Hybrid RAG-Empowered Multimodal LLM for Secure Data Management in Internet of Medical Things: A Diffusion-Based Contract Approach

Cheng Su, *Graduate Student Member, IEEE*, Jinbo Wen, Jiawen Kang, *Senior Member, IEEE*, Yonghua Wang, *Senior Member, IEEE*, Yuanjia Su, Hudan Pan, Zishao Zhong, and M. Shamim Hossain, *Senior Member, IEEE*

***Abstract*—Secure data management and effective data sharing have become paramount in the rapidly evolving healthcare landscape, especially with the growing demand for the Internet of Medical Things (IoMT) integration. The advent of generative artificial intelligence (GenAI) has further elevated multimodal large language models (MLLMs) as essential tools for managing and optimizing healthcare data in IoMT. MLLMs can handle multimodal inputs and generate different kinds of data by utilizing large-scale training on massive multimodal datasets. Nevertheless, significant challenges remain in developing medical MLLMs, especially security and data freshness concerns, which impact the quality of MLLM outputs. To this end, this article proposes a hybrid Retrieval-Augmented Generation (RAG)-empowered medical MLLM framework for healthcare data management. The proposed framework enables secure data training by utilizing a hierarchical cross-chain design. Furthermore, it improves the output quality of MLLMs by using hybrid RAG that filters different unimodal RAG results using multimodal metrics and integrates these retrieval results as additional inputs for MLLMs. Furthermore, we utilize the age of information (AoI) to indirectly assess the influence of data freshness on MLLMs and apply contract theory to motivate healthcare data stakeholders to disseminate their current data, thereby alleviating information asymmetry in the data-sharing process. Finally, we employ a generative diffusion model-based deep reinforcement learning (DRL) technique to find the optimal contract for efficient data sharing. Numerical results show the effectiveness of the proposed approach in achieving secure and efficient healthcare data management.**

***Index Terms*—Contract theory, generative diffusion models (GDMs), healthcare data sharing, multimodal LLMs (MLLMs), retrieval-augmented generation (RAG).**

Received 24 November 2024; accepted 8 December 2024. Date of publication 23 December 2024; date of current version 9 May 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62102099, Grant U22A2054, and Grant 12326604; in part by the Science and Technology Research Cultivation Project of Chinese Medicine Guangdong Laboratory under Grant HQL2024PZ037; in part by the Guangzhou Basic Research Program under Grant 2023A04J1699; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515140137; and in part by the Researchers Supporting Project through King Saud University, Riyadh, Saudi Arabia, under Grant RSP2025R32. *(Corresponding author: Jiawen Kang.)*

Cheng Su, Yonghua Wang, and Yuanjia Su are with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: chengsu9251@163.com; wangyonghua@gdut.edu.cn; syj1216902331@163.com).

Jinbo Wen is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: jinbo1608@163.com).

Jiawen Kang is with the School of Automation, Guangdong University of Technology, Guangdong Basic Research Center of Excellence for Ecological Security and Green Development, Guangzhou 510006, China (e-mail: kavinkang@gdut.edu.cn).

Hudan Pan and Zishao Zhong are with the State Key Laboratory of Traditional Chinese Medicine Syndrome/The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangdong Provincial Hospital of Chinese Medicine, Guangdong Provincial Academy of Chinese Medical Sciences, Guangzhou 510006, China, and also with Chinese Medicine Guangdong Laboratory, Zhuhai 519000, China (e-mail: hdpan@gzucm.edu.cn; zhongzishao@gzucm.edu.cn).

M. Shamim Hossain is with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia (e-mail: mshossain@ksu.edu.sa).

Digital Object Identifier 10.1109/JIOT.2024.3521425

I. Introduction

THE HEALTHCARE system has seen rapid advancements with the integration of advanced technologies like cloud computing, the Internet of Things (IoT), and artificial intelligence (AI). These innovations have transformed the sector, giving rise to the Internet of Medical Things (IoMT), which is an interconnected network of medical devices and applications that collect and transmit vital healthcare data [1]. IoMT has not only paved the way for more intelligent and efficient healthcare systems, but also catalyzed the generation, storage, and analysis of vast amounts of big healthcare data [2], including omics data, clinical records, electronic health records, etc. [3]. Although the exponential growth in healthcare data volume holds the potential to revolutionize the healthcare industry by providing insights into patient care, disease patterns, and treatment effectiveness, it also requires sophisticated tools for its analysis and interpretation. Fortunately, generative artificial intelligence (GenAI) as a new branch of AI has emerged as a potent technology within IoT landscape [4], [5], enabling the effective analysis of vast datasets and generation of diverse content [6], [7]. In particular, GenAI enhances data management by analyzing complex patient records and treatment data, enabling more efficient sharing of critical healthcare information [8].

Large language models (LLMs), as a technological application of GenAI, can achieve general-purpose language generation and conventional natural language processing tasks, which hold the potential to significantly transform healthcare data management in IoMT [9]. With the integration of multimodal data into LLMs, patients can effectively comprehend many aspects of their physical health through multimodal LLMs (MLLMs) [10]. For example, the latest GPT-4, equipped with vision capabilities and exceptional performance in natural

language processing tasks, can be fine-tuned as a powerful guidance tool in the healthcare domain [11]. However, the sheer size and complexity of MLLMs necessitate efficient retrieval mechanisms to enhance their performance further. Retrieval-augmented generation (RAG) is a cutting-edge technique that boosts the reliability and accuracy of GenAI models by retrieving facts from an external knowledge base [12]. Moreover, RAG can capitalize on the similarity between the alignment vectors of the query to retrieve pertinent data, thereby enhancing user prompts by integrating relevantly retrieved data within the context, enabling MLLMs to generate accurate and contextually appropriate responses [6]. Thanks to the prominent capabilities of RAG, the integration of MLLMs and RAG has been widely used in various domains [13], [14].

Despite the advancements in RAG-empowered MLLMs, there are several persistent challenges in the application of these technologies for healthcare data management in IoMT.

1) Since healthcare data is normally multimodal and stored in different databases in a distributed manner, unimodal RAG using a single search manner, such as vector similarity search and keyword search [14], may not efficiently retrieve multimodal healthcare data to support LLM tasks that handle multiple modes.
2) The application of MLLMs in analyzing healthcare data poses significant security risks and privacy concerns [15]. Healthcare data is highly sensitive, and any breach or misuse can have severe consequences for patients and healthcare providers [2]. Thus, ensuring the confidentiality and integrity of healthcare data during MLLM processing is a critical concern.
3) Pretrained medical MLLMs can result in inaccurate inferences during task-specific fine-tuning due to biases in the dataset. Hence, incorporating fresh, high-quality healthcare data is crucial for fine-tuning MLLMs to avoid incorrect learning patterns [6].
4) Considering the problem of information asymmetry, healthcare data holders often have more data information, and appropriate incentive mechanisms need to be implemented to encourage healthcare data holders to provide accurate and up-to-date information, which is helpful to enhance the medical diagnostic quality of MLLMs empowered by RAG.

To address these challenges, we propose a hybrid RAG-empowered medical MLLM framework for healthcare data management in IoMT. Specifically, we allow participants to share data without the involvement of a central institution by implementing cross-chain techniques, which support secure and efficient data or asset transfers across multiple chains, effectively mitigating single-point-of-failure risks and enhancing overall security [16]. To enhance the diagnostic quality of MLLMs, we leverage hybrid multimodal RAG to refine the retrieval results further. Compared with RAG-empowered LLMs, we employ multimodal metrics to filter multiple unimodal RAG results and incorporate these results into MLLMs as additional inputs. Furthermore, we apply Age of Information (AoI) to evaluate the quality of healthcare data and utilize a contract theory model to encourage participants to share fresh data, thus coping with the information asymmetry of data sharing. Besides, considering the dynamic environment of data sharing, we use generative diffusion model (GDM)-based deep reinforcement learning (DRL) algorithms to efficiently find the optimal contract [17]. The key contributions of this article are summarized as follows.

1) We develop a novel hybrid RAG-empowered MLLM framework for healthcare data management in IoMT. This framework facilitates secure interactions between healthcare data holders and the MLLM service provider using a cross-chain system for secure healthcare data transmission, and MLLMs can improve their quality and complete specific tasks by employing hybrid RAG to retrieve multimodal healthcare data.
2) To optimize time-sensitive learning tasks within MLLM services, we apply AoI as a data freshness metric to indirectly assess healthcare data quality. Furthermore, we formulate a contract theory model to incentivize healthcare data holders to contribute high-quality healthcare data with small AoI, thus improving the inference performance of hybrid RAG-empowered MLLMs.
3) To tackle the high-dimensional complexity of the formulated problem, we employ GDM-based DRL algorithms to determine the optimal contract for efficient data sharing. Numerical results show the effectiveness of the proposed GDM-based scheme, showing a 20.35% performance improvement over DRL-based schemes, highlighting its superiority in this article.

The remainder of this article is structured as follows. Section II reviews the related work. Section III elaborates the proposed a hybrid RAG-empowered medical MLLM framework based on cross-chain technology to enhance data management in IoMT. In Section IV, we introduce a contract theory model to motivate healthcare data holders to provide high-quality healthcare data. In Section V, we present GDM-based DRL algorithms for optimal contract design. Section VI provides a performance analysis of our proposed approach. Finally, Section VII concludes this article. The main notations in this article are summarized in Table I.

II. Related Work

A. RAG-Empowered LLMs

RAG has incredible capabilities in enhancing the accuracy and reliability of LLM output by incorporating additional information sources, such as external knowledge bases, and augmenting user prompts with relevant retrieval data in context [12]. As a novel technique, RAG allows LLMs to bypass retraining, allowing access to the most up-to-date information to generate reliable output through retrieval-based generation [13]. Lewis et al. [12] introduced RAG, demonstrating its ability to improve the accuracy and relevance of generated text by incorporating retrieved documents into the generation process. Omrani et al. [18] proposed a hybrid RAG method that integrates sentence-window and parent–child approaches and demonstrated that the method outperforms the recent RAG techniques. Wen et al. [6] introduced a carbon emission optimization framework that integrates RAG and LLMs, significantly impacting GenAI efforts to reduce

TABLE I
KEY NOTATIONS OF THIS ARTICLE

Notation	Definition
t_{trans}	The transmission time of healthcare data
t_u	The time of completing consensus among blockchains
ℓ	The size of healthcare data
τ	The transmission rate of healthcare data between the health center and hospitals
$\overline{A}_m$	The average AoI for data sharing by data holder m
$\overline{A}_{max}$	The maximum permissible value for the AoI
R_k	Reward to the type-k healthcare data holders for the MLLM service provider
δ_k	The k-th type healthcare data holder
f_k	The update frequency of the type-k healthcare data holder
α	Overall zero-shot accuracy of MLLMs
S_k	The satisfaction function of the MLLM service provider obtained from the type-k healthcare data holder
β	The unit profit associated with type-k healthcare data holder
Q_k	The proportion of type-k healthcare data holder in healthcare industry
$\boldsymbol{\pi}_\omega$	Contract design policy with parameters ω
$\boldsymbol{\epsilon}_\omega$	Contract generation network with parameters ω
q_φ	Contract quality network with parameters φ
$\boldsymbol{\pi}_{\omega'}$	Target contract design policy with parameters ω'
$\boldsymbol{\epsilon}'_{\omega'}$	Target contract generation network with parameters ω'
$q'_{\varphi'}$	Target contract quality network with parameters φ'
Ψ^0	Optimal contract design

carbon emissions. Moreover, RAG is gradually emerging as a promising tool for healthcare applications, for example, it can optimize the interpretation of clinical guidelines for liver disease with the help of external medical knowledge [19]. In addition, Yuan et al. [20] retrieved similar image-text pairs based on image-text contrast similarity and utilized the retrieval attention module to blend the representation of images and questions with the retrieved images and texts, demonstrating effectiveness in simple biomedical visual question answering.

B. LLMs for Data Management

IoMT has significantly improved healthcare data management by enabling the seamless collection, transmission, and analysis of vast patient data through interconnected devices [8]. LLMs further enhance this process by offering advanced capabilities in processing unstructured data, extracting valuable insights, and supporting decision-making with greater efficiency and accuracy [9]. Recent advancements in LLMs have significantly contributed to data management, with numerous research efforts focusing on various aspects, such as data analysis, predictive modeling, and decision support systems. Some studies use the strong interpretative abilities of LLMs as agents to continuously improve data storage, data analysis, and additional areas [11], [21]. For instance, Achiam et al. [11] introduced GPT-4, which has proven exceptional capabilities in recognizing and generating human-like text, facilitating various data management tasks. Zhou et al. [22] presented an LLM-based database framework that leverages LLMs for automatic prompt generation and model fine-tuning, which performs highly effective in query rewriting and index tuning. Zhang et al. [23] introduced Data-Copilot, which is a data analysis agent capable of autonomously querying, processing, and visualizing vast amounts of data to meet various human needs. Existing works mainly focus on unimodal LLMs processing text data, while relatively insufficient research has been done on integrating multimodal data, such as text, images, and structured data. Addressing this gap could substantially boost the functionality of data management systems, especially in complex and data-intensive IoMT.

C. Contract Theory for Data Sharing

Contract theory is a branch of economics that studies how contractual arrangements can be designed to align incentives between parties with asymmetric information [24], and it has been widely used in wireless communication, AI, and other fields [16], [24]. In the context of data sharing, information asymmetry often arises because data holders possess more information about the data than data users. Contract theory can effectively incentivize data sharing by ensuring that both parties benefit from the exchange [16]. For example, Lim et al. [25] proposed a two-period incentive mechanism for healthcare applications, which considers the willingness to participate (WTP) of users and satisfies intertemporal incentive compatibility (IC). This dynamic contract design meets essential constraints and achieves higher profits than a uniform pricing scheme. In the context of a mobile AI-generated content network, Wen et al. [26] proposed an AoI-based contract theory model to incentivize the contribution of fresh data between mobile edge servers. Kang et al. [27] proposed an effective incentive mechanism. This mechanism integrates reputation and contract theory to motivate high-reputation mobile devices with high-quality data to engage in model learning in a federated learning scenario. In addition to the above work, several efforts have been made to develop contract theory models under prospect theory to facilitate user-centric sensing data sharing [16].

Despite these advancements, the application of diffusion-based contract theory for data sharing remains unexplored. While existing contract theory models are valuable, they often fail to address the intricate challenges associated with the dynamic and multifaceted nature of data sharing. Data sharing is not a simple transaction in many real-world contexts but a process spanning multiple stages and involving diverse participants [28]. Diffusion-based contract theory incorporates the spread and evolution of information over time and space, enabling a deeper understanding of incentives and behaviors among data holders and users [17]. By accounting for diffusion patterns, this approach allows for more accurate assessments of uncertainties and risks tied to data sharing. Consequently, diffusion-based contract models enable the design of adaptive and effective contractual arrangements that support efficient and sustainable data sharing.

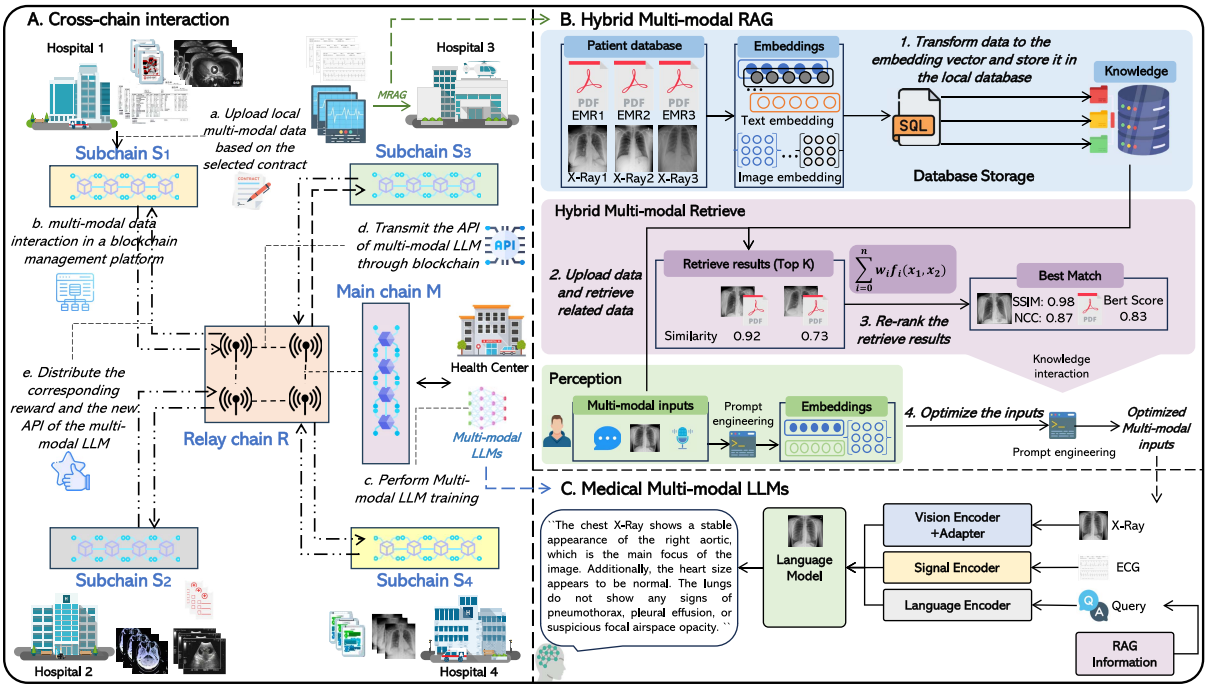


Fig. 1. Overview of the hybrid RAG-empowered medical MLLM framework for healthcare data management in IoMT. A) Shows the cross-chain interaction for secure healthcare data sharing. B) Depicts the processes of multimodal input optimization based on a hybrid multimodal RAG module. C) Presents the framework of MLLM inference based on the multimodal healthcare data.

III. Hybrid RAG-Empowered Medical MLLM Framework

This section describes our proposed hybrid RAG-empowered medical MLLM framework in IoMT. The detailed methodologies for cross-chain interaction in MLLM training and the utilization of hybrid RAG-empowered MLLM agents for data management are discussed in the following.

A. Cross-Chain Interaction in MLLM Training

In the health center, robust aggregate MLLMs are developed by training on massive, high-quality, multimodal healthcare data [20]. Due to privacy concerns and factors, such as patient willingness and incentive structures, hospitals may be reluctant to upload all healthcare data to a central health center [16]. Additionally, the significant computational power required for training MLLMs, combined with the lack of high-performance computing resources at most hospitals, limits the feasibility of using federated learning in this framework [29], [30]. These constraints necessitate alternative approaches for managing data and training models effectively.

In response to these multifaceted challenges, blockchain and cross-chain technologies have emerged as powerful solutions to facilitate secure and decentralized data sharing across healthcare networks. Blockchain technology ensures data integrity and transparency by providing an immutable ledger, while cross-chain technology enables seamless interoperability between diverse blockchain networks, allowing secure data and asset transactions across chains [8]. Recent studies further demonstrate the potential of cross-chain frameworks in significantly enhancing data security and enabling effective collaboration across distributed healthcare systems [8], [16]. To this end, we incorporate cross-chain technology to enable hospitals to securely upload sensitive healthcare data and conduct secure transactions with the health center, ensuring both data privacy and system efficiency.

As shown in Fig. 1, a robust health center utilizes a main chain to manage the comprehensive collection of healthcare data and model updates, and multiple subchains are employed to handle specific tasks from hospitals in diverse regions. These subchains utilize IoMT devices to collect real-time healthcare data from patients, such as temperature, heart rate, and blood pressure, enabling doctors to develop personalized treatment strategies based on a comprehensive analysis of multiple patient attributes [1], [6]. In addition to data collection, the subchains also manage MLLM configurations, and other workflow tasks to support effective healthcare delivery. The main chain ensures centralized oversight, while the subchains enable efficient and secure data management operations. Specifically, the main chain M sends data collection tasks to the relay chain R. When the subchains S_1, S_2, S_3, and S_4 receive the tasks, hospitals will upload local multimodal healthcare data based on the selected contracts [step 1) in Fig. 1]. Upon successful verification of cross-chain requests by the miners of the relay chain R, the relay chain R returns a readiness confirmation, allowing subchains to upload multimodal healthcare data [step 2) in Fig. 1] [16]. After transmitting the multimodal healthcare data to the main chain M, the health center initiates MLLM training [step 3) in Fig. 1]. Once MLLM training is completed, the health center and hospitals can access the MLLM application programming interface (API) or the weights file of MLLMs through the relay

chain R [step 4) in Fig. 1], and the health center rewards them with monetary compensation based on their data contribution [step 5) in Fig. 1] [16].

B. Hybrid RAG-Empowered MLLM Framework for Data Management in IoMT

Data management tasks in hospitals and the health center include data storage, analysis, and retrieval. When multimodal healthcare data is gathered into the subchains, MLLMs categorize the data by type and store it in the appropriate databases. As healthcare data is needed, MLLMs retrieve and analyze the data, thus meeting the specific requirements of data management tasks [22]. To further enhance the ability of MLLMs to analyze multimodal healthcare data, we design a hybrid multimodal RAG module [18], which is integrated with a data sharing mechanism inspired by contract theory, ensuring that the MLLM data analysis is conducted with high quality and strong privacy protection, allowing secure and effective handling of multimodal healthcare data. The workflow of hybrid RAG-empowered MLLMs for data management (shown in the right side of Fig. 1), in IoMT is presented as follows.

Step 1 (Store Multimodal Healthcare Data): Hospitals and the health center collect available multimodal healthcare data, convert them into vectors specific to each modality using an embedding model, and store these vectors in the local knowledge database with structured query language (SQL) tools [12].

Step 2 (Retrieve Multimodal Healthcare Data): When a task query is received, the hybrid multimodal RAG system uses the same embedding model from step 1 to convert the query into a vector. Then, the system calculates similarity scores between the task query vector and vectors within the knowledge database, retrieving and prioritizing the top K vectors that most closely match the task query [13].

Step 3 (Rerank the Retrieved Information): When all relevant information, particularly multimodal data, is fed directly into MLLMs, it can lead to information overload and may reduce attention to critical details due to the inclusion of irrelevant content [13]. To address this, the system further screens the results by applying our multimodal information similarity (MIS) metric, which is calculated by

$$\text{MIS} = \sum_{i=0}^{n} w_i f_i(x_1, x_2) \tag{1}$$

where $f_i(\cdot)$ represents the similarity measure function between the task query and the source data in the database and is determined freely according to the requirements of the specific task, x_1 and x_2 are the unimodal data corresponding to the task query and the source data in the database, respectively, and the weight factor w_i is used to characterize the proportion of each the similarity measure function $f_i(\cdot)$. When the results are reranked and filtered by MIS, the retrieved optimized healthcare information is then used to expand the context in the prompt.

Step 4 (Optimize the Multimodal Inputs): Upon completing the retrieval process, we employ prompt engineering based on zero-shot prompting technology to optimize and synthesize a coherent prompt that integrates the original multimodal task query with the retrieved healthcare data [13], [31]. By using "probability" as a control keyword, this refined prompt enhances the credibility of the entire prompt and improves the MLLM's capability to generate accurate and context-relevant responses.

Step 5 (Generate the Corresponding Content Based on the Inputs): Upon receiving the multimodal inputs, MLLMs connect each modal input to its respective pretrained encoder model, where a pretrained linear projection adapter is employed to unify all processed embeddings [20], [32]. This linear projection adapter, trained on 600K image-text pairs from PMC-15M, standardizes the embeddings, allowing pretrained LLMs to generate the corresponding content based on the inputs [13], [32].

In the generation of MLLMs, hybrid RAG enhances the generation quality by effectively incorporating relevant information from various sources. However, RAG does not improve the generalization ability of MLLMs, indicating that the ability of MLLMs to apply learned information to new and unseen contexts remains limited, limiting its overall learning ability. To improve the output quality of MLLMs, an essential way is to continuously incorporate new healthcare data for training. Thus, we propose an incentive mechanism to encourage data holders to share updated healthcare data.

IV. Problem Formulation

In this section, we begin by developing a metric for healthcare data quality, followed by the formulation of utility functions for both healthcare data holders and the MLLM service provider. Finally, we formulate a contract theory model to motivate healthcare data holders to contribute high-quality healthcare data.

The training of MLLMs relies heavily on a large volume of high-quality data [11]. Unfortunately, most healthcare data are stored in hospital databases in various regions. Without data sharing, these valuable resources remain untapped, hindering MLLM development. Furthermore, the effectiveness of MLLMs is directly influenced by the quality of the data used in training. Therefore, it is critical to implement an incentive mechanism that encourages hospitals to share healthcare data. Referring to [16], we consider there are multiple hospitals in diverse regions and a health center as an example. The health center acts as the MLLM service provider, and the hospitals in diverse regions serve as the healthcare data holders, represented by a set of $\mathcal{M} = \{1, \ldots, m, \ldots, M\}$. Initially, we propose a healthcare data quality metric through the AoI metric to assess the quality of healthcare data utilized for fine-tuning MLLMs. Subsequently, acting as the data task publisher, the MLLM service provider employs a contract theory model to encourage M healthcare data holders to engage in data sharing [26].

A. Healthcare Data Quality Metrics

AoI has gained broad acceptance as a metric for assessing data freshness, especially within wireless communication

TABLE II
PARAMETERS OF ACCURACY IN DIFFERENT DOMAINS OF LLAVA-MED-10K/60K

Model	Question Types		Domains					Overall
	Conversation	Description	CXR	MRI	Histology	Gross	CT	
(Question Count)	(143)	(50)	(37)	(38)	(44)	(34)	(40)	(193)
LLaVA-Med-10K	42.4	32.5	46.1	36.7	43.5	34.7	37.5	39.9
LLaVA-Med-60K	53.7	36.9	57.3	39.8	49.8	47.4	52.4	49.4

networks [33]. In this article, AoI is described as the duration between the data gathering at the hospital and the finalization of MLLM training. Lower AoI correlates with higher quality MLLM output for healthcare applications. As described in [26], we propose a healthcare data quality metric through AoI, which is relevant for scenarios involving periodic data updates.

To generalize, we define the size of healthcare data as ℓ (bytes) and the transmission rate between the health center and hospitals as τ (bytes per second). Hence, the transmission time of healthcare data is $t_{\text{trans}} = \ell/\tau$ [26]. Meanwhile, we denote t_u as the time of completing a consensus process among blockchains [16]. Therefore, we represent the length of a single time slot t as $t = t_{\text{trans}} + t_u$ [16]. To maintain data freshness, each healthcare data provider m periodically updates its healthcare data, with θ_m indicating the length of a single time slot in each update cycle. The refreshment of healthcare data happens in the initial time slot of the cycle. Referring to [34], the AoI for a data request made in the ith time slot is $(i+1)t$ for $i = 2, \ldots, \theta_m - 1$, and for requests initiated in the first or last time slot, the AoI is $2t$. Due to the Poisson process [16], [26], data requests are equally likely to occur in any time slot, with a probability of $1/\theta_m$. Therefore, the average AoI for data sharing by healthcare data holder m is given by

$$\overline{A}_m(\theta_m) = \frac{2}{\theta_m}(2t) + \sum_{i=2}^{\theta_{m-1}} \frac{(i+1)t}{\theta_m} = t\left(\frac{1}{\theta_m} + \frac{\theta_m}{2} + \frac{1}{2}\right). \quad (2)$$

Recognizing that a large AoI can degrade the quality of sensing data, we define the healthcare data quality metric $G(A_m)$ based on AoI as $G(\overline{A}_m) = \overline{A}_{\max}/\overline{A}_m$, where $\overline{A}_{\max}$ represents the maximum permissible value for the AoI. The healthcare data quality parameter plays an important role in the quality of MLLM services. Given that (2) is a convex function in relation to the update cycle θ_m, increasing θ_m results in a decrease in AoI [26]. Consequently, there is a tradeoff in managing AoI, which can be optimized by modifying the update cycle.

B. Healthcare Data Holder Utility

In the context of healthcare data sharing for MLLM services, the utility for each healthcare data holder m is the difference between the reward R_m and the cost C_m incurred by data sharing tasks, expressed as $U_m = R_m - C_m$ [16]. According to [35], the cost for healthcare data holder m is defined as $C_m = \xi_m f_m$ [26], with $f_m = (1/\theta_m)$ representing the update frequency and ξ_m denoting the cost of each update [16]. Thus, the utility of healthcare data holder m is

$$U_m = R_m - \xi_m f_m. \quad (3)$$

Due to information asymmetry, the MLLM service provider lacks precise knowledge of the update cost of each healthcare data holder. To address this, the MLLM service provider classifies data holders into discrete types by using statistical distributions derived from historical data, and its expected utility will be optimized [16]. By classifying M healthcare data holders into various types, we denote the kth type healthcare data holder as $\delta_k = 1/\xi_k$ and group them into a set $\mathcal{K} = \{\delta_k : 1 \le n \le K\}$, where a smaller update cost corresponds to a higher healthcare data holder type, and the healthcare data holder types are organized as $\delta_1 \le \delta_2 \le \cdots \le \delta_K$. Thus, the utility of the type-k healthcare data holder is given by

$$U_k(R_k, f_k) = R_k - \frac{f_k}{\delta_k}. \quad (4)$$

C. MLLM Service Provider Utility

Due to the quality of MLLMs' output being affected by healthcare data freshness, large AoI leads to poor output for MLLMs and reduces the satisfaction of the MLLM service provider. Referring to [29], the satisfaction function for the MLLM service provider, based on type-k healthcare data holders, is defined as

$$S_k = \alpha \log\left(G(\overline{A}_k) + 1\right) \quad (5)$$

where α is the overall zero-shot accuracy of MLLMs for various services. For example, the zero-shot accuracy of LLaVA-Med as a medical LLM across different domains is presented in Table II. Here, the value of α is determined by past experience when applied to various services [32].

Owing to information asymmetry, the MLLM service provider just knows the total count and type distributions of healthcare data holders, without detailed information about the type of each healthcare data holder [16], [26]. Thus, the expected utility of the MLLM service provider is calculated in the following manner [16], [35]:

$$U_s(\boldsymbol{f}, \boldsymbol{R}) = \sum_{k=1}^{K} Q_k(\beta S_k - R_k). \quad (6)$$

Here, $\beta > 0$ represents the unit profit associated with S_k, while Q_k is the probability that a healthcare data holder

is type-k, subject to the constraint that the sum of these probabilities equals 1, i.e., $\sum_{k=1}^{K} Q_k = 1$. Additionally, $\boldsymbol{R} = [R_k]_{1\times K}$ and $\boldsymbol{f} = [f_k]_{1\times K}$ represent the vectors of rewards and update frequencies for all K types of healthcare data holders, respectively.

D. Contract Formulation

To prevent rational healthcare data holders from supplying low-quality data in pursuit of higher rewards, a robust method is required to maintain MLLM service quality [26]. Given that contract theory is an economic tool for effectively designing incentive mechanisms under conditions of asymmetric information, we propose a contract theory model for the MLLM service provider. This model leverages contract theory to effectively motivate healthcare data holders to provide timely data updates, ensuring the reliability of MLLM services [16].

In this scenario, the MLLM service provider takes the lead in designing a set of contract items and offers them to K healthcare data holders. Based on its type, each healthcare data holder selects the most appropriate contract item, denoted by $\Psi_k = \{(f_k, R_k), k \in \mathcal{K}\}$, where f_k represents the update frequency for type-k healthcare data holders, and R_k is the reward given to type-k healthcare data holders as an incentive for its contribution. To guarantee that each healthcare data holder opts for the most advantageous contract item for its type, the designed contract must adhere to both (IC) and individual rationality (IR) constraints.

Definition 1 (IR): The contract item for a type-k healthcare data holder guarantees a nonnegative utility, formulated as

$$R_k - \frac{f_k}{\delta_k} \geq 0 \quad \forall k \in \mathcal{K}. \tag{7}$$

Definition 2 (IC): A healthcare data holder of type-k will choose the contract item (f_k, R_k) tailored to its type rather than any other contract item (f_i, R_i), $i \in \mathcal{K}$, and $i \neq k$, i.e.,

$$R_k - \frac{f_k}{\delta_k} \geq R_i - \frac{f_i}{\gamma_k} \quad \forall k, i \in \mathcal{K},\ k \neq i. \tag{8}$$

To maximize the expected utility of the MLLM service provider, the optimization problem can be formulated as

$$\begin{aligned} \max_{\boldsymbol{f},\boldsymbol{R}} \ & U_s(\boldsymbol{f}, \boldsymbol{R}) = \sum_{k=1}^{K} Q_k(\beta S_k - R_k) \\ \text{s.t.} \ & R_k - \frac{f_k}{\delta_k} \geq 0 \quad \forall k \in \mathcal{K} \\ & R_k - \frac{f_k}{\delta_k} \geq R_i - \frac{f_i}{\gamma_k} \ \forall k, i \in \mathcal{K},\ k \neq i \\ & f_k \geq 0, R_k \geq 0, \delta_k > 0 \quad \forall k \in \mathcal{K}. \end{aligned} \tag{9}$$

Traditional mathematical solutions often struggle to effectively adapt to the complexity and dynamic changes inherent in data-sharing environments [36]. In response, we leverage GDMs, a key component of GenAI, which excels not only in image generation but also in optimizing network performance [6], [37]. Building on similar approaches [28], [38], we employ GDMs as a more efficient solution for identifying optimal contracts. This approach capitalizes on the generative capabilities of GDMs to capture uncertainties and fluctuations in network conditions, allowing for more accurate identification of optimal contracts in real-time scenarios and effectively addressing the high-dimensional and intricate nature of the problem [17].

V. Generative Diffusion-Based Contract Design

In this section, we initially formulate the contract design between the MLLM service provider and healthcare data holders as a Markov decision process (MDP). Then, we present a GDM-based contract generation model to determine the optimal contract.

A. MDP Formulation

1) State Space: To find the optimal contract item, i.e., (f_k^*, R_k^*), $k \in \mathcal{K}$, the system first adds Gaussian noise to the initial contract sample. In the current diffusion round $t = 1, 2, \ldots, T$, the state space affecting the optimal contract design is defined as

$$\boldsymbol{s} \triangleq \{M, K, \overline{A}_{\max}, \mathcal{Q}, \mathcal{K}\} \tag{10}$$

where M and K are constant values, while $\overline{A}_{\max}$, $\mathcal{Q} = (Q_1, \ldots, Q_K)$, and $\mathcal{K} = (\delta_1, \ldots, \delta_K)$ are generated randomly in the current diffusion round t.

2) Action Space: As the MLLM service provider designs a contract Ψ to motivate healthcare data holders to provide high-quality healthcare data, the action $\boldsymbol{a}^t$ at round t is defined as

$$\boldsymbol{a}^t \triangleq \{\Psi^t\} \tag{11}$$

where $\Psi^t = \{(f_k^t, R_k^t), k \in \mathcal{K}\}$ determines the update frequency and reward for type-k healthcare data holders.

3) Immediate Reward: Following the action $\boldsymbol{a}^t$, the MLLM service provider achieves an immediate reward $r(\boldsymbol{s}, \boldsymbol{a}^t)$ aimed at maximizing the expected utility described in (6) while ensuring compliance with the IR (7) and IC (8) constraints. Thus, the reward function is defined as

$$r(\boldsymbol{s}, \boldsymbol{a}^t) = \begin{cases} U_s^t(\boldsymbol{f}, \boldsymbol{R}), & \text{if } \boldsymbol{a}^t \text{ satisfies (7) and (8)} \\ U_p, & \text{otherwise} \end{cases} \tag{12}$$

where $U_s^t(\boldsymbol{f}, \boldsymbol{R})$ denotes the expected utility of the MLLM service provider during round t and $U_p \leq 0$ serves as the penalty for violating either the IR or IC constraints.

B. GDMs for Optimal Contract Design

Compared with DRL algorithms that directly optimize model parameters [39], GDMs can enhance contract design by iterative denoising the initial distribution [28], [38]. The diffusion model network maps the environmental state to contract design, which constitutes the contract design policy represented as $\boldsymbol{\pi}_\omega(\boldsymbol{a}|\boldsymbol{s})$ with parameters ω. The policy $\boldsymbol{\pi}_\omega(\boldsymbol{a}|\boldsymbol{s})$ designed to generate an optimal contract over multiple time steps can be expressed as

$$\boldsymbol{\pi}_\omega(\boldsymbol{a}|\boldsymbol{s}) = p_\omega\left(\boldsymbol{a}^0, \ldots, \boldsymbol{a}^T|\boldsymbol{s}\right) = \mathcal{N}\left(\boldsymbol{a}^T; \mathbf{0}, \mathbf{I}\right) \prod_{t=1}^{T} p_\omega\left(\boldsymbol{a}^{t-1}|\boldsymbol{a}^t, \boldsymbol{s}^t\right). \quad (13)$$

Here, $\boldsymbol{\pi}_\omega(\cdot)$ represents the reverse process of the conditional diffusion model and $p_\omega(\boldsymbol{a}^{t-1}|\boldsymbol{a}^t, \boldsymbol{s}^t)$ is modeled as a Gaussian distribution $\mathcal{N}(\boldsymbol{a}^{t-1}; \boldsymbol{\mu}_\omega(\boldsymbol{a}^t, \boldsymbol{s}, t), \boldsymbol{\Sigma}_\omega(\boldsymbol{a}^t, \boldsymbol{s}, t))$, where the covariance matrix $\boldsymbol{\Sigma}_\omega(\boldsymbol{a}^t, \boldsymbol{s}, t)$ is formulated as [38]

$$\boldsymbol{\Sigma}_\omega\left(\boldsymbol{a}^t, \boldsymbol{s}, t\right) = \delta_t \mathbf{I} \quad (14)$$

where $\delta_t \in (0, 1)$ is a hyperparameter determined before model training and $\mathbf{I}$ is the identity matrix. Consequently, the mean $\boldsymbol{\mu}_\omega(\boldsymbol{a}^t, \boldsymbol{s}, t)$ can be given by [38]

$$\boldsymbol{\mu}_\omega\left(\boldsymbol{a}^t, \boldsymbol{s}, t\right) = \frac{1}{\sqrt{\chi_t}}\left(\boldsymbol{a}^t - \frac{\delta_t}{\sqrt{1-\bar{\chi}_t}}\boldsymbol{\epsilon}_\omega\left(\boldsymbol{a}^t, \boldsymbol{s}, t\right)\right) \quad (15)$$

where $\chi_t = 1 - \delta_t$, $\bar{\chi}_t = \prod_{i=0}^{t} \delta_j$, and $\boldsymbol{\epsilon}_\omega$ denotes the contract generation network. We first sample $\boldsymbol{a}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and then sample from the reverse diffusion chain parameterized by ω, which is given by [38]

$$\boldsymbol{a}^{t-1}|\boldsymbol{a}^t = \frac{\boldsymbol{a}^t}{\sqrt{\chi_t}} - \frac{\delta_t}{\sqrt{\chi_t(1-\bar{\chi}_t)}}\boldsymbol{\epsilon}_\omega\left(\boldsymbol{a}^t, \boldsymbol{s}, t\right) + \sqrt{\delta_t}\boldsymbol{\epsilon}. \quad (16)$$

Referring to [28], [38], we effectively train the contract design policy $\boldsymbol{\pi}_\omega$ to enhance the training quality of the contract generation network $\boldsymbol{\epsilon}_\omega$. Additionally, inspired by the concept of the Q-function [40], we introduce a contract quality network $q_\varphi(\boldsymbol{s}, \Psi)$. The training of the contract quality network utilizes the double Q-learning technique to minimize the Bellman operator, involving two critic networks $q_{\varphi_1}, q_{\varphi_2}$ and the corresponding target critic networks $q_{\varphi_1'}, q_{\varphi_2'}$. We define $q_\varphi = \min\{q_{\varphi_1}, q_{\varphi_2}\}$, and the optimal contract design policy that maximizes the expected cumulative utility of the client is expressed as [28]

$$\boldsymbol{\pi} = \arg\max_{\boldsymbol{\pi}_\omega} \mathbb{E}\left[\sum_{z=0}^{Z} \gamma^z (r(\boldsymbol{s}_z, \boldsymbol{a}_z) - \varsigma \boldsymbol{\pi}_\omega(\boldsymbol{s}_z) \log \boldsymbol{\pi}_\omega(\boldsymbol{s}_z))\right] \quad (17)$$

where γ represents the discount factor, $\boldsymbol{a}_z$ represents the action in the training step z, and ς represents the temperature coefficient controlling the strength of the entropy.

We define the target policy as $\boldsymbol{\pi}_{\omega'}$, and the optimization of φ_i for $i = 1, 2$ is performed by minimizing the following objective function [28]:

$$\mathbb{E}_{(\boldsymbol{s}_z, \boldsymbol{a}_z, \boldsymbol{s}_{z+1}, r_z)\sim\mathcal{B}_z}\left[\sum_{l=1,2} (r(\boldsymbol{s}_z, \boldsymbol{a}_z) - q_{\varphi_l}(\boldsymbol{s}_z, \boldsymbol{a}_z) + \gamma^z (1 - d_{z+1}) \boldsymbol{\pi}_{\omega'}(\boldsymbol{s}_{z+1}) q'_{\varphi'}(\boldsymbol{s}_{z+1}))^2\right] \quad (18)$$

where $\mathcal{B}_z$ is a mini-batch of transitions sampled from the experience replay memory $\mathcal{D}$ in the training step z and d_{z+1} is a 0-1 variable denoting the terminated flag.

The pseudo-code of the proposed GDM-based contract generation scheme is shown in Algorithm 1, which consists of three phases, and its computational complexity is $\mathcal{O}(|\omega|+|\varphi|+E_{\max}Z_{\max}(T|\omega| + |\varphi|))$. In the proposed GDM-based contract generation scheme, denoising techniques are employed to generate optimal contract designs [28], [38]. By integrating exploration noise into the contract design and executing it, the process accumulates exploration experience, contributing to the enhancement of contract quality.

Algorithm 1: GDM-Based Optimal Contract Design

Input: GDM's hyperparameters, e.g., diffusion step T, discount factor γ, and exploration noise ε.
Output: The optimal contract design $\boldsymbol{a}^0$.

```
###### Phase 1: Initialization
Initialize replay buffer D, contract generation network ε_ω, contract quality network q_φ, target contract generation network ε'_ω', target contract quality network q'_φ'.
###### Phase 2: Training
for Episode e = 1 to E_max do
    Initialize a random process N to facilitate contract design exploration.
    for Step z = 1 to Z_max do
        Observe the current environment s_z.
        Set a_z^T as Gaussian noise and generate contract design a_z^0 by denoising a_z^T based on (16).
        Execute contract design a_z^0 and observe the reward r_z (12).
        Store record (s_z, a_z^0, r_z, s_{z+1}) into replay buffer D.
        Sample a random mini-batch of N records (s_i, a_i^0, r_i, s_{i+1}) from replay buffer D.
        Update the contract quality network by minimizing (18).
        Update the contract generation network by computing the policy gradient (17).
        Update the target networks: ω' ← ηω + (1 − η)ω', φ' ← ηφ + (1 − η)φ'.
    end
end
return The trained contract generation network ε_ω.
###### Phase 3: Inference
Input the environment vector s (10).
Generate the optimal contract design a^0 based on (16).
return a^0 = {(f_k^*, R_k^*), k ∈ K}.
```

VI. Numerical Results

In this section, we present extensive experiments to evaluate the performance of the proposed hybrid RAG-empowered MLLM framework for healthcare analysis and the effectiveness of the proposed incentive mechanism. MLLM inference uses Python 3.10.14 on an Intel Xeon Gold 6133 CPU and an NVIDIA RTX A6000 GPU. For the implementation of GDM-based DRL algorithms, the primary parameter settings are detailed in Table III, with experiments run on an NVIDIA GeForce RTX A6000 server GPU using CUDA 11.8.

A. Case Study of Hybrid RAG-Empowered Medical MLLMs

We simulate a prototype of the hybrid RAG-empowered MMLMs with the support of LLaVA-Med [32] and llamaindex.[1] As illustrated in Fig. 2, we present two examples to

[1] https://docs.llamaindex.ai/

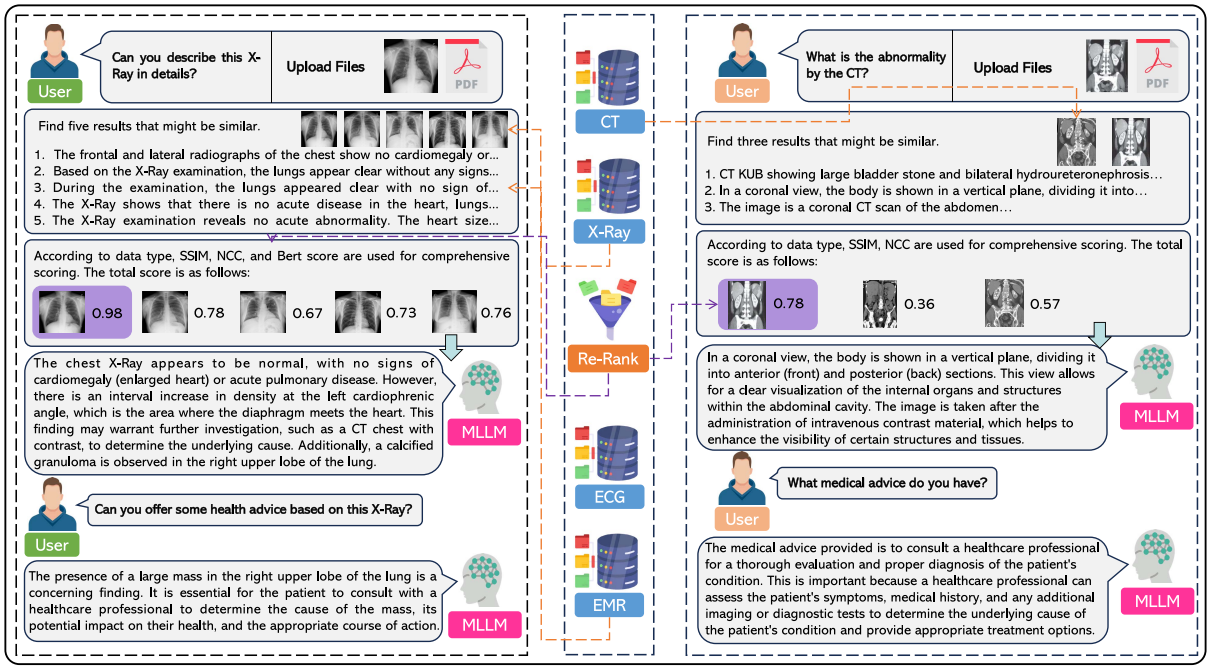


Fig. 2. Real case study of hybrid RAG-empowered medical MLLMs. In the proposed hybrid RAG-empowered medical MLLM, the RAG initially retrieves healthcare data using unimodal methods. Next, we rerank the information using metrics, such as the structural similarity index measure [41], normalized cross-correlation, and BERT score [42]. The detailed information is then combined with the task query and input into the MLLM to generate results.

TABLE III
KEY HYPERPARAMETERS IN THE SIMULATION

Hyperparameters	Setting
Learning rate of the contract generation network	1×10^{-6}
Learning rate of the contract quality network	1×10^{-6}
Soft target update parameter τ	0.005
Exploration noise ε	0.01
Batch Size N	512
Denoising steps for the diffusion model T	5
Maximum capacity of the replay buffer $\|\mathcal{D}\|$	10^6

demonstrate the functionality and application of our framework. Upon receiving a task query with multimodal healthcare data, the framework first retrieves the corresponding data from the respective modal database. It performs a preliminary screening of K results based on the cosine similarity between vectors. Next, hybrid RAG further refines these results using the MIS metric to identify the best matches, which are then used as inputs to MLLMs. Finally, the MLLMs process all this information to provide diagnostic outputs and personalized services according to the query.

We apply the criteria of responsive artificial intelligence (RAI)[2] to assess whether the outputs of MLLMs present potential risks related to morality, bias, and ethics. Additionally, we also assess the relationship between the output of MLLMs and task query with the semantic similarity (SS) [43], which reflects the diagnosis quality and serves as a crucial indicator for measuring the output of MLLMs. Due to the lack of evaluation benchmarks, we integrate LLM evaluators [44] with prompt engineering techniques [31] to measure the quality

[2]https://www.microsoft.com/en-us/ai/responsible-ai

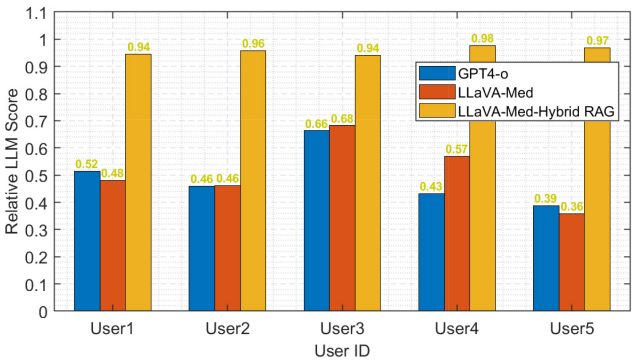


Fig. 3. Performance comparison between the proposed framework under different healthcare data cases. Note that the initial two users provide conditional healthcare data cases, while the subsequent three users provide normal healthcare data cases.

of data analysis for MLLMs under the method of GPT4-o, LLaVA, and LLaVA with hybrid RAG. The scoring is normalized to a range of [0, 1]. Higher scores denote greater reliability, lack of bias in the MLLM output, and a strong correlation to task query information. In contrast, lower scores signify a substantial gap from the anticipated results. Finally, we combine the RAI evaluation and SS into a unified score, known as the relative LLM score ζ [45], which is calculated using the formula

$$\zeta = \lambda \cdot \mathrm{RAI} + \nu \cdot \mathrm{SS} \tag{19}$$

where λ and ν are the weighting factors for RAI and SS, respectively. We assign equal weights in our approach by setting $\lambda = 0.5$ and $\nu = 0.5$.

As shown in Fig. 3, we present the performance comparison between the proposed framework under different healthcare

TABLE IV
PERFORMANCE OF DIFFERENT METHODS

Methods	RAI	SS	Relative LLM scores
GPT4-o	0.55	0.43	0.49
LLAVA-Med	0.54	0.48	0.51
LLAVA-Med-Hybrid RAG	**0.98**	**0.93**	**0.96**

data cases. Our findings indicate that hybrid RAG enables LLaVA-Med to consistently score above 0.9, particularly in *X*-Ray cases from Users 1 and 2 with known etiologies, maintaining high-quality answers and stability. In contrast, other MLLMs exhibit reduced output quality due to the interference of disease factors. In the scenarios involving Users 3 and 4, who are normal without specific causes, MLLMs achieve high scores and deliver reasonable judgments. However, in the case of User 5, who is normal but has an *X*-ray that a doctor can easily misjudge, other MLLMs exhibit a higher misjudgment rate. In contrast, hybrid RAG continues to produce high-quality outputs by matching similar disease conditions. These cases illustrate that the data retrieved by hybrid RAG provides valuable information for answering questions. We summarize all scores in Table IV, which clearly demonstrates that hybrid RAG helps LLAVA-Med maintain consistently high scores, showcasing its strong performance across different scenarios. This indicates that hybrid RAG effectively considers the quality of retrieved information by utilizing the features of multimodal data, including images and texts. The retrieved relevant healthcare data can aid MLLMs through contextual relationships, allowing MLLMs to deliver reliable and robust outputs owing to their powerful contextual learning capabilities.

B. Performance of GDM-Based Contract Theory Approach

In the proposed contract model, we employ an on-policy GDM algorithm within a double actor-critic framework for optimal contract design, and the specific settings of training hyperparameters are shown in Table III. In our setup, we consider 10 healthcare data holders divided into two types, with $M = 10$ and $K = 2$. For the two types of healthcare data holders θ_1 and θ_2, values are randomly sampled from the intervals $[1, 6]$ and $[13, 18]$, respectively. Additionally, the maximum tolerance of AoI $\overline{A}_{\max}$ is sampled randomly within the range of $[30, 60]$. For the utility of the MLLM service provider, the parameters α, β, and t are set to 39.9, 10, and 2, respectively, and Q_1 and Q_2 are randomly generated according to the Dirichlet distribution [46].

First, we compare our proposed contract-based incentive mechanism, which operates under information asymmetry, with other methods: a contract-based mechanism with complete information, a greedy scheme, and a random scheme. As illustrated in Fig. 4, we can find that our proposed contract scheme consistently outperforms the greedy and random schemes. However, for identical parameter settings, the contract-based mechanism with complete information yields higher performance than our model. This result highlights the disadvantage of information asymmetry, as the MLLM service provider gains fewer benefits without precise knowledge of the types of healthcare data holders. Although a complete information scenario allows the MLLM service provider to offer the optimal contract items to healthcare data holders by knowing their exact types, it is not a realistic environment. In practice, even with complete information, a rational healthcare data holder may provide misleading information to manipulate rewards, ultimately reducing the subjective utility for the MLLM service provider. Thus, our proposed contract model, which handles asymmetric information, proves to be more reliable and practical, achieving the highest utility in real-world scenarios.

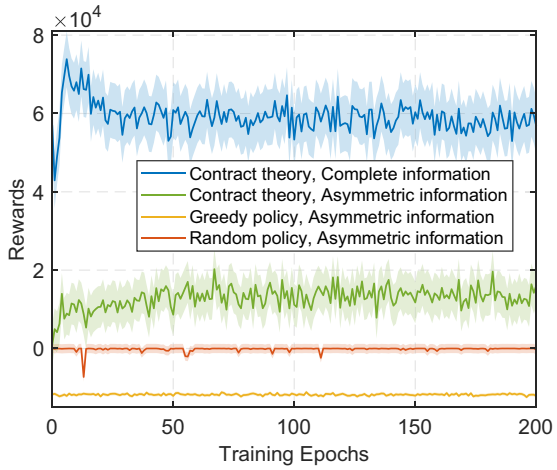


Fig. 4. Reward comparison of our scheme with other schemes, i.e., contract-based incentive mechanism with complete information, greedy, and random.

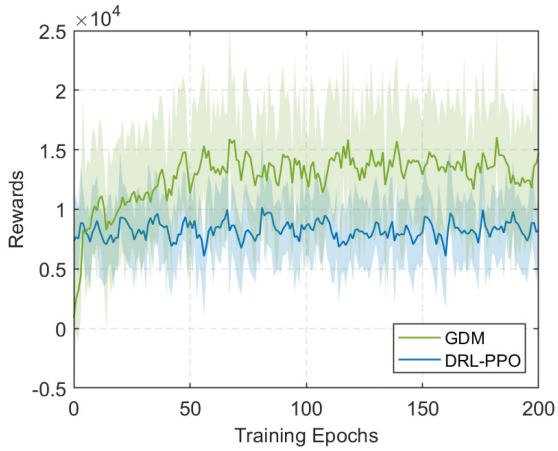


Fig. 5. Performance comparison between the GDM and DRL-PPO in optimal contract design.

In Fig. 5, we compare the performance of the GDM and DRL with proximal policy optimization (DRL-PPO) in optimal contract design. Both models are capable of continuously acquiring rewards in complex and variable environments until convergence. Notably, the final test rewards of GDMs are significantly higher than those for DRL-PPO under identical parameter settings, allowing the MLLM service provider to consistently secure greater utilities. This is attributed to

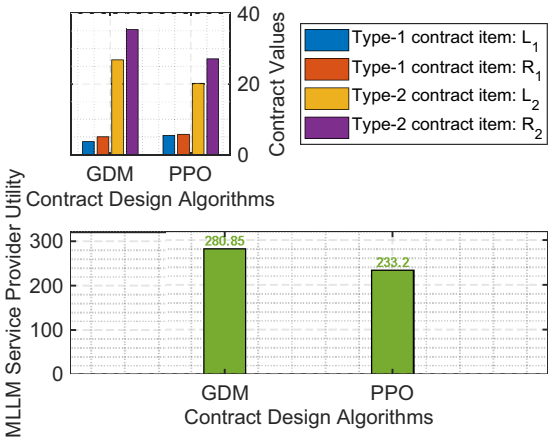


Fig. 6. Optimal contracts designed by the GDM and DRL-PPO.

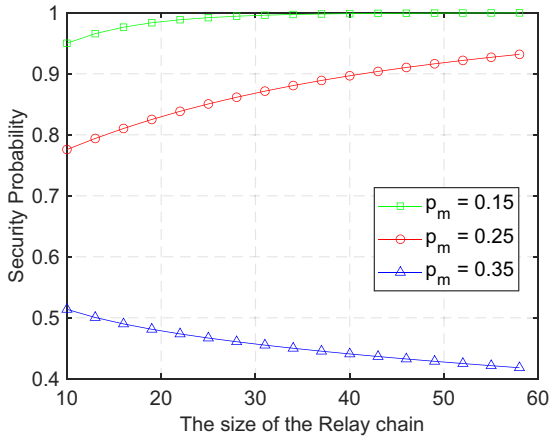


Fig. 7. Security probability at varying malicious miner probabilities.

the fine-grained policy adjustments during the diffusion process, which effectively reduces the impact of randomness and noise [46]. Additionally, exploration through diffusion enhances the flexibility and robustness of the contract design policy, preventing it from falling into suboptimal solutions. Consequently, this superior performance demonstrates the ability of GDMs to capture intricate patterns and connections among environmental observations, and it can effectively reduce the complexity of the relationship between healthcare data holders and the MLLM service provider.

In Fig. 6, we present the optimal contracts designed by the GDM and DRL-PPO. Given the environmental state, the GDM-based model, enhanced by exploration during the denoising process, produces a contract design that delivers a utility value of 280.85 for the MLLM service provider, which is higher than the 233.2 achieved by DRL-PPO. This advantage arises from GDM's capability to generate near-optimal contracts. Additionally, as the type of healthcare data holder increases, the rewards they receive also rise. However, DRL-PPO shows consistent variables for the type-1 healthcare data holders, indicating a tendency toward local optimal solutions, which may not align with global interests. Overall, this numerical analysis highlights the practical feasibility and superior performance of the proposed GDM-based scheme.

C. Secure Block Verification Performance Analysis

To assess the security of the blockchain system, we evaluate the reputation value of hospitals. Each hospital's associated subchain generates a block and broadcasts it to the relay chain for validation. If validated, the relay chain submits the block to the main chain linked to the health center. The health center then rewards each hospital according to their actions, applying a reputation-based bonus and penalty system.

As illustrated in Fig. 7, we use the practical Byzantine fault tolerance (PBFT) consensus algorithm to assess the security performance of the blockchain system. We consider that the subchains operate reliably and model the relay chain's security performance as a random sampling problem with two potential outcomes, i.e., malicious delegates and well-behaved delegates [47]. When the number of malicious delegates is no greater than $(N-1)/3$, where N represents the total number of delegates, the block verification process remains accurate [47]. Therefore, the probability of secure consensus, denoted as $P_{\text{safety}} = \sum_{z=0}^{\lfloor N/3 \rfloor} \binom{N}{z} p_m^z (1-p_m)^{N-z}$, depends on p_m that represents the probability of a delegate being malicious. Fig. 7 shows that as the size of the relay chain increases, the security probability also rises, regardless of the likelihood of malicious delegates. This improvement is due to the larger number of well-behaved delegates involved in block validation, which strengthens security in the consensus process. Thus, the proposed blockchain system with the PBFT consensus algorithm supports secure and reliable data sharing by ensuring robust block verification.

VII. Conclusion

In this article, we have studied the service quality issues of MLLMs and the design of incentive mechanisms for healthcare data management. We have proposed a hybrid RAG-empowered medical MLLM framework based on cross-chain technologies to enhance healthcare data management in IoMT. Specifically, we have utilized a cross-chain structure comprising a main chain and multiple subchains to ensure the security of healthcare data. Additionally, we have applied hybrid RAG with multimodal information similarity metrics to retrieve similar healthcare data, thereby improving the quality of MLLM services. Then, we have applied AoI to quantify healthcare data quality indirectly and utilized contract theory to incentivize healthcare data holders to contribute high-quality healthcare data with small AoI, thus enhancing the quality of MLLM services. Furthermore, we have employed GDMs to generate the optimal contracts for efficient data sharing. Finally, numerical results show the effectiveness and reliability of our proposed framework and incentive mechanism. For future work, we aim to enhance our framework's performance by integrating additional characteristics of multimodal healthcare data and developing a multidimensional contract model to address the complexities of IoMT environments better.

REFERENCES

[1] C. Huang, J. Wang, S. Wang, and Y. Zhang, "Internet of Medical Things: A systematic review," *Neurocomputing*, vol. 557, Nov. 2023, Art. no. 126719.

[2] M. Isgut, L. Gloster, K. Choi, J. Venugopalan, and M. D. Wang, "Systematic review of advanced AI methods for improving healthcare data quality in post COVID-19 era," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 53–69, 2023.

[3] N. S. Gupta and P. Kumar, "Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine," *Comput. Biol. Med.*, vol. 162, Aug. 2023, Art. no. 107051.

[4] P. Li et al., "Filling the missing: Exploring generative AI for enhanced federated learning over heterogeneous mobile edge devices," *IEEE Trans. Mobile Comput.*, vol. 23, no. 10, pp. 10001–10015, Oct. 2024.

[5] X. Tang, Q. Chen, R. Yu, and X. Li, "Digital twin-empowered task assignment in aerial MEC network: A resource coalition cooperation approach with generative model," 2024, *arXiv:2405.01555*.

[6] J. Wen et al., "Generative AI for low-carbon Artificial Intelligence of Things," 2024, *arXiv:2404.18077*.

[7] J. Chen, Y. Shi, C. Yi, H. Du, J. Kang, and D. Niyato, "Generative AI-driven human digital twin in IoT-healthcare: A comprehensive survey," 2024, *arXiv:2401.13699*.

[8] A. Bisht, A. K. Das, D. Niyato, and Y. Park, "Efficient personal-health-records sharing in Internet of Medical Things using searchable symmetric encryption, blockchain, and IPFS," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 2225–2244, 2023.

[9] X. Yuan, W. Kong, Z. Luo, and M. Xu, "Efficient inference offloading for mixture-of-experts large language models in Internet of Medical Things," *Electronics*, vol. 13, no. 11, p. 2077, 2024.

[10] B. Meskó, "The impact of multimodal large language models on health care's future," *J. Med. Internet Res.*, vol. 25, Nov. 2023, Art. no. e52865.

[11] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[12] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[13] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," 2023, *arXiv:2312.10997*.

[14] R. Zhang et al., "Interactive AI with retrieval-augmented generation for next generation networking," *IEEE Netw.*, vol. 38, no. 6, pp. 414–424, Nov. 2024.

[15] F. Ullah, G. Srivastava, H. Xiao, S. Ullah, J. C.-W. Lin, and Y. Zhao, "A scalable federated learning approach for collaborative smart healthcare systems with intermittent clients using medical imaging," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 6, pp. 3293–3304, Jun. 2024.

[16] J. Kang et al., "Blockchain-empowered federated learning for healthcare metaverses: User-centric incentive mechanism with optimal data freshness," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 1, pp. 348–362, Feb. 2024.

[17] H. Du et al., "Diffusion-based reinforcement learning for edge-enabled AI-generated content services," *IEEE Trans. Mobile Comput.*, vol. 23, no. 9, pp. 8902–8918, Sep. 2024.

[18] P. Omrani, A. Hosseini, K. Hooshanfar, Z. Ebrahimian, R. Toosi, and M. A. Akhaee, "Hybrid retrieval-augmented generation approach for LLMs query response enhancement," in *Proc. 10th Int. Conf. Web Res. (ICWR)*, 2024, pp. 22–26.

[19] S. Kresevic, M. Giuffrè, M. Ajcevic, A. Accardo, L. S. Crocè, and D. L. Shung, "Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework," *NPJ Digit. Med.*, vol. 7, no. 1, p. 102, 2024.

[20] Z. Yuan et al., "RAMM: Retrieval-augmented biomedical visual question answering with multi-modal pre-training," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 547–556.

[21] W. Chen, X. Ma, X. Wang, and W. W. Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," 2022, *arXiv:2211.12588*.

[22] X. Zhou, Z. Sun, and G. Li, "DB-GPT: Large language model meets database," *Data Sci. Eng.*, vol. 9, no. 1, pp. 102–111, 2024.

[23] W. Zhang, Y. Shen, W. Lu, and Y. Zhuang, "Data-Copilot: Bridging billions of data and humans with autonomous workflow," 2023, *arXiv:2306.07209*.

[24] Z. Hou, H. Chen, Y. Li, and B. Vucetic, "Incentive mechanism design for wireless energy harvesting-based Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2620–2632, Aug. 2018.

[25] W. Y. B. Lim et al., "Dynamic contract design for federated learning in smart healthcare applications," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16853–16862, Dec. 2021.

[26] J. Wen et al., "Freshness-aware incentive mechanism for mobile AI-Generated Content (AIGC) networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2023, pp. 1–6.

[27] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019.

[28] J. Wen et al., "Diffusion-model-based incentive mechanism with prospect theory for edge AIGC services in 6G IoT," *IEEE Internet Things J.*, vol. 11, no. 21, pp. 34187–34201, Nov. 2024.

[29] M. Xu et al., "Cached model-as-a-resource: Provisioning large language model agents for edge intelligence in space-air-ground integrated networks," 2024, *arXiv:2403.05826*.

[30] D. Yang et al., "DetFed: Dynamic resource scheduling for deterministic federated learning over time-sensitive networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 5162–5178, May 2024.

[31] Y. Liu et al., "Optimizing mobile-edge AI-generated everything (AIGX) services by prompt engineering: Fundamental, framework, and case study," *IEEE Netw.*, vol. 38, no. 5, pp. 220–228, Sep. 2024.

[32] C. Li et al., "LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems*, vol. 36. Red Hook, NY, USA: Curran Assoc., Inc., 2023, pp. 28541–28564.

[33] W. Y. B. Lim et al., "When information freshness meets service latency in federated learning: A task-aware incentive scheme for smart industries," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 457–466, Jan. 2022.

[34] S. Zhang, J. Li, H. Luo, J. Gao, L. Zhao, and X. S. Shen, "Towards fresh and low-latency content delivery in vehicular networks: An edge caching aspect," in *Proc. 10th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2018, pp. 1–6.

[35] X. Zhou, W. Wang, N. U. Hassan, C. Yuen, and D. Niyato, "Towards small AoI and low latency via operator content platform: A contract theory-based pricing," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 366–378, Jan. 2022.

[36] Q. Ye, W. Shi, K. Qu, H. He, W. Zhuang, and X. Shen, "Joint RAN slicing and computation offloading for autonomous vehicular networks: A learning-assisted hierarchical approach," *IEEE Open J. Veh. Technol.*, vol. 2, pp. 272–288, 2021.

[37] C. Su et al., "Privacy-preserving pseudonym schemes for personalized 3-D avatars in mobile social metaverses," in *Proc. 6th Int. Conf. Electron. Commun., Netw. Comput. Technol. (ECNCT)*, 2024, pp. 375–380.

[38] H. Du, J. Wang, D. Niyato, J. Kang, Z. Xiong, and D. I. Kim, "AI-generated incentive mechanism and full-duplex semantic communications for information sharing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 9, pp. 2981–2997, Sep. 2023.

[39] X. Tang et al., "Digital-twin-assisted task assignment in multi-UAV systems: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 10, no. 17, pp. 15362–15375, Sep. 2023.

[40] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double *Q*-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.

[41] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2020.

[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTscore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.

[43] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—A survey," *ACM Comput. Surveys*, vol. 54, no. 2, pp. 1–37, 2021.

[44] J. Wang et al., "Is ChatGPT a good NLG evaluator? A preliminary study," in *Proc. 4th New Front. Summarization Workshop*, 2023, pp. 1–11.

[45] Y. Huang et al., "Large language models for networking: Applications, enabling techniques, and challenges," *IEEE Netw.*, early access, Jul. 30, 2024, doi: 10.1109/MNET.2024.3435752.

[46] J. Wen et al., "From generative AI to generative Internet of Things: Fundamentals, framework, and outlooks," *IEEE Internet Things Mag.*, vol. 7, no. 3, pp. 30–37, May 2024.

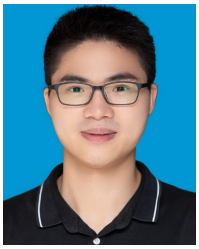
[47] Y. Zhong et al., "Blockchain-assisted twin migration for vehicular metaverses: A game theory approach," *Trans. Emerg. Telecommun. Technol.*, vol. 34, no. 12, 2023, Art. no. e4856.

Cheng Su (Graduate Student Member, IEEE) received the B.Eng. degree from Guangdong University of Technology, Guangzhou, China, in 2023, where he is currently pursuing the M.S. degree with the School of Automation.

His research interests include generative AI, AI for Internet of Medical Things, and metaverse.

Jinbo Wen received the B.Eng. degree from Guangdong University of Technology, Guangzhou, China, in 2023. He is currently pursuing the M.S. degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include generative AI, blockchain, edge intelligence, and metaverse.



Jiawen Kang (Senior Member, IEEE) received the Ph.D. degree from Guangdong University of Technology, Guangzhou, China, in 2018.

He has been a Postdoctoral Fellow with Nanyang Technological University, Singapore, from 2018 to 2021. He currently is a Full Professor with Guangdong University of Technology. His research interests mainly focus on generative AI, blockchain, security, and privacy protection in wireless communications and networking.

Yonghua Wang (Senior Member, IEEE) received the B.S. degree in electrical engineering and automation from Hebei University of Technology, Tianjin, China, in 2001, the M.S. degree in control theory and control engineering from Guangdong University of Technology, Guangzhou, China, in 2006, and the Ph.D. degree in communication and information system from Sun Yat-sen University, Guangzhou, in 2009.

He is currently an Associate Professor with the School of Automation, Guangdong University of Technology. His current research interests include machine learning, intelligent control, and cognitive radio networks.

Yuanjia Su received the B.Eng. degree from Shandong Jianzhu University, Jinan, China, in 2021. He is currently pursuing the M.S. degree with the School of Automation, Guangdong University of Technology, Guangzhou, China. His research interests include blockchain, generative AI, and federated learning.

Hudan Pan received the Ph.D. degree from Macau University of Science and Technology, Macau, China, in 2017.

She has been a Postdoctoral Fellow with Macau University of Science and Technology from 2017 to 2019. She is currently a Professor with The State Key Laboratory of Traditional Chinese Medicine Syndrome/The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China. Her research interests mainly focus on AI for Internet of Medical Things, and the treatment of rheumatic diseases with Chinese medicine.

Zishao Zhong received the Ph.D. degree from Guangzhou University of Chinese Medicine, Guangzhou, China in 2018.

He has been a Postdoctoral Fellow with Tongji University, Shanghai, China, from 2020 to 2022. He is currently an Associate Chief Physician with The State Key Laboratory of Traditional Chinese Medicine Syndrome/The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China. His research interests mainly focus on AI for Internet of Medical Things, and the integration of Chinese and western medicine in gastroenterology.

M. Shamim Hossain (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Ottawa, Ottawa, ON, Canada, in 2009.

He is currently a Professor with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He is also an Adjunct Professor with the School of Electrical Engineering and Computer Science, University of Ottawa. His research interests include cloud networking, smart environments (smart city and smart health), AI, deep learning, edge computing, Internet of Things, multimedia for health care, and multimedia big data.