A Self-Evolving Framework for Multi-Agent Medical Consultation Based on Large Language Models

基于大型语言模型的多智能体医疗咨询自我进化框架

Kai Chen [1] , Ji Qi [2] , Jing Huo [1*] , Pinzhuo Tian [4] , Fanyu Meng [3] , Xi Yang [2] , Yang Gao [1 1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2] China Mobile (Suzhou) Software Technology Co., Ltd. Suzhou, China [3] China Mobile Research Institute, Beijing, China

陈凯 [1] , 齐吉 [2] , 霍晶 [1*] , 田品卓 [4] , 孟凡宇 [3] , 杨曦 [2] , 高扬 [1 1] 南京大学新型软件技术国家重点实验室，中国南京 [2] 中国移动 (苏州) 软件技术有限公司，中国苏州 [3] 中国移动研究院，北京，中国

[4] School of Computer Engineering and Science, Shanghai University, Shanghai, China

[4] 上海大学计算机工程与科学学院，中国上海

Abstract-We propose a multi-agent approach (SeM-Agents) based on large language models for medical consultations. This framework incorporates various doctor roles and auxiliary roles, with agents communicating through natural language. Using a residual structure, the system conducts multi-round medical consultations based on the patient's treatment background and symptoms. In the final summary and output stage of the consultation, it utilizes two experience databases-the Correct Consultation Experience Database and the Chain of Thought (CoT) Experience Database-which evolve with accumulated experience during consultations. This evolution drives the framework's self-improvement, significantly enhancing the rationality and accuracy of the consultations. To ensure that the conclusions are safe, reliable, and aligned with human values, the final decisions undergo a safety review before being provided to the patient. This framework achieved accuracy rates of 89.2% and 83.1% on the MedQA and PubMedQA datasets, respectively.

摘要——我们提出了一种基于大型语言模型的多智能体医疗咨询方法 (SeM-Agents)。该框架涵盖多种医生角色及辅助角色，智能体通过自然语言进行交流。系统采用残差结构，基于患者的治疗背景和症状开展多轮医疗咨询。在咨询的最终总结与输出阶段，利用两个经验数据库——正确咨询经验数据库和思维链 (Chain of Thought, CoT) 经验数据库——随着咨询经验的积累不断进化。该进化驱动框架自我提升，显著增强咨询的合理性和准确性。为确保结论安全、可靠且符合人类价值观，最终决策在提供给患者前经过安全审查。该框架在 MedQA 和 PubMedQA 数据集上分别达到了 89.2% 和 83.1% 的准确率。

Index Terms-Large Language Model, Multi-Agent, Self-Evoling, Medical Consultation

关键词——大型语言模型，多智能体，自我进化，医疗咨询

I. INTRODUCTION

一、引言

Large Language Models (LLMs), with their extensive parameters and broad training across multiple domain

knowledge bases, exhibit remarkable task generalization capabilities [1]- [5]. Their prowess in logical reasoning and complex problem-solving positions them as potential assets in medical diagnostics [6]-[8]. However, the acquisition of real medical consultation data is challenging due to privacy constraints and irregular data preservation, and even after medical knowledge fine-tuning, LLMs still suffer from hallucination issues [9], [10]. In the medical domain, any errors induced by hallucinations are unacceptable and can lead to severe medical mishaps. Autonomous agents driven by LLMs open new avenues for medical consultations. LLM-based multi-agent technologies not only enhance medical reasoning capabilities [11] but also more effectively elicit embedded medical knowledge, which cannot be solely guided by Chain of Thought (CoT) reasoning. Moreover, the complementary actions among agents in multi-round interactions significantly reduce the risk of hallucinations [12].

大型语言模型 (Large Language Models, LLMs) 凭借其庞大的参数规模和跨多个领域知识库的广泛训练，展现出卓越的任务泛化能力 [1]-[5]。其在逻辑推理和复杂问题解决方面的能力，使其成为医疗诊断领域的潜在利器 [6]-[8]。然而，由于隐私限制和数据保存不规范，真实医疗咨询数据难以获取，即使经过医疗知识微调，LLMs 仍存在幻觉问题 [9],[10]。在医疗领域，幻觉引发的错误不可接受，可能导致严重医疗事故。基于 LLMs 的自主智能体为医疗咨询开辟了新途径。LLM 驱动的多智能体技术不仅提升了医疗推理能力 [11]，还能更有效地挖掘内嵌的医疗知识，这些知识单靠思维链 (CoT) 推理难以引导。此外，多轮交互中智能体间的互补行为显著降低了幻觉风险 [12]。

On one hand, well-designed organizational structures in Multi-Agent Systems significantly reduce system errors and enhance interaction efficiency [13]-[16]. ChatDev [13] decomposes tasks into sub-tasks, each managed by an instructor agent and an assistant agent, addressing software development and scientific discussion issues through multi-round inquiry-based collaboration, thus mitigating hallucinations. MAC-NET [16] organizes agents using a directed acyclic graph and simplifies their interactive reasoning through topological sorting to extract solutions from dialogues. MetaGPT [17] assigns agents roles within a software company and encodes Standard Operating Procedures (SOPs), demonstrating how to integrate agents' expertise to efficiently complete tasks. However, this method is primarily designed for software development, and its sequential execution process appears inefficient and unintuitive in medical consultation. Medagents [18] adopts a method where each LLM-Agent assumes a different doctor role, ultimately deciding the final solution through consensus voting. While this method is intuitive and clear, a simple voting mechanism without a robust organizational strategy could lead to collective hallucinations [19]. Moreover, these methods utilize static structures, unable to evolve, relying solely on the Zero-Shot capabilities of medical large models, which inherently limits their potential.

一方面，多智能体系统中设计良好的组织结构显著减少系统错误并提升交互效率 [13]-[16]。ChatDev[13] 将任务拆分为子任务，由指导智能体和助手智能体分别管理，通过多轮基于询问的协作解决软件开发和科学讨论问题，从而缓解幻觉。MAC-NET[16] 采用有向无环图组织智能体，通过拓扑排序简化交互推理，从对话中提取解决方案。MetaGPT[17] 为智能体分配软件公司内角色并编码标准操作流程 (SOP)，展示如何整合智能体专业知识高效完成任务。然而，该方法主要针对软件开发设计，其顺序执行流程在医疗咨询中显得低效且不直观。Medagents[18] 采用每个 LLM 智能体扮演不同医生角色，最终通过共识投票决定方案。该方法直观清晰，但简单投票机制缺乏稳健组织策略，可能导致集体幻觉 [19]。此外，这些方法采用静态结构，无法进化，仅依赖医疗大型模型的零样本能力，限制了其潜力。

On the other hand, inspired by the ways human intelligence is acquired, the phenomenon of enhancing LLM-Agents' problem-solving capabilities by granting them memory experiences for reflection and application has been proven feasible [20]. ExpeL [21] accumulates experiences from past successes and applies this experiential knowl-

edge during reasoning. ECL [22] concentrates on gathering experience-driven shortcuts derived from previous actions, which equips agents to more adeptly manage novel tasks. IER [23] allows LLM agents to iteratively refine experiences during task execution. Selfevolve [24] utilizes LLMs both as knowledge providers and as self-reflective programmers; through such reflective processes, agents can self-evolve. Agent Hospital [25] uses a Medical Record Library and an Experience Base to continuously accumulate diagnostic data, enhancing prompts for medical agents, and facilitating the evolution of medical agents.However, these efforts do not abstract, summarize, and reflect on erroneous cases, thus making it difficult to leverage valuable experiences from mistakes.

另一方面，受人类智能获取方式的启发，通过赋予大型语言模型代理 (LLM-Agents) 记忆体验以进行反思和应用，从而增强其解决问题能力的现象已被证明是可行的 [20]。ExpeL[21] 积累过去成功的经验，并在推理过程中应用这些经验知识。ECL[22] 专注于收集基于以往行动的经验驱动捷径，使代理能够更熟练地处理新任务。IER[23] 允许 LLM 代理在任务执行过程中迭代地完善经验。Selfevolve[24] 将 LLM 既作为知识提供者，也作为自我反思的程序员；通过这种反思过程，代理能够自我进化。Agent Hospital[25] 利用病历库和经验库持续积累诊断数据，提升医疗代理的提示效果，促进医疗代理的演进。然而，这些工作未能对错误案例进行抽象、总结和反思，因此难以从错误中汲取宝贵经验。

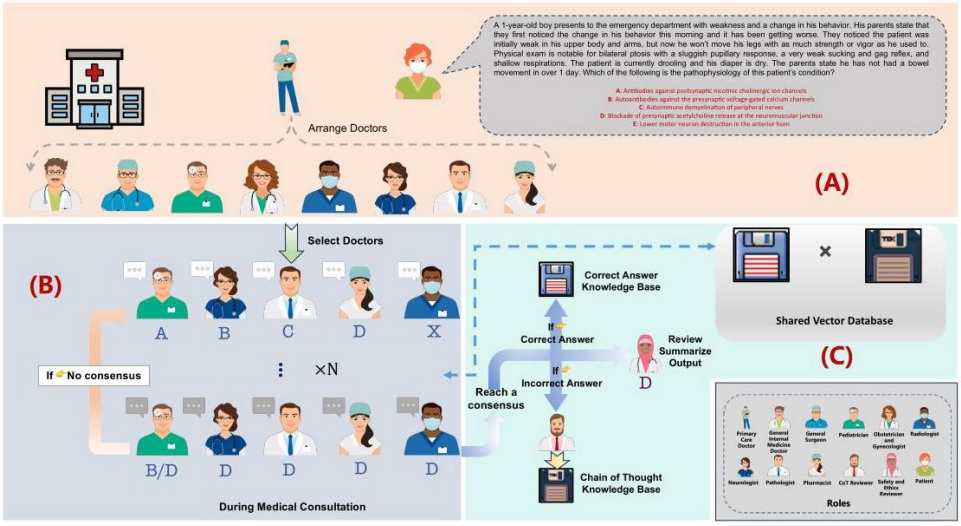

Fig. 1. Overview of the Medical Consultation Framework: (A) Arranging specialist doctors based on the specific situation of the patient; (B) Multi-round consultations of expert Agents; (C) Summary and output stage.

This work was supported in part by Nanjing University - China Mobile Communications Group Co., Ltd. Joint Institute, in part by the Science and Technology Innovation 2030 New Generation Artificial Intelligence Major Project under Grant 2021ZD0113303; in part by the National Natural Science Foundation of China under Grant 62192783, Grant 62276128; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20243051; in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

本工作部分得到了南京大学-中国移动通信集团有限公司联合研究院的支持，部分由科技创新 2030 新一代人工智能重大项目 (项目编号 2021ZD0113303) 资助；部分由国家自然科学基金 (项目编号 62192783，62276128) 资助；部分由江苏省自然科学基金 (项目编号 BK20243051) 资助；部分由新型软件技术与产业化协同创新中心支持。

图 1. 医疗咨询框架概览:(A) 根据患者具体情况安排专科医生；(B) 专家代理多轮会诊；(C) 总结与输出阶段。

In this work, we propose a self-evolving framework for multi-agent medical consultation (SeM-Agents) based on large language models. This framework incorporates various doctor roles and auxiliary roles. Patients present with their medical problems and background, and a Primary Care Doctor assigns the most appropriate specialist doctor agents based on the specific condition of the patient. The roles of Radiologist, Pathologist, and Pharmacist are essential and always included. Specialist doctor agents engage in multi-round discussions to share information, and once a consensus is reached, the consultation results are filtered through a Safety and Ethics Reviewer before the outcomes and recommendations are issued. Depending on their accuracy, the results are saved in different databases for reference in future consultations. Our contributions include: 1. We have introduced a dynamically expandable medical consultation framework that adapts to the patient's condition. 2. The framework evolves its consultation capabilities using two experience databases as the number of consultations increases. 3. The framework adopts an efficient residual discussion structure, enabling agents to efficiently access content from earlier shared speech pools, avoid information redundancy and contamination, and ensure the clarity and relevance of medical consultations.

在本工作中，我们提出了基于大型语言模型的多代理医疗咨询自我进化框架 (SeM-Agents)。该框架涵盖多种医生角色及辅助角色。患者提出其医疗问题及背景，初级保健医生根据患者具体情况分配最合适的专科医生代理。放射科医生、病理学家和药剂师为必备角色且始终包含其中。专科医生代理进行多轮讨论以共享信息，达成共识后，咨询结果经过安全与伦理审查员过滤，最终发布结果和建议。根据准确性，结果被保存至不同数据库，供未来咨询参考。我们的贡献包括:1. 引入了一个可动态扩展、适应患者状况的医疗咨询框架。2. 随着咨询次数增加，框架利用两个经验数据库进化其咨询能力。3. 采用高效的残差讨论结构，使代理能高效访问早期共享发言池内容，避免信息冗余和污染，确保医疗咨询的清晰性和相关性。

II. METHOD

二、方法

In this section, we will introduce the details of our medical consultation framework, where we have arranged multiple roles (Primary Care Doctor, General Internal Medicine Doctor, General Surgeon, Pediatrician, Obstetrician and Gynecologist, Radiologist, Neurologist, Pathologist, Pharmacist, Chain-of-Thought Reviewer, Safety and Ethics Reviewer, Patient). This arrangement enhances the framework's universality, enabling it to effectively address a wide range of complex medical scenarios. Fig. 1 shows the overview of the framework we propose, which is divided into three critical stages: (A) Assignment, (B) Consultation, and (C) Summary.

本节将介绍我们的医疗咨询框架细节，框架中安排了多种角色 (初级保健医生、普通内科医生、普通外科医生、儿科医生、妇产科医生、放射科医生、神经科医生、病理学家、药剂师、思路审查员、安全与伦理审查员、患者)。此安排增强了框架的通用性，使其能有效应对各种复杂医疗场景。图 1 展示了我们提出的框架概览，分为三个关键阶段:(A) 分配，(B) 咨询，(C) 总结。

A. Aranging specialist doctors

A. 安排专科医生

When the patient Agent comes for a consultation carrying personal background C and medical problem Q , the Primary Care Doctor Agent assigns specialist doctor Agents based on the specific circumstances of the patient. To ensure the triage doctor's output is more accurate and structured, we configure the Primary Care Doctor Agent with a few-shot example as a reference. The workflow is as follows: This description can be expressed by the following formula:

当患者代理携带个人背景 C 和医疗问题 Q 前来咨询时，初级保健医生代理根据患者具体情况分配专科医生代理。为确保分诊医生的输出更准确且结构化，我们为初级保健医生代理配置了少量示例作为参考。工作流程如下: 该描述可用以下公式表达:

$$\text{Roles } = \text{LLM}\left(\text{ Agents } \mid C, Q\right). \tag{1}$$

$$\text{Roles } \subseteq \left\{\text{ Agent }_1, \text{ Agent }_2, \ldots, \text{ Agent }_n\right\} . \tag{2}$$

Where roles, apart from Radiologist, Pathologist, and Pharmacist, are chosen based on the specific circumstances, avoiding information pollution by too many unrelated expert agents.

除放射科医生、病理学家和药剂师外，其他角色根据具体情况选择，避免过多无关专家代理导致信息污染。

B. Multi-round consultations

B. 多轮会诊

Once the specialist doctors for the consultation have been determined, the consultation process begins. In the initial round of consultation, each specialist doctor presents their views based on the patient's condition and provides an option ID and content for the issue at hand. At this stage, each specialist doctor agent cannot observe the others' remarks. All comments from this round are stored in a shared speech pool (S_1) . Once all the specialist doctors have concluded, the consultation moves to the next round of remarks.

一旦确定了咨询的专科医生，咨询过程即开始。在首轮会诊中，每位专科医生根据患者状况提出观点，并为当前问题提供选项 ID 及内容。此阶段，每位专科医生代理无法观察其他人的发言。该轮所有评论均存储于共享发言池 (S_1) 。所有专科医生发言结束后，咨询进入下一轮评论。

Starting from the second round, each specialist doctor agent can access remarks stored in the shared speech pool from the previous round. They integrate these insights to optimize prompts, formulating their own responses denoted as $S_{2,k}$ (where $S_{2,k}$ represents the response of the k -th specialist doctor in the second round), and specify an option ID along with content relevant to the current issue for further discussion and decision-making.

从第二轮开始，每位专科医生代理可以访问上一轮存储在共享发言池中的备注。他们整合这些见解以优化提示，制定自己的回答，表示为 $S_{2,k}$ (其中 $S_{2,k}$ 代表第二轮中第 k 位专科医生的回答)，并指定一个选项 ID 及与当前问题相关的内容，以便进一步讨论和决策。

From the $i+1$ round $(i \geq 2)$, specialist doctor agents can review the remarks from rounds i and $i-1$ in the shared speech pool. Incorporating the collective remarks from the previous two rounds to enhance the prompt, they articulate their own views and provide an option ID and content for the issue. The discussion continues until all the expert doctor agents reach a consensus on the answers. If consensus is not reached or the number of discussion rounds is below the maximum (set at 10), the discussion continues. If the maximum rounds are reached without achieving consensus, the decision is made by majority rule; if votes are evenly distributed, a final answer is randomly selected from the Agents' choices. This residual discussion mode reduces information pollution and enhances the efficiency of the discussion, while also reducing memory size. Furthermore, each expert doctor agent can access deeper layers of memory, which helps prevent any single expert doctor agent from being overly influenced by other agents, thereby mitigating the occurrence of hallucinations to some extent. The consultation process is defined according to Algorithm 1.

从第 $i+1$ 轮 $(i \geq 2)$ 开始，专科医生代理可以查看共享发言池中第 i 轮和第 $i-1$ 轮的备注。通过整合前两轮的集体意见来增强提示，他们表达自己的观点，并为该问题提供选项 ID 和内容。讨论持续进行，直到所有专家医生代理达成共识。如果未达成共识且讨论轮数未达到最大值 (设为 10 轮)，则继续讨论。若达到最大轮数仍未达成共识，则采用多数决；若票数均分，则从代理的选择中随机选出最终答案。这种残余讨论模式减少了信息污染，提高了讨论效率，同时也减小了内存占用。此外，每位专家医生代理可以访问更深层次的记忆，有助于防止单一专家医生代理过度受其他代理影响，从而在一定程度上减轻幻觉现象。咨询过程按照算法 1 定义。

Algorithm 1 Multi-Round Medical Consultation Process

算法 1 多轮医疗咨询流程

Initialize: speech pool $S_1 = \{s_{1,1}, s_{1,2}, \ldots, s_{1,n}\}$

初始化: 发言池 $S_1 = \{s_{1,1}, s_{1,2}, \ldots, s_{1,n}\}$

Compute for Round 2:

计算第 2 轮:

for each specialist k do

对每位专科医生 k 执行

$S_{2,k} \leftarrow f\left(S_1, C, Q\right)$
end for

结束

Subsequent Rounds:

后续轮次:

Set $i = 2$

设置 $i = 2$

while not Consensus (S_n) and $i \leq$ MaxRounds do

当未达成共识 (S_n) 且 $i \leq$ 未超过最大轮数时

for each specialist k do

对每位专科医生 k 执行

$S_{i+1,k} \leftarrow g\left(S_i, S_{i-1}, C, Q\right)$
end for

结束

Increment i

轮数递增 i

Consensus check:

共识检查:

if $\forall k, m \in \{1, \ldots, n\} : S_{n,k} = S_{n,m}$ then

如果 $\forall k, m \in \{1, \ldots, n\} : S_{n,k} = S_{n,m}$ 则

Set Consensus $(S_n) =$ True

设置共识 $(S_n) =$ 为真

end if

结束条件判断

end while

结束循环

C. Summary and output stage

C. 总结与输出阶段

During this phase, the final output(C)is subjected to a review by the Safety and Ethics Reviewer Agent, who filters and refines the consultation conclusions, identifying any unsafe aspects and finalizing the conclusions(R). These conclusions are then compared with the correct outcomes. If the consultation conclusions are accurate, the patient's background(B), the problem, and the discussions from the final consultation round are archived in the Correct Answer Knowledge Base (CorrectKB). Conversely, if the consultation results in incorrect conclusions, the session is abstracted by the Chain-of-Thought Reviewer. This abstraction includes the patient's background and problem, structured according to the initial hypotheses, analysis process, final conclusion, and reasons for error, and is then stored in the Chain of Thought Knowledge Base (ChainKB).

在此阶段，最终输出 (C) 由安全与伦理审查代理进行审核，该代理对咨询结论进行筛选和完善，识别任何不安全的方面并最终确定结论 (R)。随后将这些结论与正确结果进行比较。如果咨询结论准确，患者背景 (B)、问题及最终咨询轮次的讨论内容将被存档于正确答案知识库 (Correct Answer Knowledge Base，CorrectKB)。相反，若咨询结果得出错误结论，则该会话由思维链审查者进行抽象处理。该抽象包括患者背景和问题，按照初始假设、分析过程、最终结论及错误原因进行结构化，并存储于思维链知识库 (Chain of Thought Knowledge Base，ChainKB)。

When the next patient arrives, the background and problem of the patient are used to retrieve the most similar cases from the two databases via cosine similarity, thus enhancing the prompts(P)for the specialist doctor agents. To preserve the independent reasoning of each specialist doctor agent, references to the two knowledge bases are generally not made before the initial round of discussions. Instead, these references are utilized starting from the second round, particularly when divergent opinions arise. However, if a consensus emerges in the first round, the bases may be consulted post-discussion as a reflective measure. This process is given in Algorithm 2.

当下一位患者到来时，利用患者的背景和问题通过余弦相似度从两个数据库中检索最相似的病例，从而增强专家医生代理的提示 (P)。为了保持每位专家医生代理的独立推理，通常在初始讨论轮次之前不会参考这两个知识库。相反，这些参考从第二轮开始使用，特别是在出现分歧意见时。然而，如果第一轮达成共识，讨论后可作为反思措施查阅知识库。该过程见算法 2。

Algorithm 2 Summarization and Enhancement of Prompts

算法 2 提示的总结与增强

Input: C, B , CorrectKB, ChainKB

输入: C, B ，正确知识库，链式知识库

Review and Validate:

审查与验证:

$R \leftarrow \text{Review}\,(C)$

$$D \leftarrow \begin{cases} \text{Correct KB} & \text{if Valid}(R \\ \text{Chain KB} & \text{otherwise} \end{cases}$$

Knowledge Management:

知识管理:

if $D = \text{CorrectKB}$ then

如果 $D = \text{CorrectKB}$ 则

Store(R, CorrectKB)

存储 (R, CorrectKB)

else

否则

abstraction $\leftarrow$ Abstract (R)

抽象 $\leftarrow$ Abstract (R)

Store(abstraction, ChainKB)

存储 (抽象, ChainKB)

end if

结束条件判断

Consultation Enhancement:

咨询增强:

similarity $\leftarrow$ CosineSim(B, CorrectKB, ChainKB)

相似度 $\leftarrow$ CosineSim(B, CorrectKB, ChainKB)

$P \leftarrow$ Retrieve(similarity, CorrectKB, ChainKB)

$P \leftarrow$ 检索 (similarity, CorrectKB, ChainKB)

enhancedPrompt $\leftarrow$ Enhance(P)

增强提示 $\leftarrow$ Enhance(P)

Apply Enhanced Prompt:

应用增强提示:

round $\leftarrow 1$ $\triangleright$ Initialize round count

轮次 $\leftarrow 1$ $\triangleright$ 初始化轮次计数

while not Consensus(C)do

当未达成共识 (C) 时执行

UsePrompt(enhancedPrompt, round)

使用提示 (enhancedPrompt, round)

round $\leftarrow$ round +1

轮次 $\leftarrow$ 轮次 +1

if round $= 1$ and Consensus (C) then

如果轮次 $= 1$ 且 Consensus (C) 则

ConsultKBs (P) $\triangleright$ Reflective measure post-discussion

咨询知识库 (P) $\triangleright$ 讨论后的反思措施

end if

结束如果

end while

结束循环

III. EXPERIENCE

三、实验

A. Datasets

A. 数据集

We use the MedQA [26] and PubMedQA [27] datasets to validate our framework. The MedQA dataset consists of USMLE-style questions, each offering four or five possible answers, designed to assess medical knowledge and

practical skills. PubMedQA, based on research paper abstracts, presents questions with Yes/No/Maybe answers, aiming to evaluate the performance of natural language processing models in academic question answering. The final results are all measured on the test set of each dataset. The Correct Answer Knowledge Base and the Chain of Thought Knowledge Base only include experiences from the training set of each dataset.

我们使用 MedQA [26] 和 PubMedQA [27] 数据集来验证我们的框架。MedQA 数据集包含类似 USMLE(美国医学执照考试) 风格的问题，每个问题提供四到五个可能的答案，旨在评估医学知识和实践技能。PubMedQA 基于研究论文摘要，提出带有是/否/可能答案的问题，旨在评估自然语言处理模型在学术问答中的表现。最终结果均在各数据集的测试集上测量。正确答案知识库 (Correct Answer Knowledge Base) 和思维链知识库 (Chain of Thought Knowledge Base) 仅包含各数据集训练集中的经验。

B. Main Results

B. 主要结果

In this experimental section, we primarily explore the Zero-shot accuracy advantages of our proposed framework, SeM-Agents, in the medical consultation domain, and validate the contributions of each component within our approach. For this subset of experiments, all agents in our framework utilize gpt-4-turbo, with SeM-Agents undergoing 600 consultation rounds-a benchmark chosen after considering both performance and cost. Overall performance can be referred to in Table I, where the foundational model used across all configurations is gpt-4-turbo. 'Single-Agent' denotes performance using only gpt-4-turb as our baseline. 'Single-Agent (w/) CoT' incorporates a 'Let's think step by step' approach in the answering process. '1 Round Multi-Agent' employs a single-round voting mechanism following the majority rule principle. '10 Rounds Multi-Agent-sequential speaking' describes a scenario where specialist doctor agents speak in a sequential order after listening to the previous agent's opinion, a method that generally underperforms compared to the voting model in medical fields. Although our method shows lower accuracy on the MedQA dataset compared to Medprompt [8]-likely due to Medprompt only being tested in four-option scenarios-it achieves higher accuracy on the PubQA dataset and on average than Medprompt. Table I illustrates that the discussion modes and experiential growth proposed positively impact overall performance, with an interesting observation that correct experiences contribute more significantly to accuracy improvements than abstracted CoT experiences, which is intuitively consistent.

在本实验部分, 我们主要探讨了所提框架 SeM-Agents 在医疗咨询领域的零样本准确率优势, 并验证了方法中各组件的贡献。对于这部分实验, 框架中的所有代理均采用 gpt-4-turbo, SeM-Agents 进行了 600 轮咨询——该基准是在性能与成本权衡后选定的。整体性能见表 I，所有配置均基于 gpt-4-turbo 基础模型。“单代理”表示仅使用 gpt-4-turbo 作为基线的表现。“单代理 (含 CoT)”在回答过程中引入了“让我们一步步思考”的链式思维 (Chain-of-Thought, CoT) 方法。“一轮多代理”采用单轮投票机制，遵循多数规则原则。“十轮多代理-顺序发言”描述专家医生代理在听取前一代理意见后依次发言的场景，该方法在医疗领域通常表现不及投票模型。尽管我们的方法在 MedQA 数据集上的准确率低于 Medprompt [8]——这可能是因为 Medprompt 仅在四选一场景下测试——但在 PubQA 数据集及平均表现上优于 Medprompt。表 I 显示，所提讨论模式和经验增长对整体性能有积极影响，有趣的是，正确经验对准确率提升的贡献明显大于抽象的链式思维经验，这与直觉相符。

TABLE I
MAIN RESULTS ON ACCURACY ACROSS MEDQA AND PUBMEDQA DATASETS

MEDQA 与 PUBMEDQA 数据集上的主要准确率结果

Method	**MedQA** (%)	**PubMedQA** (%)	Average(%)
Single-Agent	77.4	75.3	76.4
Single-Agent (w/) CoT	76.6	76.9	76.8
Medprompt [8]	90.2	82.0	86.1
1 Round Multi-Agent	78.2	73.7	76.0
10 Rounds Multi-Agent	78.5	74.0	76.3
10 Rounds Multi-Agent sequential speaking	77.8	72.9	75.4
MedAgents [18]	83.7	76.8	80.3
SeM-Agents (w/o) residual discussion mode (ours)	88.6	79.3	83.95
SeM-Agents (w/o) bases	83.3	77.1	80.2
SeM-Agents (w/o) correct answer knowledge base (ours)	86.2	80.7	83.5
SeM-Agents (w/o) CoT knowledge base (ours)	89.0	82.5	85.8
SeM-Agents (ours)	89.2	83.1	86.2

方法	**MedQA** (%)	**PubMedQA** (%)	平均 (%)
单一智能体	77.4	75.3	76.4
单一智能体 (带有) 链式思维 (CoT)	76.6	76.9	76.8
Medprompt [8]	90.2	82.0	86.1
一轮多智能体	78.2	73.7	76.0
十轮多智能体	78.5	74.0	76.3
十轮多智能体顺序发言	77.8	72.9	75.4
MedAgents [18]	83.7	76.8	80.3
SeM-Agents(无) 残差讨论模式 (本方法)	88.6	79.3	83.95
SeM-Agents(无) 基础	83.3	77.1	80.2
SeM-Agents(无) 正确答案知识库 (本方法)	86.2	80.7	83.5
SeM-Agents(无) 链式思维知识库 (本方法)	89.0	82.5	85.8
SeM-Agents(本方法)	89.2	83.1	86.2

C. Self-Evolving Experiment

C. 自我进化实验

The self-evolving, growth-capable framework of multi-agent doctors, which continuously improves through consultation experiences, often aligns more closely with practical requirements than static frameworks—an intuitively appealing notion. Herein, we demonstrate how the number of consultation cases and the volume of stored case experiences influence accuracy variations on test sets across two datasets, MedQA and PubMedQA. Each dataset contributes half of the cases in two experience databases: the Correct Answer Knowledge Base and the Chain of Thought Knowledge Base. We conduct tests using foundation models gpt-3.5-turbo and gpt-4-turbo. As illustrated in Fig. 2, the overall trend shows a gradual increase in accuracy as the number of consultation samples grows (with a slight decline observed around 100 cases), and tends to plateau after reaching 600 cases.

具备自我进化和成长能力的多智能体医生框架，通过不断的咨询经验持续改进，通常比静态框架更贴合实际需求——这一观点直观且具有吸引力。本文展示了咨询案例数量和存储案例经验量如何影响两个数据集 MedQA 和 PubMedQA 测试集上的准确率变化。每个数据集各贡献一半案例，存入两个经验库: 正确答案知识库 (Correct Answer Knowledge Base) 和思维链知识库 (Chain of Thought Knowledge Base)。我们使用基础模型 gpt-3.5-turbo 和 gpt-4-turbo 进行测试。如图 2 所示，整体趋势为随着咨询样本数量增加，准确率逐步提升 (在约 100 个案例时略有下降)，并在达到 600 个案例后趋于平稳。

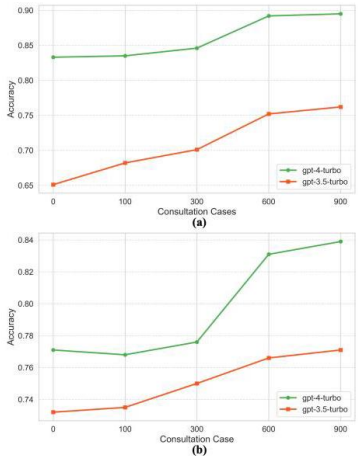

Fig. 2. Accuracy Variations: (a) Tested on MedQA (b) Tested on PubMedQA

图 2. 准确率变化:(a) MedQA 测试 (b) PubMedQA 测试

D. Impact of Different Foundation Models

D. 不同基础模型的影响

We demonstrate the accuracy of different foundational models on two datasets, each enhanced with knowledge bases from 600 consultation cases. Table II shows that gpt-4-turbo remains the optimal foundational model for the SeM-Agents framework. Meanwhile, other foundational models also demonstrate excellent performance within the SeM-Agents framework, suggesting its adaptability across different foundational models.

我们展示了不同基础模型在两个数据集上的准确率表现, 均基于 600 个咨询案例构建的知识库增强。表 II 显示, gpt-4-turbo 依然是 SeM-Agents 框架的最佳基础模型。同时, 其他基础模型在 SeM-Agents 框架中也表现出色，表明该框架对不同基础模型具有良好的适应性。

TABLE II

ACCURACY COMPARISONS ACROSS MEDQA AND PUBMEDQA USING DIFFERENT FOUNDATION MODELS

不同基础模型在 MedQA 和 PubMedQA 上的准确率比较

Backbone	MedQA(%)	PubMedQA(%)	Average
LLaMA3-8B [28]	70.1	64.9	67.5
GLM4 [29]	74.3	75.1	74.7
DeepSeek-v2 [30]	75.7	74.5	75.1
gpt-3.5-turbo	75.2	76.6	75.9
gpt-4-turbo	89.2	83.1	86.2

骨干网络	医学问答 (%)	PubMed 问答 (%)	平均值
LLaMA3-8B [28]	70.1	64.9	67.5
GLM4 [29]	74.3	75.1	74.7
DeepSeek-v2 [30]	75.7	74.5	75.1
gpt-3.5-turbo	75.2	76.6	75.9
gpt-4-turbo	89.2	83.1	86.2

IV. CONCLUSION

四、结论

In this paper, we introduce a novel multi-agent framework for medical consultation that employs a residual discussion mode to reduce information pollution and enhance discussion efficiency. By leveraging two experience databases, this framework dynamically improves overall consultation accuracy. However, the overall performance of the framework largely depends on the capabilities of the foundational model used to store and utilize consultation experiences, which may limit its performance. Despite these limitations, our approach still excels in current medical consultation scenarios. REFERENCES

本文提出了一种新颖的多智能体医疗咨询框架，采用残差讨论模式以减少信息污染并提升讨论效率。通过利用两个经验数据库，该框架动态提升整体咨询准确率。然而，框架的整体性能在很大程度上依赖于用于存储和利用咨询经验的基础模型的能力，这可能限制其表现。尽管存在这些限制，我们的方法在当前医疗咨询场景中仍表现出色。参考文献

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., ”Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat 等，“Gpt-4 技术报告”，arXiv 预印本 arXiv:2303.08774，2023 年。

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar et al., ”Llama: Open and efficient foundation language models,” arXiv preprint arXiv:2302.13971, 2023.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar 等，“Llama: 开放且高效的基础语言模型”，arXiv 预印本 arXiv:2302.13971，2023 年。

[3] J. Wang, L. Chen, A. Khare, A. Raju, P. Dheram, D. He, M. Wu, A. Stol-cke, and V. Ravichandran, ”Turn-taking and backchannel prediction with acoustic and large language model fusion,” in ICASSP 2024 - 2024 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 12121-12125.

J. Wang, L. Chen, A. Khare, A. Raju, P. Dheram, D. He, M. Wu, A. Stolcke, 和 V. Ravichandran，“基于声学与大型语言模型融合的轮次控制与回声预测”，发表于 ICASSP 2024 - 2024 年 IEEE 国际声学、语音与信号处理会议，2024 年，第 12121-12125 页。

[4] I. Malkiel, U. Alon, Y. Yehuda, S. Keren, O. Barkan, R. Ronen, and N. Koenigstein, ”Segllm: Topic-oriented call segmentation via llm-based conversation synthesis,” in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 11361-11365.

I. Malkiel, U. Alon, Y. Yehuda, S. Keren, O. Barkan, R. Ronen, 和 N. Koenigstein，“Segllm: 基于大型语言模型的主题导向通话分段”，发表于 ICASSP 2024 - 2024 年 IEEE 国际声学、语音与信号处理会议，2024 年，第 11361-11365 页。

[5] F. Chi, Y. Wang, P. Nasiopoulos, and V. C. Leung, ”Multi-modal gpt-4 aided action planning and reasoning for self-driving vehicles,” in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 7325-7329.

F. Chi, Y. Wang, P. Nasiopoulos, 和 V. C. Leung，“多模态 GPT-4 辅助自动驾驶车辆的动作规划与推理”，发表于 ICASSP 2024 - 2024 年 IEEE 国际声学、语音与信号处理会议，2024 年，第 7325-7329 页。

[6] X. Zhang, C. Tian, X. Yang, L. Chen, Z. Li, and L. R. Petzold, ”Alpacare: Instruction-tuned large language models for medical application,” arXiv preprint arXiv:2310.14558, 2023.

X. Zhang, C. Tian, X. Yang, L. Chen, Z. Li, 和 L. R. Petzold，“Alpacare: 面向医疗应用的指令调优大型语言模型”，arXiv 预印本 arXiv:2310.14558，2023 年。

[7] Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, and Z. Wei, ”Disc-medllm: Bridging general large language models and real-world medical consultation,” arXiv preprint arXiv:2308.14346, 2023.

Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, 和 Z. Wei，“Disc-medllm: 连接通用大型语言模型与真实医疗咨询”，arXiv 预印本 arXiv:2308.14346，2023 年。

[8] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu et al., ”Can generalist foundation models outcompete special-purpose tuning? case study in medicine,” Medicine, vol. 84, no. 88.3, pp. 77-3, 2023.

H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu 等，“通用基础模型能否超越专用调优？医学领域案例研究”，Medicine，卷 84，第 88.3 期，第 77-3 页，2023 年。

[9] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, ”Cognitive mirage: A review of hallucinations in large language models,” arXiv preprint arXiv:2309.06794, 2023.

H. Ye, T. Liu, A. Zhang, W. Hua, 和 W. Jia, “认知幻觉: 大型语言模型幻觉现象综述”, arXiv 预印本 arXiv:2309.06794, 2023 年。

[10] A. Pal, L. K. Umapathi, and M. Sankarasubbu, ”Med-HALT: Medical domain hallucination test for large language models,” in Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL). Singapore: Association for Computational Linguistics, Dec. 2023, pp. 314-334.

A. Pal, L. K. Umapathi, 和 M. Sankarasubbu, “Med-HALT: 大型语言模型医疗领域幻觉测试”, 发表于第 27 届计算自然语言学习会议 (CoNLL), 新加坡: 计算语言学协会, 2023 年 12 月, 第 314-334 页。

[11] R. Liu, R. Yang, C. Jia, G. Zhang, D. Yang, and S. Vosoughi, ”Training socially aligned language models on simulated social interactions,” in The Twelfth International Conference on Learning Representations, 2024.

R. Liu, R. Yang, C. Jia, G. Zhang, D. Yang, 和 S. Vosoughi, “基于模拟社交互动训练社会对齐语言模型”, 发表于第十二届国际学习表征会议, 2024 年。

[12] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, ”Improving factuality and reasoning in language models through multiagent debate,” in Forty-first International Conference on Machine Learning, 2024.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, 和 I. Mordatch, “通过多智能体辩论提升语言模型的事实性和推理能力,” 载于第 41 届国际机器学习大会, 2024 年。

[13] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, ”ChatDev: Communicative agents for software development,” in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Aug. 2024, pp. 15174-15186.

C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, 和 M. Sun, “ChatDev: 面向软件开发的交互式智能体,” 载于第 62 届计算语言学协会年会论文集 (第一卷: 长文), 计算语言学协会, 2024 年 8 月, 页 15174-15186。

[14] J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, ”Generative agents: Interactive simulacra of human behavior,” in Proceedings of the 36th annual acm symposium on user interface software and technology, 2023, pp. 1-22.

J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, 和 M. S. Bernstein, “生成式智能体: 人类行为的交互式模拟,” 载于第 36 届 ACM 用户界面软件与技术研讨会论文集, 2023 年, 页 1-22。

[15] Z. Du, C. Qian, W. Liu, Z. Xie, Y. Wang, Y. Dang, W. Chen, and C. Yang, ”Multi-agent software development through cross-team collaboration,” arXiv preprint arXiv:2406.08979, 2024.

Z. Du, C. Qian, W. Liu, Z. Xie, Y. Wang, Y. Dang, W. Chen, 和 C. Yang, “通过跨团队协作实现多智能体软件开发,” arXiv 预印本 arXiv:2406.08979, 2024 年。

[16] C. Qian, Z. Xie, Y. Wang, W. Liu, Y. Dang, Z. Du, W. Chen, C. Yang, Z. Liu, and M. Sun, ”Scaling large-language-model-based multi-agent collaboration,” arXiv preprint arXiv:2406.07155, 2024.

C. Qian, Z. Xie, Y. Wang, W. Liu, Y. Dang, Z. Du, W. Chen, C. Yang, Z. Liu, 和 M. Sun, “基于大规模语言模型的多智能体协作扩展,”arXiv 预印本 arXiv:2406.07155, 2024 年。

[17] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber, ”MetaGPT: Meta programming for a multi-agent collaborative framework,” in The Twelfth International Conference on Learning Representations, 2024.

S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, 和 J. Schmidhuber, “MetaGPT: 面向多智能体协作框架的元编程,”载于第十二届国际学习表征会议, 2024 年。

[18] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein, ”Medagents: Large language models as collaborators for zero-shot medical reasoning,” in ICLR 2024 Workshop on Large Language Model (LLM) Agents, 2024.

X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, 和 M. Gerstein, “Medagents: 作为零样本医学推理协作者的大型语言模型,”载于 ICLR 2024 大型语言模型 (LLM) 智能体研讨会, 2024 年。

[19] L. Chen, J. Q. Davis, B. Hanin, P. Bailis, I. Stoica, M. Zaharia, and J. Zou, ”Are more llm calls all you need? towards scaling laws of compound inference systems,” arXiv preprint arXiv:2403.02419, 2024.

L. Chen, J. Q. Davis, B. Hanin, P. Bailis, I. Stoica, M. Zaharia, 和 J. Zou, “更多的 LLM 调用真的是全部所需吗？复合推理系统的扩展规律探索,”arXiv 预印本 arXiv:2403.02419, 2024 年。

[20] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, ”Memorybank: Enhancing large language models with long-term memory,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, 2024, pp. 19724-19731.

W. Zhong, L. Guo, Q. Gao, H. Ye, 和 Y. Wang, “Memorybank: 通过长期记忆增强大型语言模型,”载于 AAAI 人工智能会议论文集, 第 38 卷第 17 期, 2024 年, 页 19724-19731。

[21] A. Zhao, D. Huang, Q. Xu, M. Lin, Y.-J. Liu, and G. Huang, ”Expel: Llm agents are experiential learners,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, 2024, pp. 19632-19642.

A. Zhao, D. Huang, Q. Xu, M. Lin, Y.-J. Liu, 和 G. Huang, “Expel:LLM 智能体作为经验学习者,”载于 AAAI 人工智能会议论文集, 第 38 卷第 17 期, 2024 年, 页 19632-19642。

[22] C. Qian, Y. Dang, J. Li, W. Liu, Z. Xie, Y. Wang, W. Chen, C. Yang, X. Cong, X. Che, Z. Liu, and M. Sun, ”Experiential co-learning of software-developing agents,” in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Aug. 2024, pp. 5628-5640.

C. Qian, Y. Dang, J. Li, W. Liu, Z. Xie, Y. Wang, W. Chen, C. Yang, X. Cong, X. Che, Z. Liu, 和 M. Sun, “软件开发智能体的经验共学,”载于第 62 届计算语言学协会年会论文集 (第一卷: 长文)，计算语言学协会, 2024 年 8 月, 页 5628-5640。

[23] C. Qian, J. Li, Y. Dang, W. Liu, Y. Wang, Z. Xie, W. Chen, C. Yang, Y. Zhang, Z. Liu et al., ”Iterative experience refinement of software-developing agents,” arXiv preprint arXiv:2405.04219, 2024.

C. Qian, J. Li, Y. Dang, W. Liu, Y. Wang, Z. Xie, W. Chen, C. Yang, Y. Zhang, Z. Liu 等, “软件开发智能体的迭代经验优化,”arXiv 预印本 arXiv:2405.04219, 2024 年。

[24] S. Jiang, Y. Wang, and Y. Wang, ”Selfevolve: A code evolution framework via large language models,” arXiv preprint arXiv:2306.02907, 2023.

S. Jiang, Y. Wang, 和 Y. Wang, “Selfevolve: 通过大型语言模型实现代码演化的框架,”arXiv 预印本 arXiv:2306.02907, 2023。

[25] J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, and Y. Liu, ”Agent hospital: A simulacrum of hospital with evolvable medical agents,” arXiv preprint arXiv:2405.02957, 2024.

J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, 和 Y. Liu, “Agent hospital: 一个具有可进化医疗代理的医院模拟系统,”arXiv 预印本 arXiv:2405.02957, 2024。

[26] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, ”What disease does this patient have? a large-scale open domain question answering dataset from medical exams,” Applied Sciences, vol. 11, no. 14, p. 6421, 2021.

D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, 和 P. Szolovits, “这位患者患有什么疾病? 来自医学考试的大规模开放领域问答数据集,”应用科学, 第 11 卷，第 14 期，6421 页，2021。

[27] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, ”Pubmedqa: A dataset for biomedical research question answering,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2567-2577.

Q. Jin, B. Dhingra, Z. Liu, W. Cohen, 和 X. Lu, “Pubmedqa: 一个生物医学研究问答数据集,”载于 2019 年自然语言处理实证方法会议暨第九届国际联合自然语言处理会议 (EMNLP-IJCNLP) 论文集, 2019，页 2567-2577。

[28] A. . M. Llama Team, ”The llama 3 herd of models,” arXiv preprint arXiv:2407.21783, 2024.

A. . M. Llama Team, “Llama 3 模型群,”arXiv 预印本 arXiv:2407.21783, 2024。

[29] J. Zeng, B. Zhang, Y. Ma, K. Sun, H. Zhou, Y. Liu et al., ”Chatglm: A family of large language models from glm-130b to glm-4 all tools,” arXiv preprint arXiv:2406.12793, 2024.

J. Zeng, B. Zhang, Y. Ma, K. Sun, H. Zhou, Y. Liu 等, “Chatglm: 从 glm-130b 到 glm-4 all tools 的大型语言模型家族,”arXiv 预印本 arXiv:2406.12793, 2024。

[30] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao et al., ”Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model,” arXiv preprint arXiv:2405.04434, 2024.

A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao 等, “Deepseek-v2: 一种强大、经济且高效的专家混合语言模型,”arXiv 预印本 arXiv:2405.04434, 2024。