大语言模型的自我恒定性现象及其在知识增强任务上的应用

雷艳[1,2]，庞亮[1]，魏子豪[1,2]，王元卓[1,2]，沈华伟[1,2]，程学旗[1,2]
1. 中国科学院计算技术研究所智能算法安全全国重点实验室，北京 100190；
2. 中国科学院大学计算机学院，北京 100049

摘要
近期研究指出，检索和大语言模型中存在显著的“源偏见”，即相较于真实人类数据，其更偏好模型生成的内容。人类存在类似的“自我恒定性”现象，即倾向于维护自我概念的一致性，相较于外部事实，其更倾向于相信已有的知识。探讨大模型是否表现出与类人的认知偏差，分别从显式方式（模型自身生成的内容）与隐式方式（提示词诱导模型认为内容是其生成的）两方面展开研究，并结合自评估与一致性两种置信度评价方法，在GPT-4o、DeepSeek-R1、Llama2（7B～70B）、Qwen2（7B～72B）等模型上进行了系统实验。结果表明，在隐式方式下，模型表现出明显的自我恒定性，对自身内容有显著更高的信心。利用该特性可以系统提升模型在各类知识增强任务中的表现：在TriviaQA、NQ、HotpotQA、FEVER和ZsRE等数据集上，角色提示策略带来准确率提升；在存在大量噪声干扰文档时，仍能保持稳健优势，显示出良好的鲁棒性和泛化性。最后，揭示了指令微调与人类反馈强化学习是导致模型产生此类偏见的核心因素，从训练与对齐机制层面解释了自我恒定性的形成。

关键词
大语言模型；检索增强；知识增强任务；源偏见

中图分类号：TP391 文献标志码：A doi：10.11959/j.issn.2096-0271.2026038

Self-consistency phenomenon in large language model and its application in knowledge augmentation tasks

Lei Yan[1,2], Pang Liang[1], Wei Zihao[1,2], Wang Yuanzhuo[1,2], Shen Huawei[1,2], Cheng Xueqi[1,2]
1. State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
2. Computer Science, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract
Recent studies have revealed significant “source bias” in retrieval and large language models (LLM), where these models exhibit a preference for model-generated content over authentic human-written data. Humans similarly demonstrate “self-consistency” phenomena, characterized by a tendency to maintain coherent self-concepts and favor existing knowledge over external facts. It was investigated whether LLMs exhibit human-like cognitive biases. It was conducted from two perspectives: an explicit approach (content directly generated by the LLM) and an implicit

approach (using prompts to induce the models to perceive the content as self-generated). Confidence was evaluated using both self-assessment and consistency methods. Systematic experiments were carried out on several models, including GPT-4o, DeepSeek-R1, Llama2 (7B-70B), and Qwen2 (7B-72B). The results demonstrated that under implicit conditions, LLMs exhibit pronounced self-consistency phenomena, displaying significantly higher confidence in self-attributed content. Leveraging this characteristic yields consistent performance gains: the proposed role-based prompting strategy improved accuracy on TriviaQA, NQ, HotpotQA, FEVER, and ZsRE, and maintained a clear advantage under high levels of noisy retrieved documents, indicating strong robustness and generalization. Finally, our mechanistic analysis revealed that instruction fine-tuning and reinforcement learning from human feedback were the primary factors contributing to such biases in models, explaining how self-consistency emerges during training.

Key words
large language model, retrieval-augmented generation, knowledge augmentation, source bias

0 引言

以ChatGPT为代表的大语言模型（large language model，LLM）的出现，催化了人工智能生成内容（artificial intelligence generated content，AIGC）领域的快速发展[1-2]。LLM在大规模自动化生成类人文本方面的卓越能力，使海量合成内容大量涌入信息生态，重塑上游检索及下游人机交互流程，受到工业界与学术界的广泛关注[3-4]。近期研究揭示了模型对合成内容相较于真实内容的系统性偏好[5-8]，即“源偏见”。源偏见是一个经验性的现象描述，指模型系统性地偏好AI生成内容而非人类撰写内容，其已在文本检索、多模态检索等多个场景中被观察到。Dai等[5-6]证实，神经检索模型倾向于将机器生成的文档排序优于人类撰写的文档。Xu等[7]将这些发现扩展到多模态检索，表明图像检索模型同样优先考虑AI生成的图像，确立了源偏见跨模态的普遍性。Tan等[8]进一步指出，在存在事实冲突的检索增强场景中，即便生成内容包含错误信息，LLM仍倾向于支持生成文档而非真实文档。Mao等[9]发现，相较人类撰写的文本，LLM在改写AI生成文本时施加的修改幅度显著更小，并成功利用这种不对称性进行AI生成内容检测。这种现象与心理学中的“自我恒定性”（self-constancy）理论不谋而合。该理论由Prescott Lecky提出，其认为人类倾向于维护信念与行为的一致性，缺乏这种一致性可能导致认知失调[10]。自我恒定性是借鉴心理学理论提出的解释性框架，用于从认知偏差的视角理解LLM为何对“属于自己”的内容表现出更高的信任度。源偏见回答“存在什么现象”，自我恒定性则尝试回答“为何存在该现象”。本文旨在探索大模型是否具备自我恒定性，在此基础上，进一步系统性地揭示LLM对自身生成内容的认知机制，并挖掘其在应用层面的实用特性。

为系统研究LLM的自我恒定性，本文设计了显式和隐式两种范式的实验，这两种范式以模型是否实际生成内容为区分。大模型自身生成内容的形式示例如**图1**所示，在显式范式中，模型自主生成并理解内容。相反，隐式范式采用精心设计的提示词，诱导模型相信外部提供的内容源自其自身，而不需要实际生成。本文采用两种置信度评估方法在检索增强的开放域问答中评估模型对自我归因内容的态度：显式自评估，即模型同时提供答案和置信度

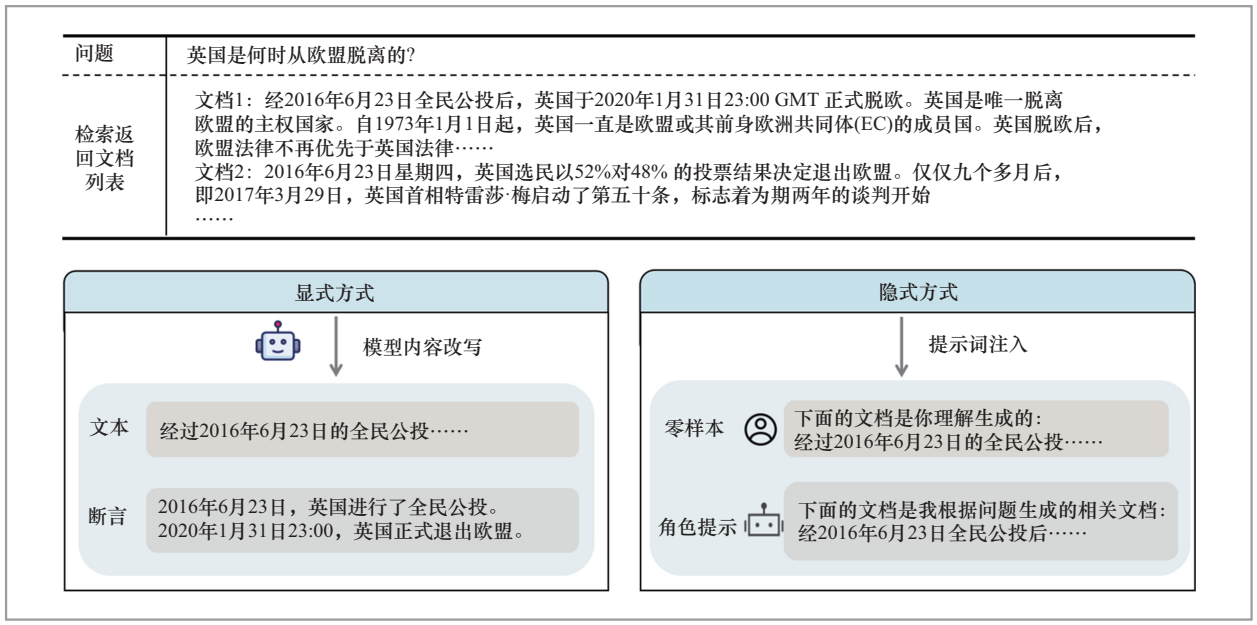


图1 大模型自身生成内容的形式示例

分数；隐式一致性评估，是指通过多个样本间的一致性来测量置信度。比较上述评估方法，能够细致理解模型在不同提示下的自信程度。

尽管在多样化任务中表现出色，LLM在知识密集型场景中仍面临挑战：参数化知识会随时间过时[11]，领域特定专业知识仍然有限[12-13]，生成内容存在幻觉问题[14-15]。检索增强生成已成为一种有前景的解决方案，但也引入了新的复杂性，即检索到的文档可能包含不相关信息，误导模型响应[16-17]。因此，增强模型识别幻觉信息的能力变得至关重要。本文提出利用LLM的自我一致性来改进外部知识整合，从而提升知识增强任务的性能。知识增强任务是指模型参数化知识不足以独立完成、需要整合外部知识源的任务类型。本文涵盖的任务如下。

（1）开放域问答，即给定问题，模型需要从外部文档中检索并整合相关信息生成答案，代表数据集包括TriviaQA数据集[18]、自然问题（natural questions，NQ）数据集[19]和HotpotQA数据集[20]。

（2）事实核查，即给定陈述句，模型需要根据外部证据判断其真实性，代表数据集为FEVER[21]。

（3）槽填充，即给定实体和关系类型，模型需要从外部知识中提取对应属性值，代表数据集为ZsRE[22]。

上述任务的共同特征在于：模型必须整合检索到的外部知识，而非仅依赖预训练阶段习得的参数化知识。

本文的核心贡献体现在以下3个方面。

（1）理论框架的构建。本文首次提出将“自我恒定性”作为解释LLM自我偏好现象的类人认知框架，将心理学中关于人类自我概念维护的理论引入LLM研究，为理解模型行为提供了新的理论视角。

（2）机制的系统揭示。本文发现隐式自我归属（尤其是角色提示词策略）相较于显式生成过程更能有效激发模型的自我恒定性，并通过在TriviaQA[18]、NQ[19]和

HotpotQA[20]等知识增强任务上的系统实验，验证了该特性在提升任务性能与增强噪声鲁棒性方面的实际应用价值。

（3）成因的初步探索。本文通过对比指令微调与人类反馈强化学习前后的模型表现，初步揭示了上述训练过程在诱导模型自我恒定性中的关键作用，为理解该现象的形成机制提供了依据。

1 相关工作

源偏见在检索模型和语言模型中被广泛讨论。本文旨在通过分析LLM的“自我恒定性”现象及其背后的机制，为改进大模型内容生成提供新的视角。

1.1 源偏见

AIGC利用先进的生成式AI技术创建内容，是一个快速发展的领域。与人类制作的传统内容不同，AIGC可以在相当短的时间内大规模生成[23-24]。随着AIGC在各个领域的应用变得越来越普遍，其潜在风险也成为关注焦点[24]。最近的研究表明，LLM存在源偏见，即LLM生成的内容优于人类撰写的文本。当前研究主要从以下4个维度展开。

（1）文本检索中的源偏见：Dai等[6]研究发现，文本神经检索模型倾向于将大模型生成的内容排在人类撰写的文本之前。该研究通过提示词方式让模型理解并改写文章，将此作为模型生成的内容。在SciFact[25]和NQ320K[19]两个数据集上的实验证实，文本神经检索模型对模型生成内容的偏好高于人类撰写的文本。

（2）多模态检索中的源偏见：Xu等[7]发现，同样的源偏见也存在于图像检索模型中。文本-图像检索模型倾向于将AI生成的图像排在高于真实图像的位置。在Flickr30k[26]和MSCOCO[27]数据集上的测试显示，该模型表现出明显的源偏见，即使AI生成的图像并未比真实图像表现出更多的与查询相关的视觉特征。

（3）检索增强场景中的源偏见：Tan等[8]研究了在大模型检索增强设置下，当检索文档与生成文档之间存在知识冲突时，生成式语言模型（如GPT-4/3.5[28]以及Llama2[29]）更倾向于相信模型生成的内容。研究结果揭示了LLM明显偏向于生成的上下文，即使生成的上下文提供了不正确的信息，而检索上下文包含正确答案。

（4）源偏见在AI内容检测中的应用：Mao等[9]利用这种源偏见来辅助AI内容检测。该研究发现，在执行重写任务时，LLM对AI生成的原始文本修改幅度较小，而对人类书写的文本修改幅度较大。该研究通过提示词方式让模型改写原始内容，通过编辑距离大小来判断内容是模型生成的文本还是人类撰写的文本，该方法在检测准确率上比当前最优模型提升了29%。

1.2 置信度

置信度是指模型对自己输出的确定性或保证程度。传统的置信度分数预测主要通过白盒模型的内部状态来判断，如词级别的概率[30]、校准策略[31]以及微调等。然而，随着商业化闭源大模型（如GPT-3.5、GPT-4[32]）的出现，用户只能通过输入获取模型的输出，无法获取中间层的词级别概率以及向量等信息。考虑上述限制，越来越多的研究开始探索黑盒方法来激发模型表达置信度分数。目前主流方法可分为以下两类。

（1）自然语言方式直接表达置信度：

通过提示词方式让大模型自己表达出确定性[33-35]，包含多种不同的类型，可按照确定性程度表述（最不自信、较不自信、中等、自信、非常自信）或按照数字表述（如85%[36-37]）。

（2）一致性间接评估确定性分数：通过多次采样，统计答案并将一致性最高的分数作为代表[34]。

1.3 检索增强

检索增强方法[38-39]将外部语料库中的相关段落纳入上下文，从而显著增强语言模型应对领域知识不足[11]、模型幻觉[14-15]等挑战的能力。检索增强范式分为两阶段：先检索，后生成。检索增强语言模型（retrieval-augmented language model，REALM）[40]、检索增强Transformer（retrieval-enhanced Transformer，RETRO）[41]等已经证明了这种范式能够有效提高生成能力。然而，检索增强也存在一些弊端：检索之后会增加大量不相关的噪声文档，误导模型生成错误的答案[16,42]。Ren等[43]发现在学习领域知识时，注入的知识与内部知识不一致会降低模型的性能。研究者提出了多种改进方案，包括基于重排序的检索优化[44]和基于意图检测的查询增强[45]等。因此，增强模型识别幻觉信息的能力变得至关重要。在指令微调阶段，决定模型性能的不是注入的领域知识，而是注入的知识与内部知识的一致性。本文通过探索大模型自我恒定性现象，期望利用这一现象提升模型理解外部知识的能力，进一步提升检索增强开放域问答的效果。

综上所述，现有的源偏见研究主要从“文档来源类别”的角度展开，即考察模型对机器生成内容与人类撰写内容的差异化处理。具体而言，Dai等[5-6]与Xu等[7]关注检索模型在排序任务中对不同来源文档的偏好差异；Tan等[8]研究生成模型在知识冲突场景下对不同来源信息的选择倾向；Mao等[9]则利用模型对不同来源文本的改写差异进行AI内容检测。上述工作的共同特点在于：将“来源”作为二元类别变量（机器生成、人类撰写）进行考察[46-48]。与之不同的是，本文将研究视角从“来源类别”转向“自我归属关系”，聚焦LLM对属于“自身内容”的认知与态度。为此，本文系统区分了两类“自身内容”的表征形式——显式（模型实际生成的内容）与隐式（通过提示词诱导模型认为属于自身的内容），并采用自评估置信度测量，多维度刻画模型的自我恒定性。这一研究视角的转换使本文能够深入探究LLM自我偏好的认知机制，而非仅停留在来源偏好的现象描述层面。

2 大模型自身知识形式

为了探讨大模型是否对其自身生成的内容表现出自我恒定性现象，本文设计了两种不同的方式：显式和隐式。显式方式要求模型对原始内容进行修改，并基于修改后的自身文档内容进行进一步生成；隐式方式则通过提示词诱导模型认为内容是其自身生成的，而实际不需要模型真正生成。这两种方法的具体实施如下。

2.1 显式方式

模型首先在阅读理解后将内容转化为自身知识并进行输出。此过程类似于人类在阅读文档理解后将内容重新组织并表达出来。该方法的优点在于：重新组织表达

的内容通常更加通顺且逻辑严密，能够更好地整合和阐述原始信息。例如，Dai等[6]通过提示词让模型对原始文档进行理解并“消化”后生成新内容，再使用模型生成的新文档进行问答，以评估其对自身生成内容的理解和态度。改写完成后，用模型自身生成的文档替换原始文档进行问答测试。然而，显式改写的一个潜在问题是模型在改写过程中可能引入语义偏移甚至幻觉，从而影响显式“自身内容”与原始文档之间的可比性。为保证后续比较的公平性，本文对改写过程进行了质量控制。具体而言，对于每个原始文档，本文采用过采样策略，在输出时采样N次（N=30），得到N个输出：$\{o_1, o_2, \cdots, o_N\}$。本文使用sentence-transformer中的Paraphrase-MiniLM-L6-v2将原文与各候选改写编码为向量，计算两者的余弦相似度，选择相似度最高的一条作为最终的改写结果，具体计算式为：

$$T^w = \arg\max_i^N \left\{ \text{cosine}(\boldsymbol{e}_r^w, \boldsymbol{e}_i^w) | \boldsymbol{e}_i^w \in E \right\} \quad (1)$$

其中，$\boldsymbol{e}_r^w$为第w个样本原始（raw）文本的向量表示，$\boldsymbol{e}_i^w$为第w个样本的第i次采样结果，$\boldsymbol{E}$表示候选样本向量集合。避免偏离原始文档的语义是探索模型对正常改写后文本自我恒定性的先决条件。改写前后文本在语义相似度和Jaccard词级别重叠度上的分布情况如**图2**所示，大多数改写后的文本与原始文本的语义相似度在0.8以上，而词级别重叠度集中在0.3～0.6，说明本文在较好地保持语义一致性的同时，获得了足够的表层表达差异。这为探究显式自身内容下的自我恒定性提供了一个相对公平且可控的实验基础。

2.1.1 模型改写为文本

模型先对原始文档进行阅读理解与改写，将内容“消化”后重新组织成自然语言文本形式，即类似于人类理解文档后用自己的语言重新表达，改写后的内容通常更通顺、逻辑更严密。

2.1.2 模型改写为断言

考虑原始文档中常包含大量无实际意义的词汇，而人类在学习记忆的过程中，通常会从原始文本信息中提取关键信息，并对应到模型中，即断言形式。

2.2 隐式方式

隐式方式的设计旨在不直接修改原始内容，而是通过提示词的隐晦方式，让模

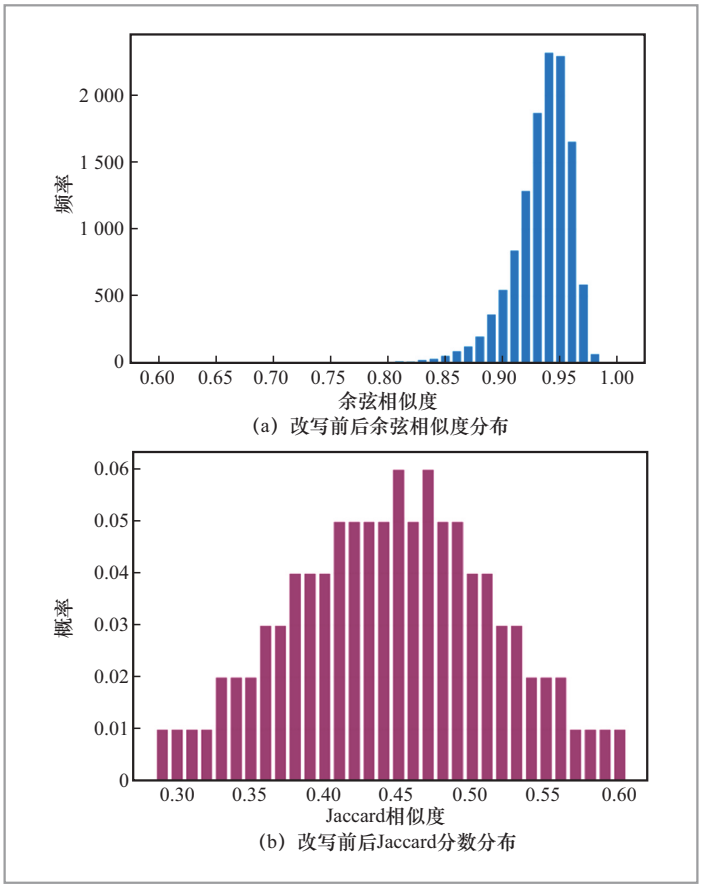


图2 改写前后文本在语义相似度和Jaccard词级别重叠度上的分布情况

型认为内容是其自身生成的，从而探索模型对该情况下的内容偏好，具体包括以下两种。

2.2.1 零样本提示诱导

零样本通过提示词让模型认为文档是自身生成的，而不需要对原始内容进行修改。

2.2.2 角色提示词

与传统的模型不同，LLM在指令微调阶段的数据组织格式中包含多种不同的角色，包括系统（system）角色、用户（user）角色以及助手（assistant）角色。在检索增强的场景中，通常本文的提示词直接将用户的查询以及原始文档一并放在用户角色中。对于大模型而言，通常以对话的形式进行使用，而助手角色中的内容相当于模型自身的输出。本文将检索到的文档内容放置在助手角色中，从而诱导模型认为这些内容是其自身的。

3 实验设置

本节详细介绍实验中使用的数据集和具体的实验设置，通过分析实验结果，评估模型在不同知识形式下的自我恒定性，以及对检索增强开放域问答任务的影响。

3.1 数据集

本文选择3类常用的知识增强任务进行评估，包括开放域问答、事实核查以及槽填充任务。各任务采用的数据集如下。

（1）开放域问答数据集：TriviaQA[18]、NQ[19]以及HotpotQA[20]。TriviaQA是一个真实场景下的问答数据集，往往需要从相应的答案证据句中进一步推理才能找到答案。NQ数据集收集了来自Google的用户查询，并在维基百科中标注了与问题对应的支撑答案。HotpotQA是根据维基百科构建的多跳问答数据集，该数据集中的问题是多跳问题，需要从多篇文章中推理得到最终的正确答案。

（2）事实核查数据集：FEVER[21]。FEVER根据文本提取事实进行验证，需要根据维基百科的内容进行人工核查，分为支持、反对以及信息不足3类。

（3）槽填充数据集：ZsRE[22]。ZsRE是零样本关系抽取数据集，也需要基于维基百科知识进行回应。

为了探索大模型对自身知识的“态度”，本文采用固定随机种子（seed=42）的均匀随机抽样方法，从每个数据集的测试集中抽取2 000条样本。该样本量的选择基于以下考虑。

（1）统计效力：根据中心极限定理，2 000条样本足以保证估计的稳定性[49-51]。

（2）计算成本：考虑需要在多个模型、多种设置下进行多次采样实验，该规模在保证结果可靠性的同时，兼顾了实验效率。具体来说，使用稠密段落检索（dense passage retriever，DPR）模型[52]来构建外部知识，从维基百科中为每个问题检索出最相关的若干文档集合。这些文档集合中既包含正确答案的文档，也包括一些相关但无法推导出正确答案的干扰文档，将检索到的文档作为提示词的一部分输入模型。

3.2 模型选择

本文选择了3类主要的LLM：ChatGPT使用GPT-4o、GPT-3.5-turbo版本；Llama2[29]使用由Meta提供的7B、13B、70B规模的开源模型；Qwen2[53]使

用由阿里提供的7B、14B、72B规模模型，是目前工业界知名的大模型。

3.3 研究问题

为深入探讨LLM在面对“自身内容”时的态度，本文旨在为理解生成模型的内部工作原理及改进其性能表现提供新的视角。基于此，本文提出以下研究问题。

问题1：LLM对“自身内容”如何看待？是否表现出类人的自我恒定性现象？

问题2：如何利用LLM的自我恒定性提升知识增强任务的能力？

问题3：LLM表现出自我恒定性的潜在原因是什么？

3.4 评估指标

（1）置信度

为了探究大模型对“自身内容”的态度，本文选择置信度作为关键指标。置信度是模型对其输出可靠性的评估，能够分析模型在不同提示下对其自身回答的信心水平。本文采用自评估黑盒方法来衡量，即通过提示词策略让大模型以自然语言的形式输出其置信度。本文选择了Top-k策略[54]，即让模型提供k个猜想以及对应的置信度，具体计算式为：

$$C(q_i)=\frac{1}{K}\max_{j=1}^{K}(M(q_i)) \tag{2}$$

其中，K表示模型预测答案的数量，$M(q_i)$表示模型根据问题输出的第i个答案的置信度。对比模型改写前后置信度差值来衡量对其生成内容的自信程度，具体计算式为：

$$\Delta C=C_{\text{after}}-C_{\text{before}} \tag{3}$$

（2）准确率

在评估大模型的开放域问答能力时，考虑到当前模型生成的文本通常为长文本，如果严格按照完全匹配标准进行评估，显然会低估模型的实际问答能力。因此，本文采用了更适合长回复的评估标准，即只要正确答案出现在生成的回复中，本文就认为该回答是正确的。对比模型改写前后问答准确率差值来衡量问答性能提升，具体计算式为：

$$\Delta\text{EM}=\text{EM}_{\text{after}}-\text{EM}_{\text{before}} \tag{4}$$

其中，EM_{after}表示应用显式方式或隐式方式后的问答准确率，$\text{EM}_{\text{before}}$表示原始问答准确率。通过这一差值，本文可以评估显式方式和隐式方式对模型问答能力的影响。

4 实验结果

本节从上述三大研究问题展开，针对每个研究问题进行实验分析。

4.1 模型对自身生成内容的置信度评估（问题1）

在显式和隐式两种形式下，本文采取自评估方式进行实验。不同模型在TriviaQA、NQ及HotpotQA数据集上的置信度变化见**表1**，每个单元格内依次展示上述3个数据集的结果。统计显著性基于配对t检验计算，其中正值表示相应方法提升了模型置信度，负值则表示降低。

（1）隐式方式（零样本提示诱导以及角色提示词）相较于显式方法表现出更明显的自我恒定性现象。在3个开放域问答的数据集上，不同的模型在隐式设置下给出的置信度比原始的置信度有明显提升。具体来说，ChatGPT在自评估方式下TriviaQA平均提升了1.84，Llama2-7B平均提升了3.63，Qwen2-7B平均提升了

表1 不同模型在TriviaQA、NQ及HotpotQA数据集上的置信度变化

模型	显式方式下的置信度变化ΔC		隐式方式下的置信度变化ΔC	
	模型改写为文本	模型改写为断言	零样本提示诱导	角色提示词
GPT-4o	0.93/0.62/0.75	−1.10/−0.94/−0.86	2.20*/1.88*/1.79*	**3.00**/2.47**/2.69****
ChatGPT	1.17/0.72/3.15	−2.90/−1.85/−0.85	1.74/**1.42**/1.65	**1.93**/0.85/**3.45****
DeepSeek-R1	0.72/0.45/0.51	−0.61/−0.54/−0.43	3.25*/2.41*/2.35*	**4.34***/3.55***/3.41****
Llama2-7B	−1.36/−1.69/−0.43	0.00/−0.29/1.48	**7.20**/4.77**/2.38***	0.07/−1.57/1.25
Llama2-13B	1.32/0.87/−1.94	1.47/0.43/−0.14	**6.41**/2.59*/2.95***	−8.56**/−13.19***/−12.93**
Llama2-70B	1.21/1.14/1.16	0.26/0.17/0.25	**4.81**/3.21*/2.04**	4.51**/2.88**/1.99*
Qwen2-7B	0.58/0.07/0.25	1.68/0.76/1.74	6.78**/**4.72****/**7.28*****	**7.02***/4.37*/6.84****
Qwen2-14B	6.87***/8.02/5.09	−4.49**/−4.80**/−3.71	7.69/ 11.49/9.09	**13.77***/16.37***/13.86****
Qwen2-72B	1.84/2.55**/1.66	−0.91/−0.84/−0.77	4.64/4.22/4.47	**6.01***/5.42***/5.61****

注：*表示$p<0.05$，**表示$p<0.01$，***表示$p<0.001$，未标注表示差异不显著($p\geq0.05$)。

6.9。零样本提示诱导设置下92.6%的结果达到统计显著水平（$p<0.05$），角色提示词设置下77.8%的结果达到统计显著水平。相比之下，显式方式（模型改写为文本和模型改写为断言）的显著比例仅为14.8%和11.1%，且多为负向效果。隐式方式的有效性在GPT-4o、DeepSeek-R1、Llama2-70B和Qwen2-72B等大模型上得到验证。GPT-4o在角色提示词设置下的自评估置信度提升达3.00/2.47/2.69，Qwen2-72B的提升达6.01/5.42/5.61，Deep-Seek-R1同样表现出显著的置信度提升（4.34/3.55/3.41），均显著高于显式方式。这说明模型对“自身内容”更相信，表现出与人类相似的自我恒定性现象。

（2）不同模型表现出的自我恒定性程度不同。其中，Qwen系列模型的自我恒定性现象比ChatGPT及Llama2更明显。在TriviaQA数据集上，角色提示词设置下，Qwen2-14B在自评估下置信度分数提高了13.77，提升的幅度显著。

（3）显式方式未能有效激发自我恒定性。在显式设置下，无论是改写为文本还是改写为断言，模型的置信度提升都不明显，甚至在某些情况下出现下降。即使在GPT-4o、Qwen2-72B等大规模模型上，显式表示的文本和断言方法仍然导致性能下降或无明显提升，进一步证实了显式方式的局限性并非由模型能力不足导致。这可能是由于模型将重新输入的自身文本视为普通上下文，且改写过程可能导致了语义损耗或引入了幻觉。

4.2 自我恒定现象对知识增强任务的影响(问题2)

不同模型在TriviaQA、NQ及HotpotQA数据集上的准确率变化见**表2**，每个单元格内依次展示上述3个数据集的结果。由**表2**可知，隐式方式下的角色提示词表现最优，在各模型上均取得稳定提升。具体而言，在中等规模模型中，Qwen2-7B在TriviaQA上提升4.0个百分点，Llama2-7B提升3.7个百分点，

表2 不同模型在TriviaQA、NQ及HotpotQA数据集上的准确率变化

模型	显式方式下的问答准确率变化ΔEM		隐式方式下的问答准确率变化ΔEM	
	模型改写为文本	模型改写为断言	零样本提示诱导	角色提示词
GPT-4o	-0.54%/-0.77%/-0.93%	-1.00%/-1.34%/0.54%	0.84%/1.24%/0.89%	**1.83%/2.25%/1.49%**
ChatGPT	-0.80%/-1.25%/-3.20%	-1.25%/-5.20%/1.90%	0.45%/2.75%/1.35%	**0.60%/3.95%/1.90%**
DeepSeek-R1	-0.32%/-0.51%/-0.65%	-0.83%/-1.13%/0.39%	1.29%/1.57%/1.32%	**2.55%/2.84%/1.87%**
Llama2-7B	0.95%/-2.90%/-2.90%	-3.50%/-8.56%/2.55%	1.15%/-2.20%/2.65%	**3.70%/5.20%/2.55%**
Llama2-13B	-1.75%/-2.21%/-2.65%	-4.70%/-8.46%/3.50%	-3.55%/-1.85%/1.15%	**0.85%/6.15%/3.50%**
Llama2-70B	-0.61%/-0.92%/-1.22%	1.24%/-1.54%/0.68%	1.52%/1.85%/1.22%	**2.87%/3.21%/2.13%**
Qwen2-7B	-1.10%/-3.37%/-5.36%	-4.10%/-10.96%/-6.33%	2.50%/2.35%/0.15%	**4.00%/4.35%/2.70%**
Qwen2-14B	-1.50%/-9.09%/-6.85%	-6.95%/-14.42%/-8.49%	-3.10%/-2.65%/-1.65%	**1.50%/-0.25%/0.20%**
Qwen2-72B	-0.42%/-0.83%/-0.91%	-0.91%/-1.25%/0.44%	1.84%/2.42%/1.54%	**3.14%/3.54%/2.34%**

Llama2-13B在NQ上提升6.15个百分点。大规模模型同样受益于该策略，GPT-4o在3个数据集上分别提升1.83/2.25/1.49个百分点，DeepSeek-R1提升2.55/2.84/1.87个百分点，Qwen2-72B提升3.14/3.54/2.34个百分点。相比之下，零样本提示诱导的提升幅度较小且不稳定，而显式方式在多数情况下导致性能下降。

为验证方法的任务泛化性，本文在FEVER数据集上进行了补充实验。不同模型在FEVER及ZsRE数据集上的检索增强开放域问答准确率变化见**表3**，每个单元格内依次展示上述两个数据集的结果。角色提示词策略在事实核查任务上同样有效：Qwen2-14B提升最显著（4.75个百分点），其次是Qwen2-72B（2.82个百分点）、ChatGPT（2.32个百分点）和DeepSeek-R1（2.11个百分点）。在ZsRE数据集上，

表3 不同模型在FEVER及ZsRE数据集上的检索增强开放域问答准确率变化

模型	显式方式下的检索增强准确率变化ΔEM		隐式方式下的检索增强准确率变化ΔEM	
	模型改写为文本	模型改写为断言	零样本提示诱导	角色提示词
GPT-4o	-0.41%/-0.65%	-0.84%/-0.71%	0.76%/1.24%	**1.84%/2.14%**
ChatGPT	-0.90%/-2.23%	-2.13%/-0.88%	0.86%/1.42%	**2.32%/2.24%**
DeepSeek-R1	-0.34%/-0.66%	-0.66%/-0.54%	0.66%/1.32%	**2.11%/2.65%**
Llama2-7B	2.76%/0.27%	-0.93%/-3.23%	-1.05%/0.65%	**1.40%/3.80%**
Llama2-13B	3.80%/0.65%	-6.67%/-0.96%	0.30%/-0.95%	**2.60%/4.85%**
Llama2-70B	4.12%/1.04%	-0.55%/-0.41%	0.54%/1.13%	**2.21%/2.81%**
Qwen2-7B	-1.26%/-3.91%	-3.47%/-0.84%	0.05%/2.70%	**2.20%/5.85%**
Qwen2-14B	5.85%/-3.62%	-6.95%/-3.10%	-0.05%/1.80%	**4.75%/1.80%**
Qwen2-72B	2.84%/0.65%	-2.72%/-0.69%	0.14%/2.54%	**2.82%/3.21%**

角色提示词策略的优势更明显。Qwen2-7B提升达5.85个百分点，Llama2-13B提升4.85个百分点，Llama2-7B提升3.80个百分点。大规模模型中，Qwen2-72B提升3.21个百分点，DeepSeek-R1提升2.65个百分点。

此外，本文还从鲁棒性以及泛化性的角度进行验证。

（1）角色提示词的鲁棒性。本文在上下文内容中设置0、0.5、1.0不同比例的干扰文档数量。在角色提示词设置下，不同模型在不同干扰文档数量下的准确率变化趋势如**图3**所示。由**图3**可知，随着干扰比例的增加，不同模型的开放域问答能力提升越来越多，说明角色提示词下，模型能够稳定地表现出抗干扰能力。

（2）角色提示词的泛化性。在主实验中，本文将改写之后的内容（包括非结构化文本和结构化断言）作为输入提供给模型，结果显示整体效果均有下降。为进一步探究角色提示词的影响，本文将改写之后的内容置于角色提示词中重新测试，不同输出格式（文本/断言）在不同提示词（思维链/角色）下的任务准确率变化见**表4**。角色提示词对不同格式的文本内容都有明显的效果提升，相比之下，思维链提示词在绝大多数情况下未见提升。

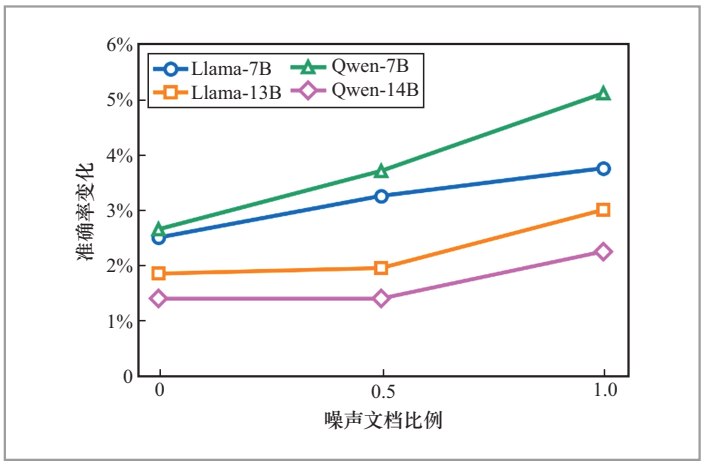


图3 不同模型在不同干扰文档数量下的准确率变化趋势

表4 不同输出格式(文本/断言)在不同提示词(思维链/角色)下的任务准确率变化

输出格式	提示词	Llama2-7B	Llama2-13B	Qwen2-7B	Qwen2-14B
文本	思维链	−0.29%	−0.46%	−0.97%	−0.79%
文本	角色	1.62%	4.45%	3.82%	2.16%
断言	思维链	0.40%	−1.25%	−1.55%	0.10%
断言	角色	5.55%	1.95%	2.10%	3.40%

4.3 大模型自我恒定性现象的原因分析(问题3)

（1）大模型为何不会对显式改写的内容产生偏好？Tan等[8]发现当模型生成文档与真实文档同时存在冲突时，模型倾向于相信自身生成的文档内容，原因有两个：一是相关性更强，即根据问题让模型生成相关的文档，生成的文本与问题是完全紧密相关的；二是完整性更好，即模型生成的段落通常更完整，而真实文档过长，在进行问答时通常会进行截断分块，导致内容不完整。Dai等[6]发现检索模型更偏好生成的内容，这是因为模型生成的内容流畅度更高，会倾向于将检索的内容放在更靠前的位置。然而，对于开放域问答而言，直接改写原始文档并不会产生过多的增益。模型在改写过程中存在两种情况：一是语义保持一致，即改写之后更有逻辑且更流畅；二是内容缺失或幻觉，即改写后出现内容缺失、幻觉等情况。由于大模型具备强大的阅读理解能力，只要正确答案出现在上下文中，流畅度等的变化对结果的影响并不大，而改写后模型出现幻觉错误等反而会导致模型问答能力下降。此外，模型将重新输入的自身改写文本视为普通上下文，缺乏明确的自我归属信号，因此，未能激发自我恒定性。

（2）为何角色提示词能够提高模型的开放域问答能力？传统的语言模型都不存在模板的概念，而是直接将问题以及相关文档输入模型，端到端地生成回复。目前流行的语言模型在海量语料进行预训练后，紧随指令微调能力的训练以及人类对齐的强化学习训练。在指令微调阶段出现了模板的概念，也就是“角色”，有系统角色、用户角色以及助理角色。通常在进行多轮请求的过程中，助理角色能够正确回答用户的问题，因此将文档内容放在助理角色中，会让模型对这部分内容更加关注，注意力分数更大。为了观察是否仅在指令微调和强化学习对齐训练后才出现该现象，本文选择预训练后的base版本（未经微调及强化学习Pre-SFT&RLHF）模型以及instruct版本（经过微调及强化学习Post-SFT&RLHF）模型进行对照实验。Qwen2-7b模型在指令微调以及人类对齐的强化学习训练前后问答性能变化如**图4**所示。从**图4**可以看出，在3个数据集上，经过指令微调和强化学习对齐之后的模型表现明显优于base版本。可见，Qwen系列模型在多个数据集上的自我恒定性表现均明显强于ChatGPT和Llama2等instruct模型。对此，本文提出若干初步的推测性解释。首先，从指令微调角度看，Qwen在技术报告中强调了更大规模、更高质量的多轮对话数据以及细粒度的系统/用户/助理数据，这类数据分布很可能促使模型在对话中形成更稳定的“角色意识”[43,55]，从而在面对隐式的自我归属提示时表现出更强的一致性偏好。其次，在对齐阶段，不同模型采用的基于人类反馈的强化学习（reinforcement learning from human feedback，RLHF）策略强度与目标设计并不一致，以更强的安全性与有用性对齐目标，尤其是对“给出清晰、

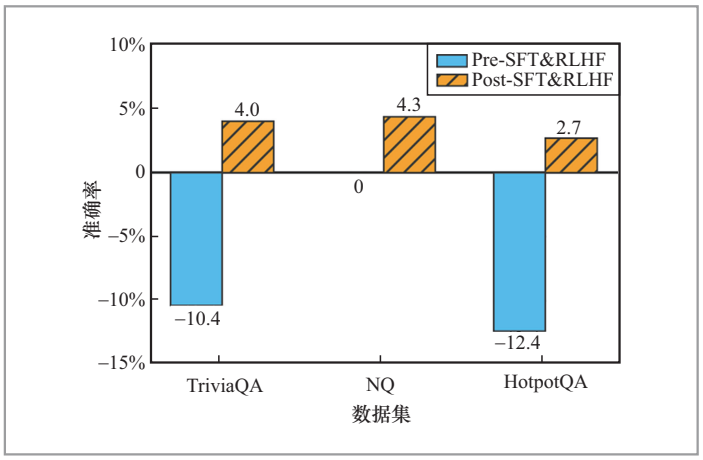


图4 Qwen2-7b模型在指令微调以及人类对齐的强化学习训练前后问答性能变化

肯定回答”的偏好，可能间接增强了模型对自身输出和自我归属内容的信任[56]。

5 结束语

本文系统性研究了LLM中的自我恒定性现象及其在知识增强任务中的应用价值。本文首次证实LLM具备类人的自我恒定性认知偏差。在隐式设置下，仅通过提示词诱导模型认为内容源自其自身，即可显著提升其置信度，将心理学中的自我一致性理论扩展至人工智能系统。激发自我恒定性的关键在于隐式归属而非显式生成。角色提示词利用模型的角色感知能力，使其对助理角色内容赋予更高注意力权重，而显式改写因语义损耗和缺乏归属信号而效果有限。自我恒定性在知识增强任务中具有显著实用价值。角色提示词策略在开放域问答任务上准确率提升，在高噪声环境下表现出卓越鲁棒性，有效增强了模型的抗干扰和知识整合能力。指令微调和RLHF是诱导自我恒定性的核心机制。对齐训练使模型形成对角色体系的深

层认知，对助理角色内容产生结构性偏好。本文在理论上深化了对大模型认知偏差的理解，在实践上为检索增强系统提供了简单、高效的优化方案。未来需要进一步探索如何平衡自我恒定性的优势与过度自信的风险，推动构建更加可靠、可控的人工智能系统。

参考文献：

[1] Cao Y H, Li S Y, Liu Y X, et al. A survey of AI-generated content (AIGC)[J]. ACM Computing Surveys, 2025, 57(5): 1-38.

[2] Wang Y T, Pan Y H, Yan M, et al. A survey on ChatGPT: AI-generated contents, challenges, and solutions[J]. IEEE Open Journal of the Computer Society, 2023, 4: 280-302.

[3] Spitale G, Biller-Andorno N, Germani F. AI model GPT-3 (dis) informs us better than humans[J]. Science Advances, 2023, 9(26): eadh1850.

[4] Koyejo S, Miranda B, Schaeffer R. Are emergent abilities of large language models a mirage [C]//Proceedings of the Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 55565-55581.

[5] Dai S H, Zhou Y Q, Pang L, et al. Neural retrievers are biased towards LLM-generated content[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024: 526-537.

[6] Dai S H, Liu W H, Zhou Y Q, et al. Cocktail: a comprehensive information retrieval benchmark with LLM-generated documents integration[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 7052-7074.

[7] Xu S C, Hou D Y, Pang L, et al. Invisible relevance bias: text-image retrieval models prefer AI-generated images[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2024: 208-217.

[8] Tan H X, Sun F, Yang W L, et al. Blinded by generated contexts: how language models merge generated and retrieved contexts when knowledge conflicts? [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 6207-6227.

[9] Mao C, Vondrick C, Wang H, et al. RAIDAR: generative AI detection via rewriting[PP]. arXiv preprint, 2024, arXiv: 2401.12970.

[10] Fatemi S H, Clayton P J. The medical basis of psychiatry[M]. Totowa: Humana Press, 2008.

[11] Wang S, Zhu Y C, Liu H C, et al. Knowledge editing for large language models: a survey[J]. ACM Computing Surveys, 2025, 57(3): 1-37.

[12] Kandpal N, Deng H K, Roberts A, et al. Large language models struggle to learn long-tail knowledge[C]//Proceedings of the 40th International Conference on Machine Learning. New York: ACM, 2023: 15696-15707.

[13] Ovadia O, Brief M, Mishaeli M, et al. Fine-tuning or retrieval? Comparing knowledge injection in LLMs[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 237-250.

[14] Min S, Krishna K, Lyu X X, et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation[C]//Proceedings of the

2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 12076-12100.

[15] Mishra A, Asai A, Balachandran V, et al. Fine-grained hallucination detection and editing for language models[PP]. arXiv preprint, 2024, arXiv: 2401.06855.

[16] Yoran O, Wolfson T, Ram O, et al. Making retrieval-augmented language models robust to irrelevant context[C]// Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). Vienna: OpenReview. net, 2024.

[17] Yu W H, Zhang H M, Pan X M, et al. Chain-of-note: enhancing robustness in retrieval-augmented language models[C]// Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 14672-14685.

[18] Joshi M, Choi E, Weld D, et al. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association forComputational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2017: 1601-1611.

[19] Kwiatkowski T, Palomaki J, Redfield O, et al. Natural questions: a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 453-466.

[20] Yang Z L, Qi P, Zhang S Z, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 2369-2380.

[21] Thorne J, Vlachos A, Christodoulopoulos C, et al. FEVER: a large-scale dataset for fact extraction and VERification[C]//Proceedings of the 2018 Conference of the North American Chapter Of the Association for Computational Linguistics: Human LanguageTechnologies, Volume 1 (Long Papers). Stroudsburg: ACL, 2018: 809-819.

[22] Levy O, Seo M, Choi E, et al. Zero-shot relation extraction *via* reading comprehension[C]//Proceedings of the 21st Conference on Computational Natural LanguageLearning (CoNLL 2017). Stroudsburg: ACL, 2017: 333-342.

[23] Foo L G, Rahmani H, Liu J. AI-generated content (AIGC) for various data modalities: a survey[J]. ACM Computing Surveys, 2025, 57(9): 1-66.

[24] Wang T, Zhang Y S, Qi S R, et al. Security and privacy on generative data in AIGC: a survey[J]. ACM Computing Surveys, 2025, 57(4): 1-34.

[25] Wadden D, Lin S C, Lo K, et al. Fact or fiction: verifying scientific claims[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2020: 7534-7550.

[26] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 2641-2649.

[27] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Computer Vision-ECCV 2014. ChamSpringer International Publishing 2014: 740-755.

[28] OpenAI, Achiam J, Adler S, et al. GPT-4 technical report[PP]. arXiv preprint, 2023, arXiv: 2303.08774.

[29] Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models[PP]. arXiv preprint, 2023, arXiv: 2307.09288.

[30] Yin Z Y, Sun Q S, Guo Q P, et al. Do large language models know what they don't know [C]//Proceedings of the

Findings of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 8653-8665.

[31] Jiang Z B, Xu F F, Araki J, et al. How can we know what language models know[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 423-438.

[32] Katz D M, Bommarito M J, Gao S, et al. GPT-4 passes the bar exam[J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2024, 382(2270): 20230254.

[33] Mielke S J, Szlam A, Dinan E, et al. Reducing conversational agents' overconfidence through linguistic calibration[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 857-872.

[34] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[C]// Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023). Kigali: OpenReview.net, 2023.

[35] Ye J, Wang Y, Huang Y, et al. Justice or prejudice quantifying biases in LLM-as-a-Judge[C]//Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025). Singapore: OpenReview.net, 2025.

[36] Yona G, Aharoni R, Geva M. Can large language models faithfully express their intrinsic uncertainty in words [C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 7752-7764.

[37] Kuhn L, Gal Y, Farquhar S. Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation[C]//Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023). Kigali: OpenReview.net, 2023.

[38] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Virtual: Curran Associates, Inc., 2020.

[39] Ram O, Levine Y, Dalmedigos I, et al. In-context retrieval-augmented language models[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1316-1331.

[40] Guu K, Lee K, Tung Z, et al. Retrieval augmented language model pre-training[C]// Proceedings of the 37th International Conference on Machine Learning (ICML 2020). Virtual: PMLR, 2020: 3929-3938.

[41] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens[C]//Proceedings of the 39th International Conference on Machine Learning (ICML 2022). Baltimore: PMLR, 2022: 2206-2240.

[42] Wang F, Wan X C, Sun R X, et al. Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models[C]// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2025: 30553-30571.

[43] Ren M J, Cao B X, Lin H Y, et al. Learning or self-aligning? Rethinking instruction fine-tuning[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 6090-6105.

[44] 孙浩然, 王志豪, 吴一帆, 等. 基于重排序和后检索反思的教育大模型问答增强方法[J]. 大数据, 2025, 11(5): 4-17.
Sun H R, Wang Z H, Wu Y F, et al. an enhanced question answering method for educational large language models based on reranking and post-retrieval reflection[J]. Big Data Research, 2025, 11(5):

4-17.

[45] 吴泽贤, 张琰彬. 基于隐性意图检测与检索模型的对话生成增强框架[J]. 大数据, 2026: 2026025.
Wu Z X, Zhang Y B. An enhanced dialogue generation framework based on implicit intent detection and retrieval models[J]. Big Data Research, 2026: 2026025.

[46] Laurito W, Davis B, Grietzer P, et al. AI-AI bias: large language models favor communications generated by large language models[J]. Proceedings of the National Academy of Sciences of the United States of America, 2025, 122(31): e2415697122.

[47] Manoranjan V, Gaikwad S S. When personas override payoffs: role identity bias in multi-agent LLM decision-making[PP]. arXiv preprint, 2026, arXiv: 2601.10102.

[48] Haez S G, Dragoni M. Neutral is not unbiased: evaluating implicit and intersectional identity bias in LLMs through structured narrative scenarios[C]// Findings of the Association for Computational Linguistics (EMNLP 2025). 2025: 15060-15088.

[49] Cochran W G. Sampling techniques[M]. 3rd ed. New York: Wiley, 1977.

[50] Cohen J. Statistical power analysis for the behavioral sciences[M]. London: Routledge, 2013.

[51] Bowyer S, Aitchison L, Ivanova D R. Position: don't use the CLT in LLM evals with fewer than a few hundred datapoints[PP]. arXiv preprint, 2025, arXiv: 2503.01747.

[52] Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2020: 6769-6781.

[53] Bai J Z, Bai S, Chu Y F, et al. Qwen technical report[PP]. arXiv preprint, 2023, arXiv: 2309.16609.

[54] Zhang M Z, Huang M Q, Shi R D, et al. Calibrating the confidence of large language models by eliciting fidelity[C]// Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 2959-2979.

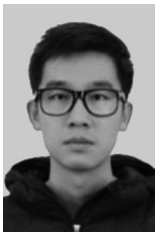
[55] Yang A, Li A, Yang B, et al. Qwen3 technical report[PP]. arXiv preprint, 2025, arXiv: 2505.09388.

[56] Agarwal S, Almeida D, Askell A, et al. Training language models to follow instructions with human feedback[C]// Proceedings of the Advances in Neural Information Processing Systems 35. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 27730-27744.

作者简介

雷艳（1999-），女，中国科学院计算技术研究所智能算法安全全国重点实验室、中国科学院大学计算机学院博士生，主要研究方向为自然语言处理、大语言模型、长文本问答。

庞亮（1990– ），男，博士，中国科学院计算技术研究所智能算法安全全国重点实验室副研究员，主要研究方向为信息检索、大语言模型。



魏子豪（1998– ），男，中国科学院计算技术研究所智能算法安全全国重点实验室、中国科学院大学计算机学院博士生，主要研究方向为大语言模型。

王元卓（1978– ），男，中国科学院计算技术研究所智能算法安全全国重点实验室研究员、博士生导师，主要研究方向为大语言模型、知识图谱。

沈华伟（1982– ），男，中国科学院计算技术研究所智能算法安全全国重点实验室研究员、博士生导师，主要研究方向为图神经网络、大语言模型。

程学旗（1971– ），男，中国科学院计算技术研究所智能算法安全全国重点实验室研究员、博士生导师，主要研究方向为图神经网络、大语言模型。

收稿日期: 2026-03-05

通信作者: 王元卓，wangyuanzhuo@ict.ac.cn；庞亮，pangliang@ict.ac.cn

基金项目: 国家自然科学基金资助项目（No.62172393）；河南省重大公益项目（No.201300311200）

Foundation Items: The National Natural Science Foundation of China (No.62172393), The Major Public Welfare Project of Henan Province (No.201300311200)