Exploring Multi-Agent Dynamics for Generative AI and Large Language Models in Mobile Edge Networks

Xiaoya Zheng[1]

[1]Affiliation not available

September 24, 2024

Abstract

The emergence of generative artificial intelligence (GenAI) marks a significant breakthrough in the realm of AI. Recently, GenAI and large language models (LLMs) have garnered tremendous attention due to their capability to automatically generate data based on the given original patterns and dataset. However, traditional GenAI-LLMs mechanisms may result in low-quality output content and considerable creation time. The Internet of agents (IoA) can address these challenges by providing a flexible and scalable platform for integrating diverse agents, enabling seamless communication and coordination in mobile environments. Therefore, in this work, we explore the integration of multi-agent GenAI-LLMs enlightened by IoA. Specifically, we first provide a brief introduction to multi-agent GenAI-LLMs and their applications in different domains. Then, we demonstrate the potential of deploying the multi-agent GenAI-LLMs in mobile edge networks. Subsequently, we discuss the emerging applications and challenges when deploying multi-agent GenAI-LLMs in mobile edge networks. In the following, we propose a novel multi-agent GenAI-LLM architecture for mobile edge networks. Moreover, we conduct a case study to show the effectiveness of the proposed architecture by applying it to generate high-quality solutions in unmanned aerial vehicle (UAV) networks. Finally, several potential research directions for GenAI-LLMs in mobile edge networks are discussed.

IEEEkeywords: Multi-agent, generative AI, large language models, mobile edge networks.

Introduction

Generative artificial intelligence (GenAI) and large language models (LLMs) represent a significant advancement over traditional AI by shifting the focus from simple data analysis to the creation of new, synthetic data. Unlike traditional AI, which excels in tasks such as pattern recognition and prediction, GenAI employs models including Transformers for LLMs (Sun et al., 2024), generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion models to produce novel content, enhance privacy, and address data scarcity. This capability allows GenAI to generate creative content, simulate realistic scenarios for training purposes, and provide personalized user experiences without compromising privacy or perpetuating biases found in original data sets. As such, GenAI not only extends the application scope of AI but also offers solutions to some of the fundamental limitations of traditional models, making it a crucial development in the AI landscape.

GenAI content creation, as a subset of GenAI, is revolutionizing the way content is created and managed within many domains. In mobile edge networks, GenAI-LLMs can automate the creation process, provide personalized user experiences, and enhance network efficiency (He et al., 2024). GenAI-LLMs allow for the tailored delivery of news, entertainment, and advertisements, adapting content in real-time to suit individual

preferences and current network conditions. Additionally, they optimizes network functionalities through predictive modeling and improves interactive experiences with dynamic and responsive virtual interfaces. Overall, GenAI-LLMs significantly boost user engagement and network performance by leveraging GenAI to meet the diverse and dynamic demands of mobile edge network users.

However, the current single-agent GenAI-LLM systems, while beneficial in many contexts, have several limitations in mobile edge networks. First, single-agent systems often lack the breadth of expertise required to handle diverse content creation needs effectively. They may excel in general areas but struggle to integrate the cross-disciplinary knowledge necessary for mobile edge networks. Second, single-agent GenAI-LLM systems suffer from scalability issues as they depend on a singular processing point, which can become a bottleneck when user demands and network loads increase. Lastly, single-agent GenAI-LLM systems may not effectively personalize content due to their limited perspective on user data and interactions, potentially leading to a less optimized user experience.

The application of the Internet of agents (IoA) can address the abovementioned challenges. Specifically, inspired by the global connectivity ability of Internet, IoA enables the seamless communication and collaboration of autonomous agents across diverse ecosystems and devices, leveraging technologies including the agent integration protocol and instant-messaging framework, etc. These features above make IoA particularly suitable for mobile edge networks by enhancing scalability through distributed processing, improving expertise with specialized collaborative agents, and delivering personalized user experiences through adaptable and real-time content management.

Building on the capabilities of IoA, multi-agent GenAI-LLMs further enhance this framework by introducing a system of collaborative agents, each specializing in content creation and optimization. This setup allows for a division of labor among agents, where each applies its specific expertise to various task components (Guo et al., 2024).

Nevertheless, implementing multi-agent GenAI-LLM systems in mobile edge networks still poses several questions. Interaction mechanisms within such systems can directly determine whether multiple agents can efficiently complete a task. Generally, the interaction mechanism can be cooperative, debating, or competitive, and their structures may be layered, shared, centralized, or decentralized. Thus, how to design the interaction mechanisms for enhancing collaboration among diverse GenAI-LLM agents? Moreover, how to update the behavior of GenAI-LLM agents based on real-time changes in the network environment? In addition, how to select a suitable interaction mechanism for a specific mobile edge network scenario? Motivated by these questions, this paper presents pioneering research into multi-agent GenAI-LLM systems and their interaction mechanism selection in mobile edge networks. To the best of our knowledge, this is the first study to investigate the multi-agent dynamics of GenAI-LLMs in mobile edge networks. The contributions of this work are summarized as follows:

- We first provide the overview of multi-agent GenAI-LLM systems. Building on this, we present the interaction mechanisms of these systems and their popular applications/methods in mobile edge networks and beyond.
- We present the potential applications of multi-agent GenAI-LLM systems in mobile edge networks, including the motivations and suitable interaction mechanisms. In addition, we present an architecture that updates the behavior of GenAI-LLM agents based on real-time changes in the network environment.
- We propose a case study to show the effectiveness of the selected interaction mechanism and proposed architecture of the multi-agent GenAI-LLM system for unmanned aerial vehicle (UAV) semantic communication. Simulation results validate that the superiority of the proposed architecture in generating high-quality content in mobile edge networks.

Overview of Multi-agent GenAI-LLMs

AI Agent and GenAI Agent

AI agents are intelligent entities implemented through models, programs, or algorithms. The primary function of these agents is to generate text, images, audio, or other forms of content based on provided instructions, data, or context. AI agents are widely applied in various fields such as natural language processing, computer vision, and audio processing.

Based on this, GenAI agents introduce advanced GenAI technologies such as diffusion models (Croitoru et al., 2023), generative adversarial networks (GANs) (Creswell et al., 2018), variational autoencoders (VAEs), flow-based generative models (Cui et al., 2023), and energy-based generative models (Bond-Taylor et al., 2021). As such, GenAI agents can generate new content similar to the training data by learning the distribution of large-scale training datasets, and thus offering the following inherent advantages:

- **Autonomy:** GenAI agents can independently perform tasks and generate content, reducing the need for constant human oversight and increasing efficiency.
- **Interactivity, Continuous Learning, and Customization:** GenAI agents can interact with users and thus update their generated behaviors continuously based on feedback. This interaction process allows the agents to cater more effectively to user needs, generating more personalized and customized content.
- **Creativity:** By leveraging their capability to understand and mimic complex patterns and structures in data, GenAI agents can produce creative and novel outputs, pushing the boundaries of content creation.
- **Scalability and Flexibility:** GenAI agents are designed to handle a broad range of tasks, from simple decision-making to complex problem-solving, making them adaptable to various industries and scalable to meet growing operational demands.

Thus, GenAI agents offer significant capabilities to transform industries by automating processes, enhancing user interactions, and creating innovative content. However, when the tasks become complex and involve multiple steps, a single GenAI-LLMs agent may face challenges such as inefficiency, error-prone, and limited reusability.

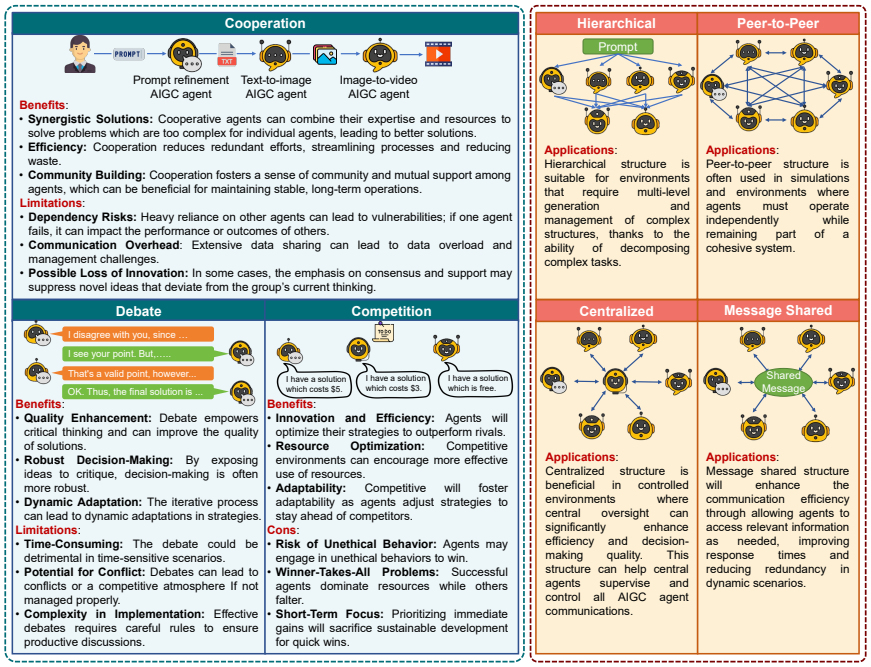

Figure 1: The interaction paradigms and structures within a multi-agent system. On the left, the diagram delineates the interaction paradigms within a multi-agent system, which are cooperation, debate, and competition, respectively. Their corresponding advantages and disadvantages are also discussed. On the right, several interaction structures are illustrated and the environment in which each structure is applicable is described.

Multi-agent GenAI-LLMs

Multi-agent GenAI-LLM systems use a set of agents specialized in content creation and optimization to cooperate in completing tasks. Compared with single GenAI-LLM agents, multi-agent GenAI-LLM systems have these advantages:

- **Expertise Diversification:** Multi-agent systems leverage the specialized skills of different agents, allowing each to handle tasks suited to their expertise, thereby enhancing the overall performance by integrating interdisciplinary knowledge. For instance, healthcare recommendation applications can introduce a team of specialist doctor agents to manage complex medical recommendations collaboratively.
- **Scalability and Efficiency:** By not depending on a single processing point, multi-agent systems distribute the workload across multiple agents. We can allocate multiple identical agents to subtasks that process large amounts of user data for parallel processing, thus avoiding bottlenecks associated with increased user demand.
- **Robust Error Handling:** The multiple agents enable cross-validation of outputs, significantly reducing the risk of errors or "hallucinations" that single-agent systems might produce. Multiple agents mitigate such errors by implementing mechanisms where multiple agents validate and cross-check each other's outputs, thereby maintaining the accuracy and reliability of the generated solution.
- **Enhanced Personalization:** Multi-agent systems can achieve a more comprehensive understanding of user preferences since each agent has a view of user interactions, which can facilitate the creation

of highly personalized content. Assigning a dedicated agent to each user to analyze their habits and preferences, thereby promoting personalized generation.

- **Cost Efficiency in Training** : These systems allow for the reuse of trained agents across similar tasks, reducing the need for repetitive training. For example, surgical and internal medicine applications can employ the same trained accurate heart rate detection agent, significantly cutting training costs and increase efficiency.
- **Knowledge Exchange** : Agents can share various types of knowledge, such as curated data, task labels, context information, and model parameters, to incorporate knowledge. For example, agents may exchange labeled data for supervised learning or share compact sets of model parameters from extensive datasets, balancing computational and privacy considerations.

The interaction mechanism within a multi-agent system is pivotal for realizing these potential advantages. Specifically, the interaction mechanism can be divided into two main components: interaction paradigm and structure. First, the interaction paradigm encompasses the styles and methods of interaction among agents, which are broadly categorized into three types (Guo et al., 2024):

- **Cooperation:** In this paradigm, agents work together towards a common goal and exchange information to develop a collective solution.
- **Debate:** This paradigm involves agents engaging in argumentative interactions where they present, defend, and critique various ideas or solutions, leading to either cooperative or conflicting outcomes.
- **Competition:** In this paradigm, agents pursue objectives that may directly conflict with those of other agents. Each agent aims to optimize its own results, potentially at the expense of others.

The benefits and limitations of these paradigms are illustrated in Fig. 1(a). As can be seen, they dictate how agents collaborate or compete and should be chosen carefully based on the desired objectives. Following this, the interaction structure refers to the organization of the multiple GenAI-LLM content creation agents, which can be categorized into four types (Guo et al., 2024):

- **Hierarchical** : This structure organizes agents into hierarchical layers, with each layer designated specific roles. Agents primarily interact within their own layer or with directly adjacent layers.
- **Centralized** : The centralized structure uses one or multiple central agents to coordinate all interactions, and other agents engage through this central node.
- **Peer-to-Peer** : In this structure, agents communicate directly with one another without the need for a central coordinating authority.
- **Message Shared** : This structure features a shared repository. Instead of interaction between agents, the agents publish and subscribe to messages based on predefined configuration files.

More details about these structures are illustrated in Fig. 1(b). Note that the structure of multi-agent GenAI-LLM systems is mainly chosen by considering the roles and places of the different agents.

Current Applications of Multi-agent GenAI-LLMs

Based on these advanced multi-agent GenAI-LLMs concepts, several multi-agent GenAI-LLM systems have excelled in content creation, strategy simulation, and problem-solving.

- **Content Creation:** Significant advancements have been made in employing multi-agent GenAI-LLM systems for content creation. For example, the authors in (Chen et al., 2023) proposed AutoAgents, which adaptively select multiple agents to form a team based on a specified text creation task. Additionally, the authors in (Wang et al., 2024) introduced AesopAgent for multimodal content creation, which transforms user prompts into scripts, images, and other formats, and subsequently integrates these contents.

- **Strategy Simulation:** Multi-agent GenAI-LLM systems are effective in simulating strategic interactions across various domains such as social sciences, gaming, psychology, economics, and policy-making. These systems reflect the complexity and diversity inherent in real-world interactions. For instance, the authors in (Xu et al., 2023) examined how multi-agent GenAI-LLM systems use past communications and experiences to enhance strategic performance. They specifically focus on the popular game *Werewolf*, illustrating the capacity of multi-agent GenAI-LLM systems for strategic reflection and improvement.
- **Problem Solving:** Multi-agent GenAI-LLM systems also excel in solving diverse problems through effective collaboration. For instance, the authors in (Xiong et al., 2023) introduced a formal debate framework to conduct a three-stage debate among GenAI-LLM agents, addressing examination and inter-consistency issues. Additionally, the authors in (Hong et al., 2023) introduced MetaGPT, a meta-programming framework designed to solve software development problems by integrating human-like standardized operating procedures.

Part I

Current Applications	Reference	Paradigm	Structure	Multi-agent Task
Content Creation	[6]	Cooperation	Hierarchical	Form customized AI teams to deal with complex tasks, e.g., software development.
	[7]	Cooperation	Hierarchical	Convert user proposals into multimodal contents, and integrate these contents into videos
Strategy Simulation	[8]	Cooperation, Debate, Competition	Peer-to-Peer	Enhance strategic performance by using past communications and experiences
Problem Solving	[9]	Debate	Centralized, Peer-to-Peer	Effectively collaborate to ultimately achieve a consensus through debate
	[10]	Cooperation	Hierarchical, Message Shared	Achieving meta-programming by integrating human-like standardized operating procedures.

Part II

	Potential Applications	Paradigm	Structure	Multi-agent Task
1	Edge-Cloud Generation Systems	Cooperation	Hierarchical	Generate the targeted content based on personal information
2	Heterogeneous Unmanned Devices Networking	Competition	Peer-to-Peer	Control and networking of heterogeneous unmanned devices
3	Near-User Content Caching	Debate	Centralized, Peer-to-Peer	Learn preferences of users, and thereby cache videos, applications, and news based on these preferences.
4	Generative Intelligence Transportation Systems	Cooperation	Hierarchical	Increasing the intelligence, coverage, and responsiveness of intelligence transportation systems
5	Semantic Communications	Debate	Peer-to-Peer	Enhance semantic communication systems by generating more contextually appropriate response

Figure 2: Summary of the applications of multi-agent GenAI-LLMs. Part I describes current applications of multi-agent GenAI-LLMs; Part II denotes potential applications of multi-agent GenAI-LLMs in mobile networks.

The interaction mechanisms of these studies are detailed in Fig. 2. The multi-agent GenAI-LLM systems examined in these studies have demonstrated strong performance in content creation, strategy simulation, and problem-solving capabilities that align well with the requirements for optimization in mobile edge networks. This situation prompts us to consider integrating multi-agent GenAI-LLM systems into mobile edge networks.

Multi-agent GenAI-LLMs in Mobile Edge Networks

In this section, we explore the application of multi-agent GenAI-LLMs in mobile edgenetworks.

Motivation

In mobile edge networks, there are several different technologies which enable the collaboration among multiple devices, such as edge computing, federated learning, and distributed machine learning. These technologies are applied to accelerate data processing and model training by distributing the computing load across multiple nodes or devices, focusing on reducing latency, and protect privacy, etc. Different from the technologies above, multi-agent generative AI models can provide innovative content by enabling real-time interactions among agents and simulating complex environments and behaviors, which have the potential to significantly enhance the efficiency, reliability, and user-centricity of mobile edge network services, offering the following benefits:

- **Reducing Information Transfer:** Deploying multi-agent systems at edge network nodes, such as edge servers, can alleviate the burden of excessive data transfer across the network. By generating user-requested content locally, latency and bandwidth usage are reduced, accelerating content delivery and enhancing responsiveness.
- **Fault Tolerance and Adaptability:** Distributing identical GenAI-LLM agents across neighboring servers within the network can improve the resilience and adaptability of mobile edge networks. This configuration ensures seamless content creation as users move across different network coverage areas, maintaining uninterrupted service even if some agents are compromised or offline, thereby protecting the user experience from potential disruptions.
- **Enhanced Understanding of User Requirements:** Multi-agent GenAI-LLM systems utilize specialized agents to continuously capture and analyze user preferences within the network. This focused learning strategy enables the GenAI-LLM agents to accurately identify and respond to user needs, swiftly generating personalized content and solutions.
- **Rapid Solution Creation:** The collaborative nature of multi-agent GenAI-LLM systems facilitates quicker processing and solution creation. This rapid response minimizes errors and delays in fluctuating network conditions, thereby maintaining high service quality and reliability.

Potential Applications of Multi-agent GenAI-LLMs in Mobile Edge Networks

Based on the motivations discussed, multi-agent GenAI-LLM systems present a promising solution for mobile edge networks, offering various valuable potential applications. Below, we illustrate some of these applications and analyze how multi-agent GenAI-LLMs enhance them.

- **Edge-Cloud Generation Systems:** Deploying different agents across edge, fog, and cloud servers enables edge-cloud computing systems to collaborate efficiently in content creation, thus reducing latency and enhancing responsiveness. For example, a cloud GenAI-LLMs agent might generate an initial text prompt, which a fog agent then enriches with regional and national details. Finally, an edge agent creates the targeted content based on personal information.
- **Heterogeneous Unmanned Device Networking:** Multiple unmanned devices can be equipped with distinct GenAI-LLM agents, each aware of the capabilities and functions of its respective device based on expert knowledge. These agents competitively handle networking tasks and receive rewards from task publishers. Compared to single-agent systems, multi-agent GenAI-LLMs can reduce costs and enhance task completion efficiency through such competitive interactions.
- **Near-User Content Caching:** Deploying GenAI-LLM agents close to users to learn their habits and preferences can optimize the caching of videos, applications, and news. In addition, near-user agents, in collaboration with content publisher agents, can use technologies like collaborative filtering to debate and determine user preferences, thereby optimizing content caching.
- **Intelligent Transportation Systems:** Drones and unmanned vehicles can be equipped with agents that handle simple responses, management, and monitoring functions. Each agent is responsible for

a specific traffic area and shares messages to facilitate experience sharing and decision-making, thus enhancing the intelligence, coverage, and responsiveness of transportation systems.

- **Semantic Communications:** Multi-agent GenAI-LLM systems improve semantic communication by generating contextually relevant responses, ensuring accurate semantic exchanges. For instance, sender GenAI-LLM agents might engage in debates to preserve essential information during transmission, while receiver GenAI-LLM agents use similar strategies for precise decoding.

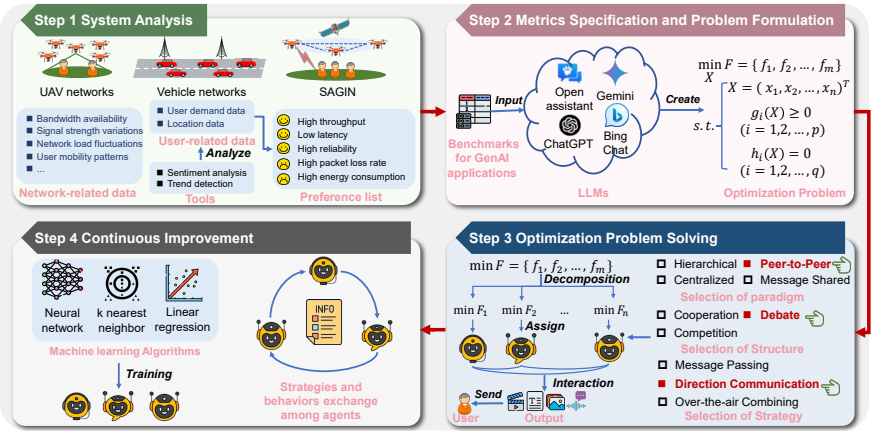

Figure 3: The framework of the proposed architecture. In step 1, the network-related and user-related data is obtained, and then tools are used to analyze the user preference and other information. In step 2, benchmarks for GenAI applications are identfied and LLMs are used to formulate the optimization problem. In step 3, the interaction paradigm and structure is first confirmed. Then, the optimization problem is decomposed into several subproblems, each of which is assigned to an agent for execution. When the interaction among multiple agents ends, the output will be forwarded to the original user. In step 5, continuous improvement is conducted via knowledge exchange among agents and continuous learning with the assistance of machine learning algorithms.

The interaction paradigms and structures of potential applications for deploying multi-agent GenAI-LLMs in mobile edge networks are illustrated in Fig. 2. These applications demonstrate how different roles and tasks are assigned to multiple GenAI-LLM agents, resulting in variations in their interaction mechanisms. Ensuring that the interaction mechanism maximizes system efficiency and adapting it to dynamic environments pose significant challenges. To address these challenges, we propose a new architecture designed to explore effective solutions.

The Proposed Architecture

Based on the characteristics of the multi-agent GenAI and mobile edge networks, we propose an architecture for building a multi-agent GenAI-LLMs system for mobile edge network applications, as shown in Fig. 3. The architecture has the following steps.

- **Step 1: System Analysis.** We begin with a comprehensive analysis of bandwidth availability, signal strength variations, and network load fluctuations, both over time and across different geographical regions. Moreover, we assess user mobility patterns to forecast changes in network demand, which involves examining user density variations during various times and events, as well as their movement between network cells. Our analysis encompasses user demand data collected from mobile devices, network traffic, and end-to-end feedback. This phase also utilizes tools such as sentiment analysis and trend detection algorithms to uncover user preferences and identify areas for improvement in mobile

edge network applications. For instance, when a user aims to offload computing tasks to nearby edge servers, minimizing service latency is a critical concern.

- **Step 2: Metrics Specification and Problem Formulation.** Based on the results from the system analysis, we define benchmarks for GenAI applications, including image resolution, video frame rates, system latency, and data throughput. Next, with the assistance of GenAI-LLMs, we formulate the optimization problem by incorporating the user requirements and performance metrics, while taking into account constraints related to wireless connectivity, computing resources, and user mobility.

- **Step 3: Optimization Problem Solving.** The process to solve the formulated problem are divided as follows:

  First , based on the desired properties of the formulated problem, we select the interaction paradigm among multiple agents. For example, a cooperation paradigm is chosen to handle efficiency-driven problems.

  Second , we determine the interaction structure among multiple agents based on the features of the optimization problem. For instance, security-related problems, such as security protocol updates, are better suited to a centralized structure since uniform implementation across the network is essential.

  Third , for implementing the interaction mechanism, we choose a specific strategy from direct communication, over-the-air combining, and message passing based on the determined interaction paradigm and structure. For instance, we adopt over-the-air combining strategy when a centralized paradigm is considered, while direct communication strategy is adopted if peer-to-peer paradigm is used. Moreover, the determination of strategy is also affected by the trade-off between the requested resources and expected performance. For instance, direct communication will be adopted if high interaction performance is required.

  Fourth , we decompose the optimization problem into multiple subproblems and assign them to different specialized agents, each of which will deal with their assigned subproblem using equipped GenAI models such as GANs and VAEs. Based on the interaction mechanism, these agents interact with each other to enhance the quality of the created content.

  Finally , if the interaction among agents is incomplete but latency nears its limit, each agent will provide the partial content they have created. This content is then aggregated and forwarded to the original user. Otherwise, the complete content of each agent is consolidated and sent instead.

- **Step 4: Continuous Improvement.** Effective strategies and learned behaviors are shared to enhance communication among agents. Moreover, we employ machine learning algorithms to support online learning, allowing agents to adjust their strategies in real-time based on current data and minimizing the need for extensive retraining downtime.

The proposed framework is highly applicable to the field of dynamic intelligent network optimization. This suitability stems from the powerful optimization capability of the proposed framework and its proficiency in multi-agent collaboration, which are particularly well-suited for addressing complex optimization problems in dynamic environments, such as edge computing resource management, dynamic bandwidth allocation, and resource scheduling.

Case Study: GenAI-LLMs for UAV-enabled Semantic Communication

In this section, we present a detailed case study on GenAI-LLMs for UAV-enabled semantic communication to demonstrate the effectiveness of the proposed architecture.

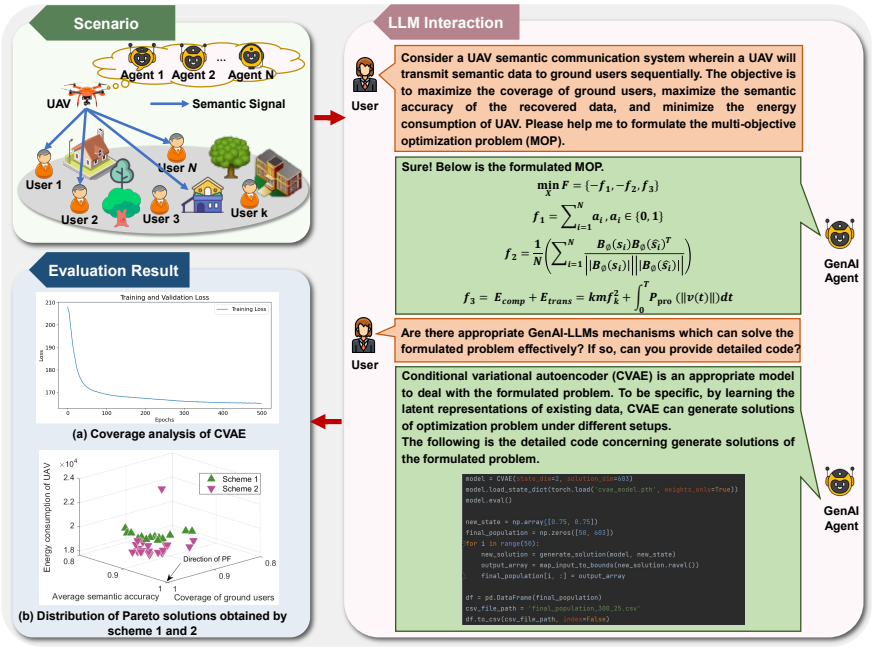

Figure 4: The experiment results of the UAV semantic communication optimization case. The scenario part shows the considered system model. The interaction part demonstrates the interaction part between user and GenAI-LLMs. The evaluation results part presents the convergence analysis and effectiveness of the proposed framework.

Scenario Description

The application of semantic communication can significantly reduce data transmission time, making it well-suited for the energy-sensitive UAV networks (Zheng et al., 2024). Therefore, we consider a UAV semantic communication system in which a hovering UAV-mounted base station equipped with LLM agents is responsible for executing semantic extraction tasks, and it then sequentially transmits the extracted data to ground users within a monitored area, as shown in Fig. 4.

In this scenario, the performance metrics to be optimized include the coverage of ground users, the semantic accuracy of the recovered data, and the energy consumption of the UAV. However, there are trade-offs among these metrics. Specifically, higher coverage of ground users implies more semantic communication tasks, resulting in the increased energy consumption of the UAV. Moreover, enhancing semantic accuracy requires additional computing and processing resources, which also leads to higher energy consumption. Therefore, we aim to formulate a multi-objective optimization problem (MOP) that simultaneously maximizes the user coverage, maximizes the semantic accuracy of the recovered data, and minimizes the total computation and communication energy consumption of the UAV for achieving the energy-efficient communication. This will be achieved by optimizing the number of transmitted semantic symbols ($\mathbb{N}$), the hovering position ($\mathbb{H}$), and the transmit power of the UAV ($\mathbb{P}$).

Leveraging the advanced content creation capabilities of GenAI-LLMs, we investigate a multi-agent GenAI-LLM system to tackle the complex optimization problem at hand. In the proposed architecture, each LLM agent is designed to specialize in optimizing distinct decision variables, thereby enhancing the overall efficiency and effectiveness of the system. To illustrate the potential of this approach, we examine two distinct GenAI-LLM schemes, with a comprehensive description of each provided below.

The scheme based on Cooperation Paradigm and Centralized Structure

In this scheme, the optimization problem is delegated to three GenAI-LLM agents, where the first agent solely optimizes the number of transmitted semantic symbols ($\mathbb{N}$), the second agent optimizes the hovering positions of the UAV ($\mathbb{H}$) and the third agent optimizes the transmit power of the UAV ($\mathbb{P}$). Then, these obtained decision variables are forwarded to a centralized controller, which aggregates the results and derives the Pareto optimal solution set by sorting the combined solutions. This approach ensures a comprehensive evaluation of the possible solutions and identifies the most effective trade-offs among the variables.

Scheme based on Debate Paradigm and Peer-to-Peer Structure

This scheme also assigns the optimization problem to three GenAI-LLM agents. Different from the scheme above, each GenAI-LLM agent will simultaneously optimize $\mathbb{N}$, $\mathbb{H}$ and $\mathbb{P}$, and a Pareto optimal solution set can be obtained accordingly. Subsequently, the multiple generative models will refine a group of solutions from the obtained Pareto solution sets in a debate manner.

Considering that most of the low-end devices, such as the UAVs discussed in this work, are not equipped with graphics processing unit (GPU) and thus may struggle to train models and process data efficiently, we define that model training is conducted on high-performance computing systems, such as servers. The trained models can then be transferred to these low-end devices for deployment.

More details about the abovementioned paradigms and structures can be found in Fig. 1.

Case Study Setup

In this section, several key parameters and architecture configuration are presented.

Key Parameters

In the system under consideration, 300 ground users awaiting semantic communication services are randomly distributed within a monitoring area of 800 $\times$ 800 m^2. Moreover, the data transmission is facilitated by DeepSC (Xie et al., 2021), which maps the data to semantic symbols for transmission. The receiver then reconstructs the original data from the received symbols. For this study, the number of symbols is constrained to a range of [1, 20], as specified in (Yan et al., 2022). In addition, the system parameters are set as follows: the bandwidth is 2 MHz, the noise power spectral density is -174 dBm/Hz, the effective switched capacitance is 10^{26}, and the frequency is 0.9 GHz. Besides, the available power for transmission is restricted to a range of [20, 30] dBm.

Architecture Configuration

The ChatGPT-4 model is employed to execute the pluggable LLM module, while the conditional VAE (CVAE) is implemented using PyTorch.

Result Analysis

Evaluation result Part (a) of Fig. 4 illustrates the convergence of the adopted generative model, i.e., CVAE, and the results show that the model reaches convergence after a period of training, indicating that it has effectively learned the underlying structure and conditional distribution of the data. In other words, the model can generate plausible data based on the given input conditions.

Evaluation result Part (b) of Fig. 4 illustrates the distribution of Pareto solutions obtained from the two schemes. It is clear that both schemes yield impressive results. Specifically, both the coverage of ground users and average semantic accuracy are maintained within the range of [0.8, 1], while the UAV energy consumption remains relatively low. This improvement is attributed to the implementation of GenAI-LLM, which leverages additional conditional information to enhance the representation of the latent space.

Consequently, this enhancement improves the ability of model to capture the data structure and distribution, leading to higher-quality generated solutions.

It can be also seen from the figure that the solutions derived from the second scheme are closer to the Pareto Frontier (PF) compared to those from the first scheme. This is because the approach in the first scheme where each GenAI-LLM agent is responsible for specific decision variables does not account for the dependencies among variables in MOP. In contrast, the second scheme which employs a debate paradigm and peer-to-peer structure, encourages multiple GenAI-LLM agents to generate decision variables simultaneously, facilitating a more comprehensive exploration of the solution space. Simulation results suggest that the second scheme is better suited for handling complex MOPs, such as the one presented in Fig. 4, as it can more effectively explores the solution space and takes advantage of parallel computing when resources are abundant.

Future Direction

In this section, we outline three key future directions for advancing and expanding multi-agent GenAI-LLM systems in mobile edge networks.

Enhanced Large-scale Hybrid Interaction Mechanism Integration

Future research could investigate the integration of diverse nested interaction paradigms and structures within large multi-agent GenAI-LLM systems. This exploration might involve developing an extensive and intricate system where multiple multi-agent GenAI-LLM systems interact in ways analogous to human social behaviors. Moreover, such a system could simulate human-like interactions and utilize the capabilities of multi-agent GenAI-LLM to effectively manage complex, large-scale network construction and communication tasks.

Mobile Edge Network Enhancing Multi-agent GenAI-LLM systems

Intensifying the relationships between mobile edge networks and multi-agent GenAI-LLM systems is another critical research direction. Specifically, researchers could investigate how mobile edge network resources can be utilized to accelerate the training of GenAI agents. This includes exploring efficient data collection and transmission mechanisms that enable agents to continuously learn and improve from real-time network data. Moreover, enhancing computational resource allocation within mobile edge networks to support rapid agent training and deployment may be essential for maintaining the high performance and adaptability of multi-agent GenAI-LLM systems.

Enhancing Security of Multi-agent GenAI-LLM systems

Security and privacy continue to be significant challenges for the multi-agent GenAI-LLM systems in mobile edge networks. Future research should focus on developing robust mechanisms to identify and mitigate the impact of malicious attacks. Attackers might attempt to manipulate user data or agent communications to disrupt model generation or produce harmful content. Thus, research could investigate advanced agent interaction mechanisms, such as the debate paradigm, to detect and counteract false or malicious information.

Conclusion

In this paper, we explored the multi-agent GenAI-LLM systems in mobile edge networks. Following this, we presented the interaction mechanisms of multi-agent GenAI-LLM systems and their popular applications/methods in mobile edge networks and beyond. Based on these, we proposed an architecture that implements the multi-agent GenAI-LLM systems in mobile edge networks. We conducted a case study on

UAV semantic communication to verify the effectiveness of the proposed architecture. Finally, we discussed several potential research directions for future extensions.

References

Generative AI for Advanced UAV Networking. (2024). *ArXiv Preprint ArXiv:2404.10556*.

Generative AI for Game Theory-based Mobile Networking. (2024). *ArXiv Preprint ArXiv:2404.09699*.

Large language model based multi-agents: A survey of progress and challenges. (2024). *ArXiv Preprint ArXiv:2402.01680*.

Diffusion models in vision: A survey. (2023). *IEEE Trans. Pattern Anal. Mach. Intell.*.

Generative adversarial networks: An overview. (2018). *IEEE Signal Process Mag*, *35*(1), 53–65.

Analysis of learning a flow-based generative model from limited sample complexity. (2023). *ArXiv Preprint ArXiv:2310.03575*.

Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. (2021). *IEEE Trans. Pattern Anal. Mach. Intell.*, *44*(11), 7327–7347.

Autoagents: A framework for automatic agent generation. (2023). *ArXiv Preprint ArXiv:2309.17288*.

AesopAgent: Agent-driven Evolutionary System on Story-to-Video Production. (2024). *ArXiv Preprint ArXiv:2403.07952*.

Exploring large language models for communication games: An empirical study on werewolf. (2023). *ArXiv Preprint ArXiv:2309.04658*.

Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. (2023). *The 2023 Conference on Empirical Methods in Natural Language Processing*.

MetaGPT: Meta programming for multi-agent collaborative framework. (2023). *ArXiv Preprint ArXiv:2308.00352*.

Energy-Efficient Semantic Communication for Aerial-Aided Edge Networks. (2024). *IEEE Trans. Green Commun.*, 1–1.

Deep Learning Enabled Semantic Communication Systems. (2021). *IEEE Trans. Signal Process.*, *69*, 2663–2675.

QoE-Aware Resource Allocation for Semantic Communication Networks. (2022). *Proc. IEEE GLOBECOM*, 3272–3277.