An Agentic Multimodal Framework for Reliable Medical Reasoning under Privacy Constraints

Hao Lin

{ hl563@sussex.ac.uk }

Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou, China

Abstract. As artificial intelligence evolves, medical question-answering systems are changing to generative artificial intelligence (AIGC). Nevertheless, the general large-scale language models cannot find clinical applications yet because of the lack of multimodal perception, hallucinations, no clear reasoning chains, and the rigid data privacy requirements, which limit the deployment to clouds. This paper proposes a Multimodal Medical Agent (MMA) framework to address these challenges. It is based on a visual encoder that is connected to the LLaVA-Med backbone and introduces Retrieval-Augmented Generation (Self-RAG) and ReAct mechanisms. Based on the 4-bit quantization technology, this research is able to realize a low-resource, offline implementation of the end-to-end system. Experimental results demonstrate that the system scores an accuracy of 74.0 per cent on the Slake multimodal medical question-answering dataset, representing a notable increase of 29 percentage points over the baseline model. Such findings confirm that the suggested framework could be successfully used to improve the multimodal medical reasoning performance and ensure the security of data, which proves that this technology could be applied to the computer-aided diagnosis context in the real-world setting.

Keywords: Multimodal Large Language Models; Medical Agents; Retrieval-Augmented Generation (RAG); Computer-Aided Diagnosis; On-device Deployment.

1 Introduction

The dearth and lack of uniformity in medical resource allocation are also a real threat to the health of the population across the world. In recent years, the Large Language Model (LLM) explosion has provided new opportunities for automated medical support. The models characterized by the Med-PaLM 2 and Med-Gemini proved to be impressive in terms of clinical knowledge retrieval and reasoning, which is the first step towards turning the AI doctor into more than a concept [1, 2]. The foundation models are slowly becoming a process of generalist medical artificial intelligence that will transform clinical diagnostic and treatment processes.

General models are very effective in text processing. Nonetheless, to adjust to the specialization of the medical industry, scientists have developed instruction-following models

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365100

like ChatDoctor and Huatuo [3, 4]. Although all these have been developed, the computer implementation of these models in real clinical diagnosis continues to encounter three fundamental limitations.

To start with, a single modality restricts the ability to understand diagnosis. Image data such as CT-scans, MRI and pathology slides are critical to clinical diagnosis. Physically, these models are incapable. First-generation medical LLMs can only receive text input and have a perceptual limitation of being blind to images [3, 4]. LLaVA-Med tries to match images with texts. Its use in complicated clinical reasoning is however at the exploratory phase [5].

Second, hallucinations that involve facts pose health risks to the medical field. Thanks to the characteristics of probabilistic generation, the generative models unproblematically create treatment plans or drug names that do not exist. This is unethical in critical medical situations [6].

Lastly, it does not have reasoning chains or barriers to deployment, which limits its application. Multi-step reasoning and expediency of outside knowledge are needed when making complex clinical decisions. Even a basic input-output mode will not be able to model the stringent Chain of Thought doctors use. More so, due to the sensitivity of medical data, high-cost models relying on cloud computing are hard to implement in this particular instance. Thus, the low-resource solutions that can be deployed locally urgently are required [7].

To deal with the issues mentioned above, the presented research suggests a medical auxiliary diagnostic system on Multimodal Large Language Models and based on an Agentic Architecture. The main contributions of the study are as follows:

Building a Multimodal Perception Module: LLaVA-Med is embraced as the basis in order to consolidate the encoding of medical images and texts. This is an overcoming of the shortcomings of the traditional models that fail to interpret images.

Presentation of the Self-RAG Reflection Mechanism: With the help of a Retrieve-Generate-Reflect workflow, the model independently ensures the validity of external knowledge and content generated. This system is useful in the suppression of factual hallucinations.

ReAct Diagnostic Closed-Loop design: The diagnostic process decomposed into four stages of Perceive-Plan-Retrieve-Reflect is then designed by interpreting APIs on drug queries and knowledge base queries. This structure increases the interpretability of clinical decision-making.

End-to-end Local operation with Low-Resource Offline Deployment: Together with 4-bit quantization technology, the system can end-to-end locally operate on consumer-grade GPUs. This application architecture provides privacy and security of data.

2 Related Work

The most advanced medical artificial intelligence studies carried out today have subsumed pure text-based Large Language Models into multimodal studies. The field is moving towards the agents with the ability to plan autonomously. Nevertheless, even models at every level

have some levels of limitations in their perception abilities, factual accuracy and logic of reasoning. This chapter also reviews and analyses the existing work on these three dimensions.

2.1 Large Language Models in the Medical Field

In order to fit the specialization in the medical field, researchers have come up with different strategies of fine-tuning. Web ChatDoctor and Huatuo (BenTsao) have previously tried preliminary efforts via Supervised Fine-Tuning (SFT) [3, 4]. However, there are two challenges that are still faced by most of these models.

The first is the issue of blindness to modality. Physically, these models are incapable of handling important imaged data like CT and MRI scans [5].

Second, when it comes to making complex decisions under high risks, there is a high tendency of generative models to experience what can be termed as factual hallucinations [6]. Moreover, they do not have self-check tools on generated materials. This lack seriously limits their use in practice in actual clinical practice [8].

2.2 Multimodal Medical Large Models

One of the research areas nowadays is to break the boundaries between images and text. Image-text alignment has found its basis in CLIP and LLaVA [9]. LLaVA-Med has also shown the realization of low cost fine-tuning in biomedicine [5]. Besides, PMC-CLIP and CHIEF have demonstrated advancements in medical image pre-training [10, 11]. Although these developments have been made, essentially the above models are still question-answering models. They do not have the long chain of logical reasoning (Chain of thought) that doctors possess [7]. Moreover, they do not have the capacity of agents to make independent calls to external tools in the course of diagnosis [12].

2.3 Medical Agents and Hallucination Suppression

The emphasis on agents with the ability to plan on their own has slowly replaced studies on Chain-of-Thought (CoT) and ReAct reasoning schemes with agents [7, 13]. Li et al. suggested an AgentHospital that builds a complete simulation in a hospital setting. This piece of work showed that self-evolution could be used by medical agents to keep enhancing their diagnostic abilities continuously [12]. On the same note, HuatuoGPT-Vision also delved into the paradigms of construction of general multimodal medical agents. This model focused on the essence of visual encoding in complicated diagnostic tasks [14]. These trials show that the harnessing of LLMs to act as the clinical reasoning agent has huge potential. It is possible to solve complex problems with the help of planning capabilities, which could be divided into sub-steps: symptom acquisition, knowledge retrieval, and diagnosis generation.

Asai et al. investigated Self-RAG to resolve the problem of hallucination, and this solution is novel as it proposes introducing a self-reflection process. This enables the model to independently analyze the topicality and genuineness of retrieved materials [8]. Further on this, Xiong et al. established MedRAG, a method to maximize the retrieval-augmented generation technology in medical contexts. It is a more focused and improved solution to the existing field [15].

The current medical agents are however largely restricted to plain-text or simulated environments. They are not able to handle multimodal clinical data of the real world. Based on this, the proposed MMA framework will fill this gap. With an intense combination of multimedia perceptions with a Retrieve-Reflect closed loop, the system itself can be viewed as a kind of auxiliary diagnostic apparatus which is capable of both interpreting images, as well as reasoning in rigorous way.

3 Method

The suggested Multimodal Medical Agent system (MMA) comprises three main components, namely, a multimodal perception layer, an agent brain, and external tools, including a knowledge base. In order to deal with fundamental issues in existing models, namely, single modality and absence of reasoning chains, this system is planned to overcome the conventional constraints. The technical approach of this system has considerably differentiated benefits compared to HuatuoGPT, which only processes texts [4], or LLaVA-Med, which is a simple image-text Q&A process [5].

To begin with, according to the ReAct (Reasoning + Acting) paradigm, a closed-loop of dynamic reasoning is created. This enables the system to get past mere input-output mapping. Rather, it modifies the entire rational thought process of a physician: having images, designing algorithms, invoking utilities, and checking outcomes [13].

Second, a dual anti-hallucination mechanism is developed by incorporating Self-RAG with the external body of knowledge. This implements a Retrieve-Reflect dual check requirement before the generation of a diagnosis. As a result, both the quality and medical advice become as low as possible [8].

Lastly, since civil service data privacy is very strict at the clinic and the deployment of the cloud is expensive, the notion of 4-bit quantization technology and a training-free architecture is presented. End-to-end offline execution on consumer-grade GPUs is effectively achieved. This will allow patient information to be kept locally and adequately overcome deployment obstacles in resource-limited primary medical settings.

3.1 Overall System Architecture

The system is built with the use of the ReAct (Reasoning + Acting) paradigm [13]. The process of the work follows the following order. The text complaints and an image of a medical problem that are presented by the user are first encoded through the multimodal perception layer. After that, a plan of diagnostics (Thought) is created by the agent's brain. According to this plan, actions (Action) are performed by the system, i.e., querying drug databases or accessing guidelines. The last issue is the production of the final diagnostic recommendation based on the Self-RAG mechanism by means of combining the observations (Observation). Figure 1 represents the general structure of the Multimodal Medical Agent.

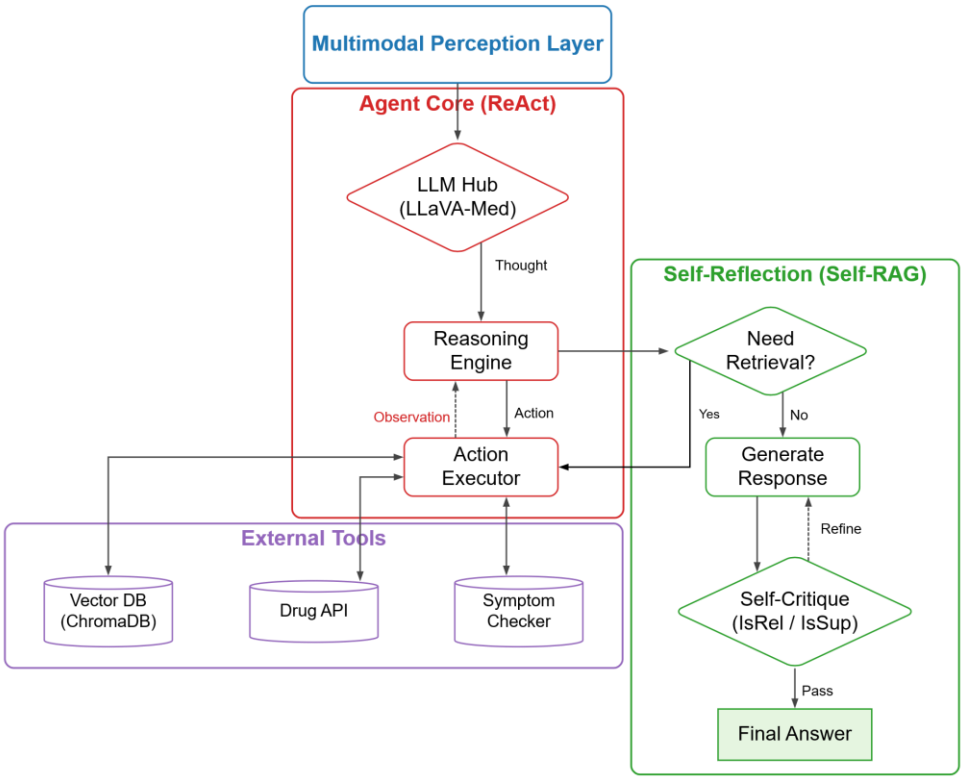

Fig. 1. Overall architecture of the Multimodal Medical Agent (MMA) (Picture credit: Original)

3.2 Multimodal Perception Module

The architecture of LLaVA-Med is designed to process patient-uploaded imaging data [5]. To extract image features X_v [9], A pre-trained visual encoder (CLIP ViT-L/14) is used. These are mapped to the text embedding space with a two-layer MLP projection layer. The visual token sequence H_v is obtained as a result of this process.

$$H_v = MLP\big(Encoder(X_v)\big) \tag{1}$$

This series is added to the input tokens of an input given by users to act as the input to the Large Language Model (LLM) to provide the model with the ability to see. LLaVA-Med has been chosen as the common backbone model in this study. Besides the incorporation of CLIP visual encoder, it has also fine-tuned in the biomedical field, to interpret complex medical images directly and can be used in verbal communication with natural language. In the case of highly specialized images like pathology slides, the multimodal alignment feature of LLaVA-Med provides the correctness of detail capture.

3.3 Agent Planning and Tool Invocation

Referring to the agent planning philosophy in AgentHospital, a library of clinical tools is presented in the present study as they facilitate complicated diagnostic and treatment actions

[12]. In particular, the system incorporates the DrugAPI to favorably access drug indications, contraindications, and dosages. At the same time, a KnowledgeRetriever is made based on the strategies of optimization of MedRAG. It is an innovative component using a vector database to optimize medical knowledge and search terms into indexed information to maximize medical knowledge [15]. In addition, the system has rule based SymptomChecker to implement standardized initial symptom screening logic. The Agent can be autonomous in the path planning of tool invocation by the clinical context. The model then, based on patient input, generates a reasoning intent of differential diagnosis (Thought), followed by access to the tool of knowledge base retrieval (Action: KnowledgeRetriever). According to the feedback (Observation) that is being retrieved, the system also invokes a drug query action (Action: DrugAPI) and eventually becomes the synthesizer of all pieces of information to produce clinical recommendations (Final Answer).

3.4 Self-RAG Reflection Mechanism

In a bid to moderate medical finding and reasoning efficiency with limited resources, this research deploys a lean version of Self-RAG. The reflection mechanism is rebuilt into a response-level (unlike the original architecture), Generate-Evaluate-Refine closed loop. It is more applicable in the on-device deployment.

To begin with, the system uses a dynamic retrieval strategy which is dynamic and key word driven. Depending on the situation in context x, the model recognizes the most important medical entities (e.g., "Nodule", "Mass"). The external knowledge base is only triggered to access the relevant criteria of differential diagnosis when the high-risk diagnostic terms are identified. This is a good strategy that prevents extraneous noise created by other irrelevant retrievals.

This is followed by the introduction of a Skeptic Medical Auditor module to carry out post-hoc checking on the entire initial draft $\hat{y}$ produced by the model. This module assesses the likeness of the answer with the evidence that is retrieved ($IsRel$) and the level of support by visual facts ($IsSup$). The last output decision comes about decisions made in accordance with a threshold correction, and it is formally defined as follows:

$$\hat{y}_{\text{final}} = \begin{cases} \hat{y}_{\text{refined}}, & \text{if } IsSup(\hat{y}) < \tau \\ \hat{y}, & \text{otherwise} \end{cases} \quad (2)$$

In this case, τ is the confidence threshold (in this trial taken to 90). This algorithm will automatically detect and remove low-confidence (hallucinated) content through this mechanism. Although this is done by guaranteeing the rigor of the diagnostic recommendations, it minimizes the inference overhead of the system.

4. Experimental Design and Expected Results

4.1 Datasets

Instead of thoroughly testing the clinical reasoning and multimodal understanding abilities, two representative datasets were chosen.

The MedQA-USMLE dataset consists of questions from the United States Medical Licensing Examination. It is mainly applied to evaluate the proficiency of clinical logical reasoning in a pure text situation [16]. In the meantime, the Slake dataset is a semantically-labeled, knowledge-enhanced medical visual question answering benchmark. It is used in testing the multimodal comprehension skills of Chinese-English bilingual settings [17].

As to the evaluation framework, a hybrid system is taken with quantitative metrics and manual assessment. In particular, Accuracy can be used as a measure of multiple-choice tasks. At the same time, there is an introduction of a Hallucination Rate measure. The judge model used to examine the responses to identify fallacies in facts is known as Gemini 3 Pro. Moreover, an Invocation Success Rate of the tool is developed to measure the accuracy with which the Agent invokes external APIs autonomously.

4.2 Experimental Setup

The NVIDIA RTX 3080 Laptop was chosen as a test platform to check the possibility of achieving system deployment in the primary medical facility. The backbone model was adopted as LLaVA-Med-7B. Using the 4-bit quantization technology, video memory usage was reduced in the area of operation that could be implemented within a local setting. This environment is like a high-privacy offline deployment environment as found in the real world. This study adopted a training-free architecture. ChromaDB vector database was used to build a knowledge base, whereas the BGE model was used to retrieve localized vectors.

4.3 Experimental Results

The Slake dataset was set to test the generalization ability of the MMA architecture in managing the complex case of multimodal. Taking into account the requirements on the validation to deploy on the device, Sampled Subset of 50 cases was picked randomly out of the test set. Several imaging modalities are included in this subset, such as CT, MRI and X-Ray.

The Closed-ended questions (e.g., confirmation on the anatomical location of the disorder, yes/no judgment) and Open-ended ones (e.g., diagnosis of the disease, etiology analysis) were evaluated. The control point was the LLaVA-Med-7B version (Zero-shot) that does not include the Agent mechanism.

As shown by the experimental findings (Table 1), once the MMA architecture was introduced, the system had an aggregate accuracy of 74.0 on the Slake data. This is profoundly better than the 29 percentage points as the baseline.

Table 1. Main Results on the Slake Dataset

Method	Overall Acc.	Closed-ended Acc.	Open-ended Acc.	Latency/Item
Baseline (LLaVA-Med-7B Zero-shot)	45.0%	52.0%	38.0%	~3.2s
MMA-Agent (Ours)	74.0%	81.2%	66.0%	~12.5s
Improvement	+29.0%	+29.2%	+28.0%	-

First, in terms of several visual queries in the Slake dataset on the location of lesions or their properties, MMA is able to exhibit much better multimodal comprehension skills. Taking advantage of the perceive-reflect, closed loop of the Agent, MMA better utilizes the visual encoder to capture minute features. Therefore, it scores a high of 81.2 percent on closed-ended questions.

Second, native models tend to produce generic answers to open-ended queries in case of complex reasoning activities because of insufficient background knowledge. Contrary to this, the proposed system has been suggested to help to fill the gaps in pathological justifications, generating the images by accessing an external stock of knowledge. This makes a breach in complicated reasoning abilities. Consequently, the precision of the open-ended questions gets improved by almost 28 percentage points.

Lastly, initial experiments with these results support the usefulness of the suggested architecture. They show that with a consumer-grade marginally high-quality GPU (RTX 3080), the accuracy of diagnostic performance is close to that of large models running on clouds with a well-established Agent scheduling.

5. Conclusions

The paper suggests a concept of a medical auxiliary diagnostic system (MMA) that incorporates multimodal knowledge with the autonomous agent technology. Effectively solving two fundamental pain points of general large models applied in medical contexts via the ReAct planning architecture and the Self-RAG reflection mechanism, this system manages to solve the so-called lack of multimodal perception and the problem of factual hallucinations. These are some of the issues that are addressed both theoretically and practically. Experimental findings show that the architecture has an accuracy of 74.0% on typical samples of the Slake data. This results in a performance that is significantly better than the baseline model (29.0 percentile points) and confirms that large medical agents run on an extremely privatized architecture can be built on consumer-only GPUs.

However, there are still some limitations of this research. To begin with, being a training-free architecture, the performance of the system entirely relies on the quality of the external knowledge base and the corresponding level of retrieval algorithms. In experiments, it was found that repetitive responses can be obtained by the model when working with uncovered rare diseases when retrieval noise is used (the long-tail effect). Second, multi-step reasoning and reflection stages were introduced to be able to trade off high accuracy of reasoning. This leads to a comparably high average query latency (about 12.5 seconds), which would have to be optimized in case of emergency events with very high real-time needs. Moreover, the validation at the moment is mainly grounded on general datasets. Human-machine collaboration tests on a large scale and on real clinical set-ups are not carried out yet.

The three directions will be addressed in the future. To minimize latency, first, reasoning efficiency will be optimized through the investigation of lightweight quantization schemes and parallel reasoning schemes. Second, the library of clinical tools is to be enlarged with the possibility to include the Electronic Health Record (EHR) analysis and multidimensional image processing features. Third, application-specific multimodal knowledge bases (like the

Radiopaedia image-text library) will be built to take the place of general text knowledge bases. This is to increase the accuracy and usefulness of retrieval-augmented generation further.

References

[1] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., et al.: Toward expert-level medical question answering with large language models. Nature Medicine. pp. 943 – 950 (2025)

[2] Saab, K., Tu, T., Weng, W. H., Tanno, R., Stutz, D., Wulczyn, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)

[3] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: Chatdoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. Cureus. Vol. 15, No. 6 (2023)

[4] Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T.: Huatuo: Tuning LLaMA model with Chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023)

[5] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems. pp. 28541 – 28564 (2023)

[6] Kim, Y., Jeong, H., Chen, S., Li, S. S., Park, C., Lu, M., et al.: Medical hallucinations in foundation models and their impact on healthcare. arXiv preprint arXiv:2503.05777 (2025)

[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems. pp. 24824 – 24837 (2022)

[8] Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint (2024)

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In International Conference on Machine Learning. pp. 8748 – 8763. PMLR (2021)

[10] Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525 – 536. Springer Nature, Cham (2023)

[11] Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature. pp. 970 – 978 (2024)

[12] Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., et al.: Agent hospital: A simulacrum of hospital with evolvable medical agents. arXiv preprint arXiv:2405.02957 (2024)

[13] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. In International Conference on Learning Representations (2022)

[14] Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G. H., et al.: HuatuoGPT-Vision: Towards injecting medical visual knowledge into multimodal large language models at scale. arXiv preprint arXiv:2406.19280 (2024)

[15] Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics: ACL 2024. pp. 6233 – 6251 (2024)

[16] Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., Szolovits, P.: What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences. Vol. 11, No. 14, pp. 6421 (2021)
[17] Liu, B., Zhan, L. M., Xu, L., Ma, L., Yang, Y., Wu, X. M.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In IEEE International Symposium on Biomedical Imaging. pp. 1650 – 1654. IEEE (2021)