

GENERAL EDGE-CLOUD HETEROGENEOUS MULTI-AGENT COLLABORATIVE REASONING ARCHITECTURE WITH LOCAL-PRIVACY CONSENSUS MECHANISM

Cheng Xie[#], MianJie Huang[#], LiZhu Ye^{*}

School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou 511363, Guangdong, China.

[#]These authors contributed equally.

^{*}Corresponding Author: LiZhu Ye

Abstract: Multi-agent systems for large-language-model tasks are evolving from single-model generation toward complex-task collaborative reasoning. Conventional schemes concentrate raw data, private knowledge and reasoning workflows on the cloud, and merge outputs via majority voting among homogeneous agents, which hardly satisfy native-privacy protection, heterogeneous task division, private-knowledge enhancement and traceable-evidence requirements simultaneously. This paper proposes GEHMAR, a general edge-cloud heterogeneous multi-agent collaborative reasoning architecture oriented to local-privacy consensus. The architecture decouples data, knowledge and computation: the edge undertakes private-data parsing, local RAG retrieval, evidence compression and privacy isolation, while the cloud only receives desensitized evidence after minimal processing. A heterogeneous reasoning group consisting of Analysis Agent, Multimodal Verification Agent, Evidence Review Agent and Arbitration Consensus Agent is deployed in the cloud. Furthermore, a Local-Privacy RAG mechanism and a heterogeneous evidence-weighted consensus mechanism are put forward. The former prevents private knowledge from leaving the edge domain, and the latter constructs dynamic weights using evidence confidence, agent-role reliability and cross-agent consistency to realize traceable consensus through multi-round divergence resolution. Validated on a privacy-sensitive multi-expert medical consultation scenario without domain-specific customization, the proposed architecture is evaluated from reasoning quality, privacy exposure and system-performance perspectives. Ablation experiments verify individual contributions of core modules. Experimental results show that GEHMAR outperforms traditional multi-agent schemes in accuracy, evidence completeness and task success rate. It drastically reduces raw-data transmission volume and achieves strong data isolation with acceptable latency overhead.

Keywords: Edge-cloud collaboration; Multi-agent; Local privacy; Retrieval-augmented generation; Heterogeneous reasoning

1 INTRODUCTION

Large-language-model-based multi-agent systems have advanced from isolated single-model generation to complex collaborative reasoning tasks [1-2]. Representative multi-agent frameworks such as AutoGen and Mixture-of-Agents improve reasoning quality through interactive discussion and hierarchical output aggregation [3-4]. Nevertheless, most existing collaborative workflows run entirely on cloud servers. Raw private records and knowledge bases are uploaded to remote models, and homogeneous agents adopt simple majority voting for result fusion, which brings prominent privacy risks and ignores role-specific capability differences [5-6]. Private RAG systems are particularly vulnerable to inference-based knowledge leakage, posing critical challenges for privacy-sensitive decision-making scenarios. Even if post-hoc output filtering is adopted, it cannot block private information embedded in intermediate reasoning content. Such hidden leakage risks greatly hinder the practical deployment of multi-agent systems in industries with strict data governance requirements.

Existing studies have explored multi-agent debate mechanisms, edge-cloud collaborative inference and privacy-enhanced RAG techniques separately [7]. However, few integrated architectures jointly address local private knowledge retention, heterogeneous agent division, desensitized cross-domain communication and evidence-driven consensus. Traditional cloud-centric multi-agent designs couple raw data access with heavy-weight reasoning. Output-only sanitization cannot fundamentally limit private-data exposure, and majority-vote fusion cannot quantify evidence reliability or support traceable audit for complex decisions [8-10]. These gaps restrict multi-agent deployment in practical privacy-critical applications. Most existing solutions optimize only one single dimension such as reasoning performance or privacy protection, lacking end-to-end collaborative protocols covering the full data processing pipeline. There is still insufficient research on how to balance reasoning accuracy, privacy constraints and system overhead simultaneously.

To mitigate the above limitations, this paper proposes GEHMAR, a general edge-cloud heterogeneous multi-agent collaborative reasoning architecture with local-privacy consensus. Data, knowledge and computation are decoupled across edge and cloud layers. A Local-Privacy RAG mechanism keeps private knowledge within the edge domain, and a heterogeneous evidence-weighted multi-round consensus mechanism replaces naive majority voting. Evaluated under

a medical multi-expert consultation scenario without domain-hardcoded components, ablation studies verify the independent contribution of each core module. This work provides a feasible collaborative-reasoning paradigm for privacy-sensitive complex-decision tasks.

2 PROPOSED ARCHITECTURE

2.1 Overall Design

GEHMAR consists of the Edge Local-Privacy Private Perception Layer, the Desensitized Interactive Communication Layer, and the Cloud Heterogeneous Multi-Agent Reasoning Layer. The three layers are not merely functional partitions; instead, they correspond to three distinct trust and computing boundaries: the private-data boundary, the cross-domain communication boundary, and the cloud cognitive-computing boundary. Its core objective is to decouple “where data resides” from “where reasoning is performed” from the traditional all-in-one cloud-based paradigm. Figure 1 illustrates a general-purpose cloud–edge heterogeneous multi-agent collaborative inference architecture designed for local privacy preservation.

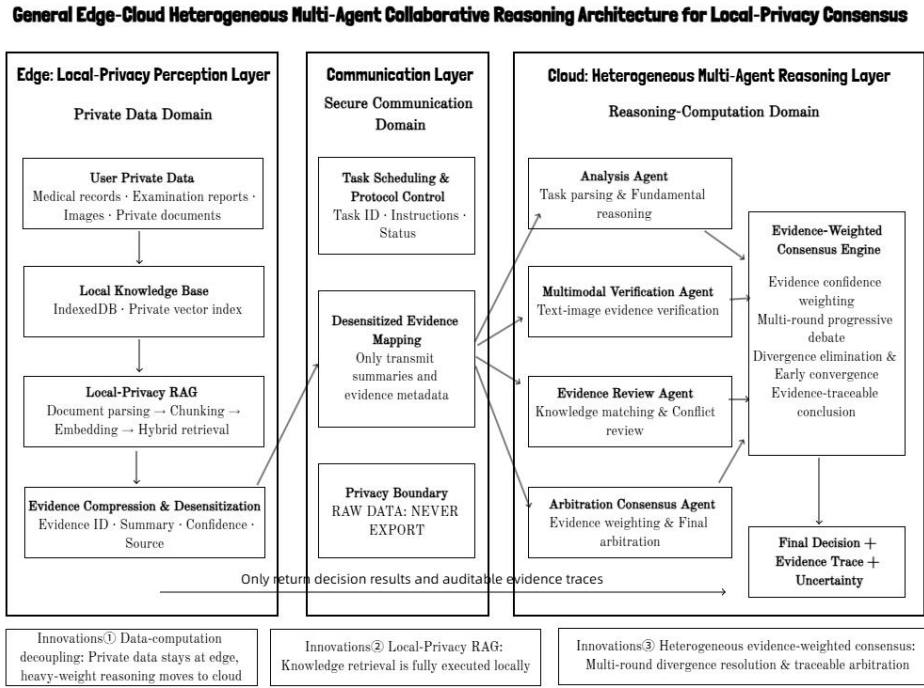


Figure 1 A General-Purpose Cloud–Edge Heterogeneous Multi-Agent Collaborative Inference Architecture for Local Privacy-Conservation Scenarios

2.2 Edge Local-Privacy Private Perception Layer

The edge layer undertakes the lifecycle management of private data, including file parsing, semantic chunking, embedding computation, vector storage, hybrid retrieval, and evidence compression. Persistent browser-side storage is adopted to preserve private data, and RAG indexing and retrieval are completed locally, so that private knowledge achieves data-boundary isolation before entering cloud-side reasoning.

Let the private knowledge base be $D = d_1, \dots, d_n$, and the user task query be q . This paper constructs the initial retrieval function via a linear combination of vector similarity and keyword similarity:

$$R(q, d) = \lambda \cdot \text{Sim}_{\text{emb}}(q, d) + (1 - \lambda) \cdot \text{Sim}_{\text{key}}(q, d) \quad (1)$$

where $\lambda \in [0, 1]$ controls the contributions of semantic retrieval and lexical retrieval. $\text{Sim}_{\text{emb}}(q, d)$ denotes embedding-based semantic similarity between query q and document fragment d , and $\text{Sim}_{\text{key}}(q, d)$ represents keyword matching similarity. After obtaining Top-K candidate segments, the edge layer performs evidence compression and generates $e_j = (id_j, s_j, c_j, r_j)$, in which id_j denotes the evidence identifier, s_j is the desensitized abstract, c_j is the evidence confidence value, and r_j is the source trace information. The original segment d_j is never transmitted to the cloud.

2.3 Desensitized Interactive Communication Layer

The design principle of the communication layer follows minimum sufficient exposure. For task q , the cloud-side required message is denoted as $M_q = \text{TaskID}, E_q, \text{Meta}_q$, where E_q represents the set of desensitized evidence and

Meta_q only contains task-control information. Original documents, raw images, complete vectors and task-irrelevant private fields are defined as prohibited outbound payloads. Accordingly, privacy protection is upgraded from “desensitization before model invocation” to “prohibition of raw-data outbound at the protocol layer”.

2.4 Cloud Heterogeneous Multi-Agent Reasoning Layer

Four role-fixed, model-replaceable agents are deployed on the cloud. The Analysis Agent is responsible for task decomposition and initial hypothesis generation. The Multimodal Verification Agent conducts text-image consistency checking and cross-modal evidence verification. The Evidence Review Agent handles knowledge matching, conflict identification, and evidence adequacy inspection. The Arbitration Consensus Agent does not generate initial outputs; instead, it aggregates evidence scores from each agent, identifies divergences, and produces final decisions. This design decouples “roles” from “models”: models can be replaced while functional responsibilities remain stable.

2.5 Local-Privacy RAG Mechanism

The key distinction between the Local-Privacy RAG adopted in this paper and conventional Cloud-RAG lies in the location of the knowledge base. Cloud-RAG usually sends private documents or their retrieval context to cloud-hosted models. By contrast, this paper keeps retrieval operations on the Edge and only transmits compressed and desensitized evidence. This strategy directly addresses RAG-oriented privacy risks, as existing studies have verified that private retrieval databases can serve as attack surfaces for information leakage [9]. Therefore, the privacy mechanism in this paper imposes exposure restrictions before data enters the cross-domain collaboration pipeline, rather than applying post-hoc remedies after result generation.

Let the Top-K evidence set be $E_q = e_1, \dots, e_K$. The edge layer calculates the comprehensive confidence based on evidence relevance, source credibility, and retrieval stability:

$$c_j = \eta_1 r_j^{\text{sem}} + \eta_2 r_j^{\text{key}} + \eta_3 r_j^{\text{src}}, \sum_{m=1}^3 \eta_m = 1 \quad (2)$$

where r_j^{sem} denotes semantic relevance, r_j^{key} denotes keyword matching degree, and r_j^{src} denotes source reliability. Only evidence whose c_j exceeds threshold τ_e is delivered in cloud-bound messages, which reduces both communication overhead and irrelevant information propagation.

2.6 Heterogeneous Evidence-Weighted Consensus

The implicit assumption of traditional majority voting is that different agents possess identical reliability, i.e., $w_i = 1/N$. Nevertheless, in heterogeneous multi-agent systems, remarkable disparities exist in role capabilities, evidence types, and reasoning quality. Accordingly, this paper defines the comprehensive evidence score of Agent i for candidate conclusion y as:

$$S_i(y) = \alpha E_i(y) + \beta R_i + \gamma C_i(y) \quad (3)$$

where $E_i(y)$ is the evidence confidence score provided by Agent i for candidate conclusion y , R_i is the prior role reliability, and $C_i(y)$ denotes the consistency between this conclusion and independent evidence from other agents, satisfying $\alpha + \beta + \gamma = 1$. Further, dynamic normalization is performed for agents according to evidence support strength:

$$w_i(y) = \frac{\exp(S_i(y)/T)}{\sum_{k=1}^N \exp(S_k(y)/T)} \quad (4)$$

where T is the temperature parameter that controls the sharpness of weight distribution. The consensus probability of the final candidate conclusion is defined as:

$$P(y|E) = \sum_{i=1}^N w_i(y) P_i(y|E_i) \quad (5)$$

To prevent false consensus formed by repeated expression of erroneous viewpoints, this paper introduces divergence elimination conditions. When the weighted support for candidate conclusion y is lower than τ_d and at least one high-confidence conflicting piece of evidence exists, y is removed from the next-round discussion. Meanwhile, when the confidence gap between the top-ranked and second-ranked candidate satisfies $\Delta P > \tau_c$, discussion terminates in advance. Hence, multi-round collaboration incorporates dual controls of “error elimination” and “early stopping”.

To assign clear probabilistic interpretation to the consensus process, let the set of candidate conclusions be Y . $P_i(y|E_i)$ represents the conditional probability of Agent i toward candidate conclusion y . Taking $S_i(y)$ derived from Equation (3) as the reliability prior, normalized weights satisfy $\sum_i w_i(y) = 1$. Substitution into Equation (5) yields $P(y|E) \in [0, 1]$. Agents with high-quality evidence obtain higher $S_i(y)$, and their marginal contributions to the final consensus increase monotonically.

This paper further defines the consensus gap of candidate conclusion y as $\Delta(y) = P(y_1|E) - P(y_2|E)$, where y_1 and y_2 are the first-ranked and second-ranked candidate conclusions in the current round. When $\Delta(y) \geq \tau_c$ and evidence coverage $\Gamma(y) \geq \tau_e$, the system terminates debate in advance; otherwise, it proceeds to the next round. This rule unifies “accuracy priority” and “computation-cost constraints” within a single stopping condition.

Regarding communication overhead, let the size of original private data be V_{raw} and the size of desensitized evidence transmitted cross-domain be V_e . The privacy-exposure compression ratio is defined as $\rho = 1 - V_e/V_{\text{raw}}$. A value closer to 1 indicates that more raw data remains confined to the edge. Together with Accuracy, Hallucination and Latency, this

metric forms the core system-evaluation dimensions of this paper. The general-purpose local privacy-aware collaborative inference workflow is shown in Figure 2.

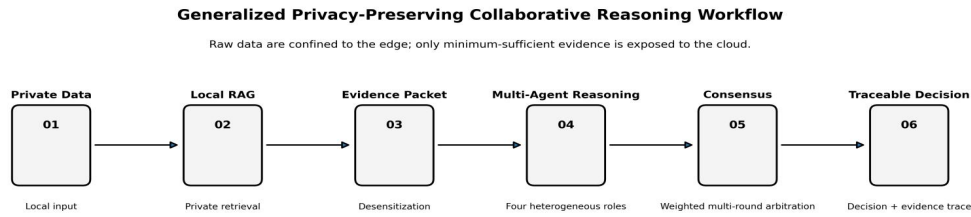


Figure 2 General Local Privacy-Sensitive Collaborative Inference Workflow

3 EXPERIMENTAL VALIDATION

3.1 Experimental Objectives and Verification Scenarios

To verify the effectiveness of GEHMAR, this paper develops an edge-cloud collaborative multi-agent medical expert consultation platform based on the above-mentioned architecture, and conducts systematic experiments concerning reasoning quality, privacy protection and system performance.

The medical expert consultation scenario involves multi-source heterogeneous information, private knowledge retrieval, complex reasoning and high-reliability requirements, which can comprehensively cover the key technical challenges targeted by the architecture in this paper. The experiments adopt 2000 text comparison test samples, 600 regression test samples, 30 Word/PDF private-knowledge documents and more than 200 multimodal images. Four systems are compared: Single LLM, Traditional Homogeneous Multi-Agent, Cloud-RAG Multi-Agent, and GEHMAR.

3.2 Experimental Setup

The experimental environment and data configuration are shown in Table 1.

Table 1 Experimental Environment and Data Configuration

Item	Configuration
Text evaluation	2 000 comparison cases + 600 regression cases
Private knowledge base	30 Word/PDF documents; approximately 3 800 chunks
Multimodal set	200 images
GPU / CPU	RTX 3090 24 GB / AMD EPYC 9754
Memory	60 GB RAM
Stress test	50 concurrent requests for 2 h
Streaming	SSE
Edge storage	IndexedDB

3.3 Baseline Methods and Evaluation Metrics

The baseline methods are shown in Table 2 and the evaluation indicators are shown in Table 3.

Table 2 Baseline Methods

Method	Key characteristics
Single LLM	Single model, direct reasoning
Traditional Homogeneous Multi-Agent	Same-role agents + majority voting
Cloud-RAG Multi-Agent	Cloud-side RAG + multi-agent reasoning
Ours	Edge local-private RAG + heterogeneous roles + weighted consensus

Table 3 Evaluation Metrics

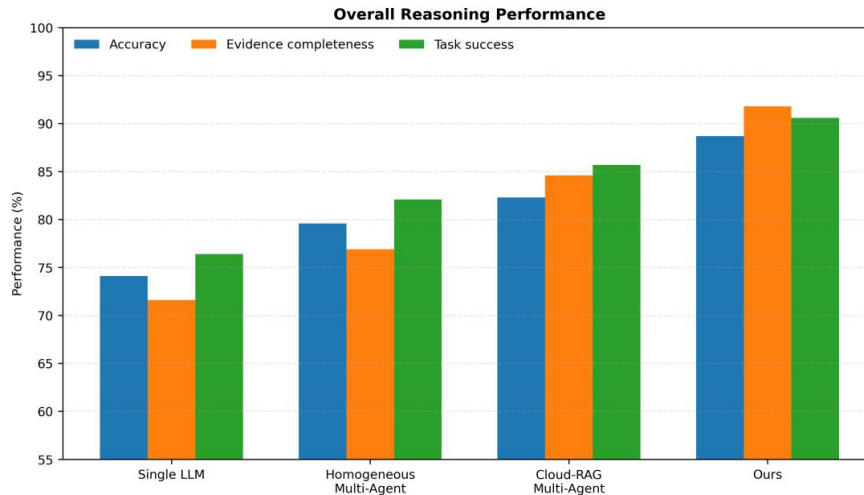
Category	Metric	Definition
Reasoning quality	Accuracy	Correct final decisions / total cases
Reasoning quality	Hallucination error	Unsupported or contradictory claims / total cases
Reasoning quality	Evidence completeness	Required evidence items adequately covered
Reasoning quality	Task success	Cases producing executable and traceable decisions
Privacy	Raw-data upload	Original/private data transmitted to cloud per task
System performance	End-to-end latency	Submission-to-decision elapsed time

3.4 Overall Performance Comparison

A comparative analysis of the overall performance of different collaboration architectures is presented in Table 4. A comparison of inference quality across different collaboration architectures is shown in Figure 3.

Table 4 Overall Performance Comparison of Different Collaborative Architectures

System	Accuracy (%)	Hallucination (%)	Evidence completeness (%)	Task success (%)	Raw upload (MB)	Latency (s)
Single LLM	74.1	16.8	71.6	76.4	1.82	1.72
Homogeneous Multi-Agent	79.6	12.4	76.9	82.1	2.35	2.48
Cloud-RAG Multi-Agent	82.3	9.8	84.6	85.7	1.96	2.71
Ours	88.7	5.9	91.8	90.6	0.12	2.92

**Figure 3** Reasoning Quality Comparison across Different Collaborative Architectures

As can be observed from Table 4 and Figure 3, GEHMAR achieves top performance on all three primary reasoning-quality metrics. Compared with the Traditional Homogeneous Multi-Agent, Accuracy increases from 79.6 % to 88.7 %, with an absolute improvement of 9.1 percentage points. Evidence completeness rises from 76.9 % to 91.8 %, improved by 14.9 percentage points. Task success increases from 82.1 % to 90.6 %, improved by 8.5 percentage points. The results demonstrate that performance gains do not stem from mechanically increasing agent quantity; instead, they mainly originate from heterogeneous role division and evidence-weighted arbitration. The comparison between the degree of privacy leakage and the hallucination error is shown in Figure 4.

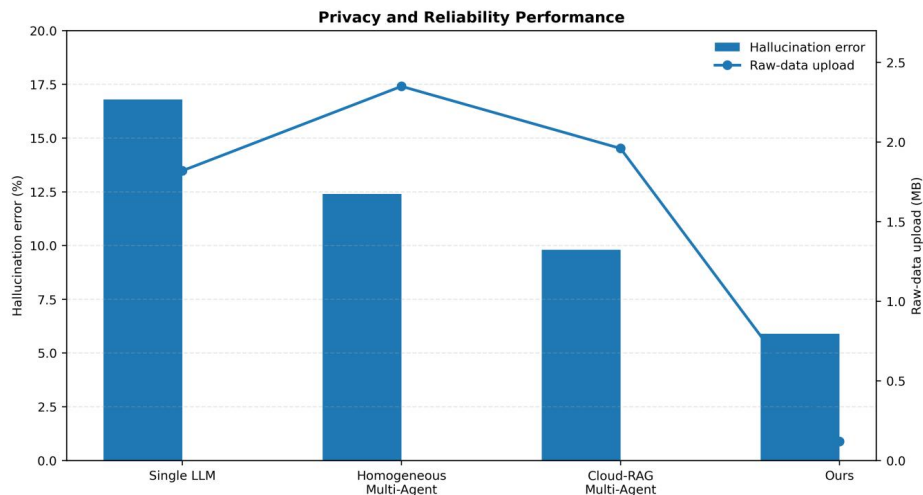
**Figure 4** Privacy Exposure and Hallucination Error Comparison

Figure 4 further reveals the most significant systematic benefits brought by the proposed architecture. The average raw-data upload volume of the Traditional Homogeneous Multi-Agent is 2.35 MB, while GEHMAR yields 0.12 MB, representing a reduction of 94.9 %. Meanwhile, the hallucination error rate decreases from 12.4 % to 5.9 %, reduced by 52.4 %. This indicates that migrating RAG retrieval and private-data processing to the Edge does not impair cloud-side reasoning capability. On the contrary, it reduces ungrounded generation via high-relevance evidence and arbitration mechanisms. A comparison of end-to-end inference latency across different architectures is shown in Figure 5.

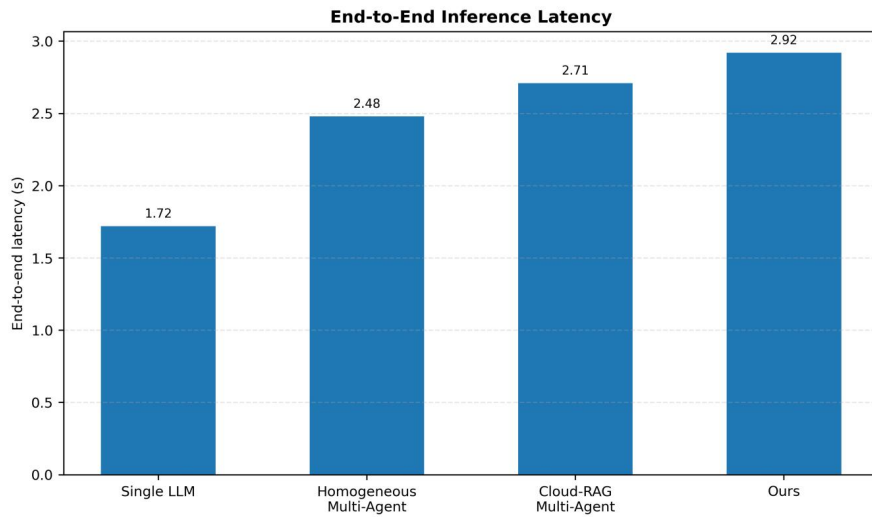


Figure 5 Comparison of End-to-End Inference Latency across Different Architectures

Privacy protection is not cost-free. GEHMAR achieves an end-to-end latency of 2.92 s, which is 0.44 s higher than the 2.48 s of the traditional homogeneous multi-agent setup. This overhead mainly comes from local retrieval, evidence compression and multi-round arbitration. For privacy-sensitive complex decision-making tasks, such latency increment remains acceptable, and it is traded for a dramatic drop in raw-data exposure. In further stress tests with 50 concurrent requests lasting 2 hours, the system TTFT P50/P95 reaches approximately 0.9 s / 2.4 s, and the request failure rate is 0.8 %. This validates that the cloud-side reasoning pipeline maintains stable service capacity under target load conditions.

3.5 Ablation Experiments

The core mechanism ablation experiments are presented in Table 5. The ablation results for the three core mechanisms are shown in Figure 6.

Table 5 Ablation Experiments for Core Mechanisms

Configuration	Accuracy (%)	Hallucination (%)
Full model	88.7	5.9
w/o Local-Privacy RAG	84.9	8.7
w/o Heterogeneous Roles	82.7	10.2
w/o Weighted Consensus	84.1	8.9

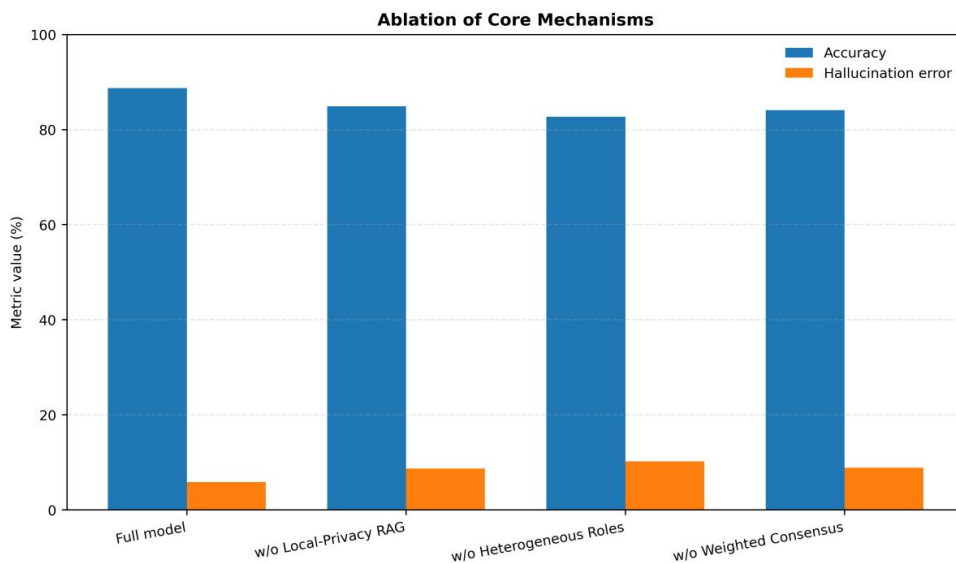


Figure 6 Ablation Results of Three Core Mechanisms

After removing Local-Privacy RAG, Accuracy drops by 3.8 percentage points. This shows that local private knowledge not only provides privacy isolation but also supplies task-relevant evidence. Removing Heterogeneous Roles reduces Accuracy by 6.0 percentage points and raises the hallucination rate to 10.2 %, which demonstrates that role heterogeneity constitutes an important source of complementary reasoning. Removing Weighted Consensus decreases

Accuracy by 4.6 percentage points, indicating that simple majority voting fails to fully exploit quality discrepancies among evidence produced by different agents. All three ablations cause performance degradation, which experimentally supports the synergistic relationship among the three core mechanisms in this paper.

4 CONCLUSIONS

This paper proposes GEHMAR, a general edge-cloud heterogeneous multi-agent collaborative reasoning architecture with local-privacy consensus. By constructing a three-layer decoupled framework of edge perception, desensitized communication and cloud-side heterogeneous reasoning, this paper realizes Local-Privacy RAG that keeps private raw data within the edge domain, together with a heterogeneous evidence-weighted multi-round consensus mechanism replacing simple majority voting. Experimental results based on a medical multi-expert consultation scenario demonstrate that the proposed architecture achieves better accuracy, evidence completeness and task success rate compared with mainstream multi-agent baselines. Meanwhile, it drastically cuts raw-data transmission volume under acceptable latency overhead. The architecture is decoupled from domain-specific logic, so it can be migrated to various privacy-sensitive complex-decision scenarios such as industrial fault diagnosis and financial risk assessment, showing favorable practical application feasibility.

Nevertheless, this work still has certain limitations. The current experiments are mainly conducted on a single dataset, and the compression of evidence abstracts may bring minor information loss. Future research will introduce multi-source cross-domain measured data for further validation. Moreover, this paper will extend the present framework to support harmonic-like interference and multi-modal complex evidence processing, so as to further improve the generalization ability of edge-cloud multi-agent collaborative reasoning in real-world complex environments.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported by the National College Students' Innovation and Entrepreneurship Training Program under Grant No. 202613667005. The authors would like to thank Guangzhou College of Commerce for providing experimental facilities and academic support. Finally, we thank all team members and reviewers for their valuable contributions to this study.

REFERENCES

- [1] Li J, Cheng X, Zhao W, et al. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. *Proceedings of EMNLP*, 2023: 6449-6464.
- [2] Cai G, Tian R, Yang L, et al. Efficient Inference for Edge Large Language Models: A Survey. *Tsinghua Science and Technology*, 2025.
- [3] Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 2025, 43(2): 42.
- [4] Chen K, Zhou X, Lin Y, et al. A Survey on Privacy Risks and Protection in Large Language Models. *Journal of King Saud University Computer and Information Sciences*, 2025, 37: 163.
- [5] Luo J, Zhang Y, Zhang Z, et al. Cryptography-Based Privacy-Preserving Large Language Models: A Lifecycle Survey of Frameworks, Methods, and Future Directions. *Artificial Intelligence Review*, 2026, 59: 64.
- [6] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine. *Nature Medicine*, 2023, 29: 1930-1940.
- [7] Wiest I C, Ferber D, Zhu J, et al. Privacy-preserving large language models for structured medical information retrieval. *npj Digital Medicine*, 2024, 7: 257.
- [8] Piccialli F, Chiaro D, Qi P, et al. Federated and Edge Learning for Large Language Models. *Information Fusion*, 2025, 117: 102840.
- [9] Aina O K, Ha S. Deep reinforcement learning for multi-agent coordination. *Artificial Life and Robotics*, 2025, 31(2): 1-11.
- [10] Maktabifard A, Földes D. Cooperative driving route planning based on asymmetric multi-agent path planning problem with limited service area constraints. *Applied Soft Computing*, 2025, 185: 113959.