Interactive AI With Retrieval-Augmented Generation for Next Generation Networking

Ruichen Zhang, Hongyang Du, Yinqiu Liu, Dusit Niyato, Jiawen Kang, Sumei Sun, Xuemin Shen, and H. Vincent Poor

Abstract

With the advance of artificial intelligence (AI), the concept of interactive AI (IAI) has been introduced, which can interactively understand and respond not only to human user input but also to dynamic system and network conditions. In this article, we explore an integration and enhancement of IAI in networking. We first review recent developments and future perspectives of AI and then introduce the technology and components of IAI. We then explore the integration of IAI into next-generation networks, focusing on how implicit and explicit interactions can enhance network functionality, improve user experience, and promote efficient network management. Subsequently, we propose an IAI-enabled network management and optimization framework, which consists of environment, perception, action, and brain units. We also design a pluggable large language model (LLM) module and retrieval augmented generation (RAG) module to build the knowledge base and contextual memory for decision-making in the brain unit. We demonstrate through case studies that our IAI framework can effectively perform optimization problem design. Finally, we discuss potential research directions for IAI-based networks.

Introduction

With the current development trajectory, artificial intelligence (AI) is moving towards the possibility of realizing artificial general intelligence (AGI). The evolution of AI from its original rule-based algorithms to the adoption of advanced learning models marks a significant shift in the field of computing technology. This process is driven by the increasing growth and complexity of data, requiring more sophisticated AI Solutions. While early AI models provided the foundation for modern data-driven environments, their limitations in handling dynamic and large-scale data have prompted the development of more advanced and novel methods. Among these advancements, such as Google Gemini[1] and OpenAI Q*,[2] human-in-the-loop (HITL) systems highlight the importance of integrating human insights into AI decision-making [1]. These systems combine human feedback with the AI's learning process, improving the accuracy and contextual understanding of AI results. For example, a 2018 Stanford study showed that AI combined with HITL outperformed human or AI analysis alone in healthcare.[3] However, HITL systems have limitations in adaptability and real-time responsiveness because they rely on human input, which can be limited, unpredictable, and erroneous.

To address these challenges and effectively develop AGI, interactive AI (IAI) has been proposed. The primary difference between IAI and HITL lies in their interaction modes, i.e., IAI emphasizes immediate and direct interaction between AI and users, whereas HITL focuses on human participation and supervision in the AI decision-making process [1]. IAI systems are capable of instantaneously understanding user inputs, such as voice commands, text messages, or other interactive commands, and intelligently responding or executing tasks based on these inputs. This ability not only enhances user experience but also increases flexibility and effectiveness of AI applications in dynamic environments. Integrating IAI with technologies such as retrieval-augmented generation (RAG) and LangChain can further personalize network operations [2]. For example, RAG enables IAI systems to extract information from vast databases to enrich their responses and decisions. Combined with LangChain, which extends AI reasoning capabilities, IAI can provide more context-aware solutions based on the existing databases. The main advantages of IAI over the limitations of HITL systems, especially when augmented with RAG and LangChain, include:

- **Customizabilty and Personalizability:** IAI, enhanced by RAG and LangChain, offers tailored solutions by aligning AI responses closely with user preferences and needs, resulting in more personalized and user-centric AI applications.
- **Better Flexibility:** IAI's direct user interaction, supported by LangChain's extended reasoning, can offer more flexibility to adapt to different network scenarios and users' requirements, enhancing the systems versatility and satisfaction.
- **Less Bias:** The combination of IAI with RAG and LangChain minimizes the reliance

[1] https://deepmind.google/technologies/gemini/#introduction

[2] https://community.openai.com/t/what-is-q-and-when-we-will-hearmore/521343

[3] https://scopeblog.stanford.edu/2018/05/08/artificial-intelligence-inmedicine-predicting-patient-outcomes-and-beyond/

Digital Object Identifier: 10.1109/MNET.2024.3401159
Date of Current Version: 18 November 2024
Date of Publication: 15 May 2024

Ruichen Zhang, Hongyang Du, Yinqiu Liu, and Dusit Niyato are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798; Jiawen Kang (corresponding author) is with the School of Automation, Guangdong University of Technology, Guangzhou 523083, China; Sumei Sun is with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632; Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada; H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA. Ruichen Zhang, Hongyang Du, and Yinqiu Liu contributed equally to this work.

0890-8044/24©2024IEEE

on human intervention, thus reducing the potential for human bias in AI decisions and leading to more objective outcomes.

It is particularly important to integrate IAI into wireless networks, given its dynamic nature and requirement to continuously adapt the dynamic changes. Fortunately, the capabilities of IAI are particularly promising in addressing these challenges. Its interactive adaptive resource management can optimize the utilization of network resources and improve network performance. For example, in a network with changing user requirements, IAI can dynamically allocate bandwidth to maintain high performance given instantaneous user experience feedback. Although the integration of IAI into networking has several potential advantages, the following issues need to be addressed:

- **Q1:** Why is IAI suitable for networking?
- **Q2:** Which networking challenges can IAI address?
- **Q3:** How can IAI with RAG be applied in networking?

Therefore, in this article, we attempt to provide forward-looking research to answer the above "Why, Which, and How" questions. *To the best of the authors' knowledge, the synergy between IAI and networking is still an open issue*. The contributions of this article are summarized as follows.

- **A1:** We first review some aspects of the development of AI, then introduce the features, technologies, and composition of IAI, and finally overview the IAI on networking.
- **A2:** We explore the integration of IAI in networking, focusing on how both implicit and explicit interactions enhance network functionalities, improve user experience, and facilitate efficient network management.
- **A3:** We construct an IAI-enabled network management and optimization framework. It consists of environment, action, brain, and perception. More importantly, we design pluggable LLM and RAG modules to build the knowledge base and contextual memory for decision-making. Simulation results based on a real network optimization case study verify the effectiveness of the proposed framework.

Organizations: the section "Overview of IAI and Networking" concludes with an overview of IAI and networking. The section "Interactive Applications in Networking" provides interactive applications in the network. The section "Case Study: IAI-Enabled Problem Formulation Framework" provides RAG's IAI framework and demonstrates its effectiveness through a case study. The section "Future Directions" outlines three main future directions for IAI. Finally, the section "Conclusion" summarizes the conclusions.

Overview of IAI and Networking

This section provides an overview of IAI and networking, where some abbreviations are summarized in Table 1.

The Development of AI

Phase 1: Traditional Artificial Intelligence (TAI) TAI consists of rule-based systems designed for specific tasks within well-defined parameters [3]. It leverages algorithms such as support vector machines (SVMs) and other fundamental methods to excel at pattern recognition and basic prediction tasks. Although TAI systems are effective at processing structured data, their limited flexibility highlights the need for more dynamic AI approaches, leading to the development of discriminative AI/predictive AI.

This process is driven by the increasing growth and complexity of data, requiring more sophisticated AI Solutions.

Phase 2: Predictive AI (PAI)/Discriminative AI (DAI) With the use of deep neural networks such as convolutional neural networks (CNNs), PAI and DAI are good at learning special paradigms from large data sets [4]. PAI/DAI, with their enhanced learning algorithms, advances beyond the capabilities of TAI. However, their reliance on extensively annotated datasets, primarily for classification and prediction, has practical limitations, leading to the emergence of GAI.

Phase 3: Generative Artificial Intelligence (GAI) GAI marks a new era where AI systems can create new data and patterns, showcasing a degree of creativity [5]. With the advent of generative diffusion models (GDMs) and generative adversarial networks (GANs), AI's role expands from analysis to creation, which can produce innovative outputs and simulations. However, GAI's reliance on pre-existing data patterns has

Abbreviation	Full name	Description
AGI	Artificial General Intelligence	AI capable of understanding, learning, and applying intelligence across a broad range of tasks at human-level proficiency
GAI	Generative Artificial Intelligence	AI systems that focus on creating novel data and patterns, often used for tasks involving innovation and design
IAI	Interactive Artificial Intelligence	AI that emphasizes immediate, direct interaction with users for decision-making and learning
MOE	Mixture of Experts	An approach in AI using multiple expert models, each specialized in different aspects of a dataset or problem
RAG	Retrieval Augmented Generation	An approach that enhances generative models by retrieving relevant information from database to improve the generation
LLM	Large Language Model	AI models designed to process and interact using human language, capable of handling extensive text data
HITL	Human-in-the-Loop	A system where human participation and oversight are integrated into the AI decision-making process
PAI	Predictive Artificial Intelligence	AI systems that focus on forecasting future events or trends based on historical data and predictive algorithms
TAI	Traditional Artificicial Intelligence	Early AI models based on rule-based algorithms and structured data analysis
GAN	Generative Adversarial Network	AI models where two neural networks compete, often used for generating new data that mimics real data
DRL	Deep Reinforcement Learning	AI models learn to make decisions by trial and error, using deep learning to process complex inputs
CNN	Convolutional Neural Network	AI models effective for analyzing visual imagery, known for pattern recognition in structured data
GDM	Generative Diffusion Model	AI models that generate data through a process of iterative refinement, often used for precise outputs

TABLE 1. Summary of abbreviations.

It is particularly important to integrate IAI into wireless networks, given its dynamic nature and requirement to continuously adapt the dynamic changes.

limitations such as generating inaccurate results, leading to the emergence of IAI.

Phase 4: Interactive Artificial Intelligence (IAI) IAI goes beyond traditional data interaction paradigms by engaging human users in dynamic and reciprocal information exchanges. The key difference between IAI and GAI is IAI's ability to interactively incorporate and learn from human input throughout the modeling process [6]. In contrast, GAI mainly focuses on generating data based on pre-learned distributions without further input. As a result, IAI systems can bridge the gap between AI and human interaction and then move towards AGI.

For clarity, the evolution of AI development is illustrated in Fig. 1.

Overview of IAI

IAI is based on the principle of dynamic information exchange to meet the needs of administrators and human end-users, where the foundation of IAI lies in the following AI technologies:

Retrieval-Augmented Generation (RAG): RAG plays a pivotal role in IAI framework by merging retrieval-based and generative AI techniques. RAG allows IAI systems to access a vast external knowledge base, retrieving relevant information to augment the generation process [2].

Large Language Models (LLMs): LLMs are a cornerstone of IAI, particularly due to their capacity to process and generate human-like text, which facilitates complex interactive dialogues. This ability is central to IAI's focus on enhancing user interaction, as LLMs enable AI systems to comprehend and respond to natural language inputs in a conversational manner, thereby significantly improving the interactive and intuitive nature of AI systems [7].

Multi-modal Interaction: IAI systems are not limited to a single modal; they are adept at understanding and responding to various input types, such as text, voice, and images [8]. This multimodal interaction capability is fundamental to IAI, embodying the principle of versatility and adaptability in interactions. It ensures that AI systems can engage with users in the most natural and intuitive ways possible, adapting to different interaction modes seamlessly.

Mixture of Experts (MoE): The MoE framework is an integral component of IAI that embodies its responsiveness to specialized situation-awareness. MoE models are composed of multiple expert sub-models, each trained to handle different aspects or subsets of the data [7]. This specialization enables IAI systems to leverage the collective expertise of these individual models to solve complex problems. The gating mechanism is the core feature of MoE, which dynamically determines the relevance of each expert to a given input, thereby directing the input to the most appropriate expert.

Deep Reinforcement Learning (DRL): DRL embodies the interactivity and adaptability of IAI.

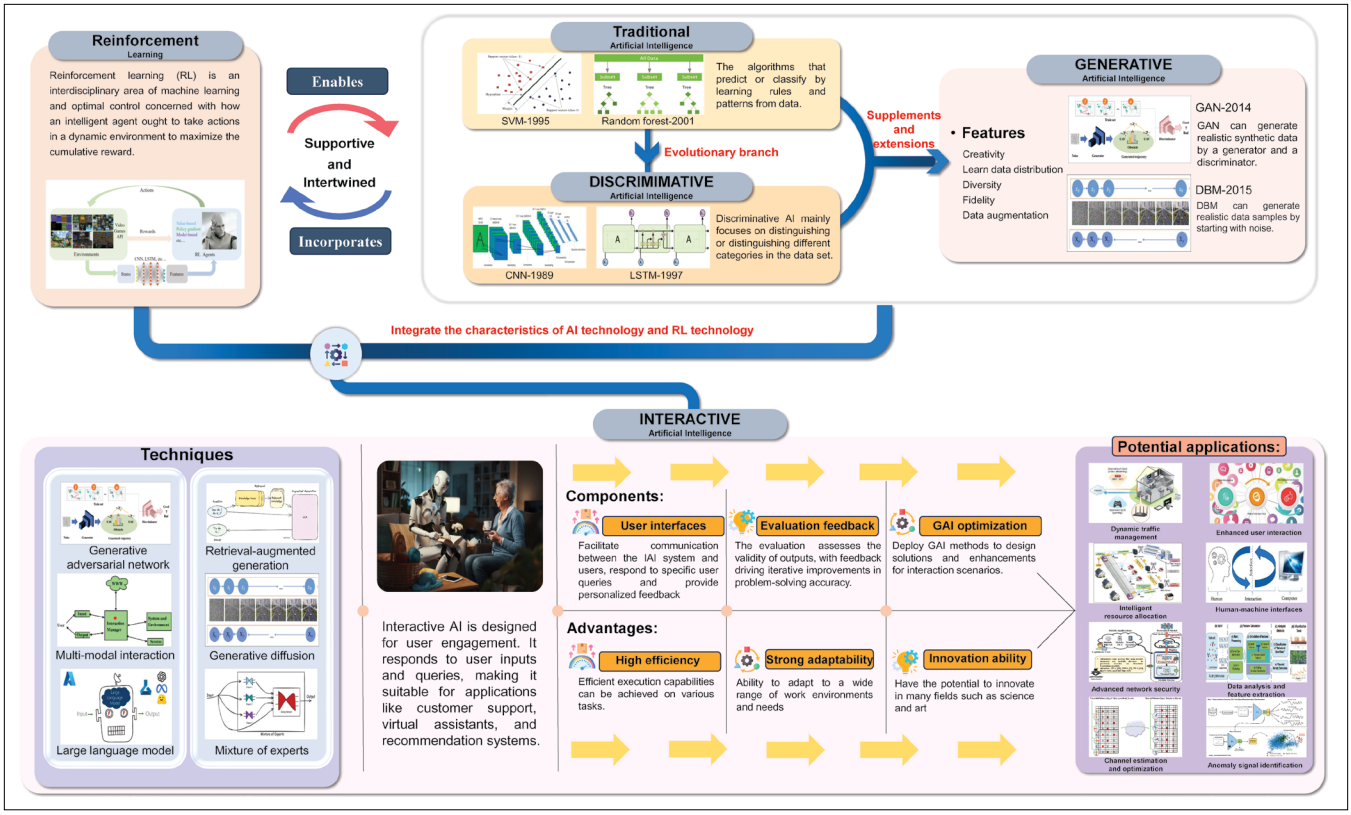

FIGURE 1. The evolution of AI. For IAI, we highlight the role of the mixture of experts, large language models, deep reinforcement learning, retrieval-augmented generation, and generative AI in promoting adaptability and interaction with users. Additionally, the components and advantages of IAI are also presented, emphasizing its efficiency and adaptability in different applications such as optimization and traffic management.

It revolves around learning from interactions with the environment and making decisions based on feedback, such as rewards and penalties [9]. This principle aligns well with the idea that IAI learns and develops through continuous interaction with the user or environment.

Generative Diffusion Models (GDMs): GDMs improve iteratively through interactions, consistent with the principles of IAI. These models generate data by learning to reverse a diffusion process, which essentially involves a series of interacting steps [5]. Each iteration in the process presents an opportunity for the model to adapt and refine its output, emphasizing the IAI principle of continuous evolution and response to new information.

Generative Adversarial Networks (GANs): GANs, with their dueling network structure, inherently align with IAI's principles of interactive learning. The continuous learning process within GANs, where each network learns from the other, enables the system to adaptively refine its outputs [8].

Features of IAI

In the networking domain, IAI systems are characterized by three main components, each of which is an integral part of its functionality as follows:

User Interfaces: Modern IAI systems have revolutionized user interfaces (UIs), extending beyond conventional text and graphical interfaces. These systems incorporate LLMs and multi-modal techniques, allowing users to interact with network management systems through conversational language. A notable example is Agent GPT,[4] which employs LLMs and multi-modal capability for interactive user interfaces in a dynamic plan-execute pattern. Unlike standard UIs, Agent GPT engages proactively with users, providing clarity for ambiguous instructions and continually modifying its approach based on user feedback.

Intelligent Evaluation and Adaptability: Network environments benefit from IAI systems that not only assess performance using real-time analytics and simulation-based testing but also adapt and evolve over time. These systems integrate cognitive network operations to create a feedback loop, continuously refining network configurations based on user inputs and environmental data. This enables AI models to develop their understanding strategies and enhance their decision-making processes. Google's VirusTotal Code Insight[5] is one such system that employs an LLM and MoE for script analysis, thus enabling human users to identify potential threats through responsive feedback during use.

Advanced Generative AI and Network Optimization: The core of IAI systems is the deployment of state-of-the-art GAI. These systems utilize algorithms like GDM and TBM for network design optimization. They facilitate the automated generation of network topologies, simulate traffic for congestion forecasting, and provide routing optimization solutions. For instance, Google's B4 network[6] employs a combination of MoE and GAN techniques for capacity planning and traffic engineering, ensuring optimal data flow and resource utilization. The generative process in these systems is iterative, leveraging both historical and real-time data to proactively adapt to evolving network conditions and user behavior trends.

GDMs improve iteratively through interactions, consistent with the principles of IAI.

IAI for Networking

IAI significantly benefits networking by integrating intelligence across multiple operational layers. It enhances networks from the physical layer to the application layer.

Physical Layer: In the physical layer, IAI can be used to improve network efficiency and reliability. A typical example is the application of channel estimation and optimization. IAI, particularly DRL approaches, leverage feedback-oriented neural network interaction to iteratively optimize beamforming patterns based on real-time environmental feedback, improving modulation techniques for robust communication [9]. This leads to more efficient use of the spectrum, reduced interference, and enhanced signal quality, particularly in dynamic environments where channel conditions frequently change.

Network Layer: In the network layer, IAI can be used to streamline network operations and enhance performance. A typical example is the application of dynamic network traffic management. IAI, particularly DRL and GDM, can dynamically manage and optimize network traffic flows by continuously learning and adapting to traffic conditions [5]. This includes predictive traffic modeling, real-time congestion management, and intelligent routing strategies, resulting in reduced latency, optimized bandwidth usage, and improved overall network resilience.

Application Layer: In the application layer, IAI can be used to enhance user interaction and experience. A typical example is the application of human-machine interfaces. IAI, particularly LLMs, can process and generate natural language for user interaction, while multi-modal methods integrate visual, auditory, and other data forms to enrich the interface [10]. This leads to more personalized and engaging user interactions, better accessibility for diverse user groups, and an overall more responsive and intelligent application environment.

For clarity, the potential issues of IAI in different layers are summarized in Table 2.

Interactive Applications in Networking

In this section, we explore the integration of IAI in networking, focusing on how both implicit and explicit interactions. For clarity, the summary of implicit and explicit interaction is shown in Table 3.

Implicit Interaction in Networking

Implicit interaction in networking refers to the process where AI systems engage and adapt to their environment, including network conditions and user behaviors, without direct external inputs or commands. The benefits of such interactions include increased system autonomy, improved adaptability to changing conditions, and enhanced overall network efficiency [9], [11], [12]. Some major applications of implicit interaction in networking are as follows:

Adaptive Signal Processing: Adaptive Signal Processing in wireless communications involves dynamically adjusting signal processing

[4] https:///agentgpt.reworkd.ai/de

[5] https://blog.virustotal.com/2023/04/introducing-virustotal-codeinsight.html

[6] https://www.b4networks.ca/

Issues \ Tech	Differences		Layers IAI advantages
	Traditional methods	Interact	
Channel Estimation and Optimization	• **Available methods:** Fourier-based techniques, CNNs • **Description:** Relies on static models for signal estimation • **Cons:** Not suited for dynamic environments, high overhead, prone to model mismatch	• **Available methods:** MOE, GANs • **Description:** Utilizes user feedback and algorithms for channel optimization • **Pros:** Resilient to changes, efficient, higher accuracy	***Potential Advantages for Physical layer:*** • **Enhanced Accuracy:** GANs provide refined channel models improving signal clarity and reducing interference. • **Predictive Modeling:** GDMs enable precise channel predictions, aligning real-world conditions. • **Real-Time Adaptation:** DRL optimizes beamforming in real-time for changing network environments. • **Reduced Complexity:** Automated processes through GANs and GDMs simplify the management of physical layer operations.
Anomaly Signal Identification	• **Available methods:** Statistical analysis • **Description:** Detects deviations using predefined criteria • **Cons:** Poor adaptability, high false-positive rates, misses subtle anomalies	• **Available methods:** Muti-modal, LLMs • **Description:** Uses LLMs for real-time anomaly detection from diverse data • **Pros:** Accurate, dynamic detection, reduced false alarms	
Adaptive Beamforming	• **Available methods:** CVX, ZF/MRT/MMSE. • **Description:** Uses mathematical modeling and fixed antenna patterns; lacks change adaptability • **Cons:** Inflexible to fast user movement and fast channel changes	• **Available methods:** RL, GANs • **Description:** Adapts beamforming in real-time environments, GANs for training scenarios. • **Pros:** Adaptable, improves spectral efficiency	
Dynamic Traffic Management	• **Available methods:** Static routing algorithms • **Description:** Uses preset routes; ignores live network status • **Cons:** Unresponsive to real-time conditions, may cause congestion	• **Available methods:** RL, GDMs • **Description:** RL for traffic flow adaptively, GDMs for traffic prediction • **Pros:** Better traffic handling, enhances network performance	***Potential Advantages for Network layer:*** • **Optimized Traffic Flow:** DRL and GDMs enable dynamic management of data traffic, reducing congestion. • **Efficient Utilization:** Intelligent allocation strategies via DRL and MOE maximize network resource usage. • **Enhanced Security:** GANs provide advanced threat simulation for robust network security measures. • **Self-Improving Systems:** DRL algorithms continuously learn and improve network performance.
Intelligent Resource Allocation	• **Available methods:** Static resource partitioning, heuristic algorithms • **Description:** Allocates based on predictions; no real-time adjustments • **Cons:** Inefficient under variable loads, risk of resource misallocation	• **Available methods:** RL, MOE • **Description:** Allocates resources smartly, MOE for traffic pattern analysis • **Pros:** More efficient use of network resources, better service quality	
Advanced Network Security	• **Available methods:** Signature-based detection systems • **Description:** Identifies threats using known signatures • **Cons:** May not identify new or variant forms of attacks quickly, leading to a window of vulnerability	• **Available methods:** GANs, LLMs • **Description:** GANs for attack simulation, LLMs for threat detection • **Pros:** Detects new threats, adapts security protocols	
Enhanced User Interaction	• **Available methods:** Rule-based command interfaces, GUIs • **Description:** Command interfaces need exact syntax; GUIs offer visuals • **Cons:** Steep learning curve for commands; GUIs may lack depth in interaction	• **Available methods:** RL, RAG • **Description:** Personalizes interactions, RAG for natural responses • **Pros:** Customized experiences, higher user interaction	***Potential Advantages for Application layer:*** • **Intuitive Interactions:** LLMs and multi-modal technologies enhance user interface responsiveness. • **Personalized Experiences:** Multi-modal capability allows for services based on user inputs and behaviors. • **In-Depth Analytics:** LLMs process vast amounts of data for analysis and extraction, improving decision- • making. • **User-Centric Design:** The LLMs and multi-modal technologies ensure interfaces cater to diverse user needs.
Human-Machine Interface	• **Available methods:** Physical controls, voice commands • **Description:** Tactile and auditory interaction options • **Cons:** Accessibility issues for some users; voice commands fail in noise	• **Available methods:** LLMs, multi-modal • **Description:** Integrates language and sensory data for interaction. • **Pros:** Intuitive interfaces, complex data handling	
Data Analysis and Feature Extraction	• **Available methods:** Batch processing, rule-based analytics • **Description:** Processes large data volumes, uses set rules for analysis. • **Cons:** Slow and intensive processing, might overlook new patterns	• **Available methods:** RAG, MOE • **Description:** RAG for unstructured data analysis, MOE for sequential data • **Pros:** Deep analysis, enhanced machine learning inputs	

TABLE 2. Summary of potential issues of IAI.

algorithms in response to changing network conditions. Utilizing IAI can significantly enhance this process by enabling the system to learn and adapt its signal-processing strategies based on real-time data and interactions. For example, in [11], the authors proposed an IAI framework for steganographic distortion learning framework. This framework employs a GAN composing two subnetworks, i.e., a steganographic generator and a steganalytic discriminator, which through adversarial training, learns the probabilities of embedding changes in pixels to minimize detectability. In this framework, implicit interactions in the form of adaptive learning allow the model to evolve from simple random embedding to advanced content-adaptive embedding, significantly improving its security performance with each iteration. Simulation results showed after 180,000 training iterations, the IAI model's system performance steadily improved, achieving an average embedding rate close to the targeted capacity.

Predictive Resource Allocation: Predictive Resource Allocation refers to the allocation of network resources like bandwidth and power based on predicted future demands and usage patterns. Implementing IAI in this domain can result in more efficient and responsive resource management. For example, in [9], the authors proposed a DRL-based IAI method to optimize system energy efficiency (EE) by adjusting beamforming vectors, power splitting ratios, and phase shifts, considering users' quality of service (QoS) and the transmitter's power constraints. The proposed IAI method implicitly interacted with the whole network environment, learning to fine-tune these parameters without direct human input to satisfy energy harvesting and communication QoS requirements. Simulation results reveal that this approach yields EE close to an upper bound scheme (i.e., about 2% performance gap) while significantly reducing computation time (i.e., five orders of magnitude), particularly in dynamic wireless conditions.

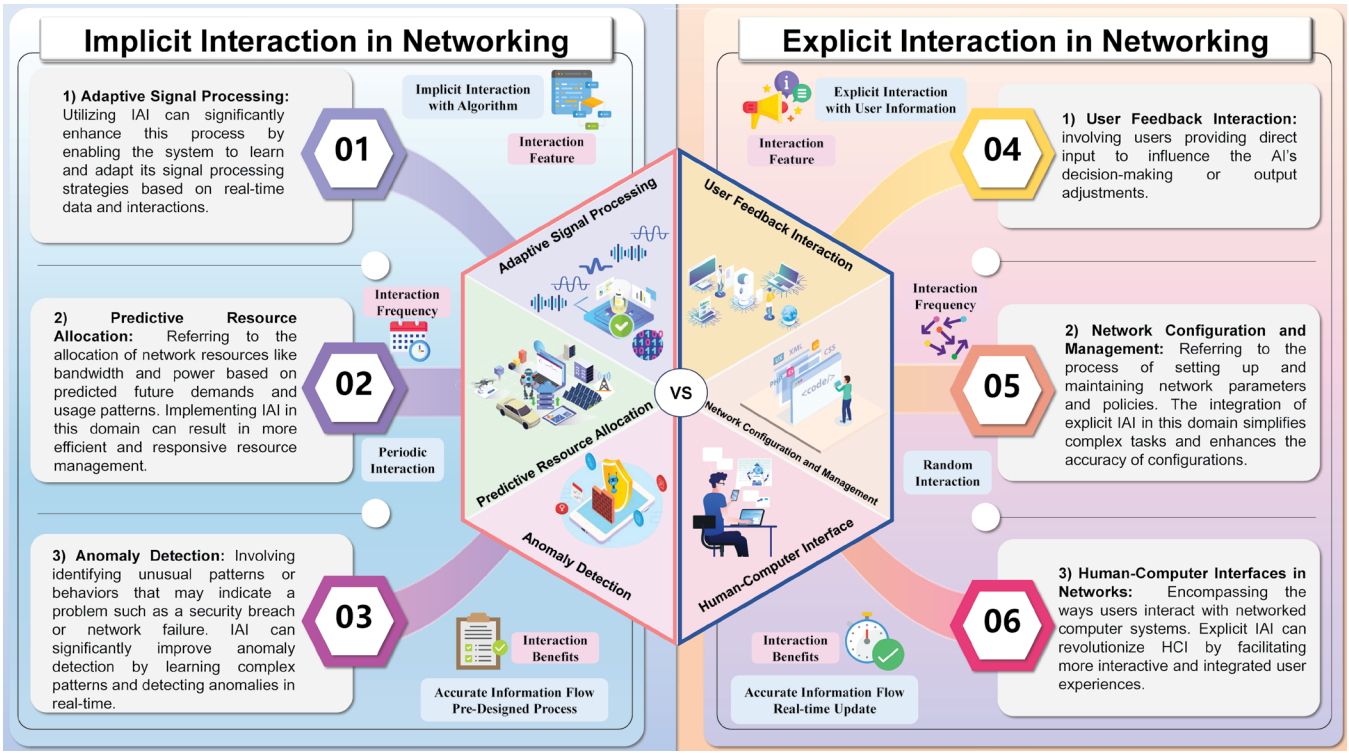

TABLE 3. Summary of implicit and explicit interaction in networking.

Anomaly Detection: Anomaly Detection in networking involves identifying unusual patterns or behaviors that may indicate a problem, such as a security breach or network failure. IAI can significantly improve anomaly detection by learning complex patterns and detecting anomalies in real-time. For example, in [12], the authors proposed an IAI framework to automatically identify dependencies between sensors for anomaly detection in multivariate time series data. The IAI framework employed a Gumbel-Softmax Sampling-based connection learning policy to automatically learn graph structures depicting sensor dependencies, and integrates this with graph convolutions and an IAI architecture for efficient anomaly detection. The framework's implicit interaction involved determining bi-directional connections among sensors, where high probability values imply strong connections, enhancing the system's ability to predict and identify anomalies. This method is effective as it incorporates user feedback and network interactions to refine its anomaly detection capability continually. Simulation results prove that the TBM-based IAI framework improves precision, recall, and F1-score by at least 6%, 16%, and 11%, respectively, compared with traditional methods in anomaly detection.

Lesson learned: The exploration of implicit IAI in networking underscores the integral role of automated adaptability and self-learning in enhancing network efficiency and security. A major benefit of implementing implicit IAI is its proficiency in navigating the complexities of dynamic network environments. Implicit IAI, through mechanisms like machine learning algorithms and predictive analytics, excels in discerning and adjusting to these changes autonomously. For instance, in adaptive signal processing, IAI has shown a notable improvement in efficiency, with performance enhancements measurable at approximately 10-15% in complex scenarios. In resource allocation, IAI has not only achieved near-optimal energy efficiency but also accelerated convergence speed by up to five times compared to traditional methods. Additionally, in anomaly detection, IAI's continuous learning from network interactions has led to a substantial increase in accuracy, outperforming traditional methods by at least 6% to 16% in key metrics. These improvements highlight IAIs real-time adaptability, faster processing, and reduced reliance on large data sets.

Explicit Interaction in Networking

Explicit interaction in networking involves deliberate and direct engagement between users or network administrators and the AI system. This type of interaction facilitates precise control over network functionalities and enhances user involvement in decision-making processes, leading to more accurate and user-centric outcomes [13], [14], [15]. Some major applications of explicit interaction in networking are as follows:

User Feedback Interaction: User feedback interaction involves users providing direct input to influence the AI's decision-making or output adjustments. IAI enhances this process by enabling more intuitive and responsive interactions, where user feedback can dynamically shape AI responses or network adjustments. For example, in [13], the authors introduced the Query Generation Assistant, a search interface that uses the LLM-based IAI method to facilitate interactive and automatic query generation in challenging search scenarios such as cross-lingual retrieval. This system allows users to actively interact in refining queries generated by LLMs, providing feedback and edits throughout the process.

It significantly improves qualitative analysis in complex search tasks and is a valuable tool for conducting human-in-the-loop simulations.

Network Configuration and Management: Network configuration and management refers to the process of setting up and maintaining network parameters and policies. The integration of explicit IAI in this domain simplifies complex tasks and enhances the accuracy of configurations. For example, in [14], the authors explored the use of the LLM-based IAI method to manage network configuration tasks. They introduce NETBUDDY, a system that translates high-level natural language requirements into low-level network configurations. This method involves explicit interactions where NETBUDDY decomposes the translation process into multiple steps, ensuring more accurate and efficient configurations, and demonstrating considerable improvement in simplifying and automating network management tasks. The simulation results showed that the method adopted not only ensures network accuracy but also increases efficiency by approximately 6 times compared with traditional ones.

Human-Computer Interfaces: Human-computer interfaces (HCI) in networking encompasses the ways users interact with networked computer systems. Explicit IAI can revolutionize HCI by facilitating more interactive and integrated user experiences. For example, in [15], the authors investigated machine-in-the-loop creative writing using a multi-modal based IAI method that suggested improvements through sight, sound, and language, marking an innovative approach in AI-assisted creativity. This method engaged human users in explicit interactions, where they incorporated AI-generated suggestions into their writing, a process involving adaptability and cognitive integration. Simulations effectively demonstrated the potential of IAI method to enhance the creativity of human-machine collaboration.

Lesson learned: Explicit IAI interactions in networking emphasize the critical role of user involvement in shaping AI-driven functionalities. This approach highlights the integration of user feedback into AI systems, resulting in more precise and customized network solutions. Unlike conventional AI, which relies on predefined algorithms, explicit IAI uses user input to refine its decision-making. This direct engagement aligns AI outputs with user-specific needs and preferences. For instance, digital twins and semantic communications can use explicit IAI to interpret and prioritize data based on its meaning and relevance to user requirements. This not only ensures accuracy but also personalizes network operations to individual contexts. Furthermore, explicit IAI is able to create a collaborative environment, merging human expertise with AI capabilities (i.e., MoE), thereby boosting the overall efficiency and effectiveness of network management.

Case Study: IAI-Enabled Problem Formulation Framework

In this section, we propose a framework that utilizes an IAI agent with RAG to help network users and designers formulate optimization problems in the network domain.

Motivation

In wireless network resource allocation, modeling complex real-world scenarios as mathematical optimization problems have traditionally required a deep understanding of complex equations and methods. This task can be challenging, particularly for newcomers or those with interdisciplinary backgrounds.

Fortunately, the IAI framework offers a transformative solution to these challenges. By interpreting the network environment and goals as defined by designers, the IAI system automatically formulates the appropriate optimization problem. This advancement speeds up the problem formulation process and importantly ensures that the optimization models accurately reflect important details of network scenarios. Acting as a cognitive intermediary, the IAI framework enhances human capabilities with AI-driven insights, reducing the complexity and manpower demand and enhancing the convenience usually needed for such tasks. Furthermore, by automating the problem formulation phase [10], IAI minimizes the need for manual input, thereby reducing the potential for common errors in traditional settings where designers must accurately define all parameters and constraints manually. While this innovation offers distinct advantages to new network designers, its main strengths are in making the optimization process more straightforward, efficiently managing network resources, and improving the accuracy and reliability of network models.

Proposed Framework

Accordingly, to effectively generate problem modeling, as shown in Fig. 2, we propose an IAI-enabled problem formulation framework. Note that the IAI framework is suitable for centralized network management systems with less constrained computational resources. In these settings, the network's central server handles complex calculations and data processing, reducinng the burden on individual devices. The IAI framework consists of the following units.

Perception: The Perception component of the IAI-enabled problem formulation framework is akin to human sensory reception, drawing from diverse sources and modalities. This multi-modal approach enables the IAI to assimilate information from text, visuals, and numerical data. Upon receiving this data, the system employs prompt engineering to transform raw information into structured embeddings. These embeddings are designed to encapsulate the complexity of the input data in a format that is readily interpretable by the IAI agents. Consequently, the Perception component prepares the IAI to comprehend the nuances of the environment and facilitates informed decision-making. In our framework, the Perception component thus serves as the foundational interface between multifaceted data inputs and the IAI's cognitive mechanisms, ensuring that the system's responses are grounded in a comprehensive understanding of the environmental context.

Brain: The Brain is the central component of the IAI system, functioning in a three-unit structure: database, storage, and decision-making.

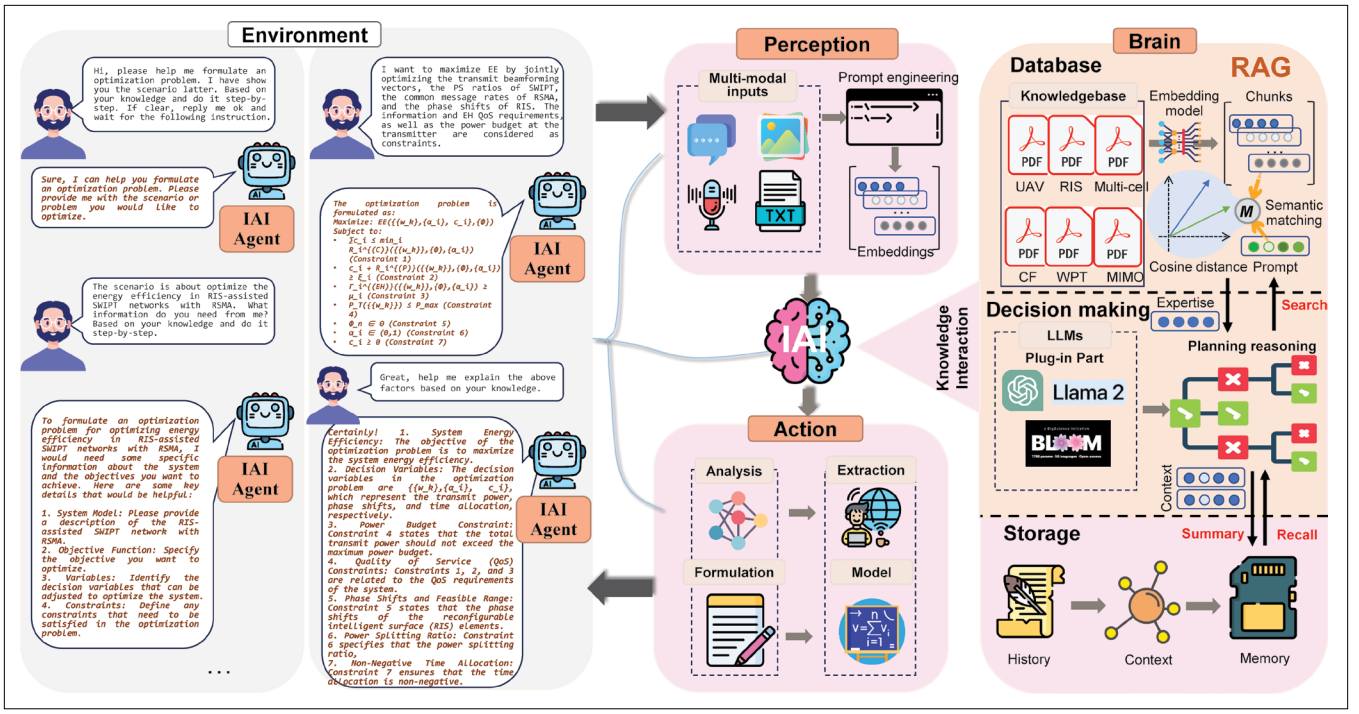

FIGURE 2. The IAI-enabled problem formulation framework presents three key components: brain, perception, and action. The brain units with LLM and RAG modules are the central processors, handling essential tasks such as memory retention, retrieval processing, and decision-making. Meanwhile, the perception module with prompt engineering is responsible for acquiring and interpreting diverse environmental data. Lastly, the action module implements responses and interacts with the environment, utilizing multi-modal extraction and dialogue analysis for execution.

The RAG database contains a wealth of searchable academic texts, such as those pertaining to unmanned aerial vehicles (UAVs), reconfigurable intelligent surfaces (RISs), wireless power transfer (WPT), and more, as its training dataset. This corpus of knowledge is transformed into text embeddings that are stored within the IAI's knowledge base. For the decision-making, the system adopts a plug-in architecture for selecting the appropriate LLM, such as GPT 4, LLAMA 2, Gemini, and Bloom. When the database unit and the decision-making unit are combined, they form the RAG module. RAG allows the brain to infer, learn, retrieve, and reason, combining its capabilities with database resources to plan and execute actions to provide appropriate strategies for generating problems. For the storage, it is a repository of the agent's historical observations, thoughts, and actions. It enables the agent to effectively recall and apply previous strategies when tackling complex reasoning tasks, similar to how humans draw on memory to navigate unfamiliar situations. The Brain's architecture is integral to the IAI's learning process, providing a contextual understanding and the ability to utilize historical insights for informed decision-making, thereby enhancing the academic rigor of the system's cognitive process.

Action: Within the IAI framework, the Action component responds to the Brain's directions and performs tasks in the network environment. This unit of the system follows a clear four-step method. It starts by analyzing network designer inputs, guided by the Brain's insights. Then, it pulls useful information from its knowledge base. Next, the module shapes the network problem into a clear mathematical form. The final step is to present the formulated optimization model. The Action module turns plans into real results, making sure that the IAI's decisions have a direct and practical impact on the network.

Environment: The Environment component of the IAI framework is pivotal in encapsulating the application domain for the IAI agent's operation. It represents the real-world network scenarios as described by the network designer, forming the foundation for the IAI's problem formulation process. For instance, when a network designer wants to formulate an optimization problem for an RIS-assisted SWIPT network with RSMA, the interaction begins with the network designer's initial request for assistance. The IAI agent then prompts the network designer for specific details, such as the system model, the optimization objective, the decision variables, and necessary constraints. Through this iterative dialogue, the network designer provides the required information step-by-step. In this process, IAI uses existing datasets and learning models to generate models, which is similar to GAI. However, unlike GAI, this is just the beginning of the process. After the initial generation, IAI presents the model to the user (i.e., network designer) for review. Then, the network designer might describe the network's configuration for optimizing EE, and the IAI agent will use the agent's information to develop an appropriate optimization problem. This interactive process allows the IAI system to thoroughly understand the network designer-defined environment and generate an optimized problem model that aligns with specific characteristics and requirements of the network scenario. It showcases how the

Environment component acts as a collaborative stage, integrating network designer inputs with IAI functionality to create tailored problem-solving approaches for complex network environments.

Computational Complexity Analysis: The computational complexity of our IAI framework is mainly determined by two processes, i.e., text data embedding and semantic similarity calculation. The text data embedding converts text data into a vector representation with a complexity of $O(n \times d)$, where n is the count of text items and d is the dimension of the embedding. Next, the RAG module needs to obtain semantic similarity, which requires a vector comparison with a complexity of $O(m \times d)$, where m represents the number of query operations. Therefore, the total computational complexity is $O((n + m) \times d)$ for our IAI framework.

Experiments

Experimental Settings: To explore the effectiveness of the proposed IAI agent, we apply it to solve a real-world network optimization problem. In our experiments, the pluggable LLM module is implemented by OpenAI APIs for calling the GPT-4 model. The network-oriented knowledge base and context memory are built on LangChain, where the chunk size, chunk overlap, and retrieval results are set as 1000, 200, and 3, respectively.

Effectiveness of the IAI Agent: As shown in Fig. 2, the user intends to optimize the energy efficiency in RIS-assisted simultaneous wireless information and power transfer (SWIPT) networks with rate splitting multiple access (RSMA) [9]. Traditionally, the network designer needs to search numerous articles, select an appropriate model, and apply it to the problem. With the help of an IAI agent, such a process can be realized automatically through four rounds of user-agent interactions. To be specific, the user first states the requirements, i.e., formulating a network optimization problem. Here, the chain-of-thought prompting is applied to enhance the reasoning ability of the IAI agent. Then, the user further illustrates the specific scenario to the IAI agent, which then returns the step-by-step guidance of the entire optimization process. Afterward, the users clarify the optimization objective. Empowered by information retrieval techniques, the IAI agent will jointly search the local knowledge base and perform the inference. The resulting problem formulation perfectly aligns with the ground truth (i.e., the standard problem formulation from academic journals). Moreover, the user can acquire detailed explanations of all the factors.

Knowledge Base Settings: With the effectiveness of the IAI agent being verified, we further analyze the influence of knowledge base settings on the interaction quality. Specifically, we adjust the chunk size to organize knowledge embeddings in the RAG module and test the number of interaction rounds required to solve the aforementioned task. As shown in Fig. 3, when $k = 1$ and $k = 3$, setting the chunk size as 2000 or 3000 can lead to the best interaction quality since the knowledge chunks can be effectively fetched. In contrast, if the chunk size decreases to 1000 , the problem cannot be completely formulated within 10 rounds of interactions because the knowledge that can be referred to is too limited. Note that extremely large chunk sizes are also undesirable. For instance,

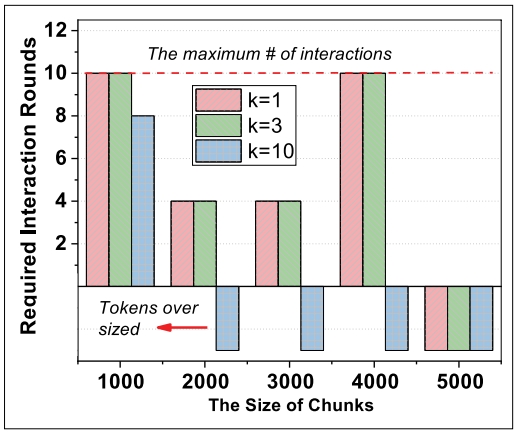

FIGURE 3. The chunk size versus the number of interaction rounds required to solve the network optimization task, where k represents the number of chunks that the LLM puts into context in each round of interaction.

when the chunk size is 4000 , the problem formulation cannot be completed eiter since the LLM can hardly extract useful knowledge from large chunks. When the chunk size reaches 5000, the IAI agent cannot even perform inference due to the limited context size. If we increase k to 10 , i.e., allowing the IAI agent to retrieve more chunks in each round, the interaction quality improves when the chunk size is 1000.

Performance Comparison: To demonstrate the efficacy of our proposed IAI framework, Fig. 4 shows an example to present a comparison of different ways for generating optimization problems. It is evident that the first step in creating an optimization problem involves providing a detailed problem description. Our IAI framework then interacts with this description, as illustrated in Fig. 2, to automatically derive the corresponding optimization problem. In contrast, network designers often rely on their experience to manually create optimization problems. However, this manual approach is less accurate and less efficient, especially when designers are new or unfamiliar with certain aspects, such as overlooking or incorrectly defining constraints (e.g., decoding RSMA constraints in our example). Such oversights can lead to flawed optimization problems. Upon applying the same implicit IAI method, i.e., the PPO-based optimization method for solving, it is observed that the performance of algorithms under our framework closely matches that of the original real problem formulation (i.e., in [9]), whereas manually designed optimization problems yield the lowest results. This difference is attributed to the strength of the IAI framework, which leverages LLM capabilities and features within LangChain to generate precise optimization problems through an interactive process. Therefore, the IAI framework ensures a more accurate and efficient problem formulation compared to traditional manual methods.

Lesson learned: Through the above experiments, we can observe that the IAI goes beyond GAI, which only considers the quality of generated content. Instead, the IAI agent is designed to help users accomplish specific tasks. Therefore, not only the core LLM but also the settings for

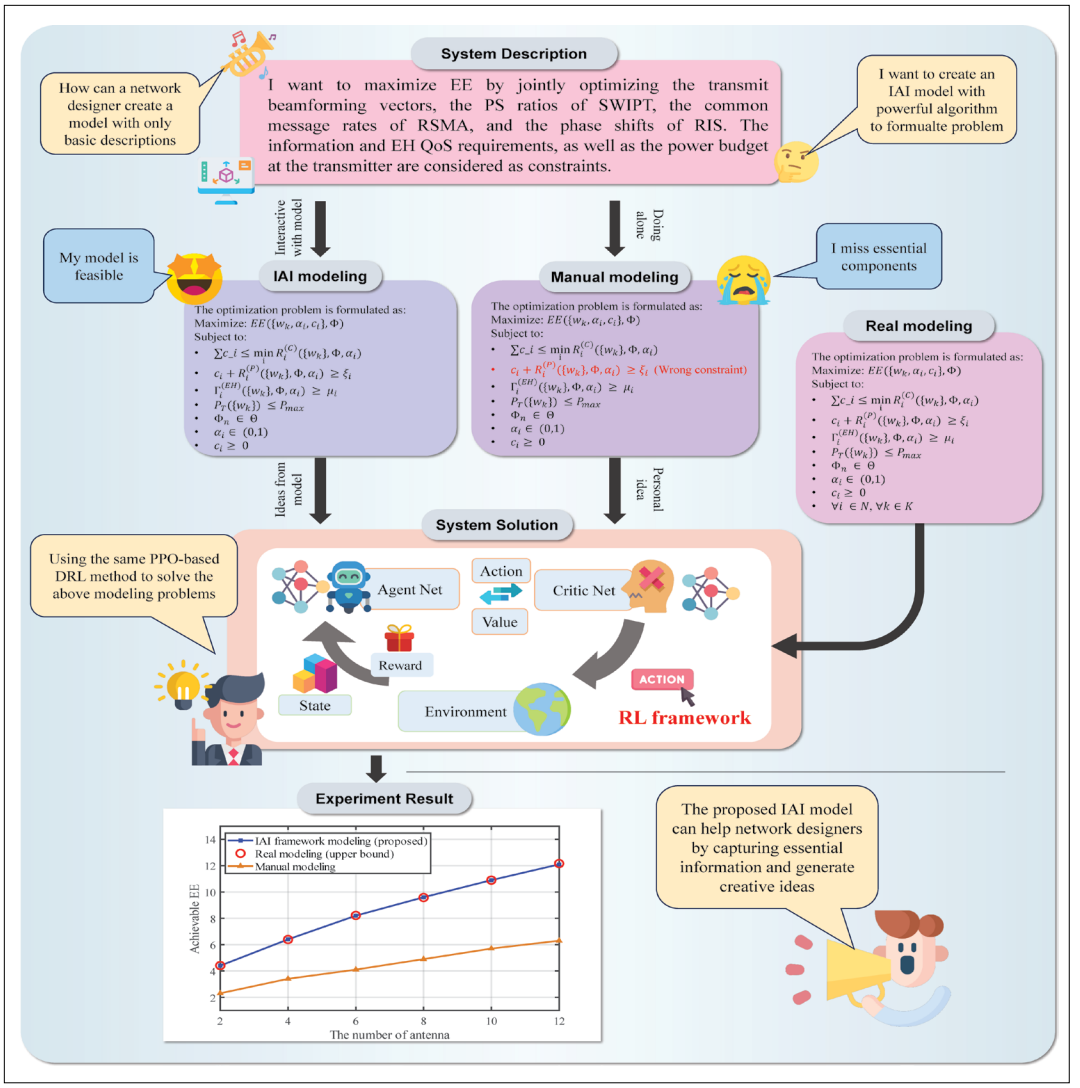

FIGURE 4. Comparison of system performance under various optimization problem generation methods. The figure displays the effectiveness of the proposed IAI framework modeling against traditional real modeling (upper bound) and manual modeling approaches, where the PPO-based DRL method is set as the solution method to demonstrate the performance results.

other modules that guarantee the smoothness and quality of interaction (e.g., the knowledge base and memory) should be well crafted.

Future Directions

In this section, we outline three main future directions for the improvement of IAI-enabled networking.

Integration With Emerging Technologies: IAI can enhance B5G/6G networks, particularly in real-time data processing, adapting to bandwidth and latency needs. Implementing a decentralized multi-agent approach where multiple IAI agents operate in a coordinated manner to manage different network segments. Moreover, edge computing frameworks can be utilized to process data locally at IAI network edges, thereby decreasing latency and minimizing the bandwidth required for central data processing. This method reduces the burden on any single agent and improves overall efficiency and responsiveness.

Security Aspects of IAI-Enabled Networks: Improving security in IAI networks is crucial. Future research should focus on IAI-driven protocols for early threat detection and adaptive response, automating security management, and evolving with emerging threats.

Evaluation of IAI Systems: It is important to design a model that can evaluate IAI. In the future, AI evaluation criteria should be given to evaluate models generated by IAI instead of human user evaluation.

Conclusion

In this article, we have explored the integration and enhancement of IAI in networking. We have proposed a problem formulation framework by using IAI with RAG, where the effectiveness of the framework was verified through simulation results. Finally, some potential research directions for IAI-based networks were outlined.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62102099 and Grant U22A2054; in part by the Pearl River Talent Recruitment Program under Grant 2021QN02S643; in part by

the National Research Foundation, Singapore, and Infocomm Media Development Authority under its Future Communications Research and Development Programme, Defence Science Organisation (DSO) National Laboratories through the AI Singapore Programme (AISG) under Award AISG2-RP-2020-019 and Award FCP-ASTAR-TG-2022-003; in part by the Singapore Ministry of Education (MOE) Tier 1 under Grant RG87/22; and in part by the U.S. National Science Foundation under Grant ECCS-2335876.

References

[1] W. Xu et al., "Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI," *Int. J. Hum.-Comput. Interact.*, vol. 39, no. 3, pp. 494–518, Feb. 2023.
[2] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.
[3] Y. Lu, "Artificial intelligence: A survey on evolution, models, applications and future trends," *J. Manage. Anal.*, vol. 6, no. 1, pp. 1–29, Jan. 2019.
[4] Z. Li et al., "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.
[5] H. Du et al., "Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization," 2023, *arXiv:2308.05384*.
[6] R. Zhang et al., "Interactive generative AI agents for satellite networks through a mixture of experts transmission," 2024, *arXiv:2404.09134*.
[7] Z. Chen et al., "Octavius: Mitigating task interference in MLLMs via LoRA-MoE," 2023, *arXiv:2311.02684*.
[8] R. Zhang et al., "Generative AI-enabled vehicular networks: Fundamentals, framework, and case study," 2023, *arXiv:2304.11098*.
[9] R. Zhang et al., "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.
[10] Z. Xi et al., "The rise and potential of large language model based agents: A survey," 2023, *arXiv:2309.07864*.
[11] W. Tang et al., "Automatic steganographic distortion learning using a generative adversarial network," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1547–1551, Oct. 2017.
[12] Z. Chen et al., "Learning graph structures with transformer for multivariate time-series anomaly detection in IoT," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9179–9189, Jun. 2021.
[13] K. D. Dhole, R. Chandradevan, and E. Agichtein, "An interactive query generation assistant using LLM-based prompt modification and user feedback," 2023, *arXiv:2311.11226*.
[14] C. Wang et al., "Making network configuration human friendly," 2023, *arXiv:2309.06342*.
[15] N. Singh et al., "Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence," *ACM Trans. Comput.-Hum. Interact.*, vol. 30, no. 5, pp. 1–57, Oct. 2023.

Biographies

Ruichen Zhang (ruichen.zhang@ntu.edu.sg) received the Ph.D. degree from Beijing Jiaotong University (BJTU), China, in 2023. He is currently working as a Post-Doctoral Research Fellow with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. His research interests include reinforcement learning-enabled wireless networks, the Internet of Things, and generative models.

Hongyang Du (hongyang001@e.ntu.edu.sg) received the B.S. degree from Beijing Jiaotong University, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include semantic communications, generative AI, and network management.

Yinqiu Liu (yinqiu001@e.ntu.edu.sg) received the B.E. degree from the Nanjing University of Posts and Telecommunications, China, in 2020, and the M.Sc. degree from the University of California, Los Angeles, 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 wireless communications, mobile AIGC, and generative AI.

Dusit Niyato (Fellow, IEEE) (dniyato@ntu.edu.sg) received the B.Eng. degree from the King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada, in 2008. He is currently a Professor with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include the Internet of Things (IoT), machine learning, and incentive mechanism design.

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

Sumei Sun (Fellow, IEEE) (sunsm@i2r.a-star.edu.sg) is currently the Executive Director of the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore. She also holds an adjunct appointment with the National University of Singapore and joint appointment with the Singapore Institute of Technology, as a Full Professor. Her current research interests include next-generation wireless communications, joint communication-sensing-computing-control design, the industrial Internet of Things, applied deep learning, and artificial intelligence. She is a member of the IEEE Vehicular Technology Society Board of Governors from 2022 to 2024 and a fellow of the Academy of Engineering Singapore.

Xuemin (Sherman) Shen (Fellow, IEEE) (sshen@uwaterloo.ca) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor of electrical and computer engineering, University of Waterloo, Canada. His research interests include wireless communication networks, including capacity analysis, mobility, and radio resource management.

H. Vincent Poor (Life Fellow, IEEE) (poor@princeton.edu) received the Ph.D. degree in EECS from Princeton University, Princeton, NJ, USA, in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990, he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor. His research interests include information theory and signal processing, and their applications in wireless networks, energy systems, and related fields.