npj | wireless technology

Perspective

https://doi.org/10.1038/s44459-025-00015-w

Ubiquitous intelligence via wireless network-driven LLMs evolution

Check for updates

Xingkun Yin, Feiran You, Hongyang Du ✉ & Kaibin Huang

We introduce *ubiquitous intelligence* as a paradigm where Large Language Models (LLMs) evolve within wireless network-driven ecosystems. Unlike static model deployments, this approach enables scalable and continuous intelligence ascension through coordination between networks and LLMs. Wireless networks support system-orchestrated lifelong learning, while LLMs drive the next-generation network development that is more adaptive and responsive. This co-evolution highlights a shift toward self-improving systems, sustaining capability growth across diverse and resource-constrained environments.

Large Language Models (LLMs) have quickly expanded from their original applications in machine translation and summarization to a wide spectrum of complex generation tasks, including code, graphics, and video[1]. Advances such as ultra-long context processing, multimodal integration, and frameworks like Retrieval-Augmented Generation (RAG)[2] have further pushed LLMs into domains such as law and medicine, traditionally relying on human expertise. Over time, LLMs have evolved from passive decision-making assistants to active participants in end-to-end processes, while the emerging Artificial Intelligence (AI) agent[3] paradigm further extends its capabilities to autonomous reasoning, planning, and execution.

Despite these advances, most LLMs still operate as cloud-centric models, relying on large clusters for inference and periodic offline retraining[4]. This architecture delivers scale but struggles to meet the growing demands for low latency, strong privacy, and adaptive personalization. Wireless networks, which connect billions of heterogeneous edge devices, offer a promising alternative. By enabling distributed inference and continual learning closer to data sources, they provide the foundation for more responsive, private, and personalized LLMs deployment. Such a transition is technically feasible and increasingly necessary to sustain the next generation of large-scale intelligence.

Large Language Models

LLMs have emerged as a primary expression of machine intelligence, demonstrating the ability to generalize across diverse tasks. Their effectiveness relies on large-scale training and inference pipelines, which are predominantly deployed in centralized data centers[4,5]. These infrastructures integrate high-performance accelerators, low-latency interconnects, and deep memory hierarchies to support large-batch optimization and high-throughput inference[6]. Leading commercial platforms, including OpenAI's GPT models on Microsoft Azure GPU clusters[7], Google's Gemini and PaLM on TPU-based infrastructures[8], and DeepSeek's multi-GPU distributed framework[9], leverage this architecture for scalable, consistent training and inference. Services, such as NVIDIA's DGX Cloud, further extend these capabilities via cloud-hosted multi-node inference, supporting enterprise-scale workloads[10]. This cloud-centric structure offers key advantages: on-demand scalability, centralized management, and streamlined maintenance, making it the standard for contemporary foundation models.

However, as LLMs become increasingly integrated into daily life, the limitations of the cloud-centric paradigm are growing more apparent. Development priorities now emphasize low latency, strong privacy, and adaptive personalization, yet transmitting data to and from remote data centers conflicts with the demands of time-sensitive applications[11,12]. Effective personalization further requires continuous, context-aware learning from the user environment, but centralized processing of such data raises significant privacy risks[13]. These challenges highlight a widening gap between centralized infrastructures and user expectations, underscoring the need to shift LLM inference closer to data sources through wireless networks.

Wireless network

Wireless networks form a crucial interface between cloud-centric AI and context-aware inference at the edge[11,14,15]. Modern infrastructures, characterized by dense connectivity, high throughput, and low latency, offer a flexible foundation for collaborative training and synchronized model updates across heterogeneous devices[16–18]. As LLMs are deployed in increasingly diverse scenarios, conventional reliance on centralized data centers imposes significant limitations. Transmission delays hinder responsiveness in time-sensitive applications[19], while personalized services require handling user-specific data previously ignored[20]. These demands exceed the capabilities of traditional network pipelines.

To meet these emerging challenges, wireless components such as Base Stations (BSs), access points, and user devices must evolve from passive relays into active participants in distributed reasoning[21–23]. Through Device-to-Device (D2D) links, adaptive edge caching, and opportunistic spectrum access, network nodes can exchange intermediate computations, propagate refined models, and adapt decision policies in real time. This architectural

Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong SAR, China. ✉e-mail: duhy@hku.hk

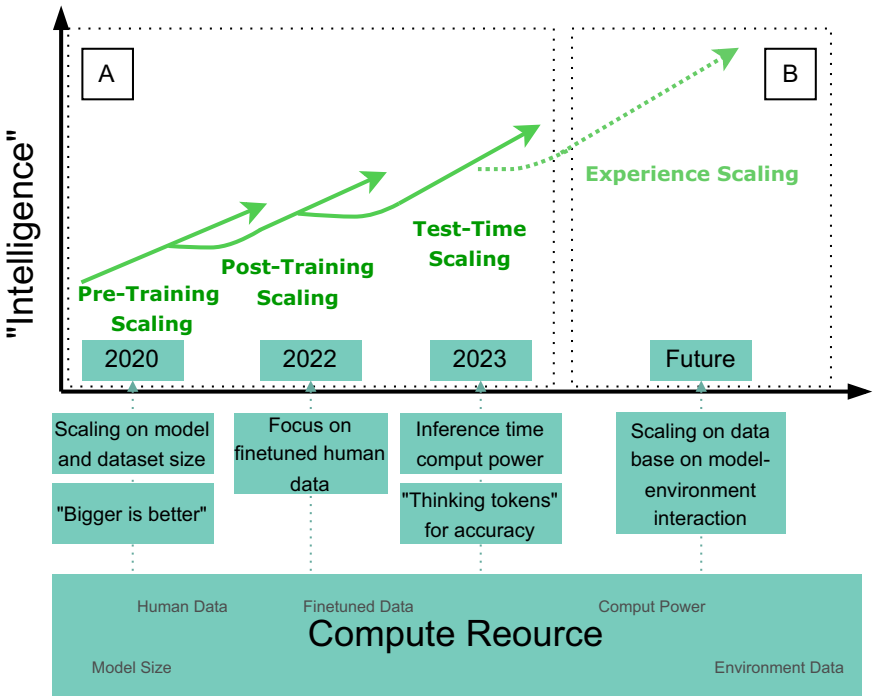

Fig. 1 | Scaling laws. Scaling in terms of parameter size, training data, and compute time can be conceptualized as scaling the amount of compute resources. LLMs have progressed through three stages of scaling laws, culminating in their current advanced stage (Part **A**). We propose a new scaling paradigm, experience scaling (Part **B**), within the concept of *ubiquitous intelligence*, aiming to push scaling to a new dimension and further enhance LLMs' capabilities.

transformation is central to the vision of Sixth-Generation (6G) networks[24], which integrate communication, computation, and learning into a unified, human-centric system. By embedding intelligence directly within the network fabric, future systems aim to support personalized, secure, and ultra-reliable services under highly dynamic conditions.

Motivation

LLMs and wireless networks have driven major advances in intelligence and connectivity, but both now face fundamental limits. Addressing these constraints is key to sustaining large-scale evolution.

Developments and challenges of large language models

The progression of LLMs has been driven by successive scaling paradigms (Fig. 1). Introduced in 2020, pre-training scaling law[25], established a power-law relationship between performance and three factors: **model size, training data, and compute power**, guided the development of GPT-3[26] and later models. Post-training scaling focuses on improving alignment and task specialization through supervised fine-tuning and Reinforcement Learning from Human Feedback (RLHF)[27,28], allowing capability gains without changing the core architecture. More recently, test-time scaling emphasizes allocating additional compute during inference, enabling step-by-step reasoning, exploring multiple solution paths, and reducing hallucinations, thereby improving overall reliability and reasoning quality[29,30]. Focusing on the concept of experience, LLMs collect and distill indigenous knowledge while, under experience scaling, they autonomously explore and absorb vast amounts of information from their environment, potentially uncovering knowledge untouched by humans.

However, current scaling laws exhibit diminishing returns: GPT-4's greatly increased scale over GPT-3 yielded diminishing per-token gains, while training data demand now surpasses human generation, detrimental for post-training methods like RLHF[28]. As shown in Fig. 1.B, LLMs performance is nearing its ceiling under this paradigm, necessitating a new scaling dimension for further advancement.

Evolution and limitations of wireless networks

Wireless networks have evolved from voice-only systems to intelligent infrastructures that support data-driven and mission-critical applications[31]. From 3G's mobile internet to 5G's broadband and massive connectivity, each generation has improved capacity and latency. Ongoing 6G research seeks to integrate sensing, communication, and intelligence into a unified architecture[17,32,33]. Key advances such as Ultra-Reliable Low-Latency Communication (URLLC)[34], D2D communication[35], and context-aware content delivery[36] enhance responsiveness and localization, while adaptive spectrum and resource management[37] improve efficiency under dynamic conditions.

However, challenges persist as LLM-based services require low latency, privacy, and personalization, which cloud-based designs often ignores[38]. Transmitting data to remote servers introduces delay and privacy risks[39–42], limiting real-time applications. Meanwhile, networks themselves face growing complexity. Device heterogeneity, mobility, and spectrum variation reduce the effectiveness of centralized control, causing resource contention and unstable performance[37,43]. Scaling distributed AI workloads further amplifies these limitations. Importantly, these issues are not only bottlenecks for deploying LLMs but also opportunities for co-optimization. LLMs can contribute to the network's optimization, enabling context-aware scheduling, semantic compression, and adaptive coordination. Fully unlocking this potential, however, requires architectural alignment and real-time integration between learning models and communication protocols.

Motivation for coevolution between LLMs and wireless networks

The advances and limitations of LLMs and wireless networks highlight the opportunity for mutual reinforcement. The co-evolution of LLMs with wireless networks involves leveraging edge computing and decentralized learning[14,44,45], where models are deployed closer to users through local processing on edge devices or base stations, allowing LLMs to continuously learning without relying on centralized cloud servers.

This perspective motivates the vision of ***ubiquitous intelligence* via AI-enabled networks and network-enabled AI**, where intelligence is embedded throughout the wireless network infrastructure and adaptively distributed across heterogeneous devices to support real-time, context-aware, and personalized services at scale, as illustrated in Fig. 2.

Synergistic evolution of LLMs and wireless networks

Achieving *ubiquitous intelligence* requires tight integration between LLMs and network infrastructure. This section explores their co-evolution and mutual reinforcement in enabling pervasive cognition.

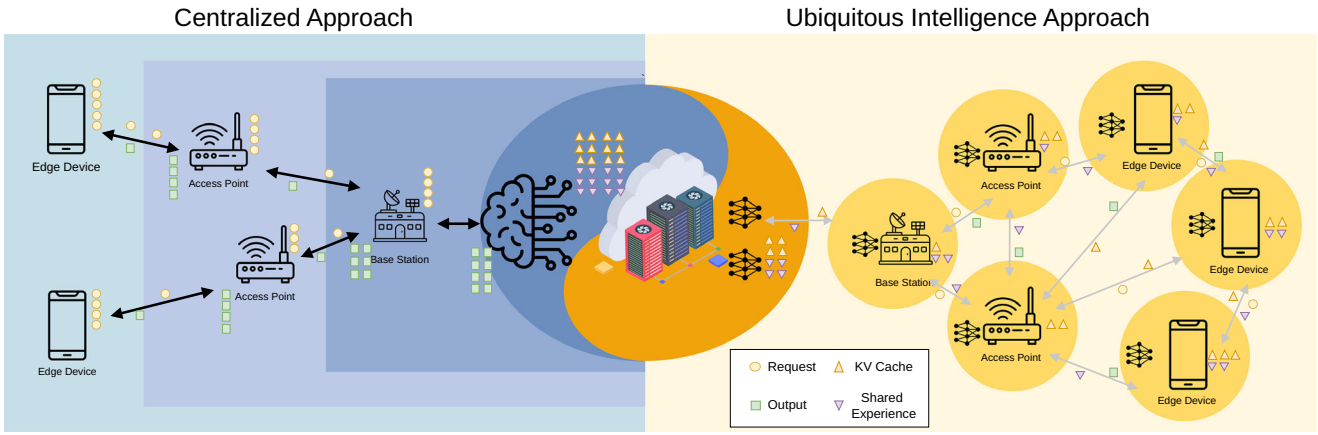

Fig. 2 | Comparison between conventional cloud-centric approach and *ubiquitous intelligence* approach. In the traditional approach, all data and computation are routed through centralized cloud infrastructure, leading to latency, privacy risks, and limited personalization. The *ubiquitous intelligence* framework leverages BSs, access points, and edge devices as active intelligence nodes, enabling local inference, cooperative learning, and D2D knowledge sharing. The distributed structure supports low-latency adoption and enhances resilience and personalization in dynamic wireless environments.

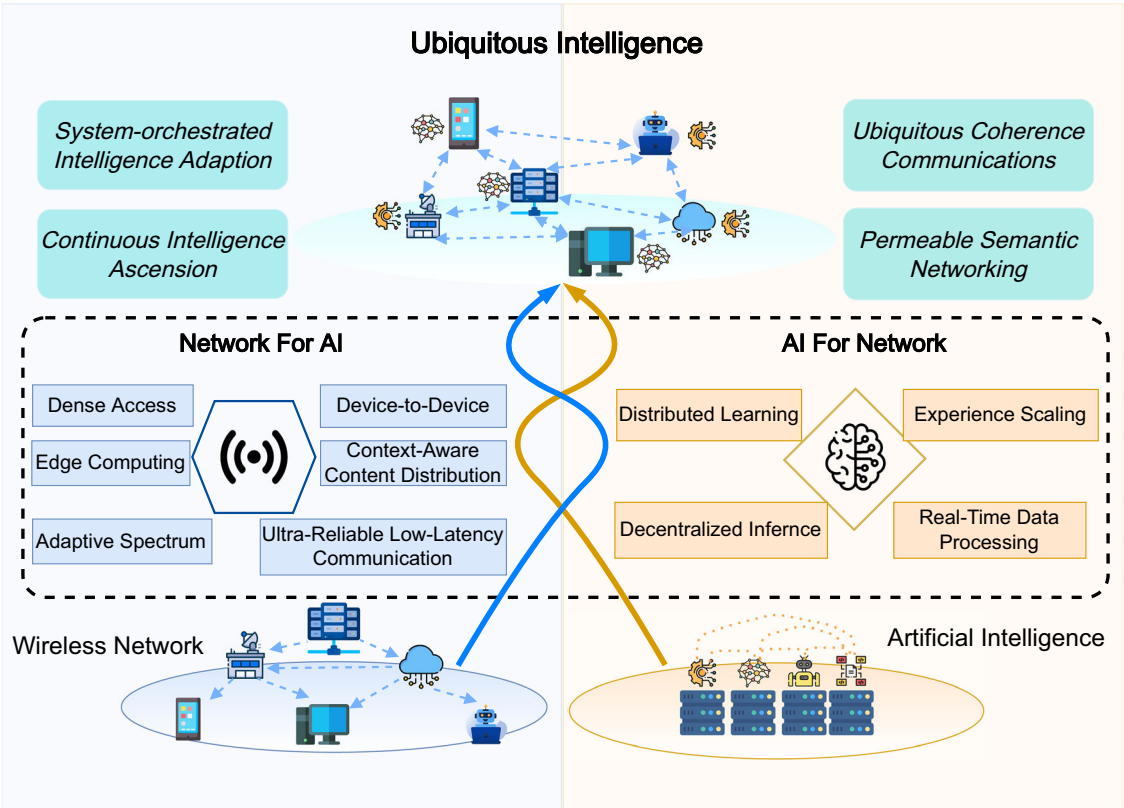

Fig. 3 | Coevolution of wireless network and AI. Adaptive, continuously improving AIs are supported through wireless network-assisted experience scaling and AI-powered, context-aware, densely connected networks, forming the foundation of *Ubiquitous Intelligence*.

Network-empowered LLMs evolution

Although scaling laws have driven significant advances in LLMs, progress is limited as human-generated data can no longer keep pace with LLMs consumption[46], signaling the need for a new paradigm. Current advances remain primarily focused on imitating humans, consuming human-created data. While LLMs are increasingly capable of performing existing tasks on behalf of humans, they still lack the capacity to deliver breakthroughs in scientific and technological domains where no human-generated data exists.

A new scaling paradigm is required to bypass the bottleneck of human-generated data by enabling LLMs to collect and create data directly from their environment. Advancements in AIs and wireless networks (Fig. 3) enable autonomous knowledge acquisition through environmental interaction, extending model understanding beyond human-derived domains. This continuous, multimodal experience accumulation supports lifelong learning. When shared across wirelessly interconnected nodes, these heterogeneous data streams foster collective intelligence at scale, unlocking access to massive decentralized environmental data and paving the way for intelligence that transcends the limits of human experience.

LLM-enhanced cognitive networking

The advancement of LLMs now hinges not only on computational scaling but also on their integration with communication infrastructures.

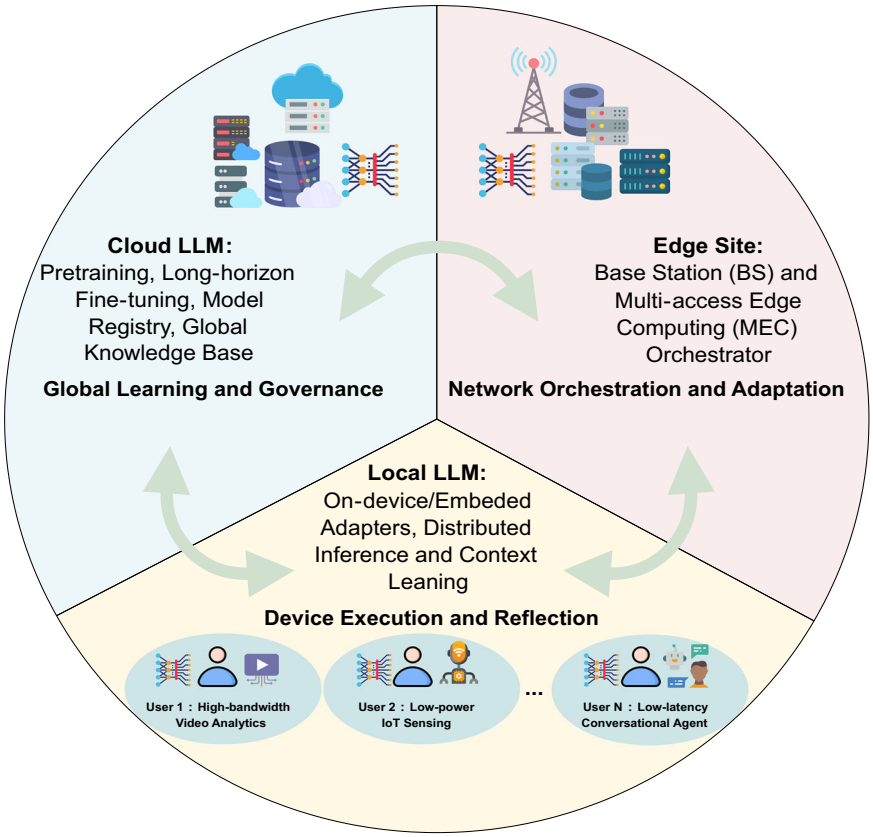

Fig. 4 | Collaboration of intelligence under ubiquitous intelligence. LLMs evolve into adaptive, networked ecosystems integrated with wireless edge environments, enabling scalable, resilient, and continuously improving intelligent services.

Wireless networks, once passive conduits for data, are increasingly envisioned as distributed platforms where models and network elements co-evolve in a tightly integrated ecosystem. This transformation enables ubiquitous intelligence, where computation and communication are jointly executed across heterogeneous, spatially distributed devices. The fusion of dense radio access, Multi-access Edge Computing (MEC), diverse terminals, and latency-aware scheduling provides a dynamic substrate for in-network learning[47]. Network nodes become intelligent agents, capable of contextual reasoning and localized model refinement. Devices actively participate in knowledge exchange and environmental adaptation, forming a collaborative learning system embedded within the communication fabric. Modern wireless architectures account for user behavior and localized demands[17,36,38], guiding the selective delivery of models or data to where they are most needed. This localized responsiveness is central to the realization of *ubiquitous intelligence*.

Figure 4 illustrates the framework of *ubiquitous intelligence* with the deployment of cloud LLMs, edge sites, and local LLMs. Each layer participates in distributed inference, model refinement, and real-time collaboration. The close integration between radio access points and edge computation platforms[48] creates a tightly coupled infrastructure where models operate near the data source. This structure supports high availability, rapid adaptation, and robust performance under dynamic wireless conditions. As LLMs increasingly depend on the environments they inhabit, the wireless network becomes a partner in learning and reasoning. This shift marks a departure from isolated model scaling toward a co-evolutionary paradigm, where intelligence arises through continuous interaction between the models and the underlying wireless infrastructure.

Ubiquitous intelligence

The trajectories of AIs and networks are increasingly converging[49], as demonstrated in Fig. 3. Their integration marks a transition from incremental advances in each domain to a co-evolutionary paradigm in which intelligence and connectivity are inseparable. This convergence provides the conceptual and technical foundation for *ubiquitous intelligence*, where adaptive learning and resilient communication form a unified, continuously evolving ecosystem. The shift towards *ubiquitous intelligence* is grounded in four core principles:

- **Continuous Intelligence Ascension**. Continuous intelligence ascension denotes the sustained enhancement of LLMs capabilities through interaction with dynamic environments. Unlike static deployments, LLMs refine reasoning, adaptability, and autonomy from live experiences, while wireless networks interconnect heterogeneous sensing devices and edge nodes to provide the bandwidth and low latency needed for real-time experience exchange. This integration transforms learning into a distributed, lifelong process, enabling intelligence to scale continuously alongside wireless network development.
- **System-Orchestrated Intelligence Adaption**. Driven by wireless network breakthroughs in D2D and context-aware content distribution, *System-Orchestrated Intelligence Adaptation* supports heterogeneity in the frameworks, refines LLMs selectively by applying only the relevant subsets of experience according to each LLMs' size and function, thereby avoiding redundant updates and enhancing overall system efficiency. During inference, tasks are allocated to the most suitable LLMs based on task characteristics and resource availability, optimizing both performance and utilization. Simple tasks with high time sensitivity are assigned to smaller models, while more complex tasks are allocated to larger models. Diverse data sources and compute resources can be further integrated through the Model Context Protocol (MCP)[50], which bridges communication gaps by providing a unified API access for heterogeneous LLMs. This principle embodies effective coordination and adaptive deployment of intelligence within distributed environments.
- **Permeable Semantic Networking**. AI empowers networks with semantic awareness, allowing systems to exchange not only raw data but also the underlying meaning contained in messages and services. LLMs extend this semantic capability through their pre-trained alignment with human intent, enabling interpretation, generation, and abstraction of meaning across modalities. When integrated into

communication and computation, LLMs enable semantic-level information exchange, transforming the network from a passive transport medium into an adaptive and context-aware infrastructure.
- **Ubiquitous Coherence Communications**. LLM's reasoning and generation convert disordered information flows into structured and refined knowledge. Networks under this paradigm evolve from chaotic to organized, from redundant to essential, and from diffuse to convergent, thereby sustaining coherence in distributed communication environments. The integration of these capabilities ensures that ubiquitous connectivity is matched with ubiquitous intelligence, establishing networks as adaptive ecosystems for resilient, large-scale intelligent services.

New opportunities

As LLMs evolve from static inference engines to dynamic cognitive systems, new opportunities emerge for continual refinement beyond initial training. This section examines how interaction, feedback, and decentralized adaptation within wireless networks support this ongoing evolution in real-world environments.

Wireless network empowered LLMs opportunities

Experience Scaling. While human-generated data remains central to current LLM paradigms, its scalability is limited. In contrast, vast amounts of data from multi-agent systems engaging with diverse environments remain largely untapped. System-wide gathering, processing, and leveraging such interactive experiences represent a critical direction supporting *continuous intelligence ascension* for the next evolutionary era of LLMs scaling.

Edge-cloud network-aided collaborative reasoning. Inference begins at the user terminal, where local computations often take place under constrained resources. Guided by the principle of *system-orchestrated intelligence adaptation*, clusters use low-latency interconnects to map heterogeneous inference tasks to suitable devices, enabling collaborative execution under varying workloads and hardware constraints. This first layer of intelligence leverages local context and device status to make fine-grained task decisions. Such localized reasoning forms the entry point of the system's adaptive intelligence pipeline, balancing response immediacy with offloading potential through wireless networks.

Distributed cached knowledge management and propagation. D2D communication enables peer-to-peer information exchange, distributing reasoning tasks and intermediate knowledge while reducing reliance on centralized infrastructure[51]. This localized cooperation enhances decentralization and adaptability, supporting the principle of *system-orchestrated intelligence adaptation* across heterogeneous environments. When a task enters the framework, the D2D-powered system orchestration allocates the task to the most appropriate compute unit. The computational cache related to the task is then sent directly to the compute unit. The distributed knowledge management thus refines global models and reduces redundancy through intelligent allocation[40,52]. Efficiency is further improved through edge caching, predictive prefetching, and opportunistic D2D clustering, enabling robust, low-latency cross-user knowledge management[53].

Spectrum-aware adaptation for distributed learning. Maintaining performance for resource scheduling in a shared and volatile spectral environment. Spectrum-aware strategies enhance learning stability by embedding real-time channel sensing into the update and synchronization protocols[54]. Devices can regulate their transmission power and update cadence based on interference levels and spectrum availability. These mechanisms play a pivotal role in achieving *system-orchestrated intelligence adaptation* via managing contention and preserving throughput, especially in dense deployments where spectral conditions fluctuate rapidly. Such spectral intelligence contributes directly to the robustness and efficiency of model dissemination.

LLMs enhanced wireless network opportunities

Adaptive task offloading across edge devices. Once local decisions are made, the LLMs edge devices dynamically determine whether to retain tasks locally or offload them with experience to more capable edge nodes[55]. Nearby servers or peer devices can assume responsibility for intensive post-inference workloads such as model refinement or skill module execution. LLMs assisted task offloading policies account for communication quality, processing load, and power availability[56], while experience redistribution strategies primarily focus on bandwidth. This adaptive balance between local and distributed computation not only reduces end-to-end latency but also mitigates energy consumption on user devices, ensuring continuity of reasoning under mobility and hardware heterogeneity. Additionally, dynamic allocation of bandwidth and spectrum resources[37] allows distributed learning and inference to remain reliable under fluctuating interference and load conditions. Through continuous adaptation to changing environments, wireless network infrastructures sustain the robustness and responsiveness that realize *permeable semantic networking* at scale.

Context-aware scheduling of network resources. To support seamless task distribution, wireless networks must ensure reliable and efficient delivery of collective LLMs updates. LLMs empowered resource scheduling mechanisms allocate bandwidth, compute cycles, and transmission windows in real time, guided by user behavior, link conditions, and service-level requirements[57]. The LLMs scheduler system continuously refines allocations based on updated channel state information and application context, thereby reducing delivery latency and improving system responsiveness[58]. This layer of context-awareness ensures consistent adaptation and facilitates knowledge sharing across users, even in dynamic and congested wireless environments, supporting *permeable semantic networking*.

Hierarchical orchestration across network tiers. At the core of the wireless intelligence system lies a coordinated architecture that spans multiple layers of the infrastructure. LLMs across small cells, macro base stations, and cloud servers collaborate to manage the global distribution of updates and computational resources[59] to achieve the goal of *permeable semantic networking*. The LLMs orchestration framework determines which components should be processed locally, cached regionally, or distributed globally, based on topology, load conditions, and service priorities. This hierarchical model alleviates backhaul congestion, balances workloads, and ensures scalable deployment of context-specific intelligence. Ultimately, it transforms the wireless network from a passive conduit into an active cognitive substrate.

Harnessing ubiquitous intelligence for a greener future

Data centers, the backbone of cloud-based LLMs deployment, are rapidly becoming major energy consumers, with U.S. facilities using 4.4% of national electricity consumption in 2023 and projected to consume 6.7–12% by 2028 due to escalating AI workloads[60], a trend accelerated by new mega-facilities like OpenAI's Stargate[61]. *Ubiquitous intelligence* mitigates this trajectory through redistributing computation to network edges, exploiting underutilized electricity in microgrids with surplus generation and limited storage[62]. Unlike conventional data centers that over-provision backup resources (e.g., batteries and generators) to buffer demand fluctuations but leave them idle during low traffic[63], our framework dynamically routes LLMs' workloads across intelligent edge nodes based on demand and energy availability, thereby harnessing wasted capacity and reducing carbon impact.

Primarily focusing on inference, *Ubiquitous Intelligence* can lead to potential net energy savings through the reuse of intermediate computational products, such as KV cache in transformer models, reducing

redundant calculations and lowering energy consumption. Another potential for energy savings within the paradigm lies in the improvement of inference quality. The quality of model outputs improves through accumulated knowledge sharing, users obtain satisfactory answers more efficiently. This reduces unnecessary queries and delays the need for retraining or replacing the model, possibly saving additional energy over time.

Challenges and open questions

Achieving truly *ubiquitous intelligence* over wireless networks requires overcoming several fundamental challenges arising from the interplay between communication, computation, and distributed learning. Key considerations include:

- **Scalable experience exchange**. The widespread dissemination of model updates and learned representations generates a substantial communication load. Efficient encoding, transmission prioritization, and adaptive scheduling are needed to ensure a timely and scalable experience sharing across wireless networks, aided by edge nodes.
- **Global model consistency**. Maintaining coherence alongside decentralized LLMs remains difficult in dynamic wireless network settings, particularly under intermittent connectivity and asynchronous updates[64]. Mechanisms for synchronization, alignment and reconciliation are essential to preserve learning stability and convergence.
- **Communication efficiency and latency management**. Exchanging intermediate features or knowledge modules drastically increases pressure on limited bandwidth, while the close integration of inference and communication introduces strict latency constraints. Real-time edge decision-making requires context-aware compression, progressive transmission, and synchronized delay control across computation and communication.
- **Security and robustness**. Distributed learning in open wireless environments exposes models to malicious updates, biased feedback, and privacy breaches. Safeguards such as secure aggregation, differential privacy, and anomaly detection are vital for preserving trust and integrity.
- **Efficient knowledge representation**. The diversity of experiential data across edge environments naturally gives rise to multi-modal information. Efficient representation of heterogeneous inputs poses significant challenges for unified knowledge integration, efficient retrieval, and downstream usability. Effective solutions must integrate this heterogeneous data in an LLMs-native manner.

Conclusion

Ubiquitous intelligence emphasizes the co-evolution of LLMs and wireless networks, where intelligence resides within the wireless infrastructure, and it dynamically coordinates across varied devices to ensure scalable, context-driven, and personalized service delivery. Distributed, context-aware, and adaptive learning across cloud, edge, and device tiers enables models to evolve continuously while meeting pressing demands for low latency, personalization, privacy, and energy efficiency. The intertwined development of wireless networks and LLMs establishes resilient, scalable, and environmentally responsible intelligence that expands through real-world experience. The shift from passive, static LLMs deployments to active, cognitive intelligent systems through wireless networks provides the foundation for continuous-learning capable, dynamic adaptive next-generation AI.

Data availability

No datasets were generated or analysed during the current study.

Received: 4 September 2025; Accepted: 24 November 2025;
Published online: 02 February 2026

References

1. Vaswani, A. et al. Attention is all you need https://arxiv.org/abs/1706.03762. (2023).
2. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks https://arxiv.org/abs/2005.11401. (2021).
3. Xi, Z. et al. The rise and potential of large language model based agents: a survey https://arxiv.org/abs/2309.07864. (2023).
4. Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J. & Sattar, M. A. Industrial applications of large language models. *Sci. Rep.* **15**, 13755 (2025).
5. Hu, Q. et al. Characterization of large language model development in the datacenter https://arxiv.org/abs/2403.07648. (2023).
6. Xu, M. et al. Unleashing the power of edge-cloud generative ai in mobile networks: A survey of aigc services. *IEEE Commun. Surv. Tutor.* **26**, 1127–1170 (2024).
7. Sánchez, A. G.*Azure OpenAI Service for Cloud Native Applications* ("O'Reilly Media, Inc.", 2024).
8. Alto, V.*Building LLM Powered Applications: Create intelligent apps and agents with large language models* (Packt Publishing Ltd, 2024).
9. Deng, Z. et al. Exploring deepseek: A survey on advances, applications, challenges and future directions. *IEEE/CAA J. Autom. Sin.* **12**, 872–893 (2025).
10. Li, A. et al. Evaluating modern gpu interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect. *IEEE Trans. Parallel Distrib. Syst.* **31**, 94–110 (2019).
11. Le, M. et al. Applications of distributed machine learning for the Internet of Things: A comprehensive survey. *IEEE Commun. Surv. Tutor.* (2024).
12. Blika, A. et al. Federated learning for enhanced cybersecurity and trustworthiness in 5G and 6G networks: A comprehensive survey. *IEEE Open J. the Commun. Soc.* (2024).
13. Cooper, K. & Geller, M. Advancing personalized federated learning: Integrative approaches with AI for enhanced privacy and customization https://arxiv.org/abs/2501.18174. (2025).
14. Zhang, X. et al. Beyond the cloud: Edge inference for generative large language models in wireless networks. *IEEE Trans. Wirel. Commun.* (2024).
15. Cox, B., Birke, R. & Chen, L. Y. Memory-aware and context-aware multi-DNN inference on the edge. *Pervas. Mob. Comput.* **83**, 101594 (2022).
16. Hu, S., Chen, X., Ni, W., Hossain, E. & Wang, X. Distributed machine learning for wireless communication networks: Techniques, architectures, and applications. *IEEE Commun. Surv. Tutor.* **23**, 1458–1493 (2021).
17. Khoramnejad, F. & Hossain, E. Generative ai for the optimization of next-generation wireless networks: Basics, state-of-the-art, and open challenges. *IEEE Commun. Surv. Tutor.* (2025).
18. He, Q. et al. Integrating iot and 6G: Applications of edge intelligence, challenges, and future directions. *IEEE Trans. Serv. Comput.* (2025).
19. Zhang, M., Cao, J., Shen, X. & Cui, Z. Edgeshard: Efficient LLM inference via collaborative edge computing https://arxiv.org/abs/2405.14371. (2024).
20. Liang, L. et al. Large language models for wireless communications: From adaptation to autonomy https://arxiv.org/abs/2507.21524. (2025).
21. Wen, D. et al. Integrated sensing-communication-computation for edge artificial intelligence. *IEEE Internet Things Mag.* **7**, 14–20 (2024).
22. Wu, W. et al. Split learning over wireless networks: Parallel design and resource management. *IEEE J. Sel. Areas Commun.* **41**, 1051–1066 (2023).
23. Fang, H., Xiao, Z., Wang, X., Xu, L. & Hanzo, L. Collaborative authentication for 6G networks: An edge intelligence based autonomous approach. *IEEE Trans. Inf. Forensics Secur.* **18**, 2091–2103 (2023).
24. Dang, S., Amin, O., Shihada, B. & Alouini, M.-S. What should 6G be?. *Nat. Electron.* **3**, 20–29 (2020).
25. Kaplan, J. et al. Scaling laws for neural language models https://arxiv.org/abs/2001.08361. (2020).

26. Brown, T. B. et al. Language models are few-shot learners https://arxiv.org/abs/2005.14165. (2020).
27. Chen, X. et al. P^2 law: Scaling law for post-training after model pruning https://arxiv.org/abs/2411.10272. (2025).
28. Ouyang, L. et al. Training language models to follow instructions with human feedback https://arxiv.org/abs/2203.02155. (2022).
29. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models https://arxiv.org/abs/2201.11903. (2023).
30. Zhang, Q. et al. A survey on test-time scaling in large language models: What, how, where, and how well? https://arxiv.org/abs/2503.24235. (2025).
31. Sabourin, V. *Strategic Adoption of 5G Technology: New Applications and Services: New Applications and Services* (IGI Global, 2025).
32. Vyas, A. K., Khatri, N. & Jha, S. K.*6G Communication network: architecture, security and applications* (CRC Press, 2024).
33. Zhang, Z. et al. Guest editorial: sustainable big ai model for wireless networks. *IEEE Wirel. Commun.* **31**, 18–19 (2024).
34. Ishtiaq, M., Saeed, N. & Khan, M. A. Edge computing in the internet of things: A 6G perspective. *IT Profess.* **26**, 62–70 (2024).
35. Guo, Q., Tang, F. & Kato, N. Federated reinforcement learning-based resource allocation in D2D-enabled 6G. *IEEE Netw.* **37**, 89–95 (2023).
36. Xia, Q., Jiao, Z. & Xu, Z. Online learning algorithms for context-aware video caching in D2D edge networks. *IEEE Trans. Parallel Distrib. Syst.* **35**, 1–19 (2024).
37. Sanjalawe, Y. et al. A review of 6g and ai convergence: Enhancing communication networks with artificial intelligence. *IEEE Open J. Commun. Soc.* (2025).
38. Panda, S. *Scalable Artificial Intelligence Systems: Cloud-Native, Edge-AI, MLOps, and Governance for Real-World Deployment* (Deep Science Publishing, 2025).
39. Syed, N., Anwar, A., Baig, Z. & Zeadally, S. Artificial intelligence as a service (aiaas) for cloud, fog and the edge: State-of-the-art practices. *ACM Comput. Surv.* **57**, 1–36 (2025).
40. Pivoto, D. G. S., de Figueiredo, F. A., Cavdar, C., de Lima Tejerina, G. R. & Mendes, L. L. A comprehensive survey of machine learning applied to resource allocation in wireless communications. *IEEE Commun. Surv. Tutor.* (2025).
41. Xu, M. et al. When large language model agents meet 6G networks: Perception, grounding, and alignment. *IEEE Wirel. Commun.* (2024).
42. Souza, P., Ferreto, T. & Calheiros, R. Maintenance operations on cloud, edge, and iot environments: taxonomy, survey, and research challenges. *ACM Comput. Surv.* **56**, 1–38 (2024).
43. Wang, C. et al. Understanding 5g performance on heterogeneous computing architectures. *IEEE Commun. Mag.* (2024).
44. Boateng, G. O. et al. A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions. *IEEE Commun. Surv. Tutor.* (2025).
45. Chen, H. et al. Towards edge general intelligence via large language models: Opportunities and challenges. *IEEE Netw.* (2025).
46. Villalobos, P. et al. Will we run out of data? limits of LLM scaling based on human-generated data https://arxiv.org/abs/2211.04325. (2024).
47. Zeydan, E., Arslan, S. S., Turk, Y., Hewa, T. & Liyanage, M. The role of mobile communications for industrial automation: Architecture, applications and challenges. *IEEE Open J. Commun. Soc.* (2025).
48. Agarwal, B., Irmer, R., Lister, D. & Muntean, G.-M. Open RAN for 6G networks: Architecture, use cases and open issues. *IEEE Commun. Surv. Tutor.* (2025).
49. Tao, M. et al. Federated edge learning for 6 G: Foundations, methodologies, and applications. *Proc. IEEE* (2024).
50. Hou, X., Zhao, Y., Wang, S. & Wang, H. Model context protocol (MCP): Landscape, security threats, and future research directions https://arxiv.org/abs/2503.23278. (2025).
51. Ioannou, I., Nagaradjane, P., Vassiliou, V., Pitsillides, A. & Christophorou, C. *Distributed Artificial Intelligence for 5G/6G Communications: Frameworks with Machine Learning* (CRC Press, 2024).
52. Hamdi, W., Ksouri, C., Bulut, H. & Mosbah, M. Network slicing-based learning techniques for IoV in 5 G and beyond networks. *IEEE Commun. Surv. Tutor.* **26**, 1989–2047 (2024).
53. Liu, Y. et al. Cachegen: Kv cache compression and streaming for fast large language model serving. *Proceedings of the ACM SIGCOMM 2024 Conference* 38–56 (2024).
54. Nguyen-Kha, H. et al. Dt-aided resource management in spectrum sharing integrated satellite-terrestrial networks https://arxiv.org/abs/2507.20789. (2025).
55. Yan, H., Huang, H., Zhao, Z., Wang, Z. & Zhao, Z. Accuracy-aware mllm task offloading and resource allocation in uav-assisted satellite edge computing. *Drones* **9**, 500 (2025).
56. Liu, C.-F., Bennis, M., Debbah, M. & Poor, H. V. Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing. *IEEE Trans. Commun.* **67**, 4132–4150 (2019).
57. Sun, M. et al. Llm-based task offloading and resource allocation in satellite edge computing networks. *IEEE Trans. Vehic. Technol.* (2025).
58. Ornee, T. Z., Shisher, M. K. C., Kam, C. & Sun, Y. Context-aware status updating: Wireless scheduling for maximizing situational awareness in safety-critical systems https://arxiv.org/abs/2310.06224. (2023).
59. Lin, Z. et al. Hierarchical split federated learning: Convergence analysis and system optimization https://arxiv.org/abs/2412.07197. (2025).
60. Guidi, G. et al. Environmental burden of united states data centers in the artificial intelligence era https://arxiv.org/abs/2411.09786. (2024).
61. OpenAI. Announcing The Stargate Project. https://openai.com/index/announcing-the-stargate-project/ Official OpenAI announcement of the Stargate Project. (2025).
62. Erices, C., Filis, P. & Papantonopoulos, E. Hairy black holes in disformal scalar-tensor gravity theories. *Phys. Rev. D.* **104**, 024031 (2021).
63. Shi, Y., Xu, B., Zhang, B. & Wang, D. Leveraging energy storage to optimize data center electricity cost in emerging power markets https://arxiv.org/abs/1606.01536. (2016).
64. Chen, Z., Dahl, M. & Larsson, E. G. Decentralized learning over wireless networks: The effect of broadcast with random access https://arxiv.org/abs/2305.07368. (2023).

Acknowledgements

The work described in this paper was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China under a fellowship award (HKU RFS2122-7S04), NSFC/RGC CRS (CRS_HKU702/24), the Areas of Excellence scheme grant (AoE/E-601/22-R), Collaborative Research Fund (C1009-22G), and the Grants 17212423 & 17304925, and in part by the Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) (SGDX20230821091559018).

Author contributions

X. Yin and F. You led the writing of the manuscript following extensive discussions and with input from all authors. H. Du and K. Huang discussed the idea and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Hongyang Du.