connectivity and architecture patterns

Where the GenAI model runs determine what it can safely control. There are three deployment patterns:

- ***Fully local/on-prem***: inference runs on-device or on a gateway. This gives deterministic latency, keeps sensor data private, and removes connectivity as a failure point, but the model size and the reasoning capability are constrained by the device, available memory and computing capabilities;
- ***Cloud-only***: full-scale models provide maximum interpretive capability, but round-trip latency, bandwidth cost and connectivity availability makes the control path non-deterministic;
- ***Hybrid***: a local model handles routine sensing and actuation. Low-confidence and ambiguous cases are escalated to a cloud model, while local deterministic controllers hold a safe state until a response is received from the cloud or a timeout. The escalation criteria must be explicitly defined.

The first is the only architectural pattern suitable for closed-loop control, while the second can be used for non-time-critical functions.

In a GenAI pipeline, the sensor data is a de facto prompt, so the physical environment becomes an injection vector. Adversarial physical inputs (e.g., visual or audio manipulation of camera or microphone sensors, data spoofing of industrial sensors), can steer the model toward attacker-chosen interpretation or actions. This differs from the classical sensor spoofing as the sensor data can be valid and within the bounds yet still manipulate the output of the model. When actuation depends on that output, the attack gives the attacker control over physical behaviour. IoT field end devices and gateways are physically accessible. Without HW protection, attackers can extract model weights or temper with them.

Over the Air (OTA) model updates must meet FW-grade security: cryptographic signing, on-device signature verification, versioning, and anti-rollback protection. Model payloads are much larger than typical FW images, which increases bandwidth cost, flash partition requirements, and update duration on constrained connectivity links. In this case, delta updates, compressed weights formats, and staged fleet rollouts become necessary. As a model update might change actuation behaviour, each release requires behavioural regression testing against the safety bounds before deployment.

Tooling

Embedded toolchains, such as Zephyr RTOS and FreeRTOS, provide compilation and scheduling but no native GenAI runtime. Deployment usually depends on a separate interface layer like LiteRT [78] for microcontrollers or vendor SDK like ST STM32Cube AI Studio [79], NXP eIQ [80], which remain less mature than the cloud ML stacks (i.e., offer narrower operator coverage, debugging support, and ecosystem maturity).

Deployment considerations

Before the engineering implications, several decisions should be made before and during the deployment of GenAI for IoT sensing and actuation.

© AIOTI. All rights reserved.

Should you use GenAI? Is the task actually generative or classificatory? Many IoT use-cases are better served by small discriminative models or classic sensing processing. GenAI adds value mainly when flexible interpretation is needed not just pattern matching.

What's the cost of a wrong answer? If an incorrect actuation in a critical national infrastructure could cause physical harm, financial loss, or safety violations, that changes the risk computation and pushes toward more conservative modelling and architectures.

Design time considerations: define a fails-safe state up front, right-size the model to task and device rather than defaulting to the largest, capable model.

Validation and testing: out-of-distribution and adversarial inputs matter more than in typical ML deployments, since sensor conditions in the field diverge from training data in ways that are hard to anticipate. In strongly regulated sectors, demonstrating safety and compliance for an IoT system without deterministic branch coverage is an open problem.

It is important to both monitor for sensor drift during the lifetime of the deployment, as sensors and IoT devices HW ages and varies across units, and implement field telemetry and periodic revalidation. It is also important to threat-model the sensor-to-model path specifically and to secure the model update pipeline the same way FUOTA is secured.

On-device GenAI often means higher BOM cost, more complex OTA infrastructure, and ongoing model maintenance, worth comparing against classical approaches before committing.

Open research questions

Limits of Model Compression and Architecture Selection: Identifying the mathematical limits of extreme quantization, structured pruning, and knowledge distillation before generative reasoning, coherence, and accuracy degrade unacceptably. Closely tied to this is the challenge of determining the optimal model architecture for edge deployment, specifically evaluating whether linear-time frameworks such as state-space models and recurrent configurations should supersede traditional transformer architectures to eliminate volatile memory bottlenecks.

Continuous Adaptation and Federated Optimization at Scale: Enabling continuous on-device adaptation requires finding mechanisms that allow models to learn from streaming local data in dynamic environments without suffering from catastrophic forgetting or behavioural drift. Furthermore, scaling decentralized training and federated learning across generative models remains an acute hurdle, as exchanging parameter updates for multi-billion-parameter networks quickly overwhelms the constrained bandwidth and energy budgets of local IoT communication links.

Safety Verification, Bounded Actuation, and Auditability: Integrating probabilistic generative outputs into safety-critical control loops exposes a major gap in formal certification, as classical embedded systems rely on deterministic execution and branch coverage that cannot evaluate stochastic models. Rigorous mathematical methods should be used to bound probabilistic actions and provide explainable auditability when complex physical control decisions originate from opaque, emergent neural processes rather than inspectable conditional rules.

Pipeline Security and Resilient Real-World Benchmarking: Distributing model updates to millions of edge nodes introduces severe cybersecurity vulnerabilities, requiring robust frameworks that can authenticate, cryptographically secure, and verify deployment pipelines without the delivery infrastructure itself becoming an attack vector. Finally, the research community lacks standardized benchmark suites capable of evaluating GenAI under real-world IoT degradation, such as packet loss, sensory noise, physical timing jitter, and resource exhaustion, rather than sterile laboratory conditions.

© AIOTI. All rights reserved.

6.2 From task-specific sensing to reusable sensor foundation models

6.2.1 Data Spaces as a Sovereign Context Layer for GenAIoT

GenAI are AI models capable of generating new outputs such as text, images, code, plans, architectures, explanations, predictive actions or others based on learned patterns and supplied context. GenAIoT is the integration of GenAI with IoT devices, sensors, actuators, Edge Computing and connected cyber physical systems (CPSs). Rather than IoT devices only collecting data or executing predefined ML tasks, GenAIoT enables systems to interpret context, reason over observations, generate explanations, architectural plans and potentially influence physical action. The Dataspace is a federated environment in which multiple organisations can share and access data while retaining control over how their data are accessed and used. A dataspace typically provides mechanism for participant identity, discovery, interoperability, access and usage policies, trust, provenance and governance. It doesn't necessarily require all data to be placed in one central database [81].

With the increasing integration of generative and agentic AI into IoT systems, devices may need information beyond what is available from their own sensors, nearby edge infrastructure, or within the same operational domain. For example, in a smart heritage conservation environment, sensors installed around a historic stone moment, castle or archaeological site may detect changes in temperature, humidity, moisture, vibration or surface conditions. To understand whether these changes indicate deterioration, the AI system could combine local sensor readings with previous conservation records, material information, weather data, the site DT and observations from similar heritage sites. To understand whether these changes indicate deterioration, the AI system could combine local sensor readings with conservation records, material information, weather data, DT information and observations from similar heritage sites. This is an example of multimodal and contextual data fusion, where different types of information are brought together to support a more reliable assessment. A data space can support this process by providing governed access to these distributed data sources across different organisations, while preserving information about provenance, access rights and usage conditions. The resulting richer context can help AI systems assess the condition and conservation needs of a heritage asset more reliably than relying on a single sensor source alone. Bringing these distributed data sources together can provide AI systems with richer view surrounding condition and conservation needs of heritage asset. AIOTI has already identified the integration of IoT, Edge Computing, CPSs and DT as an important part of data space architecture.

A frontier future research direction is therefore to investigate data spaces as a sovereign context layer for GenAIoT. Rather than trusting data spaces primarily as repositories from which datasets are clustered for model training, they could provide controlled runtime access during operation to distributed data and contextual information about the current conditions of the physical objects and their surrounding environment. An AI agent running at the edge could use data-space catalogues, metadata and semantic descriptions to identify relevant data and services. The search could be guided by the current task, the identity of the physical asset and the sensor observations already available. The agent could then access useful information from trusted sources such as live sensor data, DT information, maintenance documents, operating procedures and other AI services. It would then only retrieve the information needed for the task it is currently trying to solve. Accordingly, the AI system would not need to collect or store all potentially relevant data in advance; instead, it could retrieve additional context dynamically when required by the current task. It can ask for extra information only when it needs it.

Such framework would naturally lead to policy aware and provenance aware federated RAG [82], because the retrieved data and information may originate from independent organisations and remain subject to different access rules, usage conditions, provenance, quality levels and semantic descriptions. In normal RAG system, the information often comes from one database or one organisation. Conventional RAG typically presumes that once

© AIOTI. All rights reserved.

information is obtained then it can be integrated into the context of a model. However, in data spaces, information may come from different organisations. Each source may have different rules, level of quality, and interpretations of the same situation. The AI system therefore needs to understand where the information comes from, how reliable it is and whether different sources agree with each other.

For example, in a smart cultural heritage conservation environment a museum may provide conservation records, a local authority may provide weather and pollution data, a university or a sensor company may provide sensor measurement, a DT may provide model of the current condition of the site and a technology provider may provide an AI analysis service. Each of these sources may have different rules about how their data can be used. They may also have different levels of accuracy, format, meaning, different update time, or even give conflicting information. So overall, the AI system should understand who provided the data, how reliable it is, how recent it is, what rules apply to its use, and whether it agrees with other available information. Hence, future GenAIoT systems should therefore investigate retrieval mechanisms that consider not only semantic information because it is relevant, but also according to where it came from, when it was collected etc and whether the AI is authorised to use it for the intended purpose.

AI agents are increasingly being considered as active participants in data spaces. Organisations and initiates such as International Data Spaces Association (IDSA) [83], NTT DATA [84], the Data Space Support Centre (DSSC) [85], Gaia-X [86] and the innovation data spaces (i-Space) label granted by the Big Data Value Association (BDVA) [87] are exploring how agents can discover data and services, interpret machine readable SLAs, negotiate terms and invoke services on behalf of an accountable organisation using delegated identity and auditable interactions. For GenAIoT this raise a new set of research questions.

Important research challenge appears when AI agents move from simply accessing data to controlling physical systems. This position paper uses the term action sovereignty to describe the need to separate permission to access and use data from permission to carry out physical actuation in real world. For example, an AI agent may be allowed to read sensor data and recommend an action, but it should not automatically be allowed to control an actuator or change a device setting. Some low-risk actions may be allowed with redefined threshold and limitations while safety critical actions may require human approval. Standards policies [88] already exists to provide machine readable representations of permissions, prohibitions, duties and constraints over assets. A future research challenge is to investigate how these mechanisms can be extended to govern AI initiated physical actions. Such policies would need to specify which AI agents are authorised to recommend or execute an action, under what operating conditions, within which safety limits, and when human approval is required.

In this context, it is increasingly important to have sensor to actuator traceability, meaning that we should be able to trace a physical action back to the information and decisions that led to it. For example, if an AI system automatically activates a dehumidifier to protect a historic stone wall, sensor to actuator lineage would provide a record of why that action happened. It could show the humidity and moisture measurements that triggered the decision, the weather or conservation information retrieved from the data space, the AI model that analysed the information, the safety rules that were checked and the final command send to the dehumidifier. This makes the AI system's actions easier to understand, verify and investigate. Data spaces and standards such as W3C PROV [89] already support data provenance and traceability. Sensor to actuator traceability would extend this idea by tracing the complete path from sensing and AI reasoning to the final physical action. This can help with auditing, explaining decisions, investigating problems, and checking whether an autonomous GenAIoT system behaves correctly.

The EC's 2026 Dataspaces for manufacturing initiative [90] explicitly calls for collecting large scale data for real industrial environments to train or fine tune GenAI models, providing

© AIOTI. All rights reserved.

authorised AI developers with data access, and for developing interoperability and governance frameworks, also explicitly including AI powered robotics. A complementary research question for GenAIoT is therefore how the same sovereign data space infrastructure could support runtime reasoning, data access, policy enforcement, active sensing and controlled physical actuation and action. Following which how edge agents find trustworthy information, handle conflicting data, decide when more sensing data is needed, respect access rules and keep traceability from the original sensor data to final action. The goal is to enable distributed AI systems to make informed physical decision without requiring all data to be centrally collected.

6.2.2 TinyML and Edge Intelligence as the Local Foundation of GenAIoT

TinyML [91] is the process of optimising ML models to run on resource constrained devices like microcontroller. These models are typically small, efficient, and capable of performing tasks like image classification, speech recognition, and sensor data analysis.

The emergence of GenAIoT does not imply that generative foundation models will execute directly on every sensing device. Many IoT endpoints operate under severe constraints in memory, processing capabilities, energy availability, and communication bandwidth. TinyML therefore remains an important component of the future GenAIoT architecture because it allows light weight ML models to execute directly on microcontroller devices close to the physical process.

TinyML can be viewed as the first intelligence layer of the hierarchical perception, reasoning and action continuum. Rather than forwarding all physical observations to a large generative model, TinyML models can continuously process sensors data locally and identify events that require higher level interpretation. Normal operating conditions may be handled completely on the device, while anomalies, novel observations or low confidence use-cases and scenarios can trigger escalation to a more capable edge node. This creates a hierarchical intelligence model in which TinyML performs fast and energy efficient physical perception, while generative reasoning is activated selectively, when broader semantic or contextual understanding is required.

Edge Computing then provides the intermediate layer between constrained devices and more computationally intensive cloud or foundation model services. Anomaly detection may execute on a microcontroller, multimodal integration and generative reasoning may occur on an edge gateway, governed contextual information may be retrieved from data spaces and only complex or long reasoning may be pass on to the larger model deployed either cloud resources or much powerful edge compute. This should adapt dynamically w.r.t. latency requirements, energy availability, communication conditions, privacy constraints, task complexity and safety criticality.

Another promising research direction is uncertainty triggered escalation. TinyML and specialist edge models could handle routine activities locally, while low confidence or safety relevant events are passed to a more capable generative reasoning layer. The resulting system avoids the energy and latency cost of invoking a large model continuously while still benefiting from generative reasoning when it is most valuable. This can work the other way as well where a generative edge agent could request that a TinyML device increases its sampling rate or collects high resolution data sets when additional evidence is needed.

TinyML, edge computing and data spaces should therefore be treated as complementary rather than competing architectural layers. TinyML anchors intelligence at the point of sensing, Edge Computing provides local multimodal integration and generative reasoning, and federated data spaces extend the available context beyond the boundaries of an individual device or organisation. Future GenAIoT systems may dynamically move between these layers according to uncertainty, latency, energy, privacy and governance constraints, creating an

© AIOTI. All rights reserved.

adaptive continuum from physical sensing to sovereign contextual reasoning and controlled actuation. The research challenge is therefore not simply to determine whether a generative model should run at the edge or at the cloud, but how the complete AI workload should be partitioned across heterogeneous levels of the IoT-Edge-Cloud continuum. Key Future research Questions:

- How can constrained devices provide useful predictions, uncertainty and trigger escalation?
- How can data spaces be used as a runtime, sovereign context later for GenAIoT?
- How can agent authority over physical actions be defined, enforced and delegated?
- How can end-to-end traceability be maintained from sensing to final actuation?
- How can agent decide what additional observation are needed from whom and under which conditions?
- How can perception, reasoning, data access, policy enforcement and actuation be jointly optimised across the IoT-Edge-Cloud continuum?
- How can we have policy aware data retrieval of data considering relevance, provenance, quality, uncertainty and usage rights?

6.3 Foundation models

Most deployed intelligent IoT systems still use models built for a particular combination of sensor, task and operating environment. Such systems can perform well within their development domain but often generalise poorly across device types, locations, users and changing conditions.

Foundation models offer a different approach. Large-scale pretraining, frequently using self-supervised objectives, can produce reusable representations that are subsequently adapted to classification, forecasting, anomaly detection, imputation and other downstream tasks. The field is already developing foundation-model approaches specifically for time-series analysis, with architectures and pretraining strategies intended to transfer across datasets and tasks [92].

This development is particularly significant for IoT because much physical-world information is not naturally represented as language or static imagery. It consists of irregular, asynchronous and multi rate streams generated by accelerometers, physiological monitors, environmental sensors, industrial equipment and communications infrastructure. Time-series foundation models seek to learn recurring temporal structure from such data, potentially reducing the requirement to train a separate model from scratch for each deployment [93]. However, generalisation across sensor systems remains more difficult than transfer across text collections. Sensor values depend on sampling frequency, position, calibration, units, device characteristics and operating context. Similar numerical sequences can represent different physical phenomena, while the same phenomenon can produce substantially different signals across devices and users. Consequently, a sensor foundation model must represent not only signal patterns but also sensor identity, physical semantics, temporal scale and contextual metadata.

The current literature therefore suggests that foundation models should not be understood as universal substitutes for domain-specific signal processing. A more realistic architecture combines reusable pretrained representations with local calibration, domain adaptation and explicit physical knowledge.

6.4 Multimodal grounding in the physical environment

IoT environments are intrinsically multimodal. An industrial system may combine vibration, sound, temperature, images and machine-state data; a wearable system may combine movement, heart rate, temperature, location and user reports; and an autonomous system may integrate cameras, ranging sensors, proprioception and maps.

© AIOTI. All rights reserved.

Recent multimodal learning research distinguishes several levels at which such information can be combined: fusion of raw or aligned data, fusion of learned features, and fusion of model outputs. Methods include contrastive learning, attention-based fusion, generative modelling and LLM-mediated integration [94]. These techniques provide the basis for models that can relate sensor observations to textual concepts, images, plans or explanations.

For GenAIoT, however, multimodal fusion is not simply a matter of adding additional inputs. The **central problem is physical and temporal alignment**. Modalities may have different sampling rates, delays, fields of view, uncertainty characteristics and failure modes. The model must determine whether observations refer to the same event, whether one modality is stale, and whether apparent disagreement represents sensor failure or a meaningful feature of the environment.

The state of the art is moving towards shared representation spaces in which physical observations can be connected with language and domain knowledge. This allows a system to generate explanations, answer questions about current conditions, or retrieve relevant procedures. Nevertheless, there remains a significant modality gap between the well-represented text–image domain and the heterogeneous signal streams typical of IoT. Multimodal alignment, data quality and computational cost remain open challenges even in mainstream multimodal learning [94].

A credible GenAIoT system must therefore retain the provenance of generated interpretations. It should be possible to identify which observations, sensors and time windows support a conclusion. Without this connection, linguistic fluency can conceal weak or contradictory physical evidence.

6.5 Intrusion Detection and Runtime Protection for Edge AI Inference

Among the risks introduced by deploying generative intelligence in physical IoT environments, adversarial manipulation of edge inference is particularly significant. Adversarial attacks can exploit vulnerabilities in deployed models by introducing carefully crafted inputs that alter model behaviour [95]-[97], while recent work has shown that similar manipulations can also influence the runtime behaviour of multimodal generative models [98]. As intelligence is decentralised towards devices located close to sensors and actuators, compromised inputs or inference processes can directly affect local decisions and, in CPS applications, propagate into physical actions. At the same time, decentralisation expands the attack surface across heterogeneous, resource-constrained, and often physically accessible devices that may operate with intermittent connectivity and limited support from centralised security infrastructure [43][99]. Preventive protection mechanisms alone are therefore insufficient, GenAIoT systems also require continuous runtime monitoring capable of identifying suspicious inputs, anomalous model behaviour, and malicious system activity as they occur. In this context, Intrusion Detection Systems (IDS) provide a key layer of dynamic protection for edge inference, complemented by runtime protection mechanisms that can isolate, constrain, or validate critical inference operations. Such mechanisms must provide effective security while respecting the strict latency, computational, memory, and energy constraints inherent to edge devices.

The Threat Landscape

The threat landscape for edge AI inference spans multiple layers of the system, from the data entering the model to the execution environment where inference occurs. At the input level, evasion attacks manipulate sensor observations or other inference inputs to induce incorrect model behaviour, often through perturbations designed to remain difficult to distinguish from legitimate data [95][97]. In generative and multimodal models, this threat extends beyond conventional misclassification: adversarial inputs can influence generated content, override intended behaviour, or manipulate the contextual interpretation of other inputs at runtime [98]. Poisoning and backdoor attacks originate earlier in the model lifecycle, during training, fine-

© AIOTI. All rights reserved.

tuning, or model updates, but their effects can remain dormant until inference and be activated by specific inputs or conditions [95][100]. Other threats target the model itself, including model extraction through repeated queries [101], while side-channel attacks can exploit execution characteristics such as memory access patterns, timing, power consumption, or electromagnetic emissions to infer sensitive information about the deployed model [102].

These risks become particularly significant as intelligence is moved closer to sensing and actuation. An attack may affect not only an inference result but also the context inferred from multiple sensor modalities, the content generated by the model, or the actions subsequently selected by an autonomous system. Edge nodes also operate under conditions that differ substantially from protected cloud infrastructure, since they are heterogeneous, resource-constrained, widely distributed, and may be deployed in physically accessible or intermittently connected environments. As a result, the attack surface extends across the data, model, runtime, HW, and communication layers, while the consequences of successful manipulation can propagate directly into cyber-physical behaviour.

Detection and Runtime Protection Methodologies

Protecting edge inference requires mechanisms capable of detecting abnormal behaviour at different stages of the inference pipeline. At the input level, anomaly-based approaches can identify statistical outliers, distribution shifts, or patterns inconsistent with expected observations [95]. Depending on the available resources, these mechanisms may range from lightweight statistical tests to learned anomaly detectors. Prediction confidence or uncertainty can provide additional evidence, although confidence-based monitoring alone cannot reliably distinguish all adversarial inputs. More sophisticated approaches inspect intermediate model representations, for example, class-conditional feature distributions and Mahalanobis-distance-based scores have been proposed to identify both out-of-distribution and adversarial samples [103]. However, experience with adversarial-example detection has also shown that apparently effective detectors can be circumvented by adaptive attackers, reinforcing the need to avoid reliance on a single detection mechanism [104].

For GenAIoT systems, runtime monitoring should extend beyond the input itself. Open-ended generative outputs, multimodal reasoning, and locally generated decisions may require consistency checks between sensor evidence, model outputs, system state, and the actions proposed to downstream components. Detection mechanisms can therefore be complemented by runtime protection that constrains how inference results are consumed or executed. At the platform level, monitoring execution behaviour and system-level indicators can provide additional evidence of abnormal operation. Trusted Execution Environments (TEEs) provide a complementary protection mechanism by isolating sensitive model code, parameters, and inference data from untrusted SW or infrastructure components, and have already been investigated for secure edge inference [105]. Rather than replacing IDS, these mechanisms contribute to a defence-in-depth architecture in which input monitoring, model-level detection, runtime validation, and trusted execution protect different parts of the inference pipeline.

Operational Challenges

Deploying the mechanisms at the edge introduces a fundamental resource trade-off: the security mechanisms protecting inference compete for the same computation, memory, energy, and latency budgets as the inference workload itself. Edge intelligence already operates under substantially tighter resource constraints than cloud-based AI [43], and recent work on energy-efficient edge AI highlights the importance of jointly considering computation, communication, memory, and energy costs in on-device inference [99]. Generative models make this competition even more demanding. Runtime protection must therefore provide sufficient security coverage without compromising the responsiveness, availability, or energy efficiency of the application it is intended to protect.

© AIOTI. All rights reserved.

This requirement makes lightweight and HW-aware security mechanisms particularly important. The compression and optimisation principles discussed can also inform the design of compact detection models, although reducing model size alone does not guarantee lower end-to-end security overhead on every HW platform. Memory constraints may prevent devices from storing large attack-signature databases or long histories of runtime observations, favouring streaming analysis, compact state representations, and incremental processing. Real-time applications additionally impose strict limits on detection latency, particularly where inference outputs directly influence actuators or safety-critical decisions. Finally, intermittent connectivity means that protection cannot depend continuously on cloud-based analysis. Local detection and protection must remain functional during disconnection, while more computationally demanding correlation, learning, or forensic analysis can be distributed towards more capable edge or cloud nodes when connectivity and latency requirements permit.

Emerging Research Directions

A first research direction is the development of lightweight, HW-aware monitoring co-designed with the edge inference workload. The objective should not be detection accuracy alone, but the joint optimisation of security effectiveness, inference latency, memory footprint, computational overhead, and energy consumption [99]. Collaborative detection across the IoT-edge-cloud continuum is another promising direction, although federated and distributed learning mechanisms must themselves be protected against malicious participants and poisoned updates [106][107].

A further challenge is adaptation. Static detectors may fail as attacks, operating conditions, sensor distributions, and deployed models evolve, particularly against adaptive adversaries [104]. Future runtime protection will therefore require cross-layer approaches combining input validation, model and system monitoring, trusted execution, output verification, and application-level safety constraints. In safety-critical GenAIoT systems, detection should also trigger an appropriate response, such as re-evaluation, execution through an alternative model, escalation to a more capable edge or cloud node, or transition to a predefined safe operating mode.

Evaluation requirements and open challenges

Evaluation of intrusion detection and runtime protection for edge inference must go beyond conventional detection accuracy. Detection rate, false-positive and false-negative rates, and robustness should be assessed together with detection latency, computational overhead, memory footprint, energy consumption, and the impact on the protected inference workload [94][99]. Detection latency is particularly critical in CPSs, where a correct but delayed response may occur after an unsafe action has already been executed.

Evaluation should also consider generalisation beyond the attacks and datasets used during detector development, since adaptive adversaries can circumvent mechanisms that appear effective against known attacks [104]. Representative benchmarks should therefore include benign operating variations, unseen or adaptive attacks, heterogeneous edge HW, and realistic sensing and communication conditions. For GenAIoT systems, evaluation should additionally consider the impact of attacks on generated outputs, multimodal reasoning, and downstream actions [98], ultimately assessing security and operational performance as an integrated property of the edge AI architecture.

6.6 GenAI as a semantic layer over sensing

A central capability of GenAI is to transform low-level observations into higher-level semantic representations. A conventional anomaly detector may indicate that a vibration pattern is unusual; a generative system may relate the anomaly to maintenance records, operating

© AIOTI. All rights reserved.

conditions and technical documentation, then generate a possible explanation and an inspection procedure. This semantic layer can support several functions:

- Contextual interpretation of sensor events;
- Natural-language summarisation of complex system states;
- Retrieval and synthesis of relevant operational knowledge;
- Generation of labels and annotations;
- Reconstruction or imputation of missing observations;
- Creation of synthetic data for rare or hazardous scenarios;
- Explanation of predictions made by specialised models;
- Interaction between human operators and distributed sensing systems.

The important conceptual shift is from treating each sensor model as an isolated predictor to treating the IoT installation as a source of evidence for a broader reasoning system.

Surveys of on-device AI and edge–cloud generative services show that this capability is being developed through combinations of compact local models, model compression, edge inference and access to more capable remote foundation models [108][109]. The likely architecture is not a single foundation model operating independently on every sensor node. It is a hierarchy in which embedded models perform preprocessing and routine inference, edge systems integrate local context, and larger models provide more computationally intensive reasoning and knowledge access.

This arrangement also counters a common misconception: placing GenAI within an IoT architecture does not necessarily mean executing the complete model on the sensing device. The relevant question is which representation, inference and reasoning functions should be placed at each level.

6.7 Edge–cloud collaboration and selective escalation

The computational requirements of foundation models conflict directly with the memory, energy and thermal constraints of many IoT devices. Research on on-device AI consequently emphasises quantisation, pruning, distillation, efficient model architectures, HW acceleration and workload partitioning [108]. Research on edge–cloud generative services additionally consider model placement, caching, collaborative inference, fine-tuning and lifecycle management [109].

A state-of-the-art GenAIoT architecture is therefore likely to be **asymmetric and adaptive**:

- Lightweight models continuously process local observations;
- Edge nodes combine data across devices and maintain recent context;
- Larger foundation models are invoked for unfamiliar, complex or high-value cases;
- Outputs are distilled or adapted back into smaller operational models;
- The location of computation changes according to latency, connectivity, privacy and available resources.

This is more sophisticated than a fixed division between edge inference and cloud training. Recent industrial-IoT thinking frames foundation-model deployment across sensing, computing, connectivity and long-term model evolution, recognising that intelligence must be managed across both the processing pipeline and system lifecycle [110].

An especially promising pattern is **uncertainty-triggered escalation**. A local specialist model handles common cases efficiently, while uncertain or open-set observations are referred to a more capable model. This can reduce communication and energy use without assuming that a constrained local model can manage every condition. Field evidence adds a caution: in a deployed greenhouse controller, agreement among sampled outputs of a local language model flagged its indecision but missed its systematic errors, half of which were unanimous for

© AIOTI. All rights reserved.

one model [142]. Escalation triggers should therefore combine model uncertainty with independent consistency checks, such as a deterministic reference controller (see Section 5.2).

The architectural design problem is therefore not merely model compression, rather the dynamic allocation of sensing, representation, reasoning and learning functions across heterogeneous infrastructure.

6.8 From generated interpretations to generated actions

The most consequential development is the transition from models that generate information to models that generate actions. Vision–language–action models extend vision–language systems by representing actions as model outputs. Current research distinguishes between models that produce low-level control actions and models that perform high-level planning by decomposing long-horizon tasks into subtasks [5]. This distinction is directly relevant to IoT actuating systems. GenAI can participate at several levels:

- ***Observation and description:*** describing the state of a device or environment;
- ***Decision support:*** proposing a diagnosis, intervention or maintenance action;
- ***Planning:*** producing an ordered sequence of actions or allocating tasks among devices;
- ***Parameter or policy generation:*** configuring a controller or adapting an operational policy;
- ***Direct action generation:*** producing commands executed by actuators.

These levels should not be treated as equivalent. A model that produces a natural-language recommendation can be reviewed before execution. A model generating motor or actuator commands operates within a much tighter latency and safety envelope.

The emerging embodied-AI literature shows that foundation models can connect perception, language-conditioned reasoning and action generation [5]. Yet the strongest present capabilities are generally obtained in constrained task environments, using substantial training data, simulation and task-specific adaptation. General competence in open, changing physical environments remains an aspiration rather than an established capability.

In most critical IoT applications, the more defensible pattern is therefore **generative planning with constrained execution**. A foundation model may interpret the situation and propose a plan, while a separate control or safety layer checks physical constraints, permissions and operating limits before actions are executed. Conventional estimation and control remain important because they provide predictable behaviour, high-frequency response and explicit representations of system dynamics.

6.9 World models, DTs and anticipatory action

A further direction is the convergence of generative models with world models and DTs. Rather than mapping an observation directly to an action, a world model represents how the environment may evolve under alternative actions. An agent can then simulate possible futures, compare outcomes and select a plan.

This approach is potentially transformative for IoT because distributed sensing provides the observations required to maintain a local representation of the physical system, while DTs and learned dynamics models provide a substrate for prediction and planning.

The resulting architecture may combine:

- Sensor streams that update the current estimated state;
- A DT or learned world model that predicts future trajectories;
- A foundation model that interprets goals, constraints and domain knowledge;
- A planner that evaluates candidate actions;

© AIOTI. All rights reserved.

- A verified controller that executes allowable actions;
- Feedback from subsequent sensing to revise the plan.

This moves GenAIoT beyond reactive automation towards anticipatory operation. Potential applications include adaptive manufacturing, energy management, logistics, robotics, environmental control and personalised health monitoring.

However, the value of simulated futures depends on model validity. A fluent plan generated from an inaccurate world model may be systematically unsafe. Research must therefore address not only hallucination in language outputs but also **model error in imagined physical trajectories**.

6.10 Human–system interaction

Natural-language interfaces make complex IoT systems more accessible, allowing users to ask why an event occurred, what evidence supports an alert or what action is recommended. This may substantially improve the usability of large sensor installations.

At the same time, generative explanations can create an illusion of understanding. An explanation may be coherent without accurately reflecting the mechanism that produced the underlying result. GenAIoT interfaces should therefore distinguish:

- Direct observations;
- Estimated latent states;
- Model predictions;
- Retrieved knowledge;
- Generated hypotheses;
- Recommended actions.

This distinction is especially important when the user must decide whether to intervene. Rather than merely generating persuasive prose, an effective interface should expose the evidence base, conflicting observations, relevant uncertainty and the consequences of alternative actions.

Human interaction is therefore not a downstream presentation issue. It forms part of the sensing–reasoning–action architecture: human feedback may supply missing context, reject implausible interpretations, set operational goals or authorise actions.

6.11 Research challenges and emerging directions

GenAI changes the role of IoT devices more fundamentally than simply adding a larger ML model at the edge. Conventional IoT architectures largely separate sensing, inference and action: sensors produce measurements, task-specific models classify or predict, and predefined application logic determines the response. Foundation models increasingly bring these functions together by learning transferable representations across data types and tasks, interpreting observations in context, generating semantic descriptions and plans, and, in embodied systems, producing actions.

The emerging trajectory is therefore from connected sensing, through local predictive intelligence, towards a distributed perception–reasoning–action loop. This extends the evolution described earlier in this position paper, but it also introduces important technical distinctions. Generative models may be used to reason about the physical world, to produce recommendations for the physical world, or to act within the physical world. These uses create progressively stronger requirements for grounding, latency, robustness and control assurance. Several major research priorities are identified as presented below.

© AIOTI. All rights reserved.

Physical Tokenisation and Representation: Foundation models require representations that preserve units, temporal dynamics, measurement uncertainty, and sensor identity. Standard text-like tokenisation often discards the fine-grained physical relationships and numerical properties essential for accurate physical interpretation.

Cross-Device and Cross-Context Generalisation: Pretrained architectures must reliably transfer across different HW variants, individual users, physical sites, and operational environments. Research must move beyond proving performance on narrow partitions of uniform benchmark datasets to validating true cross-domain adaptability.

Multimodal Alignment under Imperfect Conditions: Real-world systems must handle sensory disruptions without catastrophic failure. Models need the resilience to process missing data streams, asynchronous timing jitter, cross-sensor calibration offsets, and conflicting sensory evidence.

Continual and Personalised Adaptation: Because operational environments change over time, models must adapt dynamically to local shifts. This requires on-device learning strategies that avoid catastrophic forgetting, prevent erratic behavioural drift, and eliminate the need to centralize sensitive raw data.

Resource-Aware Orchestration: Systems must dynamically balance computational loads across the network topology. Intelligent scheduling engines are needed to decide in real time which tasks run locally, what minimal data to transmit, and when to offload complex reasoning to larger models.

Grounded Generation: Synthesized diagnostics, operational plans, and physical descriptions must not rely on unmoored probabilities. Every output must be verifiable, auditable, and traceable back to concrete sensor observations, established domain knowledge, and the current system state.

Separation of Reasoning and Execution: Architectures must decouple probabilistic generative reasoning from real-world actuation. Generative models should propose high-level plans that are independently validated and bounded by deterministic control layers or formal mathematical monitors before reaching physical HW.

Evaluation Beyond Model Accuracy: System performance cannot be captured by accuracy metrics alone. Holistic benchmarking must evaluate end-to-end sensing-reasoning-action loops, measuring inference latency, energy budgets, fault recovery, decision robustness, and real-world physical outcomes.

Long-Duration Real-World Validation: Current research remains overly reliant on simulated environments, static benchmarks, and short pilot demonstrations. The field urgently needs empirical validation derived from long-term, continuously operating deployments running across messy, heterogeneous physical settings.

6.12 Future Research Trends

The most significant implication of GenAI for IoT devices is not that every device will host a language model. It is that sensing systems are becoming components within distributed reasoning architectures capable of relating physical observations to semantic knowledge, goals and possible actions.

The state-of-the-art points towards hybrid systems rather than wholesale replacement of established IoT methods. Specialist models and conventional signal processing provide efficient perception; foundation models provide transfer, semantic integration and planning; edge–cloud orchestration allocates computation; and verified control mechanisms protect physical execution.

© AIOTI. All rights reserved.

The resulting research frontier lies at the boundary between **generalisation and grounding**. Foundation models bring broad reusable knowledge and flexible reasoning, but IoT systems must remain anchored in specific devices, measurements, physical constraints and operational consequences. The quality of future GenAIoT systems will depend on how effectively these two properties are combined.

© AIOTI. All rights reserved.

7 Trustworthiness

7.1 Introduction

Trustworthiness is emerging as one of the most critical requirements for the successful deployment of GenAIoT and Edge Computing systems. As intelligence moves from centralized cloud infrastructures towards distributed devices, sensors, gateways, robots, autonomous vehicles, and CPSs, the consequences of AI-generated outputs extend beyond digital environments into the physical world. In such environments, incorrect predictions, hallucinated recommendations, biased decisions, or compromised models can directly impact human safety, operational reliability, privacy, and societal trust.

The growing integration of GenAI with IoT introduces a new paradigm in which AI systems are not only interpreting data but also generating actions, decisions, SW code, synthetic data, control policies, and user interactions. While these capabilities create opportunities for increased automation, efficiency, and autonomy, they also introduce significant challenges related to ethics, accountability, transparency, security, and governance. Trust in GenAIoT systems therefore cannot be based solely on performance metrics; it must be supported by evidence demonstrating that systems are lawful, ethical, robust, and aligned with human values.

The EC's framework for Trustworthy AI identifies seven core requirements [111]:

- Human agency and oversight;
- Technical robustness and safety;
- Privacy and data governance;
- Transparency;
- Diversity, non-discrimination and fairness;
- Societal and environmental well-being;
- Accountability.

These principles remain highly relevant for GenAIoT systems, but their implementation becomes considerably more complex in highly decentralized environments where thousands or millions of edge devices continuously learn, generate, and adapt.

Furthermore, recent developments in Responsible AI (RAI) emphasize that trustworthiness is not merely a technical property but a socio-technical objective that spans the entire lifecycle of AI systems, from design and deployment to monitoring, governance, and retirement [112].

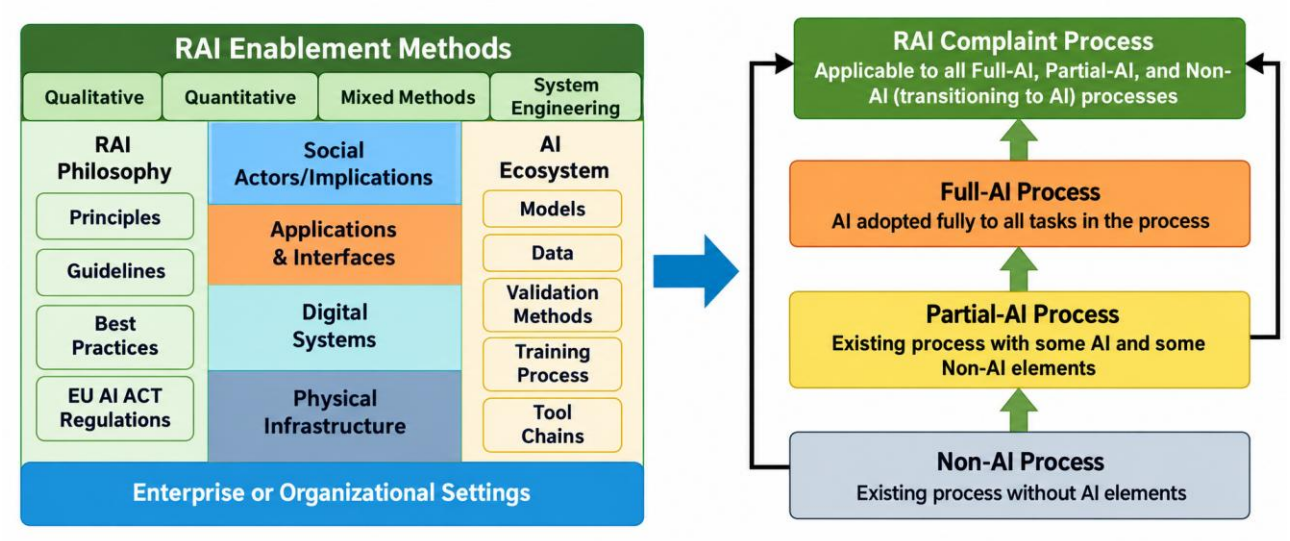


Figure 7-1: RAI Adoption/Adaptation Framework

© AIOTI. All rights reserved.

In this direction, organizations can develop and use the RAI Adoption and Adaptation framework, as illustrated in Figure 7-1 [112], at their process level to embed ethical, legal, technical, and societal considerations into AI-enabled systems from their inception, while also supporting the retrospective adaptation of existing non-AI or legacy processes to achieve RAI compliance. In the context of GenAIoT, this means developing mechanisms for continuous oversight, explainability, transparency, risk management, and stakeholder participation.

Consequently, trustworthiness should be viewed as a foundational pillar for future GenAIoT ecosystems. Without sufficient trust, citizens, enterprises, regulators, and critical infrastructure operators may be reluctant to adopt autonomous and intelligent edge systems regardless of their technical capabilities. Therefore, advancing trustworthy GenAIoT requires interdisciplinary research spanning AI, cybersecurity, human-computer interaction, ethics, law, systems engineering, and governance.

Trust is a foundational requirement for GenAIoT systems, because decision makers, edge devices, and cloud services all rely on the origin, integrity, and permitted use of the data, prompts, and generated outputs they exchange. As GenAIoT pipelines span the IoT-Edge-Cloud continuum, establishing trust and regulatory compliance between participants becomes increasingly important [113]. The following subsections examine the key trust challenges for GenAIoT systems, and outline security, privacy, and governance considerations that shape the design of trustworthy generative pipelines.

7.2 Key Challenges in Trustworthiness for GenAIoT Systems

The convergence of GenAI and IoT introduces several trustworthiness challenges that differ significantly from those encountered in traditional cloud-based AI systems. These challenges arise primarily from the decentralized nature of intelligence, limited computational resources at the edge, heterogeneity of devices, dynamic environments, and the increasing autonomy of AI-driven decision-making.

A primary challenge relates to **transparency and explainability**. Modern foundation models and LLMs often operate as "black-box" systems whose internal reasoning processes remain difficult to understand. When these models are embedded within edge environments and make decisions affecting vertical sectors, such as industrial operations, healthcare systems, transportation networks, or smart cities, stakeholders require explanations that justify system behaviour. The lack of explainability impedes trust, complicates debugging, and creates uncertainty regarding accountability when failures occur.

Another significant challenge concerns **hallucination and reliability**. Generative models can produce outputs that are syntactically convincing but factually incorrect. While hallucinations may be tolerable in low-risk applications, they become highly problematic in industrial IoT, healthcare monitoring, autonomous logistics, or energy management systems where incorrect recommendations can lead to operational disruptions or safety hazards. Ensuring factual grounding and context-aware generation therefore becomes essential.

Bias and fairness present another major concern. Training datasets often reflect existing societal, geographical, economic, and demographic biases. As generative intelligence becomes integrated into decision-support systems, such biases may propagate into resource allocation, predictive maintenance prioritization, service delivery, or automated decision-making processes. Ensuring fairness becomes especially challenging when distributed edge models are trained on geographically localized datasets that may not adequately represent broader populations.

Trustworthiness is also challenged by **accountability and governance complexity**. In decentralized architectures, responsibility for decisions may be distributed across device manufacturers, model developers, service providers, data owners, and system integrators.

© AIOTI. All rights reserved.

Determining liability becomes increasingly difficult when autonomous decisions emerge from interactions among multiple AI agents operating across heterogeneous IoT infrastructures.

Another important challenge stems from **continuous adaptation and autonomous learning**. Unlike static AI systems, future GenAIoT deployments may continuously update themselves based on local data observations. Although adaptation improves performance and personalization, it can also introduce unpredictable behaviour, model drift, and compliance risks. Establishing mechanisms for monitoring and validating continuously evolving models remains an open research problem.

Finally, trustworthiness must address broader **socio-technical impacts**. Responsible AI research has highlighted concerns regarding growing automation, concentration of power among large technology providers, workforce displacement, and digital inequalities. As intelligence becomes embedded within every connected device, ensuring equitable access, human empowerment, and societal benefit will become increasingly important dimensions of trustworthiness.

GenAIoT systems face several trust challenges as data, prompts, and generated outputs move across the continuum. Provenance is difficult to track, since it is often unclear which device or service produced a given input or output, and under what terms [114]. Following the approach of data spaces, a policy formulating the terms of use needs to be exchanged alongside the data itself, so information on how the data can be used accompanies it as it crosses IoT, edge, fog, and cloud layers [16]. Verifying participant identities is equally essential for assessing trustworthiness, so data providers, consumers, services, and devices all need to be authenticated [97]. Maintaining logs of all interactions across GenAIoT and data-space systems adds accountability and helps resolve disputes and demonstrate regulatory compliance [115]. Notably, these mechanisms need to scale without becoming a deployment bottleneck, since GenAIoT systems can span many providers, consumers, and nodes [114].

7.3 Security and Safety Considerations

Moving generative capabilities into physical environments introduces new risks for GenAIoT systems, as the threat landscape evolves with vulnerabilities such as adversarial attacks on edge inference, the dangers of model hallucination in mission-critical IoT, and the physical safety implications of autonomous generative control.

Security threats against GenAIoT systems are substantially broader than those affecting conventional IoT architectures. Adversaries may target training data, inference pipelines, communication channels, model parameters, HW accelerators, or collaborative learning mechanisms. Since edge devices often operate in untrusted or physically accessible environments, attackers may exploit device vulnerabilities to manipulate AI-generated outputs or gain unauthorized access to sensitive operational data. Below are the major security and safety concerns that involve:

- **Adversarial ML attacks:** Carefully crafted inputs can manipulate generative models into producing incorrect outputs while appearing legitimate to human observers. In autonomous transportation systems, industrial automation environments, or smart healthcare applications, such attacks could result in unsafe recommendations, erroneous actions, or operational disruptions;
- **Model poisoning and supply-chain attacks:** Distributed training and federated learning approaches enable edge devices to contribute to model improvement, but compromised devices may intentionally introduce malicious updates into collaborative models. Similarly, reliance on pre-trained foundation models sourced from external ecosystems creates risks concerning model provenance, hidden vulnerabilities, and embedded malicious functionality;

© AIOTI. All rights reserved.

- **Issue of hallucinations in safety-critical environments**: Unlike traditional cybersecurity threats, hallucinations arise from the intrinsic limitations of generative models. In industrial settings, an AI assistant may generate an incorrect equipment maintenance procedure; in healthcare, it may provide inaccurate clinical guidance; in critical infrastructure, it may recommend unsafe operational actions. Robust validation mechanisms, domain-specific constraints, and human oversight become essential safeguards;
- **Safety concerns around autonomous AI agents deployment:** As generative models increasingly interact with actuators, robots, drones, and industrial control systems, failures can translate directly into physical consequences. Consequently, future GenAIoT systems require **safety-by-design** approaches incorporating redundancy, fail-safe mechanisms, runtime monitoring, and emergency intervention capabilities.

Experience from an operational greenhouse controller driven by an edge language model shows why these mechanisms must be engineered explicitly [141][142]. Its safety interlock allowed commands when a sensor reading was missing, its decision cycle had no watchdog, and field data showed heater-off breaches of the configured temperature floor lasting from 45 to more than 75 minutes, far beyond the designed worst-case latency of 7 to 8 minutes. Runtime assurance therefore needs fail-closed handling of missing inputs, enforced deadlines and deterministic fallback actions that do not wait for the generative model.

Trustworthy GenAIoT, therefore, demands **integrated security and safety engineering frameworks** capable of addressing both cyber threats and AI-specific risks. Future research should focus on secure model architectures, trustworthy, federated learning, adversarial robustness, verifiable AI reasoning, runtime assurance mechanisms, and human-in-the-loop safety controls.

7.4 Privacy, Data Governance, and Regulatory Alignment

While processing data locally inherently improves data privacy, managing distributed GenAIoT deployments introduce new compliance complexities, including aligning localised data synthesis with evolving regulatory frameworks, ensuring data provenance, and governing IP at the edge.

Moving generative intelligence closer to the edge can offer important privacy and data-governance benefits, particularly by reducing the need to transfer raw or sensitive data to centralised infrastructures. At the same time, decentralisation creates new governance and regulatory challenges. Data processing, model inference, decision-making and actuation may be distributed across devices, edge nodes and cloud services, often involving multiple technology providers, system integrators, operators and deployers. As a result, regulatory and governance considerations need to extend beyond individual AI models and address GenAIoT systems across their full lifecycle. Relevant factors include the intended purpose of the system, the types and provenance of data being processed, the capabilities and limitations of its AI components, the degree of autonomy, human oversight mechanisms, the deployment context, and the roles and responsibilities of the different actors involved. This becomes particularly important in systems that evolve dynamically over time. Models may be updated or replaced, SW and orchestration policies may change, new data sources may be introduced, and processing may shift between device, edge and cloud environments. Such changes can alter system behaviour and risk and may affect the technical and organisational measures on which previous assessments were based.

In this context, data, model and SW provenance become key elements of trustworthy GenAIoT. Relevant information may include the origin and transformation of data, model and SW versions, configuration changes, processing locations and significant operational events. Maintaining this information can support accountability, auditability, incident investigation, intellectual-property governance and regulatory assessment. AI-assisted approaches may help organisations manage this growing complexity. Natural-language processing and LLMs can support activities

© AIOTI. All rights reserved.

such as navigating regulatory and standardisation documents, identifying potentially relevant requirements, analysing technical documentation, maintaining traceability between requirements and system characteristics, and locating evidence relevant to an assessment. These mechanisms should, however, support rather than replace accountable legal, technical and conformity-assessment expertise.

European research is already exploring such approaches. The Horizon Europe **TRUST – Supporting and Valuing EU Regulatory Compliance with a Seal of TRUST and Co-designed Online Services based on Trustable LLM** project investigates how reliable LLM-based services can support organisations in understanding and applying EU regulatory requirements. Its scope includes regulatory information retrieval, identification and operationalisation of obligations, regulatory risk analysis, preparation of mandatory reports, compliance assessment, and support for auditing and certification processes. The project also addresses the reliability of LLM-based legal guidance, including risks related to hallucinations, privacy and the handling of sensitive information. In this way, TRUST illustrates both the potential of AI-assisted regulatory support and the safeguards required for its responsible use [116].

A key future direction is therefore to strengthen traceability between **requirements, system characteristics, controls and evidence**. Common terminology, structured system descriptions, provenance mechanisms and interoperable evidence models could make it easier to assess complex GenAIoT deployments across technological and organisational boundaries. Governance and assessment mechanisms will also need to accommodate change over time. Significant modifications to models, SW, data processing or deployment configurations may require previous assessments to be revisited. Likewise, developments in legislation, regulatory guidance or relevant standards may need to be evaluated against existing deployments. Supporting such event-driven reassessment in a reliable, proportionate and auditable manner remains an important research and standardisation challenge.

One of the key advantages of edge-based generative intelligence is the ability to process sensitive information closer to its source. This reduces the need to transfer personal, industrial, or operational data to centralized cloud infrastructures, thereby lowering exposure to privacy breaches and reducing compliance concerns. However, decentralized processing does not eliminate privacy risks; rather, it shifts them towards model management, data governance, and lifecycle oversight.

A significant challenge involves **data provenance and traceability**. Organizations must be able to identify where and how training data originated, how it was processed, and whether appropriate consent, licensing, or ownership rights existed. As edge models learn from locally generated data streams, maintaining transparent data lineage becomes increasingly complex yet essential for accountability and regulatory compliance.

Another major issue concerns **IP rights.** GenAI systems can create synthetic content, SW code, operational procedures, and design artifacts. Questions remain regarding ownership of AI-generated outputs, particularly when distributed models are trained on datasets originating from multiple stakeholders across different jurisdictions. Future governance frameworks must clarify responsibilities and rights relating to generated content within decentralized ecosystems.

The emergence of regulations such as the **EU AI Act** [117], **GDPR** [118], and sector-specific compliance frameworks highlights the need for regulatory alignment throughout the AI lifecycle. Organizations deploying GenAIoT systems must demonstrate risk management processes, transparency measures, human oversight mechanisms, and accountability procedures consistent with evolving legal requirements. Compliance becomes particularly challenging when AI models continuously adapt and operate across multiple geographical regions.

Privacy-preserving technologies are expected to play a central role in future GenAIoT deployments. Techniques such as federated learning, differential privacy, secure multiparty

© AIOTI. All rights reserved.

computation, confidential computing, and homomorphic encryption provide mechanisms for extracting value from distributed data while minimizing privacy risks. However, practical deployment remains constrained by computational cost, performance trade-offs, and implementation complexity.

The **Responsible AI perspective** further emphasizes the importance of governance structures that extend beyond legal compliance. Organizations must establish policies, roles, responsibilities, audit processes, and monitoring mechanisms that ensure ongoing alignment with ethical principles and societal expectations. Trustworthy governance will therefore become a strategic differentiator for future GenAIoT ecosystems.

7.5 Future Research Trends

Future research on trustworthy GenAIoT systems must move beyond traditional AI performance optimization and adopt a holistic RAI perspective that integrates technical, ethical, societal, legal, and environmental dimensions.

A key research direction involves the development of **explainable GenAI for IoT and edge environments**. Future systems should be capable of providing human-understandable explanations for generated outputs, decisions, and autonomous actions. Explainability mechanisms must be lightweight enough for resource-constrained devices while remaining meaningful to end-users and auditors.

Another emerging area is **trust-aware autonomous agents**. As AI agents increasingly collaborate, negotiate, and coordinate across distributed IoT networks, new trust models will be needed to evaluate reliability, reputation, intent, and behavioural consistency among intelligent entities. Research is required on agent governance, accountability frameworks, and multi-agent trust architectures.

Continuous assurance and runtime governance represent another promising field. Rather than certifying systems only before deployment, future GenAIoT infrastructures should support continuous monitoring, auditing, risk assessment, and adaptive compliance verification throughout operational lifecycles.

Researchers should also explore **human-centred trust calibration** mechanisms. The goal is not to maximize trust indiscriminately but to ensure that trust is justified and aligned with actual system capabilities. This includes investigating transparency mechanisms, human oversight strategies, and collaborative decision-making models that prevent both excessive reliance and unwarranted distrust. Measurements on edge language models illustrate the gap between apparent and justified confidence: repeated runs with fixed seeds revealed essentially nothing about whether a decision was correct, and agreement among sampled outputs was poorly calibrated, with expected calibration errors of 0.16–0.35 [142]. Trustworthiness metrics for GenAIoT should therefore include outcome-oriented measures, such as the rate of unsafe actuation in situations that require restraint and the share of the physical safety margin consumed by decision latency.

Further attention is needed regarding the **socio-economic impacts of widespread GenAI adoption**. RAI frameworks raise important questions regarding labour market transformation, digital inclusion, monopolization risks, and equitable distribution of AI benefits. Future GenAIoT research should include methodologies for measuring social, economic, legal, and environmental impacts throughout adoption and adaptation processes [112].

Finally, there is a growing need for **standardized trustworthiness metrics and certification frameworks**. Current evaluations primarily focus on model accuracy and efficiency, whereas future GenAIoT ecosystems will require multidimensional metrics encompassing fairness, transparency, robustness, accountability, privacy preservation, energy efficiency, and societal

© AIOTI. All rights reserved.

impact. Such frameworks will be essential for creating trustworthy, interoperable, and globally accepted GenAIoT infrastructures.

Future research needs to explore intermediation services, as discussed in various EU regulations, that mediate GenAIoT data exchange across the continuum while ensuring trust, security, accountability, and regulatory compliance. Such services need to be optimised to minimise data-sharing latency while also reducing deployment cost, so that using data intermediation services remains sustainable [114]. Another trend is enabling GenAIoT-based decision support systems to validate and audit GenAI outputs before they reach decision makers, which can significantly enhance the reliability of GenAI-based decision support [119]. Applying data-sovereignty principles and regulatory alignment across different sectors of society is also a promising research trend, one that can strengthen user trust in GenAIoT systems and support their wider adoption [16][98][113][115][120].

© AIOTI. All rights reserved.

8 Ethics

8.1 Introduction

GenAIoT systems introduce unprecedented opportunities for intelligent automation, adaptive decision-making, and autonomous service delivery. However, the integration of generative reasoning capabilities into connected devices, cyber-physical infrastructures, and Edge Computing platforms also raises profound ethical questions. Unlike traditional AI systems that primarily support human decision-making, GenAIoT systems can increasingly generate recommendations, actions, SW agents, synthetic content, and operational strategies that influence both digital and physical environments.

A forward-looking perspective on the socio-technical impact of GenAIoT is therefore essential. Ethical considerations must extend beyond individual algorithms and encompass entire ecosystems involving data providers, AI developers, device manufacturers, infrastructure operators, regulators, and citizens. As autonomous generative agents become embedded in smart homes, healthcare facilities, industrial systems, public infrastructures, and urban environments, ethical principles must guide the design, deployment, governance, and evolution of these technologies.

The challenge is further amplified by rapid advances in foundation models, specialized edge AI accelerators, federated intelligence, and autonomous multi-agent systems. In future GenAIoT ecosystems, millions of interconnected devices may continuously learn, reason, collaborate, and act with limited human intervention. Building trustworthy and socially acceptable systems therefore requires a comprehensive ethical framework addressing transparency, fairness, privacy, accountability, sustainability, human autonomy, and societal well-being.

Drawing upon the EU AI Act [117] and Responsible/Trustworthy AI principles, ethics should not be viewed as a compliance exercise but as a foundational design philosophy [112]. Responsible GenAIoT systems must ensure that technological innovation remains aligned with human values, democratic principles, fundamental rights, and societal objectives. Failure to address ethical concerns may erode public trust, increase regulatory risks, and hinder large-scale adoption of intelligent edge ecosystems.

Ethics in GenAIoT encompasses the principles, values, and governance mechanisms that ensure intelligent systems operate in ways that respect human rights, human dignity, societal norms, and environmental sustainability. Ethical concerns emerge at multiple levels, including how data is collected and used, how models are trained, how decisions are generated, how risks are managed, and who is held accountable when failures occur.

A distinguishing feature of GenAIoT is its ability to influence physical environments. Unlike purely digital AI systems, generative edge intelligence may directly affect healthcare treatments, industrial processes, transportation operations, energy management decisions, and public services. Consequently, ethical failures can have tangible consequences for individuals and communities.

Responsible AI frameworks emphasize that ethical deployment requires balancing innovation with societal responsibility [112]. This includes preserving human agency, ensuring meaningful oversight, preventing harm, reducing bias, protecting fundamental rights, and promoting equitable access to AI benefits. Such principles become particularly important as future GenAIoT ecosystems evolve towards autonomous agentic networks capable of collaborative reasoning and decision-making.

Ethics therefore serves as a strategic enabler rather than a constraint. By embedding ethical considerations throughout the lifecycle of GenAIoT systems, organizations can foster public trust,

© AIOTI. All rights reserved.

improve adoption, strengthen regulatory readiness, and support sustainable technological development.

8.2 Challenges in Implementing Ethical Principles in GenAIoT Systems

While ethical principles are widely recognized, implementing them in real-world GenAIoT systems presents considerable technical, organizational, and societal challenges.

Human Agency and Oversight: One of the primary ethical concerns relates to preserving meaningful human control over increasingly autonomous systems. As GenAI agents become capable of independently interpreting contexts, generating strategies, and initiating actions, human operators may gradually lose visibility into system behaviour.

The challenge is particularly pronounced in industrial automation, autonomous transportation, smart healthcare, and critical infrastructure applications where AI-generated decisions may directly influence operational outcomes. Ensuring effective human oversight requires new mechanisms for intervention, monitoring, and decision validation that remain practical in highly distributed environments.

Explainability and Transparency: Most advanced generative models operate with limited interpretability. Their internal reasoning processes remain difficult to understand, especially when large foundation models are deployed across heterogeneous edge environments. Stakeholders increasingly demand explanations regarding how decisions are generated, which data sources influenced outcomes, and why specific actions were recommended. Without sufficient transparency, users may either over trust or under trust system capabilities, both of which can create operational and ethical risks.

Therefore, development of explainable, traceable, and transparent AI solutions for resource-constrained edge devices remains an important challenge for future GenAIoT systems.

Fairness and Bias: Ethical deployment requires that GenAIoT systems operate fairly across diverse individuals, communities, and organizational contexts. However, generative models frequently inherit biases present within training datasets.

In decentralized environments, local edge models may be trained on geographically or demographically limited datasets, potentially reinforcing existing inequalities. Biases may emerge in predictive maintenance prioritization, healthcare recommendations, resource allocation decisions, public service delivery, or workforce management systems.

Achieving fairness requires continuous monitoring, bias auditing, representative datasets, and governance mechanisms capable of identifying unintended discriminatory outcomes.

Accountability and Responsibility: A single AI-generated decision may depend on multiple stakeholders, including foundation model providers, device manufacturers, platform operators, data owners, system integrators, autonomous SW agents, and infrastructure operators. Therefore, determining accountability in GenAIoT ecosystems is substantially more complex than in traditional SW systems.

When harmful outcomes occur, identifying responsibility becomes difficult. Future governance models must clearly define liability, roles, responsibilities, and decision-making authority across distributed AI ecosystems.

Ethical and Transparent Use of Synthetic/Real Data: GenAI increasingly produces synthetic datasets or uses real world data for model training, testing, and privacy protection. While synthetic data can reduce privacy risks and alleviate data scarcity, important ethical questions remain regarding data authenticity, representation accuracy, hidden biases, data ownership, consent mechanisms, misuse potential, etc.

© AIOTI. All rights reserved.

Researchers and practitioners must establish methodologies to evaluate the quality, fairness, and societal implications of synthetic data generation and real-world data collection and governance practices.

Surveillance and Societal Impact: The widespread deployment of intelligent sensors and generative analytics across public and private spaces raises concerns regarding surveillance, autonomy, and civil liberties as defined under GDPR [118]. Future smart cities or factories may incorporate pervasive sensing and AI-driven behavioural analysis capabilities. While these technologies can improve public services and infrastructure management, they may also create risks related to excessive monitoring, profiling, and erosion of individual freedoms. Therefore, ethical GenAIoT design must balance innovation with proportionality, necessity, transparency, compliance to GDPR, EU AI Act and respect for fundamental rights.

Environmental and Sustainability Concerns: Training and operating advanced generative models often require significant computational, energy, and water resource (for cooling functions in AI data centres) consumption.

Although Edge Computing can reduce communication overhead and improve operational efficiency, increasing numbers of intelligent devices may contribute to rising energy demand and electronic waste (e-waste). Ethical deployment should therefore consider environmental sustainability alongside performance objectives.

Future systems should prioritize energy-efficient AI architectures, sustainable AI deployment design, circular economy principles, and carbon-aware computing strategies.

8.3 Future Research Trends and Directions

The ethical development of GenAIoT systems will require interdisciplinary collaboration across AI, engineering, social sciences, philosophy, public policy, law, and human-computer interaction.

One important research direction involves **Ethics-by-Design methodologies**, where ethical principles are embedded into system architectures from the earliest stages of development rather than addressed retrospectively. Future frameworks should provide practical tools, metrics, and engineering processes that translate abstract ethical principles into operational requirements.

Another emerging area is **Value-Aligned Autonomous Agents**. As agentic AI systems become increasingly capable of independent reasoning and action, researchers must develop mechanisms ensuring that agent behaviour remains aligned with human intentions, societal norms, and organizational objectives. This includes research on value alignment, behavioural constraints, and ethical decision-making models.

The development of **Explainable and Interpretable GenAI at the Edge** represents another critical challenge. Future solutions should provide real-time explanations that are understandable to operators, regulators, and citizens while remaining computationally efficient in resource-constrained environments.

Research is also needed on **Ethical Governance Frameworks for Multi-Agent Systems**. Future GenAIoT ecosystems may consist of large numbers of autonomous agents collaborating across multiple domains. New governance mechanisms will be necessary to coordinate agent behaviour, resolve conflicts, establish trust relationships, and ensure accountability.

Another promising direction involves **Continuous Ethical Auditing**. Traditional compliance assessments are insufficient for adaptive systems that evolve over time. Future research should explore mechanisms for ongoing ethical monitoring, automated risk assessment, fairness evaluation, transparency verification, and regulatory compliance validation throughout operational lifecycles.

© AIOTI. All rights reserved.

The emergence of next-generation AI HW also introduces opportunities for **Ethical AI Acceleration Architectures**. Specialized edge processors can support privacy-preserving computation, secure execution environments, explainability functions, and energy-efficient generative inference. Research should investigate how HW innovation can directly support ethical objectives.

Finally, there is a growing need for **Global Ethical Standards and Certification Mechanisms**. As GenAIoT technologies become globally deployed, interoperable ethical frameworks and certification methodologies will help establish consistent expectations regarding transparency, accountability, fairness, privacy, safety, and sustainability.

In the long term, the ethical success of GenAIoT will depend on achieving a balance between autonomy and oversight, innovation and responsibility, personalization and privacy, efficiency and fairness. By embedding Responsible and Trustworthy AI principles into technological development, future GenAIoT ecosystems can evolve in ways that maximize societal benefit while minimizing ethical risks. Ethics should therefore be regarded not as an external constraint on innovation, but as a fundamental enabler of sustainable, trustworthy, and human-centred intelligent systems.

© AIOTI. All rights reserved.

9 Standardisation and Regulations

9.1 Introduction

GenAIoT brings together AI, IoT, Edge Computing, distributed systems and, increasingly, autonomous and agentic capabilities. This convergence creates new challenges for both regulation and standardisation, as relevant requirements may apply not only to individual AI models, but also to the behaviour, integration and governance of complete systems operating across the IoT–Edge–Cloud continuum.

Such systems may simultaneously fall within several regulatory and technical domains. In the European context a single GenAIoT deployment may need to be compliant with a set of regulations, among which the AI Act [117], the Cyber Resilience Act (CRA) [121] as a product with digital elements, NIS2 [122] through the obligations of the operating entity, the Data Act [123] in respect of data generated by connected products, and, if the system controls physical machinery, the Machinery Regulation [124]. These regulations often overlap. For example, under Article 12 of the CRA, a high-risk AI system that meets the CRA's cybersecurity requirements is considered to meet Article 15 of the AI Act as well and is assessed through the AI Act's own conformity procedure [121]. This reduces duplication but also means that evidence generated for one regulation must be reusable for another.

Technical standardisation has a central implementing role in the AI Act: under Article 40(1), high-risk AI systems that follow harmonised standards listed in the Official Journal of the EU are presumed to meet the Act's corresponding requirements [117]. The expected deliverables are set out in the EC' standardisation request [125] and are being developed by CEN-CENELEC JTC 21 [126]. The development of these standards is therefore an important interface between regulatory objectives and engineering practice [127][128]. For GenAIoT, a central challenge is to establish clear and auditable links between:

- Regulatory and policy requirements;
- Technical and organisational measures;
- Evidence supporting assessment and assurance.

In distributed systems, such traceability may span different technologies, deployment environments and organisations. Addressing it effectively will require cooperation between industry, research organisations, standards development organisations, regulators and conformity-assessment stakeholders. Standardisation can support this process through common terminology, system-description models, provenance mechanisms, testing methodologies and approaches for documenting and exchanging assurance-relevant information. Such mechanisms should enable interoperability while remaining proportionate to system characteristics and risk, and without prescribing a single implementation architecture.

9.2 Challenges in Implementing Standardisation and Regulations in GenAIoT Systems

One of the first challenges is defining the **system boundary**. A GenAIoT deployment may include sensors, embedded or edge AI models, cloud-based foundation models, orchestration components, external services and physical actuators. Understanding how such a system should be described and assessed, and how responsibilities relate to its components, can be significantly more complex than for a centrally deployed AI application.

This complexity is further increased by the **dynamic nature of distributed AI systems,** where components may evolve over time. Edge-specific optimisation techniques may also result in different model variants operating on different devices or under different conditions. Changes of this kind can affect system behaviour, and the validity of documentation or previous

© AIOTI. All rights reserved.

assessments. Mechanisms are therefore needed to identify when a change requires additional analysis, evidence, or reassessment.

The importance of maintaining documentation and traceability throughout the AI lifecycle is increasingly recognised in research. Recent work on technical documentation for the AI Act, for example, highlights the need to document data, models and applications in ways that support traceability and reproducibility as systems evolve [129]. For GenAIoT, the challenge is broader because relevant evidence may be generated and maintained across heterogeneous devices and cloud environments. A related challenge concerns the **distribution of responsibilities across the value chain**. Different components may be developed, supplied, and operated by different organisations. Supporting regulatory assessment therefore requires appropriate information to flow between these actors while protecting security-sensitive information, confidential business data, and IP.

The same fragmentation applies to **assurance evidence**. In a distributed GenAIoT environment, evidence may originate from different sources, such as design documentation, cybersecurity controls, or operational logs, and be spread across multiple technical and organisational domains. Common information models and interoperable mechanisms for describing system characteristics, controls, provenance, and assurance evidence could help reduce this fragmentation. Standardisation should support such exchange regardless of the specific system architecture, deployment model, or compliance process. This is particularly important given that European AI standardisation must bridge technical practice with requirements relating not only to performance and safety, but also to fundamental rights and other objectives embedded in the European regulatory framework [128][129].

Another important consideration is that regulatory **and standardisation environments also evolve over time,** which may affect existing deployments. This creates a need for approaches that can identify when changes to either the system or its external requirements should trigger further analysis or reassessment. AI-assisted methods may help organisations manage this. Natural-language processing and LLMs can support regulatory information retrieval, analysis of technical documentation, traceability between requirements and evidence, and identification of system elements potentially affected by regulatory or technical changes. However, proposals for LLM auditing argue the importance of assessing governance, models, and their concrete applications together, as well as recognising the limitations of automated assessment [130]. AI-generated outputs in regulatory contexts should therefore be considered decision support rather than authoritative determinations of legal compliance.

The integration of GenAI in the IoT-Edge-Cloud continuum has an **impact at standardisation level** in terms of best practice on i) GenAI engineering: how to link prompt engineering, context engineering, harness engineering, graph engineering, and ontology engineering; ii) GenAI architecture: how to characterise GenAI capabilities and allocate them in the IoT-Edge-Cloud continuum.

The integration of GenAI in IoT-Edge-Cloud continuum also has an **impact at compliance level**, in terms of best practice on ensuring compliance with regulations at technology and supply chain level such as the AI act, the Data act, the CRA, the Ecodesign for Sustainable Product Regulation (ESPR) [131] and associated Digital Product Passport (DPP) [132], and a set of market verticals related regulations, e.g., regarding Energy flexibility [133]; best practice on GenAI governance in complex ecosystems: how to ensure that the various stakeholders operating a IoT-Edge-Cloud coordinate their actions (e.g., upon incidents).

To this end, ISO/IEC JTC 1/SC 41 (IoT and DT) started in June 2026 a Physical AI Adhoc Group (AhG36) with the objective to study physical AI use cases. The terms of reference of this AhG are the following:

- Explore and analyse use-cases of IoT and DT enabling Physical AI;
- Prepare a draft Technical Report (TR) on Use Cases of IoT and DT enabling Physical AI;

© AIOTI. All rights reserved.

- Conduct thorough research on JTC 1/SC 7, JTC 1/SC 42, ISO/TC 299, IEC/TC 65 (including its SCs), ISO/TC 184 (including its SCs) and ITU-T SG21;
- Identify further liaison needs;
- Maintain close collaboration with the Physical AI Technical Trend Report being developed by ISO/IEC JTC 1.

The next sections provide an analysis of the challenges and identify research trends related to standards and regulations.

Challenge 1: Cloud Edge IoT Ecosystems to support Agentic AI

A IoT-Edge-Cloud ecosystem involving DTs will typically be extended to support agentic AI. Figure 9-1 shows an example, based on ISO/IEC 30188:2026 (DT reference architecture):

- The twin (or DT) is twinned with a system (or target entity).
- AI agents use the twin (digital representation of the target entity, called system in the figure).
- AI agents use observed data associated with the twin, as well as trained models from the observed data of multiple twins.
- AI agents output predictions and act on the system.

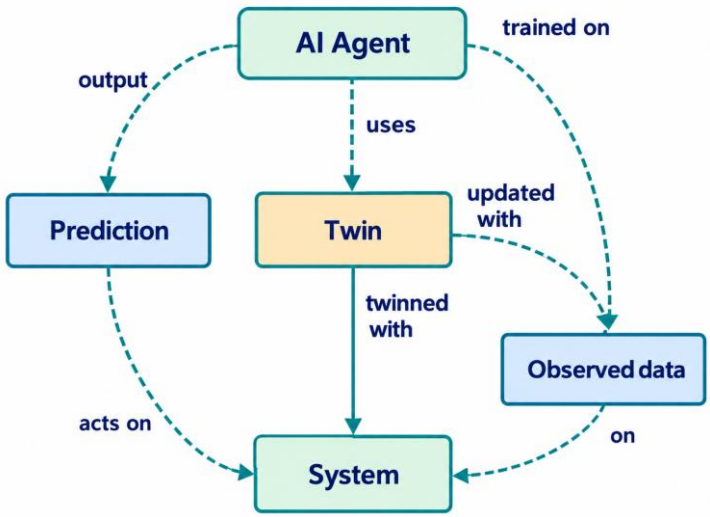


Figure 9-1 AI agent using a DT.

Figure 9-2 shows examples of four physical architecture allocations:

- In allocation 1, the AI agent and the twin are in the cloud level;
- In allocation 2, the AI agent is in the cloud while the twin is at the edge level;
- In allocation 3, both the AI agent and the twin are at the edge level;
- In allocation 4, both the AI agent and the twin are at the IoT level.

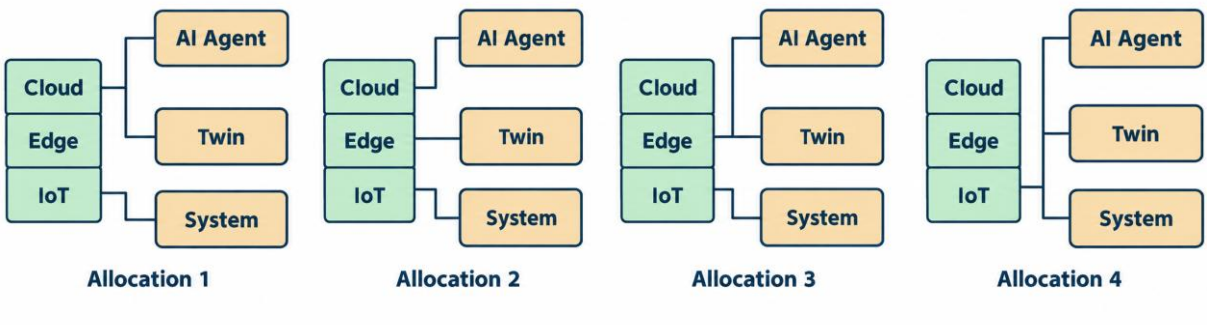


Figure 9-2 Examples of physical architecture allocations.

© AIOTI. All rights reserved.

Challenge 2: Integrate Human Machine teaming

A IoT-Edge-Cloud ecosystem integrating agents can also have to support human-machine teaming, ISO/IEC 25589 (Artificial intelligence — Framework for human-machine teaming), which is under development describes the following types of interactions:

- **Human supervisor**; the human oversees the AI application in execution;
- **Human mentor**: the human expert guides the AI application as a mentor;
- **Peer**; both the human and the AI application play equally important parts;
- **Machine mentor**: the human performs a task, and the AI application acts as a coach;
- **Machine supervisor**: the human performs the task while the AI application supervises the human.

Figure 9-3 shows two vertical twins which are managed by a local DT system above. In the figure, interoperability points are represented by green dots. Each vertical twin has two interoperability point, one inner point between a twin and a system, and one transversal point between the twin and the local DT system. The local DT system consists of a mix of human experts and agents. In Figure 9-3, there are two verticals one including twin A and system A, and one including twin B and system B.

- The local DT includes
  - Users of Twin A: Human expert 1 and Agent 2.
  - Users of Twin B: Agent 3, Human expert 5, and Agent 5.
- Human machine teaming can be specific to a vertical:
  - Focusing on Twin A, Human expert 1 teams with Agent 2.
  - Focusing on Twin B, Human expert 4 teams with Agent 5.
- Human machine teaming can be transversal to two verticals:
  - Human expert 1 using Twin A teams with Agent 3 using Twin B.
  - Agent 2 using Twin A teams with Human expert 4 using Twin B.
- Machine-machine teaming can also take place:
  - Agent 2 using Twin A teams with Agent 5 using Twin B.

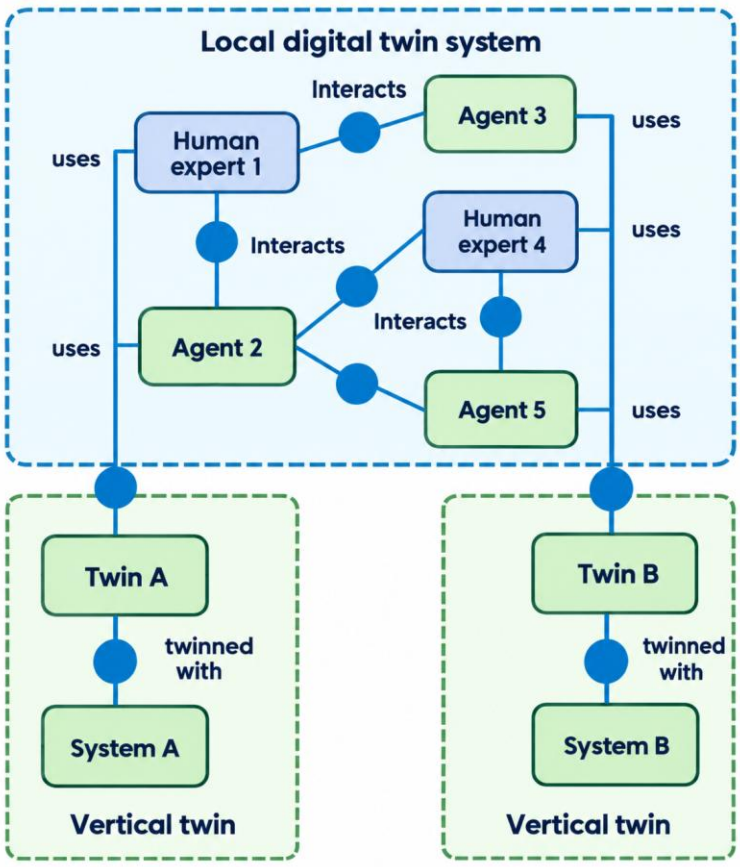


Figure 9-3 Human-machine teaming example.

© AIOTI. All rights reserved.

Challenge 3: Align with upcoming AI standards related to GenAI

It will be also important to align with upcoming AI standards. Several standards related to GenAI are anticipated but not available yet:

ISO/IEC JTC 1/SC 42 standards:
- ISO/IEC 22989 Amendment 1 on GenAI;
- ISO/IEC 25568 Artificial Intelligence — Guidance on addressing risks in GenAI systems;
- ISO/IEC 25589 Artificial intelligence — Framework for human-machine teaming;
- ISO/IEC 42109 Artificial intelligence — Use cases of human-machine teaming;
- ISO/IEC 25880 Artificial intelligence — Requirements and guidance for the organizational implementation of human-machine teaming;
- ISO/IEC 25512 Artificial intelligence — Frontier AI risk management and impact assessment (preliminary work item);

ISO/IEC JTC 1/SC 27 standards:
- ISO/IEC 25240 Evaluation of AI-based technology (preliminary work item);
- ISO/IEC 25959 Application of attack potential to deep learning-based technology;
- ISO/IEC 26160 Enhancing the ISO/IEC 15408 series and ISO/IEC 18045 for the Evaluation of Artificial Intelligence (AI) Functionality;
- Proposal underway on security evaluation and testing methods for GenAI models and for AI agents and their orchestration;
- ISO/IEC JTC 1/SC 44 standards: Consumer-centric AI agents through privacy-by-design (Preliminary work item).

Challenge 4: Ensure alignment of GenAI standards with other standards

The integration of GenAI in the IoT-Edge-Cloud continuum will require alignment in the following categories of standards:

Data standards such as:
- ISO/IEC 5207:2024 (Terminology and use cases);
- ISO/IEC 5212:2024 (Guidance for data usage);

Data space standards, such as:
- ISO/IEC 20151-1 (Concepts and characteristics);
- ISO/IEC 20151-2 (Trust frameworks);
- ISO/IEC TR 25850 (Use cases for data spaces);
- ISO/IEC 26451 (Eclipse decentralized claims protocol);
- ISO/IEC 26450 Eclipse Dataspace Protocol;

IoT and DT architecture standards such as:
- ISO/IEC 30141:2024 (IoT reference architecture);
- ISO/IEC 30188:2026 (Digital twin — Reference architecture);
- ISO/IEC 30152 (IoT and digital twins - Guidance on the connection to data spaces), while being aligned with horizontal architecture standards such as
- ISO/IEC/IEEE 42010:2022 (architecture description);
- ISO/IEC/IEEE 42024 (architecture fundamentals);
- ISO/IEC/IEEE 42042 (reference architecture);

cloud, edge, IoT and DT interoperability standards such as:
- ISO/IEC 19941 (Cloud computing interoperability);
- ISO/IEC 21823-1 (IoT interoperability framework);
- ISO/IEC 21823-3:2021 (semantic interoperability);
- ISO/IEC 21823-5 (behavioural and policy interoperability);

trustworthiness standards such as:
- ISO/IEC 5723:2022 (Trustworthiness vocabulary);
- ISO/IEC 31303 (Trustworthiness overview);
- ISO/IEC 31310 (Trustworthiness ontology);

© AIOTI. All rights reserved.

- ISO/IEC 30147 (Integration of IoT trustworthiness activities in ISO/IEC/IEEE 15288 system engineering processes);
- ISO/IEC 30149:2024 (IoT Trustworthiness principles);
- ISO/IEC 5181 (Security and privacy protection — Data provenance);
- ISO/IEC 25985 (Trusted data usage);

Complex system engineering standards such as:

- ISO/IEC 27564:2025 (Guidance on the use of models for privacy engineering);
- ISO/IEC 27568 (Security and privacy of digital twins);
- ISO/IEC 27570:2021 (Privacy guidelines for smart cities);

Cybersecurity standards on system of systems, such as:

- ISO/IEC 26601 (Guidance for Addressing Cybersecurity of System of Systems);
- ISO/IEC 27115-1 (Cybersecurity of System of Systems — Part 1: Security Architecture Description);
- ISO/IEC 27115-2 (Part 2: Security Architecture Evaluation);
- ISO/IEC 27115-3 (Part 3: Security Profiles);

Asset management standards such as:

- ISO 55000:2024 Asset management — Vocabulary, overview and principles;
- ISO 55002:2018 Guidelines for the application of ISO 55001;
- ISO 55013:2024 Asset management — Guidance on the management of data assets;
- ISO/IEC 19770-1:2017 IT asset management — Part 1: Requirements;
- ISO/IEC 19770-2:2015 Software asset management — Part 2: Software identification tag;
- ISO/IEC 19770-5 IT asset management — Overview and vocabulary — Part 5;
- ISO/IEC 19770-10:2025 IT asset management — Part 10: Guidance for implementing ITAM;

Governance standards such as:

- ISO 37000:2021 (Governance of organizations — Guidance);
- ISO/IEC 38500:2024 (Governance of IT for the organization);
- ISO/IEC 38502:2017 (Governance of IT — Model and framework);
- ISO/IEC 38503:2022 (Assessment of the governance of IT);
- ISO/IEC 38505-1:2026 (Application of ISO/IEC 38500 to the governance of data);
- ISO/IEC 38505-2:2018 (Implications of ISO/IEC 38505-1 for data management);
- ISO/IEC 38505-3:2021 (Guidelines for data classification);
- ISO/IEC 10866:2024 (Framework and concepts for organizational autonomy and digital sovereignty);

AI security and privacy standards such as:

- ISO/IEC 27563:2023 (Security and privacy in artificial intelligence use cases — Best practices);
- ISO/IEC 27090 (Guidance for addressing security threats and compromises in artificial intelligence systems);
- ISO/IEC 27091 (Artificial intelligence – Privacy protection);

AI data standards, such as:

- ISO/IEC 8183:2023 (AI — Data life cycle framework);
- ISO/IEC 42103 (AI — Overview of synthetic data in the context of AI systems).

Challenge 5: Support regulation

The integration of GenAI in the IoT-Edge-Cloud continuum will require the support of upcoming regulations:

Cyber resilience act harmonised standards. They include:

- EN 40000-1-1 Cybersecurity requirements for products with digital elements - Part 1-1: Vocabulary;
- EN 40000-1-2 Cybersecurity requirements for products with digital elements - Part 1-2: Principles, product risk management, and lifecycle activities;
- EN 40000-1-3 Cybersecurity requirements for products with digital elements - Part 1-3: Vulnerability handling;

© AIOTI. All rights reserved.

- EN 40000-1-4 Cybersecurity requirements for products with digital elements - Part 1-4: Generic security requirements.

ESPR and digital product passport harmonised standards. They include:

- EN 18239:2026 Digital product passport - Access rights management, information system security, and business confidentiality;
- EN 18246:2026 Digital product passport - Data authentication, reliability and integrity.

AI Act harmonised standards. They include:

- EN 18228 AI Risk Management;
- EN 18229-1 AI trustworthiness framework - Part 1: Logging;
- EN 18229-2 AI trustworthiness framework - Part 2: Transparency;
- EN 18229-3 AI trustworthiness framework - Part 3: Human oversight;
- EN 18229-4 AI trustworthiness framework – Part 4: Accuracy;
- EN 18229-5 AI trustworthiness framework - Part 2: Robustness;
- EN 18282 Cybersecurity specifications for AI Systems;
- EN 18283 Concepts, measures and requirements for managing bias in AI systems;
- EN 18284 Quality and governance of data sets in AI;
- EN 18285 AI Conformity assessment framework;
- EN 18286:2026 Quality management system for EU AI Act regulatory purposes.

Data act harmonised standards. They include:

- EN 18235-3 (Trusted data transactions - Part 3: Interoperability requirements);
- EN 18235-2 (Trusted data transactions - Part 2: Trustworthiness requirements);
- EN 18352 (Quality framework for internal data governance for participants in data spaces);
- EN 303 760 (Implementation framework for semantic assets Guidelines for Semantic Interoperability);
- TR 104 244 (Implementation of Data Catalogues with NGSI-LD and DCAT-AP).

9.3 Future Research Trends and Directions

The following research trends and directions can be listed.

- **Cloud Edge IoT assets integrating GenAI**: a holistic and fine grain management of capabilities is needed on the continuum. To this end AIOTI will submit to ISO/IEC JTC 1/SC 41 on behalf of its members a proposal for a Iot-Edge-Cloud asset management standard, leveraging the EU research project EUcloudedgeiot [134], with the support of the CEI-Sphere (Cloud-Edge-IoT Spheres of Large-Scale Pilots) support action [135].
- **Integration of graph engineering:** a consistent semantic engineering approach leveraging the harmonised standard on semantic assets (EN 303 760) as well as ISO standards on semantic interoperability (ISO/IEC 21823-5). To this end, AIOTI members will participate to both EU and International standards to promote innovation outcomes from EU research projects.
- **Continuum architectures:** EU research projects should continue to work on innovation architecture and solutions and support the diverse continuum architectures needs. They can leverage the recently created Continuum interest group in the ECLIPSE foundation [136].

Future research and standardisation should investigate how regulatory and assurance considerations can be more effectively integrated into the engineering and operational lifecycle of GenAIoT systems.

One important direction is the development of **structured and, where appropriate, machine-interpretable representations of system characteristics, selected requirements, controls, and evidence**. Such representations could improve traceability throughout the compliance pipeline while enabling automated analysis for suitable parts of an assessment. Research on translating provisions of the AI Act into more concrete technical requirements and evaluation criteria illustrates both the potential and the difficulty of such approaches [137].

© AIOTI. All rights reserved.

A closely related direction concerns **interoperable assurance evidence**. Common information models and interfaces could allow independent components to provide unified relevant information on configuration, provenance, testing, monitoring, and implemented controls. Research is needed to determine the appropriate level of abstraction and how such information can be exchanged securely, proportionately, and across organisational boundaries.

Another important area is **change-impact assessment**. Methods are required to determine how changes to individual components may affect system behaviour, risks, and existing assurance claims. Similar approaches could help organisations assess the potential impact of new or revised legislation, regulatory guidance, and standards on systems that are already deployed.

AI-assisted regulatory and standardisation analysis also represents a promising research direction. LLMs and related techniques may help users navigate complex regulatory environments, identify potentially relevant requirements, structure technical documentation, and maintain traceability between requirements and supporting evidence. Further research is needed to establish appropriate levels of reliability, explainability, reproducibility, provenance, confidentiality, and human validation for such applications.

Several European research and innovation initiatives are exploring complementary aspects of these challenges. The Horizon Europe **TRUST** project investigates reliable LLM-based services for navigating EU regulation and supporting compliance and certification [116]. The Horizon Europe **CERTAIN** project addresses ethical and regulatory certification of AI systems, including approaches for lifecycle management, data lineage, regulatory workflows and certification support [138]. The **FAITH** project investigates methods for measuring and improving AI trustworthiness across the lifecycle and multiple critical application domains, with an explicit objective of generating evidence relevant to regulation and standardisation [139]. These activities illustrate a broader European movement towards linking regulatory requirements, engineering processes, trustworthiness assessments, and evidence in more systematic ways.

Research is also emerging on the use of AI to support the standardisation ecosystem. The Horizon Europe **StandardAIse** project, for example, is developing an AI-supported standardisation landscape hub combining LLMs, knowledge graphs and human expert validation to improve access to and navigation of standards [140]. Such approaches may become increasingly relevant as organisations need to identify and maintain links between rapidly evolving technologies, standards, and regulatory requirements.

A further research area concerns the **policy-aware operation of distributed AI systems**. Edge–cloud orchestration is typically driven by technical factors such as latency, computational capacity, bandwidth, energy consumption and cost. Future systems may also need mechanisms for representing and enforcing constraints related to privacy, data governance, security policies, permitted processing environments and other organisational or regulatory requirements when selecting where and how AI processing is performed.

Key research and standardisation topics therefore include:

- Common terminology and system descriptions for distributed GenAIoT systems;
- Traceability between requirements, system characteristics, controls and evidence;
- Interoperable provenance and assurance-evidence models;
- Machine-interpretable representations for selected regulatory and standardisation requirements;
- Change-impact assessment for evolving AI systems;
- Assessment of the impact of regulatory and standards evolution;
- Assurance approaches for autonomous and multi-agent systems;
- Allocation and documentation of responsibilities across GenAIoT value chains;
- Testing, monitoring and documentation of distributed AI behaviour;
- Explainable and auditable AI-assisted regulatory analysis;
- AI-assisted navigation and analysis of complex standardisation landscapes; and
- Appropriate human oversight and validation of AI-supported assurance activities.

© AIOTI. All rights reserved.

Progress in these areas could support a gradual transition from predominantly document-centric and point-in-time assessment towards more integrated, evidence-based and lifecycle-oriented assurance. Such approaches should complement established regulatory and conformity-assessment processes while helping the IoT and edge community address the growing complexity of deploying trustworthy GenAI systems across distributed environments.

© AIOTI. All rights reserved.

10 Recommendations

Based on the challenges and architectures discussed, actionable guidelines for industry and academia are needed. This section outlines recommendations on the critical research directions for safely deploying the next generation of intelligent edge systems and the architectural patterns organisations should adopt today. The recommendations synthesise the implications of the shift towards GENAIoT and identify priority research trajectories for the next years.

GenAIoT implementations and applications must shift from single-tier cloud computational models to unified processing fabrics spanning the device, edge, and cloud continuum by addressing architectural hybridity and computing continuum orchestration. Foundation models cannot be treated as isolated monolithic blocks running solely in remote hyperscalers data centres or forced entirely onto resource-starved microcontrollers. Enterprise architects must construct layered operational fabrics where local nodes execute real-time sensory tokenisation, intermediate gateways handle contextual synthesis through compressed models, and centralised cloud clusters govern longitudinal tracking and continuous model evolution.

Dynamic computational offloading must become a primary infrastructural standard. Edge systems require intelligent scheduling runtimes that continuously evaluate available network bandwidth, link jitter, thermal dissipation levels, and remaining battery life. When physical conditions degrade or radio connectivity drops, operational tasks must autonomously degrade gracefully to local inferencing. As network paths stabilise, systems should opportunistically leverage upstream compute resources to handle deeper multi-modal reasoning.

Model optimisation and compression toolchains must be directly integrated into the SW deployment lifecycle. Engineering implementations should make weight quantisation, structured layer pruning, and knowledge distillation standard elements of their deployment automation. Converting dense parameters into low-bit representations ensures that context-aware reasoning can execute within milliwatt and low-wattage operational budgets without compromising real-time performance.

HW procurement strategies must transition from general-purpose microcontrollers toward silicon tailored for low-precision tensor operations and parallel processing. Fleet modernisation efforts should prioritise Edge Computing nodes featuring specialised NPUs, vector math extensions, and high-bandwidth on-chip memory arrays. Procuring HW designed to reduce data movement between memory registers and compute cores is essential to mitigating the thermal throttling and energy drain typical of intensive transformer workloads.

Microservice architectures and container runtimes deployed at the periphery must adapt to dynamic generative workloads. System architects should deploy lightweight sandboxes and deterministic virtual execution layers that can serve modular models. These environments must isolate resource-intensive model inferencing from safety-critical operating system functions, ensuring that unpredictable processing spikes never interrupt essential device operations or deterministic communication buses.

Data serialisation and transport mechanisms across the distributed architecture must be modernised to accommodate non-textual data formats. Sensor telemetry can no longer be processed using traditional text-based formats that discard physical nuance and scale unmanageably. Research and engineering teams must implement tokenisation standards that preserve temporal alignments, measurement units, sensory variance, and HW origins directly within unified multimodal feature embeddings.

System developers must actively decouple probabilistic generative reasoning from deterministic physical actuation. Generative models excel at contextual diagnosis, multi-step scenario

© AIOTI. All rights reserved.

forecasting, and synthesising broad action directives, but their inherently probabilistic nature makes them unsuitable for direct, unverified HW actuation. Edge architectures must route all generated action proposals through deterministic safety layers, mathematical verification filters, and hard-coded physical limiters before commands reach physical motors, valves, or robotic joints, addressing the physical embodiment, agentic autonomy, and real-time control.

Autonomous physical agents must be granted dynamic planning capabilities backed by formal real-time validation. When deploying embodied edge robotics, autonomous mobile platforms, or smart manufacturing arms, systems should utilize localized reasoning to continually assess their physical surroundings and adjust their task sequences mid-stride. However, all self-generated plans must operate strictly within bounded kinematic envelopes, thermal thresholds, and physical invariant boundaries enforced by dedicated real-time microkernels.

GenAIoT implementations must pair physical edge agents with dynamic virtual models and localised DTs. By running compact generative models alongside continuous physical simulations, agents can project potential future states and evaluate risk trajectories before executing physical interventions. Simulating actions in virtual environments gives edge systems a practical way to test alternatives, assess failure probabilities, and optimise physical interactions safely.

Human-machine collaboration frameworks must be redesigned around grounded natural-language interactions and transparent diagnostics. Instead of delivering cryptic numeric error codes or uncalibrated confidence percentages, edge systems must leverage localised generative capabilities to synthesise plain-language explanations of observed anomalies, suspected root causes, and recommended physical adjustments. This dynamic enables human operators to collaborate smoothly with autonomous systems during complex maintenance and operational interventions.

Real-time local context must be continually reinforced using localised RAG and situational memory buffers. Edge systems should maintain structured, bounded local stores of operating manuals, historical service logs, site-specific maintenance histories, and dynamic physical schematics. Grounding synthesised actions and diagnostics in verified operational references drastically limits the risk of model hallucinations and ensures device responses remain tied to immediate physical reality.

Continuous local adaptation must be deployed without risking catastrophic forgetting or unauthorised behavioural drift. Device fleets operating in non-stationary physical environments require adaptive learning mechanisms to handle mechanical wear, sensor drift, and seasonal baseline shifts. GenAIoT implementations must implement parameter-efficient fine-tuning frameworks that adapt isolated sub-layers of edge networks on streaming local data, holding the underlying foundational base parameters permanently frozen.

Industrial and enterprise operators must enforce strict zero-trust operational models across the entire distributed generative edge ecosystem and address security governance, trust boundaries, and lifecycle operations. Every endpoint, intermediate fog node, and centralised coordinator must mutually authenticate all data transfers, task delegations, and model updates using HW-backed roots of trust. HW-isolated secure enclaves must safeguard model weights, cryptographic keys, and intermediate activation tensors from physical bus inspection or local memory extraction attacks.

Comprehensive defence frameworks must protect edge devices from adversarial manipulation and sensor spoofing. Physical input signals, ranging from video frames and acoustic waveforms to high-frequency current logs, can be deliberately perturbed through physical and environmental interference. Edge systems must incorporate real-time signal validation layers, multi-modal cross-sensor consistency checks, and adversarial detection filters to flag and reject manipulated inputs before they enter the generative reasoning pipeline.

© AIOTI. All rights reserved.

Decentralised data governance and privacy guarantees must be maintained by enforcing edge-native data residency. The primary business case for moving generative intelligence to the periphery is keeping sensitive personal information, proprietary operational patterns, and confidential site schematics on local premises. System policies must strictly prevent the transmission of raw, uncompressed operational telemetry to centralised servers, mandating that only fully anonymised, aggregated gradient vectors or high-level summaries ever traverse wide-area networks.

Model evaluation frameworks must expand beyond basic academic accuracy metrics to assess complete real-world operational cycles. System benchmarks must evaluate complete sensing-reasoning-action loops, capturing real-time latency, peak energy demand, system recovery after unexpected power loss, thermal stability, and overall physical task performance under stress. Systems must be tested in unpredictable operational environments, with edge-case sensory corruption and network dropouts, to confirm reasoning performance remains stable over extended operational lifespans.

Automated orchestration pipelines must be instituted to manage the complex lifecycle of distributed models at scale. Fleet operators need robust OTA deployment systems featuring dual-bank memory partitioning and automated rollback routines to handle mid-transmission failures safely. Model update pipelines must incorporate automated canary testing, performance regression detection, and distributed verification steps to ensure that updated weights do not introduce subtle operational regressions or erratic behavioural shifts.

GenAIoT applications deploying autonomous physical intelligence must establish cross-disciplinary safety and regulatory governance panels. Engineering leaders, cybersecurity professionals, control engineers, and regulatory officers must jointly define and audit the behavioural boundaries of autonomous physical agents. These governance groups must ensure that all autonomous edge deployments comply with emerging artificial intelligence safety regulations, respect international data sovereignty mandates, and maintain auditable decision trails for all safety-critical physical actions.

The transition from centralised, cloud-bound GenAI to decentralised intelligence across the processing continuum is a structural, not incremental, shift. GenAIoT becomes operationally viable if research communities, industry consortia, and standards bodies jointly address a tightly coupled set of scientific, architectural, hardware, governance, and socio-technical problems.

To implement GenAIoT applications, a research priority is systematically designing hybrid architectures that treat foundation-model inference not as a remote service but as a workload spanning edge devices (micro-, deep-, meta-edge) and cloud backends. Current orchestration stacks were designed for telemetry, rule engines, or compact discriminative models. They are structurally not well adapted to generative workloads whose memory footprints, token-generation latency, and energy profiles vary by orders of magnitude depending on prompt complexity, context length, and modality. Future systems research should therefore concentrate on:

Workload-aware placement and splitting of generative pipelines. Research is needed on schedulers and runtime systems that can partition encoder–decoder stacks, retrieval modules, world-model rollouts, and actuator planners across heterogeneous nodes, under joint constraints of latency, energy, bandwidth, and privacy. Static model offloading is insufficient; adaptive, context-triggered escalation (e.g., edge AI models classify locally, compact generative reasoning runs at the edge, and large-model synthesis runs in the cloud) must become programmable.

Continuum-native serving abstractions. Serving systems must expose generative inference as a distributed, interruptible, and resumable computation rather than a remote procedure call. This

© AIOTI. All rights reserved.

includes streaming token generation across hops, cache migration, decoding across tiers, and graceful degradation when connectivity or energy budgets are not sufficient.

Interoperability as an architectural invariant. Data interoperability, semantic, syntactic, and policy-level, must be treated as a design principle. Research on shared ontologies, digital product passports, and contract-based data exchange is necessary so that generative models can reason over heterogeneous IoT streams without proprietary lock-in.

Digital twins, immersive triplets, and closed-loop world models. GenAIoT systems increasingly couple physical sensors and actuators with continuously updated digital twins and immersive triplets. Research is needed on synchronisation protocols, uncertainty-aware twin–reality alignment, and using generative world models for anticipatory action, counterfactual planning, and safe what-if simulation before physical execution.

Decentralisation is computationally very challenging if generative intelligence remains synonymous with multi-billion-parameter cloud models. A research direction should be new hardware-aware compression and specialisation of generative models for edge and device constraints. Priority directions include:

Quantisation as the compression pathway. Research must move beyond uniform post-training quantisation toward mixed-precision, activation-aware, and token-adaptive schemes that preserve generative quality under 4-bit, 3-bit, and sub-4-bit regimes. Of particular importance is quantisation that remains stable under continual local adaptation and under multimodal (vision–language–audio–sensor) inputs typical of physical environments.

Pruning, structured sparsity, execution-aware model graphs and knowledge distillation from frontier LLMs to compact generative models. The transition from large language models to small, domain-grounded generative agents is a key GenAIoT research priority. Distillation must preserve linguistic understanding, model capabilities, reasoning, safety constraints, and multimodal grounding. Research on multi-teacher distillation, preference-aligned compact models, and on-device self-distillation from local experience is needed.

Compression beyond weights: KV cache, activation, and retrieval indices. Research on cache quantisation, eviction policies, cross-request cache sharing, and retrieval-index compression will determine whether long-context generative reasoning can ever run at the edge.

Hardware-aware co-design as an infrastructure requirement. Co-design of algorithms with NPUs, in-memory computing, neuromorphic architectures, and advanced packaging (chiplets, 3D stacking) is becoming a systems-infrastructure agenda that must be addressed alongside HW-SW-AI-Data optimisation.

GenAIoT changes the nature of IoT devices themselves. Sensors and actuators cease to be passive telemetry endpoints and become physical, embodied, agentic participants in physical processes. Research must therefore re-found sensing, actuation, and local intelligence. Multi-step spatial and logical reasoning must run locally so agents can replan mid-task, interact with humans, and operate under intermittent connectivity. Recommended directions:

Reusable sensor foundation models. The field should move from task-specific sensing pipelines toward foundation models pre-trained on large multimodal sensor corpora (vibration, current, vision, acoustics, RF, environmental) and adapted with few-shot or continual local learning. Research is needed on self-supervised objectives for non-vision IoT modalities, domain generalisation across edge devices, and on-device personalisation without forgetting.

Small language models as the cognitive base. The edge models should be repositioned from anomaly detection toward the *local foundation* of GenAIoT: always-on perception, trigger

© AIOTI. All rights reserved.

generation, safety interlocks, and compact generative or retrieval-augmented responses. Research on ultra-low-power generative decoding, event-driven transformers, and duty-cycled agent loops is important. Combining local reasoning with digital twins and immersive triplets is promising but immature: research should address twin–agent co-evolution, sim-to-real transfer for generative policies, and runtime assurance of physical actions originating from stochastic models.

Multimodal grounding in the physical environment. Generative models at the edge must be grounded in geometry, physics, and device semantics, not only in language. Spatial reasoning, affordance inference, and sensor-to-action alignment constitute a distinct research programme from cloud multimodal LLMs.

From generated interpretations to generated actions. A critical and understudied transition is the move from GenAI as a *semantic layer over sensing* (explaining, summarising, diagnosing) to GenAI as a *generator of actions* (plans, control sequences, natural-language commands to actuators, or policies for multi-step tasks). This requires research on constrained decoding, formal safety envelopes, verifiable plan generation, and tight coupling with classical control and optimisation.

Decentralising generative models expands the attack surface and introduces failure modes with physical consequences. Hallucinated diagnoses, adversarial prompts injected through sensors, poisoned local adapters, and unsafe generated control actions are not merely information-security issues; they are operational and safety-critical risks. Future research should prioritise:

A threat model specific to GenAIoT. Classical IoT security (device identity, secure boot, encrypted telemetry) is necessary but not sufficient. Research must characterise generative-specific threats: prompt injection via physical side channels, model extraction at the edge, membership inference on local fine-tunes, adversarial examples in multimodal sensor streams, and cascading failures when multiple generative agents interact.

Intrusion detection and runtime protection for edge inference. Detection must operate under tight energy and latency budgets. Promising directions include lightweight attestation of model integrity, activation-level anomaly detection, output-constrained decoding against safety policies, and hardware roots of trust for generative runtimes. Evaluation methodologies are currently inadequate: new datasets, protocols, and metrics that capture physical harm, not only classification accuracy, are required.

Hallucination under physical grounding. Research is needed on calibrated uncertainty, refusal behaviour, retrieval-augmented grounding in live sensor data, and formal bounds on generated plans. In GenAIoT, an ungrounded generation can become a physical actuation.

Trustworthiness as a system property. Trust cannot be reduced to model cards. It must span data provenance, model lineage, runtime monitors, human override, and post-incident accountability across a fleet of distributed models. Research on distributed model lifecycle management, updates, rollbacks, and kill-switches at scale remains unsolved.

Privacy, data governance, and sovereign context. On-device and edge inference improves privacy in principle but complicates governance in practice. Data spaces should be investigated as a *sovereign context layer* for GenAIoT: providing policy-aware access to context, consent, purpose limitation, and auditability for generative reasoning. Alignment with GDPR, the EU AI Act, sectoral safety regulations, and emerging GenAI-specific standards must be treated as a research constraint, not a compliance afterthought.

© AIOTI. All rights reserved.

Technological capability without institutional scaffolding will stall deployment. Several non-algorithmic research and standardisation agendas are therefore critical.

Operationalising ethics in resource-constrained, physical systems. Fairness, transparency, and accountability take different forms when models run on microcontrollers, when decisions have irreversible physical effects, and when users cannot inspect a cloud API. Research is needed on edge-feasible explainability, contestability of generated actions, and value alignment of compact agents.

Standards alignment across GenAI, IoT, edge, and safety domains. GenAI standards must align explicitly with IoT, industrial automation, cybersecurity, and functional-safety standards rather than being developed in isolation. Research and pre-standardisation work should map gaps: model cards for edge generative models, interoperability profiles for agent protocols, safety integrity levels for generative control, and conformity-assessment methods for continuously updated distributed models.

Cloud–edge–IoT ecosystems for agentic AI. Platforms must support discovery, composition, and governance of agents that span devices and clouds. Research on agent communication protocols, capability advertisement, multi-agent negotiation under resource constraints, and economic models for distributed intelligence will determine whether GenAIoT becomes an ecosystem or a set of silos.

Human–machine teaming as a first-class design goal. Generative edge systems will work *with* operators, technicians, and citizens. Research should address mixed-initiative control, natural-language and multimodal interfaces that remain reliable under noise and stress, appropriate trust calibration, and designs that keep humans meaningfully in or on the loop when generated actions affect safety, rights, or critical infrastructure.

GenAIoT raises an operations-research problem of unprecedented scale: how to engineer, observe, update, certify, and retire millions of generative models running in uncontrolled physical environments. Emerging directions include:

Selective escalation and edge–cloud collaboration policies. Research should formalise when a local agent may answer, when it must retrieve, when it must escalate, and when it must refuse. These policies are simultaneously systems, safety, and governance artefacts.

Evaluation science for GenAIoT. The community lacks shared benchmarks that combine generative quality, physical-task success, latency, energy, robustness to sensor faults, adversarial resistance, and human-team performance. Establishing such evaluation regimes is itself a high-leverage research contribution.

The decentralisation paradigm as a research theme. The trajectory from telemetry ("Connect") through local compute, task-specific ML, and discriminative AI to generative reasoning is not complete. The open question is whether fully autonomous, trustworthy edge intelligence can be achieved without re-centralising data, control, or model authority. Answering it will require co-evolution of algorithms, architectures, hardware, standards, and institutions.

© AIOTI. All rights reserved.

11 Conclusions

11.1 The Decentralisation Paradigm and the Evolution Toward Autonomous Edge Intelligence

The migration of GenAI from centralised hyperscalers architectures to the physical periphery represents a decisive architectural transformation across the IoT and Edge Computing domains. For decades, distributed computing evolved through sequential milestones: transitioning from the pure observation and uncompressed streaming of the telemetry era to localised filtering within fog computing gateways, and eventually to discriminative task-specific ML.

Today, the convergence of generative reasoning with distributed networks defines an entirely new era: the GenAIoT paradigm. This transition fundamentally shifts edge nodes from passive data forwarders and rigid heuristic engines into capable, context-aware participants operating directly within physical environments.

Decentralising generative inference directly addresses the inherent structural limitations of cloud-centric AI. Routing high-frequency sensory data across wide-area networks introduces transmission latency, exposes sensitive raw observations to external interception, consumes unsustainable network bandwidth, and leaves mission-critical field operations vulnerable to unexpected network failures.

By migrating multi-modal generative inference pipelines to local execution environments, GenAIoT applications unlock deterministic real-time responsiveness, preserve operational continuity during network disconnections, maintain strict on-premises data residency, and eliminate compounding network egress expenses from continuous cloud telemetry streaming.

GenAIoT fundamentally redefines the relationship between digital intelligence and physical actuation through embodied edge AI. In legacy systems, field devices merely perceived their surroundings and responded via pre-programmed, single-step scripts.

With generative reasoning running locally, embodied autonomous systems and robotic platforms can now perform complex, multi-step spatial and logical reasoning in real time. Rather than relying on cloud communication loops, edge nodes can plan complex task chains, continuously evaluate their physical environments, and adapt dynamically to unforeseen operational disruptions.

Furthermore, integrating real-time virtual simulations, DTs, and immersive operational environments directly into the edge reasoning loop enables embodied systems to test forward scenarios before executing physical actions. By coupling localised generative reasoning with predictive dynamic modelling, physical actors can evaluate alternate actions, assess operational risks, and optimise their responses in real time.

This transformation elevates field HW from simple, isolated digital monitors to resilient, context-aware physical actors that can collaborate safely alongside human operators in dynamic, unpredictable industrial environments.

11.2 The IoT Edge–Cloud Computing interoperability by means of Data Interoperability

The GenAIoT Edge–Cloud Computing interoperability determines where data processing and intelligence can take place; semantic interoperability helps ensure that information remains meaningful as it moves between those locations.

© AIOTI. All rights reserved.

The GenAIoT interoperability defines that distribution of sensing, processing, storage, and analytics across devices, edge nodes, fog infrastructure, and cloud platforms. The trade-offs between local processing and centralised resources, including latency, bandwidth, energy consumption, and computational capacity.

In the GenAIoT interoperability approach semantic descriptions of data, services, and capabilities that support interoperability between infrastructure layers exist across the computing and data continuum.

The role of standards-based architectures in supporting distributed resources and interoperable services. The architectural foundation is supported by the OpenFog Reference Architecture adopted as IEEE 1934-2018 and by IEEE 1935-2023, which addresses edge/fog manageability and orchestration.

11.3 The Physical Realities of Silicon, Packaging, and System-Level Co-Design

Realising the full promise of decentralised generative intelligence requires confronting the unforgiving thermodynamic and physical limits of semiconductor engineering. Foundation models have historically been developed within the unconstrained computational environments of centralised enterprise data centres.

Deploying these reasoning capabilities onto edge platforms requires addressing a fundamental metric: useful work performed per unit of energy. System viability is no longer measured solely by raw clock speeds or peak theoretical calculations, but by the electrical, thermal, and economic costs of generating each individual reasoning token.

This focus on energy efficiency forces semiconductor designers to pursue deep innovations across design and manufacturing processes. Advances in transistor design, such as nanosheet architectures and extensive design-technology co-optimisation, confirm that designers cannot develop physical transistors in isolation from the SW workloads they execute.

Standard-cell architectures, power-rail distribution networks, and metal interconnect layouts must be tailored to balance performance, power consumption, and silicon area across diverse operating conditions. Because edge AI HW serves diverse roles, from low-power continuous sensing outposts to high-throughput localised inference gateways, silicon implementations cannot follow a single, monolithic blueprint.

At the same time, the physical architecture of modern computing must move beyond single-die processors to overcome the memory wall. In generative inferencing workloads, computing units frequently sit idle, stalled while awaiting data transfers from external storage registers.

To break this throughput bottleneck, advanced multi-chip packaging and three-dimensional die stacking are essential. Integrating high-bandwidth memory directly adjacent to or vertically stacked upon compute logic dramatically expands data throughput while cutting the electrical energy required to move information between registers.

As decentralised generative systems scale to handle multi-modal arrays and distributed collaborative workloads, communications between individual chips, packages, and peripheral gateways emerge as a primary performance limiter. High-bandwidth, low-power interconnects, increasingly incorporating integrated optical connections, are necessary to overcome the latency and thermal penalties inherent to high-frequency electrical wiring.

Consequently, continuous progress in GenAIoT depends on holistic system-level co-design. Raw computational capacity, memory packaging, high-speed networking, thermal dissipation, and

© AIOTI. All rights reserved.

physical power delivery must be engineered as deeply interconnected elements of a unified HW architecture.

11.4 Operational Governance, Autonomous Engineering, and the Future Research Frontier

Introducing generative reasoning across physical devices fundamentally reshapes the operational, safety, and security attack surfaces of connected infrastructure. In purely digital environments, a generative hallucination or probabilistic error produces poor text or an inaccurate image. In embodied edge environments, an ungrounded or non-deterministic output can trigger physical damage, operational downtime, or severe bodily harm.

Addressing these physical safety risks requires engineering architectures that strictly decouple probabilistic reasoning from direct, unverified HW actuation. Generative models must function as planning and diagnostic layers whose proposed actions are continuously validated by deterministic verification filters, formal mathematical monitors, and hard-coded physical limits before reaching physical HW.

Decentralised deployments also introduce severe lifecycle management hurdles. Maintaining consistent security postures, tracking model versions, and preventing behavioural drift across fleets of millions of autonomous field devices introduce unprecedented operational complexity.

GenAIoT applications must deploy robust, zero-trust HW roots of trust, tamper-proof on-chip enclaves, and authenticated OTA deployment pipelines equipped with automatic fail-safes. Furthermore, systems require continuous anomaly-detection frameworks to identify and neutralise physical sensor spoofing, electromagnetic interference, and adversarial manipulations before corrupted inputs reach the core reasoning layers.

Managing this design and operational complexity requires modernising engineering workflows through autonomous, AI-assisted design automation. The intricate trade-offs involved in configuring nanosheet layouts, multi-die advanced packaging, and 3D interconnects quickly surpass traditional human engineering bandwidth.

Deploying specialised design enablement toolkits and autonomous engineering workflows can help optimise physical layouts, detect manufacturing-rule violations, and translate novel packaging concepts into production designs. However, these automated design agents must be rigorously benchmarked against strict physical timing, thermal dissipation, and manufacturing standards to guarantee that resulting designs work reliably in production.

Ultimately, the transition toward decentralised generative intelligence marks a profound turning point in distributed computing and industrial automation. Beyond the initial era of cloud-centric experiments, the discipline's future rests on cross-disciplinary systems engineering that unites model compression, advanced packaging, deterministic control architectures, and resilient HW design.

By building upon these interconnected principles, the broader technology industry can deploy scalable, resilient, and context-aware GenAIoT systems that deliver autonomous intelligence safely and efficiently to the physical edge.

The transition from centralised, cloud-bound generative AI to distributed intelligence across the cloud–edge–IoT continuum represents a structural transformation in how intelligent systems are designed, deployed, and governed. GenAIoT extends connected devices beyond sensing and telemetry towards contextual reasoning, adaptive planning, and physical action. Its significance therefore lies in bringing generative models closer to data sources and enabling embodied, collaborative intelligence under real-world constraints. Realising this potential

© AIOTI. All rights reserved.

requires coordinated advances across algorithms, architectures, HW-SW-AI-Data, safety, governance, and human–machine interaction.

GenAIoT cannot be achieved by transferring existing cloud models and serving frameworks to resource-constrained devices. It requires continuum-native architectures that distribute generative workloads dynamically across micro-, deep-, and meta-edge tiers and cloud backends. Placement, partitioning, and escalation decisions must jointly account for latency, energy, bandwidth, privacy, and operational risk. Distributed inference must also remain interruptible, resumable, and resilient to degraded connectivity. Interoperability, across data formats, semantics, agent interfaces, and access policies, must be an architectural component.

Developing compact, specialised, and physically grounded generative models is key, as hardware-aware quantisation, structured pruning, knowledge distillation, and compression of caches and retrieval indices must advance alongside algorithm–hardware co-design. Efficiency alone is not sufficient, as smaller models must retain the capabilities, uncertainty awareness, and safety constraints needed for their intended environments. Reusable sensor foundation models and locally adapted agents offer a pathway towards intelligence grounded in multimodal observations, device semantics, geometry, and physics. Digital twins, immersive triplets, and generative world models can support anticipatory reasoning and simulation, provided that their alignment with physical reality is continuously assessed.

The most consequential transition is from generating interpretations to generating actions. Once model outputs influence actuators or operational decisions, hallucination, adversarial manipulation, and coordination failures become potential sources of physical harm. Trustworthiness must therefore be engineered as an end-to-end system property. Constrained generation, verifiable planning, classical control, runtime assurance, calibrated refusal, and human override must work together within explicit safety envelopes. Secure model updates, provenance tracking, rollback mechanisms, and fleet-wide lifecycle management are equally necessary for sustained operational reliability.

Decentralisation does not automatically deliver privacy, sovereignty, or accountability. These outcomes depend on policy-aware context access, consent management, auditable data use, and clearly assigned responsibilities across distributed actors. System design must align with data-protection requirements, AI regulation, cybersecurity, and sector-specific safety standards from the outset. Human–machine teaming is similarly fundamental: operators and affected users need appropriate explanations, meaningful opportunities to contest decisions, and effective control when generated actions affect safety, rights, or critical infrastructure.

Progress in developing GenAIoT technologies is measured through shared evaluation frameworks that combine generative quality with physical-task success, energy consumption, latency, robustness, security, and human-team performance. Research communities, industry consortia, and standards bodies must jointly establish these benchmarks and the interoperable infrastructure needed to support them. The defining challenge is not whether generative models can execute at the edge, but whether distributed intelligence can remain useful, safe, governable, and genuinely decentralised at scale. Addressing that challenge will determine whether GenAIoT becomes a dependable foundation for autonomous physical systems or remains a collection of fragmented demonstrations.

© AIOTI. All rights reserved.

12 References

[1] M. Merenda, C. Porcaro, and D. Iero, "Edge Machine Learning for AI-Enabled IoT Devices: A Review," *Sensors*, vol. 20, no. 9, p. 2533, Apr. 2020. https://doi.org/10.3390/s20092533.

[2] Y. LeCun, "A Path Towards Autonomous Machine Intelligence: Version 0.9.2," OpenReview, Tech. Rep., Jun. 2022. https://openreview.net/pdf?id=BZ5a1r-kVsf.

[3] A. Brohan *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Proc. 7th Conf. Robot Learn. (CoRL)*, vol. 229, 2023, pp. 2165–2183.

[4] P. Budzianowski *et al.*, "EdgeVLA: Efficient Vision-Language-Action Models," 2025, *arXiv:2507.14049*. https://arxiv.org/abs/2507.14049.

[5] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A Survey on Vision-Language-Action Models for Embodied AI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 37, pp. 3031–3051, 2024.

[6] European Commission, "aerOS: Autonomous, Hierarchical and Reliable Operating System for the IoT-Edge-Cloud Continuum," Horizon Europe Research and Innovation Programme, Grant Agreement 101069732, 2022. https://aeros-project.eu.

[7] European Commission, "CODECO: Cognitive Decentralized Edge-to-Cloud Orchestration Framework," Horizon Europe Research and Innovation Programme, Grant Agreement 101092582, 2023. https://he-codeco.eu.

[8] European Commission, "OPEN-CONTINUUM: Open Cloud-Edge-IoT Initiative (O-CEI / EUCloudEdgeIoT)," Horizon Europe Coordination and Support Action (CSA), Grant Agreement 101070038, 2022. https://eucloudedgeiot.eu.

[9] J. Achiam *et al.*, "GPT-4 Technical Report," 2023, *arXiv:2303.08774*. https://arxiv.org/abs/2303.08774.

[10] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023, *arXiv:2307.09288*. https://arxiv.org/abs/2307.09288.

[11] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," 2024, *arXiv:2407.21783*. https://arxiv.org/abs/2407.21783.

[12] European Commission, "Smart Networks and Services Joint Undertaking (SNS JU): 6G-SANDBOX Project," Horizon Europe Research and Innovation Programme, Grant Agreement 101096328, 2023. https://6g-sandbox.eu/.

[13] O. Vermesan and V. Frascolla, Eds., "Edge IoT Industrial Immersive Technologies and Spatial Computing Continuum," 1st ed., AIOTI, Tech. Rep., 2024. https://aioti.eu/wp-content/uploads/AIOTI-Paper-Edge-AI-IoT-Immersive-Technologies-Published.pdf.

[14] W. Shi *et al.*,"Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, Oct. 2016. https://doi.org/10.1109/JIOT.2016.2579198.

[15] K. Kawaharazuka *et al.*, "Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications," *IEEE Access*, 2025. https://ieeexplore.ieee.org/document/10850388.

[16] V. Karagiannis, "Data Sovereignty and Compliance in the Computing Continuum," in *Proc. Int. Conf. Future Internet of Things and Cloud (FiCloud)*, 2024, pp. 123–130.

[17] 17 E. Jordan, M. Serrano, A. Gyrard, E. Song, T. Roth and D. Wollman, "Semantics for Enhancing Communications- and Edge-Intelligence-enabled Smart Sensors: A Practical Use Case in Federated Automotive Diagnostics," IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society, Chicago, IL, USA, 2024, pp. 1-6, doi: 10.1109/IECON55916.2024.10905565.

[18] 18 M. Serrano, A. Zappa and W. Ashraf, "Understanding the Importance of Data Continuum Enabling the Data Exchange, and Semantic Interoperability with Trust in Future Networks," 2023 IEEE Future Networks World Forum (FNWF), Baltimore, MD, USA, 2023, pp. 1-5, doi: 10.1109/FNWF58287.2023.10520489.

[19] 19 Satyanarayanan, M. (2017). "The Emergence of Edge Computing." Computer, 50(1), 30–39.

[20] 20 Compton, M. et al. "The SSN ontology of the W3C semantic sensor network incubator group." 2012 Web Semantics: Science, Services and Agents on the World Wide Web, 17, 25–32.

© AIOTI. All rights reserved.

[21] 21 W3C. (2017). "Semantic Sensor Network Ontology." Available online: https://www.w3.org/TR/vocab-ssn/
[22] 22 ISO/IEC 21823-3:2021. Internet of Things (IoT) – Interoperability for IoT systems – Part 3: Semantic interoperability. Accessible Online: https://www.iso.org/standard/83752.html
[23] 23 T. Berners-Lee, R. Fielding, and L. Masinter. 2005. RFC 3986: Uniform Resource Identifier (URI): Generic Syntax. RFC Editor, USA.
[24] 24 Quilitz, B., Leser, U.. "Querying Distributed RDF Data Sources with SPARQL". In The Semantic Web: Research and Applications. ESWC 2008 vol 5021. Springer, https://doi.org/10.1007/978-3-540-68234-9_39
[25] 25 SOSA: Spatial Data on the Web WG, https://www.w3.org/2015/spatial/wiki/SOSA_Ontology.
[26] 26 Shi, W., et al. (2016). "Edge Computing: Vision and Challenges." IEEE Internet of Things Journal, 3(5), 637–646.
[27] IEEE, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," Version 2, IEEE, Tech. Rep., 2017, pp. 1–263. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.
[28] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, Nov. 2020. https://doi.org/10.1145/3381831.
[29] Eclipse Foundation, "Eclipse KuDECO," *projects.eclipse.org*, 2026. https://projects.eclipse.org/projects/technology.kudeco. [Accessed: Oct. 19, 2026].
[30] A. Iosifidis, V. Karagiannis, and S. Schulte, "Atlas Synchronization in the Hierarchical Federated Learning Continuum," in *Proc. IEEE 10th Int. Conf. Fog Edge Comput. (ICFEC)*, 2026, pp. 81–88.
[31] V. Karagiannis, D. Ignjatović, A. Iosifidis, and S. Schulte, "Cost-Effective Processing of IoT Data in the Computing Continuum," in *Proc. IEEE 10th Int. Conf. Fog Edge Comput. (ICFEC)*, 2026, pp. 9–16.
[32] M. Merenda, R. Chatila, and J. C. Havens, "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," in *Robotics and Well-Being*, Cham, Switzerland: Springer, 2019, pp. 11–16, doi: 10.1007/978-3-030-12524-0_2.
[33] O. Vermesan, F. Pétrot, M. Coppola, M. Schneider, and A. Höß, "Industrial AI Technologies for Next-Generation Autonomous Operations with Sustainable Performance," in *Intelligent Edge-Embedded Technologies for Digitising Industry*, Aalborg, Denmark: River Publishers, 2022, ch. 1. https://www.riverpublishers.com/pdf/ebook/chapter/RP_9788770226103C1.pdf.
[34] O. Vermesan and V. Frascolla, Eds., "Advancing Next-Generation IoT and Edge Computing Research and Innovation - Strategic Research and Innovation Agenda (SRIA), Release 1," AIOTI, Tech. Rep., Jan. 2023. https://aioti.eu/wp-content/uploads/2023/02/AIOTI-SRIA-2023-Release-1.pdf.
[35] R. C. Sofia *et al.*, "Towards Scalable Federated Container Orchestration: The CODECO Approach," 2026, *arXiv:2601.13351*. https://arxiv.org/abs/2601.13351.
[36] D. Joshi *et al.*, "AI at the Edge: Challenges, Applications, and Directions," in *IoT for Defense and National Security*, Hoboken, NJ, USA: Wiley-IEEE Press, 2023, pp. 133–160, doi: 10.1002/9781119892199.ch9.
[37] A. Munir *et al.*, "Artificial Intelligence and Data Fusion at the Edge," *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 7, pp. 62–78, Jul. 2021, doi: 10.1109/MAES.2020.3043072.
[38] Y. Matsubara, M. Levorato, and F. Restuccia, "Split Computing and Early Exiting for Deep Learning Applications: Survey and Research Challenges," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–30, Dec. 2022, doi: 10.1145/3527155.
[39] Y. Leviathan, M. Kalman, and Y. Matias, "Fast Inference from Transformers via Speculative Decoding," in *Proc. 40th Int. Conf. Machine Learn. (ICML)*, PMLR, vol. 202, 2023, pp. 19274–19286.
[40] ETSI, "Multi-Access Edge Computing (MEC); Framework and Reference Architecture," ETSI, Sophia Antipolis, France, Tech. Spec. ETSI GS MEC 003 V3.1.1, Dec. 2022.

© AIOTI. All rights reserved.

https://www.etsi.org/deliver/etsi_gs/MEC/001_099/003/03.01.01_60/gs_MEC003v030101p.pdf.
[41] European Commission, "FluidOS: Transparent, Organised, and Orchestrated Computing Continuum," Horizon Europe Research and Innovation Programme, Grant Agreement 101070222, 2022. https://www.fluidos.eu/.
[42] Eclipse Foundation, "Eclipse Zenoh: Zero Overhead Pub/sub, Shared Memory and Query Protocol," Eclipse Foundation, Tech. Rep., 2023. https://zenoh.io.
[43] Z. Zhou *et al.*, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019, doi: 10.1109/JPROC.2019.2918951.
[44] JE. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022. https://openreview.net/forum?id=nZeVKeeFYf9.
[45] *Function Blocks - Part 1: Architecture*, IEC Standard 61499-1:2012, International Electrotechnical Commission (IEC), Geneva, Switzerland, 2012.
[46] V. Vyatkin, "IEC 61499 as enabler of distributed and intelligent automation: State-of-the-art review," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 4, pp. 768–781, Nov. 2011, doi: 10.1109/TII.2011.2166785.
[47] Universal Automation Association (UAA), "Universal Automation Association Portal," 2026. https://universalautomation.org/.
[48] T. Lyu, U. D. Atmojo, and V. Vyatkin, "Complexity-aware distribution optimization for computing-intensive distributed automation systems with IEC 61499," in *Proc. IEEE Int. Conf. Industrial Informatics (INDIN)*, 2025, doi: 10.1109/INDIN64977.2025.11279613.
[49] T. Lyu, U. D. Atmojo, and V. Vyatkin, "Towards cloud-based virtual commissioning of distributed automation applications with IEC 61499 and containerization technology," in *Proc. 47th Annu. Conf. IEEE Industrial Electronics Society (IECON)*, 2021, doi: 10.1109/IECON48115.2021.9589945.
[50] W. Su *et al.*, "Cloud-edge computing-based ICICOS framework for industrial automation and artificial intelligence: A survey," *Journal of Circuits, Systems and Computers*, vol. 32, no. 10, Art. no. 2350168, Jul. 2023, doi: 10.1142/S021812662350168X.
[51] Y.-C. Wang, J. Xue, C. Wei, and C.-C. J. Kuo, "An overview on generative AI at scale with edge-cloud computing," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 2952–2971, 2023, doi: 10.1109/OJCOMS.2023.3320646
[52] Y. Zheng *et al.*, "A review on edge large language models: Design, execution, and applications," *ACM Computing Surveys*, vol. 57, no. 8, 2025, doi: 10.1145/3719664.
[53] R. Chatila and J. C. Havens, "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," in *Robotics and Well-Being*, M. I. A. Ferreira et al., Eds. Cham, Switzerland: Springer, 2019, pp. 11–16. Available: https://doi.org/10.1007/978-3-030-12524-0_2.
[54] R. C. Sofia et al., "Towards Scalable Federated Container Orchestration: The CODECO Approach," arXiv preprint arXiv:2601.13351, Jan. 2026.
[55] R. Morabito and S. Jang, "Smaller, Smarter, Closer: The Edge of Collaborative Generative AI," IEEE Internet Computing, vol. 29, no. 3, pp. 24–32, May-Jun. 2025. doi: 10.1109/MIC.2025.11022747.
[56] D. Ali and R. C. Sofia, "Experimenting with Energy-Awareness in Edge-Cloud Containerized Application Orchestration," arXiv preprint arXiv:2511.09116, Nov. 2025.
[57] Y. S. Patel, A. Choubey, A. Singh, and P. Townend, "Decentralized Multi-Agent Reinforcement Learning for the Green Serverless Cloud-Edge Continuum," in Proceedings of the 19th IEEE International Conference on Service-Oriented System Engineering (SOSE 2025), 2025, pp. 108–117, doi: 10.1109/SOSE67019.2025.00017.
[58] Z. Su, M. T. Islam, M. Goudarzi, and A. N. Toosi, "LLM-Driven Intent-Based Privacy-Aware Orchestration Across the Cloud-Edge Continuum," in Proceedings of the 2026 Australasian Computer Science Week (ACSW 2026), 2026, pp. 32–41. doi: 10.1145/3708420.3708442.
[59] Y. Sun, S. Guo, S. Huang, Z. Feng, F. Qi, and X. Qiu, "Generative Intent Prediction Agentic AI Empowered Edge Service Function Chain Orchestration," IEEE Transactions on Cognitive Communications and Networking, to be published. doi: 10.1109/TCCN.2026.3682482.

© AIOTI. All rights reserved.

[60] V. E. Bojnurdi, T. Lyu, U. D. Atmojo, and V. Vyatkin, "Optimal Function Block Distribution for IEC 61499 Distributed Control Application," in Proc. IEEE 3rd Ind. Electron. Soc. Annu. On-Line Conf. (ONCON), Beijing, China, 2024, pp. 1–6, doi: 10.1109/ONCON62778.2024.10931597.
[61] B. Yang *et al.*, "An efficient edge-cloud collaboration system with foundational models for open-set IoT applications," *IEEE Transactions on Mobile Computing*, vol. 25, no. 7, pp. 10740–10756, 2026, doi: 10.1109/TMC.2026.3666311.
[62] Z. Zhou, C. Liu, and K. Li, "Energy-Efficient Joint Task Offloading and Network Resource Allocation in Mobile Edge-Cloud Computing," *IEEE Transactions on Cloud Computing*, vol. 12, no. 1, pp. 182–195, Jan.–March 2024, doi: 10.1109/TCC.2023.3328912.
[63] J. Pei *et al.*, "Distributed Large Models Training Optimization with Real-Time Wireless Channel Feedback," in *IEEE Journal on Selected Areas in Communications*, vol. 44, pp. 2231-2243, 2026, doi: 10.1109/JSAC.2025.3640136.
[64] Z. Bencherki *et al.*, "The Landscape of IEC 61499 in Modern Automation: A Scoping Review of Research Trends and Gaps (2018–2025)," *Journal of Process Control*, vol. 162, Art. no. 103726, 2026, doi: 10.1016/j.jprocont.2026.103726.
[65] J.P. Morgan, "Overdrawn at the Memory Bank: How AI is Creating Memory Shortage, and What Happens Next," J.P. Morgan Global Research, Tech. Rep., Mar. 2024. https://www.jpmorgan.com/insights/global-research/artificial-intelligence/dram-memory-shortage-from-ai.
[66] S. S. Girija *et al.*, "Optimizing LLMs for resource-constrained environments: A survey of model compression techniques," in Proc. IEEE Int. Comput. Softw. Appl. Conf. (COMPSAC), 2025
[67] R. Gao, L. Zhang, S. Yue, and Z. Gao, "Empowering large language models to edge intelligence: A survey of edge efficient LLMs and techniques," *Computer Science Review*, vol. 57, Art. no. 100755, 2025, doi: 10.1016/j.cosrev.2025.100755.
[68] J. Chee, Y. Cai, V. Kuleshov, and C. De Sa, "QuIP: 2-bit quantization of large language models with guarantees," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 4396–4429.
[69] S. Sharify *et al.*, "Post training quantization of large language models with microscaling formats," in *Proc. 4th NeurIPS Efficient Natural Language and Speech Processing Workshop*, PMLR, vol. 262, 2024, pp. 241–255.
[70] JJ. Wang *et al.*, "The role of data spaces in the Internet of Things," *e+i Elektrotechnik und Informationstechnik*, 2025, doi: 10.1007/s00502-025-1082-x.
[71] L. Cao *et al.*, "Bitnet.cpp: Efficient edge inference for ternary LLMs," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1, 2025, pp. 9305–9322, doi: 10.18653/v1/2025.acl-long.457.
[72] S. Muralidharan *et al.*, "Compact language models via pruning and knowledge distillation," 2024, *arXiv:2407.14679*. https://arxiv.org/abs/2407.14679.
[73] J. Sander *et al.*, "Constrained edge AI deployment: Fine-tuning vs distillation for LLM compression," 2025, *arXiv:2505.18166*. https://arxiv.org/abs/2505.18166.
[74] H. Li *et al.*, "A survey on large language model acceleration based on KV cache," 2024, *arXiv:2412.19442*. https://arxiv.org/abs/2412.19442.
[75] Y. Liu *et al.*, "KV cache compression for inference efficiency in LLMs: A review," *arXiv preprint arXiv:2508.06297*, 2025.
[76] A. Chen *et al.*, "The pitfalls of KV cache compression," in *Proc. 64th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2026, pp. 41530–41553. doi: 10.18653/v1/2026.acl-long.1926.
[77] Y. Liang, H. Shi, H. Shao, e Z. Wang, "AccLLM: Accelerating Long-Context LLM Inference Via Algorithm-Hardware Co-Design," *arXiv preprint arXiv:2505.03745*, 2025.
[78] Google AI Edge, "LiteRT for Microcontrollers | Google AI Edge," 2026. https://developers.google.com/edge/litert/microcontrollers/overview IEEE, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," Version 2, IEEE, Tech. Rep., 2017, pp. 1–263. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

© AIOTI. All rights reserved.

[79] STMicroelectronics, "STM32Cube AI Studio - The next-gen stand-alone tool for embedded AI," 2026. https://www.st.com/content/st_com/en/campaigns/edge-ai-toolchain-for-mcus-z14.html. [Accessed: Oct. 09, 2026].
[80] NXP Semiconductors, "eIQ® AI Software Development Environment," 2026. https://www.nxp.com/design/design-center/software/eiq-ai-development-environment:EIQ. [Accessed: Oct. 09, 2026].
[81] AIOTI WG Standardisation, "Guidance for the Integration of IoT and Edge Computing in Data Spaces," Alliance for Internet of Things Innovation (AIOTI), White Paper, Sep. 2022. https://aioti.eu/wp-content/uploads/2022/09/AIOTI-Guidance-for-IoT-Integration-in-Data-Spaces-Final.pdf. [Accessed: Oct. 09, 2026].
[82] Google Cloud, "What is Retrieval-Augmented Generation (RAG)?," 2026. https://cloud.google.com/use-cases/retrieval-augmented-generation. [Accessed: Oct. 09, 2026].
[83] International Data Spaces Association (IDSA), "International Data Spaces Association," 2026. https://internationaldataspaces.org/. [Accessed: Oct. 09, 2026].
[84] NTT DATA Group Corporation, "NTT DATA | Trusted Global Innovator," 2026. https://www.nttdata.com/global/en. [Accessed: Oct. 09, 2026].
[85] Data Spaces Support Centre (DSSC), "Data Spaces Support Centre," 2026. https://dssc.eu/. [Accessed: Oct. 09, 2026].
[86] Gaia-X AISBL, "Gaia-X: A Federated Secure Data Infrastructure," 2026. https://gaia-x.eu/. [Accessed: Oct. 09, 2026].
[87] Big Data Value Association (BDVA), "Innovation data spaces (i-Spaces)," 2026. https://bdva.eu/what-are-i-spaces/". [Accesses Oct. 20, 2026].
[88] R. Iannella and S. Villata, Eds., "ODRL Information Model 2.2," W3C Recommendation, Feb. 15, 2018. https://www.w3.org/TR/odrl-model/. [Accessed: Oct. 09, 2026].
[89] P. Groth and L. Moreau, Eds., "PROV-Overview: An Overview of the PROV Family of Documents," W3C Working Group Note, Apr. 30, 2013. https://www.w3.org/TR/prov-overview/. [Accessed: Oct. 09, 2026].
[90] European Health and Digital Executive Agency (HaDEA), "Data Space for Manufacturing," European Commission, Brussels, Belgium, Call Document DIGITAL-2026-DSM-AI-09, 2025. https://hadea.ec.europa.eu/document/download/3dad4bf6-864e-4193-a159-517f8838b685_en?filename=DIGITAL-2026-DSM-AI-09+-+Data+Space+for+Manufacturing.pdf. [Accessed: Oct. 09, 2026].
[91] Actian Corporation, "GenAI at the Edge: The Power of TinyML and Embedded Databases," *Actian Blog*, 2026. https://www.actian.com/blog/databases/genai-at-the-edge-the-power-of-tinyml-and-embedded-databases/. [Accessed: Oct. 09, 2026].
[92] Y. Liang *et al.*, "Foundation Models for Time Series Analysis: A Tutorial and Survey," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Barcelona, Spain, 2024, pp. 6555–6565. doi: 10.1145/3637528.3671451.
[93] Y. Mao *et al.*, "A Survey on Spatio-Temporal Prediction: From Transformers to Foundation Models," *ACM Comput. Surv.*, vol. 58, no. 4, Art. no. 44, Sep. 2025. doi: 10.1145/3766546.
[94] S. Li and H. Tang, "Multimodal Alignment and Fusion: A Survey," *Int. J. Comput. Vis.*, vol. 134, no. 3, pp. 412–435, Feb. 2026. doi: 10.1007/s11263-025-02104-z.
[95] B. Tekeste *et al.*, "Adversarial Machine Learning: A 20-Year Survey of Attacks, Defenses, and Standards," *IEEE Access*, vol. 14, pp. 69778–69812, 2026. doi: 10.1109/ACCESS.2026.3690620.
[96] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, 2014, pp. 1–10.
[97] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–11.
[98] L. Bailey, E. Ong, S. Russell, and S. Emmons, "Image Hijacks: Adversarial Images Can Control Generative Models at Runtime," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, ser. Proceedings of Machine Learning Research, vol. 235, Vienna, Austria, 2024, pp. 2443–2455.
[99] Y. Mao *et al.*, "Green Edge AI: A Contemporary Survey," *Proc. IEEE*, vol. 112, no. 7, pp. 880–911, Jul. 2024. doi: 10.1109/JPROC.2024.3437365.

© AIOTI. All rights reserved.

[100] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," IEEE Access, vol. 7, pp. 47230–47244, 2019, doi: 10.1109/ACCESS.2019.2909068.
[101] F. Tramèr *et al.*, "Stealing Machine Learning Models via Prediction APIs," in *Proc. 25th USENIX Secur. Symp. (USENIX Security)*, Austin, TX, USA, 2016, pp. 601–618.
[102] W. Hua, Z. Zhang, and G. E. Suh, "Reverse Engineering Convolutional Neural Networks Through Side-Channel Information Leaks," in *Proc. 55th ACM/ESDA/IEEE Des. Autom. Conf. (DAC)*, San Francisco, CA, USA, 2018, pp. 1–6. doi: 10.1109/DAC.2018.8465773.
[103] K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 7167–7177.
[104] N. Carlini and D. Wagner, "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods," in *Proc. 10th ACM Workshop Artif. Intell. Secur. (AISec)*, Dallas, TX, USA, 2017, pp. 3–14. doi: 10.1145/3128572.3140444.
[105] Y. Li *et al.*, "Efficient and Secure Deep Learning Inference in Trusted Processor Enabled Edge Clouds," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 10, pp. 4311–4325, Oct. 2022. doi: 10.1109/TPDS.2022.3187772.
[106] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing Federated Learning through an Adversarial Lens," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, ser. Proceedings of Machine Learning Research, vol. 97, Long Beach, CA, USA, 2019, pp. 634–643.
[107] V. Mothukuri *et al.*, "A Survey on Security and Privacy of Federated Learning," *Future Gener. Comput. Syst.*, vol. 115, pp. 619–640, Feb. 2021. doi: 10.1016/j.future.2020.10.007.
[108] X. Wang *et al.*, "Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models," *ACM Comput. Surv.*, vol. 57, no. 9, Art. no. 3, Mar. 2025. doi: 10.1145/3724420.
[109] M. Xu *et al.*, "Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1127–1170, Second Quarter 2024. doi: 10.1109/COMST.2024.3353265.
[110] J. Tang *et al.*, "Toward General Industrial Intelligence: A Survey of Large Models as a Service in Industrial IoT," *IEEE Commun. Surveys Tuts.*, vol. 28, no. 3, pp. 2054–2086, Third Quarter 2025.
[111] European Commission High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," European Commission, Brussels, Belgium, Tech. Rep., Apr. 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[112] P. Singh, A. U. Haq, and M. Presser, "Towards a Responsible AI Adoption/Adaptation (RAA) Ecosystem: Vision and Model to Keep Socio-Technological Balance," in *Global Internet of Things and Edge Computing Summit - 2nd International Summit, GIECS 2025, Proceedings*, ser. Communications in Computer and Information Science (CCIS), vol. 2221. Cham, Switzerland: Springer, 2025, pp. 71–92. doi: 10.1007/978-3-032-09555-8_5.
[113] V. Karagiannis *et al.*, "The Role of Data Spaces in the Internet of Things," *Elektrotech. Inf.tech.*, Apr. 2025. doi: 10.1007/s00502-025-01320-6.
[114] V. Karagiannis, "Optimized Deployment of Data Intermediation Services," in *Proc. 21st Int. Conf. Softw. Eng. Adaptive Self-Managing Syst. (SEAMS)*, 2026, pp. 1–7.
[115] V. Karagiannis, A. Al-Akrawi, and O. Hödl, "Data Sovereignty at the Edge of the Network," in *Proc. 7th IEEE Int. Conf. Fog Edge Comput. (ICFEC)*, Bengaluru, India, 2023, pp. 33–39. doi: 10.1109/ICFEC57925.2023.00013.
[116] European Commission, "Supporting and Valuing EU Regulatory Compliance with a Seal of TRUST and Co-designed Online Services based on Trustable LLM (TRUST)," CORDIS EU Funded Projects, Grant Agreement No. 101299091, 2026. https://cordis.europa.eu/project/id/101299091. doi: 10.3030/101299091.
[117] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, OJ L 2024/1689, Jul. 12, 2024. http://data.europa.eu/eli/reg/2024/1689/oj.
[118] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data

© AIOTI. All rights reserved.

Protection Regulation), OJ L 119, 4.5.2016, pp. 1–88. http://data.europa.eu/eli/reg/2016/679/oj.
[119] D. Ignjatović *et al.*, "Crisis Management in the Era of the IoT, Edge Computing, and LLMs," in *Proc. 11th Int. Conf. Internet of Things: Syst. Manage. Security (IOTSMS)*, Malmö, Sweden, 2024, pp. 45–52.
[120] V. Karagiannis *et al.*, "A Review of Emerging Trends in Energy Data Management Systems," in *Proc. 11th Int. Conf. Internet of Things: Syst. Manage. Security (IOTSMS)*, Malmö, Sweden, 2024, pp. 74–81.
[121] *Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on Horizontal Cybersecurity Requirements for Products with Digital Elements (Cyber Resilience Act)*, OJ L 2024/2847, Nov. 20, 2024. http://data.europa.eu/eli/reg/2024/2847/oj.
[122] *Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on Measures for a High Common Level of Cybersecurity Across the Union, Amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and Repealing Directive (EU) 2016/1148 (NIS 2 Directive)*, OJ L 333, Dec. 27, 2022, pp. 80–152. http://data.europa.eu/eli/dir/2022/2555/oj.
[123] *Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on Harmonised Rules on Fair Access to and Use of Data and Amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)*, OJ L 2023/2854, Dec. 22, 2023. http://data.europa.eu/eli/reg/2023/2854/oj.
[124] *Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on Machinery and Repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC*, OJ L 165, Jun. 29, 2023, pp. 1–102. http://data.europa.eu/eli/reg/2023/1230/oj.
[125] European Commission, "Commission Implementing Decision of 22.5.2023 on a Standardisation Request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in Support of Union Policy on Artificial Intelligence," C(2023) 3215 final, May 22, 2023. https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en.
[126] CEN-CENELEC Joint Technical Committee 21, "Artificial Intelligence," 2026. https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/..
[127] J. Soler Garrido *et al.*, "Harmonised Standards for the European AI Act," European Commission, Joint Research Centre, Seville, Spain, Science for Policy Brief JRC139430, Oct. 2024. https://publications.jrc.ec.europa.eu/repository/handle/JRC139430..
[128] R. Kilian, L. Jäck, and D. Ebel, "European AI Standards – Technical Standardisation and Implementation Challenges under the EU AI Act," *Eur. J. Risk Reg.*, vol. 16, no. 3, pp. 1038–1062, 2025. doi: 10.1017/err.2025.10032.
[129] L. Lucaj *et al.*, "TechOps: Technical Documentation Templates for the AI Act," in *Proc. 8th AAAI/ACM Conf. AI, Ethics, Soc. (AIES)*, vol. 8, no. 2, Madrid, Spain, Oct. 2025, pp. 1647–1660. doi: 10.1609/aies.v8i2.36663.
[130] J. Mökander *et al.*, "Auditing Large Language Models: A Three-Layered Approach," *AI Ethics*, vol. 4, no. 4, pp. 1085–1115, Nov. 2024. doi: 10.1007/s43681-023-00289-2.
[131] Regulation (EU) 2024/1781 of the European Parliament and of the Council of 13 June 2024 establishing a framework for the setting of ecodesign requirements for sustainable products, amending Directive (EU) 2020/1828 and Regulation (EU) 2023/1542 and repealing Directive 2009/125/EC (Text with EEA relevance), OJ L, 28.6.2024, ELI: http://data.europa.eu/eli/reg/2024/1781/oj.
[132] IEEE Draft Standard for Digital Product Passport - Reference Architecture and Technical Requirements, IEEE Standard P3828, Sept. 2024. https://standards.ieee.org/ieee/3828/11761.
[133] ACER Recommendation on the Network Code on Demand Response, ACER Recommendation No. 01/2025, Mar. 2025. https://www.acer.europa.eu/sites/default/files/documents/Recommendations/ACER_Recommendation_01-2025_Demand_Response_Network_Code.pdf.

© AIOTI. All rights reserved.

[134] EUCloudEdgeIoT, "EUCloudEdgeIoT: Coordinating the European Cloud-Edge-IoT Continuum Portfolio," European Commission Horizon Europe Initiative, CSA Projects OpenContinuum (Grant No. 101070013) and Unlock-CEI (Grant No. 101070016), 2026. https://eucloudedgeiot.eu.
[135] CEI-Sphere, "Cloud-Edge-IoT Spheres of Large-Scale Pilots (CEI-Sphere)," European Commission Horizon Europe Project, Grant Agreement No. 101189683, Oct. 2024. https://ceisphere.eu.
[136] Eclipse Foundation, "Eclipse Continuum Interest Group," projects.eclipse.org, 2026. https://projects.eclipse.org/interest-groups/eclipse-continuum.
[137] P. Guldimann *et al.*, "COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act," *arXiv preprint arXiv:2410.07959*, Oct. 2024.
[138] European Commission, "Certification for Ethical and Regulatory Transparency in Artificial Intelligence (CERTAIN)," CORDIS EU Funded Projects, Grant Agreement No. 101189650, 2025. https://cordis.europa.eu/project/id/101189650. doi: 10.3030/101189650.
[139] European Commission, "Fostering Artificial Intelligence Trust for Humans towards the Optimization of Trustworthiness through Large-Scale Pilots in Critical Domains (FAITH)," CORDIS EU Funded Projects, Grant Agreement No. 101135932, 2024. https://cordis.europa.eu/project/id/101135932. doi: 10.3030/101135932.
[140] European Commission, "An AI-Powered, Human-Validated Standardisation Landscape Hub (StandardAIse)," CORDIS EU Funded Projects, Grant Agreement No. 101295503, 2026. https://cordis.europa.eu/project/id/101295503. doi: 10.3030/101295503.
[141] M. Penica, M. Bhattacharya, A. Dooley, and E. O'Connell, "Small Language Models Are Agentic but Not Autonomous in Greenhouse Climate Control," Scientific Reports, 2026, to be published, doi: 10.1038/s41598-026-78255-1.
[142] M. Penica, M. Bhattacharya, H. Du, and E. O'Connell, "Uncertainty-Gated Agentic Language Models for Physical Control: Safety and Latency Evaluation on Edge Hardware," in Proc. 2026 Int. Conf. Large Language Models (LLM 2026), Las Vegas, NV, USA, Dec. 2026, to be published.
[143] L. Sha, "Using Simplicity to Control Complexity," IEEE Software, vol. 18, no. 4, pp. 20–28, Jul. 2001. https://doi.org/10.1109/MS.2001.936213.

© AIOTI. All rights reserved.

13 Contributors

Editors:

Ovidiu Vermesan, SINTEF, Norway

Valerio Frascolla, Intel, Germany

Reviewer:

Valentina Peniche, AIOTI Secretary General

Contributors (alphabetic order):

Amelie Gyrard, Trialog, France

Antonio Kung, Trialog, France

Antonio Ruiz-Alba, ITI, Spain

David Sarabia, i2CAT, Spain

Elena Gaura, Coventry University, UK

Flemming Sveen, Hafenstrom, Norway

George Suciu, BEIA, Romania

Javier Coronel, ITI, Spain

Joan Valls, ITI, Spain

Kashif Rabbani, Digital Catapult, UK

Luis Miguel Bartolín Arnau, ITI, Spain

Martin Serrano, RINN-AI University of Galway, Ireland

Mihai Penica, University of Limerick, Ireland

Monica Florea, SIMAVI, Romania

Natalie Samovich, Enercoutim, Portugal

Nataša Krčo, DunavNET, Serbia

Ovidiu Vermesan, SINTEF, Norway

Parwinder Singh, Aarhus Universitet, Denmark

Pierre-Yves Danet, 6G IA, Belgium

Ramona Marfievici, Digital Catapult, UK

Roy Bahr, SINTEF, Norway

Rute Sofia, FORTISS, Germany

© AIOTI. All rights reserved.

Salvador Santonja Climent, ITI, Spain

Srdjan Krčo, DunavNET, Serbia

Udayanto Dwi Atmojo, Aalto University, Finland

Valerio Frascolla, Intel, Germany

Vasileios Karagiannis, Austrian Institute of Technology, Austria

© AIOTI. All rights reserved.

14 Disclaimer

The editors would like to thank all the contributors for their support in the planning, preparation, and drafting of this position paper. The recommendations and opinions expressed in the paper are those of the editors, authors, and contributors and do not necessarily reflect those of the AIOTI, EU, any organisations, employers, or companies. Several images in this publication were generated using GenAI tools. The tools were utilised to improve visual readability. The images serve an illustrative/decorative purpose and does not represent primary, empirical research data.

© AIOTI. All rights reserved.

15 Acknowledgements

All rights reserved, Alliance for AI, IoT and Edge Continuum Innovation (AIOTI). The content of this document is provided 'as-is' and for general information purposes only; it does not constitute strategic or any other professional advice. The content or parts thereof may not be complete, accurate or up to date. Notwithstanding anything contained in this document, AIOTI disclaims responsibility (including where AIOTI or any of its officers, members or contractors have been negligent) for any direct or indirect loss, damage, claim, or liability any person, company, organisation or other entity or body may incur as a result, this to the maximum extent permitted by law.

This paper was prepared by its author(s) and contributor(s) in their personal capacity. The opinions expressed in this paper are the individual author(s) and contributor(s) own and do not reflect the view of AIOTI or its members. The statements, opinions and data contained in this paper and the presentation of materials therein do not imply the expression of any opinion whatsoever by AIOTI or its members. AIOTI hold no responsibility for any breach or damage to people or property resulting from any ideas, methods, instructions, or products referred to in this paper.

16 About AIOTI

AIOTI is the multi-stakeholder platform for stimulating AI, IoT and Edge Computing Innovation in Europe, bringing together small and large companies, academia, policy makers and end-users and representatives of society in an end-to-end approach. We work with partners in a global context. We strive to leverage, share, and promote best practices in the IoT and Edge Computing ecosystems, be a one-stop point of information on all relevant aspects of IoT Innovation to its members while proactively addressing key issues and roadblocks for economic growth, acceptance, and adoption of AI, IoT and Edge Computing Innovation in society. AIOTI's contribution goes beyond technology and addresses horizontal elements across application domains, such as matchmaking and stimulating cooperation in AI, IoT and Edge Computing ecosystems, creating joint research roadmaps, driving convergence of standards and interoperability, and defining policies.

© AIOTI. All rights reserved.



Alliance for IoT, AI and Edge Continuum Innovation

2026