文章编号： 1673-3193（2025）02-0219-11　　J. North Univ. China, Nat. Sci. Ed., 2025,46(2):219-229.

大语言模型在多代理辩论中作为辩论者表现的比较分析

张立炎，梁志剑

（中北大学 计算机科学与技术学院，山西 太原 030051）

摘　要： 为了深入探索大型语言模型（Large Language Models，LLMs）在模拟人类智能，特别是辩论能力方面的潜力与局限性，将思维链（Chain-of-Thought，CoT）与检索增强生成（Retrieval-Augmented Generation，RAG）技术相结合应用到多代理辩论（Multi-Agent Debate，MAD）中，构建了一套多代理辩论框架——CoRAG-MAD，旨在模拟人类辩论比赛流程，包括开篇立论、质询环节、自由辩论和总结陈词四个阶段。设计了公平辩论（Fair Debate）、不平等辩论（Unequal Debate）和混合辩论（Mixed Debate）三种不同的辩论场景，通过自动化评估工具与人工专家评审相结合的方式，对辩论内容进行了深度分析。以OrChiD数据集为测试平台，实验结果表明，CoRAG-MAD可以有效提高LLMs在各个辩论场景中的多项能力。具体而言，在不平等辩论中，LLMs的逻辑推理得分提升57.56%，创造力得分提升49.77%；在混合辩论中，LLMs的协作能力提升23.36%，整体性能提升28.20%。本文进行了消融实验和对比实验，验证了CoT模块在增强逻辑推理能力方面、RAG模块在提升事实准确性和激发创新思维方面以及CoRAG方法在MAD中的有效性。

关键词： 多代理辩论；检索增强生成；思维链；大语言模型；NLP

中图分类号： TP391　　**文献标识码：** A　　**doi:** 10.62756/jnuc.issn.1673-3193.2024.08.0016

引用格式： 张立炎，梁志剑. 大语言模型在多代理辩论中作为辩论者表现的比较分析[J]. 中北大学学报（自然科学版），2025，46(2)：219-229.

ZHANG Liyan，LIANG Zhijian. A Comparative analysis of large language models as debaters' performance in multi-agent debates[J]. Journal of North University of China(Natural Science Edition)，2025，46(2)：219-229.

A Comparative Analysis of Large Language Models as Debaters' Performance in Multi-Agent Debates

ZHANG Liyan，LIANG Zhijian

（School of Computer Science and Technology，North University of China，Taiyuan 030051，China）

Abstract： In order to explore the potential and limitations of Large Language Models（LLMs）in simulating human intelligence，particularly in debate capabilities，a framework called CoRAG-MAD was constructed that integrated Chain-of-Thought（CoT）and Retrieval Augmented Generation（RAG）techniques into Multi-Agent Debate（MAD）. It was designed to simulate the process of human debating competition，including four stages：opening statements，attack and defence，free debate，and closing statements. It was employed in three distinct debate scenarios：fair debate，unequal debate，and mixed debate. By combining automated evaluation tools

收稿日期： 2024-08-15

作者简介： 张立炎（1998－），男，硕士生，主要从事人工智能、自然语言处理的研究

通信作者： 梁志剑（1978－），男，教授，博士，主要从事人工智能、自然语言处理等研究。E-mail：116585916@qq.com。

and human expert review, a thorough analysis of the debate content was conducted. The experiment, using the OrChiD dataset as the test platform, shows that CoRAG-MAD can effectively improve several abilities of LLMs in various debate scenarios. Specifically, in the unequality debate, LLMs' logical reasoning score improves up to 57.56% and creativity score improves up to 49.77%; in the mixed debate, LLMs' collaborative ability improves up to 23.36%, and overall performance improves up to 28.20%. This paper presented ablation and comparative experiments, which were conducted to verify the effectiveness of the CoT in enhancing logical reasoning, the RAG in enhancing factual accuracy and stimulating creative thinking, and the CoRAG approach in MAD.

Key words: multi-agent debate; retrieval-augmented generation; chain-of-thought; large language models; NLP

0 引 言

鉴于大语言模型(Large Language Models, LLMs)[1-3]已普遍融入日常社会协作环境,开发具有复杂社会智能的人工智能变得日益迫切。多代理(Multi-Agent)系统[4-8]通过集成多个LLMs并进行协作,实现了针对不同的子任务或方面的专业化处理和高效协同。但是,LLMs在模拟人类智能方面仍有着诸多挑战,如理解深层语境、推断隐含意义、历史记忆缺失等。辩论作为一种高度复杂且综合的人类智能活动,涵盖了逻辑推理、语言表达等多个关键方面,是全面评估LLMs模拟人类智能水平的理想测试场景。因此,本研究旨在通过构建CoRAG-MAD辩论框架,将思维链(Chain-of-Thought, CoT)与检索增强生成(Retrieval-Augmented Generation, RAG)技术有机结合,深入探索LLMs在辩论场景下的逻辑推理、实时准确性等方面的能力与局限。本文研究表明,在辩论场景中,CoRAG方法在多项关键评估指标上提高了LLMs的性能,彰显了其卓越的适应性与稳健性。不过,它对"创造性"的影响有限。总体而言,LLMs在CoRAG-MAD中的表现,凸显了其进一步完善的潜力。

本文的贡献如下:

1)新颖的实验设计:提出了公平辩论、不平等辩论和混合辩论三种辩论场景。通过多样化的场景设置,全面考察LLMs在不同环境下的适应性与表现,为深入理解LLMs在复杂社会交互中的行为提供了丰富视角。

2)RAG与CoT有机整合:将RAG与CoT技术有机整合并应用到多代理辩论框架中。RAG技术通过检索外部知识,有效弥补了LLMs内部知识的局限性,增强了事实准确性;CoT技术则帮助其在辩论中构建逻辑清晰的论证链条,提升了推理能力。这种技术融合的方式为提升LLMs性能提供了新的思路和方法。

3)比较分析:采用自动化评估与人工专家评审相结合的方式对LLMs的辩论表现进行了评估。自动化评估对文本生成的质量进行量化分析,人工专家评审则从专业语言学、辩论学角度进行深入评价,两者相辅相成,确保了评估结果的客观性和全面性,从而能够更全面、深入地揭示LLMs在模拟人类智能辩论过程中的能力与不足。

1 相关工作

1.1 CoT

思维链(Chain-of-Thought, CoT)推理[9]极大地增强了语言模型的推理能力,这项技术不仅展示了LLMs制定问题解决策略的能力,而且还推动了诸如least-to-most prompting[10]、zero-shot[11]、self-consistency[12]、无提示词的zero-shot CoT[13]等方法的进步。此外,思维树(Tree-of-Thought)[14]和思维图(Graph-of-Thought)[15]等推理框架的出现则克服了线性CoT的限制,拓宽了推理轨迹的范围。

但是,现有的CoT方法在处理复杂推理任务时仍然面临挑战,特别是在逻辑连贯性方面,当推理链条较长或问题较为复杂时,可能无法始终保持严密的逻辑关系,导致推理过程出现跳跃或不连贯的情况。在事实准确性方面,CoT主要依赖模型内部的知识和训练数据,如果数据存在偏差或不完整,可能会影响其生成的推理链条的准确性。

1.2　MAD

多代理辩论（Multi-Agent Debate，MAD）将LLMs中的CoT推理提升到了一个新的水平，主要提高了其事实准确性和推理能力。MAD通过为LLMs实例分配不同角色来模拟人类对话，并鼓励得出最佳答案。最近的研究证明了MAD在提高LLMs性能方面的有效性，其性能优于自我反思等技术[16]。例如，利用ChatGPT实施一种“交叉考官”方法[17]，以增强事实评估；定制多代理裁判小组[18]，用于评估NLG任务；与单代理或单LLMs的多代理形式相比，多LLMs辩论表现出更优越的性能[19]。MAD通过多视角和协作决策体现了人类推理的复杂性，提高了事实的准确性，是提高LLMs认知能力的综合方法。

不过，MAD在如何有效整合不同代理的观点以增强辩论的逻辑性和说服力方面仍有待改进。在实际辩论中，不同代理的观点可能存在冲突或不一致，如果不能有效地协调和整合这些观点，可能会导致辩论过程混乱，影响最终结论的质量。此外，MAD对计算资源的要求较高，多个代理之间的交互和推理需要大量的计算资源支持，这在一定程度上限制了其应用范围。

1.3　RAG

检索增强生成（Retrieval-Augmented Generation，RAG）利用检索到的知识来增强LLMs的能力[20]。RAG为LLMs与外部环境交互提供了一种经济有效的方式，使其无需进行昂贵的模型更新[21-24]。它的多功能性体现在各种应用中，如代码生成[25-27]、问题解答[28-29]和创意写作[30]。最近的发展，如IRCoT[31]、IRGR[32]和ITRG[33]，通过整合推理元素进一步提高了RAG的有效性。然而，在实际使用中，RAG对检索系统的依赖性较强，如果检索系统返回的结果不准确或不相关，可能会影响模型的性能。此外，由于RAG需要花费时间进行知识检索和整合，因此其在处理实时性要求较高的任务时可能存在一定的延迟。

2　CoRAG-MAD

2.1　Multi-Agent Debate

受国际华语辩论邀请赛的启发，本文设计了一个MAD流程（如图1所示），主要包括开篇立论、质询环节、自由辩论与总结陈词四个环节。

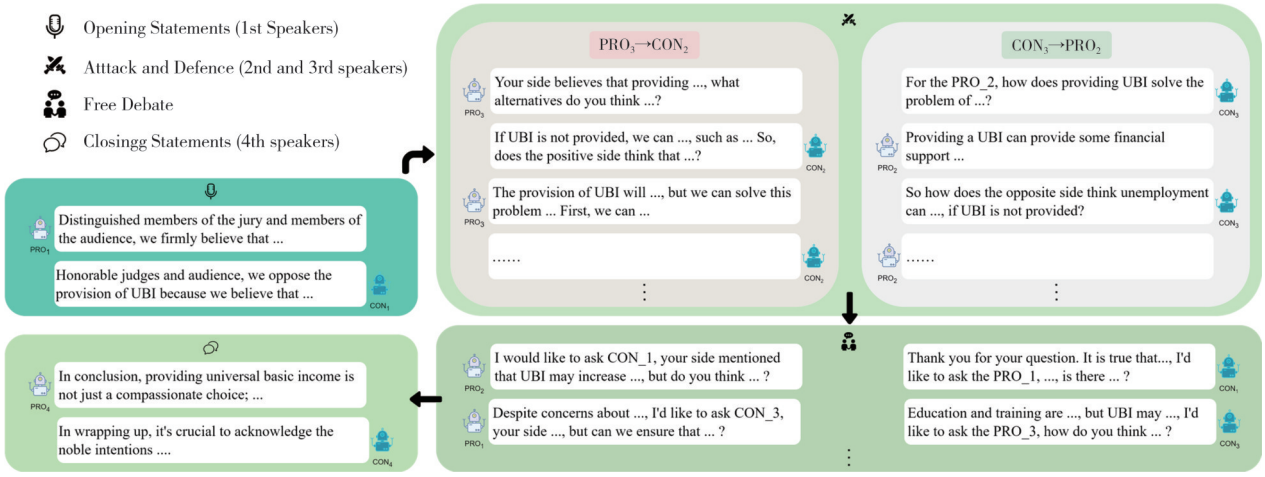

图1　多代理辩论流程

Fig. 1　The multi-agent debate rounds

为了深入探讨这一过程，使用表1中详列的符号和术语模拟了多代理辩论，它涉及8个LLMs代理，由4个正方 $\mathcal{P}=\{p_i\}_{i=1}^{4}$ 和4个反方 $\mathcal{C}=\{c_i\}_{i=1}^{4}$ 构成，记为 $\mathcal{A}=\{p_i, c_i\}_{i=1}^{4}$，每个LLMs代理的能力定义为 $\{l_{-1}, l_0, l_{+1}\}$，分别代表能力较弱、相当和较强。根据能力的高低，将其分配至图2所示的3种辩论场景 $\mathcal{S}=\{S_1, S_2, S_3\}$ 中，分别代表平等、不平等和混合辩论场景。表2根据MMLU得分[34]定义了每个代理的能力 $\mathcal{L}$，并据此分配到3种辩论场景中。

$$S_1=\{(c_i, p_i)_{i=1}^{4} \leftarrow l_0\}, \tag{1}$$

$$S_2=\{(c_i)_{i=1}^{4} \leftarrow l_{+1}, (p_i)_{i=1}^{4} \leftarrow l_{-1}\}, \tag{2}$$

$$S_3=\begin{cases}(c_1, c_3, p_1, p_3) \leftarrow l_{+1}, \\ (c_2, c_4, p_2, p_4) \leftarrow l_{-1}, \end{cases} \tag{3}$$

式中：$(c_i, p_i)\leftarrow l_k$表示代理c_i和p_i被分配到的能力等级为l_k。

表1 符号说明

Tab. 1 The description of the symbols

符号	含义
$\mathcal{A}$	代理实例集合
$\mathcal{P}$	正方
$\mathcal{C}$	反方
p_i/c_i	正方/反方的第i位代表
$\mathcal{S}$	辩论场景集合
S_j	第j个辩论场景
$\mathcal{L}$	基于MMLU的LLMs的能力分类
l_{-1}	能力较弱的LLMs
l_0	能力相当的LLMs
l_{+1}	能力较强的LLMs

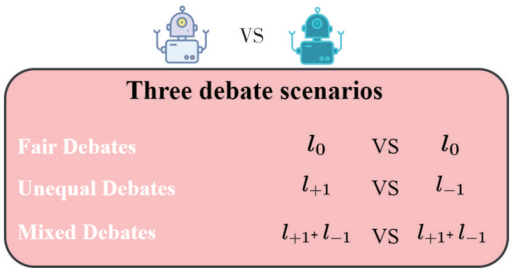

图2 3种辩论场景

Fig. 2 Three debate scenarios

表2 3种不同辩论场景的模型设置

Tab. 2 The detailed model setup for the three different debate scenarios

$\mathcal{S}$	$\mathcal{P}/\mathcal{C}$	模型	MMLU	$\mathcal{L}$
S_1	$\mathcal{P}$	Claude3-Haiku	75.2	l_0
		Qwen1.5-32B	73.4	l_0
		Mixtral-8×7B-MoE	70.6	l_0
		Phi-3-mini 3.8B	68.8	l_0
		Phi-3-small 7B	75.3	l_0
	$\mathcal{C}$	Grok-1	73.0	l_0
		GPT-3.5	70.0	l_0
		Llama3-8B-Instruct	68.4	l_0
S_2	$\mathcal{P}$	Claude3-Opus	86.8	l_{+1}
		GPT-4	86.4	l_{+1}
		Gemini-ultra	83.7	l_{+1}
		Gemini 1.5 Pro	81.9	l_{+1}
	$\mathcal{C}$	Mixtral-8×7B-MoE	70.6	l_{-1}
		GPT-3.5	70.0	l_{-1}
		Phi-3-mini 3.8B	68.8	l_{-1}
		Llama3-8B-Instruct	68.4	l_{-1}
S_3	$\mathcal{P}$	GPT－4	86.4	l_{+1}
		Mixtral-8×7B-MoE	70.6	l_{-1}
		Gemini-ultra	83.7	l_{+1}
		Phi-3-mini 3.8B	68.8	l_{-1}
	$\mathcal{C}$	Claude3-Opus	86.8	l_{+1}
		GPT-3.5	70.0	l_{-1}
		Gemini 1.5 Pro	81.9	l_{+1}
		Llama3-8B-Instruct	68.4	l_{-1}

本文的主要目的是通过这种MAD框架，深入探讨LLMs在模拟人类智能方面的潜力。基于此目的，策划了一系列横跨不同领域的主题，通过使用CoT来增强LLMs的推理能力，确保其论证方法更加细致入微，逻辑性更强，同时还利用RAG方法来减少模型幻觉，从而提高生成答案的准确性。

算法1 CoRAG

输入： 代理集合$\mathcal{A}=\{p_i, c_i\}_{i=1}^{4}$和辩论主题$t$

输出： 辩论发言集合$\mathcal{E}$

1. 将辩论发言集合$\mathcal{E}$初始化为空
2. $\mathcal{A}$. **for Each**$((a)=>\{$
   #初始化CoT
   let Initial_CoT = **CoT**(a, t);
   #依据初始化CoT生成RAG信息
   let RAG_Info = **RAG**(Initial_CoT);
   #将两者结合生成初始辩论发言
   $\mathcal{E}_{pre}$ = **Integrate**(Initial_CoT, RAG_Info)$\}$);
   #为每个代理分配初始发言
   a.**push**($\{\mathcal{E}$: $\mathcal{E}_{pre}\}$);
   #将辩论发言添加到辩论发言集合中
   $\mathcal{E}$.**append**($\mathcal{E}_{adp}$)
   $)\}$
3. **Repeat Debate**(p_i, c_j)$\{$
   #根据p_i和c_j的辩论发言生成动态p_i的CoT
   let Adapt_CoT = **CoT**($p_i.\mathcal{E}$, $c_j.\mathcal{E}$);
   #依据动态CoT生成RAG信息
   let RAG_Info = **RAG**(Adapt_CoT);
   #将两者集合生成p_i的动态辩论发言
   $\mathcal{E}_{adp}$ = **Integrate**(Adapt_CoT, RAG_Info);
   #替换p_i之前的辩论发言
   $p_i.\mathcal{E}=\mathcal{E}_{adp}$;
   #将p_i的辩论发言添加到辩论发言集合中
   $\mathcal{E}$.**append**($\mathcal{E}_{adp}$);
   if (辩论没有结束)$\{$
   #基于p_i的发言生成c_j的动态辩论发言
   Debate(c_j, p_i);
   $\}$
   $\}$ **Until** 辩论结束.
4. **return** $\mathcal{E}$

2.2 CoRAG

在人工智能领域，MAD已成为模拟类人论证和推理的强大平台。为了进一步研究LLMs模拟人类智能的能力，本文引入了一种整合了CoT和RAG技术的新方法CoRAG，旨在加强MAD中提出的论点的逻辑连贯性和事实丰富性，为智能代理之间进行更细致入微的讨论奠定基础。有关本方法所使用的LLMs提示词的实例，请参阅图3。

完整的CoRAG流程如算法1所示，具体解释如下：

1）CoT生成：每个代理根据辩论主题或双方发言使用CoT函数生成初始CoT，有助于代理初步构建论点。

2）RAG信息生成：通过RAG方法检索相关

数据，增强CoT，确保论点的事实支持。

3）信息融合：将检索到的信息与CoT整合，生成更加精炼丰富的发言。

4）动态适应：代理根据对手的论据动态调整其发言，以保持逻辑连贯性。

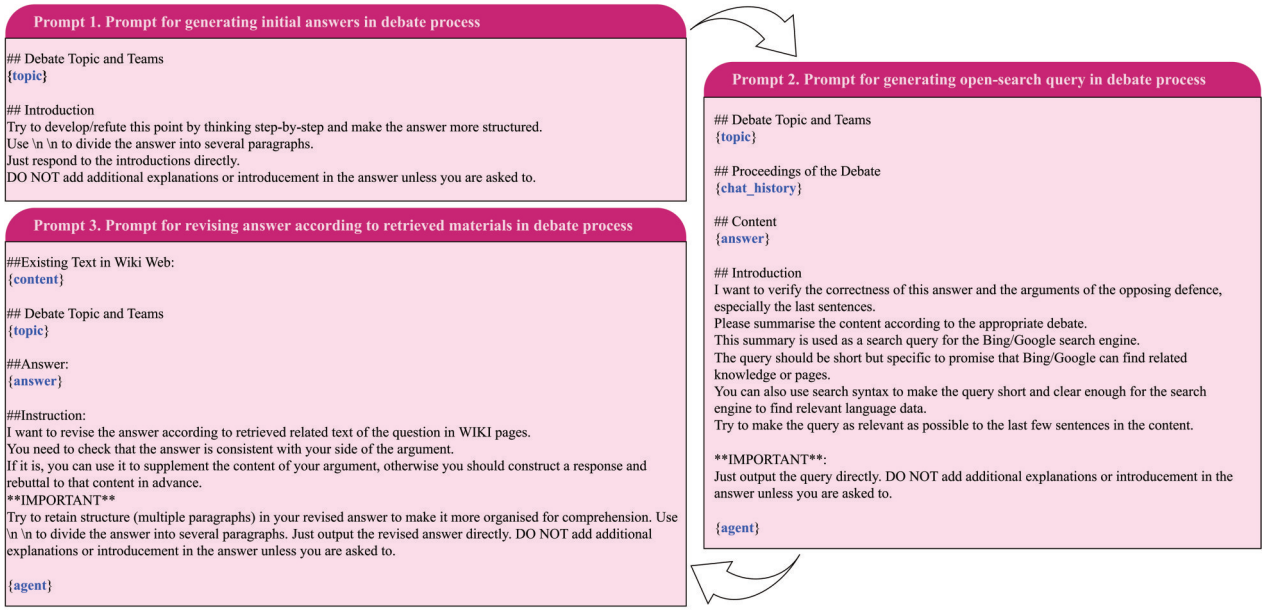

图 3　提示词实例

Fig. 3　The detailed settings for prompts

CoRAG利用RAG技术为CoT提供丰富准确的信息源，在一定程度上弥补了CoT在事实准确性方面的不足。通过信息融合机制将检索到的信息与推理链条相结合，从而增强了LLMs输出内容的质量。此外，CoRAG还通过动态适应机制使LLMs在辩论过程中实时调整其论证，优化推理流程以降低延迟，提高辩论表现的说服力。

3　实验与分析

3.1　评价指标

本文将自动化评估与人工评估相结合，全面评估LLMs在模拟人类辩论方面的优势和局限性，从而更深入地了解其模拟人类智能的能力。

3.1.1　自动评估

为了确保对代理人的辩论表现进行全面、公正的评估，本文设计了一套自动化评分标准，包括BLEU、PPL、ROUGE-2和ROUGE-L。

BLEU：BLEU是一种用于评估机器翻译文本与参考翻译文本之间相似度的指标，在本实验中用于衡量生成的辩论文本与理想文本在词汇和结构上的相似程度。其计算公式为

$$BLEU = BP \times \mathrm{e}^{\sum_{n=1}^{N} w_n \log p_n}, \tag{4}$$

式中：N为计算1—N元组的精度；w_n为权重，一般设置为$\frac{1}{N}$，使得每个n-gram的权重相等；p_n为生成文本与参考文本中共同出现的n-gram的最大数量与生成文本中n-gram的数量的比值；BP（Brevity Penalty）为简短惩罚因子，用于惩罚生成文本过短的情况。

$$BP = \begin{cases} 1, & \text{if } c > r, \\ \mathrm{e}^{1-\frac{r}{c}}, & \text{if } c \leqslant r, \end{cases} \tag{5}$$

式中：c为生成文本的长度（以单词或词元为单位）；r为参考文本中最短文本的长度。

PPL：PPL通过计算模型生成文本概率倒数的几何平均值来评估语言模型的连贯性和流畅性，其计算公式为

$$PPL = \left(\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1}, \cdots, w_1)} \right)^{\frac{1}{N}}, \tag{6}$$

式中：N为文本中的单词或词元数量；$P(w_i|w_{i-1}, \cdots, w_1)$表示给定前$i-1$个单词或词元的情况下，模型预测第$i$个单词或词元的概率。$PPL$值越低，表示模型生成的文本越流畅、合理。

ROUGE-2：ROUGE - 2主要评估生成文本与参考文本之间在bigram上的重叠程度，以衡量相关论据的引用情况。计算公式为

$$ROUGE\text{-}2 = \frac{\sum_{S\in\{R\}}\sum_{b\in S} c_m(bi)}{\sum_{S\in\{R\}}\sum_{b\in S} c(bi)}, \quad (7)$$

式中：$\{R\}$为参考文本集合；$c_m(bi)$为生成文本与参考文本中共同出现的bigram的数量；$c(bi)$为参考文本中bigram的总数量。

ROUGE－L：ROUGE-L通过计算生成文本与参考文本之间的最长公共子序列(Longest Common Subsequence，LCS)的长度来评估文本的连贯性和论据组织。计算公式为

$$ROUGE\text{-}L = \frac{LCS(X, Y)}{m}, \quad (8)$$

式中：X为生成文本；Y为参考文本；$LCS(X, Y)$为X和Y之间的最长公共子序列的长度；m为参考文本的长度(以单词或词元为单位)。ROUGE-L的值越高，表示生成文本与参考文本在结构和内容上越相似，连贯性越好。

3.1.2 人工评估

为了深入分析辩论内容，本文设计了如表3所示的人工评估指标，重点关注4个关键领域：逻辑推理(Logical Reasoning，LR)、创造力(Creativity，C_1)、协作(Collaboration，C_2)和整体性能(Overall Performance，OP)。LR评估代理人在辩论中构建逻辑论据和识别对手论据中逻辑缺陷的能力；C_1评估其提出的原创性观点和解决方案的能力；C_2衡量其作为一个团队完成任务的效率；OP综合评估其辩论技巧。

表3 人工评估指标的详细设计

Tab.3 The detailed design of manual evaluation

指标	评分维度	评分范围	评分标准
逻辑推理(LR)	论证连贯性	0～20	逻辑严密且推理过程连贯的论证
	观点明晰性	0～10	观点清晰且表达准确
	证据支撑性	0～10	充分的证据和恰当的引用
创造力(C_1)	思想独创性	0～15	所呈现的观点或论证的新颖性与独特性
	创新解决问题	0～15	以创造性方式解决问题的能力
	思维创造性	0～10	创造性思维及策略的展示
协作(C_2)	团队协作能力	0～15	与团队成员进行有效协作的能力
	团队贡献	0～10	积极参与并提供有价值的见解
	辩论协同能力	0～15	在辩论中与他人协同工作的能力
整体性能(OP)	整体表现	0～40	语言能力、逻辑推理能力与创造力
	策略适应性	0～15	根据辩论进程灵活调整策略的能力
	说服力	0～15	在说服对手和评委方面的有效性

为了确保人工评估的公正性，组建了一支多元化评估团队，成员包括语言学专家和辩论领域资深人士，团队成员根据表3提出的评分标准，对LLMs在辩论活动中的表现进行评估。评估过程中，团队成员之间不得进行任何交流，对于存在差异较大的评分，组织专家进行讨论并重新评估，以确保评价结果的独立性和可靠性。

3.2 数据集

在本研究的实验环节，采用OrChiD数据集[35]，该数据集由1 218场真实世界的辩论组成，这些辩论是在476个独特的主题上用中文进行的，包含2 436个特定立场的摘要和14 133个完全注释的话语，为评估LLMs在不同辩论场景下的表现提供了一个理想的测试平台。

3.3 实验配置

根据LLMs在MMLU中的得分将其分为3个不同的能力层级(l_{-1}, l_0, l_{+1})，并分配至3种不同的辩论场景(S_1, S_2, S_3)中，具体划分情况见表2。

为了确保辩论内容的广泛性，本文随机抽取OrChiD中的辩题，分别引入之前设计的3种不同的辩论场景中，为准确衡量LLMs代理的辩论性能提供坚实基础。

3.4 实验过程

实验步骤如下：1) 辩题分配。随机为辩论队伍分配辩题，告知辩论中的角色(正方和反方)

2) 辩论执行。LLMs使用MAD框架进行辩论。在辩论过程中，LLMs代理通过RAG技术检索相关信息，并利用CoT将这些信息融合到辩论中。

3) 结果记录。辩论过程中，将代理的输出记录下来，以便进行后续的评估和分析。

4) 评估与分析。使用自动化评估工具和人工评估指标对辩论结果进行评估。

3.5　消融实验

3.5.1　实验设置

为了验证CoRAG-MAD框架中各核心模块的作用和影响，本文设计了一系列消融实验。这些实验包括：

1）独立模块实验：本实验专注于探究CoT（或RAG）模块在辩论任务中的独立效能。本文构建了实验对照组，其中仅激活CoT（或RAG）模块，同时禁用RAG（或CoT）模块。通过对比分析启用与未启用CoT（或RAG）模块的LLMs在辩论任务中的具体表现，旨在精确量化CoT（或RAG）模块对辩论性能的独立贡献与影响。

2）CoT与RAG集成实验：在这一实验中，同时激活CoT和RAG模块，将其共同作用于LLMs的辩论任务中，以评估其对LLMs辩论性能的综合影响。

3.5.2　实验结果

在表4中，本文以Llama3-8B- Instruct为例，展示了CoRAG各个模块对辩论的影响。具体而言，仅激活CoT模块时，在3种辩论场景（S_1, S_2, S_3）中，模型在逻辑推理（LR）维度的平均得分较未启用CoT模块时呈现提升态势。其中，在不平等辩论场景（S_2）下，该提升效果尤为突出，彰显了CoT模块在增强模型逻辑推理能力方面的潜力，进一步揭示了其在应对非对称辩论环境中的独特优势。

表4　在MAD中使用CoRAG前后的自动评估指标分数比较

Tab. 4　Comparison of automated assessment metric scores before and after the use of CoRAG in MAD

辩论场景	模型	自动化评分标准			
		BLEU	*PPL*	*ROUGE-2*	*ROUGE-L*
S_1	Claude3-Haiku	0.287 (0.342)	12.45 (9.86)	0.143 (0.194)	0.291 (0.362)
	Qwen1.5-32B	0.312 (0.378)	9.87 (8.12)	0.168 (0.217)	0.324 (0.395)
	Mixtral-8×7B-MoE	0.296 (0.354)	11.02 (9.63)	0.154 (0.201)	0.305 (0.371)
	Phi-3-mini 3.8B	0.274 (0.326)	13.21 (10.98)	0.317 (0.183)	0.282 (0.352)
	Phi-3-small 7B	0.301 (0.365)	10.43 (8.94)	0.159 (0.208)	0.312 (0.387)
	Grok-1	0.293 (0.349)	11.56 (9.27)	0.149 (0.198)	0.301 (0.368)
	GPT-3.5	0.327 (0.392)	9.23 (7.45)	0.175 (0.231)	0.337 (0.413)
	Llama3-8B-Instruct	0.315 (0.379)	10.12 (8.09)	0.169 (0.219)	0.328 (0.396)
S_2	Claude3-Opus	0.281 (0.327)	25.23 (10.45)	0.149 (0.192)	0.287 (0.354)
	GPT-4	0.315 (0.368)	11.78 (9.12)	0.171 (0.213)	0.324 (0.381)
	Gemini-ultra	0.294 (0.342)	13.45 (11.67)	0.158 (0.189)	0.302 (0.361)
	Gemini 1.5 Pro	0.278 (0.324)	16.31 (12.83)	0.144 (0.177)	0.281 (0.342)
	Mixtral-×7B-MoE	0.267 (0.311)	17.56 (14.29)	0.136 (0.168)	0.273 (0.328)
	GPT-3.5	0.301 (0.353)	12.90 (10.76)	0.162 (0.198)	0.311 (0.372)
	Phi-3-mini 3.8B	0.254 (0.296)	18.73 (16.42)	0.129 (0.159)	0.261 (0.317)
	Llama3-8B-Instruct	0.289 (0.331)	14.87 (13.15)	0.153 (0.184)	0.296 (0.357)
	GPT-4	0.310 (0.354)	12.15 (10.42)	0.169 (0.197)	0.325 (0.368)
S_3	Mixtral-8×7B-MoE	0.285 (0.321)	14.78 (12.87)	0.148 (0.178)	0.296 (0.339)
	Gemini-ultra	0.302 (0.342)	13.27 (11.53)	0.161 (0.190)	0.312 (0.357)
	Phi-3-mini 3.8B	0.263 (0.298)	16.92 (15.41)	0.134 (0.159)	0.274 (0.315)
	Claude3-Opus	0.297 (0.335)	13.89 (11.94)	0.157 (0.183)	0.308 (0.349)
	GPT-3.5	0.305 (0.339)	12.64 (11.36)	0.164 (0.186)	0.319 (0.352)
	Gemini 1.5 Pro	0.280 (0.312)	15.36 (13.72)	0.145 (0.169)	0.291 (0.327)
	Llama3-8B-Instruct	0.292 (0.327)	14.23 (12.98)	0.152 (0.176)	0.301 (0.341)

当仅激活RAG模块时，3种辩论场景下模型在创造力（C_1）维度上均有所提升。其中，在公平辩论场景（S_1）下的提升效果最为显著，达到了45.13%。这一结果表明，RAG模块在为模型提供外部知识补充的过程中，能够激发模型在观点生成和论证策略制定方面的创新性思维，从而在相对公平的辩论环境中，使模型更有可能提出新颖独特的见解和解决方案。

当同时激活CoT和RAG模块（即完整的CoRAG方法）时，通过对比GPT-4.0和GPT-3.5在不同辩论场景中质询阶段使用CoRAG前后的人工评估结果（如图4所示），并结合表4中的自动评估结果可以看出，该方法在各种指标上都提高了LLMs的性能，凸显了其在辩论领域的广泛适用性和卓越性。

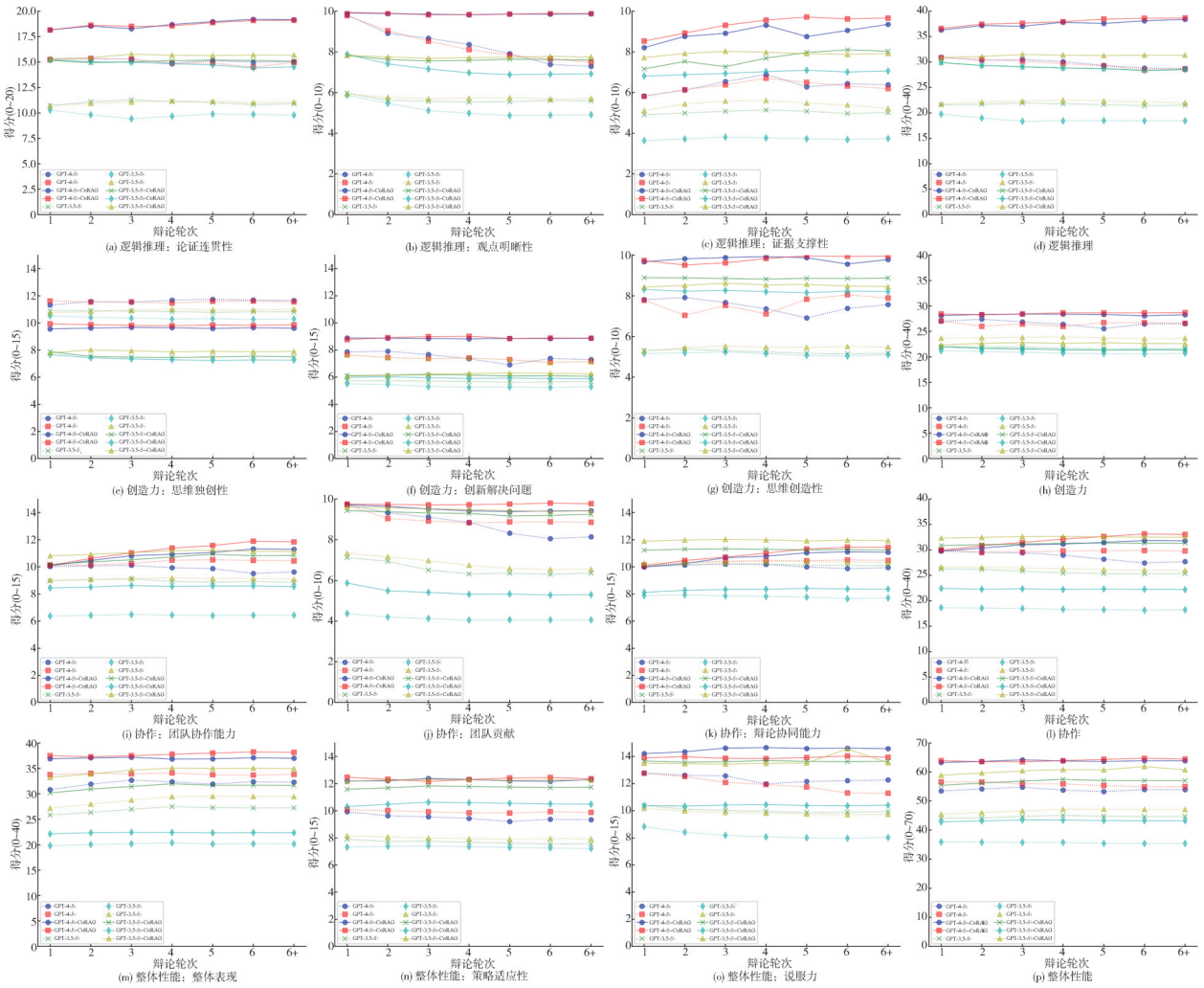

图4 在MAD的质询环节使用CoRAG前后的人工评估指标平均得分(以GPT-4.0和GPT-3.5为例)

Fig. 4 The average scores of manual metrics before and after utilizing CoRAG during the attack and defence phase of MAD (Taking GPT-4.0 and GPT-3.5 as examples)

3.6 对比实验

3.6.1 实验设置

为了更全面、准确地评估CoRAG-MAD框架的有效性，本文选取了以下几种当前优化LLMs性能的前沿方法来进行对比实验分析。

1) LaTRO[36]：隐性推理优化(Latent Reasoning Optimization)，该方法将推理过程类比为从潜在分布中采样，用变分推断方法进行优化，无需依赖外部反馈过奖励机制，基于自我奖励来提升LLMs的复杂推理能力。

2) RAP[37]：计划推理(Reasoning via Planning)，该方法通过将LLMs同时充当世界模型和推理代理，使其能够模拟世界状态并预测行动结果，并通过蒙特卡罗树搜索在探索和利用之间实现有效平衡。

3) RE2[38]：问题重读(Re-Reading the question as input)，该方法通过重新审视嵌入在输入提示中的问题信息，使LLMs能够提取更深层次的见解，识别复杂的模式，并建立更细致的联系。

3.6.2 实验结果

在以Llama3-8B-Instruct为例的对比实验中，不同方法在各项评估指标上呈现出不同的表现。根据表5数据可知，CoRAG方法在多个方面表现突出，显著优于其他方法。

具体而言，LaTRO方法，在LR上提升明显，在3种不同的辩论场景下，相比基准模型，得分分别提升7.42，7.10和5.88，但与CoRAG仍有差距；RAP方法在LR和C_2方面表现尚可，但在C_1提升幅度上不及CoRAG；RE2方法各项指标提升较为平稳，但整体提升程度相对较低，尤其在LR方面与CoRAG存在明显差距。

表 5　*LR*, C_1, C_2, *OP*的评估结果（以Llama3-8B-Instruct为例）

Tab. 5　Evaluation results on *LR*, C_1, C_2, *OP*（using Llama3-8B-Instruct as an example）

方法	*LR*	C_1	C_2	*OP*	$\mathcal{S}$
	21.51	16.13	25.86	44.68	S_1
—	18.26	15.59	18.33	35.52	S_2
	22.43	18.28	26.37	46.98	S_3
	29.76	15.92	27.31	51.03	S_1
＋CoT	25.69	15.76	20.19	41.78	S_2
	29.87	17.84	27.85	53.08	S_3
	21.47	23.41	28.10	51.18	S_1
＋RAG	18.20	21.96	20.35	41.59	S_2
	24.31	23.04	27.77	53.26	S_3
	30.14↑	23.46↑	31.18↑	56.74↑	S_1
＋CoRAG	28.77↑	22.35↑	22.31↑	43.29↑	S_2
	31.21↑	23.29↑	32.53↑	60.23↑	S_3
	28.93	16.00	28.32	48.94	S_1
＋LaTRO	25.36	15.84	20.19	41.62	S_2
	28.31	18.17	27.96	50.74	S_3
	30.07	21.45	29.24	53.68	S_1
＋RAP	27.08	15.43	21.41	42.82	S_2
	30.96	20.48	30.89	56.94	S_3
	25.76	18.48	26.19	48.71	S_1
＋RE2	22.78	16.52	19.68	41.84	S_2
	26.94	21.69	27.34	50.28	S_3

3.7　讨　论

3.7.1　辩论场景

本文通过对LLMs在不同辩论场景中的表现进行对比分析，发现能力不同的LLMs在面对不同类型的辩论场景时表现出明显的差异。

具体来说，在反映社会不平衡的不平等辩论场景S_2中，能力较强的LLMs表现出超强的适应能力和应变能力，类似于利用资源的“特权群体”，这种优势可归因于强大的信息处理能力和战略灵活性，这使得它们能够更高效地驾驭复杂的环境。相反，在竞争环境相对公平的辩论场景S_1和S_2中，能力较弱的LLMs能够利用自身的基本优势，接近人类在低压环境下的表现。

3.7.2　模型适应性

在对不同LLMs应用CoRAG-MAD的结果中发现，其在辩论技能方面的表现均有所提高，且这种方法的有效性因LLMs的基础能力而异。

对于能力较强的LLMs，CoRAG-MAD完善了LLMs本就已经十分成熟的技能，在对辩题的理解、论证策略、构建有理有据论点等方面均有所提升，这表明它们吸收和综合信息的范围更广更深刻。对于能力较弱的LLMs，CoRAG-MAD对它们能力的提升更加突出，该方法使它们的论点更加连贯，逻辑性更强，论证方法更有条理，对所讨论的问题有了更深刻的理解。同样，也使它们对不断变化的论点的反应能力有所提高，这表明它们快速适应和参与辩论的能力有所提高，反映了它们的成长心态和逻辑灵活性。

这些分析表明，CoRAG-MAD可以根据各个LLMs的具体需求进行量身定制，更精确、更有效地提高其在辩论中的表现。

3.7.3　模拟人类智能

本文通过将CoT和RAG技术结合并应用到辩论环境中，展示了LLMs在推理、构建论点、批判性思维等方面的能力。在辩论过程中，它们必须理解和分析对立的观点，并提出有说服力的论点，以支持或反驳特定的立场。

在不同的辩论场景中，能力更强的LLMs能够有效地利用现有数据并设计出巧妙的策略，这与人类在错综复杂的环境中的应变能力和适应能力不谋而合，这种适应性反映了人类在各种充满挑战的环境中茁壮成长的能力。相比之下，对于能力相对较弱的LLMs来说，它们更愿意向其他LLMs寻求合作，形成一种共生关系，在这种关系中，能力较强的LLMs向能力较弱的LLMs提供支持，使它们能够超越自身局限，释放更大潜能。这反映了人类社会中的互动关系，即经验丰富的人指导经验较少的同行或与它们合作，促进共同进步。

值得注意的是，在“创造性”方面，LLMs的表现较为有限。尽管它们在推理、构建论点等方面展现出了不俗能力，但是在需要生成全新、独特的观点时，LLMs的能力遇到了瓶颈，它们更倾向于依赖现有的知识库和语言模式，难以突破这些框架而产生真正意义上的创新思想，因此，需要继续探索新的方法和技术，以进一步提升LLMs在创造性方面的能力。

4　结　论

本文提出了一套辩论框架CoRAG-MAD，将CoT与RAG技术相结合，并运用自动化和人工评估指标，对3种辩论场景下不同能力的LLMs进行了比较分析。研究结果表明，CoRAG-MAD可以有效提高LLMs在各个辩论场景中的多项能力。具体而言，在不平等辩论中，LLMs的逻辑推理得分提升57.56%，创造力得分提升49.77%；在混合辩论中，LLMs的协作能力提升23.36%，整体

性能提升28.20%。此外，本文还通过消融实验和对比实验验证了CoT模块、RAG模块和CoRAG方法在MAD中的有效性。CoRAG-MAD在辩论场景中对LLMs的性能提升效果明显，但仍有改进空间，未来的研究工作将聚焦于拓展其应用领域，进一步挖掘潜力，更加全面地提升LLMs在复杂实际场景中的表现。

参考文献：

[1] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[DB/OL]. (2023-11-24)[2024-08-01]. https://arxiv.org/abs/2303.18223.

[2] YIN S, FU C, ZHAO S, et al. A survey on multimodal large language models[DB/OL]. (2024-11-29)[2024-08-01]. https://arxiv.org/abs/2306.13549.

[3] ZHU Y, WANG X, CHEN J, et al. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities[DB/OL]. (2024-02-22)[2024-08-01]. https://arxiv.org/abs/2305.13168.

[4] PARK J S, O'BRIEN J, CAI C J, et al. Generative agents: Interactive simulacra of human behavior[C]// Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023: 1-22.

[5] JINXIN S, JIABAO Z, YILEI W, et al. Cgmi: Configurable general multi-agent interaction framework [DB/OL]. (2023-08-28) [2024-08-01]. https://arxiv.org/abs/2308.12503.

[6] LI G, HAMMOUD H, ITANI H, et al. Camel: Communicative agents for "mind" exploration of large language model society[J]. Advances in Neural Information Processing Systems, 2023, 36: 51991-52008.

[7] DU Y, LI S, TORRALBA A, et al. Improving factuality and reasoning in language models through multiagent debate[DB/OL]. (2023-03-23)[2024-08-01]. https://arxiv.org/abs/2305.14325.

[8] LIANG T, HE Z, JIAO W, et al. Encouraging divergent thinking in large language models through multi-agent debate[DB/OL]. (2024-07-17)[2024-08-01]. https://arxiv.org/abs/2305.19118.

[9] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[10] ZHOU D, SCHÄRLI N, HOU L, et al. Least-to-most prompting enables complex reasoning in large language models[DB/OL]. (2023-04-16)[2024-08-01]. https://arxiv.org/abs/2205.10625.

[11] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners[J]. Advances in neural Information Processing Systems, 2022, 35: 22199-22213.

[12] WANG X, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models [DB/OL]. (2023-03-07) [2024-08-01]. https://arxiv.org/abs/2203.11171.

[13] WANG X, ZHOU D. Chain-of-thought reasoning without prompting [DB/OL]. (2024-05-23) [2024-08-01]. https://arxiv.org/abs/2402.10200.

[14] YAO S Y, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models [DB/OL]. (2023-12-03)[2024-08-15]. http://arxiv.org/abs/2305.10601v2.

[15] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: Solving elaborate problems with large language models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17682-17690.

[16] MADAAN A, TANDON N, GUPTA P, et al. Self-refine: Iterative refinement with self-feedback [DB/OL]. (2023-05-25)[2024-08-15]. https://arxiv.org/abs/2303.17651v2.

[17] COHEN R, HAMRI M, GEVA M, et al. LM vs LM: Detecting factual errors via cross examination [DB/OL]. (2023-05-22) [2024-08-01]. https://arxiv.org/abs/2305.13281.

[18] Chan C M, Chen W, Su Y, et al. Chateval: Towards better LLM-based evaluators through multi-agent debate [DB/OL]. (2023-08-14) [2024-08-01]. https://arxiv.org/abs/2308.07201.

[19] CHEN J C, SAHA S, BANSAL M. Reconcile: Round-table conference improves reasoning via consensus among diverse llms [DB/OL]. (2024-06-21) [2024-08-01]. https://arxiv.org/abs/2309.13007.

[20] ZHAO R, CHEN H, WANG W, et al. Retrieving multimodal information for augmented generation: A survey [DB/OL]. (2023-12-01) [2024-08-01]. https://arxiv.org/abs/2303.10868.

[21] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[22] GU J, WANG Y, CHO K, et al. Search engine guided neural machine translation [DB/OL]. (2018-03-08) [2024-08-15]. https://arxiv.org/pdf/1705.07267.

[23] KE Z, LIN H, SHAO Y, et al. Continual training of

language models for few-shot learning [DB/OL]. (2022-10-11)[2024-08-01]. https://arxiv.org/abs/2210.05549.

[24] KE Z, SHAO Y, LIN H, et al. Adapting a language model while preserving its general knowledge [DB/OL]. (2023-01-21) [2024-08-01]. https://arxiv.org/abs/2301.08986.

[25] LU S, DUAN N, HAN H, et al. ReACC: A retrieval-augmented code completion framework [DB/OL]. (2023-01-21) [2024-08-01]. https://arxiv.org/abs/2203.07722.

[26] NASHID N, SINTAHA M, MESBAH A. Retrieval-based prompt selection for code-related few-shot learning [C]//2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). IEEE, 2023: 2450-2462.

[27] ZHOU S, ALON U, XU F F, et al. Docprompting: Generating code by retrieving the docs [DB/OL]. (2023-02-18)[2024-08-01]. https://arxiv.org/abs/2207.05987.

[28] BAEK J, AJI A F, SAFFARI A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering [DB/OL]. (2023-06-07)[2024-08-01]. https://arxiv.org/abs/2306.04136.

[29] SIRIWARDHANA S, WEERASEKERA R, WEN E, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering [J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1-17.

[30] ASAI A, WU Z, WANG Y, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection [DB/OL]. (2023-10-17) [2024-08-01]. https://arxiv.org/abs/2310.11511.

[31] TRIVEDI H, BALASUBRAMANIAN N, KHOT T, et al. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions [DB/OL]. (2023-06-23) [2024-08-01]. https://arxiv.org/abs/2212.10509.

[32] RIBEIRO D, WANG S, MA X, et al. Entailment tree explanations via iterative retrieval-generation reasoner [DB/OL]. (2022-07-19)[2024-08-01]. https://arxiv.org/abs/2205.09224.

[33] FENG Z, FENG X, ZHAO D, et al. Retrieval-generation synergy augmented large language models [C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 11661-11665.

[34] DataLearner. 大模型综合能力评测对比表[EB/OL]. [2024-08-09]. https://www.datalearner.com/ai-models/leaderboard/datalearner-llm-leaderboard.

[35] ZHAO X T, WANG K, PENG W. ORCHID: A Chinese debate corpus for target-independent stance detection and argumentative dialogue summarization [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 9358-9375.

[36] CHEN H, FENG Y, LIU Z, et al. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding[DB/OL]. (2024-11-21)[2025-01-14]. https://doi.org/10.48550/arXiv.2411.04282.

[37] HAO S, GU Y, MA H, et al. Reasoning with language model is planning with world model [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 8154-8173.

[38] XU X, TAO C, SHEN T, et al. Re-reading improves reasoning in large language models[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024: 15549-15575.