Toward Effective Retrieval Augmented Generative Services in 6G Networks

Xi Huang, Yinxu Tang, Junling Li, Ning Zhang, and Xuemin Shen

Abstract

Retrieval augmented generation (RAG) empowers generative language services by integrating extensive context from external data sources (a.k.a. knowledge bases). The current RAG-enhanced generative services are predominantly hosted in cloud environments, relying on static knowledge bases without real-time sensory information which may lead to constrained scalability, responsiveness, and overall service quality. One promising opportunity is to extend the deployment of such services to the network edge, leveraging the anticipated capabilities of 6G networks. In this article, we propose a deployment framework for RAG-enhanced generative services in 6G. We address the key challenges at the convergence of service deployment, 6G networks, and user interactions. Additionally, we explore potential techniques to enhance RAG-based services through data fusion, dynamic knowledge base deployment, service customization, and interactive user experiences. Lastly, we shed light on future paths toward the effective deployment and delivery of RAG-enhanced generative services.

Introduction

Generative services based on large language models (LLMs) have been extensively adopted for various applications, such as virtual assistants in healthcare and smart-home scenarios. To enhance the accuracy and reliability of such services, retrieval augmented generation (RAG) techniques are developed by additionally integrating factual information from external data sources (a.k.a. knowledge bases or KBs) to improve service performances.

Currently, RAG-enhanced generative services opt for cloud-based deployment over fixed KBs due to their substantial resource requirements. However, the ever-growing data transfer demands and privacy concerns of managing and updating KBs in real-time have posed significant challenges in the scalability, responsiveness, and quality of such services' delivery.

As wireless communication technologies evolve toward 6G standards, it is promising to extend the deployment of generative services to network edges by utilizing the foreseen supports of 6G networks [1]. Particularly, 6G features several technological advancements. First, 6G will densely deploy smart sensors with enhanced interconnections, enabling continuous KB updates by fusing multi-modality data from various sources. Second, by integrating heterogeneous network resources, 6G will facilitate ultra-high-speed and highly reliable delivery (up to 1Tbps) of generative services [2]. Third, by orchestrating massive computing resources (~10^8 devices/km^2) at the network edge, 6G will support scalable deployments and flexible customization of RAG-enhanced generative services.

Based on such promising features, this article envisions a 6G-based deployment architecture for RAG-enhanced generative services, as shown in Fig. 1(a). Under the framework, the generative services are deployed across the cloud-edge-end continuum, responsively interacting with users and empowered by continuously updated KBs with real-time information from various sensory sources. This design enables more efficient utilization of computational and communication resources at the edge, offering a more scalable framework for RAG to leverage knowledge bases from various sources in a privacy-preserving manner. Nonetheless, several challenges remain to impede the effective realization of the framework.

The *first* challenge lies in the fusion of data from massive sensory sources in 6G for KB updates. The diversity of user interests and privacy concerns makes it challenging to incentivize their data-sharing behaviors. Meanwhile, fusing real-time data from massive sensory sources requires resolving the heterogeneity in data modalities, time scales, and perspectives of sensing. Such heterogeneity poses significant difficulties in fusing data and investigating their values for the services.

The *second* challenge involves distributing knowledge bases (KBs) for LLMs deployed on diverse edge servers in 6G networks. The heterogeneous edge environments, varying computational capabilities, and dynamic service demands make it difficult to distribute and dynamically update KBs among edge servers. Additionally, integrating edge servers with the cloud in a hybrid cloud-edge architecture presents an extra challenge for KB distribution.

Xi Huang is with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen 518172, China; Yinxu Tang is with the Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA; Junling Li (corresponding author) is with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China; Ning Zhang is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada; Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Digital Object Identifier: 10.1109/MNET.2024.3436670
Date of Current Version: 18 November 2024
Date of Publication: 6 August 2024

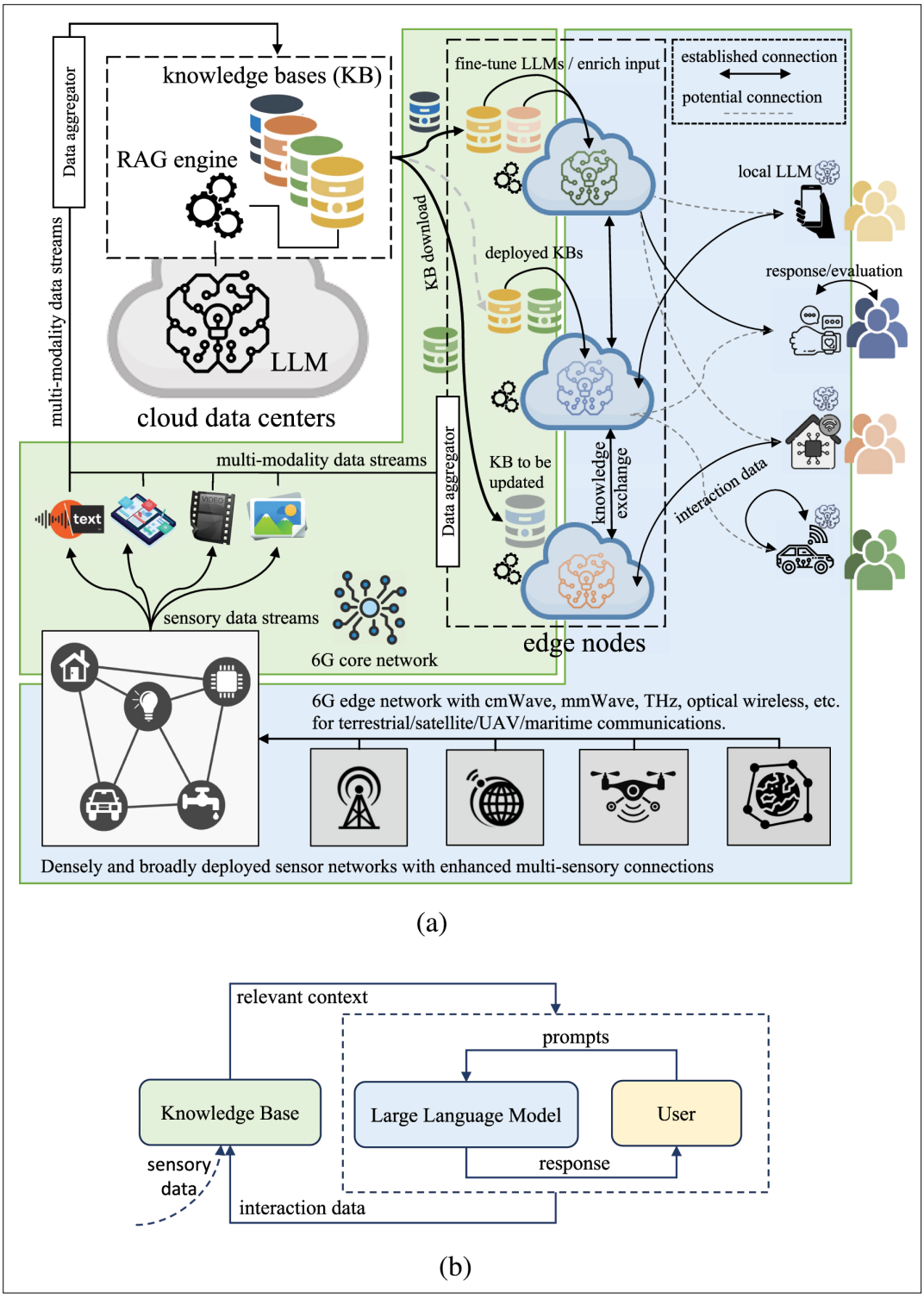

FIGURE 1. An illustration of 6G-empowered RAG-enhanced generative services. a) A deployment architecture for RAG-enhanced generative services. In the cloud, network operators pre-train LLMs and provide KBs while maintaining a platform for crowd-sourcing data. At the network edge, LLMs deliver generative services and fine-tune with stored KBs under advanced 6G networks. Users enjoy responsive generative services through compressed LLMs locally or switch to higher-quality RAG-enhanced services deployed in edge servers. b) The doubly reinforced data loops in such services where services continuously enrich KBs with data from user interactions and diverse data sources.

The *third* challenge focuses on customizing generative services using RAG-enhanced Large Language Models (LLMs) across edge servers. Implementing comprehensive RAG-enhanced LLMs with vast numbers of parameters can lead to resource-intensive training processes, resulting in latency issues for real-time applications. Deploying compact Language Models (LMs) at the network edge is crucial for low-latency service customization, requiring a strategic balance of LLMs and

Knowledge Bases (KBs). Furthermore, customizing RAG-enhanced generative services at the network edge faces a significant hurdle when multiple KBs are spread across different edge servers. Aggregating these KBs consumes substantial communication resources and raises privacy concerns.

The *fourth* challenge concerns how to associate users with customized RAG-enhanced generative services. During user-service interactions, end devices face challenges in deciding whether to process user input locally or upload it to the edge server. On the one hand, unknown dynamics in 6G networks makes it necessary to introduce online learning in the interaction process. On the other hand, the limited connection capacity of each edge server implies non-coordinated relationships among end devices, necessitating the application of game theory.

The challenges involve integrating decision-makers across the cloud-edge-end continuum in 6G networks, spanning mechanism design, game theory, network optimization, and machine learning. These decision-makers, including network operators, LLM-based generative service providers, and users, have diverse objectives related to resource efficiency and user experience. Balancing their interests to achieve a sustainable solution is complex, given uncertainties from 6G network dynamics and information disparities. Addressing these issues requires cohesive online control and learning to mitigate uncertainties and achieve scalable RAG-enhanced generative services with minimal performance trade-offs.

The remainder of the article is organized as follows. We propose the deployment framework for RAG-enhanced generative services in the section "RAG-Enhanced Generative Service Framework." Then, we discuss in sections "6G-Enhanced Data Fusion for KB Update," "6G-Empowered Cloud-Edge Collaborations for KB Deployment," "6G-Based Cloud-Edge Collaborations for RAG-Enhanced Service Customization," and "6G-Assisted End-Edge Interactions for User-Service Matching" the challenges in the enhancement and update of KBs across the cloud and the edge, as well as the customization and delivery of RAG-enhanced services. Finally, we envision future directions in the section "Implications and Potential Directions" and followed by the concluding remarks in the section "Conclusion."

RAG-Enhanced Generative Service Framework

In this article, we propose a 6G-based deployment framework (see Fig. 1(a)) for RAG-enhanced generative services. The framework is based on a promising feature of RAG-enhanced generative services – the doubly reinforced data loops (see Fig. 1(b)), where the services continuously enrich KBs with data from user interactions and diverse data sources, *e.g.*, wearable and sensory devices in 6G-based smart homes.[1]

Specifically, the framework operates through interactions between the cloud, edges, and users.

- At the cloud side, network operators leverage abundant resources to pre-train LLMs while providing a set of KBs with offline data. They also maintain a platform to crowd-source data from user interactions and sensors in 6G networks. Sourced data are utilized to update the KBs.
- At the network edge, LLMs are deployed across heterogeneous edge servers to deliver generative services. Each edge server also stores a subset of KBs to fine-tune the hosting LLMs and empower the services. The deployed services operate under advanced cell-free 6G networks [3] to interact with users.
- At the user side, empowered by on-device intelligence, users enjoy generative services responsively through locally deployed compressed LLMs but at a reduced quality. Alternatively, users can switch to RAG-enhanced services deployed in edge servers at a higher quality of experience. In this case, users should consider overheads due to 6G dynamics and matching between their preferences and services.

During the interactions, the dynamics of 6G networks and user interactions with generative services bring critical issues to realizing the framework. The following sections discuss such issues and the challenges regarding data fusion for KB updates, cloud-edge collaborative KB deployment, service customization, and end-edge cooperation for user-service interactions.

The challenges involve integrating decision-makers across the cloud-edge-end continuum in 6G networks, spanning mechanism design, game theory, network optimization, and machine learning.

6G-Enhanced Data Fusion for KB Update

RAG-enhanced generative services contain outdated knowledge, while their KBs are updated infrequently with limited user data, making it challenging to understand user intentions and serve users with personalized information. For instance, GPT-4 Turbo, the base of Microsoft's Copilot services, has knowledge of the world up to April 2023 by the time we write this article. The services refresh their KBs (*e.g.*, search indexes) globally every few days to weeks, achieving a limited timeliness and quality of services.

To this end, 6G networks will facilitate the fusion of monitored user and environment dynamics through densely deployed sensors with enhanced interconnections. This way, generative services can leverage fused information to enrich service context and improve service quality. To unleash the potential of 6G-enhanced data fusion, we present in Fig. 2 how data sharing and fusion proceed through the data fusion module's coordination. Next, we investigate them in detail.

Data Sharing

Exploiting 6G's support for sensory data acquisition incurs two major concerns. On the one hand, sharing real-time sensory data poses privacy leakage risks as the data may involve users' private information. To mitigate the concern, it is promising to preprocess the information locally on the user side and extract only the patterns of user or environment dynamics before submitting data for fusion. For example, federated computing and split learning techniques can be utilized. Mao et al. investigate how federated learning and blockchain can be integrated to ensure

[1] Although current RAG frameworks do not directly support the outer loop, researchers have practically explored similar concepts to enhance application performance. For example, state-of-the-art LLM-based applications have limited context windows, making it unaffordable to store complete interaction histories for long conversations. Some applications use LLMs to merge new user responses with conversation history in real time, extracting essential information to maintain up-to-date context.

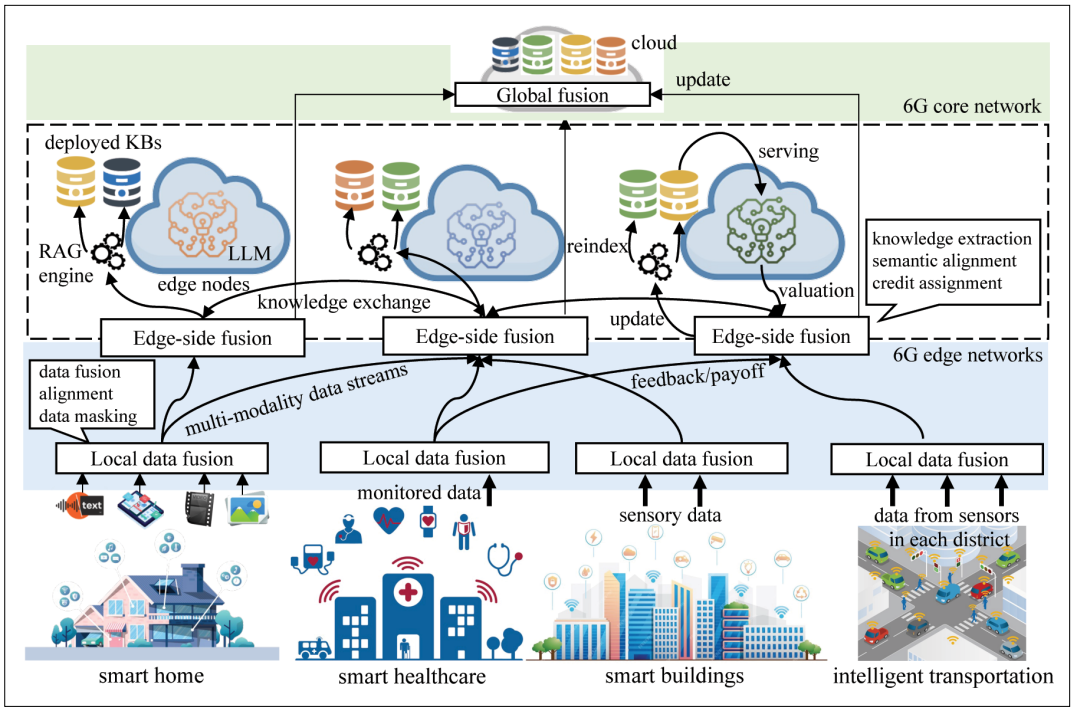

FIGURE 2. An illustration of 6G-based data fusion for RAG-enhanced generative services. Such a framework leverages the capabilities of 6G networks to gather and merge diverse data sources (e.g., local data fusion and edge-side fusion) for enhancing RAG models' performance in generating content or providing services.

In 6G networks with ultra-high-speed transmission capabilities, the primary challenge in updating knowledge bases (KBs) through data fusion is effectively integrating data from diverse sources.

secure and privacy-preserving data sharing in 6G networks.

On the other hand, users need more motivation to authorize access to their data. To this end, incentive mechanisms should be designed to optimize payoffs for users, while the values of their shared data must be properly evaluated. For example, Zhang et al. [5] proposed effective mechanisms to incentivize fresh data acquisition from massive sources with known costs and benefits. However, in RAG-enhanced services, the values of shared data cannot be determined before they are applied, making it difficult to apply existing mechanism designs in practice. For such issues, [6] initiates the study and shows that incorporating signaling schemes can reduce uncertainty in data values and improve users' payoffs of sharing data.

Data Fusion

In 6G networks with ultra-high-speed transmission capabilities, the primary challenge in updating knowledge bases (KBs) through data fusion is effectively integrating data from diverse sources. For instance, in scenarios like autonomous vehicles equipped with RAG-enhanced virtual assistants for conversation-based services, precise and secure navigation relies on continuously aggregating information from sensors and neighboring vehicles (e.g., cameras, lidar, radar). Such data, capturing environmental dynamics from different perspectives and time scales, presents difficulties in aligning and delivering critical information in real-time. One potential solution is hierarchical data fusion, where each vehicle locally combines sensor data before sharing it with others. At each fusion stage, balancing fusion quality and timeliness is crucial. While higher fusion quality demands more computation, it may impact the speed of data sharing. These concerns are generally related to data stream processing. In a recent study Huang et al. [7] addressed a data stream scheduling problem involving data streams forming a directed acyclic graph in dynamic networks. However, the broader issues pertaining to data streams encompassing diverse modalities and sampling granularities remain unexplored.

6G-Empowered Cloud-Edge Collaborations for KB Deployment

For future 6G networks, collaborating the cloud with heterogeneous edge servers for deploying KBs involves complex optimization tasks, and reinforcement learning (RL) can be a powerful tool to address these challenges [8]. The design of an RL framework (see Fig. 3) to accomplish these tasks can be done in either a cooperative or a non-cooperative environment. In either case, the framework design should be elaborated on the following common aspects.

Dynamic Environments – The RL-based KB deployment algorithm design should account for the dynamic nature of edge environments and implement mechanisms to adapt RL policies to changes in edge servers' computational capabilities, memory constraints, and network

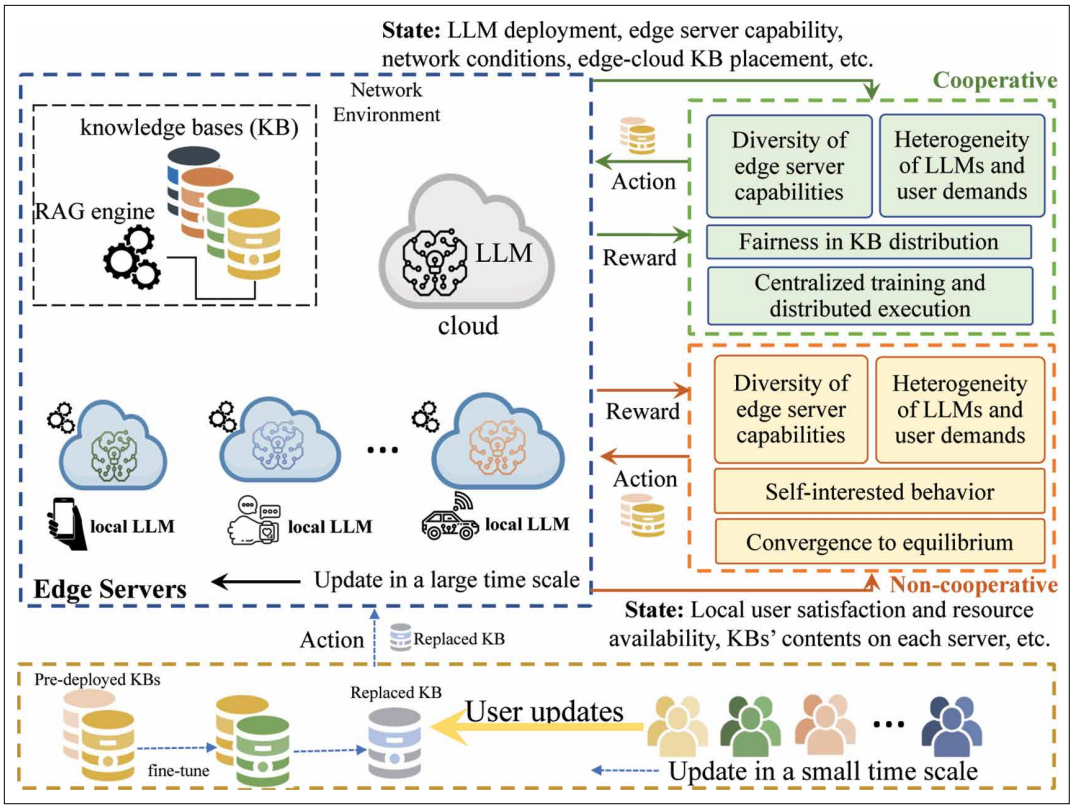

FIGURE 3. An illustration of edge-based KB deployment for RAG-enhanced generative services based on reinforcement learning in 6G networks. In cooperative scenarios, cloud and edge servers collaborate toward a common goal, while in non-cooperative scenarios, edge servers act independently to optimize their own objectives.

bandwidths, as well as LLMs' heterogeneity and network conditions.

Hybrid Cloud-Edge Architecture – Facing a hybrid cloud-edge architecture, the RL agent should decide on what parts of the KB are stored centrally in the cloud and what parts are distributed across edge servers. This decision could be made based on factors such as the computational constraints of edge servers, the response time requirements of users, and the communication costs between the cloud and edge servers.

Real-Time Adaptability – The framework should ensure that KB deployment can adapt in real-time to changes in user demands and server conditions. The frequency of KB re-deployment should consider the trade-off between the redeployment cost and the payoff for the overall system.

Cooperative KB Deployment

In cooperative scenarios, the cloud and edge servers collaborate using RL to optimize KB deployments collectively. A comprehensive state representation includes LLM deployment, server capability, network conditions, and KB content. Action spaces for each server involve requesting or sharing KBs. The reward function should reflect cooperation, considering system performance and fairness. Training stability is crucial; techniques like centralized training and distributed execution help stabilize learning in cooperative KB deployment settings.

Non-Cooperative KB Deployment

In non-cooperative scenarios, edge servers act independently, optimizing their objectives without explicit collaboration. This could include metrics like the increase in the number and diversity of local users supported by the server, and the improvement in local user satisfaction. Multi-agent reinforcement learning (MARL) can be used to model the self-interested behavior of individual edge servers [9]. First, we should model each edge server as a self-interested agent aiming to optimize its performance. The state should include information relevant to individual objectives, such as local user satisfaction and resource availability. Second, we need to define the actions for each agent related to KB sharing or requesting KBs from other servers or the cloud. Each agent decides its actions based on its objectives. Third, we have to address the potential challenges related to convergence in a non-cooperative setting. Game-theory-based MARL may be necessary to ensure convergence to equilibrium strategies.

Overall, by carefully designing an RL-based KB deployment scheme under a 6G-empowered hybrid cloud-edge architecture, we can optimize KB deployments for LLMs across the cloud and heterogeneous edge servers.

6G-Based Cloud-Edge Collaborations for RAG-Enhanced Service Customization

Deploying customized small models derived from RAG-enhanced LLMs at the edge of

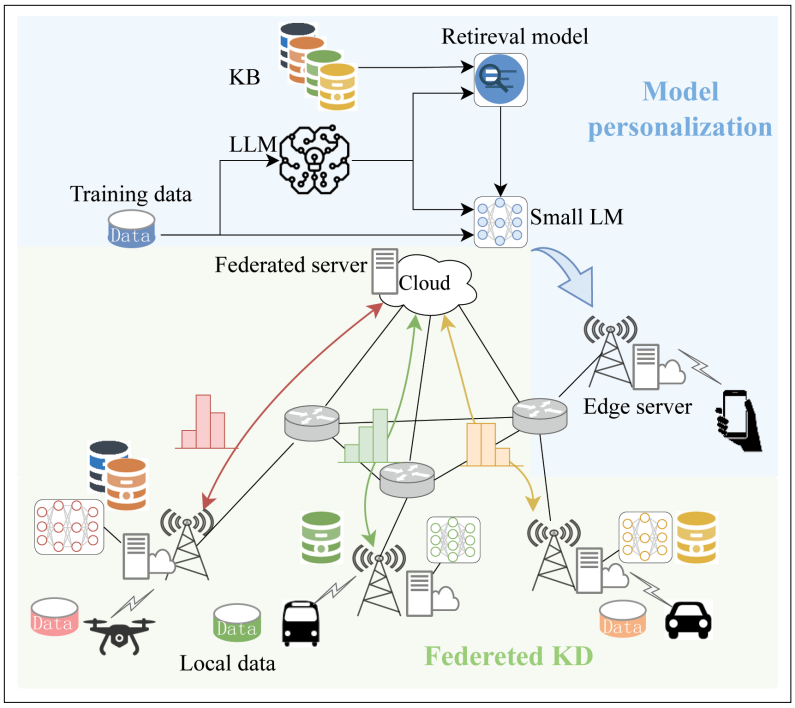

FIGURE 4. An illustration of RAG-enhanced service customization targeting 6G. Customized small models, derived from RAG-enhanced LLMs, are deployed at the network edge, while the cloud aggregates knowledge from participating edge servers in federated mode, enabling effective knowledge transfer and enhancing the resilience of RAG-enhanced services.

networks is crucial for delivering high-quality, low-latency services to users. To this end, we propose a framework for RAG-enhanced service customization in 6G networks, as shown in Fig. 4. However, under the framework, training these customized models to integrate both the inherent knowledge of the LLM and the personalized insights of users presents a formidable challenge. First, the scarcity of user-provided training data and KBs limits the performance of the personalized model. Additionally, when multiple users share a common personalized model, the non-identically distributed (non-IID) nature of data across diverse KBs complicates the creation of a cohesive and effective customization model. To address these challenges, we investigate the personalized RAG-enhanced service customization on the cloud and federated RAG-enhanced service customization at the edge in this section, to respectively address the challenges of limited data availability and non-IID data distribution.

Personalized RAG-Enhanced Service Customization on Cloud

To optimize low-latency RAG service provisioning, deploying compact models to edge servers in proximity to users holds great promise. Utilizing the user-provided KB for fine-tuning small personalized models enhances the ability to deliver task-specific services. However, the inherent limitation of user-provided data can impede model performance. Addressing this challenge, integrating high-capacity LLMs with external KBs becomes a promising strategy.

Kim et al. [10] employ the generated dataset from LLM to augment the original training dataset extracted from existing KBs. This augmented dataset significantly improves the performance of the personalized model. Simultaneously, LLM outputs for queries are employed in knowledge distillation to further enhance the personalized model. Kang et al. [11] leverage the high-quality outputs of LLM to fine-tune small personalized models. A retriever is employed to extract the most relevant knowledge from LLM's output, augmenting the KBs. This selected knowledge is then utilized in the training of the personalized model, ensuring a refined and knowledge-enriched model for improved performance.

Federated RAG-Enhanced Service Customization at Edge

In practical scenarios, numerous users may seek access to the same task through distributed edge servers, presenting an opportunity to harness the collective data of users for collaborative model training. In this collaborative framework, the cloud aggregates the knowledge from participating edge servers, enabling effective knowledge transfer and resulting in a more resilient RAG-enhanced service.

Han et al. [12] focus on the issue of non-identically distributed (non-IID) data among multiple distributed clients, specifically edge servers, during model training. Their method targets capturing the global data distribution without compromising client privacy by avoiding raw data and model parameter sharing. It allows for diverse local model architectures across edge servers influenced by local computation resources and user latency requirements. This diversity in architecture aims to balance model size with inference latency for real-time service delivery. To tackle the non-IID data challenge, they introduce a novel approach where a discriminator model, trained on the cloud, distinguishes model outputs from each edge server. Edge servers then engage in local adversarial training with the cloud server to transfer local knowledge, enabling the cloud server to align model outputs universally, enhancing model accuracy across all edge servers.

6G-Assisted End-Edge Interactions for User-Service Matching

With jointly deployed LLMs and KBs on edge servers, 6G networks can facilitate customized RAG-enhanced generative services to users [13]. The tradeoffs are illustrated in Fig. 6 based on our experiments.

Essentially, the end-edge interaction aims to optimize the matching between edge servers and edge devices. However, the uncertainty in the dynamics of 6G edge networks can bring significant challenges. In this section, we discuss how to address such uncertainties in different scenarios.

Coordinated Matching

Coordinated matching between edge devices and edge servers utilizes a centralized mechanism to optimally assign tasks or services to servers based on predefined criteria, aiming to optimize a global objective function. But edge devices often lack prior knowledge of LLM performances with their associated KBs on heterogeneous edge servers.

Meanwhile, their service selection faces uncertain communication over-heads (e.g., response latency) between end devices and edge servers. Addressing these concerns requires the coordination of service selection procedures by different end devices. Their common goal is to maximize users' satisfaction scores while minimizing performance loss in terms of communication costs due to uncertainty. For example, the problem is studied in [14] by focusing on a typical generative service – emoji prediction. By leveraging edge computing, the proposed scheme can effectively integrate online learning into the selection procedure and balances users' satisfaction, response latency, and energy budgets in an adaptive and tunable fashion.

Non-Coordinated Matching

Non-coordinated matching between edge devices and edge servers involves independent decision-making by individual devices, where devices select servers based on their own criteria without central coordination to optimize objectives locally. Due to limited service capacity, edge servers hosting RAG-enhanced generative services may struggle to accommodate a large number of users. In practical scenarios, users often have diverse quality of service requirements based on their specific applications. In uncertain and dynamic networks, end devices independently select services in an exploratory manner to maximize their own benefits while minimizing performance loss in a non-coordinated fashion. A significant challenge arises in designing a non-coordinated matching mechanism to address the matching between users and generative services under these uncertainties. In a recent study Tang et al. [15] addressed the problem of network service matching by employing a marriage between online learning and stable matching. Nonetheless, the achieved stable matching may not be socially optimal. Applying other approaches to addressing the challenge, such as the incentive mechanism design and information design, deserves investigation.

Implications and Potential Directions

In this section, we discuss the implications of our proposed framework on privacy and energy issues. We also explore potential avenues for future research.

Privacy and Energy Implications

As for privacy concerns, in cloud-based systems utilizing RAG-enhanced services, service providers must gather vast amounts of data regularly from diverse sources to enhance RAG capabilities. However, this centralized approach incurs exorbitant costs for data acquisition and poses a risk to sensitive user information due to centralized data management. In contrast, our proposed framework leverages 6G support to extend the deployment of these services to the network edge in a decentralized manner. This approach enables each RAG service to incorporate real-time data, implement unique data management strategies, and enhance privacy protection. By offering increased privacy protection options, distributing risks, and enabling tailored RAG-enhanced services, this framework provides greater flexibility and security.

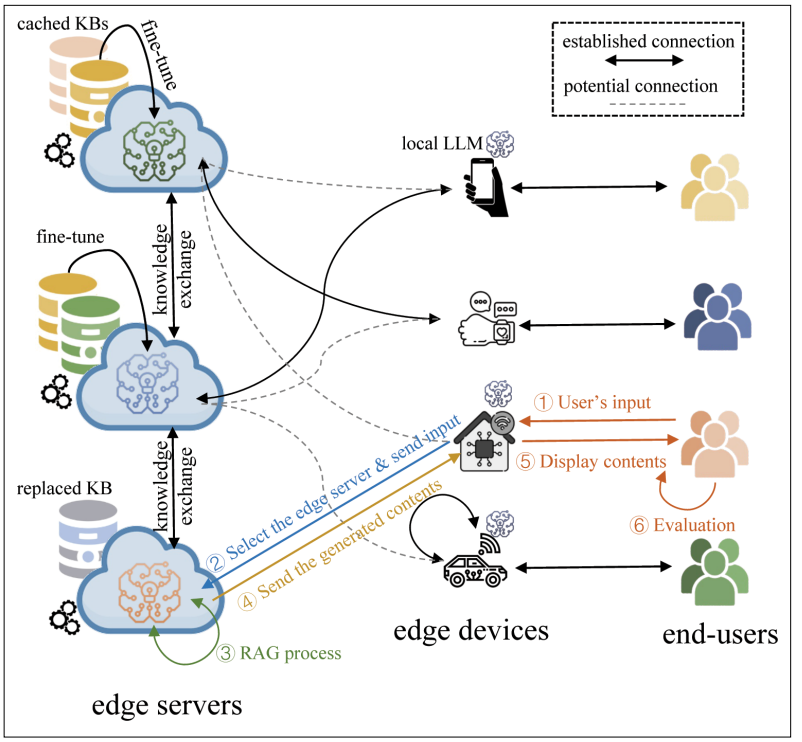

FIGURE 5. An illustration of end-edge interactions for the matching between users and generative services, where LLMs and KBs are jointly deployed. Edge devices dynamically associate with edge servers, send input, and evaluate the contents generated through the RAG process.

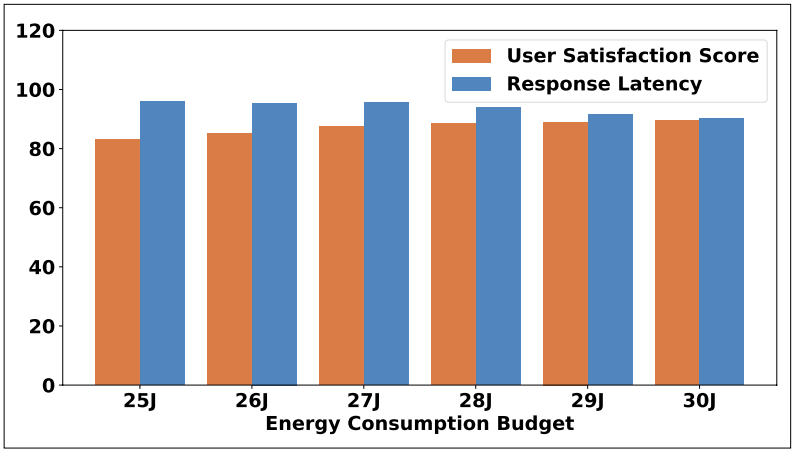

FIGURE 6. An illustration of the tradeoffs among user satisfaction score, response latency, and energy consumption in end-edge interactions for user-service matching with heterogeneous generative models, such as Bi-LSTM and BERT. As the energy consumption budget increases, user satisfaction rises while response latency decreases.

As for energy concerns, conventional approaches mainly involve either local-deployed or cloud-deployed schemes for user-service matching. The local-deployed scheme involves computation on devices, offering low response latency, energy efficiency, and data security. However, limited computation resources can hinder the development of high-performance LLM, leading to potentially unsatisfactory results for end users. On the other hand, the cloud-deployed scheme involves end-cloud data transmissions to access the rich computation resources of the cloud, offering promising services.

With the increasing application of edge intelligence systems, 6G-assisted end-edge interactions for user-service matching is a widely studied area.

In contrast to these approaches, our novel framework harnesses the power of 6G-aided network edge. This framework allows edge devices to decide whether to process input locally with on-device LLM or upload it to the edge server. While this approach can enhance RAG services by leveraging predeployed LLMs and KBs on edge servers, it may result in increased latency and energy consumption compared to the local-deployed scheme. Furthermore, RAG services can benefit from edge computing with lower latency and energy consumption compared to the cloud-deployed scheme.

Cooperative Fusion of Multi-Modality Data for KB Update

Toward 6G-enhanced data fusion for RAG-enhanced generative services, multiple directions remain to be explored. First, each sensory source may upload data to multiple data fusion modules held by distinct cloud/edge service owners. This can lead to excessive communication overheads for redundant data sharing across heterogeneous networks. A promising direction is to incentivize collaboration among service owners by regulating their data acquisition strategies from sensory sources. Second, sensory data have multiple modalities such as video, audio, and text, which encode real-time information under different representations. It is critical to develop data valuation schemes to handle such heterogeneity toward fair and incentive-compatible mechanism designs. Third, while data fusion helps to refresh KBs with timely information, it can introduce biases that affect service performances. Thus, another direction is to develop simple but effective schemes to verify bias before updating KBs, which requires coordination between operators and sensory sources.

Quantification of Performance Gain for KB Deployment

In the pursuit of 6G-empowered cloud-edge collaborations for KB deployment, ongoing research is crucial to quantify network performance improvements. RL-based KB deployment algorithms must focus on the reward function, considering deployment costs, storage resources, communication overhead, and user satisfaction with LLMs. Quantifying the impact on user satisfaction poses challenges, necessitating research on performance enhancement. Real-time KB deployment exploration is vital, requiring dynamic strategies for edge servers to meet evolving demands. This involves developing scaling, load balancing, and resource allocation policies to optimize KB deployment efficiently.

Online Service Customization Under Resource Limits

A regularly overlooked aspect of RAG-enhanced service provision is the potential for online service customization, integrating real-time model training with seamless service provisioning. In this paradigm, edge servers continuously gather user data to enhance the performance of personalized models. This concurrent operation of model training and service provisioning on edge servers poses a challenge, given the limited computational resources available. Crucially, innovations in continual learning methodologies and real-time model adaptation will play a central role, enabling personalized models to adapt dynamically to user interactions and emerging data patterns. Moreover, striking a balance in resource allocation between model training and service provisioning is imperative for delivering high-quality services using the retrained models, while ensuring uninterrupted service delivery. Furthermore, the exploration of federated online knowledge distillation becomes crucial, allowing edge servers to collaboratively update models without compromising user privacy, marking a significant stride toward the future of adaptive and privacy-preserving RAG-enhanced service customization.

Extension of Uncertainties for User-Service Matching

With the increasing application of edge intelligence systems, 6G-assisted end-edge interactions for user-service matching is a widely studied area. Regarding coordinated scenarios, a key focus is on data privacy, where the enhancement of user experience by sending inputs from end devices to edge servers raises concerns about user data protection. Differential privacy (DP) serves as a theoretical method to address such an issue, offering global DP and local DP for edge servers and end devices, respectively. As for non-coordinated scenarios, the emphasis shifts to matching based on unknown preferences of both edge servers and end devices, a critical consideration not addressed in existing works that focus on one-side unknown preferences. For example, the response latency and energy consumption are always uncertain since the edge environment is dynamic, with fluctuations in network conditions, workload intensity, and device availability. Additionally, optimizing dynamic participation of edge servers and end devices during interactions holds promise for future scalability enhancements.

Conclusion

This article proposed a deployment framework for integrating RAG-enhanced generative services into future 6G networks. Within this framework, we investigated key challenges and explored promising directions pertaining to 6G-enabled data fusion, cloud-edge collaboration for knowledge base deployment, customization of RAG-enhanced services, and user-service matching. Our discussions encompass a fusion of mechanism design, online learning, and adaptive control, laying the groundwork for interdisciplinary innovations in the effective delivery of RAG-enhanced generative services. For future research, several unresolved issues deserve further investigation, such as the valuation of multi-modality data fusion, the interaction between knowledge base deployment and service customization, and the scheme designs for improved coordination between users and services.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Project 62301336 and Project 62301151, in part by the Shenzhen Institute of Artificial Intelligence

and Robotics for Society, and in part by the Guangdong Basic and Applied Basic Research Foundation under Project 2021A1515110949.

References

[1] C.-X. Wang et al., "On the road to 6G: Visions, requirements, key technologies and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2nd Quart., 2023.
[2] T. Q. Duong et al., "Machine learning-aided real-time optimized multibeam for 6G integrated satellite-terrestrial networks: Global coverage for mobile services," *IEEE Netw.*, vol. 37, no. 2, pp. 86–93, Mar. 2023.
[3] Y. Zhang et al., "Interdependent cell-free and cellular networks: Thinking the role of cell-free architecture for 6G," *IEEE Netw.*, early access, Nov. 23, 2024, doi: 10.1109/MNET.2023.3336218.
[4] B. Mao et al., "Security and privacy on 6G network edge: A survey," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1095–1127, 2nd Quart., 2023.
[5] M. Zhang et al., "Pricing fresh data," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1211–1225, May 2021.
[6] J. Chen, M. Li, and H. Xu, "Selling data to a machine learner: Pricing via costly signaling," in *Proc. ICML*, 2022, pp. 3336–3359.
[7] X. Huang, Z. Shao, and Y. Yang, "POTUS: Predictive online tuple scheduling for data stream processing systems," *IEEE Trans. Cloud Comput.*, vol. 10, no. 4, pp. 2863–2875, Oct. 2022.
[8] J. Li et al., "Toward reinforcement-learning-based intelligent network control in 6G networks," *IEEE Netw.*, vol. 37, no. 4, pp. 104–111, Jul. 2023.
[9] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control* (Studies in Systems, Decision and Control), vol. 325, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Eds., Cham, Switzerland: Springer, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-60990-0_12
[10] B. Kim et al., "Distilling the knowledge of large-scale generative models into retrieval models for efficient open-domain conversation," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2021, pp. 1–18.
[11] M. Kang et al., "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks," in *Proc. NeurIPS*, 2023, pp. 48573–48602.
[12] P. Han, X. Shi, and J. Huang, "FedAL: Black-box federated knowledge distillation enabled by adversarial learning," 2023, *arXiv:2311.16584*.
[13] R. Gupta, D. Reebadiya, and S. Tanwar, "6G-enabled edge intelligence for ultra-reliable low latency applications: Vision and mission," *Comput. Standards Interface*, vol. 77, Aug. 2021, Art. no. 103521.
[14] Y. Tang et al., "Green edge intelligence scheme for mobile keyboard emoji prediction," *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1888–1901, Feb. 2024.
[15] Y. Tang et al., "Learning-aided stable matching for switch-controller association in SDN systems," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 2738–2743.

Biographies

Xi Huang (Member, IEEE) (huangxi@cuhk.edu.cn) received the B.Eng. degree from Nanjing University, China, in 2014, and the Ph.D. degree from ShanghaiTech University, China, in 2021. He was a Visiting Researcher at the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in 2017. He is currently a Senior Research Fellow with the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), China, and a Distinguished Researcher of the Shenzhen Pengcheng Peacock Project. His research interests include multiagent collaborations for edge intelligence and data marketplaces.

Yinxu Tang (t.yinxu@wustl.edu) received the B.Eng. and master's degrees from ShanghaiTech University, Shanghai, China, in 2016 and 2020, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Washington University in St. Louis, USA. Her current research interests include explainable planning and scheduling.

Junling Li (Member, IEEE) (junlingli@seu.edu.cn) received the B.S. degree from Tianjin University, Tianjin, China, in 2013, the M.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016, and the Ph.D. degree from the University of Waterloo, Canada. She is currently an Associate Professor with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. Her research interests include AI-based channel modeling and digital twin online channel modeling.

Ning Zhang (Senior Member, IEEE) (ning.zhang@uwindsor.ca) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in January 2015. He was a Post-Doctoral Research Fellow at the University of Waterloo and the University of Toronto. He is currently an Associate Professor and the Canada Research Chair with the Department of Electrical and Computer Engineering, University of Windsor. His research interests include connected vehicles, mobile edge computing, wireless networking, and security. He is a Distinguished Lecturer of IEEE ComSoc and a Highly Cited Researcher (Web of Science). He also serves/served as the TPC/general chair for numerous conferences. He received several Best Paper Awards from conferences and journals, such as IEEE GLOBECOM, IEEE ICC, IEEE ICCC, IEEE WCSP, and *Journal of Communications and Information Networks*. He serves as the Vice Chair for the IEEE Technical Committee on Cognitive Networks and the IEEE Technical Committee on Big Data. He serves/served as an Associate Editor for IEEE Transactions on Mobile Computing, IEEE Communications Surveys and Tutorials, IEEE Internet of Things Journal, and IEEE Transactions on Cognitive Communications and Networking.

Xuemin (Sherman) Shen (Fellow, IEEE) (sshen@uwaterloo.ca) is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include network resource management, wireless network security, social networks, 5G and beyond, and vehicular ad hoc networks. He is a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, an Engineering Institute of Canada Fellow, and a Chinese Academy of Engineering Foreign Fellow. He received the President's Excellence in Research from the University of Waterloo in 2022, the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R. A. Fessenden Award from IEEE, Canada, in 2019, and the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019.