基于领域信息分解式学习的大语言模型修辞认知增强方法

王雯[1], 于东[1,*], 刘鹏远[1,2]
1.北京语言大学信息科学学院，北京，100083
2.国家语言资源监测与研究平面媒体中心，北京，100083
wangwenblcu@gmail.com, yudong@blcu.edu.cn,liupengyuan@blcu.edu.cn

摘要

中文修辞手法多样且概念差异性大，大语言模型对部分修辞手法的认知存在缺陷。针对该问题，本文研究如何增强大语言模型的修辞认知能力，并探究其与修辞识别性能之间的关系。为此，本文提出了QAKAG框架，此框架首先引入信息分解式学习思想，通过问答形式检测大语言模型的修辞认知缺陷，然后以四种不同的知识组合方式探究最优信息补充机制，实现了大语言模型修辞认知能力的增强。本文构建了多类别中文修辞句数据集MCRSD和修辞知识库MCRKB，并在ChatGPT4等六个大语言模型上开展实验研究，验证了QAKAG框架对增强大语言模型修辞认知能力的有效性以及其各阶段的必要性。结果表明，在QAKAG框架的增强下，六个大语言模型在多类别修辞识别任务上的性能相较直接回答识别问题的平均F1值提高22.1%，优于Zero-shot-CoT、RAG-BaiKe、Few-Shot5提示策略。

关键词： 大语言模型；修辞认知；分解式学习；动态信息补充

Method for Enhancing Rhetorical Cognition of Large Language Models Based on Decomposed Learning of Field Information

Wen Wang[1],Dong Yu[1,*], Pengyuan Liu[1,2]
1.Faculty of Computer Science,Beijing Language and Culture University,Beijing,100083
2.National Language Resources Monitoring and Research Center for Print Media,Beijing,100083
wangwenblcu@gmail.com, yudong@blcu.edu.cn,liupengyuan@blcu.edu.cn

Abstract

Chinese rhetorical devices are diverse and conceptually different, weakening large language models in the cognition of some rhetorical devices. This paper aims to enhance the large language model's rhetorical cognition ability and explore the relationship between ability and performance. Therefore, the paper proposes the QAKAG framework, which firstly applies the idea of information decomposition learning and detects the rhetorical cognitive deficiencies of the large language model by asking it questions, and then explores the optimal information supplementation mechanism out of four knowledge combination methods, which enhances rhetorical cognitive ability. Based on a multi-category Chinese rhetorical sentence dataset (MCRSD) and a multi-category

*为通讯作者
基金项目： 教育部人文社科规划项目（23YJAZH184）；北京语言大学梧桐创新平台(中央高校基本科研业务费)(21PT04)；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果（24YCX114）

rhetorical knowledge base (MCRKB), we carried out experimental studies on six large language models, including ChatGPT4, to prove the effect of QAKAG framework enhancing the rhetorical cognitive ability and the indispensability of its two stages. The results show that, in average, the QAKAG framework improves the performance on multi-category rhetorical recognition tasks by 22.1% measured with F1 value, outperforming the Zero-shot-CoT, RAG-BaiKe, and Few-Shot5 prompting strategies.

Keywords: Large Language Models , Rhetorical Cognition , Decompositional Learning , Dynamic Information Supplement

1 引言

修辞识别是一个复杂且具有挑战性的任务，旨在探究和理解中文文本中的修辞结构和语义信息，从而揭示其中的情感、态度和意图，为语言理解、信息检索和文本生成等任务提供更深层次的支持和指导(刘广辉,2022)。修辞是用于各种文章或应用文在写作时语言表达方法的集合，在文本中呈现出复杂性和多样性，具有形式灵活、不易量化的特点(王荣文(2018)，修辞识别任务的关键在于对不同修辞手法进行深入理解与认知。通过在海量数据上进行大规模预训练，大语言模型（Large Language Model，LLM）具有了强大的知识表征能力和理解推理能力(崔亿萍,2023)。但在特定领域任务中，大语言模型可能缺乏某些对应的专业知识，导致在解决特定领域问题上表现不佳(Zhang et al., 2023)，而且模型的知识幻觉问题也较为突出(Li et al., 2024b)。因此，本文旨在多类别修辞识别任务上探索如何有效增强大语言模型修辞认知能力，并探究大语言模型修辞认知能力和修辞识别能力之间的关系。

修辞认知，指模型对修辞语言概念的理解、分析和运用能力(徐盛桓,2008)。模型的修辞认知水平与修辞识别能力存在密切相关性，如果模型对某种修辞认知水平较高，其修辞识别能力可能也会更强(邱文生,2015)。传统的识别方法不具备修辞认知能力，将修辞识别任务视为语言层面的分类问题，主要依赖表层特征进行分类，采用基于统计的方法识别比喻、排比等高频修辞(朱晓亮et al., 2019; 呼啸et al., 2020)，对修辞知识缺乏深层次的认知。受限于人工标注成本，修辞识别任务的训练数据集往往规模有限且数据分布不均衡，导致模型识别高频修辞的能力高于低频修辞，且由于数据稀缺问题，对通感、顶真等低频修辞缺乏深入的研究。大语言模型也存在类似问题。通过在小规模修辞语料上测试发现，大语言模型对比喻、排比等高频修辞手法的认知能力较强，但对反复、顶真等低频上的认知能力则有所不足，由此推测大语言模型在高低频修辞手法上的理解存在认知级差。图1展示了ChatGPT3.5(Dai et al., 2022)对比喻和顶真两种修辞的识别情况。可以看到，ChatGPT3.5对比喻的识别和理解能力比较出色，但对顶真缺乏基础认知，导致识别效果较差，说明大语言模型的修辞识别能力仍然是基于训练语料统计特征，对低频修辞的认知能力有所欠缺。

图 1: ChatGPT3.5对比喻和顶真修辞的识别理解情况

因此，本文深入探究了大语言模型在不同修辞手法上的认知表现，并在序列求解思想的启发下，提出了一种基于领域信息分解式学习的QAKAG（**Q**uestion **A**nswering with **K**nowledge **A**ugmented **G**eneration）框架来增强大语言模型的修辞认知能力，缩小大语言模型对于不同修辞手法的认知差异。在QAKAG框架中，首先是领域信息认知缺陷检测阶段，该阶段将测试问题分解成系列子问题，以子问题针对性问答的形式探究大语言模型对特定修辞手法

的理解情况。然后是领域信息动态补充阶段，该阶段根据前一阶段对子问题的理解情况，从修辞知识库中为大语言模型动态抽取并补充其基本理解和理解欠缺的修辞知识点。该框架结合了分解式学习的思想和知识增强生成模型的方式，可以提高大语言模型的修辞认知能力。为验证QAKAG框架的有效性，本文构建了多类别中文修辞句数据集MCRSD(**M**ulti-category **C**hinese **R**hetorical **S**entence **D**ataset)和修辞知识库MCRKB(**M**ulti-category **C**hinese **R**hetorical **K**nowledge **B**ase)，并设计了四种不同的知识组合方式进行实验探究。结果表明，QAKAG框架可以有效增强大语言模型修辞认知能力，能够为大语言模型解决复杂认知问题提供参考。

本文的贡献主要包括以下三个方面：第一，提出了QAKAG框架，能够有效增强大语言模型的修辞认知能力，提高其在修辞识别任务上的性能。第二，构建了多类别中文修辞句数据集MCRSD和修辞知识库MCRSB，为验证QAKAG框架的有效性提供了数据基础，也丰富了中文修辞识别研究领域的语言资源。第三，探究了多个大语言模型在不同修辞手法上的认知表现，并对各大语言模型对不同修辞的认知情况进行了详细分析，为大语言模型进一步提升修辞识别能力提供了新的思路和方法。

2 相关工作

修辞识别作为自然语言处理中的重要任务，其目标是理解和分析文本中所使用的修辞手法，如比喻、比拟、排比等，具有广泛的应用场景。早期研究集中在修辞学的理论层面，Mann and Thompson (1988)提出了著名的修辞结构理论（RST），为文本修辞的分析和生成提供了理论框架。在20世纪以后，修辞学的研究逐渐扩展到了更广泛的领域，研究者们开始尝试使用各种自然语言处理技术来自动化识别和分析修辞。传统的修辞识别研究主要依赖于语言规则和特征提取。例如国外的PEG(Page,1966)、IEA(Landauer et al., 2000)、E-rater(Attali and Burstein,2004)系统利用多元线性回归、统计学等方法进行特征工程并构建分类器对学生作文进行自动打分；Tsvetkov et al. (2014)提出一种基于语义特征的分类器进行隐喻识别研究。国内学者刘明杨(2015)提出利用排比标检验的启发式方法识别排比和比喻；李明峰and 贾修一(2018)则是采用多分类器集成学习策略进行中文反语识别等。然而，由于修辞的复杂多样性，基于规则和特征的方法往往面临泛化能力不足的问题。随着深度学习的兴起，基于神经网络的模型开始主导修辞识别的研究。长短时记忆（LSTM）、卷积神经网络（CNN）和递归神经网络（RNN）等被广泛应用于此任务中(付瑞吉et al., 2018; 石昀东,2019; 赵琳玲,2020)，但测试数据集规模小，难以覆盖语言的各种变体和复杂构造，限制了修辞识别的准确性和可扩展性。可以看出，传统修辞识别研究关注重点仍在于如何有效提取不同修辞手法的特征进行分类，缺少从修辞认知的角度探讨模型是否对修辞手法的深层含义具有理解能力。

具有强大语言分析理解能力的ChatGPT等大语言模型的出现，为修辞识别任务提供了新的研究途径。Ziems et al.(2024)在零样本设置下评估大语言模型，发现模型在比喻语言分类上表现出色。李玲玲and 王晓燕(2023)在对传统机器翻译和ChatGPT翻译分析对比中发现大语言模型在隐喻理解层面具备显著优势。蔡越(2024)认为可以利用大语言模型生成各种修辞手法的文本，帮助写作者理解和掌握修辞手法。Li et al.(2024a)提出的StyleChat框架可以从词汇、句法、修辞手法、修辞目的四个方面增强大语言模型风格化对话生成能力，侧面说明了大语言模型在理解和应用修辞方面存在潜力。但目前，关注如何利用大语言模型处理修辞识别任务的相关研究还比较少，缺乏对大语言修辞认识能力和修辞识别性能之间的探究分析。

如何增强大语言模型解决复杂问题的能力是近年来的研究热点。Zhou et al.(2022)提出了least-to-most prompting以解决复杂推理问题，主要思想是将一个复杂的问题分解为一系列更简单的子问题依次解决，然后用先前解决的子问题的答案来促进当前子问题的解决。Trivedi et al.(2022)在多步骤QA任务上，用CoT指导检索，再反过来用检索的结果来改进CoT。Liu et al.(2024)提出的RA-ISF框架中的问题分解模块则是可将复杂问题拆解为子问题并迭代处理，这种自反馈机制可以改善大语言模型在复杂任务中的表现。因此，本文将在现有主流的六个大语言模型上针对比喻、比拟、通感、排比、对偶、反复、设问、夸张、反问，顶真十种常用修辞开展认知能力测试，并基于序列求解的思想针对其中的不足提出解决方案，旨在有效提高大语言模型的修辞认知水平，进而提升其在修辞识别任务上的性能。

3 大模型修辞认知能力评测数据集建设

3.1 多类别中文修辞句数据集MCRSD构建

修辞数据集是修辞识别任务的基础。目前公开可用的中文修辞数据集有CMC中文比喻语料库(Li et al., 2022a)、C-SEM语义评测基准基础修饰知识检测（SLRFC）子评测项数据集[0]和CCL 2024发布的“中小学作文修辞识别与理解评测”任务数据集[1]。其余未公开数据集多以语文作文、古代诗歌、高考试卷、中文微博数据集等作为语料来源。现有的数据集语料规模小，修辞类别不均衡，限制了修辞识别任务的研究，因此建立更全面和更大规模的修辞数据集对于推动修辞识别任务的发展具有重要意义。中文修辞手法多样且概念差异大，为建立规模合理、类别平衡且修辞覆盖性广泛的修辞数据集，探究大语言模型对多种常见修辞手法的认知能力，本文采用人工标注的方式构建了一个多类别中文修辞句数据集MCRSD，构建流程如图2所示。未来，可以通过人工标注或者机器辅助生成的方式对数据集进行扩充，进一步提高该数据集的应用价值。

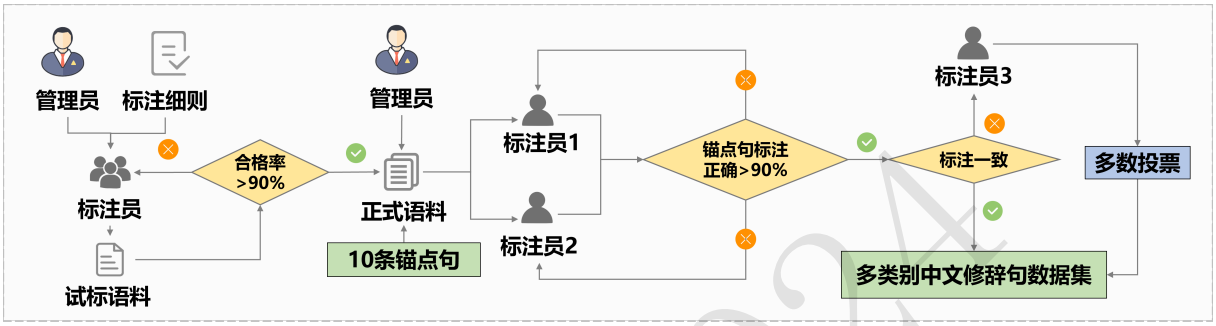

图 2: 多类别中文修辞句数据集构建流程

MCRSD数据集构建以北京语言大学BCC现代汉语语料库(荀恩东et al., 2016)、北京大学CCL语料库(詹卫东et al., 2019)和Li et al.(2022b)构建的CLGC文采优美段语料库作为基础，以句为单位进行切分作为待标数据。本次标注招募三名语言学专业的硕士研究生作为标注员，标注过程分为标注培训、预标注、正式标注、核对检验四个环节。在标注培训环节，标注员需要认真学习十种修辞的相关理论，明确标注任务和标注细则，标注过程中如果待标句子不具有任何修辞手法则标注为“无修辞”。培训完成后，每名标注员分配100条试标语料进行试标注，标注准确率达到90%方为合格。在正式标注阶段，管理员每天发放500条标注语料，两名标注员进行独立标注。为确保标注质量，在正式语料中投放10条锚点句，管理员对每日的标注数据进行整理检查，锚点句标注正确率达到90%以上且两名标注员标注一致的数据则纳入数据集。标注不一致的数据则发放给第三名标注员进行标注，采用多数投票原则选出句子最终标签，若无法选出则删除该条语料。持续以上标注过程，直至十种修辞和“无修辞”标签均具有1000条标注句子。最终，通过标注收集到了11000条标签句，MCRSD数据集的情况如表1所示。

类别	数量	平均词类符数	平均句长	平均实词数	平均虚词数	示例
比喻	1000	20.35	34.96	16.50	6.46	他沉默着，像一个木偶似的站立在林青史的面前。
比拟	1000	16.09	26.96	12.16	5.37	这会儿，你看，小草含着泪珠儿，在泣在愁。
排比	1000	22.17	48.51	25.75	7.62	夏天的夜晚，是那么的宁静，那么的美丽，那么凉爽。
夸张	1000	16.79	25.67	12.72	5.11	一个芝麻大的官放个屁，到了你们这儿也是8级地震！
反问	1000	14.50	22.92	10.76	4.60	下回有事谁还敢找你呀！
设问	1000	16.72	30.36	14.36	5.52	是谁创造了人类世界？是我们劳动群众。
对偶	1000	7.73	13.70	7.24	1.11	无尽波涛归学海，长春花木在词林。
反复	1000	12.92	23.15	12.16	5.10	冒着敌人的炮火，前进！前进！前进！
通感	1000	16.03	26.55	12.30	5.11	话语消融在拂过的轻风和升起的湿润香气之中。
顶真	1000	10.00	17.63	9.64	2.32	味摩诘之诗，诗中有画；观摩诘之画，画中有诗。
无修辞	1000	15.13	24.34	11.88	4.64	是这样，我回去替你想想办法吧，明天再告诉您，好吗?

表 1: MCRSD数据集概况

[0]https://github.com/FlagOpen/FlagEval
[1]https://github.com/cubenlp/CERRU

表1展示了MCRSD数据集中每种标签类别的数量、平均词类符数、平均句长、平均实词数、平均虚词数和示例。通过分析这些统计数据，可以看出MCRSD数据集中不同类别的修辞句子在语言使用上的特点与差异。例如，排比句的平均词类符数（22.17）、平均句长（48.51）和平均实词数（25.75）相对较高，表明了排比修辞的句子结构相对其他修辞句更加复杂多样。而对偶句的各项指标数值都相对较低，体现了其简洁对仗的特点。比喻、比拟、夸张、设问和通感修辞句的平均词类符数都相对较高，分别为20.35、16.09、16.79、16.72和16.03，说明这些修辞手法倾向于使用复杂语言结构，增强了描述的生动性和多样性。

3.2 大模型修辞认知评测集组成与测试

针对MCRSD数据集中每条修辞标签句，随机从“无修辞”标签句中抽取3条句子与其组合成选择题，题目是“请从以下四个选项中选出具有XX修辞手法的句子”，候选答案包括ABCD四个选项，其中包括1条正确修辞标签句和3条“无修辞”标签句，组建形成了包含10000个修辞问题的评测集。评测集基本统计信息如附录A表7所示。选取ChatGPT4(gpt-4-1106-preview)(Achiam et al., 2023)、ChatGPT3.5(gpt-3.5-turbo)、文心一言(ERNIE-Bot-turbo-0922)[2]、Baichuan2-7B-Chat[3]、ChatGLM3-6B[4]、Qwen-7B-Chat[5]六种不同的大语言模型进行初步测试，使用准确率指标来评估大语言模型对修辞手法的认知能力。

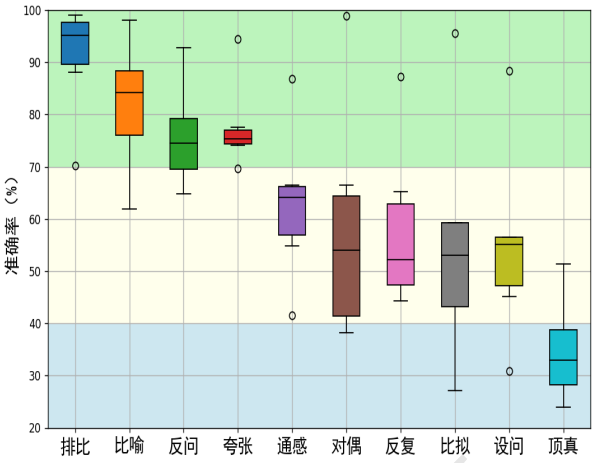

图 3: 不同模型在十种修辞上的平均识别效果

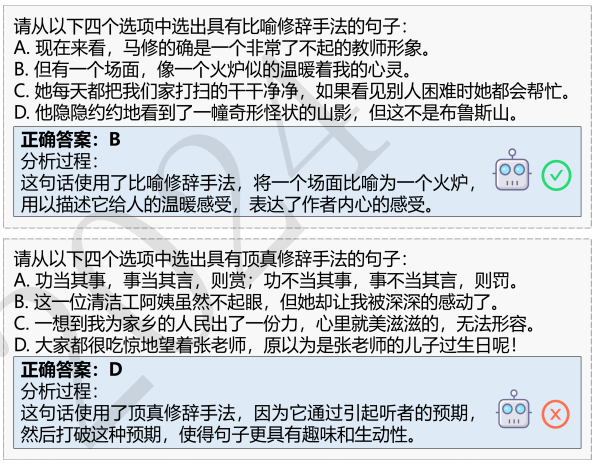

图 4: ChatGPT3.5评测示例

图3展示了六个大语言模型在十种修辞上的平均识别效果。可以看出，大语言模型对十种修辞的理解存在认知级差。对于排比、比喻等修辞，大语言模型呈现较高的认知水平，其识别准确率普遍达到70%以上；对于通感、对偶等修辞，大语言模型的修辞认知能力一般，识别准确率在70%至40%之间；而对于顶真修辞，大语言模型的认知水平相对较低，其识别准确率仅在40%以下。图4展示了ChatGPT3.5在测试过程中回答正确和回答错误的示例。可以发现，对于比喻，模型能够很好地分析出比喻的本体、喻体及语言效果；但对于顶真，模型不能准确理解该修辞手法的定义或特征，可能的原因是在模型的训练数据中某些修辞出现频率较低，导致模型受到数据偏差的影响，在这些修辞上的识别表现相对较差。这也说明了大语言模型的修辞认知能力和其修辞识别性能存在密切相关性，如果大语言模型对某种修辞具有较高的认知能力，那大语言模型对应的修辞识别性能也会更强。

3.3 多类别中文修辞知识库MCRKB构建

为确保后续为大语言模型补充领域信息的规范性和针对性，有效增强语言模型的修辞认知能力，本文参考陈望道(2008)、骆小所(2010)和谭汝为(2023)等人提出的修辞理论并结合多个知识来源构建了包括比喻、比拟、排比、夸张、设问、反问、对偶、反复、通感、顶真十种修辞手法的修辞知识库MCRKB，其中每类修辞的领域信息划分为定义、特征、变体特征、分析方

[2] https://yiyan.baidu.com/welcome
[3] https://www.baichuan-ai.com/home
[4] https://chatglm.cn/main/detail
[5] https://tongyi.aliyun.com/qianwen/

法、易混淆点、作用六个方面。MCRKB修辞知识库的基本统计信息如附录A表8所示，其中比喻修辞的知识点示例如附录A表9所示。MCRKB修辞知识库构建完成后，我们邀请两名语言学专业的硕士研究生进行人工审核和校对，确保知识库中信息的准确性、可靠性和权威性。图5展示了从MCRKB修辞知识库中提取相关修辞领域信息，增强大语言模型修辞认知能力的方式。对比图4中的模型回答错误示例发现，模型的修辞识别效果有所增强，验证了MCRKB修辞知识库的可用性。

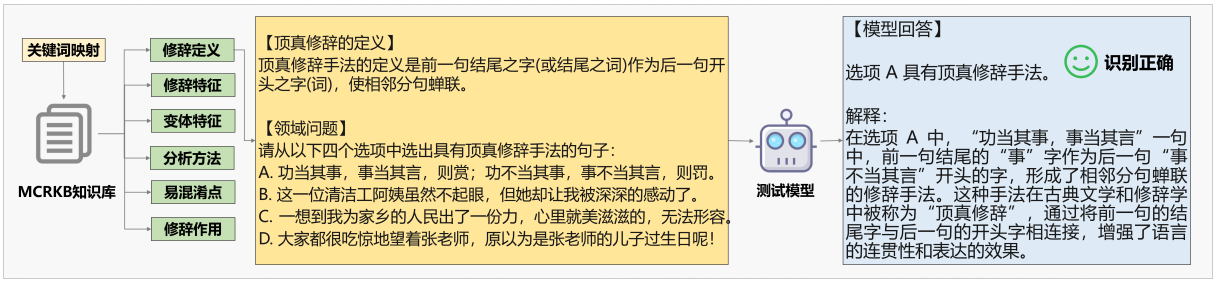

图 5: MCRKB修辞知识库为大语言模型补充领域信息的方式

4 方法

4.1 QAKAG整体框架

受限于数据集质量和模型性能，传统识别方法并不具备对修辞的理解认知能力。大语言模型通过在海量文本数据上的训练，具有较强的语言分析和语义理解能力，能够为解决复杂认知问题提供了新的可能性。在这样的背景下，本文提出了QAKAG框架，致力于系统化地识别并弥补大语言模型在特定领域知识理解方面存在的不足，通过领域信息认知缺陷检测与动态补充信息机制，增强大语言模型对领域知识的认知水平。整体框架如图6所示。

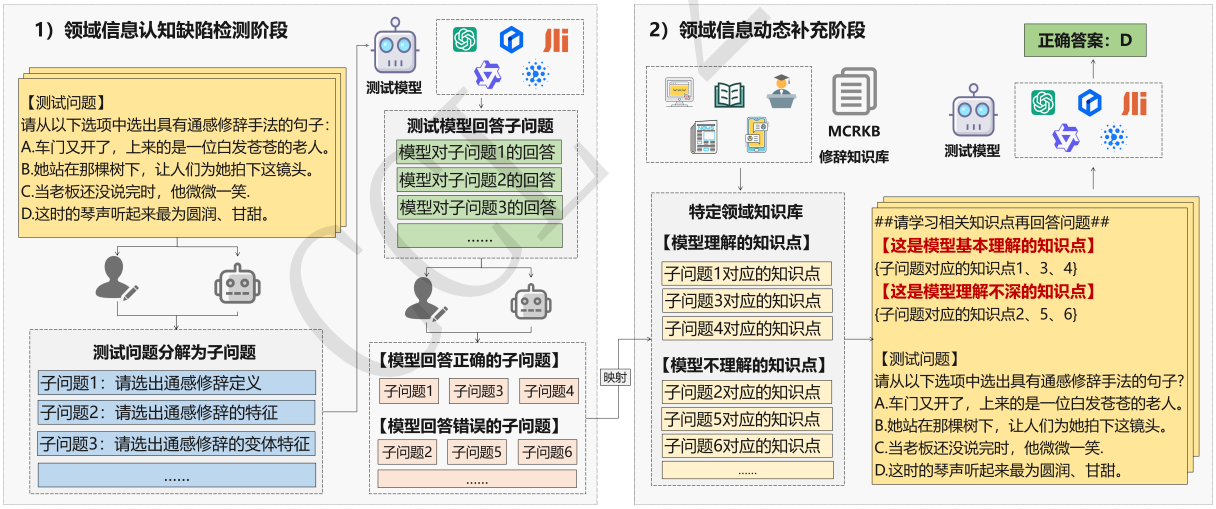

图 6: QAKAG整体框架

QAKAG框架分为：领域信息认知缺陷检测阶段和领域信息动态补充阶段。在领域信息认知缺陷检测阶段，将输入的测试问题分解为一系列子问题，判断大语言模型对每一个子问题的回答，获取大语言模型在领域知识上的理解情况。该阶段可以揭示大语言模型对哪些知识点已经基本了解，对哪些知识点需要加强学习，为后续的信息补充提供目标指向。然后，在领域信息动态补充阶段中，基于第一阶段获得的大语言模型对知识点理解情况，对大语言模型基本理解和理解不足的领域知识进行针对性的信息补充。本文所提出的QAKAG框架不仅可以强化大语言模型对已掌握知识点的理解，而且能够补充大语言模型在特定领域的知识缺陷。在QAKAG框架的增强下，大语言模型能够对不同修辞手法有更加深入的认知理解。

4.2 领域信息认知缺陷检测阶段

领域信息认知缺陷检测阶段的核心目的是评估大语言模型对子问题的认知能力。首先，通过“人工+模型”的方法，将测试问题分解为若干个具有高度针对性的子问题，并为子问题设置一项正确答案和三项错误答案。具体来说，本文通过设计子问题分解提示模板，使用ChatGPT4模型将测试问题从[定义、特征、变体特征、分析方法、易混淆点、作用]六个层面进行分解。由于大语言模型对修辞的理解深度不够，分解出来的子问题候选答案可能存在不准确的情况，因此在这一步加入人工校对步骤，对大语言模型分解出来的子问题进行审核。子问题若分解准确则进入下一步，子问题若有问题，则进入编辑模式修正该子问题，然后再进入下一步。以上阶段每类修辞仅需执行一次，审核通过的子问题将自动保存到子问题库中，后续若遇到相同类型的修辞问题，则自动从子问题库中进行抽取对应的子问题。这些子问题也分别对应MCRKB修辞知识库中六个方面的领域信息。测试模型被要求逐一回答这些子问题，以便判定其是否掌握回答测试问题所需的领域知识。子问题库形式如表2所示，每个子问题包括一项正确答案和三项错误答案，答案列是子问题的正确答案。

子问题	答案
请从以下选项中选出“顶真”修辞手法的定义: A.把物当作人写，或把人当作物写，或把甲物当作乙物来写。 B.为了加强语气，用疑问的形式表示确定的意思，无疑而问，明知故问。 C.把三个或三个以上结构相同或相似的句子或句子成分排列起来。 D.前一句结尾之字(或结尾之词)作为后一句开头之字(词)，使相邻分句蝉联。	D

表 2: 子问题示例

测试模型对子问题的回答是否正确将被用作衡量其领域知识理解度的指标。如果测试模型能够正确回答子问题，可以推断测试模型已经掌握了该子问题对应的知识点，具备利用该知识点进行有效推理和应用的能力；反之，如果测试模型未能正确回答子问题，则推断测试模型尚未理解相关知识点或存在理解偏差，可能会造成测试模型在后续推理过程中的逻辑错误。通过这种分解和评估机制，我们能够检测出测试模型对解答测试问题所需领域知识的掌握水平，为进一步优化测试模型在领域知识上的理解和应用提供明确方向。

4.3 领域信息动态补充阶段

在给大语言模型输入领域信息时，不同的知识组合方式可能会对其输出产生不同的影响。本文设计了四种不同的知识组合方式，分别是全知直授法、断点补教法、诊断增强法和辩知强化法，如图7所示。对评测集中第n个问题，通过领域信息认知缺陷检测阶段分解为m个子问题。根据模型对子问题的回答判断其对知识点的理解情况，如果子问题回答正确则认为模型基本理解该知识点，反之则推断模型对该知识点理解不足。将子问题回答情况映射到知识库中，可以将K个相关知识点划分为$K_{i(\text{true})}$基本理解和$K_{i(\text{false})}$理解不足两种类别。

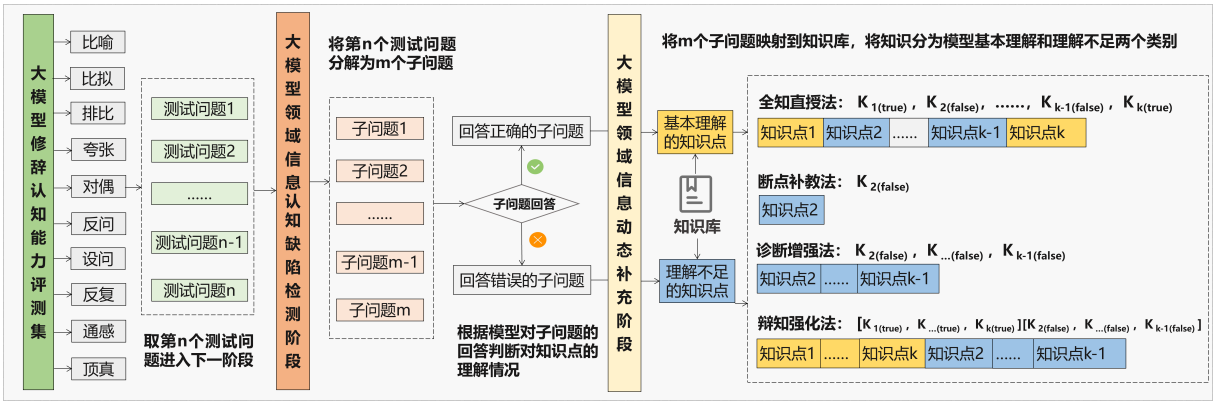

图 7: 在领域信息动态补充阶段设计四种知识组合方式

全知直授法，指不区分大语言模型对知识点的理解情况，一次性向大语言模型补充相关知识点。断点补教法，指仅补充大语言模型回答错误的第一个子问题所对应的知识点。诊断增强法，指补充大语言模型回答错误的所有子问题所对应的知识点。辩知强化法，指根据大语言模型对知识点的理解情况，将基本理解和理解不足的知识点分别进行组合，然后再与测试问题结合输入给大语言模型。通过尝试不同的知识组合方式，让大语言模型进行知识的理解和补全，有助于确定出最佳的知识组合方式，更好地提高模型的认知能力。

为确保新补充的知识能够被模型有效理解，本文利用以上设计的四种知识组合方式进行实验，最终选定辩知强化法作为QAKAG框架领域信息动态补充阶段的知识组合方式，补充形式如：$[K_{1(\text{true})}, K_{...(\text{true})}, K_{k(\text{true})}][K_{2(\text{false})}, K_{...(\text{false})}, K_{k-1(\text{false})}][n]$。这种方式将引导模型对不同知识点的关注方向，强化对基本理解知识点的掌握，并提高对理解不足知识点的认知。

领域信息动态补充阶段的目标在于设计有效的信息补充机制。根据认知缺陷检测阶段中模型对子问题的回答情况，以关键词映射的方式从知识库中提取出子问题对应的知识点，以此优化模型的知识结构并扩展其知识边界，帮助模型达到更高层次的认知和分析水平。关键词映射方式，即测试问题在分解为系列子问题后，子问题的题目形如“请从以下四个选项中选出比喻修辞的定义”。首先，检测该子问题的题目部分中是要求回答[比喻，比拟，夸张，排比，反复，反问，设问，对偶，通感，顶真]中的哪一修辞类别，检测出来的类别即作为该测试问题的关键词1。然后，继续检测该子问题的题目部分中要求回答[定义，特征，变体特征，分析方法，易混淆点，作用]中的哪一知识点类型，检测出来的类型即作为该测试问题的关键词2。最后，以关键1和关键2映射到MCRKB知识库中，即可提取到子问题对应的知识点。

5 实验与结果

5.1 实验设计

实验环节选用的测试模型包括ChatGPT4(gpt-4-1106-preview)、ChatGPT3.5(gpt-3.5-turbo)、文心一言(ERNIE-Bot-turbo-0922)、Baichuan2-7B-Chat、ChatGLM3-6B、Qwen-7B-Chat六个大语言模型。其中ChatGPT4、ChatGPT3.5、文心一言通过API调用的方式进行测试，Baichuan2-7B-Chat、ChatGLM3-6B、Qwen-7B-Chat则是从https://huggingface.co/网站下载开源代码，部署到本地服务器上进行测试。实验数据是包含10000个修辞问题的评测集。实验采用精确率（P）、召回率（R）和F1值作为评价指标。

大语言模型修辞认知能力探究 探究大语言模型在未经任何特定训练或提示优化的条件下在修辞问题上的认知表现。在不附加任何形式的指导信息下，将修辞识别问题直接输入大语言模型，探究大语言模型在原始状态下处理修辞识别问题的表现，为后续的优化提供基准，提示形式如下：[原始测试问题]→模型输出。

提示策略增强大语言模型修辞认知能力探究 选取提示工程策略中最具有代表性的零样本思维链（Zero-shot-CoT）、检索增强生成（RAG）、少样本学习（FSL）方法来评估大语言模型修辞认知能力增强效果。在大语言模型解决修辞识别问题时，Zero-shot-CoT方法是在测试问题结尾附加“让我们一步步思考(let’s think step by step)”，提示形式如下：[原始测试问题][Zero-shot-CoT]→模型输出。RAG方法是以关键词检索的方式，从百度百科[6]数据源中检索信息辅助大语言模型生成答案，提示形式如下：[百度百科信息提示][原始测试问题]→模型输出。FSL方法是以关键词映射的方式，从自建的修辞问题样本库中抽取5个示例样本融入提示中，增强大语言模型快速学习能力，提示形式如下：[示例1-5] [原始测试问题]→模型输出。

QAKAG框架增强大语言模型修辞认知能力探究 采用本文提出的QAKAG框架，在六个大语言模型上进行测试。通过QAKAG框架，探究大语言模型对测试问题分解出的系列子问题的理解情况，并将映射得到的领域信息组合后与原始测试问题合并输入大语言模型，以此验证QAKAG框架对大语言模型修辞认知能力的增强效果，该实验方法即为Ours。同时，采用全知直授法、断点补教法、诊断增强法进行对比实验。

5.2 实验结果及分析

ChatGPT等大语言模型在文本分类任务中具有巨大的潜力，可以有效解决流派识别、情感分析、立场检测等问题(Liu et al., 2023)，但尚未有研究工作关注如何利用大语言模型处理多类别中文修辞认知问题。因此，本文利用组建的评测集深入探索了不同大语言模型在不同修辞

[6] https://baike.baidu.com/

手法上的认知能力。表3展示了不同方法下大语言模型认知结果。从表中可以看出，对比其他方法，六个大语言模型均在Ours上取得了最优的性能，表明本文所提出的QAKAG框架对增强大语言模型的修辞认知能力具有更好的效果。其中，ChatGPT4在不同方法下都取得了最优异的表现。在直接回答的情况下，ChatGPT4的F1值就达到了87.5%，且在QAKAG框架的增强下其F1值提高到了95.8%。这说明ChatGPT4本身具有比较高的基础性能，在QAKAG框架下，能够对修辞进行更加全面、深入的认知理解。

分析方法	ChatGPT4	ChatGPT3.5	文心一言	Baichuan2-7B-Chat	ChatGLM3-6B	Qwen-7B-Chat
	P/R/F1	P/R/F1	P/R/F1	P/R/F1	P/R/F1	P/R/F1
直接回答	.877/.875/.875	.645/.612/.610	.621/.612/.698	.618/.532/.552	.600/.589/.591	.725/.686/.698
Zero-shot-CoT	.894/.867/.879	.672/.609/.637	.583/.389/.361	.636/.578/.592	.684/.639/.658	.774/.727/.737
RAG-BaiKe	.920/.912/.915	.753/.694/.722	.628/.468/.460	.705/.370/.334	.697/.666/.664	.628/.538/.564
Few-Shot5	.913/.912/.911	.869/.868/.868	.712/.332/.256	.723/.575/.566	.815/.796/.797	.784/.778/.777
全知直授法	.958/.956/.957	.874/.870/.870	.682/.610/.620	.672/.621/.614	.709/.681/.680	.725/.717/.717
断点补教法	.932/.929/.930	.835/.820/.818	.658/.575/.597	.714/.631/.620	.748/.687/.709	.801/.784/.788
诊断增强法	.957/.955/.956	.885/.874/.874	.660/.590/.602	.741/.630/.630	.709/.679/.685	.718/.696/.702
Ours	**.960/.957/.958**	**.888/.885/.885**	**.747/.743/.745**	**.737/.649/.646**	**.819/.801/.800**	**.844/.829/.829**

表 3: 不同方法下大语言模型在修辞识别任务上的表现

从表3的实验结果中，对比提示策略可以发现，六个大语言模型在修辞识别任务上的整体认知表现：Few-Shot5>Zero-shot-CoT>RAG-BaiKe。相较直接回答，在Few-Shot5和Zero-shot-CoT提示策略下，除文心一言的F1值有所下降外，其余大语言模型的整体认知表现均有一定程度的提升，这可能是文心一言的学习推理能力稍有不足。而RAG-BaiKe提示策略对各大语言模型的增强效果不佳，则可能是因为从百度百科返回的领域信息长且冗余，缺少针对性和规范性，存在噪音问题，导致模型性能下降。对于不同的知识组合方式，对比全知直授法、断点补教法、诊断增强法三种不同的知识组合方式，采用辩知强化法的Ours可以帮助模型具有更好的识别性能，表明在解决复杂认知问题时，将知识点按照大语言模型基本理解和理解不足进行区分后再进行补充，有利于模型对知识点进行针对性学习，从而有更深入的理解。

从图8中可以看出，ChatGPT3.5在QAKAG框架下的F1值相对于直接回答增幅为45.1%，提升幅度是六个模型里面最显著的，说明ChatGPT3.5的训练数据中可能存在修辞知识不足或不平衡问题，但模型对上下文信息的敏感度和学习性能很高，QAKAG框架可以帮助模型更好地掌握并应用新的知识信息。值得注意的是，ChatGLM3-6B的增幅为35.4%，仅次于ChatGPT3.5，表明ChatGLM3-6B具有较强的学习推理能力，能快速分析领域信息并完成自我增强，在处理认知问题上具有很大的潜力。其余四个模型也都在QAKAG框架下有了不同幅度的提升，且对比其他提示策略和知识组合方式均达到了最优，验证了QAKAG框架的有效性。图9展示了ChatGPT3.5在采用QAKAG框架前后的回答样例，可以看出ChatGPT3.5通过QAKAG框架的增强，可以很好地学习到顶真修辞的相关知识，从而正确回答测试问题。

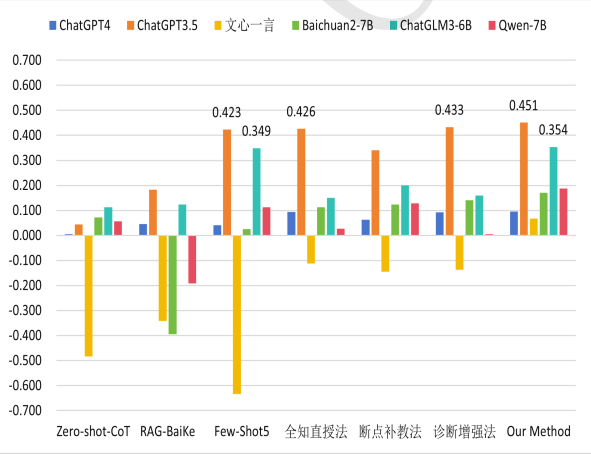

图 8: 各方法相对于直接回答的变化幅度对比

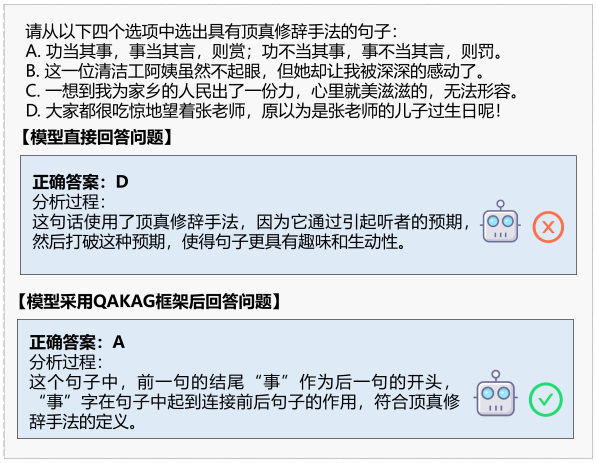

图 9: 模型应用QAKAG框架前后回答对比

5.3 探究实验分析

(1) QAKAG框架各组成部分对大语言模型修辞认知能力的影响

为深入探究QAKAG框架各个组成部分对实验结果的作用和影响，本文设计了三个消融实验进行验证。在第一个消融实验中，我们消除了领域信息动态补充阶段，即将大语言模型对子问题的回答情况补充在输入提示中辅助大语言模型生成答案，提示形式如下：[子问题][模型回答正误][子问题正确答案][原始测试问题]→模型输出，从而验证领域信息动态补充阶段的有效性。在第二个消融实验中，我们消除了领域信息缺陷检测阶段，即不通过子问题分解检测，直接以关键词映射的形式从知识库中将问题对应的知识点补充给大语言模型，提示形式如下：[测试问题相关知识点][原始测试问题]→模型输出，从而验证领域信息缺陷检测阶段的有效性。在第三个消融实验中，为验证QAKAG框架中MCRKB修辞知识库的有效性，我们将MCRKB修辞知识库分别替换为ChatGPT4知识库和百度百科知识库，比较不同来源的知识库对框架效果的影响。ChatGPT4知识库，即将第一阶段分解出的子问题输入给ChatGPT4，将模型生成的答案作为补充的领域信息来源；百度百科知识库，则是利用第一阶段分解出的子问题，搜索百科百科相关条目作为补充的领域信息来源，若百度百科没有对应的条目，则该子问题则跳过，不补充相关知识。表4展示了不同的消融实验下大语言模型在修辞识别任务上的认知表现结果。

分析方法	ChatGPT4	ChatGPT3.5	文心一言	Baichuan2-7B-Chat	ChatGLM3-6B	Qwen-7B-Chat
	P/R/F1	P/R/F1	P/R/F1	P/R/F1	P/R/F1	P/R/F1
Ours	**.960/.957/.958**	**.888/.885/.885**	**.747/.743/.745**	.737/**.649/.646**	**.819/.801/.800**	**.844/.829/.829**
消融领域信息动态补充	.943/.939/.940	.885/.880/.881	.690/.483/.477	.724/.593/.582	.797/.768/.769	.814/.807/.808
消融领域信息缺陷检测	.958/.956/.957	.874/.870/.870	.682/.610/.620	.672/.621/.614	.709/.681/.680	.725/.717/.717
替换MCRKB知识库-ChatGPT4	.902/.900/.900	.745/.741/.741	.697/.505/.511	.637/.464/.440	.705/.687/.689	.686/.641/.650
替换MCRKB知识库-BaiKe	.911/.910/.909	.871/.863/.864	.719/.715/.716	**.759**/.497/.476	.770/.742/.738	.808/.791/.791

表 4: 各消融实验下大语言模型在修辞识别任务上的表现

可以发现，对比QAKAG框架的增强效果，如果消融领域信息动态补充阶段，六个大语言模型在修辞识别性能的平均F1值下降了6.77%，表明模型在某些修辞手法的理解上缺乏相关领域信息，对不同修辞手法存在认知不全或不平衡问题。如果消融领域信息缺陷检测，模型修辞识别性能的平均F1值下降了6.75%，这可能是大语言模型在学习相关知识时缺乏针对性，出现了一定程度的信息损失。如果将MCRKB知识库替换为ChatGPT4知识库，模型修辞识别性能的平均F1值则下降了15.60%，说明ChatGPT4对修辞知识的理解和掌握还不足够准确；如果将MCRKB知识库替换为百度百科知识库，模型修辞识别性能的平均F1值则下降了6.15%，说明百度百科对修辞知识的内容组织、质量或者覆盖范围上与MCRKB知识库存在一定的差距。故无论消融或替换任一个模块，各大语言模型的修辞识别能力均有所下降，验证了QAKAG框架中两个阶段的有效性和人工构建MCRKB修辞知识库的合理性。

(2) QAKAG框架增强大语言模型对不同修辞手法认知能力的影响

大语言模型对不同修辞手法的理解程度存在差异。表5和表6分别展示了Baichuan2-7B-Chat和文心一言在不同方法下对十种修辞手法的识别性能，括号内的数字代表相对于直接回答，模型性能的变化情况。可以看到，除了Ours外，其余方法下Baichuan2-7B-Chat和文心一言的修辞识别性能均有不同程度的提升和下降。附录A中的表10、11、12、13也展示了其他四个大语言模型的表现情况。综合所有数据发现，在QAKAG框架的增强下，六个大语言模型在十种修辞手法上的识别性能均有提升，表明了QAKAG框架的鲁棒性和普适性。

修辞类别	Zero-shot-CoT	RAG-BaiKe	Few-Shot5	全知直授法	断点补教法	诊断增强法	Ours
排比	.800(.098↑)	.763(.061↑)	.772(.070↑)	.887(.195↑)	.909(.217↑)	.673(.019↓)	.790(.098↑)
比喻	.759(.022↑)	.394(.343↓)	.717(.020↓)	.765(.028↑)	.761(.024↑)	.791(.054↑)	.787(.050↑)
夸张	.693(.064↓)	.499(.258↓)	.691(.066↓)	.626(.131↓)	.679(.078↓)	.696(.061↓)	.772(.015↑)
反问	.769(.018↑)	.513(.238↓)	.702(.049↓)	.802(.051↑)	.682(.069↓)	.740(.011↓)	.772(.021↑)
比拟	.499(.029↑)	.393(.077↓)	.673(.203↑)	.435(.035↓)	.633(.163↑)	.638(.168↑)	.653(.183↑)
对偶	.500(.118↑)	.261(.121↓)	.586(.204↑)	.631(.249↑)	.498(.116↑)	.673(.291↑)	.754(.372↑)
通感	.586(.171↑)	.634(.219↑)	.508(.093↑)	.559(.144↑)	.685(.270↑)	.673(.258↑)	.605(.190↑)
反复	.463(.007↓)	.345(.125↓)	.421(.049↓)	.513(.043↑)	.486(.016↑)	.603(.133↑)	.478(.008↑)
设问	.414(.105↑)	.477(.168↑)	.411(.102↑)	.431(.122↑)	.546(.237↑)	.400(.091↑)	.466(.157↑)
顶真	.296(.032↓)	.358(.030↑)	.273(.055↓)	.362(.034↑)	.433(.105↑)	.410(.082↑)	.408(.080↑)

表 5: Baichuan2-7B-Chat在不同方法下对十种修辞手法的认知表现

	Zero-shot-CoT	RAG-BaiKe	Few-Shot5	全知直授法	断点补教法	诊断增强法	Ours
排比	.856(.025↓)	.830(.051↓)	.809(.072↓)	.856(.025↓)	.830(.051↓)	.809(.072↓)	.953(.072↑)
比喻	.655(.036↑)	.554(.065↓)	.636(.017↑)	.655(.036↑)	.554(.065↓)	.636(.017↑)	.858(.239↑)
夸张	.757(.060↑)	.703(.006↑)	.718(.021↑)	.757(.060↑)	.703(.006↑)	.718(.021↑)	.753(.056↑)
反问	.676(.005↓)	.683(.002↑)	.680(.001↓)	.676(.005↓)	.683(.002↑)	.680(.001↓)	.852(.171↑)
比拟	.640(.047↑)	.537(.056↓)	.601(.008↑)	.640(.047↑)	.537(.056↓)	.601(.008↑)	.668(.075↑)
对偶	.328(.252↓)	.364(.216↓)	.402(.178↓)	.328(.252↓)	.364(.216↓)	.402(.178↓)	.694(.114↑)
通感	.626(.078↑)	.589(.041↑)	.545(.003↓)	.626(.078↑)	.589(.041↑)	.545(.003↓)	.636(.088↑)
反复	.445(.002↑)	.406(.037↓)	.438(.005↓)	.445(.002↑)	.406(.037↓)	.438(.005↓)	.747(.304↑)
设问	.612(.048↑)	.600(.036↑)	.620(.056↑)	.612(.048↑)	.600(.036↑)	.620(.056↑)	.732(.168↑)
顶真	.501(.013↓)	.483(.031↓)	.450(.064↓)	.501(.013↓)	.483(.031↓)	.450(.064↓)	.541(.027↑)

表 6: 文心一言(ERNIE-Bot-turbo-0922)在不同方法下对十种修辞手法的认知表现

图10展示了QAKAG框架对大语言模型在不同修辞手法认知上的增强效果。可以看出，测试的大语言模型对比喻、对偶等直观形象的修辞手法认知提升较为明显，而对于通感、反复等更依赖语境和逻辑关系的修辞手法的认知虽有所提升，但幅度略低。这可能是因为评测集采用了以句为单位的选择题形式，而对于逻辑丰富的修辞手法，大语言模型需要更多的上下文语境来增强认知。因此，对于大语言模型来说，要在这种片段化的评测中完全把握修辞手法的细微差别可能会更加困难。此外，实验结果也表明，通过QAKAG框架增强大语言模型修辞认知能力，大语言模型对不同修辞手法的理解和应用能力也会相应增强，其修辞识别性能也会同步提升，验证了大语言模型修辞认知能力和修辞识别性能之间的正相关性。

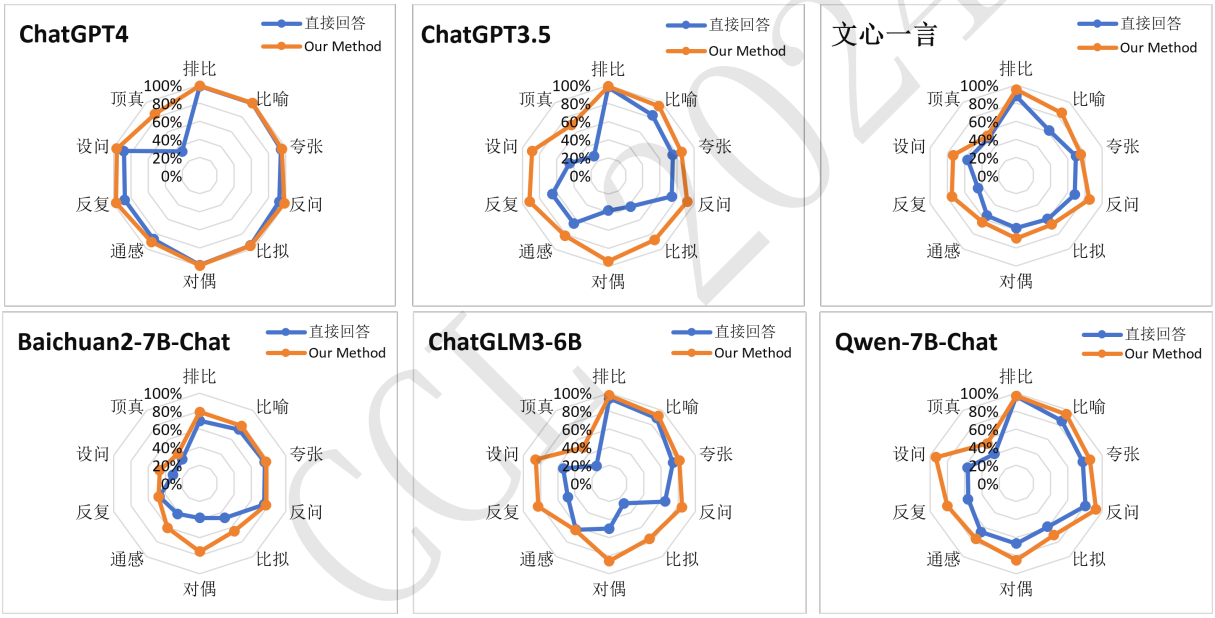

图 10: QAKAG框架对大语言模型在不同修辞手法识别能力增强情况

6 结论

针对大语言模型对不同修辞手法存在的认知级差问题，本文提出了QAKAG框架，提高大语言模型对不同修辞手法的认知能力，缩小其在不同修辞手法上呈现的认知差异，提升其修辞识别性能。为验证QAKAG框架的合理性，本文构建了多类别中文修辞句数据集MCRSD和修辞知识库MCRKB，并设计了四种不同的知识组合方式进行实验对比。实验证明，相比于直接回答和Zero-shot-CoT、RAG-BaiKe、Few-Shot5提示策略，QAKAG框架对大语言模型的修辞认知增强效果最优，而且各大语言模型对不同修辞手法的认知能力也均有不同程度的提升，表明QAKAG框架有效性的同时，也说明了大语言模型在处理复杂认知问题上具有强大潜力。但QAKAG框架也存在改进的空间，如减少问题分解阶段的人工介入、提高因提示信息的增加导致模型推理速度变慢的问题。未来，我们将以此为研究重点，持续优化QAKAG框架，提高其自动化程度和推理速度，并探索其在更广泛的修辞类别和语言理解任务中的适用性。

参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report."arXiv preprint arXiv:2303.08774(2023).

Yigal Attali and Jill Burstein. "Automated essay scoring with e-rater® V. 2." The Journal of Technology, Learning and Assessment 4.3 (2006).

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, Furu Wei. "Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers." arXiv preprint arXiv:2212.10559 (2022).

Thomas K. Landauer, Darrel Laham, Peter W. Foltz. "Automated scoring and annotation of essays with the Intelligent Essay Assessor." Automated essay scoring: A cross-disciplinary perspective (2003): 87-112.

Yucheng Li, Chenghua Lin, and Frank Guerin. "Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling." Proceedings of the 29th international conference on computational linguistics. 2022.

Yi Li, Dong Yu, and Pengyuan Liu. "CLGC: A corpus for Chinese literary grace evaluation." Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022.

Jinpeng Li, Zekai Zhang, Quan Tu, Xin Cheng, Dongyan Zhao, Rui Yan. "StyleChat: Learning Recitation-Augmented Memory in LLMs for Stylized Dialogue Generation." arXiv preprint arXiv:2403.11439 (2024).

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, Yong Jiang. "Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, et al. "Summary of chatgpt-related research and perspective towards the future of large language models." Meta-Radiology (2023): 100017.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, Tianyu Du. "RA-ISF: Learning to Answer and Understand from Retrieval Augmentation via Iterative Self-Feedback." arXiv preprint arXiv:2403.06840 (2024).

William C. Mann and Sandra A. Thompson. "Rhetorical structure theory: Toward a functional theory of text organization." Text-interdisciplinary Journal for the Study of Discourse 8.3 (1988): 243-281.

Ellis B. Page. "The imminence of... grading essays by computer." The Phi Delta Kappan 47.5 (1966): 238-243.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, Chris Dyer. "Metaphor detection with cross-lingual model transfer." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, Ashish Sabharwal. "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions." arXiv preprint arXiv:2212.10509 (2022).

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, Shuming Shi. "Siren's song in the AI ocean: a survey on hallucination in large language models." arXiv preprint arXiv:2309.01219 (2023).

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, Ed Chi. "Least-to-most prompting enables complex reasoning in large language models." arXiv preprint arXiv:2205.10625 (2022).

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, Diyi Yang. "Can large language models transform computational social science?." Computational Linguistics 50.1 (2024): 237-291.

蔡越著.AI赋能写作AI大模型高效写作一本通[M].北京：北京大学出版社, 2024

陈望道著.修辞学发凡[M].上海：复旦大学出版社, 2008

崔亿萍.大模型在自然语言处理的应用和研究[J].中国科技期刊数据库工业A, 2023, (12): 57-61

付瑞吉, 王栋, 王士进, 胡国平, 刘挺.面向作文自动评分的优美句识别[J].中文信息学报, 2018, 第32卷(6): 88-97

呼啸, 付瑞吉, 宋巍, 王士进, 胡国平, 秦兵, 刘挺.一种比喻句识别方法、装置、设备及存储介质[P].中国专利：CN111914544A, 2020.11.10

李玲玲, 王晓燕.人工智能翻译对文学文本处理的对比分析[J].时代报告, 2023, (46): 10-12

李明峰, 贾修一.基于多分类器集成学习的中文反语识别技术[J].计算机与数字工程, 2018, 第46卷(9): 1790-1795

刘广辉.基于修辞句识别的中学语文作文评价系统的设计与实现[D].中国科学院大学(中国科学院沈阳计算技术研究所), 2022

刘明杨.高考作文自动评分关键技术研究[D].哈尔滨工业大学, 2015

骆小所著.现代修辞学[M].昆明：云南人民出版社, 2010

邱文生.修辞的认知性研究[J].重庆理工大学学报(社会科学), 2015, (12): 122-126

石昀东.基于修辞使用的小学作文自动分类评价方法研究[D].华中师范大学, 2019

谭汝为著.汉语修辞学指要[M].天津出版传媒集团；天津人民出版社, 2023

王荣文著.现代汉语中的修辞艺术剖析[M].中国原子能出版社, 2018

徐盛桓.修辞研究的认知视角[J].西安外国语大学学报,2008,第16卷(2): 1-5

荀恩东, 饶高琦, 肖晓悦, 臧娇娇.大数据背景下BCC语料库的研制[J].语料库语言学, 2016, (1): 93-109, 118

詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙.北京大学CCL语料库的研制[J].语料库语言学, 2019, (1): 71-86, 116

赵琳玲.面向高考鉴赏类问题的明喻与拟人识别方法研究[D].山西大学, 2020

朱晓亮, 谯宇同, 龚颖, 殷姿, 董鑫, 石昀东, 刘三女牙, 杨宗凯.一种小学语文作文排比句自动识别的方法[P].中国专利：CN110598202A, 2019.12.20

A 附录

标签	数量	平均长度	答案	问题示例
夸张	1000	144.952	A	请从以下四个选项中选出具有夸张修辞手法的句子：A. 江浩被她的惨叫吓得三魂冲天，七魂出窍。B. 卢老师告诉了我们一个小诀窍就是，先读一句然后背一句。C. 其实我觉得要留待下一期再和大家分享一下。D. 在我看来，马修是一个才华横溢的音乐家，不过在1949年的法国乡村，他没有发展自己才华的机会，最终成为了一间男子寄宿学校的助理教师。
通感	1000	145.081	C	请从以下四个选项中选出具有通感修辞手法的句子：A. 差5分零点时，庙前已聚集了数百人，他们在等待着那新旧交替的难忘一刻。B. 虽然我是她的得力助手，但她做事永远都是毫不留情的啊。C. 她的声音很透明。D. 我一怔，抬头正想回答，看见妈妈只穿了一件衣服，身子在寒风中瑟瑟发抖。
比拟	1000	144.095	C	请从以下四个选项中选出具有比拟修辞手法的句子：A. 我们班的李老师就是其中一位，她教语文，她对自己的每一节课都认真备课，仔细研究如何才能让学生更容易理解自己所讲的知识。B. 我很小的时候爸爸曾经给我讲过圣诞节的故事，那可真是一个美妙的故事啊！C. 落日挥毫泼墨，把我们的世界装扮得美丽动人。D. 人的劳动能力是生产力中最活跃、最重要的因素。
反复	1000	145.201	B	请从以下四个选项中选出具有反复修辞手法的句子：A. 我们在离开的时候，爸爸去了公平秤，称了那条鱼的重量，公平秤显示的是3斤。B. 英国人永远，永远，永远不会成为奴隶。C. 陈春泉答应了表妹王采玉的请求后，左思右想、反复比较，才选定了毛鼎和家。D. 前者倾向于突破现有规则，后者更愿意维护传统。
对偶	1000	130.715	B	请从以下四个选项中选出具有对偶修辞手法的句子：A. 虽然我是她的得力助手，但她做事永远都是毫不留情的啊。B. 把问题找实，把根源挖深。C. 我忍不住咬了几口，可它的汁液太少了，一颗不过瘾，又拿起第二颗，第三颗.....D. 我上前地问了一句：“需要我帮忙吗？”
设问	1000	150.901	C	请从以下四个选项中选出具有设问修辞手法的句子：A. 我没有回答，我又想起了妈妈的话：将梦想进行到底。B. 我踏着轻松的步子来到教室里，同学们高高兴兴地纷纷跑上前来送我生日礼物。C. 女人同样喜欢帅哥，但又有多少男人愿意为了女人的一句赞美而天天待健身房，天天苦练肌肉？说到底这还是男女性社会地位问题不平等导致的。D. 然后，他又告诫自己，要冷静，不能感情用事。
比喻	1000	153.126	B	请从以下四个选项中选出具有比喻修辞手法的句子：A. 这时天空下起了绵绵细雨，但是人们一点也不在乎。B. 你就像辛勤的园丁，细心培育着祖国花朵，使这些花朵茁壮成长。C. 我生气极了，“不，你骗人，那不是故事，是真实的。”D. 明明是你暗恋别人，那个人从来察觉，也许知道了，却无动于衷。
反问	1000	140.766	B	请从以下四个选项中选出具有反问修辞手法的句子：A. 我爸爸最大的特点是：让我去挑战一些我不敢做的事情。B. 若非亲眼目睹，谁能相信水能变成“石头”呢？C. 就连她自己也说：“画动物是我的强项呦！”D. 从小到大可从来没有人送过礼物给我呀！
顶真	1000	135.800	D	请从以下四个选项中选出具有顶真修辞手法的句子：A. 作为马上要小学毕业的我，我们要好好珍惜我们之间的友谊，不要动不动就吵架。B. 每当看到大哥哥、大姐姐上学时那快乐的神情，我就羡慕极了。C. 大家都很吃惊地望着张老师，原以为是张老师的儿子过生日呢！D. 雪洁林，林护雪，雪影静谧。
排比	1000	173.492	D	请从以下四个选项中选出具有排比修辞手法的句子：A. 没有啊，我以为你喂了。B. 张老师高兴地让全班齐唱《祝你生日快乐》的歌。C. 这事情虽然看起来不可思议，但你们真的是冤枉潘武了。D. 延安的歌声它是黑夜的火把，雪天的煤炭，大旱的甘霖。

表 7: 大模型修辞认知评测集概况

修辞格	平均长度	最大长度	最小长度	知识点类型
反复	97.500000	192	41	定义、特征、变体特征、分析方法、易混淆点、作用
反问	95.833333	109	73	
夸张	109.500000	176	39	
对偶	94.666667	143	48	
排比	201.166667	532	55	
比喻	152.666667	228	48	
比拟	120.333333	183	66	
设问	94.333333	184	56	
通感	83.666667	96	57	
顶真	110.500000	288	45	

表 8: MCRKB知识库概况

修辞	类型	知识点
比喻	定义	比喻修辞手法的定义是根据心理联想，抓住和利用不同事物的相似点，用另一个事物来描绘所要表现的事物。
比喻	特征	比喻修辞手法的特征包括：1.一般由三部分组成，即：本体(被比喻的事物或情境)、喻词(表示比喻关系的词语)、喻体(比喻的事物或情境)。2.本体和喻体必须是性质不同的两类事物。3.在比喻句中，喻体必须出现。4.本体和喻体之间必须有相似点。比喻通过将不同的事物或情境进行类比，从而传达出相似之处，使得文学作品更加生动形象。
比喻	变体特征	比喻修辞手法的变体主要有三种：明喻、借喻、暗喻。明喻要求本体和喻体同时出现，但连接词为比喻词如“像”、“好似”、“仿佛”等，有时还会有配合词如“似的”、“一样”；暗喻要求本体和喻体同时出现，但使用了“是”、“成了”、“变成”等词来连接，或者以偏正式词组的形式出现，用本体修饰限制喻体，有时还会使用复指成分来构成暗喻；借喻则是直接叙述喻体，本体和喻词都不出现。
比喻	分析方法	比喻修辞手法的分析方法如下：例1：思想的春风吹遍油田。通过将“思想”与“春风”相比较，揭示出了它们之间的相似性，从而达到比喻的效果。例2：岸边的华灯倒映在湖中，宛如颗颗宝石缀在湖面之上。通过将“华灯”与“宝石”相比较，也实现了比喻的目的。例3：我就知道，我们之间已经隔了一层可悲的厚障壁了，我再也说不出话。通过描述“厚障壁”来暗示人们之间的隔阂，虽然没有直接提及喻词，但通过暗示和隐喻的方式也实现了比喻的效果。
比喻	易混淆点	比喻修辞手法易和比拟、夸张、通感修辞手法混淆。1.比喻与比拟之间的区别：比喻强调的是两个事物的相似性，而比拟则是利用它们的不同特性使它们融为一体，这一点在比喻中喻体是必不可少的，而在比拟中必须出现拟体。2.比喻与夸张之间的区别：夸张是放大或缩小事物的形象特征以增强表达效果，而比喻则是根据相似点将某一事物比作另一事物，使抽象的事物具体化，道理更加浅显易懂。3.比喻与通感之间的区别：比喻强调相似性，离不开本体、喻体和相似点，而通感更注重感觉器官之间的转移。
比喻	作用	比喻修辞手法的作用是通过将一个事物与另一个不同但相关的事物相比较，使深奥的道理更加浅显易懂，帮助人们加深体味；将抽象的事物具体化，使人们更容易接受；将概括的东西形象化，给人们留下鲜明的印象。

表 9: MCRKB知识库中比喻修辞的知识点示例

	Zero-shot-CoT	RAG-BaiKe	Few-Shot5	全知直授法	断点补教法	诊断增强法	Ours
排比	.995(.005↑)	.994(.004↑)	.995(.005↑)	.995(.005↑)	.994(.004↑)	.995(.005↑)	.995(.005↑)
比喻	.996(.006↑)	.987(.003↓)	.994(.004↑)	.996(.006↑)	.987(.003↓)	.994(.004↑)	.993(.003↑)
夸张	.955(.010↑)	.959(.014↑)	.959(.014↑)	.955(.010↑)	.959(.014↑)	.959(.014↑)	.957(.012↑)
反问	.987(.059↑)	.968(.040↑)	.982(.054↑)	.987(.059↑)	.968(.040↑)	.982(.054↑)	.984(.056↑)
比拟	.960(.005↑)	.940(.015↓)	.956(.001↑)	.960(.005↑)	.940(.015↓)	.956(.001↑)	.957(.002↑)
对偶	.995(.006↑)	.998(.009↑)	.997(.008↑)	.995(.006↑)	.998(.009↑)	.997(.008↑)	.995(.006↑)
通感	.915(.047↑)	.921(.053↑)	.916(.048↑)	.915(.047↑)	.921(.053↑)	.916(.048↑)	.908(.040↑)
反复	.970(.098↑)	.965(.093↑)	.957(.085↑)	.970(.098↑)	.965(.093↑)	.957(.085↑)	.974(.102↑)
设问	.964(.081↑)	.965(.082↑)	.964(.081↑)	.964(.081↑)	.965(.082↑)	.964(.081↑)	.965(.082↑)
顶真	.825(.493↑)	.590(.258↑)	.833(.501↑)	.825(.493↑)	.590(.258↑)	.833(.501↑)	.842(.510↑)

表 10: ChatGPT4(gpt-4-1106-preview)在不同方法下对十种修辞手法的认知表现

	Zero-shot-CoT	RAG-BaiKe	Few-Shot5	全知直授法	断点补教法	诊断增强法	Ours
排比	.989(.009↑)	**.979(.001↓)**	.990(.010↑)	.989(.009↑)	**.979(.001↓)**	.990(.010↑)	.989(.009↑)
比喻	.945(.117↑)	.896(.068↑)	.922(.094↑)	.945(.117↑)	.896(.068↑)	.922(.094↑)	.953(.125↑)
夸张	.847(.096↑)	.808(.057↑)	.818(.067↑)	.847(.096↑)	.808(.057↑)	.818(.067↑)	.852(.101↑)
反问	.890(.151↑)	.885(.146↑)	.950(.211↑)	.890(.151↑)	.885(.146↑)	.950(.211↑)	.920(.181↑)
比拟	.835(.416↑)	.821(.402↑)	.773(.354↑)	.835(.416↑)	.821(.402↑)	.773(.354↑)	.874(.455↑)
对偶	.943(.558↑)	.859(.474↑)	.966(.581↑)	.943(.558↑)	.859(.474↑)	.966(.581↑)	.946(.561↑)
通感	.805(.154↑)	.812(.161↑)	.765(.114↑)	.805(.154↑)	.812(.161↑)	.765(.114↑)	.817(.166↑)
反复	.915(.263↑)	.865(.213↑)	.917(.265↑)	.915(.263↑)	.865(.213↑)	.917(.265↑)	.917(.265↑)
设问	.858(.407↑)	.770(.319↑)	.887(.436↑)	.858(.407↑)	.770(.319↑)	.887(.436↑)	.887(.436↑)
顶真	.676(.408↑)	.501(.233↑)	.754(.486↑)	.676(.408↑)	.501(.233↑)	.754(.486↑)	.695(.427↑)

表 11: ChatGPT3.5(gpt-3.5-turbo)在不同方法下对十种修辞手法的认知表现

	Zero-shot-CoT	RAG-BaiKe	Few-Shot5	全知直授法	断点补教法	诊断增强法	Ours
排比	.975(.036↑)	.974(.035↑)	.966(.027↑)	.975(.036↑)	.974(.035↑)	.966(.027↑)	.974(.035↑)
比喻	.920(.028↑)	.880(.012↓)	.950(.058↑)	.920(.028↑)	.880(.012↓)	.950(.058↑)	.923(.031↑)
夸张	.840(.099↑)	.795(.054↑)	.788(.047↑)	.840(.099↑)	.795(.054↑)	.788(.047↑)	.815(.074↑)
反问	.885(.237↑)	.815(.167↑)	.763(.115↑)	.885(.237↑)	.815(.167↑)	.763(.115↑)	.847(.199↑)
比拟	.648(.376↑)	.549(.277↑)	.437(.165↑)	.648(.376↑)	.549(.277↑)	.437(.165↑)	.756(.484↑)
对偶	.903(.402↑)	.567(.066↑)	.725(.224↑)	.903(.402↑)	.567(.066↑)	.725(.224↑)	.859(.358↑)
通感	.644(.012↑)	.670(.038↑)	.629(.003↓)	.644(.012↑)	.670(.038↑)	.629(.003↓)	.639(.007↑)
反复	.820(.336↑)	.650(.166↑)	.651(.167↑)	.820(.336↑)	.650(.166↑)	.651(.167↑)	.828(.344↑)
设问	.871(.334↑)	.669(.132↑)	.608(.071↑)	.871(.334↑)	.669(.132↑)	.608(.071↑)	.857(.320↑)
顶真	.453(.214↑)	.296(.057↑)	.272(.033↑)	.453(.214↑)	.296(.057↑)	.272(.033↑)	.492(.253↑)

表 12: ChatGLM3-6B在不同方法下对十种修辞手法的认知表现

	Zero-shot-CoT	RAG-BaiKe	Few-Shot5	全知直授法	断点补教法	诊断增强法	Ours
排比	.941(.024↓)	.978(.013↑)	.891(.074↓)	.941(.024↓)	.978(.013↑)	.891(.074↓)	.967(.002↑)
比喻	.943(.086↑)	.933(.076↑)	.782(.075↓)	.943(.086↑)	.933(.076↑)	.782(.075↓)	.945(.088↑)
夸张	.822(.047↑)	.845(.070↑)	.815(.040↑)	.822(.047↑)	.845(.070↑)	.815(.040↑)	.857(.082↑)
反问	.935(.129↑)	.867(.061↑)	.765(.041↓)	.935(.129↑)	.867(.061↑)	.765(.041↓)	.929(.123↑)
比拟	.744(.152↑)	.712(.120↑)	.644(.052↑)	.744(.152↑)	.712(.120↑)	.644(.052↑)	.708(.116↑)
对偶	.893(.228↑)	.788(.123↑)	.750(.085↑)	.893(.228↑)	.788(.123↑)	.750(.085↑)	.848(.183↑)
通感	.713(.048↑)	.747(.082↑)	.693(.028↑)	.713(.048↑)	.747(.082↑)	.693(.028↑)	.757(.092↑)
反复	.757(.197↑)	.691(.131↑)	.605(.045↑)	.757(.197↑)	.691(.131↑)	.605(.045↑)	.803(.243↑)
设问	.884(.319↑)	.747(.182↑)	.568(.003↑)	.884(.319↑)	.747(.182↑)	.568(.003↑)	.935(.370↑)
顶真	.490(.084↑)	.529(.123↑)	.446(.040↑)	.490(.084↑)	.529(.123↑)	.446(.040↑)	.545(.139↑)

表 13: Qwen-7B-Chat在不同方法下对十种修辞手法的认知表现