

# Towards Edge General Intelligence via Large Language Models: Opportunities and Challenges

Handi Chen<sup>†</sup>, Weipeng Deng<sup>†</sup>, Shuo Yang, Jinfeng Xu, Zhihan Jiang *Student Member, IEEE*,  
Edith C.H. Ngai<sup>\*</sup>, *Senior Member, IEEE*, Jiangchuan Liu, *Fellow, IEEE*, and Xue Liu, *Fellow, IEEE*

**Abstract**—Edge Intelligence (EI) has been instrumental in delivering real-time, localized services by leveraging the computational capabilities of edge networks. The integration of Large Language Models (LLMs) empowers EI to evolve into the next stage: Edge General Intelligence (EGI), enabling more adaptive and versatile applications that require advanced understanding and reasoning capabilities. However, systematic exploration in this area remains insufficient. This survey delineates the distinctions between EGI and traditional EI, categorizing LLM-empowered EGI into three conceptual systems: centralized, hybrid, and decentralized. For each system, we detail the framework designs and review existing implementations. Furthermore, we evaluate the performance and throughput of various Small Language Models (SLMs) that are more suitable for development on edge devices. This survey provides researchers with a comprehensive vision of EGI, offering insights into its vast potential and establishing a foundation for future advancements in this rapidly evolving field.

**Index Terms**—Mobile edge computing, edge general intelligence, large language models, small language models.

## I. INTRODUCTION

Edge computing has emerged as a crucial network paradigm, processing data closer to its source to reduce latency and resource demands compared to traditional cloud-centric systems. The integration of Artificial Intelligence (AI) into edge devices enables local data analysis, eliminating the need for continuous cloud communication and facilitating devices to make rapid, independent decisions. AI-enhanced Edge Intelligence (EI) advances the adaptability of distributed systems, allowing for faster responses in dynamic environments. This evolution unlocks the potential for latency-sensitive applications, transforming industries such as autonomous vehicles, smart homes, industrial automation, and healthcare by introducing more intelligent and responsive technologies.

Tracing the evolution of AI, which aims to emulate human cognitive abilities, we can categorize its development into three stages: narrow, broad, and general intelligence. Following this development trajectory, EI can likewise be divided into Edge Narrow Intelligence (ENI), Edge Broad Intelligence (EBI), and Edge General Intelligence (EGI). ENI, empowered

by extensive training, excels at performing specific, well-defined tasks, such as facial recognition. EBI transcends individual tasks, enabling systems to manage multiple interconnected functions, such as optimizing traffic flows in smart cities.

Over the past decades, traditional EI has been limited to narrow and broad intelligence within specific tasks or domains. In contrast, EGI expands these capabilities, allowing for greater autonomy and flexibility. However, achieving EGI poses significant challenges, requiring edge devices to autonomously reason, learn, and adapt across diverse knowledge, tasks, and dynamic environments. Recent advancements in Large Language Models (LLMs), such as GPT-4<sup>1</sup> and LLaMA<sup>2</sup>, represent breakthroughs in AI, which demonstrates impressive multi-modal capabilities that bridge computer vision and Natural Language Processing (NLP). These developments lay the foundation for general intelligence [1]. Leveraging LLMs to achieve EGI opens new avenues by enabling edge devices to handle complex tasks. As shown in Fig. 1, this integration paves the way for real-time, context-aware applications in various use cases, such as personalized healthcare, smart assistants, and customer service with edge networks [2].

To inspire innovation and explore the potential of LLMs-empowered EGI, this paper proposes three conceptual EGI systems, offering a comprehensive framework for researchers to integrate LLMs into the edge computing ecosystem. We begin by outlining the three levels of EI, followed by a systematic review of recent advancements in LLM-enhanced EGI. Upon this basis, we introduce three system architecture designs, including centralized, hybrid, and decentralized EGI, accompanied by a review of their implementation strategies and an analysis of their respective advantages, disadvantages, and feasibility. We finally conclude by discussing future directions, highlighting the transformative opportunities this survey holds for the evolution of EGI.

## II. BACKGROUND OF EI AND LLMs

In this section, we provide a brief overview of the background knowledge about EI and LLMs.

### A. Edge Computing and EI

Edge computing proposes a paradigm shift by relocating computational resources from distant, centralized cloud servers

H. Chen, W. Deng, S. Yang, J. Xu, Z. Jiang and E. C.H. Ngai are with the Department of Electrical and Electronic Engineering, the University of Hong Kong, Hong Kong, China (hdchen, dengf330, shuo.yang, jinfeng, zhjiang@connect.hku.hk, chngai@eee.hku.hk).

J. Liu is with the School of Computing Science, Simon Fraser University, British Columbia, Canada (jcliu@sfu.ca).

X. Liu is with the School of Computer Science, McGill University, Montreal, Quebec, Canada (xueliu@cs.mcgill.ca).

<sup>†</sup> These authors contributed equally to this work.

<sup>\*</sup> The corresponding author.

<sup>1</sup>GPT-4: <https://openai.com/index/gpt-4>

<sup>2</sup>LLaMA: <https://github.com/meta-llama>

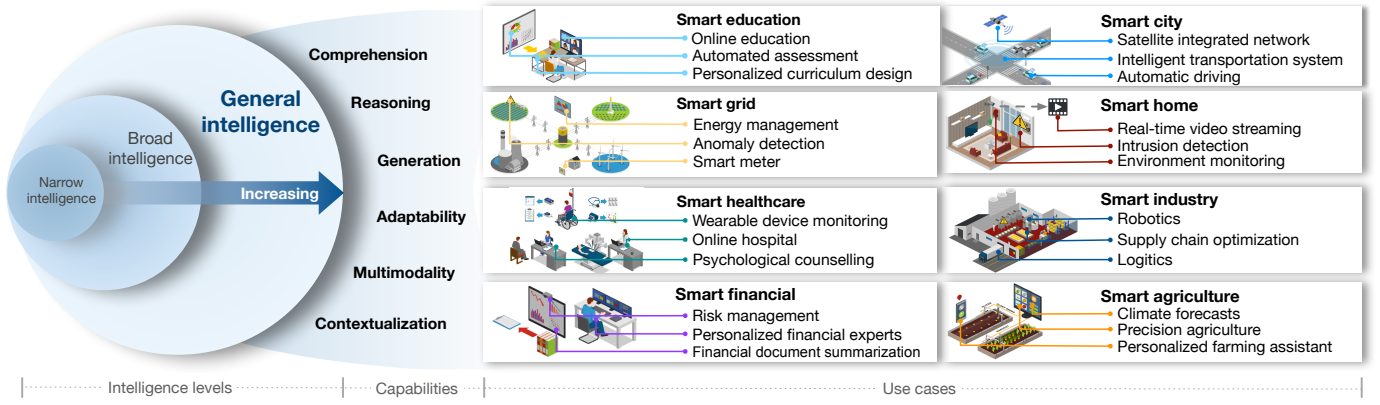


Fig. 1. An illustration of increasing intelligence from narrow to general intelligence. The main capabilities of general intelligence include comprehension (e.g., information summarization), reasoning (e.g., logic inference), generation (e.g., writing a story), adaptability (e.g., compatibility with various scenarios), multimodality (e.g., images and audios), and contextualization (e.g. context-aware dialogue), supporting various use cases.

to edge devices closer to end-users. These devices may include smartphones, industrial machines, sensors, video cameras, or any other data collection devices. Edge computing is designed to enhance the Quality of Service (QoS) for heterogeneous mobile devices and infrastructures through wireless networks by mitigating substantial transmission latency, bandwidth limitations, and connectivity issues caused by cloud computing. It facilitates faster and more efficient data processing while enhancing privacy and security measures. For example, smart traffic lights based on edge computing process sensor data in real-time to manage traffic flow more efficiently. Instead of sending data to a central server for processing, the computation is done locally at the traffic intersection, reducing latency and improving the responsiveness of the traffic management system. With advances in AI efficiency, the increasing number of IoT devices, and the ascendancy of edge computing, the potential of EI has now been actualized. The evolution of AI is shaping the future of edge computing towards EI.

### B. Large Language Models

Recent advancements in LLMs have showcased a spectrum of impressive emergent abilities, leading to significant paradigm shifts in AI. These models have displayed an exceptional capacity for understanding human instructions and demonstrating cognitive capabilities that closely mirror human thinking, paving the way for numerous opportunities across various fields. LLMs have made significant strides in complex task planning and reasoning, skills that are indispensable for problem-solving and decision-making. For instance, ChatGPT, developed by OpenAI, excels in general problem-solving, assisting users with a variety of tasks, including solving mathematical problems, devising travel plans, and analyzing investment data to guide decisions. As LLMs continue advancing towards general AI, they demonstrate an increasing ability to generate complex, contextually relevant responses across diverse domains, aligning more closely with human-like reasoning.

There has been a surge in developing domain-specific LLMs based on generic LLMs to integrate expert knowledge across

various fields. Building on the capabilities of LLMs, Retrieval-Augmented Generation (RAG) has emerged as a powerful method, enabling LLMs to dynamically access external knowledge bases during response generation. By integrating relevant information from diverse sources, RAG significantly enhances the accuracy and contextual relevance of outputs, especially for specialized tasks. Furthermore, to adapt LLMs to specific domains, fine-tuning techniques are employed to transform generic models into domain-specific experts. This strategy has proven effective in models such as MedicalGPT<sup>3</sup> and FinGPT<sup>4</sup>, which serve as valuable tools in the medical and financial sectors.

## III. LLMs: EVALUATING THE NEXT LEVEL OF EI

In this section, we first provide a concise definition of general intelligence to delineate it from traditional EI. Following this, we discuss the motivations driving the integration of general intelligence into edge computing, highlighting the potential benefits and transformative impacts on these systems. Finally, we present an overview of the three system architectures proposed in this paper.

### A. Hierarchical Cognitive Abilities of EI

Human intelligence is the cognitive ability to learn, reason, solve problems, adapt to new situations, and understand complex ideas. With this as the ultimate goal, AI systems can be constructed in a hierarchical structure, including three levels of cognitive ability: narrow, broad, and general intelligence, as shown in Fig. 1. The explanations of these three-level AI are outlined below:

- **Narrow Intelligence:** AI systems designed to perform specific tasks or solve particular problems.
- **Broad Intelligence:** AI systems that can perform a wider range of tasks across different domains but still lack the full versatility of human-like understanding.

<sup>3</sup>MedicalGPT: <https://github.com/shibing624/MedicalGPT>

<sup>4</sup>FinGPT: <https://github.com/AI4Finance-Foundation/FinGPT>

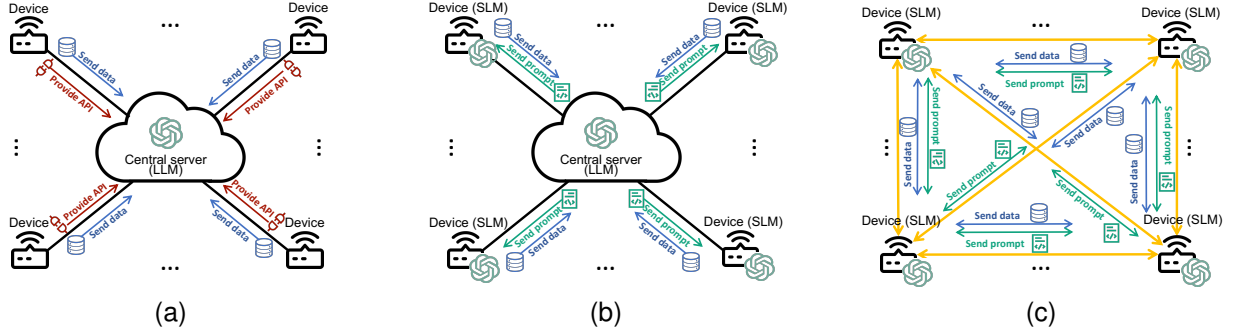


Fig. 2. Architectures of centralized, hybrid, and decentralized EGI systems (LLM and SLM represent large language model and small language model, respectively). (a) Centralized EGI system. (b) Hybrid EGI system. (c) Decentralized EGI system.

- **General Intelligence:** AI systems that possess the ability to understand, learn, and apply knowledge across a wide range of tasks, similar to human intelligence.

The development of EI is closely linked to advancements in AI, as it relies on AI algorithms for real-time data processing, efficient decision-making, and self-optimization through machine learning, and supports diverse applications in areas like smart transportation and industrial automation. To be specific, the initial level of EI implements narrow intelligence, focusing on specific tasks. For example, an edge device equipped with narrow intelligence might analyze network traffic to detect unusual patterns or potential security threats. These models excel in their specialized tasks due to extensive, task-specific training.

Expanding from narrow capabilities, edge devices with broad intelligence are able to handle a range of interconnected tasks within a domain, constructing a more versatile and intelligent system. An intelligent transportation system serves as a prime example, where edge devices integrate various functions—such as monitoring traffic flow, controlling traffic lights, and detecting congestion—to optimize traffic patterns and enhance efficiency.

However, traditional EI, whether narrow or broad, remains limited to specific tasks or domains. In contrast, EGI leverages high-level cognitive abilities to tackle a wide array of tasks across various domains. EGI is characterized by its capacity for abstract thinking, complex problem-solving, and learning from accumulated knowledge while adapting to diverse scenarios. The goal of EGI is to enhance edge computing by integrating advanced problem-solving skills, enabling systems to understand and respond to complex instructions, manage dynamic situations, and ultimately improve overall flexibility and intelligence. In this way, EGI aspires to align more closely with human intelligence, bridging the gap between specialized task execution and comprehensive cognitive abilities.

#### B. LLM-enhanced EGI

LLMs have revolutionized various fields by performing tasks that previously required human intelligence, such as coding assistance (e.g., Codex and Copilot), math problem-solving, travel planning, robot operations, and complex reasoning. Their ability to learn from diverse data sources grants

them a level of general intelligence, making them valuable in dynamic real-world settings. For instance, LLMs can recognize that both a factory worker and a motorcyclist without helmets violate safety regulations, thus eliminating the need for additional data collection. In summary, Fig. 1 illustrates the six primary capabilities of LLMs essential for constructing EGI-enhanced edge computing.

Moreover, LLMs' adaptability and compatibility with various edge sensors and hardware help address deployment challenges, enabling effective coordination of edge devices in complex environments. Enhanced by LLMs, EGI offers more flexible user interactions to manage complex tasks, leveraging strengths in NLP, task generalization, and adaptability. Fig. 1 presents a vision of several promising real-world use cases for LLM-enhanced EGI. The development of LLMs provides a viable paradigm for general intelligence, enabling EI to evolve into EGI. Furthermore, LLMs mitigate the limitations of traditional EI, which often requires specific data for distinct tasks and scenarios, thereby reducing network communication costs and enhancing data privacy and security.

### IV. EDGE GENERAL INTELLIGENCE SYSTEMS

In this section, we elaborate on the integration of LLMs into edge computing systems to achieve EGI through three conceptual architecture designs: centralized, hybrid, and decentralized systems. We provide a comprehensive analysis of these architectures, detailing their system designs, implementation strategies, and key discussions. Table I lists several representative EGI systems.

#### A. Centralized EGI System

1) *System Framework:* Due to the extensive training requirements and large parameter scales of LLMs, significant computational resources are necessary for both training and inference tasks. To effectively implement an EGI system, a straightforward approach is to deploy LLMs in a centralized cloud server, illustrated in Fig. 1(a). This configuration allows the system to fully leverage the powerful computational resources available in the central server.

In the centralized framework, all intelligence-related capabilities reside on the central server, while edge devices serve as tools accessible by the centralized LLM. By analyzing

TABLE I  
OVERVIEW OF REPRESENTATIVE LLM-BASED EGI SYSTEMS ACROSS CENTRALIZED, HYBRID, AND DECENTRALIZED ARCHITECTURES.

Archi.	Ref.	Scenarios	Datasets	Description	Cloud		Edge	
					Model	Tasks	Model	Tasks
Centralized	[3]	Digital twin	-	Use LLM to simulate a population group's arrivals and distribution across stores for optimization.	GPT-3.5, GPT-4	Inference	-	-
	[4]	Robotic	Benchmark dataset	Use LLM to convert high-level task instructions into a multi-robot task plan.	GPT-4, GPT-3.5, Llama-2 (70B), Claude-3-Opus	Inference	-	-
	[1]	Edge network	User request dataset	Use GPT to interpret user intentions and generate code in a cloud-edge-client framework.	GPT-3 (350M, 6.7B, 175B), GPT-3 IT (175B)	Inference, fine-tuning	-	-
	[2]	IoT	-	Use LLM to transform high-level verbal instructions into a control script.	ChatGPT	Inference	-	-
	[5]	Mobile device	-	Propose M4, manage a foundation model serving diverse tasks via lightweight adapters in mobile AI.	RoBERTa, BERT, DistilBERT and so on	Inference, fine-tuning	-	-
Hybrid	[6]	IoT	MixInstruct	Uses a router to assign queries to SLMs or LLMs based on predicted difficulty and desired quality.	GPT-3.5-turbo, Llama-2 (13B)	Inference	FLAN-T5 (800M), Llama-2 (7B,13B)	Inference
	[7]	Edge network	WikiText-2	Propose a collaborative edge computing framework using dynamic programming to optimize inference latency and throughput.	Llama-2 (7B, 13B,70B)	Inference	Llama-2 (7B, 13B,70B)	Inference
	[8]	Edge network	GSM8K, HumanEval, NaturalQuestion	Combine SLMs on edge devices with cloud-based LLMs to reduce costs and optimize task performance.	Llama-2 (7B)	Inference	TinyLlama (1.1B)	Inference
Decentralized	[9]	Edge network	-	Present a visual framework to explore coordination strategies in multi-agent collaboration.	-	-	GPT-4, Mistral (8*7B)	Inference, fine-tuning
	[10]	6G networks	Cornell Movie-Dialogs Corpus dataset	Propose an LLM-enhanced multi-agent system with tailored communication tools to optimize task-solving through collaboration in 6G networks.	-	-	GPT-3.5	Inference
	[11]	Edge network	WikiText, MMLU	Propose an Edge-LLM framework, featuring layer-wise compression, adaptive tuning, and hardware scheduling for optimal performance.	-	-	Llama (7B)	Inference
	[12]	IoT	TDW-House	Introduce CoELA, a multi-agent cooperation framework that leverages LLMs for decentralized control in complex environments.	-	-	GPT-4, Llama-2 (13b)	Inference, fine-tuning

data collected from each edge device, the central server gains insights into various scenarios and tasks. Each edge device provides an API, enabling the LLM to generate code that integrates and utilizes these devices collectively to complete tasks. This approach fosters a scalable and adaptable system, where adding more edge devices enhances the LLM's toolkit, thereby improving its efficiency in managing and executing tasks.

2) *Implementation*: To implement the centralized EGI system within edge networks, LLMs can be deployed on a central cloud or powerful edge servers, such as base stations. For effective collaboration, the APIs of edge devices must be readily accessible to the LLMs. Several studies demonstrate practical implementations of this framework. Yang *et al.* [3] propose an LLM-powered, autonomous agent-based digital twin that employs GPT to simulate customer behaviors and preferences in a mall with multiple stores. In this system, GPT acts as a simulator, queried iteratively to evaluate the arrival patterns and distribution of groups of people, facilitating reinforcement learning-based optimization for adjusting environmental conditions like temperature. Kannan *et al.* [4] introduce the centralized SMART-LLM, which enhances the efficiency of multi-robot task planning through a structured four-stage process: task decomposition, coalition formation,

task allocation, and task execution. This approach ensures comprehensive consideration of the environmental context and the capabilities of each robot involved. In the context of smart homes, the Sasha framework [13] is designed to respond to loosely defined commands, such as “make it cozy.” It generates an action plan in JSON format to achieve specified goals. Similar to SMART-LLM, Sasha also divides the task planning process into four components: clarifying the goal, filtering devices, planning actions, and iterative refinement. Shen *et al.* [1] propose a framework that uses GPT's capabilities in language understanding, planning, and code generation, combined with task-oriented communication and federated learning to coordinate edge AI models and meet diverse user requirements. Furthermore, LLMind [2] utilizes LLMs as a central orchestrator to coordinate domain-specific AI modules and IoT devices, thereby executing complex tasks with improved capabilities, accuracy, and performance that extend beyond the general knowledge of LLMs.

### 3) Discussion:

a) *Advantages*: Among the three proposed architectures, the centralized EGI system is the most cost-effective to construct and scale. Leveraging the high processing power of a central server allows edge devices to function as tools without incurring significant additional power consumption.

This centralized approach enables the server, with its comprehensive view of the network, to make optimal decisions by analyzing global data collected from all edge devices. As a result, the system can efficiently coordinate tasks and resources, enhancing overall performance.

*b) Disadvantages:* However, the centralized EGI system also presents notable disadvantages. As the intelligence core of the edge network, the central server bears a heavy workload in managing all edge devices. The simultaneous transmission of data from multiple devices can lead to network congestion, which ultimately limits scalability. For instance, an hour of video recorded at 1080p can range from approximately 2.25 GB to 3.6 GB. With thousands of cameras deployed in urban environments, the concurrent data transmission could easily overwhelm the network, resulting in data loss and degraded performance.

Additionally, privacy and security pose significant concerns in centralized EGI networks. An attacker only needs to breach the central server to access all data collected from connected edge devices, resulting in a single point of failure that can disrupt the system's functionality, cripple the entire network and compromise critical services.

*c) Feasibility Discussion:* Recent advancements in LLMs have made the development of centralized EGI systems increasingly practical. Numerous major corporations now offer ready-to-use APIs that provide access to these advanced capabilities, making it easier to integrate LLMs into existing edge systems. Upgrading existing edge systems can be as simple as deploying the appropriate APIs.

Furthermore, recent studies on LLM-based agents have demonstrated that employing LLMs as central schedulers can effectively coordinate collaboration among multiple tools [1], [4], [13]. This orchestration enables the completion of complex tasks that previously required human intervention. Alongside these advancements, multi-modal Vision-Language Models (VLMs) have emerged, enabling LLM-based agents to process both visual and textual inputs, thereby expanding their capacity to handle tasks like image and document interpretation for more complex decision-making. Moreover, research into function-calling and tool utilization has shown that LLMs can autonomously select and use external tools as needed. This multi-modality not only enhances the efficiency of client interactions but also allows for a more comprehensive collaboration among diverse tools on a single task. Table II presents a list of API providers and their abilities, illustrating the diverse options available for organizations looking to implement centralized EGI systems.

## B. Hybrid EGI System

*1) System Framework:* As shown in Fig. 1(b), both the central server and edge devices are equipped with language models to enhance intelligence. In a hybrid EGI system, the central server hosts a more advanced and powerful LLM, while the edge devices run Small Language Models (SLMs) due to their limited computational capacity. Recent advancements in SLMs have made such deployments possible. While these device-side SLMs may not rival the capabilities of cloud-based powerhouses like GPT-4 as demonstrated in the results

TABLE II  
LLM API PROVIDERS AND THE SUPPORTED MODALITIES.

Provider	Language	Tool	Vision	Audio
OpenAI <sup>1</sup>	✓	✓	✓	✓
Anthropic <sup>2</sup>	✓	✗	✗	✗
Cohere <sup>3</sup>	✓	✗	✗	✗
Google <sup>4</sup>	✓	✓	✓	✓
Microsoft Azure <sup>5</sup>	✓	✓	✓	✓
Mistral AI <sup>6</sup>	✓	✗	✗	✗

<sup>1</sup> OpenAI: <https://openai.com/api/>

<sup>2</sup> Anthropic: <https://docs.anthropic.com>

<sup>3</sup> Cohere: <https://docs.cohere.com/reference/about>

<sup>4</sup> Google: <https://ai.google/>

<sup>5</sup> Microsoft Azure: <https://learn.microsoft.com/en-us/azure/>

<sup>6</sup> Mistral AI: <https://docs.mistral.ai/api/>

shown in Table III. Integrating SLMs capable of basic GI tasks, such as understanding natural language instructions and answering common-sense questions, into a Mixture of Experts (MoE) framework significantly enhances their functionality. Within this framework, each SLM is fine-tuned as an expert for specific tasks and dynamically selected by a gating mechanism, enabling the efficient execution of more complex environments requiring direct human interaction, such as smart homes, mobile phone assistants, and vehicular systems.

*2) Implementation:* Hybrid EGI systems provide a more flexible implementation, although they complicate the deployment. Ding *et al.* [6] propose a “hybrid LLM” architecture that combines different models. Their approach relies on a quality-aware router, which directs queries to the most cost-effective LLM (large for complex tasks, small for simpler ones) without sacrificing response quality. This work achieves significant cost savings by leveraging the diverse strengths of different-sized LLMs. To enhance edge-cloud cooperation, Yang *et al.* [14] propose EdgeFM, an edge-cloud cooperative system that leverages foundation models (FMs) to enhance the generalization capabilities of on-device deep learning models on resource-limited IoT devices. By selectively uploading data to the cloud for FM querying and dynamically switching models based on data uncertainty and network conditions, EdgeFM significantly improves accuracy and reduces end-to-end latency. Zhang *et al.* [7] propose a collaborative edge computing framework that enables efficient LLM inference by dynamically partitioning the LLM model and distributing it across edge devices and cloud servers. This framework incorporates an adaptive device selection and model partitioning strategy optimized through a dynamic programming algorithm to minimize inference latency and maximize throughput. Hao *et al.* [8] present a dynamic token-level Edge-Cloud collaboration framework for LLMs, utilizing an SLM like TinyLlama on edge devices that interacts with cloud-based LLMs to achieve high-quality performance at reduced cost. Together, these advancements illustrate the potential of hybrid EGI systems to balance computational demands while enhancing performance and efficiency across various applications.

### 3) Discussion:

*a) Advantages:* In the hybrid EGI system, deploying edge devices with SLMs enables low-latency decision-making,

TABLE III  
THE DESCRIPTION OF REPRESENTATIVE SLMs.

Model	Provider	Language	Parameter Size	Context Length	Throughput*		Release Date
					requests/s	tokens/s	
Qwen-1.5-0.5B	Alibaba	Multilingual	0.5B	32k	34.36	14406.54	2024.02
Tinyllama	jzhang38	English	1.1B	2K	31.64	15226.31	2024.04
stablelm-2-1_6b-chat	Stability-AI	English	1.6B	8K	23.66	9929.77	2024.04
Qwen-1.5-1.8B	Alibaba	Multilingual	1.8B	8K	22.84	9575.21	2024.02
Gemma	Google	Multilingual	2B	8K	24.84	10945.68	2023.09
phi-2	Microsoft	English	2.7B	2K	13.87	6305.34	2023.12
stablelm-3b-4e1t	Stability-AI	English	3B	4K	13.98	6293.27	2023.09
phi-3-mini	Microsoft	English	3.8B	4K&128K	9.72	4675.83	2024.04
Qwen-1.5-4B	Alibaba	Multilingual	4B	32K	12.04	5048.83	2024.02
Qwen-1.5-7B	Alibaba	Multilingual	7B	8K	4.66	1953.71	2024.02
Mistral-7B	Mistral AI	English	7B	8K	10.18	4809.72	2023.09
Llama-2-7b	Meta	Multilingual	7B	4K	4.74	2280.35	2023.08
gemma-7b	Google	Multilingual	7B	8K	2.71	1192.78	2024.02
phi-3-small	Microsoft	English	7B	8K	10.82	4483.98	2024.04
Llama-3-8B	Meta	Multilingual	8B	8K	10.87	4496.12	2024.04

\* Throughput measured in one 4090 GPU with 24 GB of memory. The max context window is 4096 for all models.

allowing quick responses to local events without relying on central server processing. This immediacy is crucial for real-time services, such as smart homes and autonomous vehicles. Additionally, the system enhances load balancing by distributing tasks between edge devices and the central server, preventing any single device or server from becoming overburdened, and enhancing overall system efficiency and reliability. The architecture's flexibility is evident in its ability to handle diverse tasks with varying constraints. Edge devices can preprocess and determine which information is essential for transmission. By using SLMs for semantic communication rather than transmitting raw data, the system conserves network resources and reduces communication overhead, further optimizing performance.

*b) Disadvantages:* However, the hybrid EGI system also comes with notable disadvantages. Deployment costs are higher due to the requirement to set up both a central LLM and edge SLMs. Maintenance costs can be significant as well, driven by the complexity involved in updating, securing, and managing both the edge devices and the central server. Furthermore, since SLMs cannot match the processing and decision-making capabilities of LLMs, efficient task scheduling and resource allocation become essential to balance QoE with efficiency.

*c) Feasibility Discussion:* Recent advancements in SLMs have paved the way for deploying these sophisticated models on edge devices. Research has demonstrated that models with a relatively small number of parameters, such as 1.1 billion, can still possess a substantial level of GI. These models can understand human language, engage in fluent communication, and answer questions that do not necessitate complex reasoning or in-depth world knowledge. In addition to the remarkable progress in SLM development, advancements in model optimization techniques have facilitated their deployment on edge devices. Techniques such as quantization and efficient inference frameworks, which incorporate optimizations like

kernel fusion, have been crucial. These technological innovations make deploying SLMs on consumer-grade hardware feasible, as exemplified by smartphones equipped with a Snapdragon 888 processor and 8GB of RAM, achieving impressive performance levels.

### C. Decentralized EGI System

*1) System Framework:* In the decentralized EGI framework, each edge device has its own SLM for autonomous decision-making, enabling local data processing and task execution. As depicted in Fig. 1(c), this decentralized architecture significantly reduces the need for human intervention in managing device interactions, thereby increasing the efficiency of decision-making processes. Furthermore, while these devices operate independently, their ability to collaborate with others is essential in a decentralized EGI architecture. This collaborative capability allows them to share insights, improve overall system performance, and address complex tasks that require collective intelligence.

*2) Implementation:* Although research on decentralized EGI systems is still in its early stages, several studies have demonstrated that by combining the knowledge and processing capabilities of multiple intelligent agents, decentralized EGI systems can achieve better collective intelligence performance than single agents. To coordinate cooperation among multiple agents, Zhang *et al.* [12] introduce the Cooperative Embodied Language Agent (CoELA), a cognitive-inspired modular framework that leverages the reasoning, language comprehension, and text generation abilities of LLMs in decentralized control scenarios with costly communication. Experiments demonstrate that CoELA, driven by GPT-4, outperforms traditional planning-based methods and can be fine-tuned for improved performance using collected data, thereby enhancing human-agent interaction through natural language communication. Chan *et al.* [15] introduce ChatEval, a multi-agent framework that utilizes diverse communication strategies

and unique agent personas to collaboratively evaluate text, aligning more closely with human preferences than single-agent approaches. Drawing inspiration from collective intelligence and cognitive synergy, ChatEval enhances evaluation accuracy by incorporating multiple perspectives and exhibiting human-like behavior in interactive natural language dialogue. Jiang *et al.* [10] propose a multi-agent system enhanced with LLMs that comprises three components: Multi-agent Data Retrieval (MDR), Multi-agent Collaborative Planning (MCP), and Multi-agent Evaluation and Reflection (MER). These components work together to refine communication knowledge, generate feasible solutions, and evaluate and improve current solutions for communication-related tasks using natural language. The authors also address concerns regarding resource constraints at the edge and the interaction delay associated with LLMs. Yu *et al.* [11] develop a framework to facilitate the efficient adaptation of LLMs on edge devices, addressing the challenges of high computation and memory demands. The framework incorporates a layer-wise unified compression technique to optimize computation through adaptive pruning and quantization, coupled with an adaptive layer tuning and voting mechanism that reduces memory usage by curtailing the depth of backpropagation. Additionally, a dedicated hardware scheduling strategy efficiently manages the irregular computation patterns that arise from these optimizations.

### 3) Discussion:

*a) Advantages:* In the decentralized intelligence system, intelligent edge devices can communicate directly with each other, allowing for collaborative task execution that reduces manual costs. In environments with extremely poor communication conditions, such as oceans and disaster areas, distributed collaborative intelligence can adapt to various scenarios and complete tasks effectively. This system is robust avoiding a single point of failure compared to the centralized and hybrid intelligence systems. Processing data locally minimizes the risk of privacy breaches associated with data transit or central storage. Additionally, the decentralized architecture enhances scalability and improves the collective intelligence of the network. Real-time responses are facilitated without the need to communicate with a central server, optimizing network efficiency through direct device-to-device communication, conserving bandwidth, and reducing the overall load. This structure is particularly suitable for applications requiring speed, scalability, robustness, and data privacy. Moreover, unlike traditional decentralized edge systems that rely on minimal collaboration or predefined rules, EGI systems equipped with SLMs possess individual decision-making capabilities, enabling them to actively collaborate and adapt to the dynamic real-world environment, an essential feature for edge computing.

*b) Disadvantages:* Despite its advantages, the decentralized EGI system also presents several challenges. The frequent interactions between individual edge devices increase the complexity of the trust mechanism, making it difficult to enforce consistent policies across the network due to its decentralized nature. Additionally, the computational capacities of each device's SLM may be limited, hindering their ability to handle complex tasks compared to a centralized LLM. There

is also significant coordination overhead, requiring advanced and complex protocols to manage peer-to-peer interactions and ensure efficient network performance. Given the computational power needed to run even an SLM, upgrading the hardware of edge devices becomes necessary, potentially leading to higher costs.

*c) Feasibility Discussion:* The NLP community has begun to explore the use of multiple LLMs to simulate group intelligence behaviors. This would enable collaboration among multiple LLM-based agents that can work together or engage in discussions. However, this area of research is still in its infancy, presenting a novel topic for further investigation and development.

## V. FUTURE DIRECTIONS

In this section, we outline several key future directions for implementing LLM-empowered EGI to fully unlock the potential of LLMs in resource-constrained edge environments.

### A. Efficient SLM Deployment on Resource-limited Edge Devices

Deploying efficient general intelligence on resource-limited edge devices for EGI is challenging due to the substantial RAM and computational resources required by LLMs and SLMs. Solutions can be explored from two perspectives: algorithms and hardware. Algorithmically, reducing memory consumption and computational costs through quantization—lowering the precision of model data—shows promise, but even optimized models still require at least 500MB of RAM and high-end CPUs, like the Snapdragon 888, for effective inference. While low-bit-rate inference is feasible, training models similarly remain difficult, limiting on-device refinement. Split learning/inference, which distributes tasks between edge devices and central servers, offers another approach, though it depends on optimal model partitioning and secure data transmission. On the hardware side, boosting processing power and memory within the constraints of edge devices is critical. AI-specific processors, such as GPUs, FPGAs, and ASICs (e.g., Google's TPU), alongside system-level optimizations like distributed inference across device clusters, offer potential solutions. Addressing these challenges is key to enabling the broader adoption and efficient use of LLM-empowered EGI systems in edge computing.

### B. Optimizing Latency of Providing LLM-enhanced EGI Services

Response latency, encompassing both inference and communication delays, serves as a pivotal determinant of Quality of Service (QoS) in LLM-augmented EGI systems. Inference latency is driven by the hardware's processing capabilities and the efficiency of inference algorithms. Improvements require advancements in computational methods like model optimization and quantization, alongside hardware innovations such as AI accelerators and optimized processors to reduce delays and enhance throughput. Communication latency stems from data transmission delays over the internet. While text-based LLMs



handle small data packages, multimodal LLMs, which process larger inputs like images and videos, face greater challenges. For instance, a one-hour 4K video can range from 9GB to 27GB, straining 5G networks with uplink speeds of around 50Mbps. Addressing this requires efficient data compression and transmission techniques, with future 6G advancements potentially easing these limitations and enabling more practical multimodal LLM-based EGI services.

### C. Adapting LLMs for Domain-Specific Applications in Edge Networks

Current LLMs are trained on broad, general-domain data, while edge systems often require domain-specific knowledge. To meet this need, LLMs must be adapted through continued pre-training or fine-tuning on specialized datasets, enhancing their ability to address niche queries. This adaptation is essential for edge computing, where rapid and accurate responses are critical. Additionally, managing time-varying network conditions to avoid negative impacts caused by network fluctuations. Solutions like dynamically adjusting model complexity and offloading heavier tasks to the cloud help maintain low-latency responses while ensuring essential functions remain on edge devices.

### D. Security for LLM-enhanced EGI Services

Deploying LLMs on edge networks poses significant security challenges that must be addressed to protect user privacy and ensure data integrity. Transmitting sensitive data to external servers in cloud-based models risks data exposure, making advanced encryption methods like differential privacy, homomorphic encryption, and zero-knowledge proofs essential for safeguarding information. The decentralized EGI architecture eliminates the dependence on a central server, thereby reducing the risk of data breaches. However, securing data transmission channels remains a critical challenge. Furthermore, maintaining LLM integrity poses significant challenges due to threats such as model tampering, poisoning, and adversarial attacks, where adversaries manipulate the model or its training data to produce erroneous outputs. Additionally, updating models across distributed networks introduces further complexity, as updates may be intercepted or altered by malicious actors. Addressing these security concerns is essential for the reliable construction of EGI systems.

### E. Collaborative LLM-enhanced Edge Systems

LLM-empowered edge devices bring a remarkable ability to understand and adapt to dynamic environments, making them valuable in various applications. However, leveraging multiple LLM-based agents to collaborate and complete tasks [11] introduces unique challenges, especially in resource-constrained extreme environments such as mountainous regions, earthquake zones, and oceans. These scenarios face issues like limited connectivity, low bandwidth, and unreliable power supplies, which complicate the seamless interaction among agents. Moreover, the unpredictable behaviors of LLM-based agents add another layer of complexity to cooperate. Coordinating tasks efficiently requires not only robust communication

protocols and resource management but also a system capable of handling these unpredictable responses. Overcoming these challenges is crucial for deploying EGI systems to effectively support remote tasks like disaster recovery and environmental exploration.

## VI. CONCLUSION

In this survey, we explored the evolution of LLM-enhanced EGI and differentiated it from traditional EI. The LLM-enhanced EGI systems are categorized into three conceptual designs, namely centralized, hybrid, and decentralized, each reflecting distinct architectures and operational strategies. For each system, this survey summarized the framework designs, reviewed representative research, and discussed the advantages, disadvantages, and feasibility. Moreover, we compared various SLMs based on their characteristics and throughput, offering valuable insights for orchestrating EGI systems. Looking ahead, we summarized promising future directions for EGI in terms of resource efficiency, service optimization, domain adaptation, interaction security, and multi-agent collaboration. This survey presents a comprehensive vision of EGI and outlines key future directions that will contribute to its ongoing advancement.

## REFERENCES

- [1] Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, J. Zhang, and K. B. Letaief, "Large language models empowered autonomous edge ai for connected intelligence," *IEEE Communications Magazine*, 2024.
- [2] H. Cui, Y. Du, Q. Yang, Y. Shao, and S. C. Liew, "LLMind: Orchestrating ai and iot with LLM for complex task execution," *IEEE Communications Magazine*, pp. 1–7, 2024.
- [3] H. Yang, M. Siew, and C. Joe-Wong, "An LLM-based digital twin for optimizing human-in-the loop systems," in *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*, pp. 26–31, 2024.
- [4] S. S. Kannan, V. L. Venkatesh, and B.-C. Min, "Smart-llm: Smart multi-agent robot task planning using large language models," *arXiv preprint arXiv:2309.10062*, 2023.
- [5] J. Yuan, C. Yang, D. Cai, S. Wang, X. Yuan, Z. Zhang, X. Li, D. Zhang, H. Mei, X. Jia, et al., "Mobile foundation model as firmware," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pp. 279–295, 2024.
- [6] D. Ding, A. Mallick, C. Wang, R. Sim, S. Mukherjee, V. Rühle, L. V. S. Lakshmanan, and A. H. Awadallah, "Hybrid LLM: Cost-efficient and quality-aware query routing," in *The Twelfth International Conference on Learning Representations*, 2024.
- [7] M. Zhang, J. Cao, X. Shen, and Z. Cui, "Edgeshard: Efficient LLM inference via collaborative edge computing," *arXiv preprint arXiv:2405.14371*, 2024.
- [8] Z. Hao, H. Jiang, S. Jiang, J. Ren, and T. Cao, "Hybrid slm and LLM for edge-cloud collaborative inference," in *Proceedings of the Workshop on Edge and Mobile Foundation Models*, pp. 36–41, 2024.
- [9] B. Pan, J. Lu, K. Wang, L. Zheng, Z. Wen, Y. Feng, M. Zhu, and W. Chen, "Agentcoord: Visually exploring coordination strategy for LLM-based multi-agent collaboration," *arXiv preprint arXiv:2404.11943*, 2024.
- [10] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6g communications," *IEEE Wireless Communications*, pp. 1–8, 2024.
- [11] Z. Yu, Z. Wang, Y. Li, H. You, R. Gao, X. Zhou, S. R. Bommur, Y. K. Zhao, and Y. C. Lin, "Edge-LLM: Enabling efficient large language model adaptation on edge devices via layerwise unified compression and adaptive layer tuning & voting," in *Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC '24)*, 2024.
- [12] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, "Building cooperative embodied agents modularly with large language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.



- [13] E. King, H. Yu, S. Lee, and C. Julien, “Sasha: creative goal-oriented reasoning in smart homes with large language models,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–38, 2024.
- [14] B. Yang, L. He, N. Ling, Z. Yan, G. Xing, X. Shuai, X. Ren, and X. Jiang, “Edgefm: Leveraging foundation model for open-set learning on the edge,” in *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, SenSys ’23, (New York, NY, USA), p. 111–124, Association for Computing Machinery, 2024.
- [15] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, “Chateval: Towards better LLM-based evaluators through multi-agent debate,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.