Towards zero-touch autonomy: A vision and agentic GenAI architecture for intent-based self-configuration

Yosra Njah[1], Wael Jaafar[1], Rami Langar[1,2], Bassant Selim[1], Mohamed Cheriet[1]

[1] École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada
[2] LIGM-CNRS UMR 8049, University Gustave Eiffel, F-77420 Marne-la-Vallée, France

Corresponding author: Yosra Njah, yosra.njah@etsmtl.ca

Abstract - Next-generation networks are evolving toward zero-touch autonomy to manage increasing architectural complexity and dynamic service demands. Intent-Based Networking (IBN) has emerged as a key paradigm in this transition, enabling high-level business objectives, or intents, to be automatically orchestrated across heterogeneous resources. This paper presents a vision, reference architecture, and research roadmap for agentic Generative Artificial Intelligence (GenAI)-driven intent-based self-configuration in autonomous networks, positioning IBN as the core of self-configuration upon which higher-order self-X capabilities, such as self-optimization, self-healing, and self-protection, are built. Building on an analysis of existing IBN frameworks across the intent lifecycle, we identify persistent challenges in intent-policy translation, conflict-free activation, and closed-loop assurance across multi-domain, multi-stakeholder environments, and argue that current IBN implementations lack the dynamic cognitive reasoning required to cope with the scale and heterogeneity of the 6G era. To bridge this gap, we introduce an end-to-end agentic GenAI-driven IBN reference architecture leveraging specialized Large Language Model (LLM) agents that orchestrate intent fulfillment through Knowledge Graphs (KGs) and sufficiency-aware Retrieval-Augmented Generation (RAG-S). The architecture is instantiated within a 6G Open Radio Access Network (O-RAN) ecosystem and illustrated through an Industry 5.0 use case, advancing IBN toward practical, reasoning-aware autonomy and outlining key research challenges and future directions toward trustworthy zero-touch autonomous networks.

Keywords: Autonomous networks, generative artificial intelligence, intent-based networking, large language models, zero-touch networks

1. INTRODUCTION

Next-generation networks are undergoing a fundamental architectural transformation, moving beyond 5G toward 6G's vision of integrated, global hyperconnectivity. This shift is propelled by converging pressures: (1) the stringent performance requirements of emerging digital ecosystems, such as Industry 5.0, smart cities, and digital healthcare, demanding ultra-high bandwidth, extreme latency sensitivity, high reliability, and real-time adaptability across heterogeneous environments; (2) the exploding complexity of managing integrated, multi-layered infrastructures, particularly with the advent of Space-Air-Ground Integrated Networks (SAGIN). This complexity is fueled by the convergence of terrestrial, non-terrestrial (e.g., satellite, aerial), and dense cyber-physical systems, alongside the proliferation of heterogeneous Internet-of-Everything (IoE) devices and distributed edge/cloud resources, introducing unprecedented management challenges.

These pressures, coupled with stringent global sustainability mandates, expose the fundamental inadequacy of existing network architectures, reactive performance analysis, and manual, human-driven management. Industry reports indicate that over 95% of configuration and orchestration tasks still depend on manual expertise, leading to operational inefficiency, delayed responses, and limited robustness [1]. Collectively, these limitations create an urgent need for new paradigms capable of enabling intelligent and autonomous network management.

In this context, *Autonomous Networks* (ANs) have emerged as the foundation of Zero-Touch Networks (ZTNs), enabling self-managing systems that reduce human intervention while improving efficiency, service quality, and cost-effectiveness. Built on core self-X principles (i.e., self-configuration, self-optimization, self-healing, and self-protection), ANs ensure reliability, resilience, and security under dynamic conditions. The strategic importance of ANs is further underscored by their alignment with major industrial transformations, including the shift toward human-centric and sustainable Industry 5.0. From an economic perspective, industry analysis estimates that Communications Service Providers (CSPs) could achieve annual savings of approximately $300M in CAPEX and $350M in OPEX, along with a revenue increase of about $144M. Together, these benefits represent nearly $794M per year, around 5% of total CSP revenues, highlighting network autonomy as a critical business imperative [2].

The aspiration to automate network management is not new. It reflects a decades-long evolution spanning Policy-Based Network Management (PBNM) in the late 1990s, automatic computing and networking in the early 2000s, most notably IBM's Monitor-Analyze-Plan-Execute over a shared knowledge (MAPE-K) model, and, more recently, Zero-Touch Network and Service Management (ZSM) [3]. This evolution is underpinned by a software-driven foundation comprising programmable networking, virtualization, and cloud-native architectures, complemented by a cognitive layer powered by Artificial Intelligence (AI). While the former provides the flexibility and programmability required for automation, the latter enables adaptive intelligence and autonomous decision-making. Nevertheless, achieving true zero-touch autonomy remains an open challenge.

Intent-Based Networking (IBN) represents the latest evolution of this trajectory, shifting network management from low-level configuration ("how") to high-level, declarative objectives ("what"). An *intent* specifies desired network states or service goals and is translated into enforceable policies for activation and assurance through Closed-Control Loops (CCLs) [4]. IBN has recently attracted significant attention from both academia and standardization bodies, including the 3rd Generation Partnership Project (3GPP), the Internet Engineering Task Force (IETF), ETSI, and the TM Forum [4]. However, existing efforts remain fragmented, addressing only partial aspects of the IBN workflow rather than orchestrating the full intent lifecycle, from high-level intent expression to continuous assurance and refinement. This limitation is particularly critical in complex, multi-domain, and multi-stakeholder deployments, manifesting in core challenges such as accurately translating intent across heterogeneous domains, resolving dynamic conflicts among competing intents, and ensuring robust closed-loop assurance with adaptive refinement. These challenges highlight the need for an intelligent framework that unifies capabilities into a coherent End-to-End (E2E) management process.

Our analysis of the state of the art indicates that bridging this gap requires a fundamental transition from conventional automation toward cognitive capabilities, such as semantic understanding, contextual reasoning, and adaptive generation. Recent advances in Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), offer transformative capabilities to realize this cognitive shift, as their ability to process unstructured language, reason over complex knowledge, and generate structured outputs aligns naturally with intent fulfillment. However, a single, monolithic LLM is insufficient to handle the distributed, hierarchical, and continuously evolving nature of the intent lifecycle.

Building on this synthesis, we derive a vision, reference architecture, and research roadmap for agentic, E2E GenAI-driven IBN, in which specialized LLM agents collaboratively orchestrate the complete intent lifecycle within a scalable and trustworthy cognitive framework. The proposed framework employs four coordinated LLM-based agents: (1) a cross-domain semantic intent orchestration agent, (2) a multi-layered translation and refinement agent, (3) a dual-Knowledge Graphs (KG) conflict-aware activation agent, and (4) a key value and performance indicator (KVI/KPI)-driven assurance agent operating within CCLs. Each agent leverages adaptive learning strategies, such as fine-tuning and few-shot learning, while grounding real-time reasoning through a sufficiency-aware Retrieval-Augmented Generation (RAG-S) mechanism and structured knowledge via KGs, following recent KG-centric 6G design methodologies [5].

The architecture is instantiated within a 6G Open Radio Access Network (O-RAN) ecosystem and illustrated through an Industry 5.0 use case. We further outline key research challenges and future directions toward scalable, trustworthy, and explainable agentic GenAI-driven autonomy in 6G networks.

The remainder of the paper is organized as follows. Section 2 introduces foundational concepts of ANs and their self-X capabilities. Section 3 analyzes state-of-the-art IBN frameworks across the intent lifecycle and identifies their limitations. Section 4 presents the proposed agentic GenAI-driven IBN architecture and discusses key research challenges and future directions. Section 5 concludes the paper.

2. AUTONOMOUS NETWORKS: TOWARDS ZERO-TOUCH MANAGEMENT & ORCHESTRATION

ANs are cognitive, self-managing systems designed to operate with minimal human intervention. This zero-touch paradigm shifts network operations from static, rule-based processes to dynamic, intent-based automation, addressing the complexity of next-generation networks.

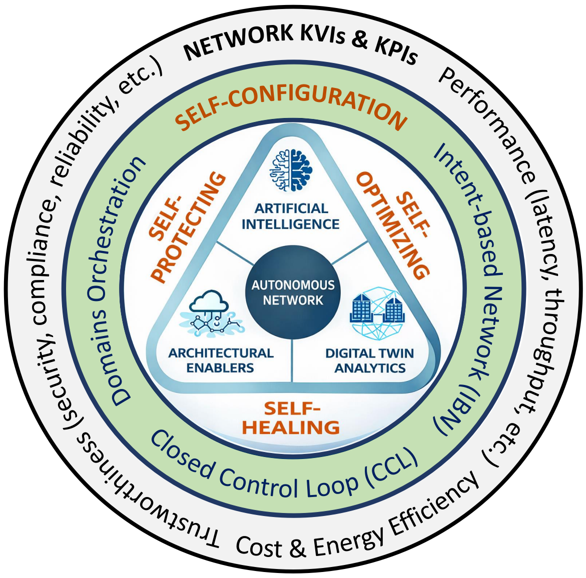


Figure 1 – Multi-layer Autonomous Network (AN) architecture: architectural enablers and cognitive technologies form the inner core, driving surrounding self-X capabilities through intent-based orchestration, while overall network performance and business objectives are evaluated through Key Performance Indicators (KPIs) and Key Value Indicators (KVIs).

Fig.1 illustrates a conceptual multi-layer AN architecture, structured from enabling technologies to measurable impact. At the innermost core sit the enabling technologies, which drive the surrounding self-X capabilities. These are steered by intent-based orchestration, specifically enabling self-configuration, and are ultimately measured against the outermost layer of KVIs and KPIs.

2.1 Architectural enablers and cognitive technologies

As the innermost core of the AN architecture, *architectural enablers and cognitive technologies* provide the foundational intelligence and flexibility for autonomy. Network softwarization, such as Software-Defined Networking (SDN), Network Function Virtualization (NFV), and cloud-native architectures, provides the programmable, scalable infrastructure necessary for flexible resource allocation and service orchestration. Complementing this, *Big Data Analytics and AI/Machine Learning (ML) models*, including agentic AI, act as the "brain," providing cognitive intelligence for predictive analytics and adaptive decision-making. Finally, *Digital Twins* (DTs) create high-fidelity virtual replicas of Network Entities (NEs), enabling "what-if" proactive verification and risk reduction to support trustworthy and strategic decision-making. Together, these enablers provide the technological substrate required to operationalize the high-level self-X functions described in the following section.

2.2 Goal-oriented self-X capabilities

ANs are defined by four core self-X capabilities, which collectively enable the transition from manual to fully autonomous operations:

- *Self-configuration*: Automatically translates high-level intents into initial network policies, providing the operational base for all higher-order self-X functions. This capability is directly enabled by the IBN paradigm described in the following section.
- *Self-optimization*: Continuously enhances network performance and efficiency through intelligent provisioning and adaptation, balancing service quality and cost.
- *Self-healing*: Ensures service continuity through proactive failure detection and automated recovery.
- *Self-protection*: Proactively detects, analyzes, and neutralizes threats to maintain integrity and compliance.

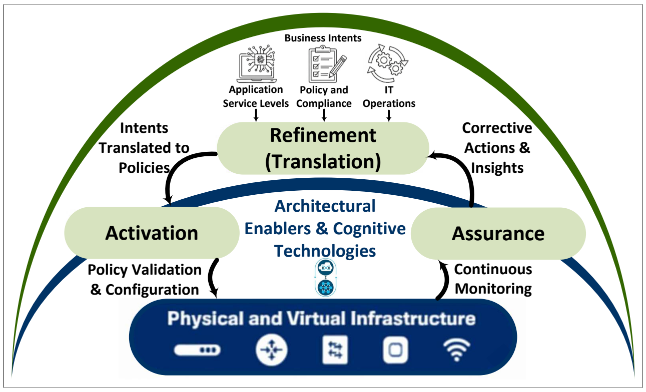
2.3 E2E intent-driven self-configuration

Central to AN is *intent-based networking*, which enables self-configuration by translating high-level business objectives (or intents), declarative specifications of requirements, goals, and constraints, into automated network actions. These intents can originate from human operators via voice (e.g., natural language processing (NLP)-powered assistants) or graphical interfaces, or autonomously through programmatic Application Programming Interfaces (APIs) and monitoring agents. For instance, if a smart factory operator specifies this intent: *"guarantee RAN latency below 2 ms for robotic arm control"*, the IBN framework automatically activates the necessary policies and continuously assures compliance, bridging high-level objectives with programmable infrastructure.

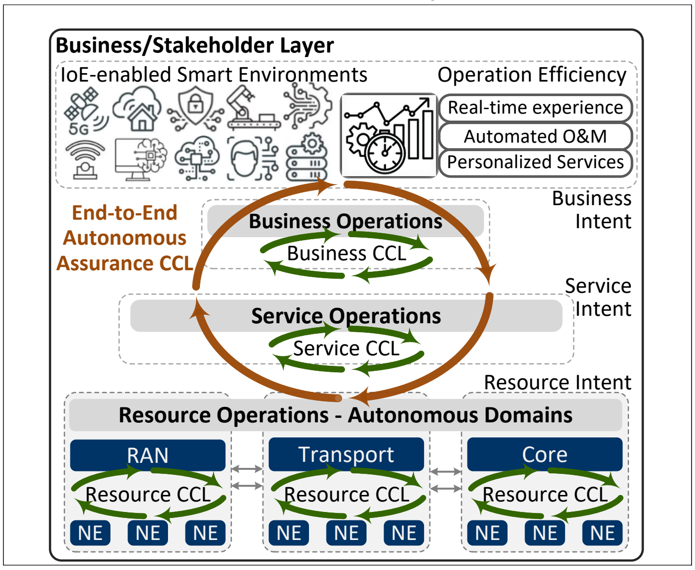
As illustrated in Fig. 2(a), this intent lifecycle is operationalized through CCLs that ensure continuous alignment between goals and the network state via a perpetual cycle of refinement, activation, and assurance. To manage complexity in multi-stakeholder environments, these loops are organized into a multi-layered architecture aligned with Standard Development Organizations (SDOs), as shown in Fig. 2(b). In this framework, IBN acts as the vertical glue, i.e., it translates high-level *business intents* into service-layer requirements, which are then enforced as *resource intents* across autonomous domains (e.g., RAN, transport, core). This hierarchical coordination of CCLs enables real-time E2E assurance across the entire infrastructure.

2.4 Objectives and indicators: KVI & KPI

Autonomous decision-making in ANs is guided by measurable objectives. The latter are captured through KVIs, which assess business and user value (e.g., cost, energy efficiency, compliance), and KPIs, which quantify operational health (e.g., latency, availability), defined in the outermost governance ring of Fig. 1. KVIs & KPIs enable intent-aware operations, supporting closed-loop optimization to meet technical service-level requirements and strategic business goals.



(a) Intent-based networking CCL.



(b) E2E intent orchestration with multi-layered CCLs.

Figure 2 – IBN and E2E orchestration in ANs.

3. INTENT-BASED NETWORKING: CORE ENABLER FOR SELF-CONFIGURATION IN AUTONOMOUS NETWORKS

This section analyzes IBN components and recent advances across key intent lifecycle mechanisms within single and multi-domain 5G/6G architectures.

Standardization bodies have significantly shaped the IBN landscape. IETF defines core terminology and architectural principles for IBN, including intent concepts and definitions (RFC 9315) [6], intent classification (RFC 9316) [7], and service assurance for intent-based architectures (RFC 9417) [8]. ETSI's ZSM and Experiential Networked Intelligence (ENI) groups focus on closed-loop automation and AI-driven management, providing the ZSM reference architecture (GS ZSM 002) [9] and the ENI system architecture (GS ENI 005) [10] on which contemporary intent-based frameworks build. In parallel, the TM Forum provides operational models and ontologies for intent translation and orchestration through IG1253 [11] and IG1161 [12], while 3GPP incorporates intent-based management into 5G/6G specifications (TR 28.812) [13] to simplify the deployment and assurance of complex services such as network slicing. These initiatives establish a foundation for refinement, activation, and assurance, converging toward intent-native 6G architectures.

The following analysis examines the key contributions and limitations of existing IBN research from single and multi-domain perspectives, including considerations for multi-stakeholder environments.

3.1 Key IBN components

The literature typically structures IBN around the core stages of refinement, activation, and assurance, while some work introduces a more granular lifecycle comprising decomposition, translation, negotiation, activation, and assurance. These finer stages represent subprocesses of the broader components, offering deeper insight into conflict management and intent processing. For instance, decomposition and translation fall under refinement, whereas negotiation may appear either as a distinct stage or within activation to ensure consistency before enforcement. Assurance remains the most consistently defined phase across all models. Foundational work by Mwanje et al. [25] established end-to-end definitions and modeling of intent-driven management, providing a structured basis for the intent lifecycle on which subsequent learning-based contributions build.

Table 1 evaluates recent work against seven taxonomy dimensions: *Intent Translation and Refinement* (ITR), *Intent Conflict Resolution* (ICR), *Conflict-Free Intent Activation* (CFIA), *Intent Assurance* (IAS), *Closed-Control Loop* (CCL), *Multi-Domain Orchestration* (MDO), and *Multi-Stakeholder* (MSH). A checkmark (✓) denotes coverage, (X) denotes the specific technique used, and a dash (–) indicates the component is not addressed.

3.1.1 *Intent translation and refinement*

ITR translates high-level intents into actionable, machine-readable network policies. These intents range from technical (expert-defined) to non-technical (e.g., natural-language user requests), creating significant challenges in complex multi-domain environments where traditional heuristic methods (e.g., hierarchical finite-state machine, FSM) lack generalization and scalability [20]. Hence, recent work has explored learning-based approaches. Early models used Bi-LSTMs for intent processing [21, 24], while more recent efforts employed LLMs for semantic parsing [16]–[20]. Specifically, authors of [20] applied few-shot LLMs for single-domain refinement, while authors of [16, 19] extended this work to cross-domain and Internet-wide scenarios. In parallel, [17] explored fine-tuned transformer models such as BERT with Named Entity Recognition (NER) for improved accuracy. More recently, Arafat et al. [18] employed a lightweight pre-trained BERT-based LLM with few-shot prompting for intent classification and keyword extraction (intent type and magnitude), feeding the parsed output directly into

Table 1 – Summary of existing architectures in the road of IBN-driven autonomous networks.

Ref.	Scope	End-To-End Architecture						
		ITR	ICR	CFIA	IAS	CCL	MDO	MSH
[14]	A survey and GenAI/LLM-driven architecture for interactive wireless networks with human-in-the-loop RL.	✓ LLM	–	✓	✓	✓ RLHF	✓	✓
[15]	High-level IBN architecture with conflict handling and orchestration; lacks AI algorithmic design.	✓	✓	✓	✓	✓	✓	✓
[16]	6G-INTENSE: Native AI-driven service and resource federation for multi-domain/stakeholder networks.	✓ LLM	✓ RLHF	✓	✓	✓	✓	✓
[17]	Layered IBN with fine-tuned LLM parsing and DRL/MAS deployment; emerging MDO support.	✓ LLM	✓ DRL	✓ MAS	✓ MAS	✓ Reward	✓	–
[18]	LLM intent parsing with transformer-based forecasting and attention-HRL xApp orchestration in O-RAN.	✓ LLM	✓ Forecast	✓ HRL	✓	✓	–	–
[19]	Few-shot LLM (CodeLlama) for multi-domain intent decomposition and translation to ILIs (NSD/RAND).	✓ LLM	✓	✓	✓	✓ HF	✓ LLM	–
[20]	Few-shot LLM-based refinement with policy trees & FSM-guided control loops for intent fulfillment.	✓ LLM	✓	✓	✓ LLM	✓ FSM	–	–
[21]	Decoupled IBN with BiLSTM for translation, NN/FSM for conflict-to-policy, and an SAI control loop.	✓ BiLSTM	✓ NN	✓ FSM	✓	✓ SAI	✓	–
[22]	Full-lifecycle verification; conflict resolution via policy graph; NFV feasibility with feedback (no AI/MDO).	✓ Rule	✓ PGA	✓	✓ Hybrid	✓	–	–
[23]	Static rule-based intent negotiation for conflict resolution & QoS-aware alternate intent generation; no AI.	✓	✓ Rule	✓	✓	✓	✓	–
[24]	Three-layer SMART architecture for full-lifecycle automated intent management.	✓ BiLSTM	✓ LT	✓ DQN/DT	✓ DNE-Net	✓	✓	–

Legend: ITR: Intent Translation/Refinement, ICR: Intent Conflict Resolution (Negotiation), CFIA: Conflict-Free Intent Activation, IAS: Intent Assurance, CCL: Closed-Control Loop, MDO: Multi-Domain Orchestration, MSH: Multi-Stakeholders.
✓/(X) Fully addressed with a specific X technique (e.g., LLM, DRL); ✓ Partially addressed without a specific technique; – Not addressed.

downstream validation and control modules.

Despite these advances in translating high-level intents into structured policy configurations, several challenges remain, such as ensuring semantic consistency, mitigating LLM hallucinations, and developing standardized policy models that generalize across multi-domain multi-stakeholder environments, where high-quality domain datasets are scarce. Moreover, the widespread reliance on few-shot learning raises reproducibility issues when studies omit details on example selection, potentially leading to overfitting and poor generalization.

3.1.2 *Intent activation with conflict resolution*

Intent activation enforces refined policies on network elements using self-X capabilities. A key challenge is resolving conflicts that arise from competing intents, shared resources, and divergent stakeholder objectives. These conflicts typically manifest as (i) *inter-policy* conflicts (overlapping or contradictory rules) and (ii) *policy–state* conflicts (incompatibilities with resource availability or operational conditions).

Conventional frameworks, such as the Metro Ethernet Forum (MEF) and Yet Another Next Generation (YANG) models, provide structured mappings from policies to configurations but remain static and are inadequate for dynamic conflict resolution in evolving environments [20]. Recent research addressed this limitation with varying degrees of automation. For instance, the SMART framework in [24] used a label tree for conflict handling, augmented by a Deep Q-Network (DQN) and Fuzzy Decision Tree (DT)-based policy optimization. Other efforts include a rule-based Service Level Agreement (SLA) negotiation framework [23] and the use of Policy Graph Abstraction (PGA) for conflict detection and resolution [22]. Subsequently, [21] integrated Neural Network (NN)–based conflict handling with FSM-driven activation to extend PGA. Alternatively, Njah et al. leveraged in [17] Multi-Agent Reinforcement Learning (MARL) to coordinate policy activation toward global objectives, while authors in [16] employed hierarchical RL with Human Feedback (RLHF) to reconcile multi-stakeholder priorities. At a broader architectural level, Alemany et al. highlighted the need for cross-domain conflict handling but provided limited detail on concrete AI implementations [15]. In parallel, Arafat et al. [18] addressed policy–state conflicts by introducing a transformer-based predictive validation step that filters intents against forecasted traffic before activation, coupled with an attention-based hierarchical reinforcement learning controller that orchestrates combinations of network-optimization xApps in alignment with the validated intent.

Although LLMs are promising for intent refinement [16, 20], existing activation and negotiation mechanisms often rely on static rules or human intervention, thus creating an adaptability gap that limits fully autonomous self-configuration.

3.1.3 Intent assurance and closed-control loops

Intent assurance relies on CCLs to ensure that deployed network configurations continuously reflect business intent, while automatically correcting configuration drift detected through telemetry and key indicators. Human Feedback (HF) can further refine system adaptability.

In this context, learning-based methods have been pioneering such efforts. For instance, the SMART framework used a Deep Neural Evolutionary Network (DNE-Net) for fault prediction and repair [24]. Moreover, Emergence approach in [20] monitored KPIs and employed LLMs for intent-driven adjustments. Alternatively, Njah et al. used a MARL-based technique with goal-aligned reward signals for continuous assurance [17]. Also, authors in [21] exemplified this shift by implementing a continuous telemetry-driven State, Action, Intent (SAI) loop for autonomous control. Finally, Abdelkader et al. applied LLM-based activation (CodeLLaMA) with HF-validated assurance in [19], providing implicit performance tracking but without autonomous KPI adaptation. Broader architectural views such as [15] emphasized the importance of CCLs for 6G, but remained conceptual and lacked concrete AI implementations.

3.2 Intent fulfillment across multi-domain environments

Extending intent enforcement from single to multi-domain environments remains challenging, requiring semantic alignment and coordinated CCLs to maintain E2E service consistency while preserving domain autonomy.

3.2.1 Architectural approaches

Various architectural models have been proposed for multi-domain IBN. *Decentralized frameworks*, as in [15], deployed an *Intent Management Function (IMF)* within each domain as the core entity for autonomy, managing assurance loops and coordinating provisioning, optimization, protection, and healing. The IMF acts as both an intent owner, defining and propagating refined intents to peers, and as a handler, enforcing incoming intents locally. Its modules typically include intent translation, policy modeling, conflict resolution, and situational awareness, enabling scalable multi-domain orchestration. *Centralized models*, such as those surveyed in [16], employed a top-down orchestrator to manage fulfillment across domains, aligning with standardization efforts (e.g., ETSI ZSM) and ensuring global consistency and visibility. Finally, *hybrid architectures* combined both approaches, as depicted in Fig. 3. Specifically, a central orchestrator decomposes high-level intents for distributed IMFs, while it aggregates telemetry (e.g., KVIs and KPIs) from CCLs to align local actions with global objectives.

3.2.2 Multi-domain CCL coordination

Coordinating CCLs is essential for real-time intent assurance but challenged by policy misalignment and conflicting objectives across layers, domains, and stakeholders. Addressing these challenges relies on standardized interfaces and real-time telemetry. Solutions such as the *Distributed Intent Inventory (DII)* maintain a synchronized record of intent states and policies across domains, enabling policy alignment and conflict-aware coordination. This allows IMFs to negotiate trade-offs between local policies and shared service objectives.

Overall, the literature shows a shift from rule-based refinement to learning-based intent processing, yet lacks a coherent multi-domain orchestration model unifying refinement, conflict resolution, activation, and assurance within a closed-loop framework. In particular, while recent contributions such as 6G-INTENSE [16] address

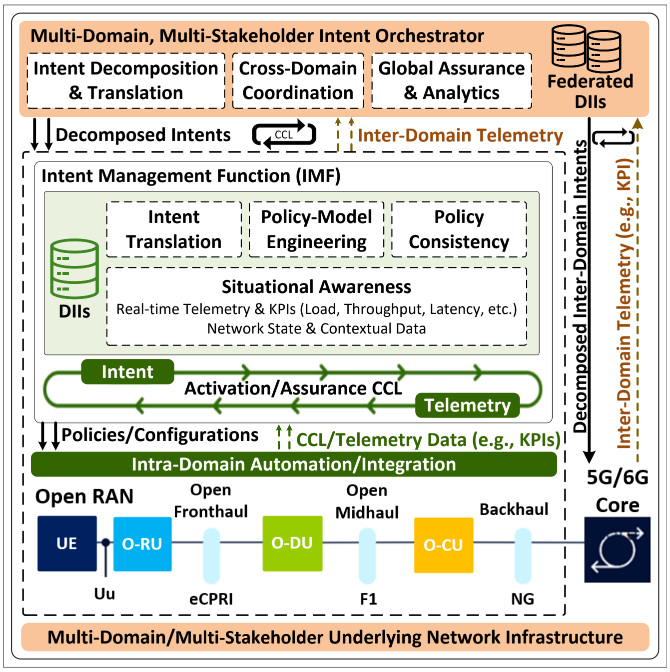


Figure 3 – Intent-Based Multi-Domain Orchestration.

structural infrastructure abstraction for multi-domain resource federation, they do not provide a cognitive reasoning lifecycle unifying these stages. To address this gap, Section 4 derives an enriched reference model for agentic GenAI-driven IBN, grounded in the state-of-the-art, enabling cross-domain E2E intent fulfillment.

4. TOWARD AGENTIC GENAI-DRIVEN IBN FOR SELF-CONFIGURATION: VISION AND CHALLENGES

GenAI is reshaping network automation by enabling semantic processing, knowledge-grounded decision support, and generative policy workflows. It leverages foundational models that move beyond traditional analytics. Among the most pivotal are LLMs. Built on the transformer architecture, LLMs use self-attention mechanisms to capture dependencies across input tokens. While originally designed for text, this mechanism extends to multimodal data, from image patches to sensor readings, enabling complex cross-domain contextual understanding. Complementing LLMs, Generative Adversarial Networks (GANs) employ adversarial learning to synthesize realistic data, while diffusion models iteratively denoise inputs to produce coherent outputs [14]. Building on this foundation, we employ LLMs for natural language intent fulfillment as the core mechanism for self-configuration in E2E ANs.

4.1 Closed-loop agentic GenAI-driven IBN vision

This section outlines our agentic GenAI vision for autonomous intent lifecycle management, wherein a hierarchy of LLM-driven agents collaborates within CCLs to orchestrate E2E intent refinement, activation, and assurance across multi-domain, multi-stakeholder environments.

To ground this agentic framework within a standardized architecture, we instantiate it within the 6G O-RAN system. Particularly, intent-based closed-loop control is distributed across the O-RAN RAN Intelligent Controllers (RICs), where the Non-Real-Time RIC (Non-RT RIC), typically residing within the Service Management and Orchestration (SMO) layer, orchestrates network intelligence by deploying rApps for strategic intent reasoning, policy generation, and long-term assurance analysis. In contrast, the Near-Real-Time RIC (Near-RT RIC) operates at sub-second timescales and employs xApps for time-sensitive policy activation and continuous intent optimization at the RAN level [26]. Consistent with this functional split, the LLM-based agents reside in the Non-RT RIC, while the Near-RT RIC xApps enforce the validated, conflict-free policies. Lightweight domain-specialized LLMs are assumed to be deployed at the SMO/edge for local inference, while larger models are reserved for offline tasks.

The core components of this architecture, illustrated in Fig. 4, are detailed in the following subsections. The Industry 5.0 vision drives stringent requirements for network automation across heterogeneous IIoT environments [27], motivating the use case explored in this section. To facilitate understanding the system's operation, we introduce three high-level operational intents for an autonomous smart factory, each targeting one of its core objectives (performance, sustainability, and intelligence). These intents are intentionally expressed at different abstraction levels to demonstrate the framework's decomposition and orchestration capabilities:

- Intent 1 (Factory Operator): *"Ensure RAN latency below 2 ms for robotic arm control traffic."*
- Intent 2 (Energy Manager): *"Reduce RAN energy consumption by 25% during off-hours, while preserving performance for critical systems."*
- Intent 3 (Automation Supervisor): *"Enable real-time predictive maintenance for critical robotic assets with low-latency and high-reliability operational analytics."*

4.1.1 *Cross-domain semantic intent orchestration agent*

The multi-domain orchestrator performs Semantic Intent Decomposition (SID), transforming high-level intents from the business input layer into coordinated, domain-specific atomic sub-intents. Guided by the LLM's reasoning and domain knowledge, SID integrates semantic parsing, capability grounding, and constraint propagation to interpret functional goals, operational needs, and requirement constraints. It first identifies atomic functional tasks (e.g., real-time control, data streaming, analytics storage) and maps each to the most suitable

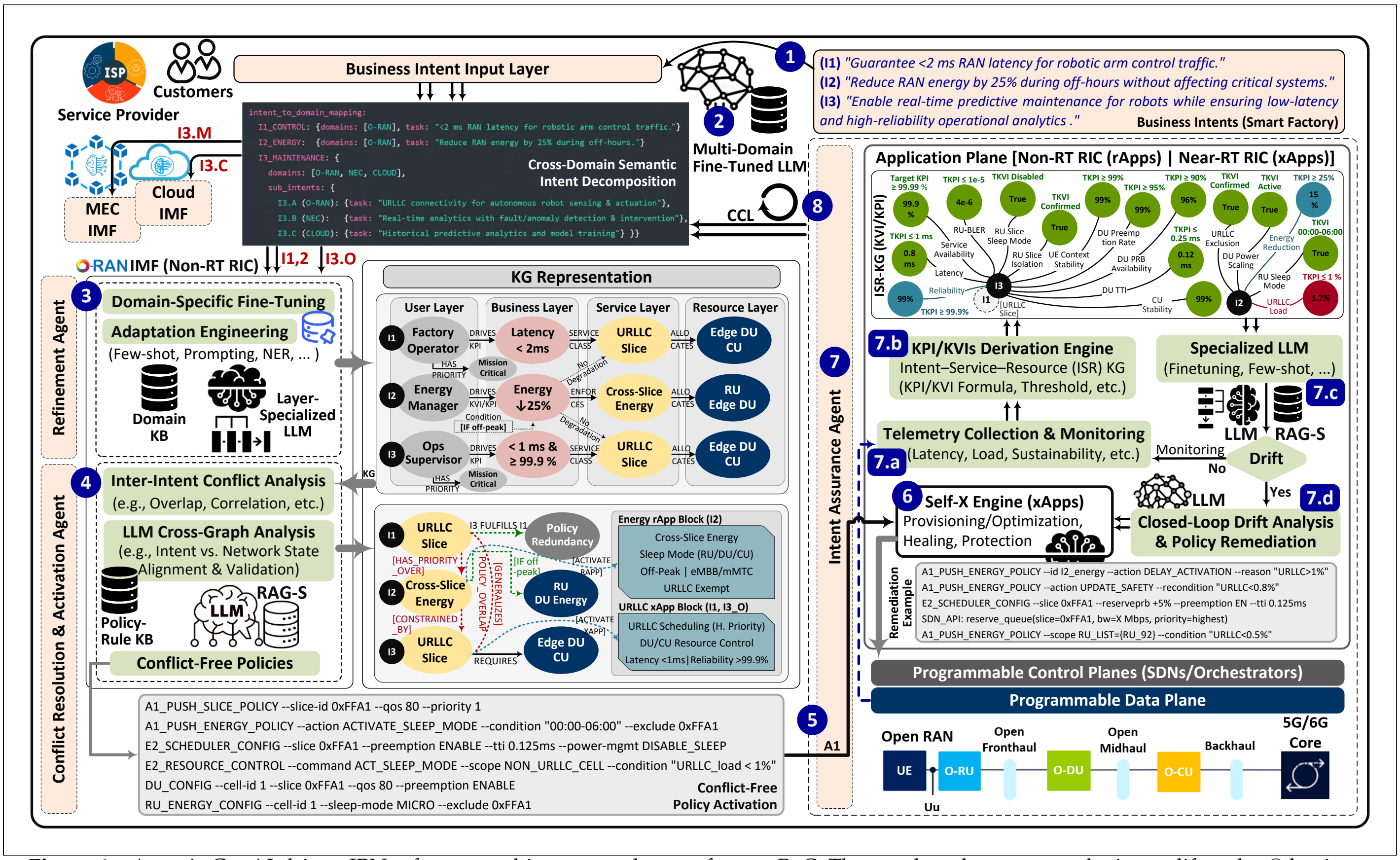


Figure 4 – Agentic GenAI-driven IBN reference architecture and smart factory PoC. The numbered steps trace the intent lifecycle: ① business intent capture; ② cross-domain semantic orchestration via multi-domain fine-tuned LLM; ③ domain-specific refinement and KG construction; ④ inter-intent conflict analysis and LLM cross-graph validation grounded by RAG-S; ⑤ conflict-free policy activation via A1; ⑥ Self-X xApp enforcement (provisioning, optimization, healing, protection); ⑦ closed-loop assurance, comprising (a) telemetry collection, (b) KPI/KVI derivation, (c) drift analysis, and (d) policy remediation; ⑧ closed control loop (CCL) feedback for intent re-evaluation and continuous lifecycle adaptation.

domain (e.g., RAN, core, edge/cloud) while propagating E2E performance constraints (e.g., latency, reliability, availability) into domain-specific performance budgets.

As illustrated in Fig. 4 (step ②, Cross-Domain Semantic Intent Orchestration), Intents 1 and 2 are fulfilled within the O-RAN domain IMF, focusing on latency, reliability, and energy optimization. Intent 3 is decomposed into three coordinated sub-intents: I3.O (O-RAN sub-intent, URLLC connectivity for robotic sensing and actuation), I3.M (MEC sub-intent, real-time analytics with fault and anomaly detection), and I3.C (Cloud sub-intent, historical analytics and model training). The outcome is a structured set of sub-intents that collectively ensure E2E intent fulfillment while preserving autonomy and cross-domain assurance traceability.

4.1.2 Refinement agent with multi-layered translation

This specialized LLM refines user intents into enforceable policies via a two-stage pipeline (Fig. 4, step ③, Refinement Agent): (i) multi-layer decomposition across user/business/service/resource layers, and (ii) semantic parsing with KG grounding for operational feasibility and policy traceability. For instance, Intent 1 is refined from a user-layer priority (factory operator) into business KPIs, then mapped to a service-layer policy (URLLC slice), and finally grounded in resource-layer configurations (DU/CU scheduling). Similarly, Intent 2 is translated into business objectives (RAN energy reduction up to 25%), mapped to a service-layer cross-slice energy policy while preserving URLLC guarantees, and grounded in resource-level actions such as RU sleep-mode activation, as illustrated by the layered KG representation in Fig. 4.

This LLM-driven approach overcomes traditional ML limitations in processing unstructured inputs while improving structure and controllability of the generated policies. To ensure reliability, mitigate hallucinations in LLMs, and prevent overgeneralization, the agent leverages complementary techniques, such as: *domain-specific fine-tuning* for semantic adaptation, *few-shot learning* for rapid policy formulation, *layer-aware prompting* to preserve abstraction integrity, and *entity recognition* for precise policy extraction. The final output is a structured, KG-based representation of the refined intent, explicitly linking all layers from user requirements to resource mappings.

4.1.3 Conflict resolution and activation agent

This agent leverages three core capabilities for autonomous policy activation: (1) LLMs for adaptive reasoning,

(2) KGs for structured, domain-aware policy representations, and (3) RAG-S for real-time grounded inference. The agent operates on two KGs per domain: an *intent KG*, derived from the refinement agent, modeling intent enti-ties and relationships to detect structural conflicts (e.g., redundancy, shadowing), and a *network inventory KG*, capturing the dynamic state of services and resources for feasibility validation against infrastructure constraints. Leveraging this dual-KG architecture, the agent performs a two-stage validation process within RAG-S loops.

Before detailing these stages, we specify the underlying RAG-S mechanism. RAG-S extends standard retrieval-augmented generation by verifying the sufficiency and completeness of retrieved network context prior to LLM inference, ensuring that generated actions are grounded in a fully validated operational state rather than partial or inconsistent evidence. Sufficiency requires two conditions: (i) *structural coverage* of all entity and attribute slots required by the query, retrieved from the intent KG and network inventory KG, and (ii) *semantic consistency* across retrieved items, assessed via LLM self-evaluation. On failure, the agent triggers bounded re-retrieval; on persistent insufficiency, it blocks policy generation and escalates to the closed-loop controller rather than emitting an ungrounded policy. This default-deny behavior distinguishes RAG-S from Self-RAG [28] (self-critique during generation) and Corrective RAG [29] (post-retrieval correction via external sources). In contrast, RAG-S performs this check before inference, grounded by the domain-specific KGs (intent and network inventory), rather than during or after.

Concretely, the two-stage validation loop operates across both KGs as follows. First, the agent detects inter-intent conflicts and proposes prioritized resolutions via intent KG analysis. Then, it validates policy feasibility through cross-graph comparison with the network inventory KG. For instance, as shown in Fig. 4 (step ④, Conflict Resolution & Activation block, KG Representation panel), a policy generalization is detected between I1 and I3, where I3's URLLC service satisfies I1's latency requirement, leading to the deactivation of the redundant I1 policy. A contradiction is also identified between I2's energy-saving objective and the URLLC constraints of I1/I3. The agent resolves this through priority enforcement: I2 applies only during off-peak hours and excludes URLLC slices, ensuring that I1/I3 retain high-priority scheduling.

Rather than optimizing intents in isolation, the agent assesses their collective impact to maintain network-wide stability. For unresolvable conflicts, the CCL triggers administrator alerts with LLM-generated diagnostics. Validated, conflict-free policies are deployed as A1 policies from the Non-RT RIC to the Near-RT RIC, where they are enforced by xApps supporting self-optimization, self-healing, and self-protection across the programmable infrastructure, as depicted in step ⑤ (Conflict-Free Policy Activation block) and step ⑥ (Self-X Engine xApps) of Fig. 4.

4.1.4 Closed-loop assurance agent

This agent maintains continuous alignment between network operating conditions and business objectives through a closed-loop assurance process. It integrates two complementary mechanisms: (1) KVI/KPI-driven monitoring to quantify strategic value and operational health, and (2) LLM-based analysis enhanced with KGs and RAG-S. The KVIs/KPIs evaluate the state of intent-fulfilling services and resources using a signed impact vector $[-1, 0, +1]$, representing degraded (red), threshold-level (blue), and optimal (green) performance, as depicted in the *ISR-KG (KVI/KPI)* block of Fig. 4 (step ⑦b). These indicators (e.g., throughput, latency, packet loss) are continuously monitored via the *Telemetry Collection & Monitoring module* (Fig. 4, step ⑦a).

The assurance process operates as follows. The collected measurements are mapped into an ISR-KG, encoding persistent relationships between intents, services, and resources to ensure semantic traceability. The ISR-KG links each intent (e.g., I1/I3 and I2) to corresponding DU/RU/CU resources and their validated KPIs/KVIs across performance, reliability, resource, and energy dimensions, forming a unified, machine-verifiable representation of policy fulfillment across the O-RAN architecture. The KG-enhanced LLM, operating within the RAG-S loop (Fig. 4, Specialized LLM + Drift block, step ⑦c), performs graph traversal and KVI/KPI subgraph matching to detect policy fulfillment deviations (drift). Upon detection, it generates evidence-backed corrective recommendations. For example, in Fig. 4 step ⑦d (Closed-Loop Drift Analysis & Policy Remediation block), a URLLC load drift triggers the LLM agent to issue corrective configuration commands. These actions block conflicting energy-saving modes, enforce safety preconditions (e.g., URLLC load $\leq 0.8\%$), and prioritize radio and transport resources for the affected slice. This restores safe operating conditions while preserving intent alignment, with enforcement spanning the O-RAN infrastructure, from RIC policies down to O-CU, O-DU, and O-RU elements. This semantic traceability enhances explainability and ensures that all remediation actions remain aligned with high-level business objectives.

Our proposed framework establishes a synergistic multi-agent system where LLM agents operate through hierarchical CCLs (Fig. 4, step ⑧). This structure is essential for scalable reliability, enabling local domain autonomy while ensuring global oversight through a multi-domain orchestrator. The system's assurance is grounded in the following validation rule: E2E intent satisfaction is confirmed when all decomposed domain-specific sub-intents are fulfilled.

4.2 Challenges and future research directions

Achieving fully autonomous, GenAI-powered IBN presents challenges ranging from establishing robust single-domain autonomy to enabling seamless orchestration across multi-domain, multi-stakeholder environments. The following discusses the key open issues toward realizing this vision.

4.2.1 Semantic fidelity in intent translation

Accurately translating high-level intents into enforceable policies remains a key challenge for autonomous IBN. The difficulty arises from supporting multi-dimensional intelligent services with diverse QoS expectations, personalized user demands, and stringent vertical-specific requirements, all of which require precise semantic interpretation. This complexity is further amplified by heterogeneous infrastructures where resources, constraints, and policies differ across domains and vendors. Because intent representation and policy specification models are largely non-standardized, reliable LLM-driven intent translation remains an open research problem.

4.2.2 Conflict-free intent activation

Achieving conflict-free intent activation is hindered by competing services, resource contention, and misaligned stakeholder objectives in multi-domain environments. Heterogeneous policies and platform dependencies further complicate enforcement. Although GenAI-based models offer promising capabilities for conflict analysis and adaptive policy adjustment, developing scalable and standardized negotiation mechanisms that balance local autonomy with global objectives remains under-investigated.

4.2.3 Scalable intent assurance

Reliable intent assurance depends on robust single-domain autonomy. GenAI models can enhance CCLs by analyzing assurance telemetry (e.g., KPIs/KVIs) to detect intent drift, quantify its impact, and recommend corrective configurations. However, in multi-domain environments, local CCLs may initiate corrective actions that are locally optimal yet suboptimal at the cross-domain level. A drift detected in one domain may originate in another, leading to inefficient or counterproductive remediation. Designing GenAI-driven orchestrators capable of coordinating cross-domain corrective actions while preserving domain autonomy through minimal, trusted data exchange remains an open research challenge.

4.2.4 Trustworthy and secure GenAI-driven intent fulfillment

The integration of GenAI, particularly LLMs as core components for intent fulfillment, faces several challenges. First, ensuring robust reasoning across diverse network domains and stakeholder perspectives is complex, as single generalist models often lack the contextual awareness required for fine-grained decisions, such as validation and drift detection. This limitation is further compounded by the scarcity of sensitive operational data for LLM adaptation, the complexity of fusing heterogeneous feedback (e.g., QoS, QoE, KPI/KVIs, and telemetry), and the need to generalize reliably to unseen intents in dynamic heterogeneous environments. Second, automated intent lifecycles introduce new attack surfaces and confidentiality risks. Indeed, LLMs are vulnerable to adversarial prompts and policy manipulation, and sensitive operational data may be exposed during inference or cross-domain coordination phases. Ensuring trustworthy operation requires robust defenses and privacy-preserving mechanisms across the GenAI pipeline, which are currently immature.

Addressing these challenges requires robust approaches for trustworthy automation. A promising direction lies in leveraging digital twins (DTs) for proactive trust assurance, enabling "what-if" pre-deployment validation and adversarial testing. Moreover, GANs and diffusion models can generate realistic, privacy-preserving synthetic data to overcome fine-tuning scarcity, while LLMs enable explainable and secure automation. Future research should integrate these capabilities into scalable, cross-domain solutions for 6G-era ANs.

5. CONCLUSION

ANs represent the shift toward self-managing systems capable of configuration, optimization, healing, and protection with minimal human intervention. This paper presented a vision, reference architecture, and research roadmap for agentic GenAI-driven intent-based self-configuration in autonomous networks. Building on a focused analysis of recent IBN frameworks across the intent lifecycle, we identified persistent gaps in intent translation, conflict-free activation, and closed-loop assurance, and introduced an end-to-end agentic architecture leveraging specialized LLM agents, dual KGs, and the RAG-S mechanism. The architecture was instantiated within a 6G O-RAN ecosystem and illustrated through an Industry 5.0 use case, outlining key research challenges and future directions toward trustworthy, cognitive, zero-touch 6G networks.

REFERENCES

[1] Cisco Systems, Inc. *What Is Network Automation?* Accessed: 2025-09-16. URL: https://www.cisco.com/site/us/en/learn/topics/networking/what-is-network-automation.html.

[2] Nokia. *Using AI Driven BSS, Autonomous Networks, and Digital Twins to Open New Revenues: 10 Use Cases*. Accessed: 2025-09-17. Nokia. URL: https://pf.content.nokia.com/t007no-cns-ops-digital-operations-center/white-paper-using-ai-driven-bss-autonomous-networks-and-digital-twins.

[3] Estefanía Coronado, Rasoul Behravesh, Tejas Subramanya, Adriana Fernàndez-Fernàndez, Muhammad Shuaib Siddiqui, Xavier Costa-Pérez, and Roberto Riggio. "Zero Touch Management: A Survey of Network Automation Solutions for 5G and 6G Networks". In: *IEEE Communications Surveys Tutorials* 24.4 (2022), pp. 2535–2578. DOI: 10.1109/COMST.2022.3212586.

[4] Yao Wang et al. "A Survey on Intent-Driven End-to-End 6G Mobile Communication System". In: *IEEE Commun. Surv. Tuts.* (2025), pp. 1–1. DOI: 10.1109/COMST.2025.3575041.

[5] Akshay Jain et al. "Knowledge Graph-Based Approach for Sustainable 6G End-to-End System Design". In: *arXiv preprint arXiv:2507.08717* (2025).

[6] A. Clemm, L. Ciavaglia, L. Z. Granville, and J. Tantsura. *Intent-Based Networking - Concepts and Definitions*. Internet Request for Comments RFC 9315. IETF, Oct. 2022. DOI: 10.17487/RFC9315. URL: https://www.rfc-editor.org/info/rfc9315.

[7] C. Li, O. Havel, A. Olariu, P. Martinez-Julia, J. Nobre, and D. Lopez. *Intent Classification*. Internet Request for Comments RFC 9316. IETF, Oct. 2022. DOI: 10.17487/RFC9316. URL: https://www.rfc-editor.org/info/rfc9316.

[8] B. Claise, J. Quilbeuf, D. Lopez, D. Voyer, and T. Arumugam. *Service Assurance for Intent-Based Networking Architecture*. Internet Request for Comments RFC 9417. IETF, July 2023. DOI: 10.17487/RFC9417. URL: https://www.rfc-editor.org/info/rfc9417.

[9] ETSI. *Zero-touch network and Service Management (ZSM); Reference Architecture*. Tech. rep. GS ZSM 002 V1.1.1. European Telecommunications Standards Institute (ETSI), Aug. 2019. URL: https://www.etsi.org/deliver/etsi_gs/ZSM/001_099/002/01.01.01_60/gs_ZSM002v010101p.pdf.

[10] ETSI. *Experiential Networked Intelligence (ENI); System Architecture*. Tech. rep. GS ENI 005 V3.1.1. European Telecommunications Standards Institute (ETSI), June 2023. URL: https://www.etsi.org/deliver/etsi_gs/ENI/001_099/005/03.01.01_60/gs_ENI005v030101p.pdf.

[11] TM Forum. *Intent in Autonomous Networks*. Tech. rep. IG1253 v1.2.0. TM Forum, 2022. URL: https://www.tmforum.org/resources/introductory-guide/ig1253-intent-in-autonomous-networks-v1-2-0/.

[12] TM Forum. *Autonomous Networks Technical Architecture*. Tech. rep. IG1161 v1.0.0. TM Forum, 2020. URL: https://www.tmforum.org/resources/standard/ig1161-autonomous-networks-technical-architecture-v1-0-0/.

[13] 3GPP. *Telecommunication management; Study on scenarios for Intent driven management services for mobile networks*. Tech. rep. TR 28.812 V17.1.0. 3rd Generation Partnership Project (3GPP), 2020. URL: https://www.3gpp.org/ftp/Specs/archive/28_series/28.812/.

[14] Xiaoqi Qin et al. "Generative AI Meets Wireless Networking: An Interactive Paradigm for Intent-Driven Communications". In: *IEEE Trans. Cogn. Commun. Network.* (2025), pp. 2056–2077. DOI: 10.1109/TCCN.2025.3567613.

[15] Pol Alemany et al. "Defining Intent-Based Service Management Automation for 6G Multi-Stakeholders Scenarios". In: *IEEE Op. J. of the Commun. Soc.* 6 (2025), pp. 2373–2396. DOI: 10.1109/OJCOMS.2025.3554250.

[16] Akram Boutouchent et al. "6G-INTENSE: Intent-Driven Native Artificial Intelligence Architecture Supporting Network-Compute Abstraction and Sensing at the Deep Edge". In: *IEEE Veh. Technol. Mag.* 20.1 (2025). DOI: 10.1109/MVT.2024.3525001.

[17] Yosra Njah, Aris Leivadeas, and Matthias Falkner. "An AI-Driven Intent-Based Network Architecture". In: *IEEE Commun. Mag.* 63.4 (2025), pp. 146–153. DOI: 10.1109/MCOM.001.2400143.

[18] Md Arafat Habib, Pedro Enrique Iturria Rivera, Yigit Ozcan, Medhat Elsayed, Majid Bavand, Raimundus Gaigalas, and Melike Erol-Kantarci. "LLM-Based Intent Processing and Network Optimization Using Attention-Based Hierarchical Reinforcement Learning". In: *2025 IEEE Wireless Communications and Networking Conference (WCNC)*. 2025, pp. 1–6. DOI: 10.1109/WCNC61545.2025.10978505.

[19] Abdelkader Mekrache, Adlen Ksentini, and Christos Verikoukis. "Intent-Based Management of Next-Generation Networks: an LLM-Centric Approach". In: *IEEE Netw.* 38.5 (2024), pp. 29–36.

[20] Kristina Dzeparoska, Ali Tizghadam, and Alberto Leon-Garcia. "Emergence: An Intent Fulfillment System". In: *IEEE Commun. Mag.* 62.6 (2024), pp. 36–41. DOI: 10.1109/MCOM.001.2300270.

[21] Tong Li et al. "Architecting Autonomous Network Management and Control via Intent-Driven Decoupled Network". In: *IEEE Netw.* 38.6 (2024), pp. 361–369. DOI: 10.1109/MNET.2024.3399189.

[22] Yanbo Song, Chungang Yang, Jiaming Zhang, Xinru Mi, and Dusit Niyato. "Full-Life Cycle Intent-Driven Network Verification: Challenges and Approaches". In: *IEEE Netw.* 37.5 (2023), pp. 145–153. DOI: 10.1109/MNET.124.2200127.

[23] Yogesh Sharma, Deval Bhamare, Andreas Kassler, and Javid Taheri. "Intent Negotiation Framework for Intent-Driven Service Management". In: *IEEE Commun. Mag.* 61.6 (2023), pp. 73–79. DOI: 10.1109/MCOM.001.2200504.

[24] Chungang Yang, Xinru Mi, Ying Ouyang, Ru Dong, Junjie Guo, and Mohsen Guizani. "SMART Intent-Driven Network Management". In: *IEEE Commun. Mag.* 61.1 (2023), pp. 106–112. DOI: 10.1109/MCOM.002.2200119.

[25] Stephen S. Mwanje, Anubhab Banerjee, Juergen Goerge, Abdelrahman Abdelkader, Gabor Hannak, Peter Szilagyi, Tejas Subramanya, Julian Goser, and Tobias Foth. "Intent-Driven Network and Service Management: Definitions, Modeling and Implementation". In: *ITU Journal on Future and Evolving Technologies* 3.3 (2022), pp. 639–654. DOI: 10.52953/QSPC9266.

[26] Michele Polese et al. "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges". In: *IEEE Commun. Surv. Tuts.* 25.2 (2023), pp. 1376–1411. DOI: 10.1109/COMST.2023.3239220.

[27] H. R. Chi, C. K. Wu, N. F. Huang, K. F. Tsang, and A. Radwan. "A Survey of Network Automation for Industrial Internet-of-Things Toward Industry 5.0". In: *IEEE Transactions on Industrial Informatics* 19.2 (Feb. 2023), pp. 2065–2077. DOI: 10.1109/TII.2022.3215231.

[28] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection". In: *Proc. International Conference on Learning Representations (ICLR)*. 2024.

[29] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling. "Corrective Retrieval Augmented Generation". In: *arXiv preprint arXiv:2401.15884* (2024).

AUTHORS

Yosra Njah (yosra.njah@etsmtl.ca) is a postdoctoral researcher at École de technologie supérieure (ÉTS), Montreal, Canada. Her research interests include zero-touch networks, intent-based networking, programmable networks, 5G/6G networks, multi-domain orchestration, and AI for networking. She has been contributing to several R&D projects with industry and government partners, including Cisco, Ciena, and the DND, Canada.

Wael Jaafar [SM] (wael.jaafar@etsmtl.ca) is an associate professor at the Software and IT Engineering Department of École de Technologie Supérieure (ÉTS), Montreal, Canada. His research interests include aerial networks, 5GB technologies, digital twin, and artificial intelligence.

Rami Langar (rami.langar@etsmtl.ca) has been a full professor of computer science at École de Technologie Supérieure (ÉTS), Montréal, Canada, since 2021, and has been affiliated with University Gustave Eiffel, France, since 2016. His research interests include resource management in 5G/6G and O-RAN networks, green networking, machine learning for communications and networks, cybersecurity, digital twin, and quantum AI.

Bassant Selim (bassant.selim@etsmtl.ca) is an assistant professor at École de technologie supérieure, Montreal, Canada. Her research interests include intelligent/artificial intelligence-enabled wireless communications, sustainable information and communication technologies, and quantum-inspired machine learning.

Mohamed Cheriet [SM] (mohamed.cheriet@etsmtl.ca) is a full professor at École de Technologie Supérieure (ÉTS), Montreal, Canada. He is the founder and Director of the Synchromedia Laboratory. He held a Tier 1 Canada Research Chair on Sustainable Smart Eco-Cloud. His research interests include computational intelligence, machine learning, AI, and their applications to networking and image processing.