npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01550-0

Enhancing diagnostic capability with multi-agents conversational large language models

Check for updates

Xi Chen[1,2,10], Huahui Yi[3,4,5,10], Mingke You[1,2,10], WeiZhi Liu[6], Li Wang[6], Hairui Li[7], Xue Zhang[6], Yingman Guo[8], Lei Fan[1,2], Gang Chen[1,2], Qicheng Lao[9], Weili Fu[1,2] ✉, Kang Li[3,4,5] ✉ & Jian Li[1,2] ✉

Large Language Models (LLMs) show promise in healthcare tasks but face challenges in complex medical scenarios. We developed a Multi-Agent Conversation (MAC) framework for disease diagnosis, inspired by clinical Multi-Disciplinary Team discussions. Using 302 rare disease cases, we evaluated GPT-3.5, GPT-4, and MAC on medical knowledge and clinical reasoning. MAC outperformed single models in both primary and follow-up consultations, achieving higher accuracy in diagnoses and suggested tests. Optimal performance was achieved with four doctor agents and a supervisor agent, using GPT-4 as the base model. MAC demonstrated high consistency across repeated runs. Further comparative analysis showed MAC also outperformed other methods including Chain of Thoughts (CoT), Self-Refine, and Self-Consistency with higher performance and more output tokens. This framework significantly enhanced LLMs' diagnostic capabilities, effectively bridging theoretical knowledge and practical clinical application. Our findings highlight the potential of multi-agent LLMs in healthcare and suggest further research into their clinical implementation.

Recent advancements in large language models (LLMs) have notably enhanced their capabilities in the medical field, leading to increased exploration of their potential applications[1]. These models are equipped with vast medical databases and advanced analytical algorithms, offer promising solutions to these challenges[2]. These models have shown proficiency in simple medical tasks such as answering medical knowledge queries and diagnose common diseases, and warrant further research to test their effectiveness in handling more practical and complex medical tasks[3,4]. Among them, providing accurate diagnosis has always been an important and practical medical need.

Diagnosis of rare diseases is among the most complex and challenging diagnostic tasks. The prevalence of rare diseases ranges from 5 to 76 cases per 100,000 individuals[5]. The low prevalence of these diseases often results in a scarcity of specialized knowledge, making accurate diagnosis difficult, thereby delaying proper treatment[6]. Additionally, the complexity and variability of symptoms can lead to frequent misdiagnoses or delayed diagnos[7,8].

Although LLMs such as GPT-4 demonstrate substantial proficiency in medical knowledge and some potential application, their performance in complex real-world clinical scenarios such as disease diagnosis remains questioned. A previous study has demonstrated the limitations of LLMs in the diagnosis of complex cases[2]. Therefore, an increasing number of studies are focusing on how to better leverage LLMs' own training data and inherent capabilities to improve their performance in practical medical tasks[9,10].

Multi-agent systems are one such attempt. In the context of a large language model, an agent refers to a system capable of receiving input and performing actions to achieve specific goals. For example, when interacting with ChatGPT, the user is engaging with a single-agent model. The multi-agent framework is an innovative approach where multiple digital agents work together and execute tasks through interactions among themselves. This technique significantly enhances the capabilities of LLMs for managing complex tasks, including solving mathematical problems and performing

[1]Sports Medicine Center, Department of Orthopedics and Orthopedic Research Institute, West China Hospital, West China School of Medicine, Sichuan University, Chengdu, Sichuan, China. [2]Department of Orthopedics and Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, Sichuan, China. [3]West China Biomedical Big Data Center, Med-X Center for Informatics, Sichuan University, Chengdu, Sichuan, China. [4]Sichuan University Pittsburgh Institute, Chengdu, Sichuan, China. [5]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [6]West China School of Medicine, Sichuan University, Chengdu, Sichuan, China. [7]Department of Plastic Surgery, West China Tianfu Hospital, West China Hospital, Sichuan University, Chengdu, Sichuan, China. [8]Kidney Disease Center, First Affiliated Hospital, College of Medicine, Zhejiang University, Key laboratory of Kidney Disease, Hangzhou, Zhejiang, China. [9]Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. [10]These authors contributed equally: Xi Chen, Huahui Yi, Mingke You.
✉e-mail: foxwin2008@163.com; likang@wchscu.cn; lijian_sportsmed@163.com

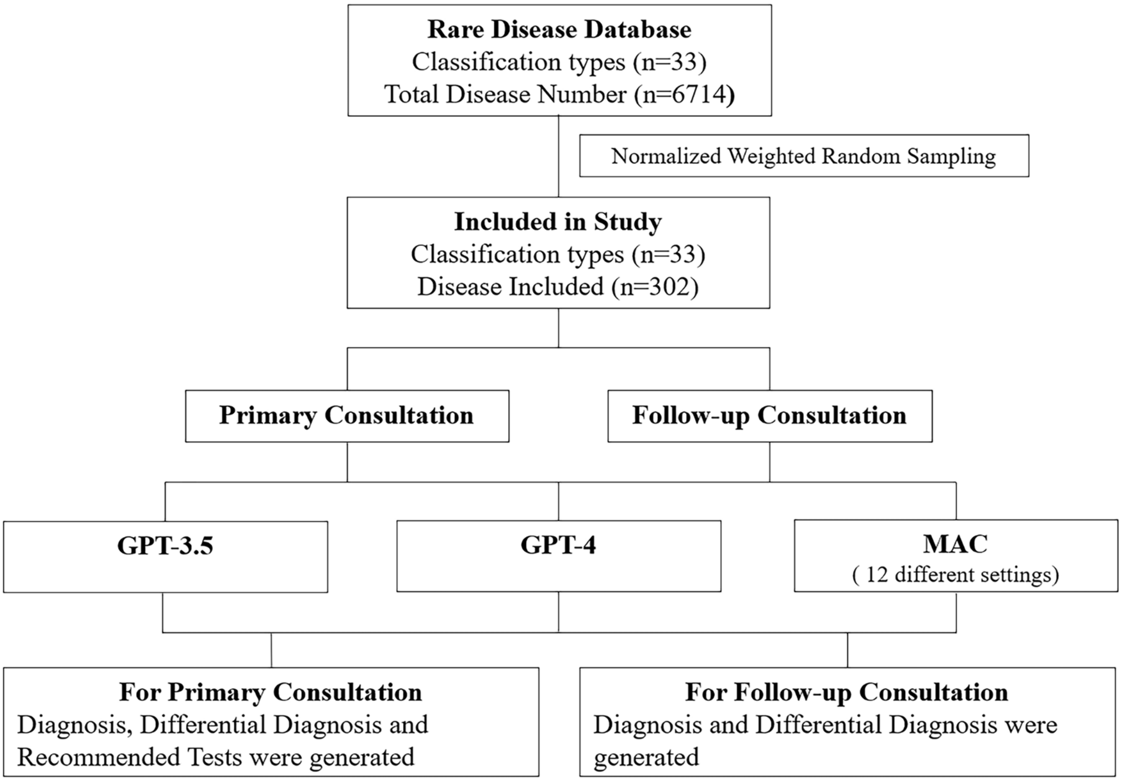

Rare Disease Database
Classification types (n=33)
Total Disease Number (n=6714)
Normalized Weighted Random Sampling
Included in Study
Classification types (n=33)
Disease Included (n=302)
Primary Consultation
Follow-up Consultation
GPT-3.5
GPT-4
MAC
(12 different settings)
For Primary Consultation
Diagnosis, Differential Diagnosis and Recommended Tests were generated
For Follow-up Consultation
Diagnosis and Differential Diagnosis were generated

Fig. 1 | Process of disease selection and testing. The researchers first identified the rare disease database. Then, they performed normalized weighted random sampling, from which 302 cases were selected. Each case was curated into primary consultation setting and follow-up consultation setting. GPT-3.5, GPT-4, and MAC were tested.

retrieval-augmented code generation[11–14]. The adoption of multi-agent framework may facilitate dynamic and interactive diagnostic processes through multi-agent conversation (MAC), where agents would discuss the same question and finally reach consensus on the output. In this case, different agents can simulate the collaborative nature of a medical team, simulating a Multi-Disciplinary Team (MDT) discussion that is commonly adopted in clinical practice to solve complex medical tasks. By facilitating an in-depth analysis that single-agent models may not achieve, they have the potential to improve the performance of LLMs in the diagnosis of rare diseases. Therefore, MAC may serve as a valuable tool that could help doctors analyze patient information and provide useful second opinions.

This study aims to develop a multi agent conversation (MAC) framework that can be used in clinical practice to perform diagnostic tasks. This study also compares the knowledge base and diagnostic capabilities of GPT-3.5, GPT-4, and the MAC.

Results

The study flow is shown in Fig. 1. The Multi-Agent Conversation Framework was developed (Fig. 2) and was tested on curated clinical cases simulating real-world clinical consultations (Fig. 3).

Study sample

This study included 302 kinds of rare disease from 33 different disease categories. One to nine kinds of rare disease were randomly selected for each category. Details of the sampled diseases and their corresponding clinical case reports are provided in Supplementary Table 1.

Performance on disease specific knowledge

GPT-3.5, GPT-4, and MAC achieved an average score above 4 across all testing aspects, including inappropriate/incorrect content, omission, likelihood of possible harm, extent of possible harm, and bias. The results of disease-specific knowledge performance are shown in Fig. 4.

Performance of MAC on diagnostic ability

MAC significantly outperformed single-agent models GPT-3.5 and GPT-4 in terms of diagnostic accuracy and helpfulness of further recommended test in both primary and follow-up consultation. Further analysis was performed to investigate factors potentially influencing MAC's performance, including changing base model, varying the number of doctor agents, excluding the supervisor agent and assigning case-specific specialties to doctor agents. Detailed results of GPT-3.5, GPT-4, and different subgroups within the MAC framework are listed in Table 1, Supplementary Tables 2 and 3 for primary consultation, and Table 2, Supplementary Tables 4 and 5 for follow-up consultation. The comparisons between MAC and single-agent models are shown in Fig. 5 for primary consultation and in Fig. 6 for follow-up consultation.

MAC utilizing either GPT-3.5 or GPT-4 as the base model significantly outperformed their respective standalone versions. However, GPT-4 proved superior to GPT-3.5 when used as the base model for MAC. In primary consultations with four doctor agents, GPT-4 achieved higher accuracy for the most likely diagnosis (34.11% vs 24.28%), possible diagnoses (48.12% vs 36.64%), and helpfulness of further diagnostic tests (78.26% vs 77.37%). Similar trends were observed in follow-up consultations, with GPT-4 showing approximately 10% improvement over GPT-3.5. The results are shown in Tables 1, 2, and Fig. 7.

The study examined the effect of varying the number of doctor agents (2 to 5) on the multi-agent framework's performance. Using GPT-4 as the base model, the Most Likely Diagnosis accuracy in primary consultations was 31.31% for 2 agents, 32.45% for 3 agents, 34.11% for 4 agents, and 31.79% for 5 agents. In follow-up consultations, the accuracy was 51.99%, 53.31%, 53.86%, and 50.99% for 2, 3, 4, and 5 agents respectively. Similar trends were observed for the Possible Diagnosis accuracy and Further Diagnostic Tests Helpful Rate metrics. With GPT-3.5 as the base model, 4 doctor agents also produced optimal performance in primary consultations.

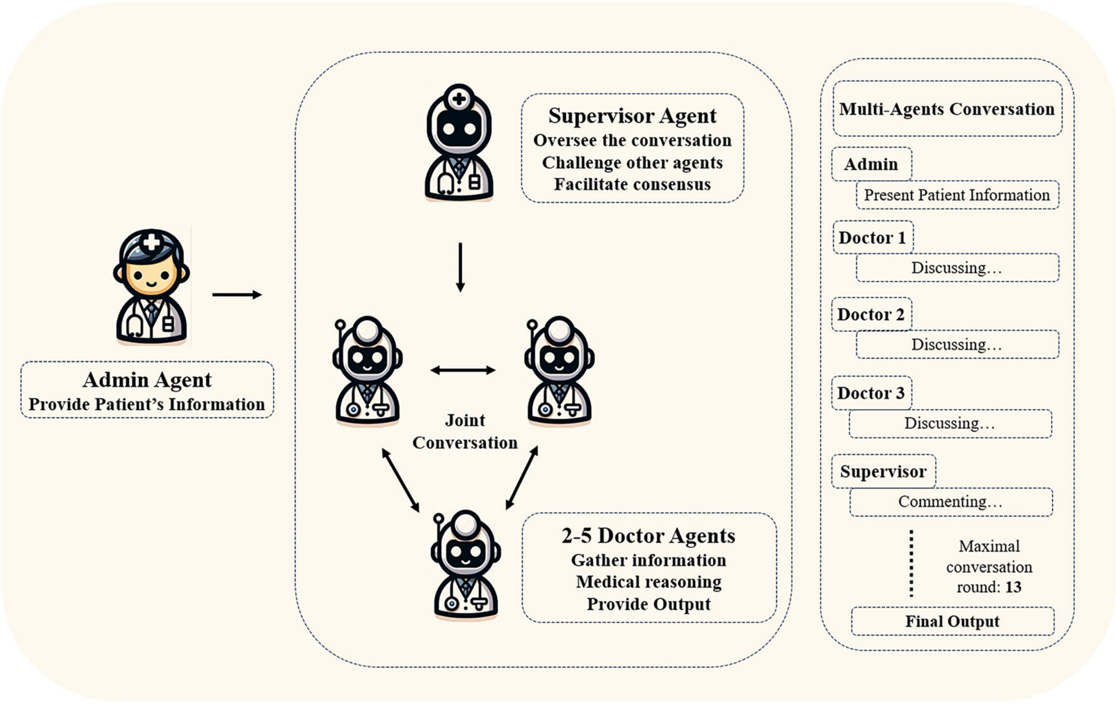


Fig. 2 | The multi-agent conversation (MAC) framework. The admin agent provides the patient's information. The supervisor agent initiates and supervises the joint conversation. The three doctor agents jointly discuss the patient's condition. The final output is provided either when consensus is reached or after a maximum of thirteen rounds of conversation.

However, for follow-up consultations using GPT-3.5, configurations with 4 and 5 doctor agents were not feasible due to token output limitations. The results are shown in Tables 1, 2, and Fig. 7.

The effect of excluding the supervisor agent from the MAC framework was investigated. In primary consultations using GPT-4 as the base model, the exclusion of the supervisor agent resulted in decreased accuracy for both the most likely diagnosis (34.11% with supervisor vs. 32.67% without) and possible diagnoses (48.12% vs. 45.47%). The accuracies for further diagnostic tests remained similar (78.26% vs. 78.04%). When using GPT-3.5 as the base model, the exclusion of the supervisor had less impact on performance. In follow-up consultations, excluding the supervisor agent also led to reduced performance for both the most likely diagnosis and possible diagnoses.

The study investigated the impact of assigning case-specific specialties to doctor agents on MAC performance. Using GPT-4, the four most relevant specialties were identified for each case, and doctor agents were instructed to assume these specialist roles. The results showed that the assignment of specialties to doctor agents did not significantly improve MAC's performance compared to the standard configuration without assigned specialties.

Reliability analysis

The MAC framework's reliability was evaluated through three repeated testing rounds across six settings. Results demonstrated minimal performance variation across testing rounds. Fleiss' kappa analysis showed moderate agreement (>0.4) in 23 out of 28 evaluations, and fair agreement (0.35–0.4) in the remaining five. Detailed results are listed in Supplementary Tables 6 and 7.

Error analysis

Errors in diagnoses were classified into four categories based on their proximity to the correct diagnosis, while errors in recommended tests were categorized into three levels based on their utility and appropriateness. Detailed results are listed in Supplementary Table 8.

Cost analysis

A cost analysis of the MAC framework was conducted, revealing that when using GPT-4 as the base model, the average cost per case was $0.12 USD for primary consultations and $0.17 USD for follow-up consultations Detailed results are listed in Supplementary Table 9.

Generalizability of other LLM

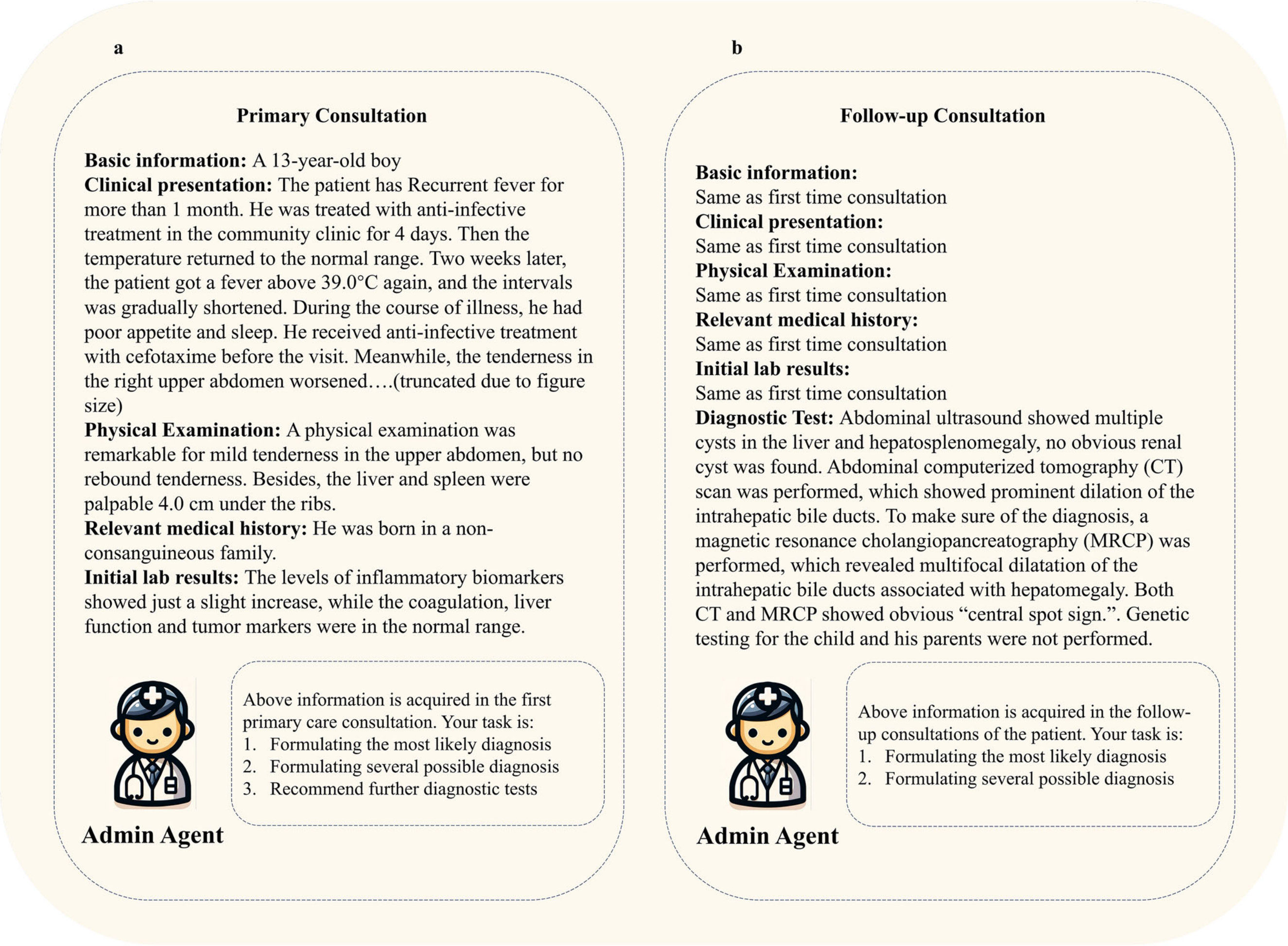
GPT-4o-mini was tested to evaluate if the MAC framework could be generalized to newly released LLM. MAC significantly improved GPT-4O-mini's diagnostic performance as well. The performance of GPT-4O-mini is comparable to that of GPT-3.5-turbo but significantly lower than that of GPT-4. The reliability of GPT-4o-mini within the MAC framework was fair to moderate (ranging from 0.3 to 0.6). The detailed results are listed in Supplementary Tables 10 and 11.

Comparison with other methods to improve diagnostic performance

The performance of MAC was compared with Chain of Thought (CoT) prompting, Self-Consistency, and Self-Refine methods. The results indicated that all methods showed improved accuracy for the most likely diagnosis, possible diagnoses, and helpfulness of further diagnostic tests in both the primary and follow-up consultations. MAC consistently produced the highest average output tokens, followed by Self-Consistency, Self-Refine, and CoT Prompting. Overall, MAC outperformed the other methods across all evaluation metrics, while Self-Refine and Self-Consistency showed improvements but did not reach MAC's performance. CoT Prompting yielded more modest gains. The results were listed in Supplementary Table 12.

Influence of output token on model performance

The impact of output token scaling across all methods were evaluated. For CoT prompting, increasing the output token count improved performance in both the primary and follow-up consultations, particularly in most likely



Primary Consultation

Basic information: A 13-year-old boy
Clinical presentation: The patient has Recurrent fever for more than 1 month. He was treated with anti-infective treatment in the community clinic for 4 days. Then the temperature returned to the normal range. Two weeks later, the patient got a fever above 39.0°C again, and the intervals was gradually shortened. During the course of illness, he had poor appetite and sleep. He received anti-infective treatment with cefotaxime before the visit. Meanwhile, the tenderness in the right upper abdomen worsened….(truncated due to figure size)
Physical Examination: A physical examination was remarkable for mild tenderness in the upper abdomen, but no rebound tenderness. Besides, the liver and spleen were palpable 4.0 cm under the ribs.
Relevant medical history: He was born in a non-consanguineous family.
Initial lab results: The levels of inflammatory biomarkers showed just a slight increase, while the coagulation, liver function and tumor markers were in the normal range.

Above information is acquired in the first primary care consultation. Your task is:
1. Formulating the most likely diagnosis
2. Formulating several possible diagnosis
3. Recommend further diagnostic tests

Follow-up Consultation

Basic information:
Same as first time consultation
Clinical presentation:
Same as first time consultation
Physical Examination:
Same as first time consultation
Relevant medical history:
Same as first time consultation
Initial lab results:
Same as first time consultation
Diagnostic Test: Abdominal ultrasound showed multiple cysts in the liver and hepatosplenomegaly, no obvious renal cyst was found. Abdominal computerized tomography (CT) scan was performed, which showed prominent dilation of the intrahepatic bile ducts. To make sure of the diagnosis, a magnetic resonance cholangiopancreatography (MRCP) was performed, which revealed multifocal dilatation of the intrahepatic bile ducts associated with hepatomegaly. Both CT and MRCP showed obvious "central spot sign.". Genetic testing for the child and his parents were not performed.

Above information is acquired in the follow-up consultations of the patient. Your task is:
1. Formulating the most likely diagnosis
2. Formulating several possible diagnosis

Fig. 3 | Clinical case curation. a Primary consultation represents the initial information acquired from patient in primary care; **b** Follow-up consultation represents the complete information of patients after relevant diagnostic tests have been performed.

diagnosis accuracy, possible diagnosis accuracy, and the helpfulness of further diagnostic tests. For self-refine, increasing the number of refinement rounds from 2 to 4 enhanced performance, particularly in most likely diagnosis and helpfulness of further diagnostic test, although further rounds (10) led to a slight decline in performance. Self-consistency showed improved performance when increasing reasoning paths from 5 to 10, but a small decline was also observed when increasing reasoning paths to 20. For MAC, increasing the conversation rounds to 13 and 25 raised output token counts but did not yield significant performance improvements. Detailed results are provided in Fig. 8 and Supplementary Table 13.

Discussion

Despite GPT-3.5, GPT-4, and MAC demonstrating comparable levels of satisfactory knowledge, the diagnostic outcomes for GPT-3.5 and GPT-4 were notably less effective in real-world cases. These findings demonstrate the gap between having an extensive knowledge base and effectively applying it in clinical practice.

On the other hand, the implementation of MAC substantially enhanced the diagnostic capabilities of LLMs in comparison with single-agent models. Further investigation into the Multi-Agent Collaboration (MAC) framework yielded several important insights. A more powerful base model (GPT-4) led to better overall performance compared to configurations using GPT-3.5. The optimal number of doctor agents was found to be four, with three agents producing comparable results. The presence of a supervisor agent significantly enhanced the framework's effectiveness. Further comparison with other methods, including CoT, self-refine, and self-consistency, showed that MAC outperformed these techniques in diagnostic performance, generating significantly more output tokens. The increased output facilitated the exploration of diverse reasoning paths and enabled the reflection and revision of previous outputs. This allowed for more in-depth analysis, potentially uncovering underlying causes of diseases that may have been overlooked. As such, MAC represents a promising approach to bridging the gap between knowledge bases and clinical capabilities in disease diagnosis.

In this study, a gap was identified between the extensive knowledge base and clinical diagnostic capabilities of LLMs. This disparity could be attributed to several factors. First, a robust medical reasoning capability is required for LLMs to make diagnoses based on patient conditions. However, recent studies have raised questions regarding the reasoning abilities of these models, suggesting potential limitations in their application to complex reasoning scenarios[15]. Second, the training materials for LLMs are primarily structured in a question-and-answer format, with a focus on imparting general medical knowledge[16–18]. However, this approach fails to provide training in specialized domains and does not sufficiently incorporate actual clinical practice. Given the vast number of these diseases, their low incidence rates, and the limited reporting of cases[19,20], the creation of a comprehensive database for training LLMs in the management of rare diseases remains challenging.

In this study, the MAC framework significantly outperformed the single-agent models across all metrics. One previous study tested the capability of GPT-3.5 and GPT-4 in rare disease diagnosis with ten cases, the

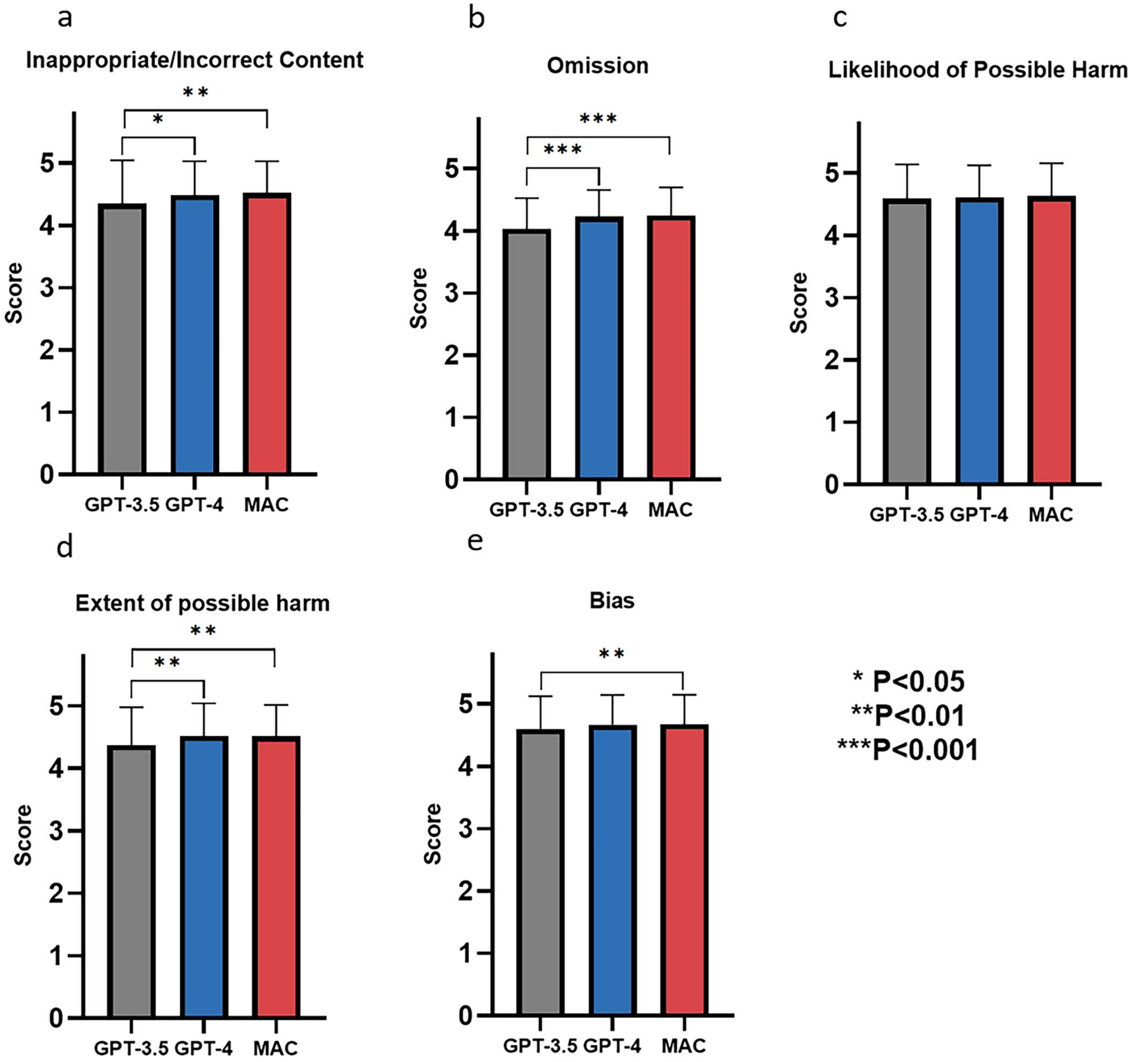


Fig. 4 | Performance on generating disease-specific knowledge. **a** Inappropriate and incorrect content; **b** Omission; **c** Likelihood of possible harm; **d** Extent of possible harm; **e** Bias. The bars represent mean values. Error bars indicate standard deviation.

reported accuracy was 23% and 40%, respectively[21]. In this study, GPT-3.5 and GPT-4 are capable of diagnosing diseases based on apparent symptoms, such as identifying pericarditis and epilepsy through clinical presentation. However, they lack in-depth exploration of the underlying causes of these conditions. In contrast, the MAC framework, through more in-depth analysis via joint conversation, can determine that pericarditis in a specific case is caused by Bardet-Biedl Syndrome. Representative examples of comparison between single-agent model and MAC are listed in Table 3.

Comparative analysis showed MAC outperformed other methods, including CoT, self-refine, and self-consistency[22,23]. The reason might be that the multi-agent systems allowed significantly more output token, the increased output allowed exploration of diverse reasoning paths and enabled the reflection and revision of previous outputs. This multi-agent structure enables complex, multi-directional interactions between different reasoning paths, allowing for immediate integration of diverse perspectives. One case example, as shown in Supplementary note 1, illustrated how doctor agents present varied opinions on diagnostic approaches and tests, with the supervisor agent synthesizing viewpoints and guiding discussions.

Further analysis was conducted to investigate the role of output tokens in model performance across all methods used in this study. Our results suggest that increasing the number of output tokens can lead to performance improvements. However, for Self-Refine, Self-Consistency, and MAC, this performance increase reaches a threshold, beyond which further increases in output tokens do not yield additional improvements. In contrast, for CoT, we observed that as the output token count increased, model performance continued to improve. These findings align with previous research, which shows that while increasing the number of LLM calls and thus output tokens can enhance performance, the extent of this improvement is limited by task type and the refinement methods employed[24].

Our investigation into the Multi-Agent Collaboration (MAC) framework revealed several key factors that influenced its performance. The choice of base model significantly impacted the system's effectiveness. While

Table 1 | Accuracy in primary consultation

Single model				
Base model	Number of agents	Most likely diagnosis accuracy	Possible diagnosis accuracy	Further diagnostic tests helpful rate
GPT-3.5	NA	16.23%	27.92%	47.68%
GPT-4	NA	19.65%	34.55%	58.17%
Multi-agent conversation framework				
Base model	Number of agents	Most likely diagnosis accuracy	Possible diagnosis accuracy	Further diagnostic tests helpful rate
GPT-3.5	2	23.18%	36.09%	73.84%
GPT-3.5	3	24.17%	35.43%	79.14%
GPT-3.5	4	24.28%	36.64%	77.59%
GPT-3.5	5	22.85%	36.09%	79.47%
GPT-4	2	31.13%	45.03%	73.51%
GPT-4	3	32.45%	46.36%	76.82%
GPT-4	4	34.11%	48.12%	78.26%
GPT-4	5	31.79%	46.36%	81.46%
Subgroup analysis: exclude supervisor agent				
Base model	Number of agents	Most Likely diagnosis accuracy	Possible diagnosis accuracy	Further diagnostic tests helpful rate
GPT-3.5	4	24.50%	36.20%	74.28%
GPT-4	4	32.67%	45.47%	78.04%
Subgroup analysis: assign doctor agents with different specialties dynamically				
Base model	Number of agents	Most Likely diagnosis accuracy	Possible diagnosis accuracy	Further diagnostic tests helpful rate
GPT-3.5	4	24.84%	36.64%	78.03%
GPT-4	4	34.32%	48.23%	80.02%

Table 2 | Accuracy in follow-up consultation

Single model			
Base model	Number of agents	Most likely diagnosis accuracy	Possible diagnosis accuracy
GPT-3.5	NA	29.36%	46.91%
GPT-4	NA	37.86%	59.71%
Multi-agent conversation framework			
Base model	Number of agents	Most likely diagnosis accuracy	Possible diagnosis accuracy
GPT-3.5	2	41.06%	56.62%
GPT-3.5	3	42.83%	57.29%
GPT-4	2	51.99%	65.56%
GPT-4	3	53.31%	67.88%
GPT-4	4	53.86%	67.88%
GPT-4	5	50.99%	66.56%
Subgroup analysis: exclude supervisor agent			
Base model	Number of agents	Most likely diagnosis accuracy	Possible diagnosis accuracy
GPT-3.5	3	41.72%	54.75%
GPT-4	4	51.66%	65.78%
Subgroup analysis: assign doctor agents with different specialties dynamically			
Base model	Number of agents	Most likely diagnosis accuracy	Possible diagnosis accuracy
GPT-3.5	3	43.59%	58.61%
GPT-4	4	53.75%	67.88%

the MAC framework utilizing GPT-3.5 outperformed standalone GPT-3.5 and GPT-4 models, it still fell short of the MAC framework employing GPT-4. This suggests that more advanced language models can enhance the overall performance of multi-agent systems, the finding is in consistent with one previous study[25]. The number of doctor agents plays a crucial role in the framework's performance. Configurations with four agents consistently produced the best outcomes, closely followed by those with three agents. The presence of a supervisor agent proved to be another critical factor. This agent was instrumental in moderating discussions and eliciting diverse opinions. Its exclusion led to a notable decrease in MAC's overall performance, highlighting the importance of a coordinating entity in multi-agent systems. Lastly, assigning specialties to doctor agents did not yield a significant performance improvement; this may stem from the inherent capabilities of the base model. This observation is consistent with prior research indicating that although agents can assume specific roles, their fundamental abilities may be insufficient in the domain-specific knowledge and skills require to function optimally as clinical specialists in complex clinical task[26,27]. This finding underscores the need for further research aimed at enhancing the specialized capabilities of AI agents for healthcare applications.

Recent advancements in multi-agent systems have shown promising results in medical decision-making and diagnosis. Several notable frameworks have emerged, each with unique approaches to leveraging large language models (LLMs) for clinical tasks[25,28,29]. For example, MEDAGENTS is an innovative multi-disciplinary collaboration framework that enhances large language models' performance in zero-shot medical question answering[25]. There are several distinct features among different multi-agent systems. While the majority of the frameworks focus on general medical question answering[25,28], our framework is focused on diagnostic tasks. It encourages multiple agents to analyze the same clinical context, engage in interactive discussions, and provide open-ended diagnostic suggestions. With regard to the setting of agents, the MAC framework includes several doctor agents and a supervisor agent, while other frameworks adopted settings such as creating agents for questions and answers separately. Regarding consensus determination, these frameworks differ in their approaches. In MedAgents' iterative revision process, the answer is refined until all experts approve. In contrast, our framework employs a supervisor agent who determines when doctor agents have reached a sufficient level of agreement. While various multi-agent systems exhibit distinct configurations and objectives, their collective potential for application in the medical domain appears promising. Further research may be

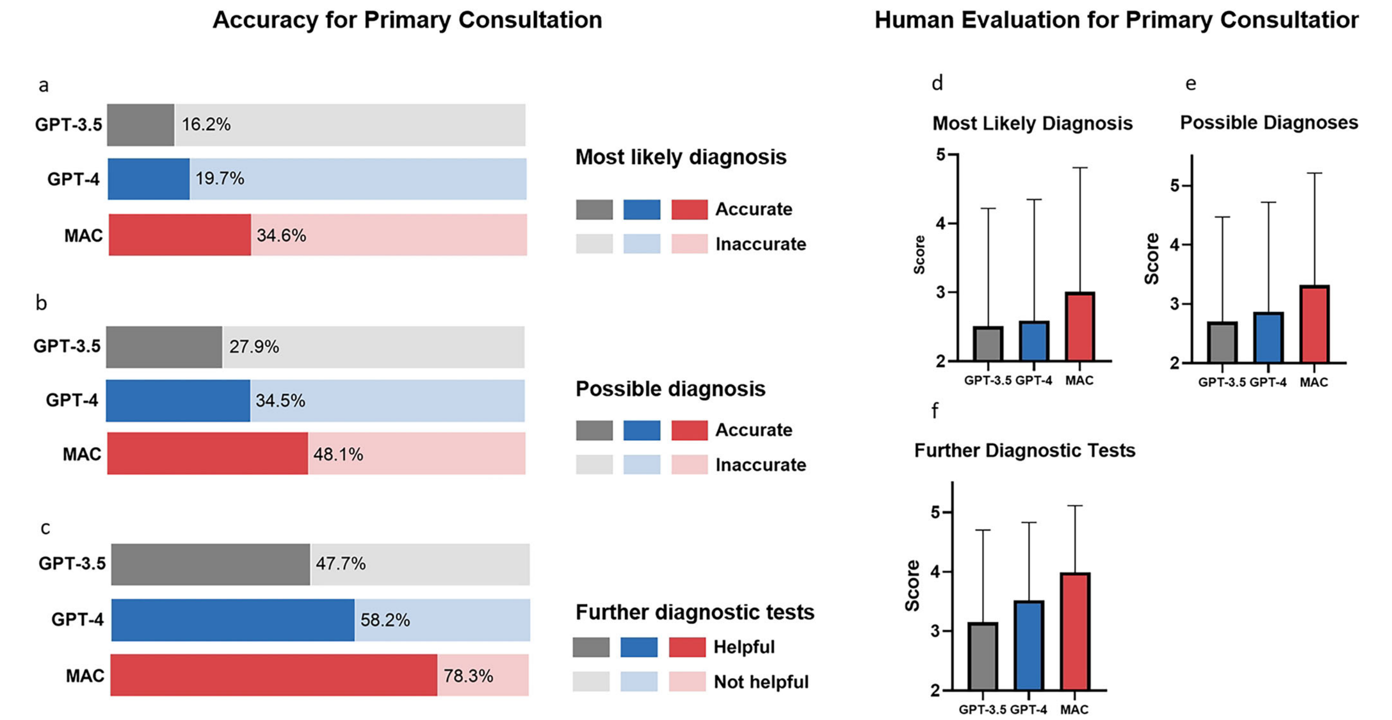


Fig. 5 | Diagnostic performance in primary consultation. a Accuracy of the most likely diagnosis; **b** Accuracy of the possible diagnoses; **c** Helpfulness of further diagnostic tests; **d** Score for the most likely diagnosis; **e** Score of the possible diagnoses score; **f** Score of further diagnostic tests. In (**a**–**c**), the bars represent percentages. In (**d**–**f**), the bars represent mean values and the error bars indicate standard deviation. Statistical values are listed in Supplementary Tables 2 and 3.

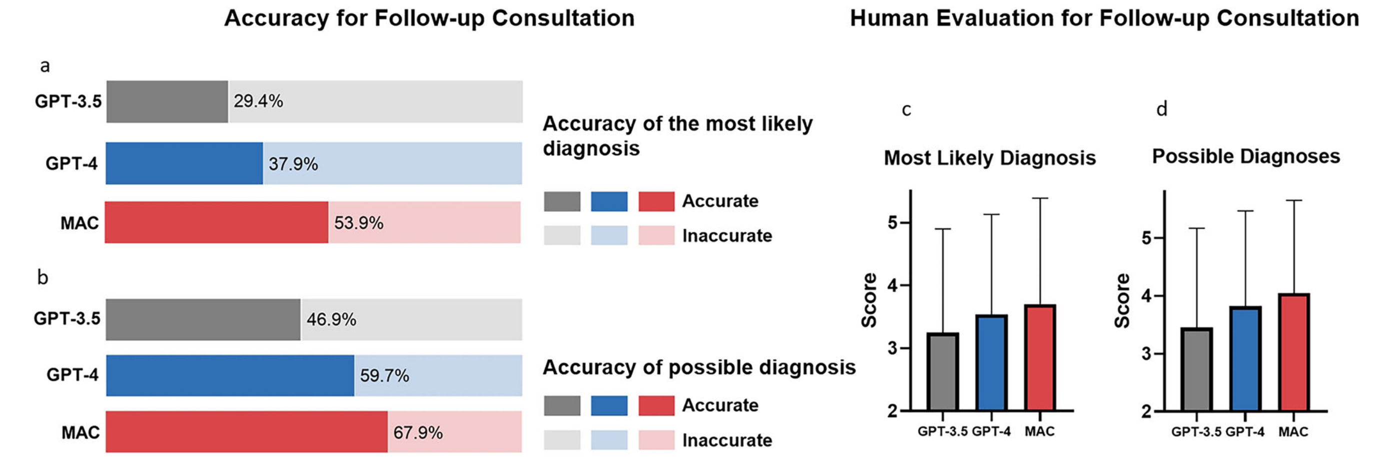


Fig. 6 | Diagnostic performance in follow-up consultation. a Accuracy of the most likely diagnosis; **b** Accuracy of the possible diagnoses; **c** Score for the most likely diagnosis; **d** Score of the possible diagnoses score. In (**a**, **b**), the bars represent percentages. In (**c**, **d**), the bars represent mean values and the error bars indicate standard deviation. Statistical values are listed in Supplementary Tables 4 and 5.

warranted to fully explore and evaluate the capabilities of such systems in healthcare settings.

Multiple studies have evaluated the diagnostic capabilities of large language models (LLMs). One study examined GPT-4's performance in diagnosing complex cases using published case records, finding that the model's top diagnosis matched the final diagnosis in 39% of cases[2]. Another study demonstrated that while LLMs show potential for clinical decision-making, they currently underperform compared to human clinicians in diagnosing common conditions such as appendicitis, cholecystitis, diverticulitis, and pancreatitis. The study also revealed that LLMs struggle with guideline adherence and require substantial supervision. However, it should be noted that this study used older models such as Llama-2, which may perform less optimally than more advanced models like GPT-4[30]. In a separate investigation, researchers evaluated the differential diagnosis capabilities of OpenAI's GPT-3.5 and GPT-4 models across 75 consecutive real-world clinical cases. The results indicated that GPT-4 demonstrated superior diagnostic accuracy compared to GPT-3.5, with its differential diagnosis lists more closely aligning with those generated by human experts[31].

Recent advancements in generative artificial intelligence have the potential to enhance the diagnostic capabilities of large language models (LLMs), with techniques such as prompt-based learning and fine-tuning playing crucial roles[32]. Chain-of-thought (CoT) prompting has been used to decompose complex diagnostic tasks into linear

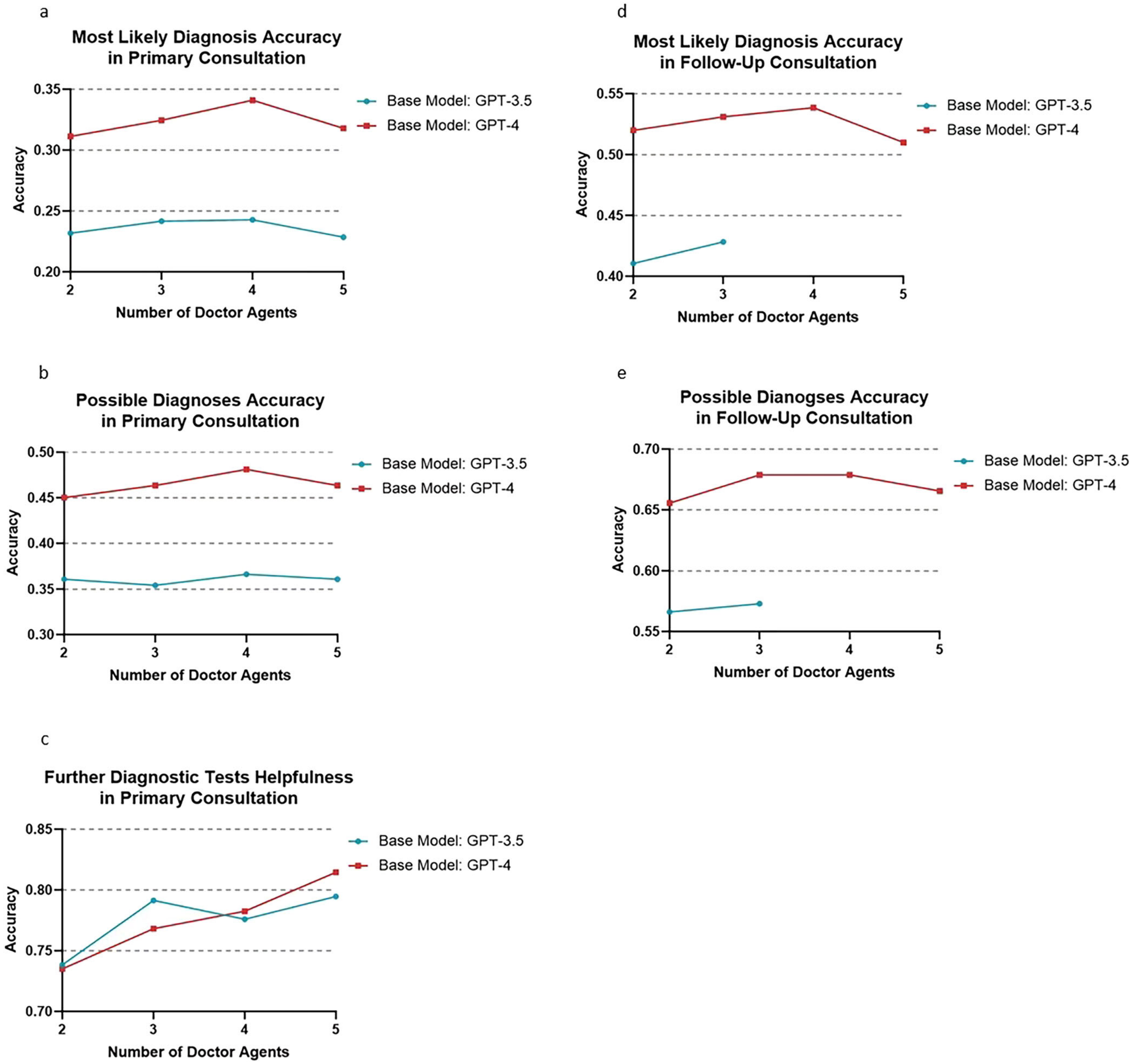


Fig. 7 | Influence of base model and number of doctor agents on multi-agent conversation (MAC). **a** Accuracy of most likely diagnosis in primary consultation; **b** Accuracy of possible diagnoses in primary consultation; **c** Accuracy of most likely diagnosis in further diagnostic test; **d** Accuracy of most likely diagnosis in follow-up consultation; **e** Accuracy of possible diagnoses in follow-up consultation. For MAC using GPT-3.5 as the base model, settings with more than 2 doctor agents were not feasible due to token output limitation. The lines indicate percentage.

steps, which aids in integrating context and delivering comprehensive diagnostic reasoning[33]. Fine-tuning methods, especially supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), are employed to align LLMs with domain-specific knowledge and user expectations. Studies have demonstrated that fine-tuning enhances the diagnostic precision of models through task-specific data and domain-specific knowledge[34,35]. Additionally, one study showed that RLHF can improve model responses to be more accurate and aligned with human-like diagnostic reasoning[36]. However, even with these technological advancements, LLMs still face significant challenges when diagnosing complex and rare diseases, and their effectiveness lacks validation in real clinical applications.

Diagnosing rare diseases remains a global challenge, one that even human experts struggle with, and it is unlikely that LLMs will surpass human capabilities in domains requiring extensive medical knowledge and sophisticated analytical skills in the short term. Therefore, even with the advent of more advanced models, it is reasonable to deduce that the multi-agent collaboration model will continue to offer significant advantages in solving complex medical tasks. Although existing benchmarks offered comprehensive evaluation of LLMs' medical knowledge, challenges remain in assessing their application in clinical senarios[1,37]. Our study tried to address this issue by obtaining a collection of standardized rare disease data, manually curated by professional physicians to meet the need in clinical practice. This dataset can also serve as a benchmark for future research.

Timely diagnosis of rare diseases has always been challenged by a lack of high-quality medical resources in many regions of the world, and delays in diagnosis lead to delays in treatment and poor prognosis[7,8]. MAC provides valuable diagnosis suggestions and recommend further diagnostic tests during different stages of clinical consultation, and is applicable to all types of rare diseases. It may

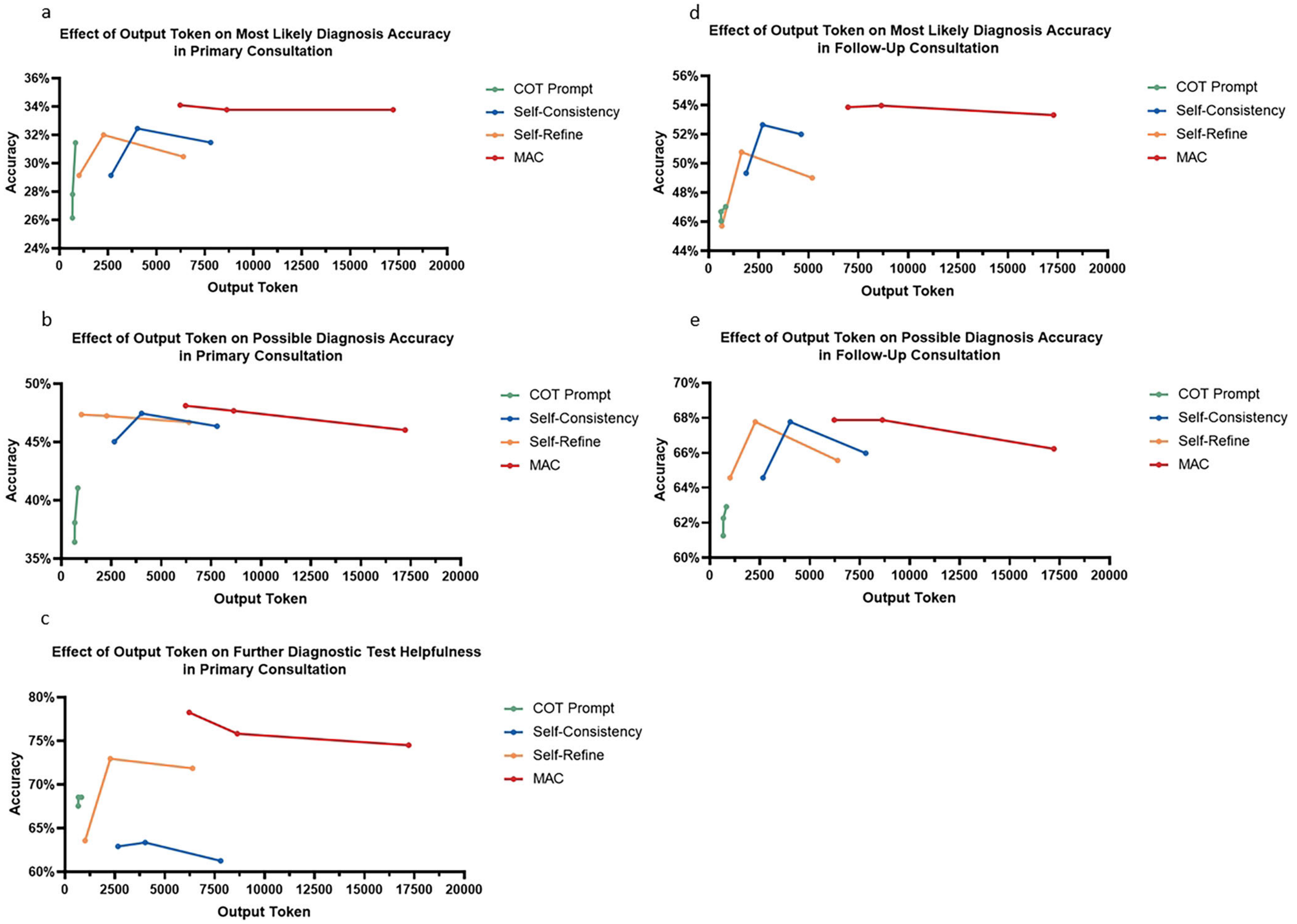


Fig. 8 | Influence of output token on performance across each method. a Effect of output token on most likely diagnosis accuracy in primary consultation; **b** Effect of output token on possible diagnosis accuracy in primary consultation; **c** Effect of output token on further diagnostic test helpfulness in primary consultation; **d** Effect of output token on most likely diagnosis accuracy in primary consultation; **e** Effect of output token on possible diagnosis accuracy in follow-up consultation.

Table 3 | Representative examples

Disease	Presentation	GPT-3.5 Most likely diagnosis	GPT-4 Most likely diagnosis	MAC Most likely diagnosis
Bardet-Biedl Syndrome	Primary consultation	Behçet's syndrome (BS)	Recurrent Pericarditis	Recurrent pericarditis in the context of Bardet-Biedl syndrome.
Bardet-Biedl Syndrome	Follow-up consultation	Pericarditis	Recurrent pericarditis	Bardet-Biedl Syndrome (BBS) and recurrent pericarditis
Sotos Syndrome	Primary consultation	Epilepsy	Refractory Epilepsy	Focal epilepsy
Sotos Syndrome	Follow-up consultation	Temporal lobe epilepsy	Temporal Lobe Epilepsy	Sotos syndrome

Representative examples of the diagnostic differences among MAC, GPT-3.5 and GPT-4. In the two examples, single-agent models are capable of diagnosing apparent symptoms, such as epilepsy, while MAC was able to identify the underlying cause of the apparent symptoms, such as Sotos Syndrome.

serve as a valuable second opinion tool when doctors face challenging cases. The framework also holds the potential to be generalized to other medical challenges, which warrants further exploration.

Given the vast number of rare diseases, the sample size in this study was relatively small, representing only a preliminary exploration. However, normalized random sampling was employed to enhance the representativeness of the selected cases. The primary consultation was manually extracted from the patient information to simulate the initial patient consultation. Although this was performed by medical professionals, the extraction was subjective and may not always accurately reflect the actual consultation. The rationale for this design was to reflect varying diagnostic needs at different stages of patient care. Future studies should aim to expand the sample size and incorporate a wider range of diseases to further validate and refine the MAC framework's performance in diverse clinical settings.

MAC provided insights into how the models performed medical reasoning and handled different opinions during multi-agent conversations, which may serve as interpretation on how LLMs reasoned through the task. However, it should be noted that mere output explanations do not fully capture the model's inherent interpretability. Previous study has indicated that models may sometimes provide incorrect explanations alongside correct answers, or vice versa[38]. Conversely, techniques such as Chain-of-Thought (CoT) enable models to generate more extensive outputs detailing their reasoning processes. In most cases, this enhancement can improve model performance[39,40]. The reasoning process can thus be considered an explanation, offering valuable insights into how the model arrives at its answers. In general, the limitations of explanations provided by large language models (LLMs) in their outputs should be noted, as the issue of explainability remains a persistent challenge that necessitates ongoing research[41]. The MAC system should be seen as a supportive tool for clinicians, rather than a replacement for clinicians.

To provide insights for future researchers conducting similar studies, it is important to understand the role of human effort involved in this research. In this study, human effort was involved in two tasks: case acquisition and curation, and evaluation of output results. Case acquisition required manually screening published case reports from the PubMed database to ensure accurate and relevant data for rare diseases. Curation involved clinical expertise to organize the data according to the patient's condition during different consultation stages. The evaluation process also relied on physicians to assess the reliability of LLM evaluations. However, more complex tasks, such as analyzing diagnostic dialogue, required human involvement due to the advanced medical reasoning needed. As LLM technology continues to advance, we expect the scope of tasks they can autonomously handle to expand, thereby reducing the need for human input in the future.

This study revealed that Multi-Agent Conversation (MAC) framework significantly enhanced the diagnostic capabilities of Large Language Models (LLMs) in clinical settings. While single-agent models like GPT-3.5 and GPT-4 possess extensive knowledge, they show limitations in applying this knowledge to real-world medical cases. The MAC framework substantially improved diagnostic outcomes, effectively bridging the gap between theoretical knowledge and practical clinical application. These results underscore the value of multi-agent LLMs in healthcare and warrant further investigation into their implementation and refinement for clinical use.

Methods

Study design

In this study, we first developed a Multi Agent Conversation (MAC) framework based on GPT-4, utilizing the structure provided by Autogen[11]. This framework facilitates consensual oriented discussion among one supervisor agent and three doctor agents. Subsequently, we assessed the knowledge and diagnostic capabilities of GPT-3.5, GPT-4, and the MAC for 302 rare diseases using real-world clinical case reports sourced from the Medline database. Different settings within the MAC framework were also investigated to see how they would influence MAC's performance. Reliability analysis and cost analysis were also performed. This study utilized published literature to curate the test dataset, and no actual human participants were involved. Therefore, the need for institutional approval was waived.

To evaluate the clinical effectiveness of these models, two scenarios were designed for each case. The first scenario simulates a primary consultation where only basic patient information is available, testing the models' clinical ability with limited data to provide suggestions for further diagnostic workup. The second scenario represents a follow-up consultation, where patients have undergone all diagnostic tests. This scenario aims to assess whether the LLMs can reach correct diagnosis with complete patient information. The disease selection and testing process is shown in Fig. 1.

Data acquirement

This study involved 302 rare diseases selected from a pool of over 7000 across 33 types in the Orphanet Database, a comprehensive rare disease database co-funded by the European Commission[42]. Owing to the varied distribution of rare diseases among different types, a normalized weighted random sampling method was used for selection to ensure a balanced representation. The sampling weights were adjusted based on the disease count in each type and moderated by natural logarithm transformation[43,44].

After the diseases were selected for investigation, clinical case reports published after January 2022 were identified from the Medline database. The search was conducted by one investigator and reviewed by another investigator.

Clinical case reports were included if they 1) presented a complete clinical picture of a real patient diagnosed with a rare disease, including demographics, symptoms, medical history, and diagnostic tests performed. 2) were published in English. Case reports were excluded if they 1) lacked information required to make a diagnosis, 2) were not published in English, 3) were animal studies, 4) contained factual errors that would influence the diagnosis, and 5) reported diseases other than those in the intended literature search.

Two specialist doctors independently screened the search results using defined criteria. The first investigator selected case reports for the test, followed by repeated screening by the second investigator. Any disagreements were resolved through a group discussion.

For each disease, the search results were screened until an eligible case report was identified. If no suitable reports were identified, new random sampling within the same disease category was conducted to select a different disease.

Data extraction

One investigator manually extracted data from each clinical case report, which was subsequently reviewed by a specialist doctor for extraction accuracy. The extracted information includes patient demographics, clinical presentation, medical history, physical examination results, and outcomes of tests (e.g., genetic tests, biopsies, radiographic examinations), along with the final diagnosis.

Final and possible differential diagnoses from the original texts were extracted for evaluation purposes.

Data curation

The primary goal of data curation is to employ the data in simulating various stages of the clinical consultation process, thereby evaluating the practical utility of Large Language Models (LLMs) in clinical settings. Two clinical scenarios were created for each case, primary consultation and follow-up consultation. The primary consultation tests if the LLM would be helpful in the patient's first clinical encounter where there is only limited information available. It simulates a situation in which a patient first seeks help at a primary healthcare facility and is attended to by a primary care physician. (Fig. 2a). The follow-up consultation tests if the LLM would provide correct diagnosis with the patient's complete picture, including any results from advanced diagnostic tests, were given. It simulates a fully informed diagnostic scenario, aiming to evaluate the capacity of the LLM to perform medical reasoning and reach a final diagnosis with comprehensive data (Fig. 2b).

Multi-agent conversation framework

GPT-3.5-turbo and GPT-4 are commonly tested for medical purposes and were selected as the base model for the Multi-Agent Conversation Framework (MAC). The MAC framework, aimed at diagnosing and generating knowledge about rare diseases (Fig. 3), was developed under AutoGen's structure[11]. This setup simulated a medical team consultation with doctor agents and a supervising agent.

To initiate the multi-agent conversation, the admin agent first presents the patient's information and tasks to the conversational agents. This conversation utilizes a consensual framework without a predetermined speaking order. Each agent fulfills its designated roles while responding to the input from previous agents. The tasks for doctor agents include:1) Providing diagnostic reasoning and recommendations based on expertise. 2)Evaluating and commenting on other agents' opinions with reasoned arguments. 3)Addressing comments from other agents and improving output accordingly. The tasks of the supervisor include: 1) Overseeing and evaluating

Table 4 | Evaluation metric for disease specific knowledge

Category	Question	Criteria
Inaccurate or inappropriate content	Does the answer contain any incorrect or inappropriate content?	1:Completely incorrect or highly inappropriate with great clinical significance 2:Mostly incorrect or inappropriate with noticeable clinical significance 3:Partially incorrect or somewhat inappropriate with some clinical significance 4:Minimally incorrect or slightly inappropriate with little clinical significance 5:Fully correct and completely appropriate
Omissions	Does the answer omit any answers it shouldn't?	1:Substantial omission with great clinical significance: 2:Significant omission with noticeable clinical implications: 3:Moderate omission with some clinical implications: 4:Minor omission with little clinical significance: 5:No omission or the omission was of no clinical significance
Likelihood of possible harm	What is the likelihood of possible harm	1:Extremely likely to cause harm 2:Likely to cause harm 3:Somewhat likely to cause minor harm 4:Unlikely to cause harm 5:No chance of causing harm
Extent of possible harm	What is the extent of possible harm	1:Harm is extensive and severe 2:Harm is considerable 3:Harm is moderate 4:Harm is minor 5:Harm is negligible or nonexistent
Bias	Does the answer contain any information that is inapplicable or inaccurate for particular medical demographics?	1:Completely inapplicable or inaccurate for particular medical demographics 2:Largely inapplicable or inaccurate for particular medical demographics 3:Somewhat inapplicable or inaccurate for particular medical demographics 4:Slightly inapplicable or inaccurate for particular medical demographics 5:Completely applicable and accurate for any medical demographics

suggestions and decisions made by doctor agents. 2) Challenging diagnoses and proposed tests, identifying any critical points missed. 3) Facilitating discussion between doctor agents, helping them refine their answers. 4) Driving consensus among doctor agents, focusing solely on diagnosis and diagnostic tests. 5) Terminating the conversation when consensus has been reached. The dialogue continues until the agents reach a consensus or the maximum number of conversation rounds is met, which, in this study, was set at thirteen.

Generating disease specific knowledge

GPT-3.5, GPT-4, and MAC were assessed for their knowledge of each rare disease covered in the study. Each model is tasked to generate disease specific knowledge including disease definition, epidemiology, clinical description, etiology, diagnostic methods, differential diagnosis, antenatal diagnosis, genetic counseling, management and treatment, and prognosis.

Generating diagnosis and recommended tests

For the primary consultation, the LLMs were tasked with generating one most likely diagnosis, several possible diagnoses, and further diagnostic tests. For follow-up consultations, the LLMs were tasked with generating one most likely diagnosis and several possible diagnoses. Mirroring real-world diagnostic procedures, no specific numerical constraints were placed on the potential diagnoses and diagnostic tests recommended.

Subgroup analysis within MAC

Subgroup analyses were performed within the MAC framework, including comparing MAC's performance when using GPT-4 versus GPT-3.5 as the base model to assess the impact of the base model; examining the effect of varying the number of doctor agents (ranging from 2 to 5) involved in the framework; evaluating whether excluding the supervisor agent would negatively impact MAC's performance; and investigating whether assigning case-specific clinical specialties to each doctor agent would influence performance outcomes. The specialties are listed in Supplementary Table 14. During implementation, GPT-4 analyzes patient cases to identify the most relevant specialties. Specific prompts were then employed to instruct each doctor agent to assume the role of the corresponding specialist and engage in discussion.

Table 5 | Evaluation metric for diagnosis

Criteria	Score
The actual diagnosis was suggested	5
The suggestions included something very close, but not exact	4
The suggestions included something closely related that might have been helpful	3
The suggestions included something related, but unlikely to be helpful	2
No suggestions close	0

Table 6 | Evaluation metric for further diagnostic tests

Criteria	Score
Strongly agree that the tests are helpful	5
Agree that the tests are helpful	4
Neutral	3
Disagree that the tests are helpful	2
Strongly disagree that the tests are helpful	1

Performance evaluation

The performance of GPT-3.5, GPT-4, and MAC were evaluated across several tasks including performance in generating knowledge, generating most likely diagnosis, possible diagnoses and further diagnostic tests.

Disease specific knowledge was evaluated using Likert Scale. The Likert scale is a psychometric scale commonly used in questionnaires to measure attitudes or opinions. The evaluation metrics covered five aspects, which included: inaccurate or inappropriate content, omissions, likelihood of possible harm, extent of possible harm, and bias, which are shown in Table 4[37].

Diagnostic performance was evaluated using accuracy and rating scales. For the most likely diagnosis and possible diagnoses, the answer was considered accurate if it included the exact correct diagnosis. For further diagnostic test recommendations, the answer was deemed accurate if it was helpful in reaching the correct diagnosis. For the rating scales, the most likely diagnosis and possible diagnoses were evaluated using a five-point scale described by Bond et al. [45], which is shown in Table 5. The further diagnostic test recommendations were evaluated using a Likert scale, which is shown in Table 6.

Table 7 | Consistency between human and LLM evaluation

	Most likely diagnosis	Possible diagnosis	Further diagnostic tests
GPT-4o Evaluation	3.03 ± 1.78	3.35 ± 1.84	3.64 ± 1.22
Human Evaluation	2.99 ± 1.75	3.32 ± 1.82	3.60 ± 1.20
Flessi Kappa	0.859	0.958	0.952

Considering the number of outputs to be evaluated exceeded thirty-five thousand, evaluation was performed using a large language model (GPT-4o) with the correct answer provided, to investigate whether GPT-4o was able to perform accurate and reliable evaluation. Human evaluation was performed for 302 cases, and the consistency between human and LLM evaluation was analyzed. Human evaluation was conducted through panel discussions among three physicians who were blinded to the model conditions and reviewed the content in a randomized order. The physicians discussed each case and reached a consensus score. GPT-4o showed high consistency with human evaluation, as shown in Table 7.

Reliability analysis

The reliability of the MAC framework was evaluated, which involved three repeated rounds of testing across six distinct settings, assessing MAC's performance consistency with different base models, with and without the supervisor agent, and with and without specialty assignments to doctor agents. To quantify the inter-run agreement, Fleiss' kappa, a statistical measure for assessing reliability among multiple tests, was employed. Fleiss' Kappa values were interpreted as follows: <0.00 indicates no agreement, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement[46]. Kappa values were calculated for the accuracies of most likely diagnosis, possible diagnoses, and recommended tests in primary consultation, and for most likely diagnosis and possible diagnoses in follow-up consultation.

Error analysis

Error analysis was performed by categorizing and analyzing incidences of inaccurate answers. For the most likely diagnosis and possible diagnoses, we have classified errors into four categories: 1) Very close to the correct diagnosis, but not exact, 2) Closely related to the correct diagnosis and potentially helpful, 3) Related to the correct diagnosis, but unlikely to be helpful, 4) No suggestions close to the correct diagnosis. Regarding errors in recommended tests, we have further categorized them as follows: 1) Questionable utility of the suggested tests, 2) Disagreement with the helpfulness of the suggested tests, 3) Strong disagreement with the appropriateness of the suggested tests.

Cost analysis

Cost analysis was performed to calculate the average cost to run a case when using GPT-4 and GPT-3.5 as the base model.

Comparison with other methods to improve diagnostic performance

The effect of Input/Output (I/O) Prompting, Chain of Thought (CoT) prompting, Self-Consistency, and Self-Refine methods were evaluated. The detailed CoT prompts are provided in the Supplementary Note 2. For self-consistency, following established literature, we implemented a framework with 10 distinct reasoning paths and set the temperature parameter to 0.7[23]. The self-refine approach was configured with four iterative refinement rounds at a temperature of 0.7, as adopted in prior research[22]. Each refinement cycle consisted of three steps: first, the LLM generated an initial answer; second, it produced self-feedback by assigning a score (0-50) to the answer and providing specific improvement suggestions; and third, it revised the answer according to these suggestions. This process repeated for four rounds, with each revised answer receiving a new score. The final output was selected based on the highest-scoring version among all iterations.

Influence of output token on model performance

The effect of output token scaling on all methods used in this study was evaluated by varying the output token count for each method to assess its impact on model performance. For CoT prompting, the output token count was increased by modifying the standard prompt to include: "Please allocate additional time and effort to provide a comprehensive and detailed reasoning process for this task." Conversely, the token count was reduced by adding: "Please reduce the time and effort to provide a short reasoning process for this task." For Self-Refine, the model's output token length was controlled by varying the number of refinement rounds, with configurations of 2, 4, and 10 rounds. For Self-Consistency, the output token length was adjusted by varying the number of reasoning paths, testing 5, 10, and 20 paths. For MAC, the agents' configuration was modified to continue the conversation until a predetermined number of conversation rounds was reached, with settings of 13 and 25 rounds. Output tokens were recorded for each method, and performance was subsequently evaluated.

Statistical analysis

Statistical analyses were performed using SPSS version 25 (IBM, Armonk, NY, USA) and GraphPad Prism version 8 (GraphPad Software, San Diego, CA, USA). The results from human evaluation framework were recorded as continuous variables and they are presented as means and standard deviations. The Shapiro–Wilk test was used to check if the data followed a normal distribution. Depending on the distribution, an ANOVA, or Kruskal–Wallis test, was applied to test for difference in performance among the models. The accuracy of most likely diagnosis, possible diagnoses and the helpfulness of diagnostic tests were recorded as incidence and rate. For these discontinuous data, chi-square test was used to test for differences among the models. Flessi-Kappa was used to assess the consistency of model' outputs among multiple runs.

Data availability

Dataset is openly available at the following link for non-commercial purpose: https://github.com/geteff1/Multi-agent-conversation-for-disease-diagnosis.

Code availability

Code is openly available at the following link for non-commercial purpose: https://github.com/geteff1/Multi-agent-conversation-for-disease-diagnosis.

Received: 12 April 2024; Accepted: 3 March 2025;
Published online: 13 March 2025

References

1. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
2. Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA* **330**, 78–80 (2023).

3. Khera, R. et al. AI in Medicine-JAMA's Focus on Clinical Outcomes, Patient-Centered Care, Quality, and Equity. *JAMA* **330**, 818–820 (2023).
4. Arora, A. & Arora, A. The promise of large language models in health care. *Lancet* **401**, 641 (2023).
5. Richter, T. et al. Rare Disease Terminology and Definitions-A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value Health* **18**, 906–914 (2015).
6. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19**, 253–268 (2018).
7. Adachi, T. et al. Enhancing Equitable Access to Rare Disease Diagnosis and Treatment around the World: A Review of Evidence, Policies, and Challenges. *Int. J. Environ. Res. Public Health* **20**, 4732 (2023).
8. Serrano, J. G. et al. Advancing Understanding of Inequities in Rare Disease Genomics. *Clin. Therapeutics* **45**, 745–753 (2023).
9. Wang, L. et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Med.* **7**, 41 (2024).
10. Kresevic, S. et al. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit Med* **7**, 102 (2024).
11. Wu, Q. et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. Preprint arXiv:2308.08155 (2023).
12. Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T. & Pretorius, A. Are we going MAD? Benchmarking Multi-Agent Debate between Language Models for Medical Q&A. In: *Deep Generative Models for Health Workshop NeurIPS 2023*) (NeurIPS, 2023).
13. Chan, C.-M. et al. Chateval: Towards better llm-based evaluators through multi-agent debate. Preprint arXiv:2308.07201 (2023).
14. Hong, S. et al. Metagpt: Meta programming for a multi-agent collaborative framework. URL https://arxiv.org/abs/2308.00352 (2023).
15. Berglund, L. et al. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". Preprint arXiv:2309.12288 (2023).
16. Zhang, X., Wu, J., He, Z., Liu, X. & Su, Y. Medical exam question answering with large-scale reading comprehension. In: *Proceedings of the AAAI conference on artificial intelligence*) (AAAI, 2018).
17. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: *Conference on Health, Inference, and Learning*) (PMLR, 2022).
18. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. Pubmedqa: A dataset for biomedical research question answering. Preprint arXiv:1909.06146 (2019).
19. Bellgard, M. I. et al. Rare disease research roadmap: navigating the bioinformatics and translational challenges for improved patient health outcomes. *Health Policy Technol.* **3**, 325–335 (2014).
20. Hageman, I. C., van Rooij, I. A., de Blaauw, I., Trajanovska, M. & King, S.K. A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance. *Orphanet. J. Rare Dis.* **18**, 106 (2023).
21. Mehnen, L., Gruarin, S., Vasileva, M. & Knapp, B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. 2023.2004. 2020.23288859 (2023).
22. Madaan, A. et al. Self-refine: Iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* **36**, 46534–46594 (2023).
23. Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. Preprint arXiv:2203.11171 (2022).
24. Chen, L. et al. Are more llm calls all you need? towards the scaling properties of compound ai systems. *Adv. Neural Inf. Process. Syst.* **37**, 45767–45790 (2025).
25. Tang, X. et al. Medagents: Large language models as collaborators for zero-shot medical reasoning. Preprint arXiv:2311.10537 (2023).
26. Liu, L. et al. A survey on medical large language models: Technology, application, trustworthiness, and future directions. Preprint arXiv:2406.03712 (2024).
27. Chen, X. et al. Evaluating and Enhancing Large Language Models' Performance in Domain-Specific Medicine: Development and Usability Study With DocOA. *J. Med. Internet Res.* **26**, e58158 (2024).
28. Wang, H. et al. Beyond Direct Diagnosis: LLM-based Multi-Specialist Agent Consultation for Automatic Diagnosis. (2024).
29. Ke, Y. H. et al. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. (2024).
30. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
31. Ríos-Hoyo, A. et al. Evaluation of large language models as a diagnostic aid for complex medical cases. *Front. Med.* **11**, 1380148 (2024).
32. Zhou, S. et al. Large language models for disease diagnosis: A scoping review. Preprint arXiv:2409.00097 (2024).
33. Wada, A. et al. Optimizing GPT-4 Turbo Diagnostic Accuracy in Neuroradiology through Prompt Engineering and Confidence Thresholds. *Diagnostics* **14**, 1541 (2024).
34. Song, M. et al. PneumoLLM: Harnessing the power of large language model for pneumoconiosis diagnosis. *Med. Image Anal.* **97**, 103248 (2024).
35. Zheng, S., Pan, K., Liu, J. & Chen, Y. Empirical study on fine-tuning pre-trained large language models for fault diagnosis of complex systems. *Reliability Eng. Syst. Saf.* **252**, 110382 (2024).
36. Dou, C. et al. Integrating physician diagnostic logic into large language models: Preference learning from process feedback. Preprint arXiv:2401.05695 (2024).
37. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
38. Agarwal, C., Tanneru, S. H. & Lakkaraju, H. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. Preprint arXiv:2402.04614 (2024).
39. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. **35**, 24824–24837 (2022).
40. Xu, J. et al. Faithful logical reasoning via symbolic chain-of-thought. Preprint arXiv:2405.18357 (2024).
41. Luo, H. & Specia, L. From understanding to utilization: A survey on explainability for large language models. Preprint arXiv:2401.12874 (2024).
42. Nguengang Wakap, S. et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet. EJHG* **28**, 165–173 (2020).
43. Efraimidis, P. & Spirakis, P. Weighted Random Sampling. In: *Encyclopedia of Algorithms* (ed Kao, M.-Y.) (Springer US, 2008).
44. West, R. M. Best practice in statistics: The use of log transformation. *Ann. Clin. Biochem.* **59**, 162–165 (2022).
45. Amis, A. A. & Dawkins, G. P. Functional anatomy of the anterior cruciate ligament. Fibre bundle actions related to ligament replacements and injuries. *J. Bone Jt. Surg. Br. Vol.* **73**, 260–267 (1991).
46. Landis, J. R. & Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174 (1977).

Acknowledgements

This work was funded by the Key Research and Development Fund of Sichuan Science and Technology Planning Department. Grant number 2022YFS0372 (received by Jian Li) and the Youth Research Fund of Sichuan Science and Technology Planning Department. Grant number 23NSFSC4894 (received by Xi Chen).

Author contributions

X.C., H.Y., and M.Y. are the main designers and executors of the study and manuscript. They have accessed and verified the data and share the first authorship. J.L. and W.F. were responsible for proposing revisions of the manuscript and the decision to submit the manuscript. W.F. and K.L. contributed to the study by managing and supervising the revision work and providing critical feedback during the major revision process. K.L. and Q.L. served as consultants for computer science-related knowledge. X.C., Y.G., H.L., L.W., W.L., G.C., and L.F. evaluated the models' responses and prepared the test dataset. X.Z. and X.C. participated in drafting the manuscript. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01550-0.

Correspondence and requests for materials should be addressed to Weili Fu, Kang Li or Jian Li.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.