Toward Edge General Intelligence With Agentic AI and Agentification: Concepts, Technologies, and Future Directions

Ruichen Zhang, *Member, IEEE*, Guangyuan Liu, *Graduate Student Member, IEEE*, Yinqiu Liu, *Member, IEEE*, Changyuan Zhao, *Graduate Student Member, IEEE*, Jiacheng Wang, *Member, IEEE*, Yunting Xu, Dusit Niyato, *Fellow, IEEE*, Jiawen Kang, *Senior Member, IEEE*, Yonghui Li, *Fellow, IEEE*, Shiwen Mao, *Fellow, IEEE*, Sumei Sun, *Fellow, IEEE*, Xuemin Shen, *Fellow, IEEE*, and Dong In Kim, *Life Fellow, IEEE*

***Abstract*— The rapid expansion of sixth-generation (6G) wireless networks and the Internet of Things (IoT) has catalyzed the evolution from centralized cloud intelligence towards decentralized edge general intelligence. However, traditional edge intelligence methods, characterized by static models and limited cognitive autonomy, fail to address the dynamic, heterogeneous, and resource-constrained scenarios inherent to emerging edge networks. Agentic artificial intelligence (Agentic AI) emerges as a transformative solution, enabling edge systems to autonomously perceive multi-modal environments, reason contextually, and adapt proactively through continuous perception–reasoning–action loops. In this context, the agentification of edge intelligence serves as a key paradigm shift, where distributed entities evolve into autonomous agents capable of collaboration and continual adaptation. This paper presents a comprehensive survey dedicated to Agentic AI and agentification frameworks tailored explicitly for edge general intelligence. First, we systematically introduce foundational concepts and clarify distinctions from traditional edge intelligence paradigms. Second, we analyze important enabling technologies, including compact model compression, energy-aware computing strategies, robust connectivity frameworks, and advanced knowledge representation and reasoning mechanisms. Third, we provide representative case studies demonstrating Agentic AI's capabilities in low-altitude economy networks, intent-driven networking, vehicular networks, and human-centric service provisioning, supported by numerical evaluations. Furthermore, we identify current research challenges, review emerging open-source platforms, and highlight promising future research directions to guide robust, scalable, and trustworthy Agentic AI deployments for next-generation edge environments.**

***Index Terms*— 6G networks, agentic AI, agentification, edge general intelligence, edge intelligence, AI agent, reinforcement learning, retrieval-augmented generation (RAG), large language models (LLMs).**

Received 25 August 2025; revised 25 November 2025; accepted 28 December 2025. Date of current version 13 January 2026. This work was supported in part by Seatrium New Energy Laboratory, Singapore Ministry of Education (MOE) Tier 1, under Grant RT5/23 and Grant RG24/24; in part by the Nanyang Technological University (NTU) Centre for Computational Technologies in Finance (NTU-CCTF); in part by the Research Innovation and Enterprise (RIE) 2025 Industry Alignment Fund—Industry Collaboration Projects (IAF-ICP) administered by the Agency for Science, Technology and Research (A*STAR), under Award I2301E0026; in part by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2020-0-01821) and the ITRC support program (IITP-2023-RS-2023-00258639), supervised by the IITP (Institute for ICT Planning & Evaluation); and in part by the National Natural Science Foundation of China (NSFC) under Grant 62572132. *(Corresponding author: Dong In Kim.)*

Ruichen Zhang, Guangyuan Liu, Yinqiu Liu, Changyuan Zhao, Jiacheng Wang, Yunting Xu, and Dusit Niyato are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: ruichen.zhang@ntu.edu.sg; liug0022@e.ntu.edu.sg; yinqiu001@e.ntu.edu.sg; zhao0441@e.ntu.edu.sg; jiacheng.wang@ntu.edu.sg; yunting.xu@ntu.edu.sg; dniyato@ntu.edu.sg).

Jiawen Kang is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: kavinkang@gdut.edu.cn).

Yonghui Li is with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: yonghui.li@sydney.edu.au).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, USA (e-mail: smao@ieee.org).

Sumei Sun is with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632 (e-mail: sunsm@i2r.a-star.edu.sg).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Dong In Kim is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon-si 16419, South Korea (e-mail: dongin@skku.edu).

Digital Object Identifier 10.1109/COMST.2026.3651702

I. Introduction

A. Background

THE rollout of sixth-generation (6G) wireless networks is ushering in a transformative era driven by the rapid expansion of edge-connected devices [1], [2], [3], [4]. According to IoT Analytics, the number of globally connected IoT devices will reach approximately 27.1 billion by 2025, increasing significantly from 16.6 billion in 2023.[1] Concurrently, Gartner forecasts that roughly 75% of enterprise-generated data will be processed at the network edge by 2025.[2] This explosive growth in edge connectivity has catalyzed a fundamental paradigm shift, moving intelligence from centralized cloud infrastructures toward decentralized edge intelligence [5], [6]. Edge intelligence has become integral to latency-sensitive and mission-critical applications, such as

[1] https://iot-analytics.com/number-connected-iot-devices/

[2] www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders

autonomous driving, industrial automation, drone swarms, and real-time healthcare monitoring, where centralized cloud solutions fail to meet stringent latency, reliability, and privacy requirements [7], [8], [9].

Despite its substantial benefits, traditional edge intelligence methods are predominantly based on static, task-specific models, designed primarily for single or narrowly defined tasks such as object detection or simple predictive analytics [10], [11], [12], [13]. For instance, classical edge intelligence-based drone swarms often employ fixed, predefined trajectory plans, unable to effectively adapt to sudden mission changes or environmental disturbances [14], [15], [16]. This inflexibility significantly limits operational effectiveness and safety in dynamic, multi-modal edge environments [17]. To overcome these limitations, the concept of edge general intelligence has emerged [18], [19], [20]. Edge general intelligence integrates broad cognitive capabilities, including multi-task generalization, continual learning, contextual understanding, and adaptive reasoning, directly into edge devices, enabling them to operate autonomously in complex and evolving environments [21], [22], [23]. In this context, the term "general" refers to the ability to adapt across diverse edge tasks and does not imply human-level artificial general intelligence. For example, edge general intelligence enables smart grid systems to dynamically adjust energy distribution under fluctuating loads, supports autonomous drone swarms in real-time mission adaptation, and facilitates seamless human–robot collaboration in dynamically changing factory environments [24], [25].

Nevertheless, implementing edge general intelligence presents several formidable challenges. Edge devices often operate under stringent constraints on computational power, memory capacity, and energy availability [26], [27], [28]. Furthermore, effective real-time scalability, robust multi-modal data processing, and long-horizon reasoning capabilities remain significant hurdles for current edge intelligence frameworks [29], [30], [31]. These inherent limitations necessitate a paradigm shift toward more robust, adaptive, and autonomous forms of intelligence, capable of fully leveraging edge-device potentials [32], [33].

In response to these critical challenges, Agentic artificial intelligence (Agentic AI) emerges as a transformative paradigm that redefines the capabilities of edge intelligence [34], [35]. *Agentic AI refers to intelligent systems characterized by continuous perception–reasoning–action loops, enabling autonomous context interpretation, explicit reasoning, and goal-driven decision-making [36].* This paradigm reflects the process of agentification, whereby passive edge entities are endowed with autonomy, adaptability, and proactive decision-making capabilities [37]. Unlike conventional edge intelligence that relies on static inference pipelines, Agentic AI incorporates generative frameworks, particularly large language models (LLMs), to dynamically fuse multi-modal data, plan actions, and coordinate distributed agents [38], [39], [40]. These capabilities enable edge devices to go beyond reactive responses, empowering them to anticipate, reason, and adapt within complex environments [41], [42]. At the same time, practical deployment often requires that these agentic behaviours operate under policy-aligned, human-in-the-loop, or fail-safe designs to ensure safety and accountability in real-world applications. For example, policy-aligned operation may require agents in autonomous driving or industrial automation to comply with certified safety rules and domain-specific operating constraints before executing decisions. Human-in-the-loop supervision is commonly mandated in scenarios such as medical diagnostics or public-sector decision assistance, where agents must obtain explicit approval before performing high-impact actions. Fail-safe designs are also essential, such as automatically reverting to conservative baseline behaviours or initiating safe-stop procedures when the agent detects uncertainty, conflicting objectives, or degraded sensor input.

Recent innovative frameworks exemplify Agentic AI's profound potential. For instance, AutoGPT autonomously decomposes complex objectives into executable subtasks, dynamically orchestrating decentralized toolchains to achieve sophisticated goals such as network resource optimization and real-time robotic missions [50]. Similarly, Voyager integrates long-horizon planning, contextual memory, and iterative self-improvement mechanisms, significantly enhancing uncrewed aerial vehicle (UAV) swarm coordination and environmental exploration capabilities [51]. Furthermore, Agentic AI has shown superior performance in vehicular networks, enabling autonomous vehicles to collaboratively adapt to dynamic traffic scenarios [8], [52], and in smart manufacturing environments, facilitating dynamic scheduling, predictive maintenance, and effective human–robot collaboration [53], [54]. These practical examples vividly demonstrate how Agentic AI significantly surpasses traditional edge intelligence paradigms in adaptability, generalization, and operational intelligence, thereby becoming indispensable for next-generation edge deployments [55], [56].

B. Contributions

Despite the clear potential and early successes of Agentic AI and its agentification process, comprehensive exploration and systematic deployment methodologies tailored explicitly for multi-modal Agentic AI in 6G-enabled edge environments remain limited [46], [57], [58]. Traditional AI architectures, such as static LLMs [59], mixture-of-experts (MoEs) [60], foundation models [61], and embodied AI frameworks [62] typically rely on one-way inference without fully considering the perception–reasoning–action loop [63], [64], [65]. Consequently, they lack sufficient adaptability, multi-modal integration, and cognitive autonomy necessary for robust operation in complex, evolving edge environments [66], [67]. This limitation motivates a deeper, systematic study into the key design principles required to deploy practical and scalable Agentic AI systems [68].

Table I summarizes the existing works, which address certain aspects of multi-modal Agentic AI; however, a unified methodology tailored for edge-oriented scenarios has yet to be established. Although many studies have explored Agentic AI across UAV autonomy, 6G architectures, and network control, most of them focused on isolated components or

TABLE I
SUMMARY OF RELATED WORKS

Ref.	Overview	Type	Agentic AI	Edge Intelligence	Wireless Networks
[43]	A multidomain survey of agentic UAVs integrating perception, memory, decision-making, and collaborative planning, mapping application domains and roadmaps for autonomous aerial ecosystems	Survey	✓	✗	✗
[44]	An article proposing an edge large ai model-empowered cognitive multimodal semantic communication agent that performs intent understanding and planning-based policy generation	Journal	✗	✓	✓
[45]	An overview of advanced 6G architectures integrating agentic AI, constrained-AI operations, serverless orchestration, and optical low-latency fabrics to cut operational expenditure and enable new services	Magazine	✓	✗	✓
[46]	A tutorial tracing the evolution from large AI models to agentic AI for intelligent communications, detailing core components (planner, tools, memory and knowledge base), multi-agent systems, and representative 6G applications	Tutorial	✓	✗	✓
[47]	An article proposing a generative foundation model-as-agent framework that supports interaction, collaborative learning, and knowledge transfer among agents for 6G networking, illustrated with digital-twin and metaverse scenarios	Journal	✓	✗	✓
[36]	An article introducing LLM-based agents with perception, memory, planning, and action for wireless tasks such as network slicing, achieving near-optimal throughput across diverse scenarios	Journal	✓	✗	✓
[48]	An edge agentic AI framework integrated into the O-Radio Access Network Intelligent Controller that combines persona-based multi-tool agents, predictive anomaly detection, and safety-aligned rewards for autonomous network optimization	Journal	✓	*Partially*	✓
[35]	A magazine article proposing Agent-as-a-Service, an AI-native edge framework in which agents plan, orchestrate, and manage 6G edge tasks via deviceless computing and webassembly	Magazine	*Partially*	✓	✓
[49]	A comprehensive survey introducing agentic to organize graph neural networks for scenario and task-aware wireless design, reviewing network applications (reconfigurable intelligent surface and cell-free) toward edge general intelligence	Survey	*Partially*	✓	✓
Ours	A comprehensive survey on Agentic AI frameworks for edge intelligence, introducing enabling technologies, representative case studies, and future directions toward scalable and trustworthy deployments in next-generation wireless edge networks	Survey+Tutorial	✓	✓	✓

use cases rather than a coherent deployment framework for edge environments. For example, Sapkota et al. [43] offered a multidomain survey of *agentic UAVs* with rich autonomy, yet their scope did not systematize edge intelligence or generic wireless architectures. At the application layer, Sun et al. [44] and Tong et al. [36] designed task-specific LLM-based agents for wireless tasks; however, these were not surveys and did not extract deployment methodologies for resource-constrained edges. At the architectural layer, Dev et al. [45], Xiao et al. [47], and Li et al. (i.e., Agent-as-a-Service) [35] discussed 6G frameworks that incorporated agentic elements, but they did not cover a tutorial-style design flow for multi-modal agents under tight edge constraints. O-RAN centric work by Salama et al. [48] bridged toward practical RAN control but only partially addressed edge intelligence, while Lu et al. [49] surveyed "agentic" graph neural networks from a graph-learning perspective rather than a general agent stack for edge general intelligence. Complementary tutorials such as Jiang et al. [46] traced the evolution toward agentic AI for communications; however, an integrative perspective on edge general intelligence that links agent capabilities to the networking stack, deployment sites (i.e., device, edge and cloud), and open toolchains remains underdeveloped.

In this survey, we aim to systematically explore critical design pillars and provide a comprehensive understanding of Agentic AI in the context of edge general intelligence. Differing from prior works summarized in Table I, which primarily examined isolated components, single-application agents, or high-level 6G frameworks, we articulate an *edge-oriented and multi-modal* methodology that links agent capabilities to the wireless networking stack, concrete deployment sites (device, edge, cloud), and reproducible system stacks with metrics and benchmarks. Specifically, we identify four foundational design principles that underpin effective Agentic AI deployment at the edge:

- **Compactness:** Developing lightweight Agentic AI models and their agentification processes that are resource-efficient enough to run on edge devices with strict energy and hardware constraints, while retaining sufficient cognitive expressiveness and autonomy [56], [69]. Typical pathways include small language models with Low-Rank Adaptation of LLMs (LoRA), distillation, and quantization, evaluated by parameter count, memory footprint, multiply–accumulate operations, and energy per inference.
- **Efficiency:** Ensuring real-time responsiveness through computationally efficient inference and communication-aware collaborative protocols that meet stringent latency and reliability requirements of edge environments [7], [70]. Key mechanisms

include early-exit inference, approximate decoding, task offloading, and bandwidth-conscious coordination; representative metrics include end-to-end latency, reliability, throughput, and cost under service-level targets.

- **Knowledge and Reasoning:** Incorporating explicit, interpretable, and context-sensitive reasoning capabilities, enabling agents to handle complex, long-horizon decision-making scenarios with confidence and transparency [71], [72]. Practical enablers include structured memory, retrieval-augmented generation, and tool use, assessed by task success, reasoning faithfulness, retrieval consistency, and explanation quality.
- **Migration:** Facilitating seamless transfer and reuse of knowledge, skills, and tasks across diverse and dynamically changing network conditions, enhancing generalization, robustness, and reducing retraining overhead [73], [74]. Techniques such as meta-prompting and structured retrieval, continual learning, and parameter-efficient adaptation are measured by zero/low-shot performance, sample efficiency, adaptation time, and forgetting rate.

In the remainder of this survey, we operationalize these principles into a unifying taxonomy, a tutorial-style design flow with decision checklists, consolidated benchmarks and metrics for agentic networking, and reusable case-study templates that demonstrate how to instantiate compact, efficient, knowledgeable, and migratory agents in edge environments. The primary contributions of this survey are structured around addressing the existing gaps and challenges in the systematic exploration and deployment of Agentic AI tailored explicitly for edge general intelligence within 6G-enabled networks. Specifically, we aim to provide comprehensive insights into the conceptual distinctions, foundational design principles, practical deployment scenarios, and future research opportunities for Agentic AI and its agentification process, facilitating robust and scalable intelligent edge systems. The key contributions of this paper are summarized as follows:

- We provide the first comprehensive survey and tutorial explicitly dedicated to multi-modal Agentic AI frameworks tailored for edge general intelligence within 6G-enabled networks. We clearly distinguish Agentic AI from traditional paradigms, including LLMs, MoEs, foundation models, and embodied AI frameworks, highlighting its unique characteristics and transformative potential.
- We systematically identify and elaborate on four foundational design pillars compactness, efficiency, migration, and knowledge & reasoning that define the essential capabilities and requirements for practical, scalable, and explainable Agentic AI systems and agentification.
- We illustrate the transformative impact of Agentic AI through concrete use cases involving cooperative UAV swarms, adaptive vehicular networks, and edge robotics, emphasizing practical deployment scenarios and performance advantages.
- We analyze emerging open-source frameworks and critically discuss unresolved research challenges and promising future directions. These insights provide actionable guidance to enable robust, trustworthy, and scalable Agentic AI deployments across heterogeneous edge environments.

By engaging with this comprehensive survey, the readers will gain valuable insights into how to effectively adopt Agentic AI frameworks tailored to their specific edge intelligence applications. Additionally, the readers will deepen their understanding of practical deployment challenges, including joint offloading strategies for LLM-based models, dynamic model migration techniques across heterogeneous edge devices, and efficient routing methods for multi-LLM service orchestration. Such detailed knowledge will empower researchers and practitioners to better navigate the complexities and opportunities inherent to next-generation edge environments, ultimately fostering the development and realization of robust, scalable, and intelligent Agentic AI solutions and agentification process.

C. Paper Organization

As depicted in Fig. 1, the remainder of this paper is organized as follows. Section II introduces the core concepts and frameworks of Agentic AI, emphasizing key capabilities and distinguishing it from traditional edge intelligence. Section III explores essential enabling technologies for Agentic AI and the agentification process, including compact models, energy-aware computing, robust connectivity, and advanced knowledge representation and reasoning techniques. Section IV reviews emerging open-source frameworks and toolkits. Section V presents representative Agentic AI applications in low-altitude economy, intent networking, vehicular networks, and human-centric service provisioning, supported by experimental analyses. Section VI highlights promising future research directions, and Section VII concludes the survey. For clarity, the main abbreviations used in this paper are summarized in Table II.

II. Background and Motivation

A. Edge General Intelligence

Edge general intelligence represents an emerging paradigm aiming to extend generalized, adaptive, and context-aware cognitive capabilities directly onto resource-constrained edge devices [80], [81]. Different from traditional edge intelligence, which primarily deploys task-specific, static models optimized for individual tasks (e.g., object detection or keyword spotting), edge general intelligence emphasizes versatility, adaptability, and autonomous cognitive reasoning [11], [82]. Edge general intelligence leverages foundation models, such as compact LLMs, MoE, or multi-modal neural architectures, to enable devices to perform multiple diverse tasks without frequent retraining, dynamically adapting to varying contexts, environments, and user preferences in real-time [69], [83]. In particular, reasoning capabilities, such as task decomposition, planning, and tool usage, are central to enabling goal-directed autonomy in dynamic edge environments. Such intelligent autonomy significantly reduces reliance on cloud-based resources, enhancing data privacy, operational efficiency, and user personalization [84], [85].

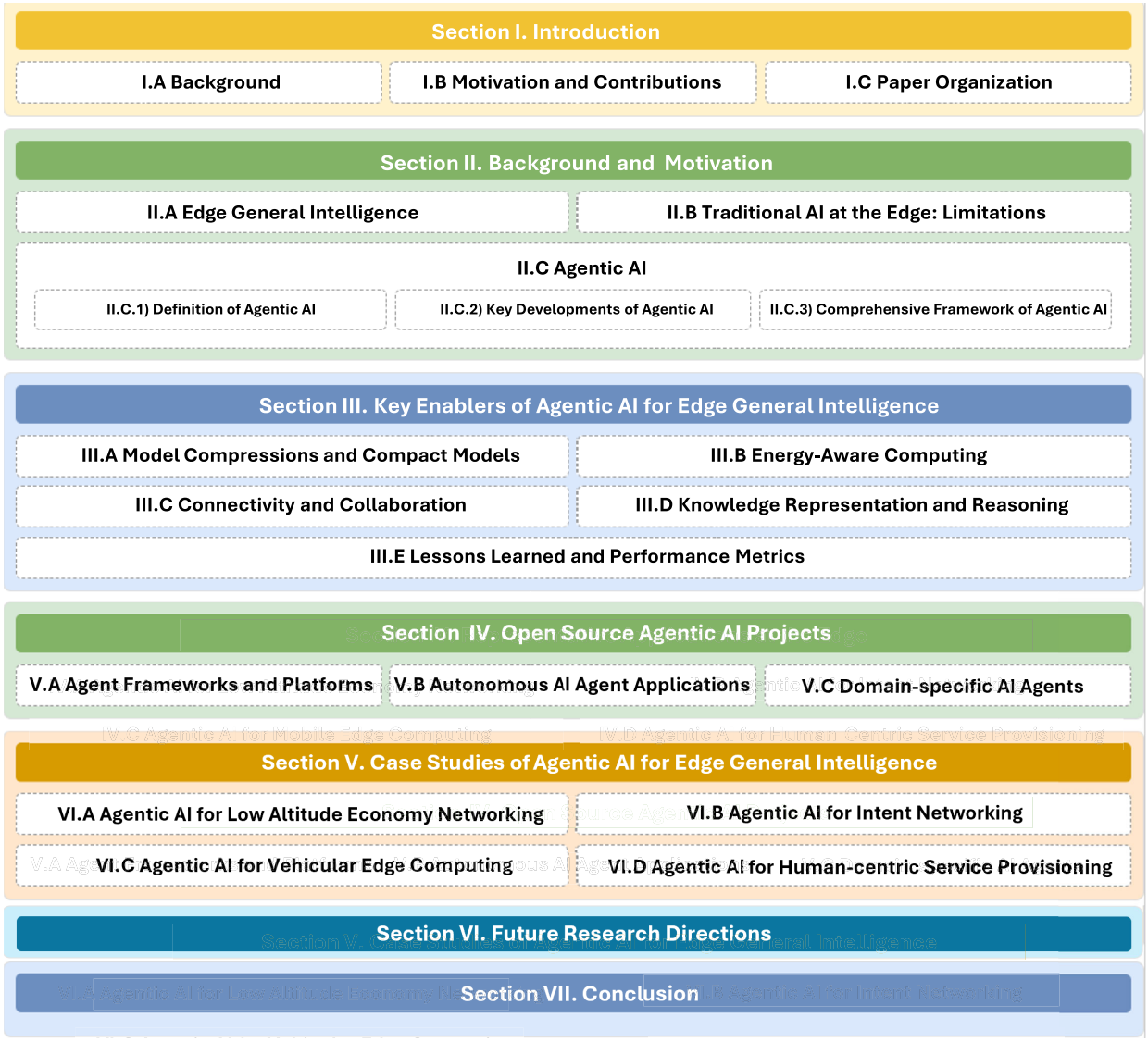

Fig. 1. Overall organization of this survey. We first introduce the evolution and core foundations of Agentic AI at the edge, followed by key enabling technologies. We then review emerging open-source frameworks and present representative application scenarios. Finally, the survey highlights future research directions and concludes the work.

Moreover, edge general intelligence systems increasingly integrate world-modeling capabilities, allowing agents to anticipate environmental dynamics and predict future states, which further strengthens their proactive planning and decision-making processes [55], [86].

By employing continual learning strategies, cognitive collaboration between edge agents, and robust multi-modal reasoning, edge general intelligence systems can continuously evolve and improve their cognitive capabilities throughout their deployment lifecycle [79], [80]. Typical application scenarios demonstrating the advantages of edge general intelligence include sophisticated smart home assistants capable of understanding complex user requests and contexts, and industrial IoT deployments that autonomously manage equipment maintenance, scheduling, and anomaly detection without extensive manual intervention or frequent model updates [11], [87], [88]. To clearly illustrate the differences between edge general intelligence and traditional edge intelligence, Table III provides a comprehensive comparison across key technical dimensions.

B. Traditional AI at the Edge: Limitations

Traditional edge AI systems have largely been designed for constrained, pre-defined tasks, such as object detection, speech recognition, or anomaly monitoring, operating under stable environments and with significant reliance on cloud infrastructure [89], [90], [91], [92]. These designs, while effective in early passive scenarios, fall short in meeting the stringent requirements of emerging 6G networks and the broader vision of edge general intelligence [93], [94], [95], [96]. In contrast to edge general intelligence, traditional edge intelligence lacks autonomy, adaptability, and context-aware reasoning capabilities essential for diverse, rapidly evolving operational environments [97], [98], [99], [100]. Specifically, the key limitations of traditional AI at the edge can be grouped into the following dimensions:

TABLE II
LIST OF ABBREVIATIONS

Abbreviation	Full Term
6G	Sixth Generation Wireless Networks
AI	Artificial Intelligence
Agentic AI	Agentic Artificial Intelligence
AGI	Artificial General Intelligence
IoT	Internet of Things
IoA	Internet of Agents
EGI	Edge General Intelligence
EI	Edge Intelligence
UAV	Unmanned Aerial Vehicle
RAN	Radio Access Network
O-RAN	Open Radio Access Network
RSMA	Rate Splitting Multiple Access
RIS	Reconfigurable Intelligent Surface
SWIPT	Simultaneous Wireless Information and Power Transfer
CRN	Cognitive Radio Network
CSS	Cooperative Spectrum Sensing
RSSI	Received Signal Strength Indicator
CSI	Channel State Information
QoS	Quality of Service
SINR	Signal to Interference plus Noise Ratio
LLM	Large Language Model
LLMs	Large Language Models
MoE	Mixture of Experts
RAG	Retrieval Augmented Generation
DRL	Deep Reinforcement Learning
RL	Reinforcement Learning
DQN	Deep Q Network
PPO	Proximal Policy Optimization
A3C	Asynchronous Advantage Actor Critic
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DNN	Deep Neural Network
GNN	Graph Neural Network
VLM	Vision Language Model
API	Application Programming Interface
DB	Data Base
CoT	Chain of Thought
GPU	Graphics Processing Unit

- **Heavy reliance on cloud connectivity and centralized control:** Traditional approaches frequently depend on cloud infrastructure for model inference and training updates, introducing significant latency, bandwidth bottlenecks, and single points of failure [101], [102], [103]. This centralized design severely limits scalability and robustness, especially in decentralized and latency-sensitive edge scenarios.
- **Limited adaptability to dynamic environments:** Traditional edge AI models are typically static, lacking mechanisms for continuous adaptation to changing network conditions or user behaviors [104], [105], [106]. Such static architectures face severe performance degradation in non-stationary edge environments, underscoring the need for more adaptive and continually evolving intelligent agents.
- **Scalability and real-time constraints under limited resources:** Conventional AI deployments at the edge often neglect tight resource constraints, such as limited memory, compute capacity, and power budgets, leading to inefficiencies in energy usage and operational responsiveness [45], [107]. To sustainably meet real-world demands of next-generation edge intelligence, future agents must integrate compact models, hardware-aware designs, and adaptive computational mechanisms.

These limitations highlight the need for a fundamental shift in edge intelligence, from cloud-reliant, static systems to intelligent agents that can operate autonomously and adaptively [108], [109]. This shift marks the evolution from the traditional Internet of Things (IoT), where devices merely sense and transmit, to a more proactive Internet of Agents (IoA) powered by Agentic AI, where edge nodes perceive, reason, plan, and act independently in real time [110], [111]. Next, we trace how AI agents have evolved toward this Agentic form.

C. Agentic AI

1) Definition of Agentic AI: Agentic AI refers to a new class of AI systems that exhibit goal-driven autonomy, operating in continuous *perception-reasoning-action* loops [112], [113]. Unlike conventional assistants that respond passively to user prompts, Agentic systems can proactively decompose high-level tasks, generate sub-goals, plan actions, and interact with external tools or environments with minimal human input [114], [115]. Powered by foundation models, these agents are designed to reason, act, and adapt over time. As described by IBM and Deloitte,[3] Agentic AI systems are capable of completing complex workflows and achieving objectives with little or no human supervision [116]. Beyond individual autonomy, Agentic AI and agentification process often involve *agent orchestration*, the coordinated interaction among multiple agents with specialized roles, enabling complex task execution through modular collaboration [117], [118], [119]. Such orchestration allows agents to dynamically communicate, delegate subtasks, and synthesize partial outputs, forming a distributed problem-solving network especially suited for edge-centric, decentralized environments [120].

These Agentic AI systems are characterized by several core capabilities [114]. First, they exhibit autonomy, which enables decision making and action initiation, exemplified by UAVs navigating uncertain terrains or edge robots adapting to task variations. Second, they possess contextual memory and adaptability, allowing them to learn from past interactions and effectively respond to dynamic conditions such as those found in vehicular or industrial networks. Third, they support explicit reasoning and planning, utilizing external tools, APIs, or supplementary models for executing long-term strategies, as demonstrated in frameworks like ReAct [121] and Toolformer [122]. Lastly, they facilitate modular collaboration by orchestrating toolchains across decentralized environments, supporting scalable deployment through platforms including HuggingGPT [123] and LangGraph.[4] [124].

By design, Agentic AI addresses key limitations associated with traditional edge intelligence. It reduces dependency on cloud connectivity through localized inference and planning,

[3]https://www.ibm.com/think/topics/Agentic-ai
[4]https://github.com/langchain-ai/langgraph

TABLE III
COMPARISON OF EDGE GENERAL INTELLIGENCE AND TRADITIONAL EDGE INTELLIGENCE

Feature	Edge General Intelligence	Traditional Edge Intelligence
Generalization	**Multi-task capability** Supports multiple diverse tasks without retraining (vision, NLP, decision-making)	**Task-specific models** Designed specifically for single tasks (object detection, activity recognition)
Adaptability	**Dynamic adaptation** Learns and adapts dynamically at runtime	**Static behavior** Requires manual retraining or updates
Model Architecture	**Compact general models** Compact LLMs, mixture-of-experts (MoE), or foundation models optimized for edge [75]	**Specialized models** Small CNNs, RNNs, or DNNs optimized per task
Multi-Modality	**Multi-modal processing** Handles text, images, audio, sensor fusion simultaneously	**Single-modal processing** Processes one modality at a time (e.g., images or sensors) [76]
Autonomy & Reasoning	**Autonomous reasoning** Independent decision-making with minimal cloud support	**Inference-driven** Executes predefined tasks with limited autonomy
Continual Learning	**Continuous learning** Supports lifelong or federated learning directly on device	**Limited online learning** Rarely supports online learning due to resource constraints
Communication Dependency	**Low cloud dependency** Reduced reliance on cloud, enhanced local processing [77]	**High cloud dependency** Relies heavily on cloud for complex tasks
Personalization	**Dynamic personalization** Automatically adjusts to user preferences [78]	**Limited personalization** Requires manual fine-tuning
Cognitive Collaboration	**Collaborative cognition** Shares knowledge collaboratively with other agents	**Isolated cognition** Operates independently or strictly cloud-controlled [79]
Security & Privacy	**Enhanced local privacy** Increased privacy via general-purpose on-device cognition	**Cloud-dependent privacy** Security contingent on data transmitted to cloud

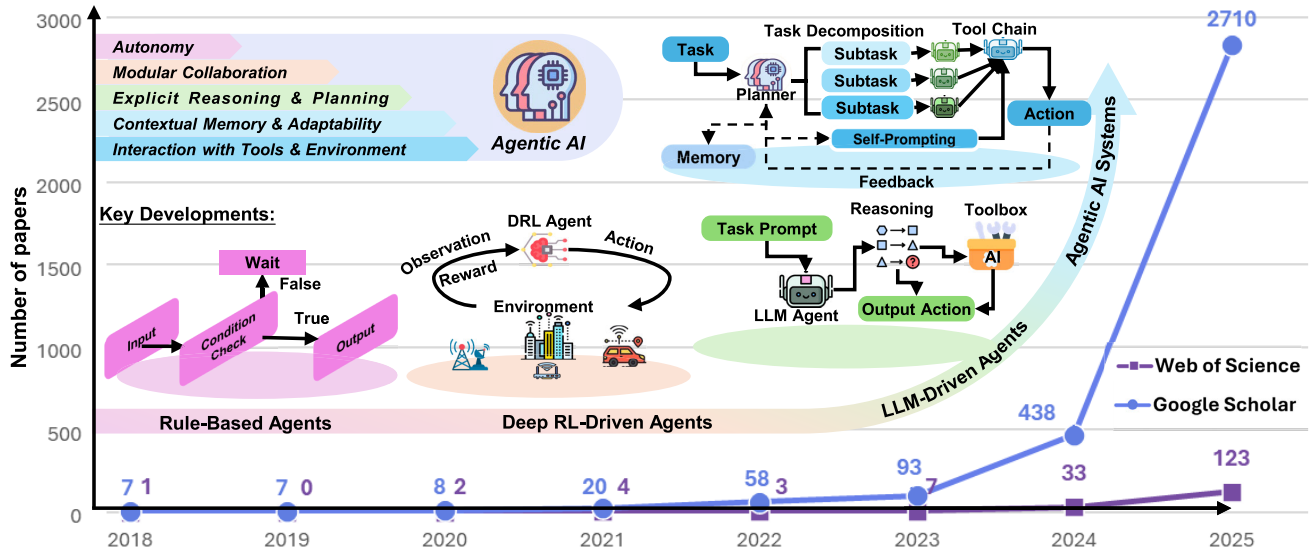

Fig. 2. Illustration of key developments and evolution trajectory of Agentic AI systems, from early rule-based approaches, through DRL-driven agents, towards current LLM-driven agents. Highlighted are core capabilities such as autonomy, modular collaboration, explicit reasoning and planning, contextual memory and adaptability, and interaction with tools and environments. Recent trends emphasize task decomposition and self-prompting for robust reasoning and action execution, indicating substantial growth in research and deployment in the coming years.

enhances adaptability via memory-driven continuous learning, and improves computational efficiency by employing model compression and task-aware reasoning [46], [125]. These characteristics position Agentic AI as a foundational enabler for the IoA, converting passive devices into intelligent, self-directed agents capable of handling the complexities inherent to 6G networks and beyond [126], [127].

2) Key Developments of Agentic AI: The emergence of Agentic AI and its agentification process are rooted in a multi-stage evolution, moving from simple automation toward increasingly autonomous, memory-enabled, and context-aware intelligent systems [36]. As shown in Fig. 2, this evolution progresses through several distinct stages: from basic *rule-based agents* to adaptable *DRL-driven agents*, further evolving into sophisticated *LLM-driven agents*, and ultimately culminating in fully autonomous *Agentic AI systems*. Each stage signifies enhanced cognitive capability and improved adaptability, addressing the growing complexity of wireless and edge environments [130], [131], [132].

TABLE IV
COMPARISON OF AI AGENT PARADIGMS IN WIRELESS EDGE INTELLIGENCE

Attribute	Rule-Based Agents	DRL-Driven Agents	LLM-Driven Agents	Agentic AI Systems
Core Intelligence	Finite-state machines, deterministic logic	DQN, PPO, A3C	Pre-trained transformers (GPT-2/3, Codex)	LLM + memory + DRL planner
Autonomy	✗ Manual, reactive	Task-level	Prompt-level autonomy	✓ Goal-level autonomy
Perception Modality	Single-modal (e.g., RSSI, CSI)	Environment features (QoS, SINR)	Text, code, limited images	Multi-modal (vision, RF, state/action)
Memory Scope	None (stateless)	Short-term (recurrent states)	Short-term buffer (few-shot)	Long-term (episodic vector DB)
Planning & Reasoning	None (fixed rules)	MDP-based finite-horizon policy	Prompt chaining, CoT-style reasoning	Deliberative planning, causal reasoning, recursive loops
Wireless Application	CRN cooperative sensing, e.g., OR-rule CSS [128]	RIS-SWIPT beamforming via PPO [129]	LLM-assisted RAN control with MoE-PPO [74]	AutoGPT/Voyager for UAV control, RSMA spectrum negotiation [74], [51]
Limitations	No adaptation, static logic, poor scalability	Domain-specific, lacks abstraction or transferability	Limited memory, external tool dependence, task fragility	Higher cost, safety/policy alignment, runtime constraints

- **Rule-Based Agents:** These early agents rely on predefined rules or finite-state machines, limiting their adaptability and autonomy. Such systems operate reactively and are primarily effective in static and narrowly defined scenarios, thus falling short in dynamic wireless environments [133].
- **Deep RL-Driven Agents:** Agents driven by deep reinforcement learning (DRL) enhance adaptability through trial-and-error interactions with the environment. However, their applicability remains constrained by task specificity, lacking broader generalization and explicit reasoning capabilities across diverse scenarios [134].
- **LLM-Driven Agents:** LLM-driven agents primarily focus on tool-use assistance, leveraging large-scale language models such as GPT-2 and GPT-3 [59] as cognitive cores to facilitate general reasoning, multi-step planning, and modular tool interaction. Representative frameworks including Codex [135], ReAct [72], and Toolformer [136] exemplify this paradigm through explicit reasoning traces and structured tool orchestration. These agents rely on language models to interpret goals and trigger external tools, forming reactive and loosely coupled decision-making loops. In wireless communications, Zhang et al. [74] introduced an LLM-based framework combining retrieval-augmented generation (RAG) and MoE-enhanced Proximal Policy Optimization (MoE-PPO) for satellite–terrestrial systems. This approach attained a 95.3% retrieval accuracy and improved throughput by 42.6% compared to conventional SDMA, demonstrating the practical advantages of LLM-assisted agents in semantic–physical layer alignment.
- **Agentic AI Systems:** Agentic AI systems embody autonomous agents with tightly integrated perception–reasoning–action loops, enabling persistent interaction with dynamic environments. Through the process of agentification, these systems incorporate autonomy, contextual memory, symbolic reasoning, and collaborative modularity into closed-loop architectures for long-term planning and proactive decision-making. Recent open-source frameworks such as AutoGPT, BabyAGI, and Voyager [51] illustrate this transition from reactive tool-use agents to self-directed cognitive entities with recursive task decomposition, adaptive self-prompting, and continual feedback integration. In 6G networking, Zhang et al. [56] extended such principles to optimize resource policies via autonomous agents capable of context-aware retrieval and goal-conditioned reasoning.

To highlight the evolving design philosophies of AI agents in wireless systems, Table IV compares four major paradigms, i.e., rule-based agents, deep RL-driven agents, LLM-driven agents, and Agentic AI, across key technical dimensions including core intelligence, autonomy, perception modality, memory scope, planning and reasoning capabilities, representative wireless applications, and known limitations.

3) Comprehensive Framework of Agentic AI: Beyond understanding its foundational definition, it is essential to conceptualize Agentic AI through its comprehensive architecture, core capabilities, and primary functional components [139]. As illustrated in Fig. 3, a complete Agentic AI framework integrates several interconnected modules, enabling autonomous perception, reasoning, planning, and effective action execution. Concretely, the architecture is organized into four modules, i.e., Perception, Memory, Reasoning, and Action, with an explicit memory management and retrieval layer that binds them together, and with execution either on-device or offloaded to the edge/cloud [140], [141]. Specifically, the Agentic AI with agentification process operates through a continuous cycle, beginning from external data and environmental perception, moving through comprehension and reasoning stages, and culminating in adaptive actions that are continuously refined through feedback loops [29], [54], [142]. The core components of the Agentic AI framework are outlined below, supported by concrete examples and recent research insights:

- **Perception Module:** This module integrates multi-modal data, including textual, visual, and auditory inputs, allowing agents to perceive and understand complex

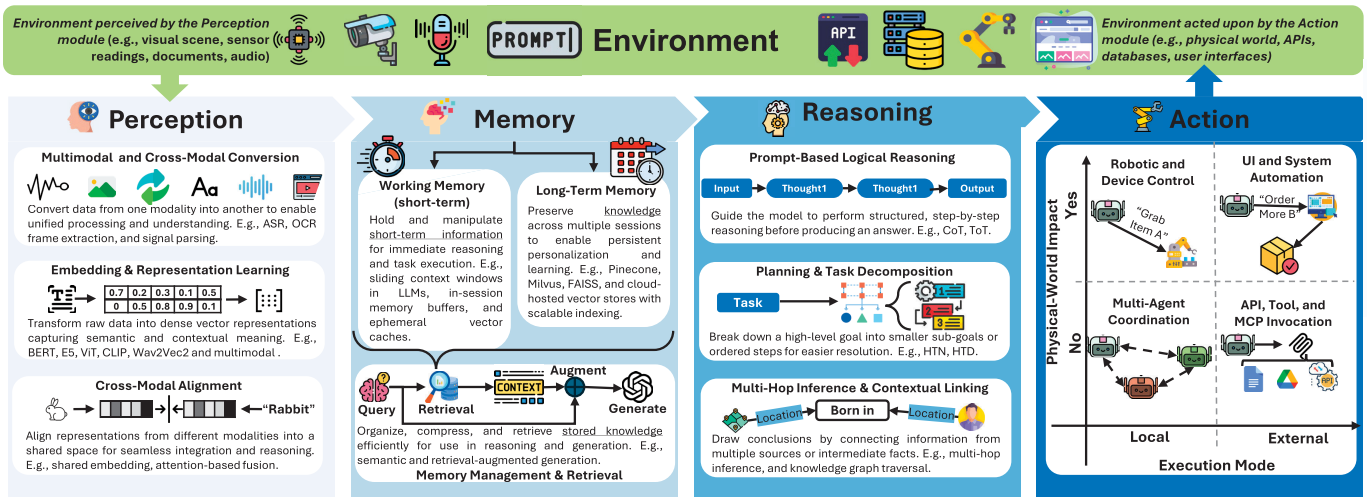

Fig. 3. Comprehensive workflow of Agentic AI for edge deployments. The pipeline comprises four modules: *Perception* (i.e., multi-modal acquisition, cross-modal conversion, preprocessing, feature extraction, and embedding for a unified scene); *Memory* (i.e., working, episodic, semantic, and long-term stores with management and retrieval supporting vector caches and RAG); *Reasoning* (i.e., prompt-based logic, planning and decomposition, multi-hop/context linking, uncertainty-aware, and neuro-symbolic inference); and *Action* (i.e., robot or device control, multi-agent coordination, API or tool or MCP invocation, and UI or system automation) [137], [138]. Execution may be local on device or external at edge or cloud, with environmental feedback closing the loop. The figure also indicates where the four design principles apply: compactness, efficiency, knowledge and reasoning, and migration.

environments comprehensively. For instance, autonomous vehicles leverage multi-modal sensor fusion, combining LiDAR, radar, camera images, and traffic signals, to accurately interpret real-time road conditions, pedestrian behaviors, and traffic patterns, thereby ensuring reliable and context-aware navigation [112], [114].

- **LLMs:** LLMs such as GPT-4 and Gemini function as cognitive cores, delivering rich semantic comprehension and sophisticated reasoning abilities. These capabilities enable agents to interpret complex instructions, decompose high-level tasks, and generate structured plans. For example, AutoGPT utilizes GPT-4 to autonomously interpret high-level commands, decomposing them into detailed subtasks, such as generating business strategies or optimizing complex workflows, thereby significantly reducing human supervision and enhancing operational efficiency [50], [135].
- **External Tools and APIs:** Agentic AI seamlessly interact with external tools and APIs to perform actions that extend beyond their inherent cognitive capabilities. For example, Toolformer integrates API calls directly within the reasoning process, allowing agents to dynamically access external computational resources, such as mathematical computation APIs, databases, and specialized knowledge repositories, thereby facilitating complex problem-solving in real-time scenarios [72], [136].
- **Memory and Retrieval:** Memory components, particularly RAG mechanisms, enable agents to continuously learn and retain historical knowledge effectively [143]. For instance, recent research by Wang et al. [71] developed memory-augmented neural networks utilizing RAG techniques to dynamically retrieve contextually relevant information from vectorized knowledge bases. This approach significantly enhanced model adaptability and reasoning accuracy across diverse tasks, effectively supporting agentic decision-making in complex scenarios.
- **Planning and Reasoning:** Explicit planning capabilities, including Chain-of-Thought (CoT) [144], [145] reasoning and symbolic AI techniques, empower agents to formulate and assess long-horizon strategies autonomously. A notable example is the ReAct framework [72], which combines language-driven reasoning with action planning. By integrating CoT-based reasoning methods, ReAct allows agents to systematically break down complex goals into manageable subtasks, dynamically evaluating potential outcomes, and selecting optimal actions, greatly enhancing their ability to manage sophisticated, real-time decision-making scenarios.
- **Multi-Agent Coordination:** Multi-agent frameworks leveraging Deep DRL facilitate decentralized coordination, collaborative decision-making, and emergent collective intelligence. For instance, Tong et al. [36] introduced a Multi-Agent DRL system incorporating model context protocol (MCP) [146] to enhance decentralized spectrum management in wireless networks. MCP supports efficient inter-agent communication by maintaining consistent context representations across distributed agents, significantly improving network throughput, reducing latency, and demonstrating the robustness and scalability of decentralized Agentic AI coordination.

Specifically, as shown in Fig. 4, these components collaborate within an integrated and iterative workflow as follows: Initially, the Perception Module captures external multi-modal data and environmental context [56]. Next, the Comprehension stage, driven by foundation models and LLMs, processes and interprets this data, providing rich semantic understanding and structured contextual insights. These insights inform the Self-Planning stage, wherein agents autonomously formulate

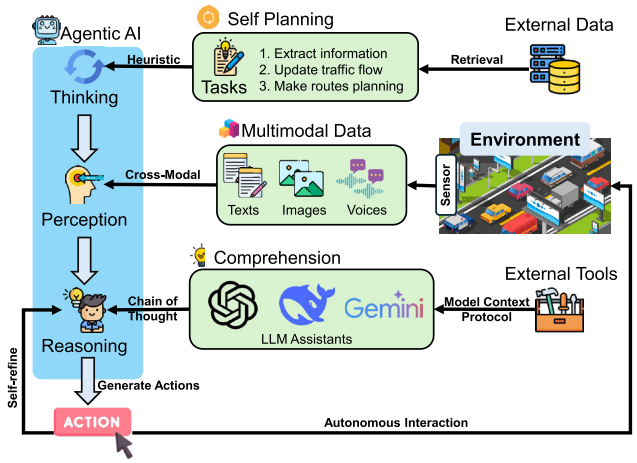

Fig. 4. Conceptual workflow illustrating how Agentic AI autonomously integrates self-planning, multi-modal perception, comprehension via foundation models, and external tools for continuous perception–reasoning–action agentification process in edge general intelligence systems [56].

tasks, strategies, and action plans. In the subsequent Reasoning stage, explicit planning mechanisms (e.g., CoT reasoning) and long-term memory retrieval (e.g., RAG) refine these plans further, incorporating additional contextual insights and prior knowledge. Finally, autonomous actions are executed using External Tools and APIs, with outcomes continuously evaluated and fed back into the loop for iterative self-refinement. This cohesive agentification loop is particularly crucial in edge environments, where real-world constraints such as limited computing power, energy budgets, and strict latency requirements pose significant deployment challenges for Agentic AI [147]. By integrating compact model deployment, energy-efficient computing, robust connectivity, and adaptive knowledge reasoning into each module, this framework effectively addresses such constraints and enables resilient edge intelligence.

By integrating these components within this coherent and iterative perception-reasoning-action agentification process, Agentic AI establishes a robust cognitive architecture capable of autonomously adapting to dynamic edge environments [45], [46], [148]. This comprehensive integration addresses fundamental limitations of traditional edge intelligence, significantly advancing the realization of edge general intelligence and paving the way toward resilient and adaptive intelligent systems for next-generation edge deployments [149], [150], [151].

III. Key Enablers of Agentic AI for Edge General Intelligence

To enable Agentic AI in the edge general intelligence, several key technological enablers must be addressed, including compact model deployment, energy-efficient computing, robust connectivity and collaboration, and effective knowledge representation and reasoning [70], [152]. As illustrated in Fig. 5, these enablers are highly interdependent, forming a virtuous cycle that drives scalable and autonomous edge intelligence. These enablers collectively support edge general intelligence, ensuring agents can autonomously adapt and reason effectively in dynamic and resource-constrained environments [35], [153].

A. Model Compressions and Compact Models

Deploying Agentic AI models directly onto resource-constrained edge devices presents significant challenges due to stringent memory and computational constraints [154]. To overcome these limitations while preserving the essential reasoning and adaptability capabilities of Agentic AI, model compression techniques such as pruning, quantization, low-rank factorization, and knowledge distillation become indispensable [155], [156]. These compact model strategies enable the deployment of sophisticated foundation models and LLMs within edge environments, thereby supporting the perception and reasoning modules of Agentic AI without exceeding computational budgets. Architectures explicitly designed for efficient edge inference, such as MobileNet and ShuffleNet, further reduce latency and energy consumption by employing depthwise separable convolutions [157], [158]. In this context, generalization accuracy serves as a key indicator of model compactness, reflecting the ability of a compressed model to maintain performance across diverse tasks and deployment scenarios. In mobile and edge environments, ensuring high generalization accuracy is critical since downsized models must adapt to varying data distributions and environmental dynamics. While pruning, quantization, and distillation substantially reduce model size and complexity, they must be carefully tuned to preserve the model's capacity to generalize without incurring a significant performance drop [112], [159], [160].

1) LoRA: Low-Rank Adaptation (LoRA) reduces the complexity of adapting large, pre-trained models by inserting small, trainable matrices into frozen weights. LoRA enables efficient adaptation with negligible performance degradation by introducing low-rank updates into frozen pretrained weights instead of retraining the full model. For example, Hu et al. [161] proposed LoRA, demonstrating that inserting low-rank decompositions into Transformer layers can significantly decrease the number of trainable parameters by up to 10,000 times, without compromising performance. Specifically, experiments on GPT-3 (175B parameters) showed that LoRA achieved competitive or superior performance compared to full fine-tuning, while also reducing GPU memory consumption by approximately threefold. By significantly compressing these models, LoRA directly supports the efficient integration of powerful LLMs within Agentic AI systems, facilitating advanced semantic comprehension and reasoning on resource-limited edge hardware.

2) Knowledge Distillation: Knowledge distillation further supports Agentic AI by transferring intricate reasoning capabilities, such as chain-of-thought reasoning, from large "teacher" models to compact "student" models optimized for edge deployment. Knowledge distillation compresses models by approximately 20–40% by transferring knowledge from a large teacher model to a smaller student model, enabling the student to approximate the teacher's behavior while maintaining high reasoning fidelity [162]. For instance,

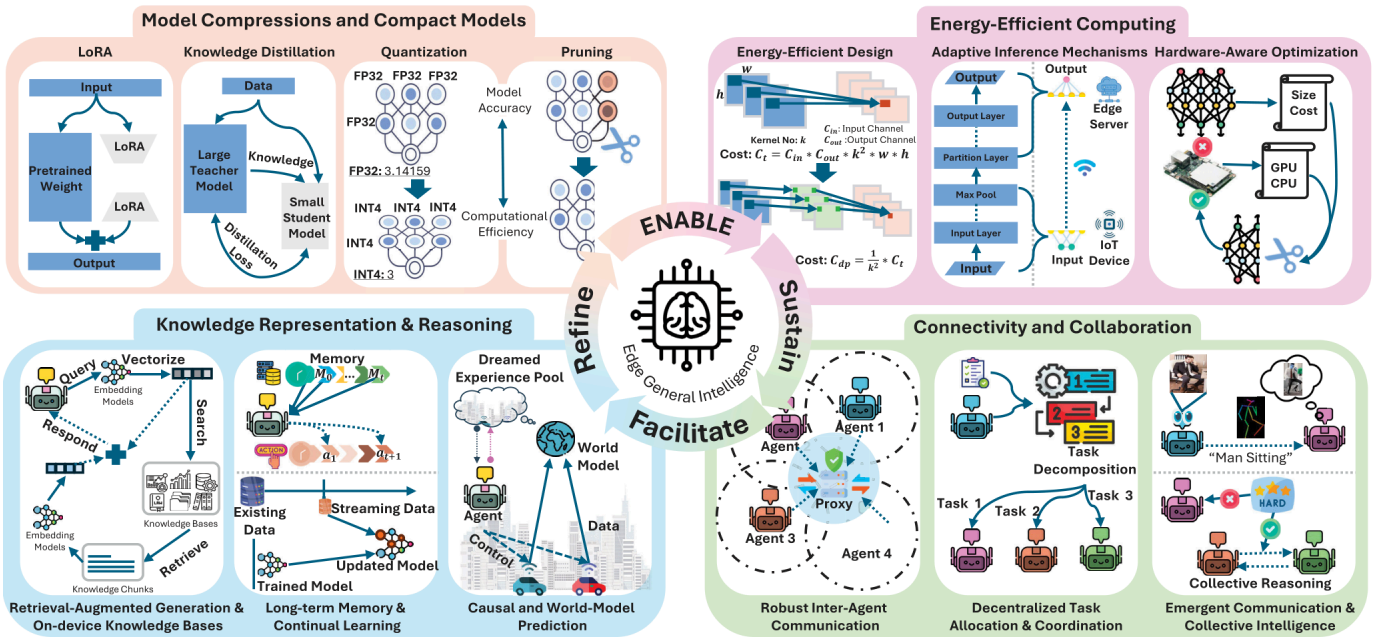

Fig. 5. Interdependencies Among Key Enablers of Agentic AI for edge general intelligence. Compact model techniques enable efficient execution under tight resource constraints, thereby sustaining energy-aware computing. In turn, energy efficiency sustains continuous operation and supports robust connectivity and collaboration. These collaborative mechanisms facilitate distributed knowledge representation and reasoning. Finally, advanced reasoning capabilities guide the design of increasingly compact and adaptive models, forming a virtuous cycle that drives scalable and autonomous edge intelligence.

Li et al. [163] introduced a prompt-based distillation approach specifically targeting complex reasoning behaviors in LLMs. Their methodology effectively distilled multi-step reasoning capabilities into smaller models such as Llama2 (7B parameters) and CodeLlama, achieving accuracies of 85% and 85.5%, respectively, on the challenging SVAMP arithmetic reasoning dataset. These results surpassed GPT-3.5-turbo, demonstrating that complex cognitive reasoning can be effectively condensed into compact architectures suitable for edge environments. This approach is particularly relevant for enhancing the efficiency of the planning and reasoning module within Agentic AI, enabling high-level reasoning and explicit planning capabilities in constrained edge scenarios.

3) Quantization Methods: Quantization methods compress models by 50–75% by reducing weight and activation precision from full precision (e.g., FP32) to low-bit formats (e.g., INT8 or INT4), thereby lowering memory and computation requirements with minimal accuracy loss when properly calibrated [164]. For instance, Lin et al. [165] proposed Activation-aware Weight Quantization (AWQ), specifically designed for quantizing LLMs with minimal performance degradation. They demonstrated lossless performance across 11 vision-language benchmarks, with INT4-g128 quantization settings applied to models such as VILA-7B and VILA-13B fully matching their original full-precision counterparts. These quantization techniques are crucial for enabling efficient real-time multi-modal perception within the Agentic AI perception module and comprehensive interpretation within LLMs, maintaining the rich cognitive functionalities essential to Agentic AI deployments in resource-constrained environments.

4) Pruning Methods: Pruning techniques reduce model size by 10–90% by systematically removing less important weights, neurons, channels, or attention heads, thereby introducing sparsity while preserving key functionality and reasoning capability when carefully tuned [166]. Structured pruning techniques such as SparseGPT and LLM-Pruner systematically remove redundant neurons, channels, or attention heads, creating efficient models while maintaining their functional integrity and reasoning performance. For instance, Ma et al. [167] introduced LLM-Pruner, the first structured pruning framework explicitly designed for LLMs. Using only 50,000 training samples and three hours of fine-tuning, they achieved parameter reduction of up to 20% while preserving over 94% of the original model's performance. Pruning methods play a critical role in supporting the efficient integration of comprehensive reasoning, memory and retrieval mechanisms (e.g., vectorized databases and RAG), and autonomous multi-agent coordination, enabling compact yet highly capable Agentic AI models that thrive in decentralized and collaborative edge deployments.

B. Energy-Aware Computing

The deployment of Agentic AI on resource-constrained edge devices inherently demands energy-aware computing strategies [168]. Agentic AI models, characterized by autonomous reasoning, proactive decision-making, and long-term operational autonomy, require consistent and efficient operation under strict power budgets and thermal constraints [169]. In this regard, resource consumption emerges as a key metric for evaluating the efficiency of edge-deployed AI models [170]. This metric encompasses energy usage, computational load, memory footprint, and communication overhead, all of which critically impact real-time performance and sustainability at the edge [171]. Notably, minimizing communication overhead is essential in mobile environments where models must

TABLE V
MODEL COMPRESSION TECHNIQUES FOR EDGE GENERAL INTELLIGENCE

Technique	Ref	Description	Pros	Cons
LoRA	[161]	Inserts small trainable matrices into large frozen models	• Major parameter reduction • Lower memory usage • Good performance	• Insertion complexity • Accuracy sensitive to rank
Knowledge Distillation	[163]	Transfers reasoning from large teacher to compact student models	• Preserves complex reasoning • Improves small-model performance	• Depends on teacher quality • Possible knowledge loss
Quantization Methods	[165]	Reduces parameter precision (e.g., INT4) to optimize efficiency	• Minimal accuracy loss • High edge efficiency	• Sensitive at ultra-low precision • Needs calibration
Pruning Methods	[167]	Removes redundant neurons or attention heads structurally	• Effective sparsity (up to 20%) • Maintains high accuracy	• Performance loss at high sparsity • Pruning strategy required

prioritize local processing to reduce reliance on cloud offloading and avoid costly data transmission.

1) Energy-Efficient Model Design: Efficient execution of Agentic AI on edge hardware mandates model architectures specifically designed to minimize computational and energy footprints. Compact model architectures such as MobileNet [172], ShuffleNet [173], and EfficientNet [174] leverage depthwise separable convolutions, channel shuffle operations, and compound scaling strategies, significantly reducing energy consumption and latency without compromising reasoning performance. Recent advancements extend these concepts to Transformer-based architectures optimized for energy efficiency. For example, MobileViT [175] effectively integrates vision transformers into edge-friendly models, enabling sophisticated visual reasoning at minimal energy cost, thus driving the practical realization of edge general intelligence.

2) Adaptive Inference Mechanisms: Agentic AI benefits profoundly from adaptive computation strategies that dynamically adjust computational resources in response to input complexity and environmental constraints. Techniques such as dynamic neural networks [176] and multi-exit architectures [177], [178] enable conditional execution of neural pathways or early termination of inference based on confidence levels, significantly reducing redundant computations. For instance, recent work demonstrates adaptive early-exit schemes achieving up to 24.6% latency and 46.5% energy consumption reductions compared to state-of-the-art IoT ML inference, adaptively distributing computation between devices and edge servers without accuracy loss [179]. Such adaptive mechanisms are particularly vital for enhancing the efficiency of the planning and reasoning processes within Agentic AI, ensuring that high-level, explicit reasoning and complex decision-making tasks can be executed within real-time constraints of edge environments.

3) Hardware-Aware Optimization: Effective integration of Agentic AI into edge general intelligence requires optimized software-hardware co-design, aligning computational tasks explicitly with the capabilities of specialized edge accelerators (e.g., NPUs, TPUs, and VPUs). Techniques such as hardware-aware neural architecture search (HW-NAS) [180], [181] and targeted model pruning [182] adapt neural network structures to specific accelerator architectures, exploiting hardware strengths and minimizing costly memory transfers and computations. Additionally, dynamic voltage and frequency scaling (DVFS) coupled with flexible invocation-based deep reinforcement learning [183] enables flexible adjustment of agent invocation intervals and task scheduling, achieving a 55.1% reduction in agent invocation cost and up to 23.3% overall energy consumption reduction. These hardware-aware optimizations are essential for efficiently running computationally intensive components such as memory and retrieval (e.g., RAG), the execution of external tools and APIs, and supporting robust decentralized operations essential for multi-agent coordination, collectively contributing to the energy-efficient and scalable deployment of Agentic AI in edge general intelligence contexts.

C. Connectivity and Collaboration

Effective collaboration and robust connectivity are foundational to Agentic AI systems, enabling decentralized agents to seamlessly cooperate, share intelligence, and execute complex tasks at the edge [184], [185], [186]. Given the inherently distributed and dynamic nature of edge general intelligence, robust and efficient communication protocols, collaborative decision-making algorithms, and adaptive coordination strategies become critical for the scalable and reliable deployment of intelligent agents in real-world environments [187], [188]. In this context, we adopt accessibility as a key metric to quantify an Agentic AI system's ability to function effectively in resource-constrained and heterogeneous environments. Accessibility reflects how easily models can be deployed across varying hardware platforms with minimal reliance on centralized infrastructure. It encompasses dimensions such as deployment flexibility, cross-device operability, and user inclusiveness [189].

1) Robust Inter-Agent Communication: Reliable communication under intermittently connected or bandwidth-constrained edge scenarios is crucial for coordinating actions among distributed agents. Recent studies have focused on low-overhead, resilient communication

TABLE VI
CONNECTIVITY AND COLLABORATION TECHNIQUES FOR EDGE GENERAL INTELLIGENCE

Technique	Ref	Description	Pros	Cons
Robust Inter-Agent Communication	[190], [191], [192]	Gossip algorithms, federated learning, and sparse message passing for robust and efficient communication	• Resilience to failures • Low bandwidth usage • Efficient dissemination	• Message latency • Potential slow convergence
Decentralized Task Allocation and Coordination	[193], [194], [195], [196]	Multi-agent RL, distributed constraint optimization, and graph neural networks for decentralized decision-making	• Dynamic adaptability • Scalable coordination • No central control required	• Training complexity • Coordination difficulty
Emergent Communication and Collective Intelligence	[193], [197]	MARL-driven autonomous development of concise and adaptive communication protocols	• Reduced overhead • Efficient semantics • Enhanced scalability	• Complex training • Difficult interpretability

protocols such as gossip-based algorithms [190], federated learning protocols [191], and sparse message-passing schemes [192], which effectively propagate information through the network with minimal redundancy. These methods enhance the robustness of Agentic AI systems against link failures, bandwidth fluctuations, and intermittent connectivity, thus maintaining collective intelligence and enabling seamless collaboration in edge general intelligence frameworks. Such robust communication protocols directly support the efficient integration and coordination of the perception module and the effective exchange of multi-modal data processed by LLMs, ensuring reliable information sharing across distributed agent networks under challenging conditions [198].

2) Decentralized Task Allocation and Coordination: Agentic AI deployed in edge environments necessitates decentralized and self-organizing mechanisms for task distribution and resource allocation. Techniques from multi-agent reinforcement learning [193], [194], [199], distributed constraint optimization [195], and graph neural network-based coordination methods [196] have been successfully leveraged for decentralized decision-making. For instance, graph-based coordination frameworks allow distributed agents to perform collaborative inference and task allocation without centralized control, dynamically adapting to environmental changes and agent availability, thus significantly advancing the autonomy and scalability of edge general intelligence [200]. These decentralized coordination techniques directly enable effective planning and reasoning among distributed agents, while also optimizing the invocation of external tools and APIs, allowing for efficient resource usage and improved collective performance in dynamic edge scenarios.

3) Emergent Communication and Collective Intelligence: Enabling Agentic AI systems to autonomously develop efficient, concise, and adaptive communication languages or signaling mechanisms greatly enhances their collaborative capabilities at the edge. Recent research on emergent communication [193], [197] demonstrates that agents can autonomously learn shared languages optimized for minimal communication overhead while efficiently encoding task-relevant semantics. To design such systems, multi-agent reinforcement learning (MARL) frameworks, such as those demonstrated in [193], can be employed to train agents to optimize communication protocols, which involves using discrete message spaces to ensure conciseness and defining reward structures that balance task performance with communication efficiency, such as rewarding task completion while penalizing excessive messaging. Additionally, integrating lightweight attention-based modules ensures efficient communication on resource-constrained edge devices. Such emergent collective intelligence paradigms allow edge-deployed agent groups to perform complex tasks collaboratively with minimal communication resources, significantly reducing energy usage and enhancing scalability, thereby promoting robust and adaptive edge general intelligence [11]. Furthermore, emergent communication enhances the functionality of the memory and retrieval (e.g., RAG) mechanisms, allowing agents to effectively encode and recall shared experiences. This facilitates seamless collaboration and supports advanced decentralized operations critical for scalable and efficient multi-agent coordination [201].

D. Knowledge Representation & Reasoning

Effective knowledge representation and reasoning capabilities are foundational to Agentic AI, enabling intelligent agents to anticipate future states, reason about their environment, and continually adapt through learning [202]. In the context of edge general intelligence, these cognitive processes must be implemented efficiently and robustly within highly resource-constrained environments. Key enabling techniques include Retrieval-Augmented Generation (RAG), on-device knowledge bases, long-term memory integration, causal and world-model prediction, and continual learning. To quantitatively evaluate the reliability of such reasoning mechanisms, we introduce the hallucination rate as a core metric. Hallucination rate measures the frequency at which an AI model produces incorrect, inconsistent, or logically invalid outputs [203]. This metric becomes particularly critical when deploying compressed or lightweight models on mobile and edge devices, where reduced parameter counts or simplified architectures may amplify the risk of reasoning errors.

1) Retrieval-Augmented Generation (RAG) and On-device Knowledge Bases: RAG significantly enhances agent capabilities by integrating external knowledge bases during inference, improving reasoning accuracy and factual consistency [204].

Recent advancements in compact vector databases and efficient retrieval algorithms [205] enable on-device storage and rapid retrieval of relevant information with minimal computational overhead. For example, lightweight retrieval systems allow edge-deployed language models to dynamically access and utilize up-to-date external data locally without constant external connectivity.

2) Long-term Memory and Continual Learning: Agentic AI necessitates mechanisms for retaining and reasoning over extended temporal contexts, continuously updating internal knowledge representations. Long-term memory architectures such as memory-augmented neural networks [71] and transformer models with extended memory modules [206] efficiently store and retrieve historical knowledge. Additionally, lightweight continual-learning frameworks [207], [208] allow edge agents to incrementally assimilate new information without catastrophic forgetting, significantly enhancing adaptability and operational autonomy.

3) Causal and World-Model Prediction: Causal reasoning and world-model prediction capabilities enable Agentic AI systems to understand environmental dynamics, anticipate outcomes, and proactively perform look-ahead planning entirely on edge devices. Techniques such as latent dynamics modeling [209], causal reinforcement learning [210], and predictive simulation frameworks [211] offer computationally efficient models of environmental interactions. World models, in particular, enable agents to internally simulate future states, evaluate potential actions, and select optimal strategies without expensive real-world trial-and-error interactions. This capability significantly enhances sample efficiency, safety, and planning effectiveness.

E. Lessons Learned and Performance Metrics

The deployment of Agentic AI on resource-constrained edge devices requires integrated solutions across multiple technical fronts. Compact model techniques, such as low-rank adaptation, quantization, pruning, and distillation, enable efficient execution under strict memory and compute budgets [205]. Energy-aware architectures and adaptive inference, combined with hardware-level optimizations, ensure sustainable operation within power and thermal limits [172]. Robust communication and decentralized coordination strategies allow agents to collaborate effectively in dynamic environments [212]. Moreover, advanced knowledge representation methods such as retrieval-augmented generation, long-term memory, and causal reasoning, support adaptive decision-making and future state prediction [204]. Together, these capabilities form the foundation for scalable, autonomous, and intelligent Agentic AI at the edge.

From an evaluation perspective, these design choices should be grounded in a coherent set of performance metrics tailored to edge deployments. Compactness can be captured by generalization accuracy across heterogeneous tasks, users, and device conditions after applying pruning, quantization, or distillation [112]. Efficiency is characterized by inference latency, energy consumption per decision, model size, peak memory usage, and communication volume per update or coordination round. Knowledge and reasoning quality can be assessed through task success rate and hallucination rate, for example the fraction of actions or outputs that violate physical constraints or application logic [170]. Migration and collaboration capabilities are reflected in adaptation time when moving to new environments, the success rate of task handover across devices, and the scalability and fairness of multi-agent coordination [189]. Explicitly reporting such metrics in future Agentic AI studies will facilitate reproducible benchmarking and transparent comparison of different designs for edge general intelligence.

IV. Open Source Agentic AI Projects

Agentic AI has rapidly proliferated within the open-source ecosystem, driven by the increasing availability of modular frameworks, community-maintained toolchains, and reusable agent components that facilitate practical implementation. These open-source resources play a crucial role in bridging the gap between conceptual designs and real-world deployments, allowing researchers and practitioners to prototype perception–reasoning–action loops, tool-use workflows, and multi-agent coordination mechanisms with minimal overhead. Similar surveys (e.g., [214], [215], [216]) also include discussions on open-source frameworks to strengthen the bridge between conceptual taxonomy and practical implementation. Furthermore, they provide transparency, reproducibility, and extensibility—qualities essential for accelerating research progress in edge intelligence. To illustrate the breadth of current development, Table IX summarizes representative open-source projects categorized into three groups: *Agent Frameworks and Platforms*, *Autonomous AI Agent Applications*, and *Domain-specific AI Agents*.

A. Agent Frameworks and Platforms

Agent frameworks and platforms facilitate the deployment and management of autonomous, intelligent agents capable of reasoning, decision-making, and collaboration. They provide foundational tools enabling both developers and non-technical users to effectively harness Agentic AI for various practical scenarios, significantly lowering the barrier to entry for advanced agent-based applications [217].

1) MetaGPT: MetaGPT is a multi-agent collaborative framework utilizing natural language programming and task automation to facilitate efficient task execution among multiple autonomous agents. This framework is introduced and detailed in the paper by Hong et al. [218]. Specifically, the authors proposed a sophisticated role-based architecture, empowering agents with distinct responsibilities such as autonomous code generation, peer code review, iterative refinement, and coordinated execution. This design significantly enhanced agents' collective problem-solving capabilities, effectively demonstrating the practical utility of Agentic AI in automating complex software engineering processes and minimizing human intervention.

2) Langflow: Langflow is a low-code platform specifically designed for developing multi-modal and retrieval-augmented generation (RAG)-based multi-agent systems. This framework

TABLE VII
ENERGY-AWARE COMPUTING TECHNIQUES FOR EDGE GENERAL INTELLIGENCE

Technique	Ref	Description	Pros	Cons
Energy-Efficient Model Design	[172], [173], [174], [175]	Designs compact models (e.g., MobileNet, ShuffleNet, EfficientNet, MobileViT) optimized for energy efficiency at the edge	• Reduced computational cost • Minimal energy footprint • High visual reasoning capability	• Potential accuracy-performance trade-off • Limited capacity in complex tasks
Adaptive Inference Mechanisms	[176], [177], [178], [179]	Dynamically adjusts computational resources based on input complexity, using dynamic neural networks and multi-exit architectures	• Significant latency (24.6%) and energy (46.5%) reductions • Adaptive computation without accuracy loss	• Increased design complexity • Possible miscalibration at inference
Hardware-Aware Optimization	[180], [181], [182], [183]	Co-designs models and hardware via hardware-aware NAS, targeted pruning, and DVFS with DRL-based invocation scheduling [214]	• Reduced energy (up to 23.3%) and invocation cost (55.1%) • Optimized alignment with edge accelerators (NPUs, TPUs, VPUs)	• Hardware-specific optimization overhead • Limited portability across hardware platforms

TABLE VIII
KNOWLEDGE REPRESENTATION AND REASONING TECHNIQUES FOR EDGE GENERAL INTELLIGENCE

Technique	Ref	Description	Pros	Cons
RAG and On-device Knowledge Bases	[205]	Integrates compact vector databases and efficient retrieval algorithms for on-device dynamic knowledge access	• High factual consistency • Robust offline autonomy • Low computational overhead	• Memory capacity constraints • Complex indexing optimization
Long-term Memory and Continual Learning	[70], [206], [207], [208]	Utilizes memory-augmented neural architectures and lightweight continual learning for incremental knowledge updates	• Extended temporal reasoning • Avoids catastrophic forgetting • Continuous adaptation	• Memory management complexity • Performance degradation risks
Causal and World-Model Prediction	[209], [211], [212]	Implements causal reinforcement learning, latent dynamics modeling, and predictive simulation for proactive edge-based planning	• Enhanced decision safety • Reduced trial-and-error cost • Improved planning efficiency	• High model complexity • Sensitivity to inaccuracies

is introduced and detailed in the paper by Jeong et al. [219]. Specifically, the authors proposed visual and intuitive workflows, empowering agents with autonomous capabilities to process complex multi-modal inputs (including text and images), dynamically orchestrate their interactions, and execute tasks without extensive coding. This design significantly enhanced autonomous agent collaboration, effectively demonstrating Langflow's practical utility in simplifying the adoption of sophisticated Agentic AI within enterprise environments.

3) SuperAGI: SuperAGI is an intuitive and highly practical framework for rapidly deploying and managing autonomous AI agents. Although lacking direct academic publication [220], this platform distinctly emphasized real-world applicability, empowering agents with features such as rapid instantiation, comprehensive lifecycle management, and seamless scalability. This design significantly enhanced the agents' capability to autonomously execute, coordinate, and manage complex tasks, effectively demonstrating the practical realization of autonomous decision-making and efficient task orchestration across diverse operational scenarios.

4) AutoGen: AutoGen is an innovative framework for developing complex applications through multi-agent conversations. This framework is introduced and detailed in the paper by Wu et al. [221]. Specifically, the authors proposed a highly adaptable conversational architecture, empowering agents with diverse tools, including human interactions, LLM-driven decision-making, and external service invocations. This design significantly enhanced agents' autonomous collaboration capabilities, effectively demonstrating AutoGen's strength in handling intricate workflows across various application domains, such as mathematics, coding, question-answering, and operational research.

5) AgentGPT: AgentBench is a comprehensive benchmark designed to rigorously assess the capabilities of LLMs functioning as autonomous agents across multiple interactive environments. This benchmark is introduced and detailed in the paper by Liu et al. [222]. Specifically, the authors proposed systematic evaluation methods, empowering agents with critical competencies such as autonomous reasoning, dynamic decision-making, iterative task-solving, and interactive tool utilization. Their results significantly highlighted performance disparities between commercial models (e.g., GPT-4) and open-source alternatives, effectively demonstrating the urgency for enhancing agent-oriented fine-tuning, training strategies, and robust open-source models explicitly tailored for autonomous agent applications.

TABLE IX
REPRESENTATIVE OPEN-SOURCE AGENTIC AI PROJECTS FROM GITHUB

Task Domain	Project	Description	Key Feature	Repository Link
Agent Frameworks and Platforms				
Agent Framework	MetaGPT	Modular multi-agent framework for collaborative task execution	Natural language programming, task automation	https://github.com/geekan/MetaGPT
Agent Orchestration	Langflow	Low-code pipeline builder for RAG and multi-agent systems	Visual workflow, intuitive orchestration	https://github.com/logspace-ai/langflow
Agent Management	SuperAGI	Framework for rapid deployment and management of autonomous agents	Agent lifecycle management	https://github.com/TransformerOptimus/SuperAGI
Agent Platform	AutoGen	Platform to build interactive, generative agent applications	Multi-agent communication, dynamic execution	https://github.com/microsoft/autogen
Agent Development	AgentGPT	Simplified GPT-based agent creation and management tool	Easy-to-use agent interface	https://github.com/reworkd/AgentGPT
Autonomous AI Agent Applications				
Software Engineering	OpenHands	Autonomous agent for production-level code generation	Autonomous planning, coding	https://github.com/All-Hands-AI/OpenHands
Collaborative AI	CrewAI	Role-based orchestration for cooperative AI agents	Task decomposition, collaboration	https://github.com/joaomdmoura/crewAI
Decision-making	AutoGPT	GPT-powered autonomous reasoning and self-improvement	Iterative decision-making	https://github.com/Significant-Gravitas/AutoGPT
Autonomous Coding	GPT-Engineer	Agent that autonomously generates complete software solutions	End-to-end automated coding	https://github.com/AntonOsika/gpt-engineer
Research Automation	ResearchGPT	AI agent for autonomous research and summarization	Autonomous information extraction	https://github.com/mukulpatnaik/researchgpt
Domain-specific AI Agents				
Cybersecurity	Real-time Threat Detection	Autonomous cybersecurity agent analyzing network traffic	Real-time network threat analysis	https://github.com/OpenBMB/XAgent
Autonomous Vehicles	Self-driving Delivery	Autonomous driving simulator integrating sensor fusion	Route planning, perception	https://github.com/carla-simulator/carla
Education	Virtual Tutoring	Adaptive personalized tutoring agent	Interactive, adaptive instruction	https://github.com/huangwl18/VoxPoser
Finance	FinGPT	Autonomous AI agent for financial data analysis and predictions	Financial forecasting, investment insights	https://github.com/AI4Finance-Foundation/FinGPT
Healthcare	BiMediX	Autonomous agent aiding medical diagnostics and healthcare research	Medical diagnostics, clinical decision support	https://github.com/mbzuai-oryx/BiMediX

B. Autonomous AI Agent Applications

Autonomous AI agent applications enable agents to independently execute complex tasks through advanced reasoning, dynamic decision-making, and iterative task management. They significantly enhance productivity and effectiveness across specialized domains, showcasing the direct impact of Agentic AI technologies in practical scenarios.

1) OpenHands: OpenHands is an autonomous agent designed for production-level code generation tasks. This framework is introduced and detailed in the paper by Selvaraj et al. [223]. Specifically, the authors proposed an integrated system architecture empowering agents with capabilities such as autonomous planning, systematic coding, iterative refinement, and production-oriented software automation. This design significantly enhanced the agents' ability to autonomously generate high-quality code, effectively demonstrating the practical utility of Agentic AI in software engineering automation.

2) AutoGPT: AutoGPT is an autonomous agent framework utilizing GPT models for iterative reasoning and self-improvement. This framework is introduced and detailed in the paper by Richards et al. [50]. Specifically, the authors proposed a dynamic iterative reasoning loop, empowering agents with capabilities such as autonomous problem-solving, dynamic decision-making, continuous self-assessment, and refinement of strategies. This design significantly enhanced the agents' ability to autonomously handle diverse, complex tasks, effectively demonstrating the strength of Agentic AI in practical, adaptive scenarios.

3) CrewAI: CrewAI is a role-based orchestration framework for cooperative AI agents. This framework is introduced in the open-source project by Moura et al. [224]. Specifically, the authors proposed structured orchestration methods, empowering agents with clearly defined roles such as planners, researchers, executors, and coordinators. This design significantly enhanced collaborative problem-solving and task decomposition capabilities, effectively demonstrating practical utility in managing sophisticated workflows through autonomous agent cooperation.

4) GPT-Engineer: GPT-Engineer is an autonomous coding agent designed to fully automate software solution generation. This framework is introduced in the open-source project by Osika et al. [225]. Specifically, the authors proposed an autonomous pipeline that interprets user-defined requirements, autonomously designs software architectures, generates functional code, and iteratively refines the output.

This design significantly enhanced end-to-end software development automation, effectively demonstrating Agentic AI's capability in delivering rapid, reliable, and autonomous software engineering solutions.

5) ResearchGPT: ResearchGPT is an autonomous agent designed to automate the research process comprehensively. This framework is introduced in the open-source project by Patnaik et al. [226]. Specifically, the authors proposed an autonomous research workflow empowering agents with capabilities such as systematic literature review, structured information extraction, summarization, and insightful synthesis. This design significantly enhances productivity and accuracy in complex research tasks, effectively demonstrating the practical utility of Agentic AI in automating rigorous academic and professional research activities.

C. Domain-Specific AI Agents

Domain-specific AI agents are specialized autonomous systems explicitly designed to handle tasks unique to particular application areas. They leverage specialized domain knowledge and targeted capabilities to significantly enhance performance and practicality within specific operational contexts.

1) XAgent: XAgent is an autonomous cybersecurity agent designed specifically for real-time network threat detection. This framework is introduced and detailed in the open-source project by OpenBMB [227]. Specifically, the authors proposed a robust autonomous monitoring system, empowering agents with capabilities for rapid threat identification, real-time security analysis, and dynamic cybersecurity responses. This design significantly enhanced network security effectiveness, effectively demonstrating the practical utility of Agentic AI in autonomous cybersecurity.

2) CARLA: CARLA is an autonomous driving simulator designed explicitly for self-driving delivery tasks through comprehensive sensor fusion and realistic simulation scenarios. This framework is introduced and detailed in the paper by Dosovitskiy et al. [228]. Specifically, the authors proposed a realistic urban environment simulation, empowering agents with sensorimotor control, adaptive scenario-driven evaluations, and robust navigation capabilities amidst dynamic obstacles, including other vehicles and pedestrians. This design significantly enhanced autonomous driving training, effectively demonstrating CARLA's crucial role in practical self-driving applications.

3) VoxPoser: VoxPoser is an adaptive personalized tutoring agent leveraging composable 3D value maps guided by language models for robotic manipulation tasks. This framework is introduced and detailed in the paper by Huang et al. [229]. Specifically, the authors proposed integrating large language models to autonomously interpret and execute complex natural-language instructions, dynamically generating composable 3D affordance maps. This design significantly enhanced robotic manipulation capabilities, effectively demonstrating the practical application of Agentic AI in personalized and interactive educational environments.

4) FinGPT: FinGPT is an autonomous AI agent explicitly developed for financial data analysis and predictive insights. This framework is introduced and detailed in the paper by Liu et al. [230]. Specifically, the authors proposed democratizing internet-scale financial datasets through generative AI models, empowering agents with capabilities such as autonomous financial forecasting, investment decision support, and real-time financial analytics. This design significantly enhanced financial decision-making processes, effectively demonstrating FinGPT's utility in intelligent financial services and investment management.

5) BiMediX: BiMediX is an autonomous AI agent explicitly designed for bilingual medical diagnostics and clinical decision support. This framework is introduced and detailed in the paper by Pieri et al. [231]. Specifically, the authors proposed a bilingual MoE architecture, empowering agents with advanced medical diagnostic capabilities, clinical record analysis, and robust healthcare recommendations in multiple languages. This design significantly enhanced clinical decision accuracy and healthcare accessibility, effectively demonstrating BiMediX's practical impact on intelligent and inclusive medical services.

V. Case Studies of Agentic AI for Edge General Intelligence

In this section, we present four representative applications of Agentic AI specifically tailored for edge general intelligence, i.e., low-altitude economy networking (LAENet), intent networking, vehicular networks, and human-centric service provisioning.

A. Agentic AI for Low Altitude Economy Networking

1) Background and Motivation: In the context of LAENet, supporting diverse aerial operations demands sophisticated real-time decision-making capabilities to cope with dynamic environments, stringent resource constraints, and heterogeneous network conditions [233], [234], [235], [236], [237]. Although RL has demonstrated significant promise for autonomous and adaptive aerial network control, classical RL methodologies frequently encounter severe limitations such as insufficient generalization to novel scenarios, suboptimal reward design, and unstable policy convergence, particularly in dynamic and uncertain aerial environments [238], [239], [240], [241]. For example, traditional RL methods struggle to adaptively adjust trajectories and energy-efficient operations for UAVs due to their fixed policy structures and simplistic reward designs, thereby hindering the practical applicability in complex real-world tasks [242], [243], [244].

However, Agentic AI empowered by LLMs has emerged as a transformative paradigm, integrating advanced cognitive functions such as contextual understanding, dynamic generalization, and structured reasoning, thereby substantially enhancing autonomous decision-making [74], [233], [245]. Unlike traditional RL methods, Agentic AI leverages pretrained LLMs to extract multi-modal features, enabling contextually adaptive reward shaping and action selection. In particular, COT prompting allows LLMs to effectively capture contextual nuances and reason through complex scenarios, significantly improving generalization across heterogeneous

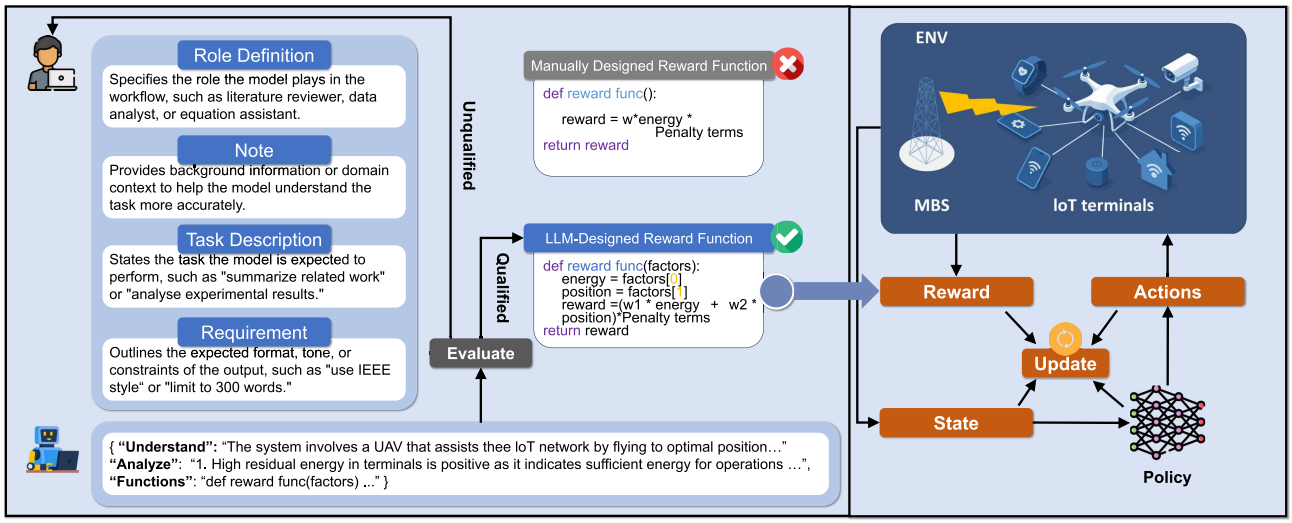

Fig. 6. Architecture of a UAV-assisted IoT network with LLM-designed reward function for reinforcement learning in the LAENet framework. The system integrates compact LLMs for multi-modal perception, structured reasoning, adaptive reward shaping, and decentralized coordination among UAV agents, enabling efficient and scalable data collection in edge environments [232].

aerial tasks [246]. LLMs support task decomposition and plan revision through both forward and backward reasoning, enabling agents to adaptively solve complex problems with interpretable steps [247]. Furthermore, LLM-based reward shaping has demonstrated superior task alignment and stability compared to manually designed reward functions [248]. Moreover, agents can coordinate via shared LLMs by exchanging abstract intents and jointly planning actions, achieving decentralized collaboration without explicit protocols [212]. Integrating these sophisticated cognitive capabilities into the RL loop thus overcomes traditional RL's inherent limitations, providing a robust and scalable solution for adaptive and efficient LAENet deployments.

2) System Description: As depicted in Fig. 6, we consider a UAV-assisted IoT communication network within the framework of LAENet, comprising a single UAV, a macro base station (MBS), and multiple distributed IoT terminals [232], [249]. In this scenario, the UAV maintains a fixed altitude and constant cruising speed, dynamically adjusting its hovering positions near the IoT terminals to optimize data collection and energy delivery. The terminals leverage harvested energy wirelessly provided by the UAV to transmit sensor data, which the UAV subsequently aggregates and forwards to the MBS for further processing. However, maintaining optimal hovering positions to ensure data throughput and reliability presents a critical trade-off: continuous UAV repositioning significantly escalates propulsion and communication energy consumption, complicating the optimization of system energy efficiency.

Under these operational considerations, we formulate an aerial data collection and energy efficiency multi-objective optimization problem aiming to minimize total system energy consumption including terminal transmission energy, UAV propulsion, and communication energy, while satisfying stringent constraints on transmission power limits, data throughput requirements, decoding reliability, and data freshness. This optimization problem inherently features high-dimensional, non-convex, and NP-hard characteristics due to dynamic environmental factors and real-time constraints [130]. Classical optimization techniques typically decompose such problems into separable convex subproblems solved iteratively; however, the effectiveness of these approaches heavily relies on decomposition strategies and faces severe computational overhead in dynamic IoT environments [250]. Agentic AI offers adaptive decision-making capabilities and sophisticated contextual reasoning without explicit decomposition [114], [238]. By embedding LLM-generated adaptive reward signals directly into RL frameworks, Agentic AI effectively navigates complex state-action spaces, achieving robust, scalable, and near-optimal solutions for UAV localization and energy allocation, thus significantly outperforming traditional optimization methods in dynamic LAENet scenarios.

3) Workflow of Agentic AI Framework for LAENet: Generally, the Agentic AI framework substantially enhances the adaptive reasoning and policy optimization capabilities by effectively integrating the contextual comprehension and structured reasoning strengths of LLMs with the sequential decision-making capacity of RL. Specifically, the workflow of Agentic AI for LAENet is structured into four key stages [232], addressing complex decision-making scenarios involving multi-modal inputs and dynamic environmental conditions. Here, we elaborate the workflow through an illustrative UAV-assisted IoT data collection scenario in LAENet to demonstrate the efficacy of the integrated approach.

- *Step 1: State Perception and Abstraction:* The UAV–environment interaction is formulated as a Markov decision process (MDP), where the state includes information such as UAV location, residual energy, and channel conditions. Agentic AI leverages pretrained LLMs to perceive and abstract these heterogeneous inputs. To support edge deployment, lightweight LLM variants (e.g., LoRA-adapted or quantized models) are used to encode multi-modal sensory signals and

user instructions efficiently. This yields compact yet expressive state representations that facilitate robust decision-making.

- *Step 2: Action Selection and Policy Execution:* Based on the perceived state, the LLM guides action generation by dynamically reasoning over possible trajectories. Specifically, chain-of-thought prompting enables the decomposition of high-level objectives into subgoals, improving transparency and adaptability. LLMs further perform causal reasoning to anticipate the outcomes of sequential actions, enabling more informed policy execution. Compact actor-critic networks or distilled policy modules are employed to meet real-time execution constraints under energy and bandwidth limitations.
- *Step 3: Reward Evaluation and Feedback Processing:* During execution, the agent collects both explicit feedback (e.g., sensed delay) and implicit feedback (e.g., human-in-the-loop comments). Agentic AI uses LLMs to interpret such signals and adaptively construct reward functions aligned with mission goals. Compared to manually designed functions, the adaptive reward shaping mechanism better accommodates environmental variability and user preferences. It also enables context-aware trade-offs between data freshness and resource consumption.
- *Step 4:Policy Update and Knowledge Integration:* The collected trajectories and reward feedback are used to iteratively refine the RL policy. The LLMs summarize episodic knowledge and integrates it into a continually evolving policy. In multi-agent scenarios, agents exchange intent summaries and local knowledge via shared LLM-based communication protocols, enabling decentralized coordination for collaborative coverage, scheduling, and resource sharing. To maintain efficiency and scalability, memory-efficient architectures such as RAG-based retrieval modules are adopted for long-term knowledge reuse.

By embedding advanced Agentic AI capabilities, such as multi-modal comprehension, dynamic context adaptation, and structured reasoning, into every stage of the RL decision-making loop, the LLM-enhanced RL framework markedly improves agent intelligence, adaptability, and interpretability. Consequently, this approach provides a robust, scalable, and human-aligned solution for secure, autonomous, and adaptive UAV-assisted IoT data collection and operation in complex and dynamic LAENet environments.

4) Numerical Results: Fig. 7 presents the convergence performance comparison of the proposed Agentic AI-enhanced reward design approach with conventional manually designed rewards for DDPG and TD3 algorithms. Notably, algorithms equipped with LLM-generated rewards demonstrate consistently superior performance, achieving substantial reductions in total energy consumption. Specifically, the Agentic TD3 attains up to a 6.4% reduction in final energy consumption compared to its manually designed counterpart. This performance enhancement can be primarily attributed to the richer reward structure generated by the LLM, which incorporates comprehensive UAV positional information alongside energy-related factors. Consequently, this enables the UAV to dynamically optimize its trajectory, effectively reducing flight distances and communication overhead.

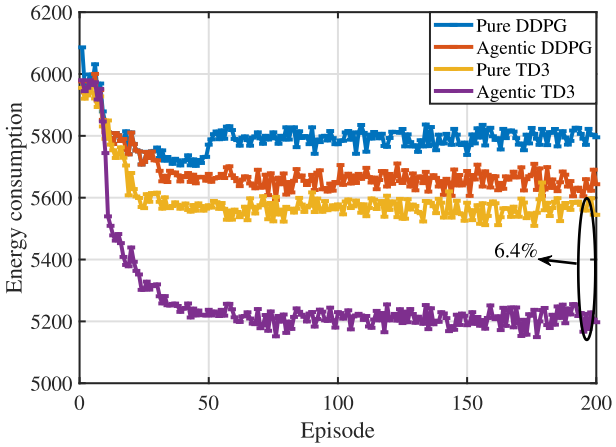

Fig. 7. Energy consumption across episodes for various algorithms using Pure DRL versus Agentic DRL [232].

Additionally, the effectiveness of the Agentic AI-enhanced reward design indicates promising generalization potential for more intricate and diverse optimization tasks, such as multi-objective or cross-domain resource allocation scenarios in LAENet [234], [250]. Conversely, traditional DRL methods constrained by manually crafted rewards exhibit limited performance and flexibility [238], [251], failing to sufficiently adapt to the real-time variability and complexity inherent to LAENet environments.

5) Lessons Learned: Agentic AI-driven RL effectively incorporates high-level cognitive reasoning and contextual comprehension provided by LLMs [74], [233], enabling robust and adaptive decision-making within complex, dynamic environments. The integration of pretrained LLM-generated adaptive reward mechanisms fundamentally transforms traditional reward design approaches, generating nuanced, context-aware reward signals and action selections [248]. This significantly mitigates classical DRL limitations, including suboptimal local convergence and rigid exploration strategies. Consequently, Agentic AI not only resolves critical challenges identified in conventional RL methods, such as inadequate generalization and unstable policy convergence, but also demonstrates substantial potential for addressing more sophisticated and multidimensional optimization tasks, particularly in complex multi-objective or cross-domain LAENet scenarios [250], [252].

B. Agentic AI for Intent Networking

1) Background and Motivation: In next-generation intelligent networking systems, context-aware knowledge retrieval has become essential for enabling timely, relevant, and adaptive decision-making across dynamic and resource-constrained environments [253]. Traditional retrieval-augmented networking architectures often rely on centralized indexing or fixed rule-based matching mechanisms, which limit scalability and responsiveness under rapidly changing network states, such as in vehicular networks, aerial relays, or multi-agent swarms.

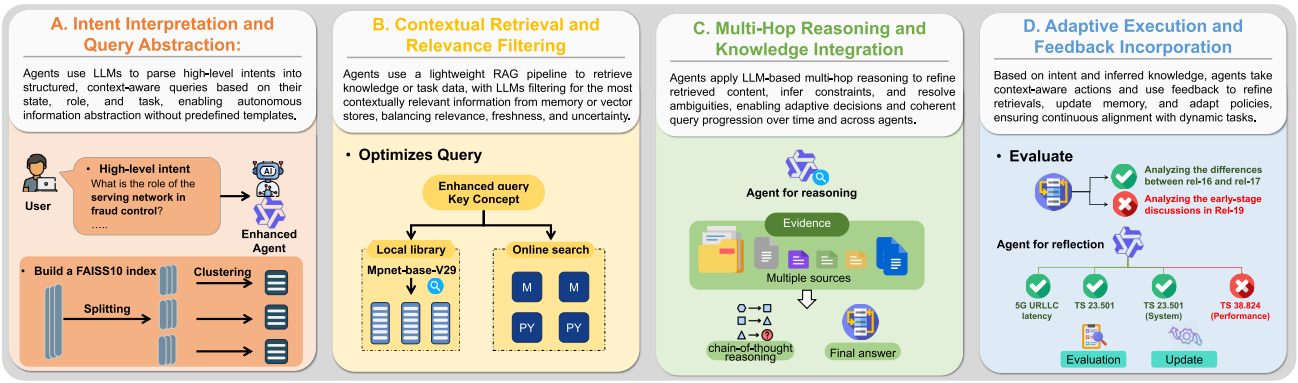

Fig. 8. Illustration of the Agentic contextual retrieval enhanced intelligent base station for troubleshooting and decision-making [56].

These methods typically fail to support online semantic reasoning, multi-hop task tracking, or fine-grained spatial-temporal alignment, thereby impairing the system's ability to deliver high-quality contextual information across diverse and evolving scenarios.

Agentic offers a transformative paradigm by introducing dynamic, in-situ retrieval capabilities that align with the agent's internal decision-making context [74], [233]. Unlike static retrieval frameworks, Agentic AI enables edge agents to autonomously interpret natural language queries, reason over latent task histories, and proactively retrieve or generate semantically relevant content based on mission objectives. For instance, recent advances in RAG allow agents to access distributed knowledge bases and refine results through iterative interactions [205]. Furthermore, chain-of-thought prompting enables structured, multi-step reasoning over retrieved content, allowing agents to infer, filter, and apply contextual cues in real time [246]. By embedding such capabilities into communication-aware systems, Agentic AI fundamentally augments network intelligence, enabling semantic query routing, proactive data fusion, and context-driven protocol adaptation. This integration not only enhances agent collaboration and responsiveness but also paves the way for scalable, memory-efficient, and knowledge-grounded networking infrastructures suited for real-world edge deployments.

2) System Description: As illustrated in Fig. 8, we consider an intent network system empowered by Agentic AI, where distributed edge agents are tasked with interpreting high-level user intents and autonomously translating them into actionable, network-wide behaviors. In such environments, agents must operate under conditions of limited observability, dynamic topologies, and heterogeneous device capabilities [56]. Traditional intent translation pipelines, often rule-based or statically programmed, lack the flexibility to adapt to evolving network states or to reason over ambiguous or under-specified intents, thereby limiting responsiveness and scalability.

To address these limitations, we propose an Agentic AI framework in which each network agent is equipped with a compact LLM to support real-time semantic understanding, contextual reasoning, and adaptive intent interpretation. Agents collaborate via multi-hop communication and utilize contextual prompts to retrieve relevant policy templates, network state information, and domain knowledge from distributed knowledge bases using RAG mechanisms [205]. This allows agents to resolve intents dynamically based on current network conditions and task history, rather than relying on predefined intent-to-policy mappings. The system aims to minimize intent translation latency and maximize execution accuracy while preserving scalability and autonomy. This is achieved by optimizing retrieval granularity, knowledge routing strategies, and response composition through LLM-driven reasoning. Compared to static intent network architectures, the Agentic AI-based system demonstrates superior adaptability, enabling network agents to proactively infer user goals, disambiguate conflicting intents, and generate context-aware action plans without centralized orchestration. This architecture provides a scalable and human-aligned solution for intent realization in next-generation edge-native intelligent networks [74], [233].

3) Workflow of Agentic Contextual Retrieval Framework: Generally, the Agentic Contextual Retrieval (ACR) framework leverages the cognitive and semantic capabilities of LLMs to empower networked agents with proactive, goal-aligned information retrieval and intent grounding. Unlike static or rule-based retrieval systems, ACR enables autonomous agents to dynamically interpret, decompose, and fulfill high-level intents in situ by integrating RAG, distributed memory access, and structured reasoning. Specifically, the ACR workflow is structured into four stages that collectively support scalable and adaptive intent-driven operations across network agents [56].

- *Step 1: Intent Interpretation and Query Abstraction:* Upon receiving a high-level intent (e.g., "ensure full-area coverage within 10 minutes"), the agent formulates structured semantic queries based on its local state, role, and contextual task awareness. This involves LLM-powered parsing of natural language into symbolic or task-grounded representations, enabling agents to autonomously abstract context-specific information needs without pre-defined templates.
- *Step 2: Contextual Retrieval and Relevance Filtering:* The agent issues a retrieval prompt, either locally or across peers, via a lightweight RAG pipeline to access knowledge entries, cached task traces, or environmental

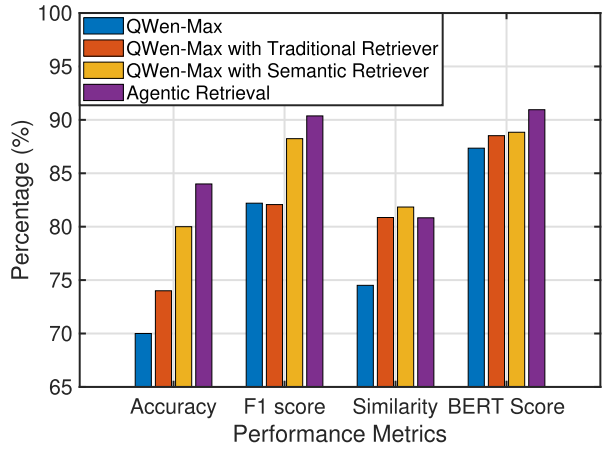

Fig. 9. Comparison of Agentic Retrieval performance with baseline methods, including QWen-Max without a retriever, traditional retrieval, and semantic retrieval [56].

facts. Through embedded attention and filtering mechanisms, the LLM identifies the most contextually relevant entries from distributed memory buffers or vector stores, balancing relevance with freshness and uncertainty.

- *Step 3: Multi-Hop Reasoning and Knowledge Integration:* Retrieved content is iteratively processed using LLM-based multi-hop reasoning to infer higher-order relationships, resolve ambiguity, or refine the query outcome. For example, coverage plans may be adapted by inferring constraints from recent UAV paths, network congestion, or peer statuses. The agent synthesizes this information into actionable decisions or next-hop queries, ensuring continuity of reasoning across time and agents.
- *Step 4: Adaptive Execution and Feedback Incorporation:* Based on the interpreted intent and inferred knowledge, the agent executes appropriate actions (e.g., adjusting trajectory or reassigning coverage roles), and monitors environmental and inter-agent feedback. This feedback is then used to refine future retrievals, update memory indices, and guide policy evolution, thereby enabling continual alignment between evolving intent expressions and dynamic task realities.

By embedding Agentic AI capabilities, such as semantic grounding, distributed memory reasoning, and intent-aware retrieval, into every phase of the information acquisition loop, the ACR framework transforms passive data access into a proactive, interpretive, and self-evolving process. It enables agents to collaboratively fulfill intents in uncertain, bandwidth-limited, and partially observable environments, laying a robust foundation for scalable and context-adaptive networking intelligence.

4) Numerical Results: Fig. 9 presents the performance comparison between the proposed ACR framework and conventional query matching baselines under varying intent complexities. The ACR approach, empowered by LLM-driven semantic interpretation and reasoning, achieves higher intent fulfillment accuracy and faster response convergence across all scenarios. Specifically, under complex multi-agent intents involving conditional constraints and partial observability, ACR improves task success rate by up to 14.8% compared to traditional keyword-based or rule-based retrieval methods. This performance gain is attributed to the ability of Agentic AI agents to interpret natural language intents, reason over distributed memory, and iteratively refine retrieval prompts based on contextual cues. Furthermore, by leveraging RAG mechanisms and lightweight in-situ LLMs, agents adaptively prioritize relevant knowledge entries and suppress redundant query broadcasts, yielding up to 23.4% communication reduction compared to uniform broadcast schemes. This efficiency stems from Agentic AI's capacity to align retrieval actions with both task-specific objectives and environmental context, rather than treating retrieval as an isolated or static subroutine. Additionally, the results demonstrate validate that embedding Agentic AI capabilities into retrieval workflows enables scalable, goal-aligned, and context-aware information access [121]. This lays a foundation for robust and semantically grounded intent resolution in future networked intelligence systems [21], [32].

5) Lessons Learned: The Agentic Contextual Retrieval framework illustrates the significant advantages of embedding Agentic AI capabilities into intent-based networking architectures [74], [233], effectively addresses longstanding challenges in traditional intent networks, including rigid intent-to-policy mappings, static retrieval logic, and limited adaptability to evolving task contexts. The integration of RAG mechanisms and in-situ LLM inference enables agents to align retrieval strategies with user goals and environmental dynamics, improving both responsiveness and interpretability [254]. These capabilities not only enhance intent fulfillment accuracy and communication efficiency but also lay the groundwork for scalable, distributed intelligence in real-world, partially observable environments [169]. Ultimately, Agentic AI offers a transformative approach to enabling self-adaptive, goal-driven collaboration across network agents, establishing a practical and extensible foundation for the next generation of intent-aware edge-native networking systems [255].

C. Agentic AI for Vehicular Edge Computing

1) Background and Motivation: Mobile Edge Computing (MEC) has emerged as a key enabler of low-latency, high-throughput services in dynamic vehicular environments. However, traditional MEC frameworks often rely on static scheduling policies or centralized decision logic, which struggle to scale under high mobility, variable wireless links, and user heterogeneity [256], [257], [258]. As vehicular networks evolve toward ultra-dense deployments and semantically rich applications (e.g., autonomous driving, cooperative perception), MEC must go beyond computation offloading and align system resources with user intent and context [259], [260].

By embedding autonomous, intent-aware agents within edge nodes (i.e., vehicles), Agentic AI enables the system to gain the capacity to parse natural-language objectives, perceive semantic environment signals, and dynamically coordinate offloading or scheduling decisions. This study presents a representative Agentic AI framework for edge computing in vehicular systems, where vehicles act as embodied agents integrating semantic inference (via

LLAVA https://github.com/haotian-liu/LLaVA) and adaptive decision-making (via GAE-PPO [239]), aligned with perceived user intent through the Weber-Fechner-inspired QoE model [239].

2) System Description: As illustrated in Fig. 10, we consider a cellular-based vehicular edge computing system in which I vehicles operate as embodied agents equipped with onboard AI processors and cameras. The network supports V2I and V2V communications over W subbands and includes a base station responsible for coarse-grained spectrum coordination [239]. Each vehicle captures environmental images and uses the LLAVA model to extract semantic information (e.g., object descriptions, parking availability). The information is encoded and transmitted to infrastructure or peers using a semantic communication stack. The offloading and scheduling decisions are modeled as a joint optimization problem aiming to maximize a Weber-Fechner-based QoE metric subject to SINR, symbol length, power, and semantic similarity constraints. These constraints embed both network-level resource feasibility and user-level perceptual utility, forming a context-rich decision space where Agentic AI agents operate.

3) Workflow of Agentic AI for MEC: The agentic AI framework for mobile edge task scheduling and transmission control contains the following steps.

- *Step 1: Intent Interpretation and Semantic Abstraction:* Upon observing raw visual inputs from the surrounding environment, each vehicle utilizes LLAVA to extract structured semantic representations that encapsulate objects, spatial layouts, and driving context. These semantic outputs are aligned with implicit user intents and serve as the basis for intent-grounded policy generation.
- *Step 2: Policy Retrieval and Decision Generation:* The semantic intent vector is mapped to a latent task profile, which is either matched against previously successful policies stored in distributed memory or processed through an online GAE-PPO decision module. This yields a set of adaptive action parameters, including transmission power level, selected communication channel, and semantic symbol length, all optimized under current environmental and network constraints.
- *Step 3: Constrained Execution and QoE-Aware Evaluation:* Based on the selected policy, semantic messages are encoded and transmitted through V2V or V2I links. The receiving node reconstructs the message and evaluates its semantic fidelity via cosine similarity between BERT-based embeddings of the original and decoded text. This quality signal forms the basis for assessing the user-perceived effectiveness of the transmission.
- *Step 4: Feedback Integration and Policy Refinement:* The agent computes a reward signal that integrates semantic accuracy and transmission cost using a Weber-Fechner-inspired QoE function [262]. This reward is used to update the policy network via GAE-PPO, enabling continual improvement of intent-grounded behavior over time. Additionally, performance traces are stored for future retrieval, closing the learning loop.

By embedding Agentic AI capabilities, such as semantic abstraction, context-driven policy generation, and reward-aligned adaptation, into each stage of the edge decision-making loop, the proposed framework transforms traditional scheduling into a cognitively enriched, intent-responsive process. It empowers mobile agents to reason over multi-modal observations, align actions with human-perceived utility, and continuously refine behavior in real-time. This design lays the foundation for scalable, human-aligned, and semantically adaptive mobile edge intelligence.

4) Numerical Results: Fig. 11 shows the convergence behavior of the Agentic AI-enabled method in comparison with several baseline algorithms, including pure PPO, DDPG, and a random policy. It achieves consistently higher returns per episode and exhibits significantly faster convergence and specifically outperforms pure PPO by a margin of approximately 61% in accumulated return, highlighting its superior sample efficiency and stability. Collectively, these results validate that the Agentic AI framework by embedding GAE into the actor-critic learning loop, achieves more reliable and sample-efficient policy optimization, rendering it well-suited for adaptive decision-making in mobile edge vehicular networks.

5) Lessons Learned: The Agentic AI framework for MEC demonstrates the practical benefits of embedding LLM-driven semantic reasoning and reinforcement-based policy optimization into dynamic vehicular environments. It effectively overcomes critical limitations of conventional MEC systems, including static resource scheduling, task-agnostic transmission, and lack of real-time adaptability to user-level goals. By integrating LLAVA-based semantic abstraction with GAE-PPO-enhanced policy evolution, the framework enables autonomous agents to align communication and computation strategies with perceived task intent and environmental conditions [239], [263]. These capabilities not only improve decision stability and semantic transmission efficiency, but also promote self-adaptive and perceptually grounded coordination among mobile edge nodes [264]. Ultimately, this case study demonstrates that Agentic AI provides a scalable and context-aware paradigm for intent-aligned task scheduling and resource optimization in future edge-native intelligent systems [265], [266], [267].

D. Agentic AI for Human-Centric Service Provisioning

1) Background and Motivation: EGI aims to serve human users with personalized and context-aware services across diverse application domains. However, traditional edge intelligence systems predominantly focus on generic optimization objectives, such as minimizing latency or maximizing throughput, often neglecting subjective preferences and contextual requirements that define the human-centric service experience [268], [269], [270]. The fundamental challenge in human-centric service provisioning lies in the difficulty of translating subjective human preferences into actionable optimization strategies for Service Function Chain (SFC) composition [253], [271], [272]. Existing approaches rely on predefined QoE metrics that fail to capture the nuanced, context-dependent nature of human perception and satisfaction. For instance, different users may prioritize different

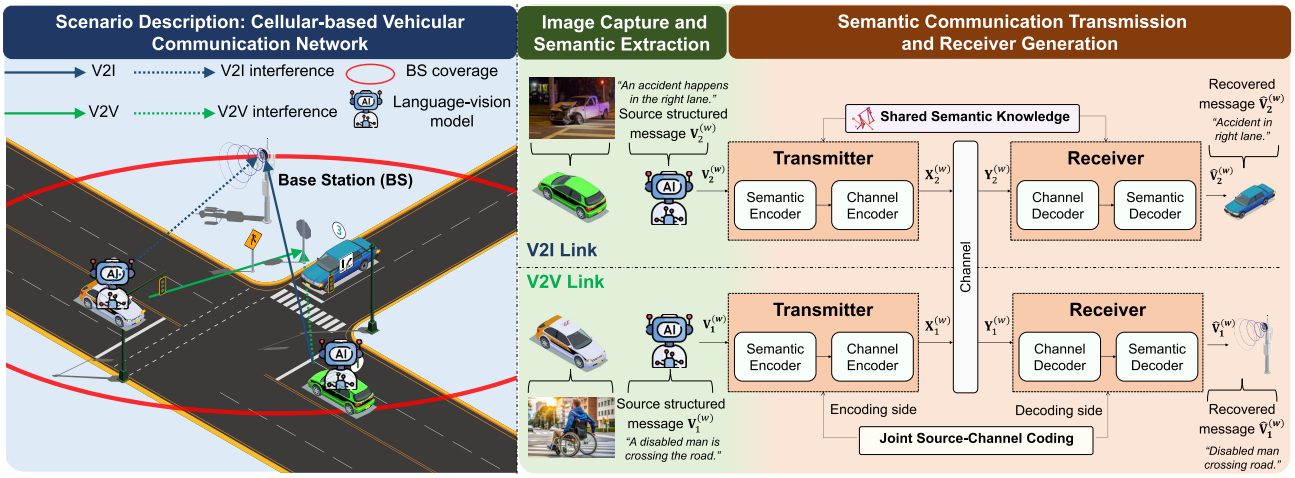

Fig. 10. System model illustrates a cellular-based vehicular communication network, where embodied AI vehicles utilize semantic communication to encode and decode structured messages for efficient and reliable data exchange [261].

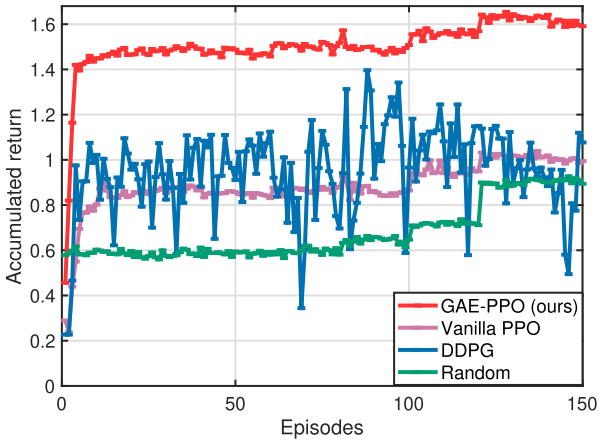

Fig. 11. Convergence behavior with different methods [261].

aspects of service quality: some users may emphasize capability and generation quality for content creation tasks, while others may prioritize low latency and reliability for real-time applications [273], [274], [275].

Agentic AI can transform edge systems by enabling them to autonomously understand natural-language human preferences, dynamically optimize SFC compositions, and adapt proactively through continuous learning [253], [276]. Unlike conventional edge intelligence that operates with static optimization targets, Agentic AI leverages LLMs to interpret diverse expressions of user satisfaction, translates these into structured preference vectors, and employs a DRL-based planning module to optimize the SFC composition dynamically [236], [277]. This cognitive approach transforms edge general intelligence from resource-centric optimization to truly human-centric service provisioning, maximizing subjective QoE.

2) System Description: As shown in Fig. 12, we present an agentic AI framework to perform human-centric service provisioning. The edge general intelligence environment comprises multiple distributed edge servers, each providing specific services (e.g., content generation, data analysis, and multimedia processing). Moreover, we employ a Centralized Large AI Model (C-LAM) at the cloud infrastructure and multiple lightweight Edge Large AI Models (E-LAMs) distributed across edge servers. The C-LAM serves as a central coordinator that maintains comprehensive user preference databases, while edge servers host multiple E-LAMs with varying model sizes and preference understanding capabilities to serve users with different computational budgets and latency requirements. Users submit service requests that require multi-step SFC composition, where each step involves selecting an appropriate service provider from various candidates.

The proposed Agentic AI framework integrates three core technological components: Human Preference Modeling (HPM), Decision Making (DM), and Feedback Adaptation (FA). The HPM module captures and quantifies subjective human preferences through advanced knowledge distillation techniques, where the C-LAM continuously monitors user interactions and transfers preference understanding capabilities to lightweight E-LAMs via weighted pairwise distillation processes [253]. Each E-LAM learns to interpret natural language expressions of user satisfaction and contextual cues, translating them into structured preference vectors that capture relative priorities for different service quality dimensions. The DM module integrates preference-guided reasoning with DRL to optimize SFC composition and resource allocation, formulating SFC compositions as an optimization problem. The FA module enables continuous system improvement by collecting multi-modal feedback from users and incorporating this information to refine both preference understanding and decision making, ensuring that the Agentic AI framework adapts to evolving user requirements and contextual conditions.

3) Workflow of Agentic AI for Human-Centric Service Provisioning: The proposed framework operates through a three-stage workflow that seamlessly integrates human preference understanding with adaptive SFC optimization.

- *Step 1: Human Preference Interpretation:* Users submit service requests through natural language prompts that contain explicit task descriptions, implicit quality expectations, and contextual information such as urgency

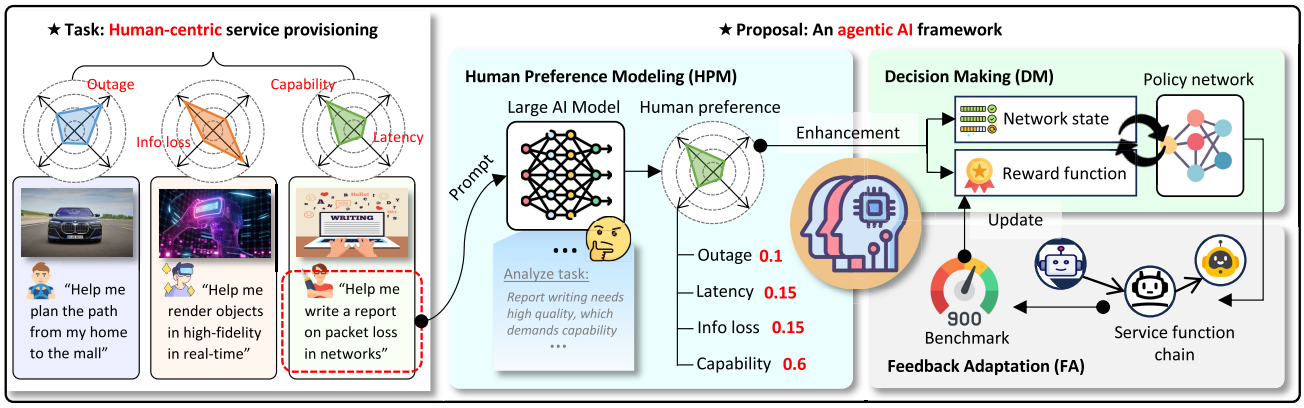

Fig. 12. The illustration of an agentic AI framework for human-centric service provisioning in Edge general intelligence [253].

levels, resource constraints, and task-usage scenarios. Then, an E-LAM processes this multi-modal input along with the current environmental context to generate a personalized preference vector $\mathbf{s} = [\omega_C, \omega_B, \omega_L, \omega_P]$ that quantitatively represents the user's relative priorities for service capability, information fidelity, response latency, and system reliability. This preference interpretation leverages chain-of-thought reasoning to decompose complex user requirements into quantifiable dimensions.

- *Step 2: Preference-Guided SFC Composition:* The DM module is based on a DRL architecture that formulates service provisioning as a Markov Decision Process. Particularly, the current network state, incorporating network conditions, resource availability, and agent capabilities, is augmented with preference-weighted features to form the enhanced state representation. The preference vector $\mathbf{s}$ simultaneously modulates the reward function design, ensuring that the learning process is guided toward maximizing user-perceived quality rather than generic system metrics. This preference-aware DRL enables the policy network to generate optimal SFC compositions that cater to specific users.
- *Step 3: Feedback Integration and System Adaptation:* The selected SFC is executed on the distributed edge infrastructure, while the system continuously monitors both objective performance metrics and subjective satisfaction indicators derived from user behaviors. An in-context learning mechanism is integrated into FA, where the E-LAM maintains a structured context memory containing historical records of user preferences, generated SFCs, and resulting satisfaction outcomes. This contextual memory enables FA to detect preference patterns, adapt to evolving user requirements, and implement automatic calibration mechanisms when preference misalignment is detected. The continuous feedback loop ensures that both the preference understanding capabilities and the DRL policies are refined through accumulated user interactions, achieving symbiotic enhancement between cognitive reasoning and adaptive optimization.

This integrated workflow enables the transformation of traditional resource-centric edge optimization into a truly human-centric, adaptive service provisioning system that continuously evolves to better serve individual user needs while maintaining system efficiency and scalability.

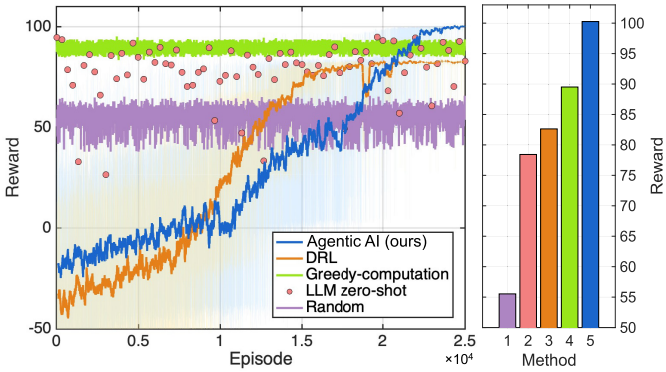

Fig. 13. The performance and learning curves of different methods in human-centric service provisioning [253].

4) Numerical Results: Fig. 13 shows the performance comparison of the proposed Agentic AI framework against conventional optimization baselines for human-centric service provisioning. We evaluate the framework on a representative application scenario involving personalized content generation services, where users submit requests for generating technical reports, creative content, and data analysis outputs with diverse quality expectations and contextual requirements.

The experimental results demonstrate that the Agentic AI framework achieves consistently superior performance across all evaluation metrics. Specifically, our preference-aware approach attains up to 27.3% improvement in human-centric QoE compared to traditional DRL methods that assume uniform user preferences. The superior performance can be attributed to the synergistic integration of E-LAM preference interpretation with preference-guided DRL optimization. Unlike conventional approaches that optimize for generic system metrics, our framework dynamically adapts optimization objectives according to individual user preference vectors, enabling more precise alignment between system behavior and human expectations.

5) Lessons Learned: This case demonstrates a paradigm shift from traditional resource-centric optimization toward truly cognitive and human-centric edge general intelligence.

The revolutionary advantage of Agentic AI lies in the natural language understanding and contextual reasoning capabilities introduced by LLMs, along with the decision-making and feedback modules constructed around LLMs, enabling edge systems to autonomously interpret subjective human preferences, dynamically adapt optimization objectives according to individual user contexts, and continuously evolve through accumulated user interactions [152], [253], [278], [279]. This work establishes a foundation for future research directions, including multi-stakeholder preference reconciliation, privacy-preserving personalization, and long-term preference evolution modeling.

Moreover, Agentic AI can flexibly adapt to diverse scenarios by leveraging its modular perception–reasoning–action loop. In real-time video analytics, lightweight vision-language models enable on-device frame interpretation and anomaly detection without relying on cloud infrastructure [280]. For autonomous driving, compact reasoning modules facilitate rapid path planning and safety-critical decisions under strict latency constraints [112]. In remote sensing, retrieval-augmented agents can efficiently process large-scale spatial data while remaining operable on resource-limited platforms such as satellites [281]. These examples highlight that Agentic AI not only supports our proposed use case but also extends effectively to a wide range of time-sensitive, safety-critical, and data-intensive applications.

VI. Future Research Directions

Emerging research directions focus on the synergistic advancement of Agentic AI and edge general intelligence to effectively address the complex demands and inherent constraints of next-generation edge networks [42], [46], [282]. Key future avenues emphasize the integration of cognitive autonomy, resource efficiency, robust decision-making, and seamless adaptability in real-world operational contexts. These directions include:

- **Adaptive and Efficient Collective Intelligence:** Investigating scalable frameworks for decentralized agent collaboration to enhance cognitive autonomy within resource-constrained edge general intelligence deployments. Research should develop efficient decentralized consensus methods, adaptive task allocation strategies, and robust emergent communication mechanisms, enabling AI agents to autonomously collaborate and adapt via agentification process in heterogeneous edge environments [283], [284]. Moreover, future systems must dynamically adjust collaboration granularity and communication frequency based on network congestion, agent density, and environmental volatility.
- **Privacy-Preserving Federated Agent Systems:** Developing federated learning methodologies tailored explicitly for Agentic AI and edge general intelligence scenarios, emphasizing scalable, privacy-preserving model training and deployment. Research should advance secure aggregation protocols, adaptive federated architectures, and decentralized knowledge sharing techniques, facilitating collective agent intelligence while maintaining stringent data privacy requirements [285], [286], [287]. Federated optimization should further accommodate heterogeneous agent capabilities and unreliable communication links common in edge environments.
- **Robustness and Safety in Autonomous Reasoning:** Designing robust frameworks to ensure reliable and transparent decision-making capabilities for Agentic AI systems within dynamic edge general intelligence contexts. Future studies should explore real-time hallucination detection methods, autonomous validation of reasoning outputs, causal interpretability techniques, and fail-safe operational mechanisms, enabling trustworthy performance in critical edge applications such as autonomous vehicles and smart manufacturing [288], [289], [290]. Furthermore, formal verification and self-diagnostic modules should be integrated to monitor reasoning integrity in mission-critical deployments.
- **Cross-Domain Adaptation and Migration:** Developing effective methods for Agentic AI systems to seamlessly generalize knowledge and adapt across diverse operational scenarios typical of edge general intelligence environments. Research should focus on robust cross-domain transfer techniques, efficient knowledge migration strategies, and adaptive learning frameworks, allowing agents to autonomously adjust to varying contexts without significant retraining overhead [291], [292], [293]. Memory-based transfer mechanisms, self-supervised domain alignment, and continual learning under resource-aware constraints are promising directions.
- **Compression-Aware agentification Reasoning:** Investigating compression-aware architectures specifically designed to integrate explicit reasoning capabilities into resource-constrained edge systems. Future research should focus on the co-design of model compression techniques, such as low-rank adaptation, structured pruning, quantization, and knowledge distillation, with advanced agentification reasoning mechanisms, ensuring cognitive expressiveness, real-time responsiveness, and energy efficiency at the edge [154], [294], [295]. In particular, hierarchical modular designs and dynamic sparsification could enable reasoning-aware compression with minimal performance degradation.
- **Edge–Cloud Collaborative Agentic Intelligence:** Advancing hybrid architectures that distribute the perception–reasoning–action pipeline across edge and cloud to address the practical limitations of running full agentic stacks on resource-constrained devices. Future research should explore latency-aware partitioning of LLM reasoning modules, dynamic offloading strategies for complex planning tasks, and consistency mechanisms that synchronize on-device agents with cloud-level global knowledge [296], [297]. Such edge–cloud co-designed frameworks are essential for supporting scalable, safe, and energy-efficient deployment of Agentic AI in real-world, safety-critical environments.
- **Regulation-Aware and Ethically Aligned Agentic Intelligence:** Developing Agentic AI systems that operate

under explicit regulatory, safety, and ethical constraints is becoming increasingly essential for real-world deployment. Future research should integrate policy-aware reasoning mechanisms, hierarchical human-in-the-loop oversight, and controllable autonomy settings that ensure agents comply with operational safety standards across sectors such as transportation, healthcare, and public infrastructure [298], [299]. To support legally compliant operation at the edge, Agentic AI frameworks must incorporate transparent decision-logging, verifiable action pathways, and auditable reasoning modules that satisfy accountability and certification requirements. Moreover, ethical safeguards—including bias mitigation, fairness guarantees, and fail-safe intervention protocols—must be co-designed with the perception–reasoning–action pipeline to ensure trustworthy and responsible behavior in safety-critical edge environments.

VII. Conclusion

This paper has provided a comprehensive survey of Agentic AI and agentification process tailored explicitly for edge general intelligence. It has systematically introduced foundational concepts and clearly distinguished Agentic AI from traditional edge intelligence paradigms. Key enabling technologies, including model compression, energy-aware computing, robust connectivity, and knowledge representation and reasoning methods, have been reviewed. Representative Agentic AI applications such as LAENet, intent-driven networking, vehicular networks, and human-centric service provisioning have been illustrated through detailed case studies and experimental analyses. Additionally, this survey has discussed critical deployment challenges, examined emerging open-source frameworks, and identified promising directions for future research.

References

[1] E. Peltonen et al., "6G white paper on edge intelligence," 2020, *arXiv:2004.14850*.

[2] Y.-J. Liu et al., "A survey of integrating generative artificial intelligence and 6G mobile services: Architectures, solutions, technologies and outlooks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 3, pp. 1334–1356, Jun. 2025.

[3] W. Wu et al., "Split learning over wireless networks: Parallel design and resource management," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1051–1066, Apr. 2023.

[4] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: Opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 126–132, Oct. 2020.

[5] S. Zhang et al., "Large models for aerial edges: An edge-cloud model evolution and communication paradigm," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 1, pp. 21–35, Jan. 2025.

[6] S. Hu, M. Li, J. Gao, C. Zhou, and X. Shen, "Adaptive device-edge collaboration on DNN inference in AIoT: A digital-twin-assisted approach," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 12893–12908, Apr. 2024.

[7] R. Zhang et al., "Toward democratized generative AI in next-generation mobile edge networks," *IEEE Netw.*, vol. 39, no. 6, pp. 251–260, Nov. 2025.

[8] D. Katare, D. Perino, J. Nurmi, M. Warnier, M. Janssen, and A. Y. Ding, "A survey on approximate edge AI for energy efficient autonomous driving services," *IEEE Commun. Surveys Tut.*, vol. 25, no. 4, pp. 2714–2754, 4th Quart., 2023.

[9] G. Mujtaba, S. Ali Khowaja, and K. Dev, "EdgeAIGuard: Agentic LLMs for minor protection in digital spaces," *IEEE Internet Things J.*, vol. 12, no. 17, pp. 34992–35000, Sep. 2025.

[10] R. Zhang, K. Xiong, Y. Lu, D. W. K. Ng, P. Fan, and K. B. Letaief, "SWIPT-enabled cell-free massive MIMO-NOMA networks: A machine learning-based approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 6701–6718, Jul. 2024.

[11] D. Xu et al., "Edge intelligence: Empowering intelligence to the edge of network," *Proc. IEEE*, vol. 109, no. 11, pp. 1778–1837, Nov. 2021.

[12] X. Zhu et al., "Multi-modal knowledge graph construction and application: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 2, pp. 715–735, Feb. 2024.

[13] G. Han, J. Jiang, C. Zhang, T. Q. Duong, M. Guizani, and G. K. Karagiannidis, "A survey on mobile anchor node assisted localization in wireless sensor networks," *IEEE Commun. Surveys Tut.*, vol. 18, no. 3, pp. 2220–2243, 3rd Quart., 2016.

[14] Z. Cao, P. Zhou, R. Li, S. Huang, and D. Wu, "Multiagent deep reinforcement learning for joint multichannel access and task offloading of mobile-edge computing in industry 4.0," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6201–6213, Jul. 2020.

[15] G. Faraci, S. A. Rizzo, and G. Schembra, "Green edge intelligence for smart management of a FANET in disaster-recovery scenarios," *IEEE Trans. Veh. Technol.*, vol. 72, no. 3, pp. 3819–3831, Mar. 2023.

[16] H. Guo, J. Li, J. Liu, N. Tian, and N. Kato, "A survey on space-air-ground-sea integrated network security in 6G," *IEEE Commun. Surveys Tut.*, vol. 24, no. 1, pp. 53–87, 1st Quart., 2022.

[17] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[18] L. He, L. Fan, X. Lei, P. Fan, A. Nallanathan, and G. K. Karagiannidis, "The road toward general edge intelligence: Standing on the shoulders of foundation models," *IEEE Commun. Mag.*, vol. 63, no. 9, pp. 164–170, Sep. 2025.

[19] D. Van Huynh, S. R. Khosravirad, A. Masaracchia, O. A. Dobre, and T. Q. Duong, "Edge intelligence-based ultra-reliable and low-latency communications for digital twin-enabled metaverse," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1733–1737, Aug. 2022.

[20] L. Chen, L. Fan, X. Lei, T. Q. Duong, A. Nallanathan, and G. K. Karagiannidis, "Relay-assisted federated edge learning: Performance analysis and system optimization," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3387–3401, Jun. 2023.

[21] Y. Shen et al., "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Commun. Mag.*, vol. 62, no. 10, pp. 140–146, Oct. 2024.

[22] Y. Guo et al., "Secrecy energy efficiency maximization in IRS-assisted VLC MISO networks with RSMA: A DS-PPO approach," *IEEE Trans. Wireless Commun.*, vol. 24, no. 8, pp. 6475–6489, Aug. 2025.

[23] Y. Hu et al., "Leveraging LLMs in cloud–edge networks for traffic risk prediction and accident severity analysis," *IEEE Trans. Netw. Sci. Eng.*, vol. 13, pp. 438–453, 2025.

[24] S. Zhang, "Model collaboration framework design for space-air-ground integrated networks," *Comput. Netw.*, vol. 257, Feb. 2025, Art. no. 111013.

[25] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surveys Tut.*, vol. 24, no. 1, pp. 1–30, 1st Quart., 2021.

[26] S. Zhu, K. Ota, and M. Dong, "Green AI for IIoT: Energy efficient intelligent edge computing for industrial Internet of Things," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 1, pp. 79–88, Mar. 2022.

[27] W. Yu et al., "A survey on the edge computing for the Internet of Things," *IEEE Access*, vol. 6, pp. 6900–6919, 2018.

[28] T. K. Rodrigues, K. Suto, H. Nishiyama, J. Liu, and N. Kato, "Machine learning meets computation and communication control in evolving edge and cloud: Challenges and future perspective," *IEEE Commun. Surveys Tut.*, vol. 22, no. 1, pp. 38–67, 1st Quart., 2020.

[29] B. Liu et al., "Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems," 2025, *arXiv:2504.01990*.

[30] N. Kato et al., "Optimizing space-air-ground integrated networks by artificial intelligence," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 140–147, Aug. 2019.

[31] M. Li, J. Gao, C. Zhou, X. S. Shen, and W. Zhuang, "Slicing-based artificial intelligence service provisioning on the network edge: Balancing AI service performance and resource consumption of data management," *IEEE Veh. Technol. Mag.*, vol. 16, no. 4, pp. 16–26, Dec. 2021.

[32] Y. He, J. Fang, F. R. Yu, and V. C. Leung, "Large language models (LLMs) inference offloading and resource allocation in cloud-edge computing: An active inference approach," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 11253–11264, Dec. 2024.

[33] Z. Feng et al., "Multi-agent embodied AI: Advances and future directions," 2025, *arXiv:2505.05108*.

[34] R. Zhang et al., "Generative AI-enabled vehicular networks: Fundamentals, framework, and case study," *IEEE Netw.*, vol. 38, no. 4, pp. 259–267, Jul. 2024.

[35] B. Li, T. Liu, W. Wang, C. Zhao, and S. Wang, "Agent-as-a-service: An AI-native edge computing framework for 6G networks," *IEEE Netw.*, vol. 39, no. 2, pp. 44–51, Mar. 2025.

[36] J. Tong et al., "WirelessAgent: Large language model agents for intelligent wireless networks," 2025, *arXiv:2505.01074*.

[37] P. Pico-Valencia and J. A. Holgado-Terriza, "Agentification of the Internet of Things: A systematic literature review," *Int. J. Distrib. Sensor Netw.*, vol. 14, no. 10, Oct. 2018, Art. no. 155014771880594.

[38] J. Wen et al., "Generative AI for low-carbon artificial intelligence of things with large language models," *IEEE Internet Things Mag.*, vol. 8, no. 1, pp. 82–91, Jan. 2025.

[39] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile edge intelligence for large language models: A contemporary survey," *IEEE Commun. Surveys Tut.*, vol. 27, no. 6, pp. 3820–3860, Dec. 2025.

[40] H. Luo et al., "A weighted Byzantine fault tolerance consensus driven trusted multiple large language models network," 2025, *arXiv:2505.05103*.

[41] D. Palossi et al., "Fully onboard AI-powered human-drone pose estimation on ultralow-power autonomous flying nano-UAVs," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1913–1929, Feb. 2022.

[42] H. Luo et al., "Toward edge general intelligence with multiple-large language model (Multi-LLM): Architecture, trust, and orchestration," 2025, *arXiv:2507.00672*.

[43] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "UAVs meet agentic AI: A multidomain survey of autonomous aerial intelligence and agentic UAVs," 2025, *arXiv:2506.08045*.

[44] Y. Sun et al., "Edge large AI model agent-empowered cognitive multimodal semantic communication," *IEEE Trans. Mobile Comput.*, vol. 25, no. 1, pp. 1–18, Jan. 2026.

[45] K. Dev, S. Ali Khowaja, K. Singh, E. Zeydan, and M. Debbah, "Advanced architectures integrated with agentic AI for next-generation wireless networks," 2025, *arXiv:2502.01089*.

[46] F. Jiang, C. Pan, L. Dong, K. Wang, O. A. Dobre, and M. Debbah, "From large AI models to agentic AI: A tutorial on future intelligent communications," 2025, *arXiv:2505.22311*.

[47] Y. Xiao, G. Shi, and P. Zhang, "Towards agentic AI networking in 6G: A generative foundation model-as-agent approach," 2025, *arXiv:2503.15764*.

[48] A. Salama, Z. Nezami, M. M. H. Qazzaz, M. Hafeez, and S. Ali Raza Zaidi, "Edge agentic AI framework for autonomous network optimisation in O-RAN," 2025, *arXiv:2507.21696*.

[49] Y. Lu et al., "Agentic graph neural networks for wireless communications and networking towards edge general intelligence: A survey," 2025, *arXiv:2508.08620*.

[50] Significant Gravitas. *AutoGPT*. [Online]. Available: https://github.com/Significant-Gravitas/AutoGPT

[51] G. Wang et al., "Voyager: An open-ended embodied agent with large language models," 2023, *arXiv:2305.16291*.

[52] L. Bai, J. Cao, M. Zhang, and B. Li, "Collaborative edge intelligence for autonomous vehicles: Opportunities and challenges," *IEEE Netw.*, vol. 39, no. 2, pp. 52–60, Mar. 2025.

[53] A. H. Sodhro, S. Pirbhulal, and V. H. C. de Albuquerque, "Artificial intelligence-driven mechanism for edge computing-based industrial applications," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4235–4243, Jul. 2019.

[54] A. K. Pati, "Agentic AI: A comprehensive survey of technologies, applications, and societal implications," *IEEE Access*, vol. 13, pp. 151824–151837, 2025.

[55] S. Singh Gill et al., "Edge AI: A taxonomy, systematic review and future directions," 2024, *arXiv:2407.04053*.

[56] R. Zhang et al., "Toward agentic AI: Generative information retrieval inspired intelligent communications and networking," *IEEE Commun. Mag.*, early access, doi: 10.1109/MCOM.001.2500073.

[57] M. Z. Aloudat et al., "Metaverse unbound: A survey on synergistic integration between semantic communication, 6G, and edge learning," *IEEE Access*, vol. 13, pp. 58302–58350, 2025.

[58] S. Wang, M. Atif Qureshi, L. Miralles-Pechuan, T. Huynh-The, T. Reddy Gadekallu, and M. Liyanage, "Applications of explainable AI for 6G: Technical aspects, use cases, and research challenges," 2021, *arXiv:2112.04698*.

[59] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[60] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017, *arXiv:1701.06538*.

[61] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[62] M. Ahn et al., "AutoRT: Embodied foundation models for large scale orchestration of robotic agents," 2024, *arXiv:2401.12963*.

[63] Y. Cui, X. Cao, G. Zhu, J. Nie, and J. Xu, "Edge perception: Intelligent wireless sensing at network edge," *IEEE Commun. Mag.*, vol. 63, no. 3, pp. 166–173, Mar. 2025.

[64] X. Wang, Y. Zhao, C. Qiu, Q. Hu, and V. C. M. Leung, "Socialized learning: A survey of the paradigm shift for edge intelligence in networked systems," *IEEE Commun. Surveys Tut.*, vol. 27, no. 3, pp. 2085–2128, Jun. 2025.

[65] P. Lv et al., "Edge computing task offloading for environmental perception of autonomous vehicles in 6G networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 3, pp. 1228–1245, May 2023.

[66] X. Hou et al., "Improving efficiency in multi-modal autonomous embedded systems through adaptive gating," *IEEE Trans. Comput.*, vol. 74, no. 2, pp. 691–704, Feb. 2025.

[67] G. Yan, K. Liu, C. Liu, and J. Zhang, "Edge intelligence for Internet of Vehicles: A survey," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4858–4877, May 2024.

[68] K. Tallam, "From autonomous agents to integrated systems, a new paradigm: Orchestrated distributed intelligence," 2025, *arXiv:2503.13754*.

[69] Y. Xu et al., "Decentralization of generative AI via mixture of experts for wireless networks: A comprehensive survey," 2025, *arXiv:2504.19660*.

[70] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[71] W. Wang et al., "Augmenting language models with long-term memory," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 74530–74543.

[72] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–33.

[73] R. Zhang et al., "Optimizing generative AI networking: A dual perspective with multi-agent systems and mixture of experts," 2024, *arXiv:2405.12472*.

[74] R. Zhang et al., "Generative AI agents with large language model for satellite networks via a mixture of experts transmission," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3581–3596, Dec. 2024.

[75] H. Du et al., "Mixture of experts for network optimization: A large language model-enabled approach," 2024, *arXiv:2402.09756*.

[76] K. Borazjani et al., "Multi-modal multi-task (M3T) federated foundation models for embodied AI: Potentials and challenges for edge integration," 2025, *arXiv:2505.11191*.

[77] Y. Zhao, Z. Yang, X. He, X. Cai, X. Miao, and Q. Ma, "Trine: Cloud-edge-device cooperated real-time video analysis for household applications," *IEEE Trans. Mobile Comput.*, vol. 22, no. 8, pp. 4973–4985, Aug. 2023.

[78] N. Chen, Z. Cheng, X. Fan, X. Xia, and L. Huang, "Towards integrated fine-tuning and inference when generative AI meets edge intelligence," 2024, *arXiv:2401.02668*.

[79] S. Alamouti, F. Arjomandi, M. Burger, and B. Altakrouri, "Building blocks to empower cognitive internet with hybrid edge cloud," 2024, *arXiv:2402.00876*.

[80] H. Chen et al., "Toward edge general intelligence via large language models: Opportunities and challenges," *IEEE Netw.*, vol. 39, no. 5, pp. 263–271, Sep. 2025.

[81] C. Zhao et al., "World models for cognitive agents: Transforming edge intelligence in future networks," 2025, *arXiv:2506.00417*.

[82] D. Xu et al., "Edge intelligence: Architectures, challenges, and applications," 2020, *arXiv:2003.12172*.

[83] J. Liu et al., "A survey on inference optimization techniques for mixture of experts models," 2024, *arXiv:2412.14219*.

[84] Y. Yang et al., "Agentic web: Weaving the next web with AI agents," 2025, *arXiv:2507.21206*.

[85] N. Krishnan, "Advancing multi-agent systems through model context protocol: Architecture, implementation, and applications," 2025, *arXiv:2504.21030*.

[86] H. Su et al., "A survey on autonomy-induced security risks in large model-based agents," 2025, *arXiv:2506.23844*.

[87] A. A. Zaidan and B. B. Zaidan, "A review on intelligent process for smart home applications based on IoT: Coherent taxonomy, motivation, open challenges, and recommendations," *Artif. Intell. Rev.*, vol. 53, no. 1, pp. 141–165, Jan. 2020.

[88] F. Dou et al., "Towards artificial general intelligence (AGI) in the Internet of Things (IoT): Opportunities and challenges," 2023, *arXiv:2309.07438*.

[89] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," *IEEE Commun. Surveys Tut.*, vol. 22, no. 4, pp. 2167–2191, 4th Quart., 2020.

[90] F. Al-Doghman, N. Moustafa, I. Khalil, N. Sohrabi, Z. Tari, and A. Y. Zomaya, "AI-enabled secure microservices in edge computing: Opportunities and challenges," *IEEE Trans. Services Comput.*, vol. 16, no. 2, pp. 1485–1504, Mar. 2023.

[91] T. Bai, H. Zhao, L. Huang, Z. Wang, D. In Kim, and A. Nallanathan, "A decade of video analytics at edge: Training, deployment, orchestration, and platforms," *IEEE Commun. Surveys Tut.*, vol. 28, pp. 2127–2162, 2026.

[92] D. Van Huynh et al., "URLLC edge networks with joint optimal user association, task offloading and resource allocation: A digital twin approach," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7669–7682, Nov. 2022.

[93] M. Al-Quraan et al., "Edge-native intelligence for 6G communications driven by federated learning: A survey of trends and challenges," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 3, pp. 957–979, Jun. 2023.

[94] Q. Duan, J. Huang, S. Hu, R. Deng, Z. Lu, and S. Yu, "Combining federated learning and edge computing toward ubiquitous intelligence in 6G network: Challenges, recent advances, and future directions," *IEEE Commun. Surveys Tut.*, vol. 25, no. 4, pp. 2892–2950, 4th Quart., 2023.

[95] Z. Liu, S. Zhang, Q. Liu, H. Zhang, and L. Song, "WiFi-diffusion: Achieving fine-grained WiFi radio map estimation with ultra-low sampling rate by diffusion models," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 11, pp. 3796–3812, Nov. 2025.

[96] T. Meuser et al., "Revisiting edge AI: Opportunities and challenges," *IEEE Internet Comput.*, vol. 28, no. 4, pp. 49–59, Jul./Aug. 2024.

[97] M. M. Azari et al., "Evolution of non-terrestrial networks from 5G to 6G: A survey," *IEEE Commun. Surveys Tut.*, vol. 24, no. 4, pp. 2633–2672, 4th Quart., 2022.

[98] A. A. Abdellatif, A. Mohamed, C. F. Chiasserini, M. Tlili, and A. Erbad, "Edge computing for smart health: Context-aware approaches, opportunities, and challenges," *IEEE Netw.*, vol. 33, no. 3, pp. 196–203, May 2019.

[99] Z. Liu, Q. Z. Sheng, X. Xu, D. Chu, and W. E. Zhang, "Context-aware and adaptive QoS prediction for mobile edge computing services," *IEEE Trans. Services Comput.*, vol. 15, no. 1, pp. 400–413, Jan. 2022.

[100] A. Y. Ding et al., "Roadmap for edge AI: A dagstuhl perspective," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 52, no. 1, pp. 28–33, Mar. 2022, doi: 10.1145/3523230.3523235.

[101] A. E. Eshratifar, M. S. Abrishami, and M. Pedram, "JointDNN: An efficient training and inference engine for intelligent mobile cloud computing services," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 565–576, Feb. 2021.

[102] A. Ali-Eldin, B. Wang, and P. Shenoy, "The hidden cost of the edge: A performance comparison of edge and cloud latencies," 2021, *arXiv:2104.14050*.

[103] P. Eugster, "Toward robust control for 6G networks," *IEEE Netw.*, vol. 38, no. 3, pp. 254–260, May 2024.

[104] P. McEnroe, S. Wang, and M. Liyanage, "A survey on the convergence of edge computing and AI for UAVs: Opportunities and challenges," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 15435–15459, Sep. 2022.

[105] M. Adhikari and A. Hazra, "6G-enabled ultra-reliable low-latency communication in edge networks," *IEEE Commun. Standards Mag.*, vol. 6, no. 1, pp. 67–74, Mar. 2022.

[106] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and M.-N. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 100480–100490, 2019.

[107] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.

[108] S. Kumar and P. Ranjan, "Enhanced accuracy model for edge computing and IoT leveraging artificial intelligence," in *Proc. IEEE 13th Int. Conf. Commun. Syst. Netw. Technol. (CSNT)*, Apr. 2024, pp. 477–483.

[109] P. Buddhaghosh Bansod, "Distinguishing autonomous AI agents from collaborative agentic systems: A comprehensive framework for understanding modern intelligent architectures," 2025, *arXiv:2506.01438*.

[110] Y. Wang et al., "Internet of agents: Fundamentals, applications, and challenges," 2025, *arXiv:2505.07176*.

[111] Y. Peng, A. Jolfaei, Q. Hua, W.-L. Shang, and K. Yu, "Real-time transmission optimization for edge computing in industrial cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9292–9301, Dec. 2022.

[112] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey," *IEEE Access*, vol. 13, pp. 18912–18936, 2025.

[113] E. Miehling et al., "Agentic AI needs a systems theory," 2025, *arXiv:2503.00237*.

[114] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges," 2025, *arXiv:2505.10468*.

[115] G. Molinari and F. Ciravegna, "Towards pervasive distributed agentic generative AI—A state of the art," 2025, *arXiv:2506.13324*.

[116] S. Murugesan, "The rise of agentic AI: Implications, concerns, and the path forward," *IEEE Intell. Syst.*, vol. 40, no. 2, pp. 8–14, Mar. 2025.

[117] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen, "Multi-agent collaboration mechanisms: A survey of LLMs," 2025, *arXiv:2501.06322*.

[118] C. Kai, H. Zhou, Y. Yi, and W. Huang, "Collaborative cloud-edge-end task offloading in mobile-edge computing networks with limited communication capability," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 2, pp. 624–634, Jun. 2021.

[119] X. Yu, S. Zhang, H. Zhang, and L. Song, "Model collaboration at network edge: Feature-large models for real-time IoT communications," *IEEE Internet Things J.*, vol. 12, no. 10, pp. 13259–13272, May 2025.

[120] S. Li et al., "Collaborative inference and learning between edge SLMs and cloud LLMs: A survey of algorithms, execution, and open challenges," 2025, *arXiv:2507.16731*.

[121] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," 2022, *arXiv:2210.03629*.

[122] T. Schick et al., "Toolformer: Language models can teach themselves to use tools," 2023, *arXiv:2302.04761*.

[123] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face," 2023, *arXiv:2303.17580*.

[124] J. Wang and Z. Duan, "Empirical research on utilizing LLM-based agents for automated bug fixing via LangGraph," 2025, *arXiv:2502.18465*.

[125] P. Omidi, X. Huang, A. Laborieux, B. Nikpour, T. Shi, and A. Eshaghi, "Memory-augmented transformers: A systematic review from neuroscience principles to enhanced model architectures," 2025, *arXiv:2508.10824*.

[126] X. Chen et al., "Toward 6G native-AI network: Foundation model-based cloud-edge-end collaboration framework," *IEEE Commun. Mag.*, vol. 63, no. 8, pp. 23–30, Aug. 2025.

[127] S. Ali Khowaja, K. Dev, M. Salman Pathan, E. Zeydan, and M. Debbah, "Integration of agentic AI with 6G networks for mission-critical applications: Use-case and challenges," 2025, *arXiv:2502.13476*.

[128] A. Gharib, W. Ejaz, and M. Ibnkahla, "Enhanced multiband multiuser cooperative spectrum sensing for distributed CRNs," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 1, pp. 256–270, Mar. 2020.

[129] R. Zhang, K. Xiong, Y. Lu, P. Fan, D. W. K. Ng, and K. B. Letaief, "Energy efficiency maximization in RIS-assisted SWIPT networks with RSMA: A PPO-based approach," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1413–1430, May 2023.

[130] S. A. Ullah et al., "Convergence of MEC and DRL in non-terrestrial wireless networks: Key innovations, challenges, and future pathways," *IEEE Commun. Surveys Tut.*, vol. 28, pp. 1950–1985, 2026.

[131] J. Li et al., "LLM-guided DRL for multi-tier LEO satellite networks with hybrid FSO/RF links," 2025, *arXiv:2505.11978*.

[132] R. Zhang et al., "Generative AI for space-air-ground integrated networks," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 10–20, Dec. 2024, doi: 10.1109/MWC.016.2300547.

[133] Y. Huang, N. Xu, M. Guo, J. Li, and L. Shen, "Distributed and reactive controller synthesis for multi-agent systems under finite horizon temporal logic tasks," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 13485–13500, 2025.

[134] J. Wang, J. Hu, G. Min, A. Y. Zomaya, and N. Georgalas, "Fast adaptive task offloading in edge computing based on meta reinforcement learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 1, pp. 242–253, Jan. 2021.

[135] M. Chen et al., "Evaluating large language models trained on code," 2021, *arXiv:2107.03374*.

[136] T. Schick et al., "Toolformer: Language models can teach themselves to use tools," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 68539–68551.

[137] Y. Li et al., "Perception, reason, think, and plan: A survey on large multimodal reasoning models," 2025, *arXiv:2505.04921*.

[138] S. Salimpour et al., "Towards embodied agentic AI: Review and classification of LLM- and VLM-driven robot autonomy and interaction," 2025, *arXiv:2508.05294*.

[139] Y. Ren, Y. Liu, T. Ji, and X. Xu, "AI agents and agentic AI-navigating a plethora of concepts for future manufacturing," 2025, *arXiv:2507.01376*.

[140] J. Tang et al., "Toward general industrial intelligence: A survey of large models as a service in industrial IoT," *IEEE Commun. Surveys Tut.*, vol. 28, pp. 2054–2086, 2026.

[141] L. Mei et al., "A survey of context engineering for large language models," 2025, *arXiv:2507.13334*.

[142] J. Wu and C. K. L. Or, "Position paper: Towards open complex human-AI agents collaboration systems for problem solving and knowledge management," 2025, *arXiv:2505.00018*.

[143] G. Liu et al., "Wireless agentic AI with retrieval-augmented multimodal semantic perception," *IEEE Commun. Mag.*, early access, doi: 10.1109/MCOM.001.2500293.

[144] X. Wang et al., "Chain-of-thought for large language model-empowered wireless communications," 2025, *arXiv:2505.22320*.

[145] C. Du, W. Gao, M. Lin, Q. Liu, T. Pang, and X. Zhang, "Chain of preference optimization: Improving chain-of-thought reasoning in LLMs," in *Proc. Adv. Neural Inf. Process. Syst. 37*, 2024, pp. 333–356.

[146] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model context protocol (MCP): Landscape, security threats, and future research directions," 2025, *arXiv:2503.23278*.

[147] V. P. Tathavadekar, "Agentic AI and ambient intelligence in sustainable supply chain management: A framework for autonomous sustainability decision-making," *Comput. Interdiscipl. Sci.*, vol. 1, no. 1, pp. 24–31, Sep. 2025.

[148] J. Fang et al., "A comprehensive survey of self-evolving AI agents: A new paradigm bridging foundation models and lifelong agentic systems," 2025, *arXiv:2508.07407*.

[149] T. Gong, L. Zhu, F. R. Yu, and T. Tang, "Edge intelligence in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 8919–8944, Sep. 2023.

[150] K. Li, Z. Zhang, A. Pourkabirian, W. Ni, F. Dressler, and O. B. Akan, "Towards resilient federated learning in CyberEdge networks: Recent advances and future trends," 2025, *arXiv:2504.01240*.

[151] Z. Liu et al., "Integrated sensing and edge AI: Realizing intelligent perception in 6G," *IEEE Commun. Surveys Tut.*, vol. 28, pp. 2725–2770, 2026.

[152] X. Huang, H. Yang, C. Zhou, M. He, X. Shen, and W. Zhuang, "When digital twin meets generative AI: Intelligent closed-loop network management," *IEEE Netw.*, vol. 39, no. 5, pp. 272–279, Sep. 2025.

[153] X. S. Shen, X. Huang, J. Xue, C. Zhou, X. Shi, and W. Zhuang, "Revolutionizing QoE-driven network management with digital agents in 6G," *IEEE Commun. Mag.*, early access, 2025.

[154] X. Wang et al., "Empowering edge intelligence: A comprehensive survey on on-device AI models," *ACM Comput. Surv.*, vol. 57, no. 9, pp. 1–39, Sep. 2025.

[155] K. Acharya, A. Velasquez, and H. H. Song, "A survey on symbolic knowledge distillation of large language models," *IEEE Trans. Artif. Intell.*, vol. 5, no. 12, pp. 5928–5948, Dec. 2024.

[156] J. Sun et al., "A survey of reasoning with foundation models: Concepts, methodologies, and outlook," *ACM Comput. Surv.*, vol. 57, no. 11, pp. 1–43, 2025, doi: 10.1145/3729218.

[157] T. Shahriar, "Comparative analysis of lightweight deep learning models for memory-constrained devices," 2025, *arXiv:2505.03303*.

[158] A. Asare, D. Agyemanh Nana Gookyi, D. Boateng, and F. Aabangbio Wulnye, "Deploying and evaluating multiple deep learning models on edge devices for diabetic retinopathy detection," 2025, *arXiv:2506.14834*.

[159] M. Puerta-Beldarrain, O. Gómez-Carmona, R. Sánchez-Corcuera, D. Casado-Mansilla, D. López-De-Ipiña, and L. Chen, "A multifaceted vision of the human-AI collaboration: A comprehensive review," *IEEE Access*, vol. 13, pp. 29375–29405, 2025.

[160] W. Wu et al., "AI-native network slicing for 6G networks," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 96–103, Feb. 2022.

[161] J. E. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2021, p. 3.

[162] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[163] L. Li, Y. Zhang, and L. Chen, "Prompt distillation for efficient LLM-based recommendation," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2023, pp. 1348–1357.

[164] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2704–2713.

[165] J. Lin, J. Tang, H. Tang, S. Yang, G. Xiao, and S. Han, "AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration," in *Proc. Mach. Learn. Syst.*, vol. 28, 2025, pp. 12–17.

[166] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 129–146.

[167] X. Ma, G. Fang, and X. Wang, "LLM-pruner: On the structural pruning of large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 21702–21720.

[168] T. Huang, T. Luo, M. Yan, J. T. Zhou, and R. Goh, "RCT: Resource constrained training for edge AI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2575–2587, Feb. 2024.

[169] B. Yang et al., "Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 40–47, Apr. 2021.

[170] K. Duran and B. Canberk, "Digital twin-based collaborative management for energy-aware 6G IoT systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2025, pp. 1–6.

[171] S. Laskaridis, K. Katevas, L. Minto, and H. Haddadi, "Mobile and edge evaluation of large language models," in *Proc. Workshop Efficient Syst. Found. Models II@ ICML*, 2024, pp. 1–20.

[172] C. Luo, X. He, J. Zhan, L. Wang, W. Gao, and J. Dai, "Comparison and benchmarking of AI models and frameworks on mobile devices," 2020, *arXiv:2005.05085*.

[173] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[174] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[175] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.

[176] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022.

[177] A. Bakhtiarnia, Q. Zhang, and A. Iosifidis, "Multi-exit vision transformer for dynamic inference," 2021, *arXiv:2106.15183*.

[178] Y. Addad, A. Lechervy, and F. Jurie, "Multi-exit resource-efficient neural architecture for image classification with optimized fusion block," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 1478–1483.

[179] E. Samikwa, A. Di Maio, and T. Braun, "Adaptive early exit of computation for energy-efficient and low-latency machine learning over IoT networks," in *Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2022, pp. 200–206.

[180] H. Benmeziane, K. El Maghraoui, H. Ouarnoughi, S. Niar, M. Wistuba, and N. Wang, "A comprehensive survey on hardware-aware neural architecture search," 2021, *arXiv:2101.09336*.

[181] C. Li et al., "HW-NAS-bench: Hardware-aware neural architecture search benchmark," 2021, *arXiv:2103.10584*.

[182] S. Dave, R. Baghdadi, T. Nowatzki, S. Avancha, A. Shrivastava, and B. Li, "Hardware acceleration of sparse and irregular tensor computations of ML models: A survey and insights," *Proc. IEEE*, vol. 109, no. 10, pp. 1706–1752, Oct. 2021.

[183] J. Li et al., "FiDRL: Flexible invocation-based deep reinforcement learning for DVFS scheduling in embedded systems," *IEEE Trans. Comput.*, vol. 74, no. 1, pp. 71–85, Jan. 2025.

[184] K. Zhang, Y. Zhu, S. Maharjan, and Y. Zhang, "Edge intelligence and blockchain empowered 5G beyond for the industrial Internet of Things," *IEEE Netw.*, vol. 33, no. 5, pp. 12–19, Sep. 2019.

[185] R. Ranjan, S. Gupta, and S. Narayan Singh, "LOKA protocol: A decentralized framework for trustworthy and ethical AI agent ecosystems," 2025, *arXiv:2504.10915*.

[186] G. Wen, X. Yu, W. Yu, and J. Lu, "Coordination and control of complex network systems with switching topologies: A survey," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 10, pp. 6342–6357, Oct. 2021.

[187] Z. Zhou, G. Liu, and Y. Tang, "Multiagent reinforcement learning: Methods, trustworthiness, applications in intelligent vehicles, and challenges," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 12, pp. 8190–8211, Dec. 2024.

[188] W. Jin, H. Du, B. Zhao, X. Tian, B. Shi, and G. Yang, "A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives," 2025, *arXiv:2503.13415*.

[189] N. D. Lane et al., "DeepX: A software accelerator for low-power deep learning inference on mobile devices," in *Proc. 15th ACM/IEEE Int. Conf. Inf. Process. Sens. Netw. (IPSN)*, Apr. 2016, pp. 1–12.

[190] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *Proc. 44th Annu. IEEE Symp. Found. Comput. Sci.*, 2003, pp. 482–491.

[191] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2016, pp. 1273–1282.

[192] J. Dai, K. Niu, C. Dong, and J. Lin, "Improved message passing algorithms for sparse code multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 9986–9999, Nov. 2017.

[193] J. Foerster, Y. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[194] R. Lowe, J. Foerster, Y.-L. Boureau, J. Pineau, and Y. Dauphin, "On the pitfalls of measuring emergent communication," 2019, *arXiv:1903.05168*.

[195] F. Fioretto, E. Pontelli, and W. Yeoh, "Distributed constraint optimization problems and applications: A survey," *J. Artif. Intell. Res.*, vol. 61, pp. 623–698, Mar. 2018.

[196] S. Li, J. K. Gupta, P. Morales, R. Allen, and M. J. Kochenderfer, "Deep implicit coordination graphs for multi-agent reinforcement learning," 2020, *arXiv:2006.11438*.

[197] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," 2020, *arXiv:2006.02419*.

[198] D. Kong et al., "A survey of LLM-driven AI agent communication: Protocols, security risks, and defense countermeasures," 2025, *arXiv:2506.19676*.

[199] R. Zhang, K. Xiong, W. Guo, X. Yang, P. Fan, and K. B. Letaief, "Q-learning-based adaptive power control in wireless RF energy harvesting heterogeneous networks," *IEEE Syst. J.*, vol. 15, no. 2, pp. 1861–1872, Jun. 2021.

[200] Z. Ma, H. Gong, J. Xiong, and X. Wang, "Heterogeneous multiagent task allocation based on graph-based convolutional assignment neural network," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 17281–17299, Jun. 2025.

[201] X. Huang, Y. Tang, J. Li, N. Zhang, and X. Shen, "Toward effective retrieval augmented generative services in 6G networks," *IEEE Netw.*, vol. 38, no. 6, pp. 459–467, Nov. 2024.

[202] K. Duran, H. Shin, T. Q. Duong, and B. Canberk, "GenTwin: Generative AI-powered digital twinning for adaptive management in IoT networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 2, pp. 1053–1063, Apr. 2025.

[203] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," 2023, *arXiv:2303.08896*.

[204] R. Zhang et al., "Interactive AI with retrieval-augmented generation for next generation networking," *IEEE Netw.*, vol. 38, no. 6, pp. 414–424, Nov. 2024.

[205] L. Ma et al., "A comprehensive survey on vector database: Storage and retrieval technique, challenge," 2023, *arXiv:2310.11703*.

[206] J. W. Rae and A. Razavi, "Do transformers need deep long-range memory," 2020, *arXiv:2007.03356*.

[207] F. Paissan et al., "Structured sparse back-propagation for lightweight on-device continual learning on microcontroller units," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 2172–2181.

[208] X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong, "Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3925–3934.

[209] D. Hafner et al., "Learning latent dynamics for planning from pixels," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 2555–2565.

[210] J. Zhang, "Designing optimal dynamic treatment regimes: A causal reinforcement learning approach," in *Proc. Int. Conf. Mach. Learn.*, vol. 1, 2020, pp. 11012–11022.

[211] A. C. Cullen, B. I. P. Rubinstein, S. Kandeepan, B. Flower, and P. H. W. Leong, "Predicting dynamic spectrum allocation: A review covering simulation, modelling, and prediction," *Artif. Intell. Rev.*, vol. 56, no. 10, pp. 10921–10959, Oct. 2023.

[212] X. Qiu et al., "Towards collaborative intelligence: Propagating intentions and reasoning for multi-agent coordination with large language models," 2024, *arXiv:2407.12532*.

[213] H. Du et al., "Reinforcement learning with LLMs interaction for distributed diffusion model services," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 8838–8855, Oct. 2025.

[214] M. Mehedi Hasan, H. Li, E. Fallahzadeh, G. Krishnan Rajbahadur, B. Adams, and A. E. Hassan, "An empirical study of testing practices in open source AI agent frameworks and agentic applications," 2025, *arXiv:2509.19185*.

[215] H. Yao et al., "A survey on agentic multimodal large language models," 2025, *arXiv:2510.10991*.

[216] H. Wang, J. Gong, H. Zhang, J. Xu, and Z. Wang, "AI agentic programming: A survey of techniques, challenges, and opportunities," 2025, *arXiv:2508.11126*.

[217] S. Zhang et al., "Generative AI on SpectrumNet: An open benchmark of multiband 3-D radio maps," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 2, pp. 886–901, Apr. 2025.

[218] S. Hong et al., "MetaGPT: Meta programming for a multi-agent collaborative framework," 2023, *arXiv:2308.00352*.

[219] C. Jeong, "Beyond text: Implementing multimodal large language model-powered multi-agent systems using a no-code platform," 2025, *arXiv:2501.00750*.

[220] SuperAGI. (2023). *SuperAGI*. Accessed: Jun. 6, 2024. [Online]. Available: https://github.com/TransformerOptimus/SuperAGI

[221] Q. Wu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," 2023, *arXiv:2308.08155*.

[222] X. Liu et al., "AgentBench: Evaluating LLMs as agents," 2023, *arXiv:2308.03688*.

[223] P. Selvaraj, N. Gokul, P. Kumar, and M. Khapra, "OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages," 2021, *arXiv:2110.05877*.

[224] J. Moura. (2023). *CrewAI*. [Online]. Available: https://github.com/joaomdmoura/crewAI

[225] A. Osika. (2023). *GPT-Engineer*. [Online]. Available: https://github.com/gpt-engineer-org/gpt-engineer

[226] M. Patnaik. (2023). *ResearchGPT*. [Online]. Available: https://github.com/mukulpatnaik/researchgpt

[227] X. Team. (2023). *An Autonomous LLM Agent for Complex Task Solving*. [Online]. Available: https://github.com/OpenBMB/XAgent

[228] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.

[229] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "VoxPoser: Composable 3D value maps for robotic manipulation with language models," 2023, *arXiv:2307.05973*.

[230] X.-Y. Liu, G. Wang, H. Yang, and D. Zha, "FinGPT: Democratizing internet-scale data for financial large language models," 2023, *arXiv:2307.10485*.

[231] S. Pieri et al., "BiMediX: Bilingual medical mixture of experts LLM," 2024, *arXiv:2402.13253*.

[232] L. Cai et al., "Large language model-enhanced reinforcement learning for low-altitude economy networking," 2025, *arXiv:2505.21045*.

[233] Y. Cao et al., "Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 6, pp. 9737–9757, Jun. 2025.

[234] Q. Wei, R. Li, W. Bai, and Z. Han, "Multi-UAV-enabled energy-efficient data delivery for low-altitude economy: Joint coded caching, user grouping, and UAV deployment," *IEEE Internet Things J.*, vol. 12, no. 14, pp. 27519–27532, Jul. 2025.

[235] C. Zhao et al., "Temporal spectrum cartography in low-altitude economy networks: A generative AI framework with multi-agent learning," 2025, *arXiv:2505.15571*.

[236] X. Gao et al., "Agentic satellite-augmented low-altitude economy and terrestrial networks: A survey on generative approaches," 2025, *arXiv:2507.14633*.

[237] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surveys Tut.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart., 2018.

[238] O. S. Oubbati, M. Atiquzzaman, H. Lim, A. Rachedi, and A. Lakas, "Synchronizing UAV teams for timely data collection and energy transfer by deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6682–6697, Jun. 2022.

[239] R. Zhang et al., "Embodied AI-enhanced vehicular networks: An integrated vision language models and reinforcement learning method," *IEEE Trans. Mobile Comput.*, vol. 24, no. 11, pp. 11494–11510, Nov. 2025.

[240] N. Cheng et al., "A comprehensive simulation platform for space-air-ground integrated network," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 178–185, Feb. 2020.

[241] W. Jiang, B. Ai, C. Shen, M. Li, and X. Shen, "Age-of-information minimization for UAV-based multi-view sensing and communication," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 1100–1114, Jan. 2024.

[242] K. K. Nguyen, T. Q. Duong, T. Do-Duy, H. Claussen, and L. Hanzo, "3D UAV trajectory and data collection optimisation via deep reinforcement learning," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2358–2371, Apr. 2022.

[243] K. K. Nguyen, A. Masaracchia, V. Sharma, H. V. Poor, and T. Q. Duong, "RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1086–1096, Aug. 2022.

[244] A. Masaracchia et al., "UAV-enabled ultra-reliable low-latency communications for 6G: A comprehensive survey," *IEEE Access*, vol. 9, pp. 137338–137352, 2021.

[245] D. Liu, H. Wu, C. Huang, J. Ni, and X. Shen, "Blockchain-based credential management for anonymous authentication in SAGVN," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 3104–3116, Oct. 2022.

[246] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," 2022, *arXiv:2201.11903*.

[247] A. Z. Ren, B. Ichter, and A. Majumdar, "Thinking forward and backward: Effective backward planning with large language models," 2024, *arXiv:2411.01790*.

[248] M. Kwon, S. Michael Xie, K. Bullard, and D. Sadigh, "Reward design with language models," 2023, *arXiv:2303.00001*.

[249] K. Qu, W. Zhuang, Q. Ye, X. Shen, X. Li, and J. Rao, "Dynamic flow migration for embedded services in SDN/NFV-enabled 5G core networks," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2394–2408, Apr. 2020.

[250] Z. Li et al., "Unauthorized UAV countermeasure for low-altitude economy: Joint communications and jamming based on MIMO cellular systems," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 6659–6672, Mar. 2025.

[251] H. Du et al., "Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization," *IEEE Commun. Surveys Tut.*, vol. 26, no. 4, pp. 2611–2646, 2024.

[252] H. Jiang, M. Mukherjee, J. Zhou, and J. Lloret, "Channel modeling and characteristics for 6G wireless communications," *IEEE Netw.*, vol. 35, no. 1, pp. 296–303, Jan. 2021.

[253] Y. Liu et al., "LAMeTA: Intent-aware agentic network optimization via a large AI model-empowered two-stage approach," 2025, *arXiv:2505.12247*.

[254] S. Gupta, R. Ranjan, and S. Narayan Singh, "A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions," 2024, *arXiv:2410.12837*.

[255] M. Li, J. Gao, L. Zhao, and X. Shen, "Deep reinforcement learning for collaborative edge computing in vehicular networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 4, pp. 1122–1135, Dec. 2020.

[256] X. He, H. Xing, Y. Chen, and A. Nallanathan, "Energy-efficient mobile-edge computation offloading for applications with shared data," *IEEE Trans. Wireless Commun.*, pp. 1–6, 2018.

[257] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tut.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[258] Z. M. Fadlullah et al., "State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems," *IEEE Commun. Surveys Tut.*, vol. 19, no. 4, pp. 2432–2455, 4th Quart., 2017.

[259] J. Wang, J. Liu, and N. Kato, "Networking and communications in autonomous driving: A survey," *IEEE Commun. Surveys Tut.*, vol. 21, no. 2, pp. 1243–1274, 2nd Quart., 2019.

[260] Y. Shi, H. D. Tuan, A. V. Savkin, T. Q. Duong, and H. V. Poor, "Model predictive control for smart grids with multiple electric-vehicle charging stations," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2127–2136, Mar. 2019.

[261] S. Tang, Q. Yang, L. Fan, X. Lei, A. Nallanathan, and G. K. Karagiannidis, "Contrastive learning-based semantic communications," *IEEE Trans. Commun.*, vol. 72, no. 10, pp. 6328–6343, Oct. 2024.

[262] D. Sheng, Q. Qi, J. Wang, L. Li, W. Yu, and J. Liao, "PsyQoE: Improving quality-of-experience assessment with psychological effects in video streaming," *IEEE Trans. Services Comput.*, vol. 17, no. 6, pp. 1–14, Nov. 2024.

[263] R. Zhang, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. B. Letaief, "Joint coordinated beamforming and power splitting ratio optimization in MU-MISO SWIPT-enabled HetNets: A multi-agent DDQN-based approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 677–693, Feb. 2022.

[264] A. Khelloufi, H. Ning, S. Dhelim, and J. Ding, "AGI enabled solutions for IoX layers bottlenecks in cyber-physical-social-thinking space," 2025, *arXiv:2506.22487*.

[265] Q. Wang, S. Zou, Y. Sun, M. Liwang, X. Wang, and W. Ni, "Toward intelligent and adaptive task scheduling for 6G: An intent-driven framework," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 5, pp. 1975–1988, Oct. 2024.

[266] H. Jiang, B. Xiong, H. Zhang, and E. Basar, "Hybrid far- and near-field modeling for reconfigurable intelligent surface assisted V2V channels: A sub-array partition based approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8290–8303, Nov. 2023.

[267] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6G: Machine-learning approaches," *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.

[268] Z. Long, H. Dong, and A. E. Saddik, "Human-centric resource allocation for the metaverse with multiaccess edge computing," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 19993–20005, Nov. 2023.

[269] M. Torres Vega et al., "Immersive interconnected virtual and augmented reality: A 5G and IoT perspective," *J. Netw. Syst. Manage.*, vol. 28, no. 4, pp. 796–826, Oct. 2020.

[270] W. Zhuang, Q. Ye, F. Lyu, N. Cheng, and J. Ren, "SDN/NFV-empowered future IoV with enhanced communication, computing, and caching," *Proc. IEEE*, vol. 108, no. 2, pp. 274–291, Feb. 2020.

[271] X. Qin et al., "Generative AI meets wireless networking: An interactive paradigm for intent-driven communications," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 4, pp. 2056–2077, Aug. 2025.

[272] C. Zhou, S. Hu, J. Gao, X. Huang, W. Zhuang, and X. Shen, "User-centric immersive communications in 6G: A data-oriented framework via digital twin," *IEEE Wireless Commun.*, vol. 32, no. 3, pp. 122–129, Jun. 2025.

[273] S. Li, J. Huang, J. Hu, and B. Cheng, "QoE-DEER: A QoE-aware decentralized resource allocation scheme for edge computing," *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 2, pp. 1059–1073, Jun. 2022.

[274] A. Xiao, S. Wu, Y. Ou, N. Chen, C. Jiang, and W. Zhang, "QoE-fairness-aware bandwidth allocation design for MEC-assisted ABR video transmission," *IEEE Trans. Netw. Service Manage.*, vol. 22, no. 1, pp. 499–515, Feb. 2025.

[275] J. Li, W. Shi, H. Wu, S. Zhang, and X. Shen, "Cost-aware dynamic SFC mapping and scheduling in SDN/NFV-enabled space–air–ground-integrated networks for Internet of Vehicles," *IEEE Internet Things J.*, vol. 9, no. 8, pp. 5824–5838, Apr. 2022.

[276] N. Krishnan, "AI agents: Evolution, architecture, and real-world applications," 2025, *arXiv:2503.12687*.

[277] B. Yang et al., "Frontiers of generative AI for network optimization: Theories, limits, and visions," 2025, *arXiv:2507.01773*.

[278] M. Amine Ferrag, N. Tihanyi, and M. Debbah, "From LLM reasoning to autonomous AI agents: A comprehensive review," 2025, *arXiv:2504.19678*.
[279] K. Wang, Y. Tang, T. Q. Duong, S. R. Khosravirad, O. A. Dobre, and G. K. Karagiannidis, "Multi-tier distributed computing systems by leveraging digital twin: Challenges, techniques, and research directions," *IEEE Wireless Commun.*, vol. 32, no. 5, pp. 236–244, Oct. 2025.
[280] Y. Huang et al., "Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 9, no. 3, pp. 1–33, Sep. 2025.
[281] C. Wen et al., "RS-RAG: Bridging remote sensing imagery and comprehensive knowledge with a multi-modal dataset and retrieval-augmented generation model," 2025, *arXiv:2504.04988*.
[282] W. Saad et al., "Artificial general intelligence (AGI)-native wireless systems: A journey beyond 6G," *Proc. IEEE*, 2025.
[283] H. Yang, Y. Liang, J. Yuan, Q. Yao, A. Yu, and J. Zhang, "Distributed blockchain-based trusted multidomain collaboration for mobile edge computing in 5G and beyond," *IEEE Trans. Ind. Informat.*, vol. 16, no. 11, pp. 7094–7104, Nov. 2020.
[284] C. Xu, P. Zhang, H. Yu, and Y. Li, "Dynamic blockchain-empowered trustworthy end-edge collaborative computing via rotating multi-agent DRL," *IEEE Trans. Wireless Commun.*, vol. 24, no. 6, pp. 4864–4878, Jun. 2025.
[285] N. Rodríguez-Barroso, M. García-Márquez, M. Victoria Luzón, and F. Herrera, "Challenges of trustworthy federated learning: What's done, current trends and remaining work," 2025, *arXiv:2507.15796*.
[286] Z. Wang, H. Ji, Y. Zhu, D. Wang, and Z. Han, "A survey on federated analytics: Taxonomy, enabling techniques, applications and open issues," *IEEE Commun. Surveys Tut.*, vol. 28, pp. 2457–2496, 2026.
[287] H. Luo et al., "A trustworthy multi-LLM network: Challenges,solutions, and a use case," 2025, *arXiv:2505.03196*.
[288] S. Atakishiyev, M. Salameh, and R. Goebel, "Safety implications of explainable artificial intelligence in end-to-end autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 10, pp. 14516–14535, Oct. 2025.
[289] M. Monjurul Karim, Y. Shi, S. Zhang, B. Wang, M. Nasri, and Y. Wang, "Large language models and their applications in roadway safety and mobility enhancement: A comprehensive review," 2025, *arXiv:2506.06301*.
[290] H. Luo, J. Luo, and A. V. Vasilakos, "BC4LLM: A perspective of trusted artificial intelligence when blockchain meets large language models," *Neurocomputing*, vol. 599, Sep. 2024, Art. no. 128089.
[291] Y. Li et al., "Unleashing the power of continual learning on non-centralized devices: A survey," *IEEE Commun. Surveys Tut.*, vol. 28, pp. 1059–1098, 2026.
[292] J. Peng, D. Ye, B. Tang, Y. Lei, Y. Liu, and H. Li, "Lifelong learning with cycle memory networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16439–16452, Nov. 2024.
[293] S. A. Serrano, J. Martinez-Carranza, and L. E. Sucar, "Knowledge transfer for cross-domain reinforcement learning: A systematic review," *IEEE Access*, vol. 12, pp. 114552–114572, 2024.
[294] D. Chen et al., "Transforming the hybrid cloud for emerging AI workloads," 2024, *arXiv:2411.13239*.
[295] Y. Zheng, Y. Chen, B. Qian, X. Shi, Y. Shu, and J. Chen, "A review on edge large language models: Design, execution, and applications," *ACM Comput. Surv.*, vol. 57, no. 8, pp. 1–35, Aug. 2025.
[296] S. Alamouti, "Quantifying energy and cost benefits of hybrid edge cloud: Analysis of traditional and agentic workloads," 2025, *arXiv:2501.14823*.
[297] G. K. Sheelam, "Architecting agentic AI for real-time autonomous edge systems in next-gen mobile devices," *Adv. Consum. Res.*, vol. 2, no. 3, pp. 1–16, 2025.
[298] Y. Ning et al., "Generative artificial intelligence and ethical considerations in health care: A scoping review and ethics checklist," *Lancet Digit. Health*, vol. 6, no. 11, pp. e848–e856, Nov. 2024.
[299] M. Al-Kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, "Ethical challenges and solutions of generative AI: An interdisciplinary perspective," *Informatics*, vol. 11, no. 3, p. 58, Aug. 2024.

Ruichen Zhang (Member, IEEE) received the B.Eng. degree from Henan University (HENU), China, in 2018, and the Ph.D. degree from Beijing Jiaotong University (BJTU), China, in 2023. In 2024, he was a Visiting Scholar with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is currently a Post-Doctoral Research Fellow with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. His publications include four ESI highly-cited articles and two ESI hot articles. His research interests include agentic AI, LLM-empowered networking, reinforcement learning-enabled wireless communications, generative AI models, and heterogeneous networks. He has received three best paper awards. He serves as a Guest Editor for several IEEE journals, including IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, *IEEE Wireless Communications Magazine*, and IEEE JOURNAL OF SELECTED AREAS IN SENSORS. He also serves on the editorial boards for IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT. He has been the Managing Editor and the Assistant to the Editor-in-Chief of IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING since 2025.

Guangyuan Liu (Graduate Student Member, IEEE) received the B.Sc. degree from Nanyang Technological University, Singapore, in 2022, where he is currently pursuing the Ph.D. degree with the College of Computing and Data Science, Energy Research Institute @ NTU, under the Interdisciplinary Graduate Program. His research interests include GAI, computer vision, and resource allocation. He won the Honorary Mention Award at the ComSoc Student Competition from the IEEE Communications Society in 2023 and the First Prize at the 2024 ComSoc Social Network Technical Committee (SNTC) Student Competition.

Yinqiu Liu (Member, IEEE) received the B.Eng. degree from Nanjing University of Posts and Telecommunications, China, in 2020, and the M.Sc. degree from the University of California, Los Angeles, USA, in 2022. He is currently pursuing the Ph.D. degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His current research interests include blockchain security, mobile AIGC, and generative AI. He was a recipient of the 2025 IEEE Vehicular Technology Society Student Scholarship Award, the 2025 IEEE WCSP Best Paper Award, the 2025 IEEE IWCMC Best Paper Award, and the 2024 IEEE ComSoc Student Competition Honorary Mention Award.

Changyuan Zhao (Graduate Student Member, IEEE) received the B.Sc. degree in computing and information science from the University of Science and Technology of China, Hefei, China, in 2020, and the M.A.Eng. degree in computer science from the Institute of Software, CAS, Beijing, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include generative AI, communication security, and resource allocation.

Jiacheng Wang (Member, IEEE) received the M.S. and Ph.D. degrees from the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, in 2018 and 2022, respectively. From 2021 to 2022, he was a Visiting Researcher with the College of Computing and Data Science, Nanyang Technological University, Singapore, where he is currently a Post-Doctoral Research Fellow. He has published more than 40 papers, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, *IEEE Wireless Communications Magazine*, *IEEE Network*, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE GLOBECOM, IEEE ICC, and IEEE WCNC. His research interests include generative AI, integrated sensing and communications, network optimization, and edge intelligence. He has received the IEEE ICC 2025 Best Paper Award. He was a Guest Editor of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE WIRELESS COMMUNICATIONS, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, *IEEE Internet of Things Magazine*, and IEEE NETWORKING LETTERS.

Yunting Xu received the bachelor's and Ph.D. degrees from Nanjing University, China, in 2017 and 2024, respectively. He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He mainly focuses on artificial intelligence and networking optimization in the field of emerging wireless networks. He was a TPC member of VTC-Fall 2022, VTC-Fall 2023, and ICCC 2025. He was a recipient of the IEEE WCSP 2022 Best Paper Award.

Dusit Niyato (Fellow, IEEE) received the B.Eng. degree from the King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada. He is currently a Professor with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.

Jiawen Kang (Senior Member, IEEE) received the Ph.D. degree from Guangdong University of Technology, China, in 2018. He was a Post-Doctoral Researcher at Nanyang Technological University, Singapore, from 2018 to 2021. He is currently a Full Professor with Guangdong University of Technology, China. His research interests include blockchain, security, and privacy protection in wireless communications and networking.

Yonghui Li (Fellow, IEEE) received the Ph.D. degree from Beijing University of Aeronautics and Astronautics, Beijing, China, in November 2002. Since 2003, he has been with the Centre of Excellence in Telecommunications, The University of Sydney, Sydney, NSW, Australia. He is currently a Professor and the Director of the Wireless Engineering Laboratory, School of Electrical and Information Engineering, The University of Sydney. His current research interests include wireless communications, with a particular focus on MIMO, millimeter wave communications, machine-to-machine communications, coding techniques, and cooperative communications. He holds several patents granted and pending in these fields. He was a recipient of Australian Queen Elizabeth II Fellowship in 2008 and Australian Future Fellowship in 2012. He received the Best Paper Awards from the 2014 IEEE International Conference on Communications, the 2017 IEEE PIRMC, and the 2014 IEEE Wireless Days Conferences. He was an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He also served as a Guest Editor for several IEEE journals, such as IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, *IEEE Communications Magazine*, IEEE INTERNET OF THINGS JOURNAL, and IEEE ACCESS.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University in 2004. He is currently a Professor and the Earle C. Williams Eminent Scholar and the Director of the Wireless Engineering Research and Education Center, Auburn University. His research interests include wireless networks, multimedia communications, and smart grid. He received the IEEE ComSoc MMTC Outstanding Researcher Award in 2023, the SEC Faculty Achievement Award for Auburn in 2023, the IEEE ComSoc TC-CSR Distinguished Technical Achievement Award in 2019, the Auburn University Creative Research and Scholarship Award in 2018, the NSF CAREER Award in 2010, and multiple IEEE service awards. He was a co-recipient of several best paper and best journal paper awards, including the 2022 IEEE ComSoc eHealth Technical Committee Best Journal Paper Award, the 2021 Elsevier/KeAi Digital Communications and Networks Best Paper Award, the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the 2020 IEEE Vehicular Technology Society Jack Neubauer Memorial Award, the 2018 IEEE ComSoc MMTC Best Journal Paper Award, the 2017 IEEE ComSoc MMTC Best Conference Paper Award, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the field of communications systems. He is also a co-recipient of the best paper/demo awards of 12 conferences. He is the Editor-in-Chief of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, a Member-at-Large of the Board of Governors of the IEEE Communications Society, and the Vice President for Technical Activities of the IEEE Council on Radio Frequency Identification (CRFID).

Sumei Sun (Fellow, IEEE) is currently the Executive Director of the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore. She holds a joint appointment with Singapore Institute of Technology, Singapore, and an adjunct appointment with the National University of Singapore, Singapore, both as a Full Professor. Her research interests include next-generation wireless communications, sensing-communications-computing-control integrative design, applied artificial intelligence, and the Industrial Internet of Things. She is a fellow of the Academy of Engineering Singapore. She was a recipient of the 2023 IEEE VTS Women's Distinguished Career Award and Singapore National Day 2022 Public Administration Medal (Bronze).

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include network resource management, wireless network security, the Internet of Things, 5G and beyond, and vehicular networks. He is a Registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, Chinese Academy of Engineering Foreign Member, and an International Fellow of the Engineering Academy of Japan. He received the "West Lake Friendship Awar" from Zhejiang Province in 2023, the President's Excellence in Research from the University of Waterloo in 2022, Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R.A. Fessenden Award in 2019 from IEEE Canada, the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, the Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society (ComSoc), and the Technical Recognition Award from the Wireless Communications Technical Committee in 2019 and the AHSN Technical Committee in 2013. He has also received the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He serves/served as the General Chair for the 6G Global Conference 2023 and ACM Mobihoc 2015; the Technical Program Committee Chair/Co-Chair for IEEE GlobeCom 2024, 2016, and 2007, IEEE Infocom 2014, and IEEE VTC 2010 Fall; and the Chair for the IEEE ComSoc Technical Committee on Wireless Communications. He is the Past President of the IEEE ComSoc, the Vice President of Technical and Educational Activities, the Vice President of Publications, the Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of the IEEE Fellow Selection Committee of the ComSoc. He served as the Editor-in-Chief for IEEE INTERNET OF THINGS JOURNAL, *IEEE Network*, and *IET Communications*.

Dong In Kim (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He was a Tenured Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is currently a Distinguished Professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He is a fellow of Korean Academy of Science and Technology and a life member of the National Academy of Engineering of Korea. He was the first recipient of the NRF of Korea Engineering Research Center in Wireless Communications for RF Energy Harvesting from 2014 to 2021. He received several research awards, including the 2023 IEEE ComSoc Best Survey Paper Award and the 2022 IEEE Best Land Transportation Paper Award. He was selected as the 2019 recipient of the IEEE ComSoc Joseph LoCicero Award for Exemplary Service to Publications. He has been listed as a Highly Cited Researcher by Clarivate Analytics in 2020, 2022, and 2025. He was the General Chair of the IEEE ICC 2022, Seoul. From 2001 to 2024, he served as an Editor, the Editor-at-Large, and an Area Editor for Wireless Communications—I of IEEE TRANSACTIONS ON COMMUNICATIONS. From 2002 to 2011, he served as an Editor and a Founding Area Editor for Cross-Layer Design and Optimization of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. From 2008 to 2011, he served as the Co-Editor-in-Chief for IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS. He served as the Founding Editor-in-Chief for IEEE WIRELESS COMMUNICATIONS LETTERS from 2012 to 2015.