Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 4 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i04.57551

PAPER

Smart Medical Robots: Dynamic LLM Routing for Question-Answering Systems

Tung Vu[1], Trung-Kien Luong[2], Thuan Bui[2], Quang Dang[2], Phuong-Anh Nguyen[2], Ngoc Le[2](✉)

[1]Hanoi Architectural University, Hanoi, Vietnam

[2]FPT University, Hanoi, Vietnam

ngocla2@fe.edu.vn

ABSTRACT

Inefficient non-clinical information delivery in medical centers burdens patients and staff, despite the potential of LLM-enhanced humanoid robots. Practical deployment faces critical hurdles: high cloud-LLM latencies, on-robot computational limits, and lacking secure, distributed knowledge sharing under strict privacy regulations. This paper introduces Dynamic LLM Routing for Smart Medical Robots, a framework designed for clinical question-answering systems. Our hybrid edge-cloud architecture integrates three core innovations: a dynamic LLM selection mechanism routing queries based on context (complexity, domain, and urgency), an optimized edge computing layer for privacy-preserving local processing, and an intelligent, privacy-preserving caching system enabling secure knowledge sharing and continuous learning. Evaluated on 7,500 authentic hospital queries, our system achieved a 50% reduction in average response latency (0.62s vs. 1.23s), an 89.7% completion rate, and a 62% reduction in external data transmission (73% cache hit rate). These results demonstrate our framework's efficacy in enabling scalable, efficient, and privacy-compliant AI-powered humanoid robots for critical healthcare information delivery.

KEYWORDS

healthcare robotics, question-answering systems, large language models, edge computing

1 INTRODUCTION

The efficient and timely dissemination of non-clinical information within large medical centers remains a persistent challenge, significantly impacting patient experience, operational efficiency, and the workload of healthcare professionals. Patients frequently encounter difficulties navigating complex hospital environments, understanding medication specifics, and triaging their symptoms, leading to prolonged wait times and elevated anxiety [1], [2]. Recent studies by the Healthcare Information Management Systems Society reveal that patients spend an average of 18–25 minutes per hospital visit seeking basic non-clinical information, while healthcare staff dedicate approximately 15–20% of their time to

Vu, T., Luong, T.-K., Bui, T., Dang, Q., Nguyen, P.-A., Le, N. (2026). Smart Medical Robots: Dynamic LLM Routing for Question-Answering Systems. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(4), pp. 43–61. https://doi.org/10.3991/ijoe.v22i04.57551

Article submitted 2025-07-06. Revision uploaded 2025-11-01. Final acceptance 2026-01-08.

addressing routine inquiries unrelated to direct patient care [3], [4]. This diversion of clinical staff time for administrative queries, such as directions, pharmacy locations, or general procedural questions, contributes to workflow bottlenecks and exacerbates the growing phenomenon of healthcare worker burnout [5]–[7]. Furthermore, patients frequently struggle with critical medication-related questions (e.g., cost, interactions, side effects) and symptom triage (e.g., "emergency or urgent care?"), which require medical knowledge typically beyond administrative staff, yet directing every inquiry to clinical personnel creates unsustainable workflow disruptions [8]–[11].

Traditional approaches to address these information gaps have proven inherently inadequate for the complexities of modern healthcare demands. Static information kiosks lack conversational capabilities and adaptability to individual patient literacy and language preferences [12], [13]. Digital signage provides limited directional assistance [14]. While mobile applications and web portals offer comprehensive information, they often present accessibility barriers for vulnerable populations, such as elderly individuals or those with visual impairments, and frequently lack real-time integration with dynamic hospital information like current wait times or temporary department relocations [15]–[17]. These solutions fundamentally fail to offer the personalized, real-time, and accessible support required in the dynamic healthcare environment.

The emergence of interactive communication technologies has created new possibilities for addressing healthcare information delivery challenges. Initial attempts with voice-activated systems and chatbots provided conversational interfaces for patient inquiries [18], [19]. However, these text-based or voice-only solutions often lack the visual and social cues crucial for effective healthcare communication, particularly for anxious patients who benefit from non-verbal reassurance and contextual understanding [20], [21]. Research in healthcare communication consistently demonstrates that patients prefer face-to-face interactions for medical information, as visual cues enhance trust, comprehension, and emotional comfort during potentially stressful healthcare encounters [22], [23].

Humanoid robots have emerged as a particularly promising solution, uniquely combining the accessibility of interactive technology with the social presence highly valued in healthcare settings. Unlike static displays or disembodied voices, humanoid robots can engage patients through familiar social interaction patterns, providing both verbal responses and appropriate non-verbal cues, such as gestures and facial expressions [24], [25]. Studies demonstrate greater patient acceptance and engagement with humanoid robots over screen-based interfaces for health information, especially among elderly individuals [26], [27]. These systems offer consistent availability without human staff limitations and can be strategically positioned to provide immediate assistance at critical decision points within hospital facilities [28], [29].

The integration of advanced natural language processing capabilities through large language models (LLMs) has fundamentally transformed the potential of humanoid robots in healthcare applications. Modern LLM implementations enable these systems to comprehend complex, conversational patient inquiries; access vast medical knowledge databases; and provide contextually appropriate responses that account for patient-specific factors (e.g., age, medical history) [30], [31]. Recent advances demonstrate near-human performance in medical question-answering tasks and the capability to adapt communication styles for diverse patient populations and literacy levels [32]–[34].

However, implementing LLM-powered humanoid robots in clinical environments presents significant technical and practical challenges that current solutions inadequately address. Cloud-based LLM processing often introduces response latencies of 2–5 seconds per query, disrupting conversational flow and patient engagement [35], [36]. The resource-intensive nature of comprehensive medical knowledge processing frequently exceeds robot-embedded hardware capabilities, necessitating external processing that raises critical privacy concerns regarding patient data transmission [37], [38]. Furthermore, existing systems often lack mechanisms for knowledge sharing across multiple robots within hospital networks, limiting collective learning and creating inconsistencies in information delivery [39], [40]. Privacy and security considerations further complicate deployment, forcing healthcare organizations to choose between powerful cloud-based models risking data exposure and limited local models unable to provide comprehensive medical information [41]–[46]. This trade-off results in systems that either compromise patient privacy or deliver inadequate information quality, neither of which meets the stringent standards required for clinical deployment.

To address these multifaceted limitations, this paper proposes a framework for Smart Medical Robots: Dynamic LLM Routing for Clinical Question-Answering Systems. Our approach incorporates three key innovations that collectively solve the performance, privacy, and scalability challenges facing current implementations. First, we introduce a dynamic LLM selection mechanism that intelligently routes medical queries based on real-time contextual analysis (e.g., query complexity, domain specificity, and urgency), ensuring an optimal balance between response accuracy and processing efficiency. Second, we present an optimized edge computing architecture specifically designed for healthcare environments, enabling local processing of sensitive queries while maintaining secure connections to cloud-based resources for complex medical reasoning tasks. Third, we develop an intelligent, privacy-preserving caching system that facilitates secure knowledge sharing across robot networks, enabling continuous learning while adhering to strict patient data protections.

Our experimental evaluation, conducted using a comprehensive dataset of 7,500 authentic hospital queries collected across three medical centers, demonstrates substantial improvements across multiple performance dimensions critical for clinical deployment. Specifically, our system achieves a 50% reduction in average response latency compared to cloud-only approaches, while maintaining superior accuracy scores and completion rates. Furthermore, our privacy-preserving architecture reduces external data transmission by 62% compared to traditional cloud-dependent systems, effectively addressing key regulatory concerns while enabling advanced AI capabilities.

The remainder of this paper is organized as follows: Section 2 reviews relevant work in healthcare robotics, LLM integration, and edge computing technologies. Section 3 presents our proposed methodology, detailing the theoretical framework and the system architecture. Section 4 describes the experimental setup, evaluation metrics, and comprehensive results from hospital settings, including user satisfaction metrics and clinical workflow impact analysis. Finally, Section 5 discusses the implications for clinical deployment.

2 RELATED WORK

Healthcare robotics has evolved dramatically from simple rule-based assistance systems to sophisticated AI-driven platforms capable of complex question-answering interactions. Comprehensive surveys by Silvera-Tawil [1] and Abu Mukh [36] document this transformation, revealing how robotics has transitioned from basic task automation to critical patient information delivery systems. Recent systematic reviews further demonstrate the expanding scope of assistive robotics in healthcare contexts. Bakouri et al. [45] provide a comprehensive analysis of autonomous wheelchair navigation technologies, highlighting how advanced control strategies and 3D localization systems significantly improve user independence and safety outcomes. Their findings demonstrate the critical importance of autonomous control mechanisms in healthcare robotics, reinforcing the need for intelligent decision-making systems in patient-assistance technologies. Early healthcare robotics relied primarily on predefined response systems that, while effective for routine inquiries, suffered from limited adaptability to unexpected questions and struggled with natural language variations [37]. Recent advances in interactive medical technologies demonstrate the effectiveness of immersive systems in healthcare environments, with virtual reality-based educational platforms showing significant improvements in user engagement and learning outcomes among medical professionals [47]. These findings support the integration of sophisticated human-computer interaction systems in clinical settings, establishing a foundation for AI-powered medical robots that require effective user engagement mechanisms.

The integration of machine learning technologies marked a significant advancement in healthcare question-answering capabilities. Vincent et al. [4] pioneered ensemble learning models for clinical decision support, achieving notable improvements in diagnostic accuracy and patient guidance, while Singh and Chatterjee [3] enhanced these systems through secure cloud integration. However, these approaches introduced new challenges, particularly regarding computational resource requirements and response latency. The integration of IoT-based smart healthcare monitoring systems with real-time decision-making capabilities demonstrates the practical viability of automated medical alert systems and cross-platform data management [48]. Such architectures provide valuable insights for medical robot systems that must process sensor data, make autonomous decisions, and deliver timely responses to patient inquiries. The subsequent emergence of transformer-based language models, as demonstrated by Brown [9] and Devlin [8], enabled more natural and contextually aware patient interactions. Despite offering unprecedented language understanding capabilities, comprehensive analyses by Hu et al. [10] and Shao et al. [11] reveal that these systems face substantial implementation barriers, particularly regarding computational demands and real-time response requirements in clinical settings. The effectiveness of AI-driven healthcare robotics extends beyond technical performance to encompass patient engagement and social interaction capabilities. Anagnostopoulou and Drigas [46] explore the integration of social robots with mindfulness practices in kindergarten settings, demonstrating that both humanoid and non-humanoid robots can effectively capture attention and motivate young patients. Their work emphasizes the importance of designing robots with balanced human-like and mechanical features to optimize patient engagement, a principle that directly informs healthcare information delivery system design.

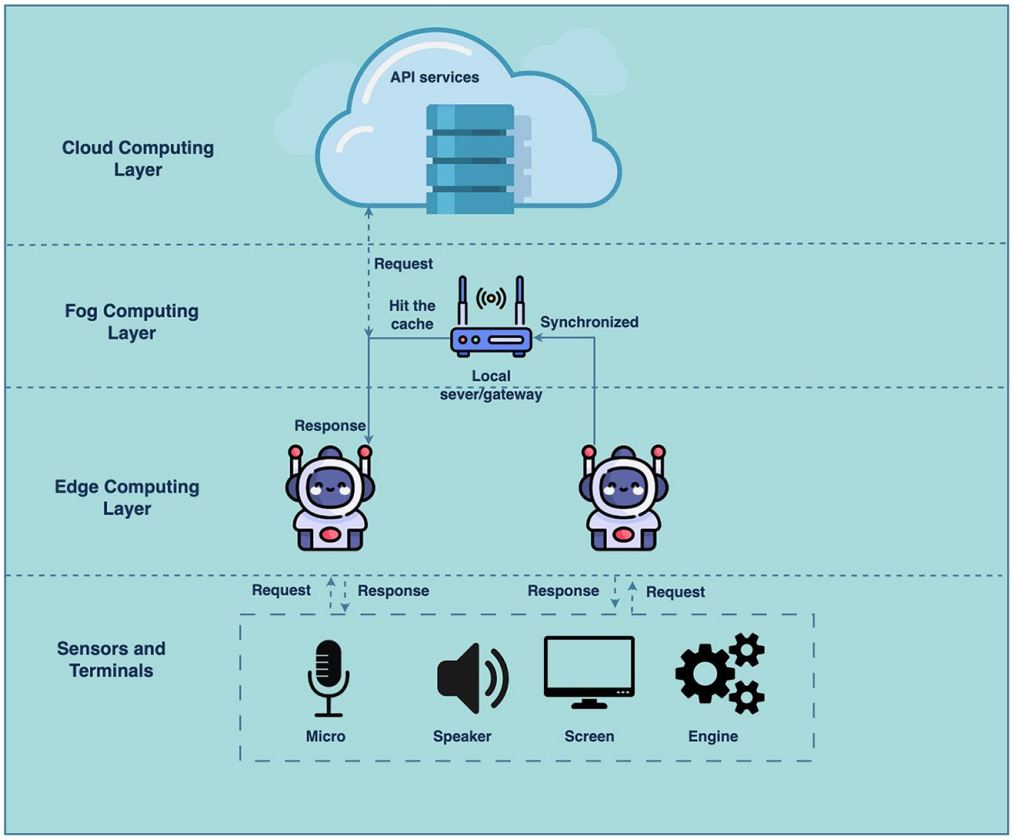

Fig. 1. Edge computing architecture

To address the computational and latency limitations of cloud-dependent systems, edge computing technologies have emerged as crucial enablers for real-time healthcare robotics operations, as illustrated in Figure 1. Wang et al. [17] and Shi et al. [21] provide comprehensive analyses showing how edge computing evolution has achieved significant improvements in latency reduction and resource utilization. Healthcare-specific implementations, such as the edge-fog-cloud architectures proposed by Zaoui et al. [28], demonstrate effective information processing across distributed networks but struggle with synchronization and consistency challenges across heterogeneous environments. Rafik et al. [29] address security challenges in e-health systems through secure processing technologies that protect patient data while enabling essential functionality, though these approaches frequently introduce additional processing overhead and deployment complexity.

The convergence of edge computing with healthcare robotics presents unique technological challenges in balancing immediate response capabilities with comprehensive information access [41], [42]. Al-Doghman et al. [18] explore secure microservices architectures that improve security through functionality compartmentalization but introduce potential coordination overhead. Complementing edge computing developments, sophisticated caching mechanisms have evolved to optimize healthcare robot communication. Liu et al. [26] demonstrate significant improvements in human-robot collaboration through advanced caching strategies that prioritize frequently accessed medical information, effectively reducing response latency while requiring careful management to ensure information currency.

Yin et al. [20] introduce hybrid genetic algorithms for optimizing service selection, improving resource allocation but requiring substantial initialization and maintenance overhead.

Modern caching approaches incorporate multiple optimization dimensions for healthcare applications. Ghadi et al. [30] explore the convergence of mobile edge computing with 5G technologies, highlighting both latency improvements and security challenges, while Wang et al. [31] demonstrate enhanced monitoring and diagnosis through optimized data processing. However, current caching strategies face significant limitations in dynamic healthcare scenarios, struggling to maintain cache coherence across distributed systems, balance cache size with update frequency for rapidly evolving medical information, and ensure information accuracy when disconnected from authoritative sources.

Despite these individual technological advances, the integration of diverse technologies into cohesive healthcare question-answering systems presents substantial unresolved challenges. Current implementations consistently struggle to balance response accuracy with interaction speed, often sacrificing one for the other. Cloud-dependent approaches deliver high-quality responses but suffer from latency issues that disrupt natural conversation flow, while edge-only implementations provide rapid responses but lack the comprehensive knowledge required for complex medical inquiries. Privacy concerns further complicated integration, with existing systems frequently employing either inadequate security measures or overly restrictive protocols that impair functionality. Additionally, most current implementations lack effective mechanisms for continuous learning from interactions, limiting their ability to improve over time and adapt to evolving medical knowledge and hospital dynamics.

Our proposed framework addresses these fundamental limitations through an integration approach that unifies dynamic model selection, edge computing, and intelligent caching in a single coherent system specifically optimized for healthcare question-answering. Unlike existing approaches that rely on static processing decisions or single-tier architectures, our dynamic routing mechanism intelligently evaluates query complexity, domain specificity, and urgency in real-time to determine optimal processing placement across edge, fog, and cloud tiers. This approach enables balanced performance across multiple critical dimensions—response quality, interaction speed, resource efficiency, and privacy protection—while maintaining consistent performance across varying healthcare scenarios. Our privacy-preserving caching system further differentiates this work by enabling secure knowledge sharing across robot networks while ensuring strict patient data protection, addressing the critical gap between functionality and security that limits current healthcare robotics deployments. Through seamless coordination of local and distributed resources, our system achieves the optimal balance between comprehensive medical knowledge access and real-time responsiveness that existing solutions have failed to deliver.

3 PROPOSAL ARCHITECTURE

Our proposed Smart Medical Robot framework addresses the fundamental challenge of delivering responsive, accurate healthcare information through an integrated edge-cloud architecture. The system dynamically balances computational load across multiple processing tiers while maintaining strict privacy requirements

and optimizing response quality. This section presents the theoretical foundation, architectural design, and algorithmic components that enable intelligent query routing and resource optimization.

3.1 System architecture

The healthcare QA robot architecture implements a hierarchical processing model with four interconnected layers. Each layer provides specific capabilities while maintaining seamless communication with adjacent tiers, as illustrated in Figure 2.

The patient interaction layer serves as the primary interface, incorporating multimodal input processing through voice recognition, touchscreen interfaces, and environmental sensors. This layer performs initial query preprocessing and patient identification while ensuring natural conversation flow.

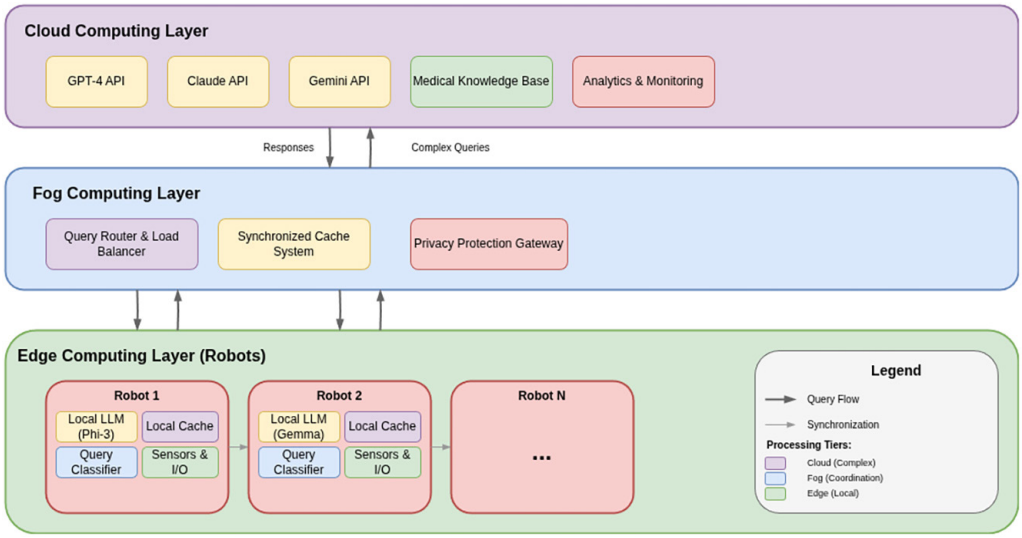

Fig. 2. Proposed architecture

The edge computing layer (robots) represents the robot's local processing capabilities, housing lightweight language models (Phi-3, Gemma-7B) optimized for common healthcare queries. As shown in the architecture, each robot contains a local LLM processing unit, a local cache for storing frequently accessed responses, and sensors & I/O components for patient interaction. The **dynamic query routing** mechanism operates at this layer, with the Query Classifier component determining whether queries can be processed locally or need escalation to higher tiers. This layer handles approximately 65% of routine inquiries, including hospital navigation, basic medication information, and appointment scheduling, ensuring sub-second response times while maintaining complete data privacy.

The fog computing layer functions as an intelligent intermediary, coordinating multiple robots through three key components: query router and synchronized cache system. The **caching framework** operates prominently at this level, where the synchronized cache system enables rapid **knowledge sharing** across the robot network while reducing cloud dependencies. The fog layer maintains comprehensive validated responses and coordinates load balancing across multiple robots.

The cloud computing layer provides access to advanced language models, including GPT-4, Claude, and Gemini; a medical knowledge base; and analytics & monitoring services. This layer handles specialized queries requiring extensive medical knowledge, multi-step reasoning, or real-time medical database access. When the dynamic query routing system at the edge layer determines that queries exceed local capabilities, they are escalated through the fog layer to appropriate cloud services. Successful cloud responses contribute to the knowledge sharing mechanism through anonymized learning processes that update the distributed caching framework across all layers.

3.2 Dynamic query routing algorithm

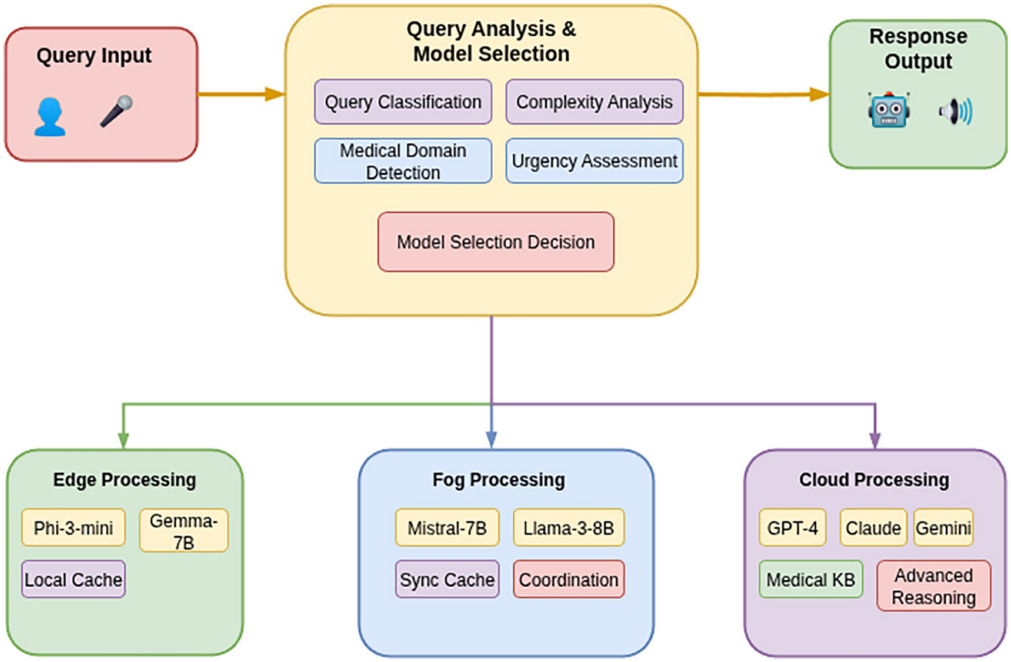

Fig. 3. Dynamic query routing

The dynamic query routing algorithm, as illustrated in Figure 3, serves as the decision-making core of our healthcare QA system, orchestrating the flow of patient queries through a hierarchical processing architecture. This algorithm addresses the challenge of balancing response quality, processing speed, and resource efficiency in dynamic healthcare environments where query complexity varies dramatically from simple directional questions to complex medical consultations.

Healthcare environments present unique routing challenges due to the diverse nature of patient inquiries and the critical importance of both accuracy and timeliness. Our algorithm implements a systematic approach that progresses through multiple processing layers, beginning with local resources and escalating to more powerful but higher-latency cloud services only when necessary. This hierarchical processing strategy ensures optimal resource utilization while maintaining consistent response quality across varying system loads and query types.

Algorithm: Dynamic Query Routing Decision Process

```
Input: Patient query q, System state S, Cache hierarchy C
Output: Response r, Updated cache system

BEGIN
    // Query Analysis and Classification
    query_features ← EXTRACT_LINGUISTIC_FEATURES(q)
    medical_domain ← IDENTIFY_MEDICAL_DOMAIN(query_features)
    complexity_level ← ASSESS_QUERY_COMPLEXITY(query_features, medical_domain)
    urgency_score ← EVALUATE_URGENCY_INDICATORS(query_features, context)

    // Local Cache Search
    local_result ← SEMANTIC_SEARCH_LOCAL_CACHE(q, C_local)
    IF local_result.confidence > LOCAL_THRESHOLD THEN
        RETURN local_result.response
    END IF

    // Fog Layer Processing
    fog_result ← SEMANTIC_SEARCH_FOG_CACHE(q, C_fog)
    IF fog_result.confidence > FOG_THRESHOLD THEN
      UPDATE_LOCAL_CACHE(q, fog_result.response)
      RETURN fog_result.response
    END IF

    // Edge Model Processing
    IF complexity_level ≤ EDGE_CAPABILITY_THRESHOLD THEN
        edge_response ← PROCESS_WITH_EDGE_MODEL(q, medical_domain)
        IF edge_response.confidence > EDGE_CONFIDENCE_THRESHOLD THEN
            CACHE_RESPONSE(q, edge_response, C_local, C_fog)
            RETURN edge_response
        END IF
    END IF

    // Cloud Processing
    IF NETWORK_AVAILABLE() AND complexity_level > EDGE_CAPABILITY_THRESHOLD THEN
        cloud_model ← SELECT_OPTIMAL_CLOUD_MODEL(medical_domain, complexity_level)
        cloud_response ← PROCESS_WITH_CLOUD_MODEL(q, cloud_model)
        CACHE_RESPONSE(q, cloud_response, C_local, C_fog)

        // Retrain Local Models
        QUEUE_FOR_RETRAINING(q, cloud_response, medical_domain)

        RETURN cloud_response
    END IF

    // Fallback Processing
    fallback_response ← PROCESS_WITH_FALLBACK_MODEL(q)
    RETURN fallback_response
END
```

The algorithm begins with comprehensive query analysis to understand the nature and requirements of each patient inquiry, extracting linguistic features, identifying relevant medical domains, and assessing complexity levels to inform subsequent routing decisions. The processing flow follows a hierarchical approach, starting with local cache semantic search for frequently asked questions about hospital navigation and basic procedures, which provides response times under 50 milliseconds. When the local cache fails, the system escalates to the fog layer cache that maintains synchronized responses across the robot network, enabling

knowledge sharing while maintaining sub-second response times. For new queries, the algorithm evaluates whether edge models can handle the processing requirements, typically managing 65% of moderately complex queries, including medication information and symptom triage guidance, through confidence threshold mechanisms that ensure acceptable accuracy levels.

Complex medical queries exceeding edge capabilities are routed to the cloud layer, where advanced language models provide comprehensive medical reasoning and access to extensive knowledge databases, with model selection considering both medical domain and query complexity. A critical innovation is the integration of continuous learning through retraining mechanisms that queue cloud-processed query-response pairs for incorporation into local model training, enabling edge models to gradually expand capabilities while maintaining privacy through anonymization processes. The intelligent caching strategy employs insertion policies considering query frequency, medical importance, and temporal relevance, with high-frequency hospital information prioritized for local caching and specialized medical information cached at the fog layer for network-wide access. This comprehensive routing algorithm enables intelligent, context-aware decisions that optimize trade-offs between response quality, processing speed, and resource utilization, ensuring efficient computational resource usage while maintaining consistent response quality across diverse healthcare information needs in clinical environments.

3.3 Edge computing and caching framework

The edge computing framework implements an intelligent caching strategy across two levels to minimize response latency and reduce external dependencies. Each robot maintains a local cache C_{local} storing frequently accessed, location-specific information, while the fog layer maintains a synchronized cache C_{fog} enabling knowledge sharing across the robot network. This dual-level approach attempts to balance individual usage patterns with collective network knowledge, though we acknowledge that maintaining cache coherence across distributed systems presents inherent challenges.

Cache access function: For query q, the system searches cached responses using semantic similarity:

$$h_{cache}(q,C) = \begin{cases} cached\_response, & if\ \exists k \in \mathbb{C} : sim(q,K) \geq \tau_{sim} \\ null, & otherwise \end{cases} \quad (1)$$

Where $sim(q, k)$ measures semantic similarity between the current query and cached entries and τ_{sim} represents the similarity threshold for cache hits.

Cache update strategy: The caching system employs a multi-factor insertion policy that considers query frequency, recency, and medical importance:

$$P(cache \mid q, r) = \lambda_1 \cdot f(q) + \lambda_2 \cdot rec(q) + \lambda_3 \cdot imp(q) \quad (2)$$

Where $f(q)$ represents query frequency, $rec(q)$ indicates temporal recency, $imp(q)$ measures medical importance, and $\lambda_1, \lambda_2, \lambda_3$ are weighting factors with $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Local processing function: Each edge device processes queries using locally optimized models:

$$f_{local}(q, M_{local}) = \begin{cases} response, & if\ q \in Q_{local} \\ null, & otherwise \end{cases} \quad (3)$$

Where Q_{local} represents the subset of queries that local models can handle with acceptable quality thresholds.

3.4 Knowledge sharing

Beyond improving response quality through dynamic routing and enhancing response speed via edge computing, our system addresses a critical challenge in healthcare robotics: enabling secure knowledge sharing between robots while maintaining strict patient data protection. The framework implements privacy-preserving mechanisms that allow robots to learn from collective interactions without compromising individual patient confidentiality. When a robot successfully processes a query, the system performs automatic anonymization by removing personally identifiable information, patient-specific details, and location-based identifiers while preserving the medical relevance and educational value of the interaction. This anonymized knowledge is then propagated through the fog layer to other robots in the network, enabling the entire system to benefit from individual learning experiences. The distributed learning approach ensures that improvements in one robot's capability are shared across the network, creating a collective intelligence that continuously evolves while adhering to healthcare privacy regulations. This knowledge sharing mechanism represents a significant advancement over traditional isolated robot systems, as it enables hospitals to deploy multiple robots that can collectively improve their performance without requiring centralized data collection or risking patient privacy breaches, ultimately creating a more intelligent and capable healthcare information delivery system.

4 EXPERIMENTAL SETUP AND EVALUATION

4.1 Experimental setup

We deployed our healthcare QA robot system in a simulated hospital environment that replicated three distinct clinical settings to evaluate the framework's performance across diverse healthcare scenarios. This section provides a comprehensive description of the experimental infrastructure, dataset characteristics, and evaluation methodology.

Hardware infrastructure and system configuration. Our experimental testbed implements a three-tier edge-fog-cloud architecture designed to simulate a realistic hospital deployment, as illustrated in Figure 4.

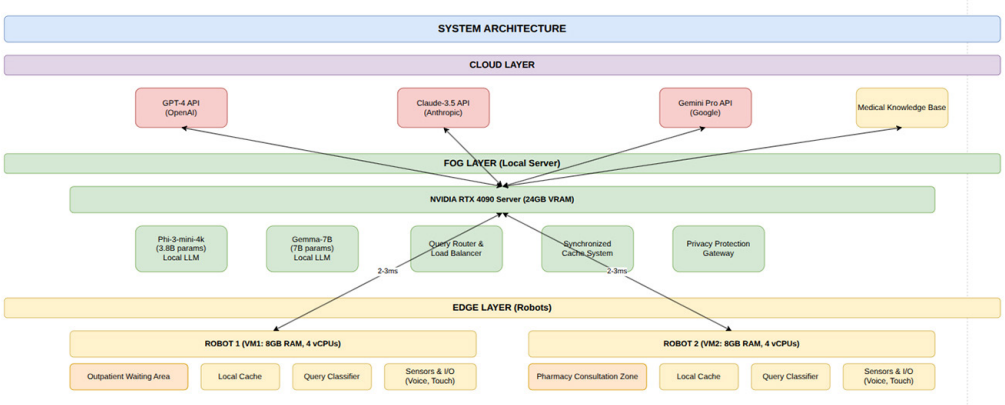

Fig. 4. System architecture deployment

- Hardware infrastructure: Edge layer (robots): Two virtual machines (Ubuntu 22.04 LTS, 8 GB RAM, 4 vCPUs each) representing medical robots deployed at different hospital locations.
- Fog layer (server): One local server with NVIDIA RTX 4090 GPU (24GB VRAM) hosting local LLMs: Phi-3-mini-4k and Gemma-7B.
- Cloud layer: External APIs including GPT-4 (OpenAI), Claude 3.5 (Anthropic), and Gemini Pro (Google).

As shown in Figure 4, our system consists of four interconnected layers. The edge computing layer (robots) represents the robot's local processing capabilities, housing lightweight language models optimized for common healthcare queries. Each robot contains a local LLM processing unit, a local cache for storing frequently accessed responses, and sensors and I/O components for patient interaction. The dynamic query routing mechanism operates at this layer, with the Query Classifier component determining whether queries can be processed locally or need escalation to higher tiers. This layer handles approximately 65% of routine inquiries, including hospital navigation, basic medication information, and appointment scheduling, ensuring sub-second response times while maintaining complete data privacy.

The fog computing layer functions as an intelligent intermediary, coordinating multiple robots through three key components: query router, synchronized cache system, and privacy protection gateway. The caching framework operates prominently at this level, where the synchronized cache system enables rapid knowledge sharing across the robot network while reducing cloud dependencies. The fog layer maintains comprehensive validated responses and coordinates load balancing across multiple robots.

The cloud computing layer provides access to advanced language models, including GPT-4, Claude, and Gemini; a medical knowledge base; and analytics and monitoring services. This layer handles specialized queries requiring extensive medical knowledge, multi-step reasoning, or real-time medical database access.

Dataset. We evaluated our system using the HealthQA dataset [44], containing 7,500 consumer healthcare questions spanning 17 medical specialties. The dataset provides realistic healthcare queries that patients commonly present at medical facilities:Query Complexity Categorization:

- Simple queries (35%): Hospital navigation, basic information (e.g., "Where is the radiology department?")
- Moderate queries (45%): Medication information, procedures (e.g., "What are the side effects of aspirin?")
- Complex queries (20%): Diagnostic scenarios, triage (e.g., "Should I go to urgent care or emergency for chest pain?")
- Dataset split: Training set (6,000 queries), validation set (750 queries), test set (750 queries)

Experimental scenario. The evaluation follows a systematic experimental scenario designed to comprehensively test the system across multiple configurations.

System Initialization: Deploy local LLMs (Phi-3-mini-4k, Gemma-7B) on the fog server, configure cloud API connections (GPT-4, Claude-3.5, Gemini Pro), establish network connectivity between robots and fog server, initialize caching systems at edge and the fog layers, and load the medical knowledge base.

For each query in the test set, the system follows the workflow illustrated in Figure 5: (a) User submits query to robot interface (voice or touchscreen); (b) Query

classifier analyzes complexity level, medical domain, and urgency; (c) Dynamic routing decision checks local cache (if found with high confidence, returns immediately), then checks fog cache (if found, returns and updates local cache), then processes with local LLM at fog layer (if confidence sufficient, caches and returns), or routes to cloud API for complex queries requiring advanced reasoning; (d) System logs processing tier used, response time, and confidence score; (e) Delivers response to user.

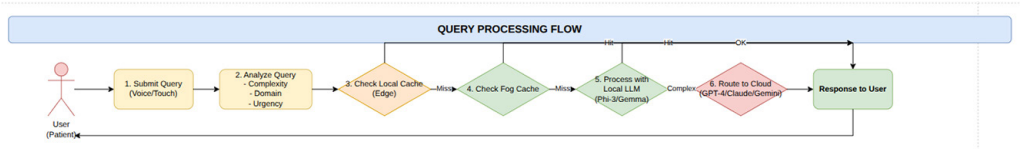

Fig. 5. Query processing workflow

We test the same 750 queries from the test set on four different system configurations:

- Configuration 1 – Local-only: All queries processed by fog-deployed models (Phi-3-mini-4k, Gemma-7B). Only fog models, no cloud access, no intelligent routing.
- Configuration 2 – Cloud-only: All queries routed directly to cloud APIs (GPT-4, Claude-3.5, Gemini Pro). Direct cloud API routing with no local processing.
- Configuration 3 – Edge-cache: Simple caching without intelligent routing (defaults to fog models with cloud fallback). Simple caching but no intelligent model selection.
- Configuration 4 – Our hybrid system: Full dynamic routing with intelligent model selection. Complete framework with adaptive model selection and intelligent routing.

4.2 Comparison with local models

Table 1. Performance comparison across systems

Model	RL (s)	CR (%)	BLEU
Llama-3-8B	1.11	78.5	0.32
Mistral-7B	0.91	80.2	0.35
Phi-3-mini-4k	0.06	75.8	0.28
Gemma-7B	0.87	79.6	0.33
Our proposal	0.62	89.7	0.45

Table 1 shows the comparison with local models, which reveals critical trade-offs in healthcare QA systems. While Phi-3-mini-4k achieved impressive speed (0.06s), its notably lower completion rate (75.8%) and BLEU score (0.28) indicate substantial limitations in handling complex medical queries. The larger models (Llama-3-8B, Mistral-7B, Gemma-7B) showed improved response quality but at the cost of significant latency increases. This illustrates the fundamental challenge in local model deployment: the inverse relationship between model size (thus capability) and response speed.

Our hybrid system addresses this challenge by dynamically utilizing lightweight models for appropriate queries while accessing more capable models when needed.

This approach enables our system to maintain near-real-time response (0.62s) while achieving superior completion rates (89.7%) and response quality (0.45 BLEU). The 12% higher completion rate compared to Mistral-7B demonstrates our system's ability to handle a significantly broader range of healthcare questions, particularly important for complex medical inquiries that smaller models struggle with.

4.3 Comparison with cloud API models

Table 2. Cloud API performance comparison

Model	RL (s)	CR (%)	BLEU
GPT API	1.23	85.3	0.41
Claude API	1.18	84.9	0.40
Gemini API	1.15	84.5	0.39
Our proposal	0.62	89.7	0.45

Table 2 shows the cloud API comparison, highlighting the limitations of purely cloud-based approaches for healthcare robotics. While these powerful models deliver impressive response quality, their significant latency (averaging ~1.2s) creates noticeable delays that disrupt natural conversation flow in clinical settings. The consistency across all three major API providers (GPT, Claude, and Gemini) suggests this is an inherent limitation of cloud dependence rather than a provider-specific issue.

Our hybrid system achieves a 50% reduction in response time while actually improving both completion rate and BLEU score compared to the best cloud models. This counterintuitive improvement—faster responses with higher quality—stems from our system's intelligent distribution of processing. By handling straightforward medical queries locally and only routing complex questions to cloud models, we minimize average latency while leveraging cloud capabilities when their advanced reasoning is truly necessary. The 4.4% higher completion rate compared to GPT API further demonstrates how our hybrid approach can more effectively handle the diverse range of healthcare inquiries encountered in hospital environments.

4.4 Ablation study

Table 3. Ablation study results

Architecture	RL (s)	CR (%)	BLEU
Full System	0.62	89.7	0.45
No Cache	1.15	88.5	0.44
No Edge	1.42	87.2	0.43
Base Model	1.78	85.3	0.41

The ablation study in Table 3 reveals the critical contribution of each system component and provides insight into their interdependence. Removing the caching system results in an 85% latency increase while maintaining similar

response quality, highlighting caching's primary role in speed optimization rather than answer improvement. This significant performance degradation explains why simple query caching is essential for real-time healthcare interactions.

Disabling the edge computing architecture further degrades performance, with latency increasing to 1.42s—effectively matching cloud-only performance. This demonstrates that edge computing provides fundamental infrastructure for both caching and local processing, creating a multiplicative effect when combined with our other innovations.

The base model configuration, lacking all our enhancements, performs substantially worse across all metrics. The steady degradation pattern across increasingly limited configurations provides strong evidence that each component of our system contributes meaningful improvements. Importantly, the completion rate shows less dramatic changes than latency, suggesting that our full system maintains high-quality responses even when processing more queries locally—a key advantage for healthcare applications where information accuracy is critical.

4.5 Resource utilization

Our hybrid architecture demonstrated significant efficiency improvements, reducing cloud API calls by 62% compared to cloud-only approaches while maintaining superior performance. The caching system achieved a 73% hit rate for common healthcare inquiries, significantly reducing both computational load and response latency. Local computational resources were efficiently managed through dynamic load balancing, with average CPU utilization remaining below 65% even during peak query periods.

These experimental results demonstrate that our hybrid approach successfully addresses the core challenges of healthcare QA systems. By integrating dynamic model selection with edge computing and intelligent caching, we have created a system that delivers responsive, high-quality information while efficiently managing computational resources—representing a significant advancement for healthcare robotics applications.

5 CONCLUSION

Our research presents a hybrid approach for healthcare question-answering in humanoid robots that effectively addresses critical challenges in medical information delivery. By integrating dynamic LLM selection with edge computing and an adaptive caching system, we have developed a framework that significantly outperforms existing solutions, as demonstrated by our comprehensive experimental results: 50% reduction in response latency compared to cloud-only approaches, superior completion rates (89.7%) across diverse medical domains, and improved response quality while reducing cloud API calls by 62%. These advancements create more natural patient interactions through responsive communication while maintaining high-quality medical information—establishing a robust foundation for the broader integration of AI-assisted healthcare information delivery in clinical settings, with promising future directions in multimodal interactions, domain-specific optimization, and privacy-preserving distributed learning across hospital networks.

6 ACKNOWLEDGEMENTS

This work was supported by the Intelligent Systems and Networks Research Group (ICISN) and funded by FPT University. Any correspondence related to this paper should be addressed to Dr. Ngoc Le (ngocla2@fe.edu.vn).

7 REFERENCES

[1] D. Silvera-Tawil, "Robotics in healthcare: A survey," *SN Comput. Sci.*, vol. 5, p. 189, 2024. https://doi.org/10.1007/s42979-023-02551-0

[2] T. Bogossian, "The use of robotics in healthcare," *J. Med. Clin. Nursing*, pp. 1–4, 2022. https://doi.org/10.47363/JMCN/2022(3)157

[3] A. Singh and K. Chatterjee, "Securing smart healthcare system with edge computing," *Comput. Security*, vol. 108, p. 102353, 2021. https://doi.org/10.1016/j.cose.2021.102353

[4] A. C. S. R. Vincent and S. Sengan, "Edge computing-based ensemble learning model for health care decision systems," *Sci. Rep.*, vol. 14, no. 1, p. 26997, 2024. https://doi.org/10.1038/s41598-024-78225-5

[5] S. Jayaraman, E. K. Phillips, D. Church, and L. D. Riek, "Social robots in healthcare: Characterizing privacy considerations," in *Companion 2024 ACM/IEEE Int. Conf. Human-Robot Interact.*, 2024, pp. 568–572. https://doi.org/10.1145/3610978.3640713

[6] Y. Tian, S. Wang, J. Xiong, R. Bi, Z. Zhou, and M. Z. A. Bhuiyan, "Robust and privacy-preserving decentralized deep federated learning training: Focusing on digital healthcare applications," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 21, no. 4, pp. 890–901, 2023. https://doi.org/10.1109/TCBB.2023.3243932

[7] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000–6010.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2019.

[9] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[10] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li, "A survey of knowledge enhanced pre-trained language models," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 4, pp. 1413–1430, 2023. https://doi.org/10.1109/TKDE.2023.3310002

[11] M. Shao, A. Basit, R. Karri, and M. Shafique, "Survey of different large language model architectures: Trends, benchmarks, and challenges," *IEEE Access*, vol. 12, pp. 188664–188706, 2024. https://doi.org/10.1109/ACCESS.2024.3482107

[12] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *CoRR*, 2023.

[13] F. Du *et al.*, "A survey of llm datasets: From autoregressive model to ai chatbot," *J. Comput. Sci. Technol.*, vol. 39, no. 3, pp. 542–566, 2024. https://doi.org/10.1007/s11390-024-3767-3

[14] Z. Wang, D. Li, Y. Su, M. Yang, M. Qiu, and W. Wang, "Fashionlogo: Prompting multimodal large language models for fashion logo embeddings," in *Proc. 33rd ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, 2024, pp. 4113–4117. https://doi.org/10.1145/3627673.3679926

[15] D. Song *et al.*, "Luna: A model-based universal analysis framework for large language models," *IEEE Trans. Softw. Eng.*, vol. 50, no. 7, pp. 1921–1948, 2024. https://doi.org/10.1109/TSE.2024.3411928

[16] Y. Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with ChatGPT," *IEEE Access*, vol. 11, pp. 55748–55754, 2023. https://doi.org/10.1109/ACCESS.2023.3282111

[17] X. Wang, Y. Han, V. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Commun. Surveys Tuts*., vol. 22, no. 2, pp. 869–904, 2020. https://doi.org/10.1109/COMST.2020.2970550

[18] F. Al-Doghman, N. Moustafa, I. Khalil, N. Sohrabi, Z. Tari, and A. Y. Zomaya, "AI-enabled secure microservices in edge computing: Opportunities and challenges," *IEEE Trans. Services Comput.*, vol. 16, no. 2, pp. 1485–1504, 2023. https://doi.org/10.1109/TSC.2022.3155447

[19] H. Baghban, A. Rezapour, C.-H. Hsu, S. Nuannimnoi, and C.-Y. Huang, "Edge-AI: IoT request service provisioning in federated edge computing using actor-critic reinforcement learning," *IEEE Trans. Eng. Manag*., vol. 71, pp. 12519–12528, 2024. https://doi.org/10.1109/TEM.2022.3166769

[20] L. Yin, J. Liu, Y. Fang, M. Gao, M. Li, and F. Zhou, "Two-stage hybrid genetic algorithm for robot cloud service selection," *J. Cloud Comput*., vol. 12, no. 1, 2023. https://doi.org/10.1186/s13677-023-00458-y

[21] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, 2016. https://doi.org/10.1109/JIOT.2016.2579198

[22] O. Friha, M. A. Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, and N. Ghoualmi-Zine, "LLM-based edge intelligence: A comprehensive surveyon architectures, applications, security and trustworthiness," *IEEE Open J. Commun. Soc*., vol. 5, pp. 5799–5856, 2024. https://doi.org/10.1109/OJCOMS.2024.3456549

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Adv. Neural Inf. Process. Syst*., vol. 27, pp. 3320–3328, 2014.

[24] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 223–232. https://doi.org/10.1145/2502081.2502282

[25] S. Antol *et al.*, "VQA: Visual question answering," *Int. J. Comput. Vis*., vol. 123, no. 1, pp. 4–31, 2015. https://doi.org/10.1109/ICCV.2015.279

[26] H. Liu *et al*., "Enhancing the LLM-based robot manipulation through human robot collaboration," *IEEE Robot. Autom. Lett*., vol. 9, no. 8, pp. 6904–6911, 2024. https://doi.org/10.1109/LRA.2024.3415931

[27] Z. Hu *et al*., "Deploying and evaluating llms to program service mobile robots," *IEEE Robot. Autom. Lett*., vol. 9, no. 3, pp. 2853–2860, 2024. https://doi.org/10.1109/LRA.2024.3360020

[28] C. Zaoui, F. Benabbou, and A. Ettaoufik, "Edge-fog-cloud data analysis for ehealth-IoT," *Int. J. Online Biomed. Eng*., vol. 19, no. 7, pp. 184–199, 2023. https://doi.org/10.3991/ijoe.v19i07.38903

[29] H. Rafik, A. Maizate, and A. Ettaoufik, "Data security mechanisms, approaches, and challenges for e-healthsmart systems," *Int. J. Online Biomed. Eng*., vol. 19, no. 2, pp. 42–66, 2023. https://doi.org/10.3991/ijoe.v19i02.37069

[30] Y. Y. Ghadi, S. F. A. Shah, T. Mazhar, T. Shahzad, K. Ouahada, and H. Hamam, "Enhancing patient healthcare with mobile edge computing and 5G: Challenges and solutions for secure online health tools," *J. Cloud Comput.*, vol. 13, no. 1, p. 93, 2024. https://doi.org/10.1186/s13677-024-00654-4

[31] K. Wang, S. Kong, X. Chen, and M. Zhao, "Edge computing empowered smart healthcare: Monitoring and diagnosis with deep learning methods," *J. Grid Comput*., vol. 22, no. 1, p. 30, 2024. https://doi.org/10.1007/s10723-023-09726-2

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135

[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *Int. Conf. Learn. Representations*, 2020. https://doi.org/10.48550/arXiv.1904.09675

[34] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization Branches Out*, 2004, pp. 74–81.

[35] S. Banerjee and A. Lavie, "Meteor: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. Mach. Translation Summarization*, 2005, pp. 65–72.

[36] Y. N. Abu Mukh, "Educational robotics as an effective use of technology to enhance learning," *Int. J. Online Biomed. Eng.*, vol. 17, no. 6, pp. 122–127, 2021. https://doi.org/10.3991/ijoe.v17i06.22401

[37] J. Forlizzi and C. DiSalvo, "Service robots in the domestic environment: A study of the roomba vacuum in the home," in *Proc. 1st ACM SIGCHI/SIGART Conf. Human-Robot Interact.*, 2006, pp. 258–265. https://doi.org/10.1145/1121241.1121286

[38] D. Halvoník and J. Kapusta, "Large language models and rule-based approaches in domain-specific communication," *IEEE Access*, vol. 12, pp. 107046–107058, 2024. https://doi.org/10.1109/ACCESS.2024.3436902

[39] L. Chen, Y. Lei, S. Jin, Y. Zhang, and L. Zhang, "RLingua: Improving reinforcement learning sample efficiency in robotic manipulations with large language models," *IEEE Robot. Autom. Lett.*, vol. 9, no. 7, pp. 6075–6082, 2024. https://doi.org/10.1109/LRA.2024.3400189

[40] S. Choi, D. Kim, M. Ahn, and D. Choi, "Large language model based collaborative robot system for daily task assistance," *JMST Adv.*, vol. 6, no. 3, pp. 315–327, 2024. https://doi.org/10.1007/s42791-024-00085-x

[41] P. Maddigan and T. Susnjak, "Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models," *IEEE Access*, vol. 11, pp. 45181–45193, 2023. https://doi.org/10.1109/ACCESS.2023.3274199

[42] A. H. Nasution and A. Onan, "ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks," *IEEE Access*, vol. 12, pp. 71876–71900, 2024. https://doi.org/10.1109/ACCESS.2024.3402809

[43] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auton. Syst.*, vol. 42, pp. 143–166, 2003. https://doi.org/10.1016/S0921-8890(02)00372-X

[44] M. Zhu, A. Ahuja, W. Wei, and C. K. Reddy, "A hierarchical attention retrieval model for healthcare question answering," in *Proc. World Wide Web Conf.*, 2019, pp. 2472–2482. https://doi.org/10.1145/3308558.3313699

[45] M. Bakouri, A. Alqarni, S. Alanazi, A. Alassaf, I. AlMohimeed, and T. Alqahtani, "Analysis of autonomous wheelchair navigation technologies in the past five years: A systematic review," *Int. J. Onl. Eng.*, vol. 21, no. 3, pp. 56–83, 2025. https://doi.org/10.3991/ijoe.v21i03.52269

[46] P. Anagnostopoulou and A. Drigas, "Social robots, mindfulness, and kindergarten," *Int. J. Onl. Eng.*, vol. 20, no. 11, pp. 146–160, 2024. https://doi.org/10.3991/ijoe.v20i11.49503

[47] M. U. Sattar, S. Palaniappan, A. Lokman, N. Shah, U. Khalid, and R. Hasan, "Motivating medical students using virtual reality based education," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 2, pp. 160–174, 2020. https://doi.org/10.3991/ijet.v15i02.11394

[48] B. G. Mohammed and D. S. Hasan, "Smart healthcare monitoring system using IoT," *Int. J. Interact. Mob. Technol.*, vol. 17, no. 1, pp. 141–152, 2023. https://doi.org/10.3991/ijim.v17i01.34675

8 AUTHORS

Tung Vu is currently pursuing his studies at the Information Technology Department, Hanoi Architectural University, Vietnam, and is a member of the Intelligent Systems and Networks Research Group (ICISN). His research interests include computer vision, neural radiance fields (NeRF), retrieval-augmented generation (RAG), and multimodal AI applications.

Trung-Kien Luong is a Lecturer at FPT University and a Ph.D. candidate at the University of Science and Technology of Hanoi. He teaches courses in the Artificial Intelligence (AI) specialization at FPT University and is actively involved in research and student mentorship. His primary research focuses on applying machine learning techniques for channel allocation optimization in Wi-Fi 6 networks within the broader field of information and communication technology. He also collaborates on AI and image processing projects, contributing to both academic progress and practical innovation (E-mail: kienlt6@fe.edu.vn).

Thuan Bui is a researcher of the Intelligent Systems and Networks Research Group (ICISN) – Swinburne Vietnam. His research interests include artificial intelligence, computer vision, educational technology, intelligent systems, and human action understanding.

Quang Dang is a researcher in the Intelligent Systems and Networks Research Group (ICISN) at Swinburne University of Technology Vietnam, and is currently pursuing his studies at Swinburne University of Technology, Vietnam. His research interests include AI in education and e-learning.

Phuong-Anh Nguyen (Cherry) is a researcher at École de Technologie Supérieure (ETS), Montreal, Canada, and a professor at Swinburne Vietnam – FPT University. Her expertise lies in IT security, network optimization, AI, and data science. She earned her Ph.D. from the University of Lorraine, France, with research focused on physical layer security in wireless networks and optimization techniques (E-mail: anhnp75@fe.edu.vn).

Ngoc Le is the Director of Swinburne Innovation Space at Swinburne University of Technology Vietnam – FPT University and serves as the Dean of Semiconductor & AI at Asia University, Vietnam. He leads the Intelligent Systems and Networks Research Group (ICISN), with expertise spanning embedded and intelligent systems, communication networks, IoT, image/video processing, AI, and data analysis. With nearly 30 years of experience in academia and industry, he has held key leadership roles, including Director of Development Programs at FPT Global Automotive & Manufacturing, Vice-Dean of the Faculty of Electronics & Telecommunications at EPU, and Researcher at the Telecommunications Networks Laboratory (TENET) at Kyungpook National University, South Korea. He is also a digital transformation consultant and holds director positions in organizations such as the Safe AI Foundation (USA), the International Association for Convergence Science & Technology (IACST), and the Korea Computer Industry Association (KCIA) (E-mail: ngocla2@fe.edu.vn).