面向情境-记忆冲突的智能体协同知识融合技术

要 鑫[1] 宋美玲[1] 孙彬弘[2] 毕 鑫[2),3)] 赵相国[1]
张奥千[4] 李博扬[5] 袁 野[5]

[1)](东北大学软件学院 沈阳 110169)
[2)](东北大学计算机科学与工程学院 沈阳 110169)
[3)](东北大学深部金属矿智能开采与装备全国重点实验室 沈阳 110819)
[4)](北京理工大学计算机学院 北京 100081)
[5)](北京理工大学人工智能学院 北京 100081)

摘 要 DB4LLM 技术虽然可以有效地集成外部知识来解决大语言模型(Large Language Model,LLM)的局限性,但也经常出现自身固有参数化知识与外部检索知识相冲突的情形,即情境-记忆知识冲突(Context-Memory Conflicts)问题。这种冲突本质上体现为 LLM 在决策过程中对外部知识与内部记忆的置信度权衡。然而,当二者置信度接近时,现有的静态阈值或单一置信度策略容易导致 LLM 陷入不确定性决策区,出现回答不稳定、准确率大幅下降等问题。为有效应对该挑战,本文提出一种基于智能体的冲突消解框架(Agent Conflict Resolution Framework,ACR),实现了对外部知识与内部知识的置信度统一建模与动态融合。ACR 由两个智能体组成。置信度校准智能体(E-Agent)量化 LLM 生成的候选答案的不确定性,并结合轻量贝叶斯校准策略,将 LLM 内部记忆与外部检索知识映射到同一度量空间,提升置信度可比性与输出稳定性;知识融合智能体(K-Agent)则通过自适应权重解析函数,同时考虑置信度差值和反事实稳定性,实现对内外部知识融合权重的动态分配。在置信度接近的高不确定性决策区域,两个智能体通过反馈机制协同合作,实现答案的自我修正。在五个公开数据集上的大量实验结果表明,在置信度高度近似的情境-记忆知识冲突任务中,ACR 在不同基础大语言模型上相较于先进方法的平均性能提升为 6.08%,验证了该方法的有效性与稳定性。

关键词 DB4LLM;情境-记忆知识冲突;智能体协同;置信度校准;知识融合
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2026.01009

Agent Collaborative Knowledge Fusion for Context-Memory Conflicts

YAO Xin[1] SONG Mei-Ling[1] SUN Bin-Hong[2] BI Xin[2),3)] ZHAO Xiang-Guo[1]
ZHANG Ao-Qian[4] LI Bo-Yang[5] YUAN Ye[5]

[1)](Software College, Northeastern University, Shenyang 110169)
[2)](School of Computer Science and Engineering, Northeastern University, Shenyang 110169)
[3)](State Key Laboratory of Intelligent Deep Metal Mining and Equipment, Northeastern University, Shenyang 110819)
[4)](School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081)
[5)](School of Artificial Intelligence, Beijing Institute of Technology, Beijing 100081)

Abstract Although DB4LLM can effectively integrate external knowledge to solve the limita-

收稿日期:2025-07-19;在线发布日期:2026-01-30。本课题得到国家自然科学基金(重大项目 No. 62394332、青年科学基金项目(A 类)No. 62225203、联合基金 No. U23A20297)、河北省创新能力提升计划项目(No. 235A0101D)、深部金属矿智能开采与装备全国重点实验室自主研究项目(No. IDMEIRI2504)、辽宁省"兴辽英才计划"项目(No. XLYC2204005)、北京市联合基金项目(No. L241010)、教育部基础学科和交叉学科突破计划(No. JYB2025XDXM108)资助。**要 鑫**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为大模型。E-mail:yaoxin@stumail.neu.edu.cn。**宋美玲**,硕士研究生,主要研究领域为大模型。**孙彬弘**,硕士研究生,主要研究领域为大模型。**毕 鑫**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为大模型、大数据管理与分析等。**赵相国(通信作者)**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为大模型、大数据管理与分析等。E-mail:zhaoxiangguo@mail.neu.edu.cn。**张奥千**,博士,副研究员,中国计算机学会(CCF)会员,主要研究领域为大数据分析。**李博扬**,博士,中国计算机学会(CCF)会员,主要研究领域为大数据、人工智能、分布式系统。**袁 野**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为大数据管理与分析、人工智能等。

tions of Large Language Models (LLM), it also often has the problem of conflict between its own intrinsic parameterized knowledge and external retrieved knowledge, which is called context-memory conflict. The conflict is essentially a confidence trade-off between the internal and external knowledge of LLM in the decision-making process. However, when the two confidence levels are close to each other, the existing static threshold or single confidence level strategy may easily cause LLM to fall into the uncertainty decision zone, which may lead to unstable answers and a significant decrease in accuracy. To address this challenge, this paper propose Agent Conflict Resolution Framework (ACR), which realizes the unified modeling and dynamic fusion of the confidence levels of external knowledge and internal memory. ACR consists of two agents. The Evidence-Confidence Agent (E-Agent) quantifies the uncertainty of candidate answers generated by LLM and combines with a lightweight Bayesian calibration strategy to map the internal memory and external retrieval knowledge of the LLM into the same metric space, which improves the comparability of confidence and output stability; the Knowledge-Fusion Agent (K-Agent), on the other hand, realizes dynamic allocation of internal and external knowledge fusion weights through continuously derivable weight resolution function, taking confidence difference and counterfactual stability into account. In the region of high uncertainty where the confidence levels are close, the two agents collaborate to realize the self-correction of the answer through the feedback mechanism. Extensive experimental results across five public datasets demonstrate that ACR achieves an average performance improvement of 6.08% over state-of-the-art methods on scenario-memory knowledge conflict tasks with highly similar confidence levels, validating the method's effectiveness and stability.

Keywords database for large language models; context-memory knowledge conflict; multi-agent collaboration; confidence calibration; knowledge fusion

1 引 言

近年来,大语言模型(Large Language Model, LLM)[1-2]因其强大的泛化能力与出色的语义理解能力,迅速成为人工智能领域的研究焦点。LLM 在预训练期间会将大量的事实知识封装在他们的参数中作为内部记忆,称为参数化知识[3]。然而,其固有的参数化知识存储机制会导致两大认知缺陷[4]:(1)知识更新滞后性[5]:模型参数固化后无法主动追踪动态演化的领域知识,如不断修订的金融政策;(2)知识覆盖有限性[6]:训练数据的长尾分布特性导致低频知识易被参数空间压缩。这种缺陷使得 LLM 容易产生"幻觉"现象,即在缺乏有效知识约束时可能输出与事实不符或误导性的内容。

为了缓解 LLM 的幻觉问题,DB4LLM(Database for Large Language Model)技术应用而生,也称检索增强生成(Retrieval-Augmented Generation,RAG)。该技术旨在通过检索高质量的外部知识库(如向量数据库、知识图谱)来为 LLM 提供更准确的事实支撑,从而提升 LLM 生成内容的可靠性[7]。然而,这种 DB4LLM 的协同架构也引发了外部检索知识(情境知识)与 LLM 内部参数化知识(内部记忆)之间的"情境-记忆知识冲突",即当 LLM 检索的外部知识与模型内部参数存储的知识存在显著差异甚至矛盾时,LLM 在决策过程中会陷入困境[8-9]。情境-记忆知识冲突增加了 LLM 偏见决策或错误推理的风险。如图 1(a)所示,对于"OpenAI 的现任 CEO 是谁"这个问题,LLM 本身的内部记忆("Mira Murati")与外部检索知识("Sam Altman")是不一致的,这会导致 LLM 面临复杂的抉择:是应该相信外部最新、更为权威但可能尚未普遍推广的知识,还是继续依赖自身参数中广泛接受但已逐渐过时的旧知识?

针对这一问题,现有研究主要有三类方法:(1)忠于记忆:即更偏向于相信 LLM 自身参数内蕴含的知识,以保持模型决策的独立性与稳定性,但该方法无法有效应对 LLM 内部知识的过时或偏差问题[10];(2)忠于情境:即无条件接受外部知识库知识以确保响应的准确性,然而这种方法容易过度依赖

外部知识源,一旦外部知识源存在误差或过时情况则可能适得其反[11];(3)情境-记忆混合:通过评估内外知识源的置信度,动态决策何时依赖外部知识库、何时信任内部参数知识,这种策略虽然灵活,却过度依赖人为设置的置信度阈值,难以适应任务、领域与时间的快速变化[12]。尽管上述方法尝试从不同角度解决情境-记忆知识冲突,但仅通过单一置信度或静态阈值来决定采信内外知识未必准确。当内外部知识置信度差异较大时,LLM 可以自信稳定地做出相应决策。但是当内外部知识置信度差异高度近似时(如图 1(b)所示),阈值裁决几乎退化为随机猜测,LLM 陷入了极度的不确定性,导致准确率骤降,同时结果也无法自我修正。换言之,当模型内外知识置信度接近且产生冲突时,如何有效地判定、融合或选择最优知识来源,即"不确定性下的知识融合"是当前研究亟待突破的关键难题。

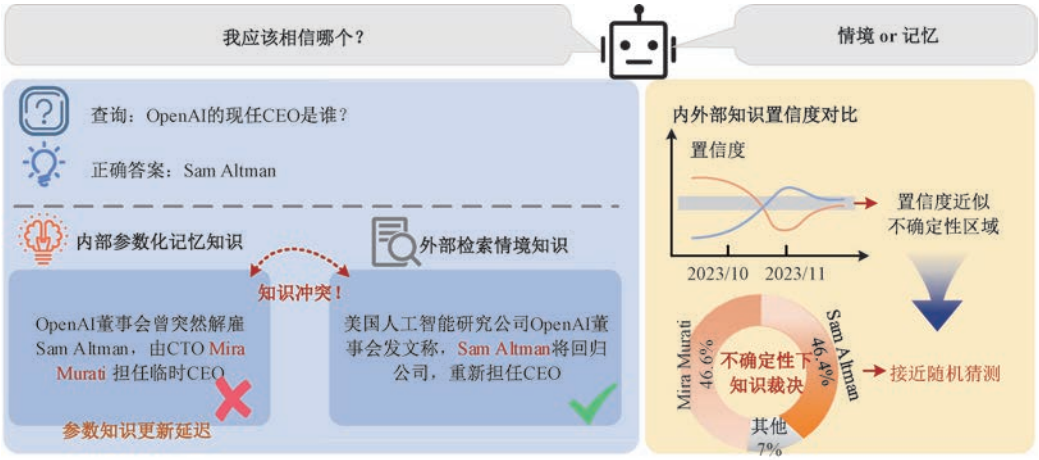

图 1　情境-记忆知识冲突示例图

为了解决上述问题,本文提出一种基于智能体的冲突消解框架(Agent Conflict Resolution Framework,ACR)。ACR 通过置信度校准智能体(Evidence-Confidence Agent,E-Agent)和知识融合智能体(Knowledge-Fusion Agent,K-Agent)协同合作,动态融合 LLM 内部固有知识与外部检索知识,提升 LLM 在两者置信度高度近似的知识冲突场景下的生成准确性。具体而言,E-Agent 并行提取 LLM 基于内部参数化知识与外部检索知识生成的多源答案,建模其不确定性,并设计了轻量化贝叶斯校准策略,将两类知识在统一概率度量空间内建模为结构化置信度分布。K-Agent 则通过自适应的权重解析函数,动态分配内外部知识的权重。若内外部知识置信度高度近似或生成答案不稳定时,K-Agent 触发反馈机制,与 E-Agent 协同合作,迭代请求补充知识,并完成答案自我修正,实现端到端的冲突消解。本文的贡献点主要包括以下 4 点:

(1)提出了一种基于智能体的冲突消解框架 ACR,通过智能体协同合作,有效地解决了 LLM 在情境-记忆知识冲突中面临的内外知识权重动态分配难题,尤其在内外知识置信度高度近似的不确定性决策区表现出明显的生成优势。

(2)置信度校准智能体 E-Agent 量化 LLM 生成的候选答案的不确定性,并将 LLM 内部记忆与外部检索知识映射到同一度量空间,提升了置信度可对比性与生成稳定性。

(3)知识融合智能体 K-Agent 同时考虑内外部知识的置信度差异与生成答案的反事实稳定性,设计自适应权重解析函数,实现了置信度高度近似时知识融合权重的平滑分配,避免硬阈值策略的随机性和不稳定性。

(4)在 5 个公开数据集上进行了大量实验,实验结果表明 ACR 在存在知识冲突(尤其是置信度高度近似)时的复杂问答任务中性能显著优于 SOTA 方法。

本文第 2 节概括介绍当前应对 LLM 知识冲突的相关研究进展及挑战;第 3 节形式化定义本文相关问题,包括情境-记忆知识冲突判定机制、LLM 生成不确定性以及冲突消解任务、反事实稳定性;第 4 节详细介绍本文提出的 ACR 框架设计思路和关键模块,包括智能体的设计细节及协同机制;第 5 节通过大量实验验证了所提框架的有效性;最后总结全文。

2　相关工作

知识冲突最初源自开放域问答(Open-Domain

Question Answering,ODQA)的研究。Longpre 等人[13]首次提出了知识冲突(Knowledge Conflict)这一概念,并重点分析了实体层面参数化知识与外部检索段落之间的矛盾。随后,多段落之间的知识差异也成为了研究关注的焦点。近年来,随着大语言模型(LLM)的迅猛发展,以及 DB4LLM 技术的兴起,知识冲突问题再次引起研究界的极大兴趣[14]。近期有研究表明,当外部检索知识与 LLM 的内部参数化知识发生冲突时,若外部检索知识较为完美时,LLM 更倾向于接受外部知识,导致其内部知识被不当修改[15]。但是,随着 LLM 参数规模的扩大及训练知识的丰富,LLM 又逐渐表现出对内部参数知识的高度依赖。此外,当外部检索知识同时包含支持 LLM 内部知识的内容时,LLM 可能出现明显的选择偏差,倾向于与自身记忆一致的知识,不论其正确性如何[16]。这种现象严重影响了 LLM 的置信度、实时准确性和稳健性。因此,深入理解和有效解决知识冲突问题成为当前研究的重要任务。

现有知识冲突解决方法可归纳为以下三类[17]:

(1)情境-记忆知识冲突(Context-Memory Conflict)

在实际应用中,用户可能提供补充提示,LLM 也可能通过外部知识库或者搜索引擎检索额外知识。这些提示、对话历史与检索文档共同构成了“情境知识”,当情境知识与模型内部参数化知识产生不一致时,即发生情境-记忆知识冲突。

(2)情境间冲突(Contextual Conflict)

现实场景中,外部检索的知识可能包含噪音甚至蓄意的误导信息,造成多个外部来源相互矛盾,给 LLM 准确理解和应用外部知识带来巨大挑战。

(3)记忆内冲突(Intra-Memory Conflict)

LLM 的参数化知识源自复杂多样的训练语料,其中存在的不一致性可能导致模型在面对语义相同但表述不同的问题时,产生相互矛盾的回答。

本文重点研究最为广泛关注的情境-记忆知识冲突。针对情境-记忆知识冲突,现有解决方法主要分为三类:

(1)忠于记忆的方法

该类方法鼓励模型谨慎对待外部检索知识,更加依赖内部参数化知识以防止被错误信息误导[18]。近年来已有研究从提示工程、查询增强以及微调等方向提出多种方法,以增强 LLM 在生成过程中的忠实性与抗干扰能力。例如,Pan 等人[19]提出了采用虚假信息检测与谨慎提示相结合的方式,要求模型主动识别和规避不可靠知识,引导 LLM 在面临潜在虚假语境时保持对内部参数知识的忠诚。Xu 等人[20]引入了记忆核查型提示,通过提示工程提醒 LLM 在作答前对自身记忆进行验证,并警示可能存在的信息干扰。Weller 等人[21]提出利用语料库中的信息冗余性来抵御虚假信息污染。他们通过设计查询增强机制,交叉验证多个来源答案的一致性以判断其可靠性,减轻知识冲突。此外,Hong 等人[22]微调了一个较小的 LMasa 鉴别器,并整合了提示工程,使模型能够区分可靠和不可靠的信息。

(2)忠于情境的方法

该类方法强调遵从外部知识以提高回答的准确性。Zhou 等人[23]则研究了基于提示策略的冲突消解方法,利用基于观点的提示和反事实提示来提高 LLM 对外部知识的依从性,而无需额外培训。Shi 等人[24]通过外部知识感知解码技术(CAD)放大了有外部知识和没有外部知识的输出概率差异,将相关外部知识置于内部记忆知识之上。Zhang 等人[25]则提出一种事实有效性预测方法,识别并丢弃 LLM 中过时的事实,确保 LLM 遵循最新的外部知识。Huang 等人[26]提出一种 PIP-KAG 方法,通过修剪 LLM 的内部参数知识以帮助 LLM 更好地利用外部知识。

(3)混合策略

该类方法则尝试结合外部检索知识和内部固有的参数化知识,以获得更加稳健和准确的答案。Li 等人[27]提出知识感知微调策略 KAFT,通过在训练过程中引入反事实和不相关外部知识微调大模型,以增强其可控性和稳健性。Zhang 等人[28]提出了 COMBO 框架,通过判别器评估外部检索知识与模型生成知识的兼容性。Jin 等人[29]提出了一种基于对比解码的算法,以最大限度地缩小知识冲突的差异并校准。Wang 等人[30]则提出了 ASTUTE RAG 框架,自适应地从 LLM 的内部知识中获取重要信息,迭代地将内部和外部知识相结合得到最终确定答案。Bi 等人[31]提出了 CK-PLUG,通过调整具有负置信度增益标记知识的概率分配,从而控制 LLM 对内部记忆的参数知识和外部检索知识的偏好。

尽管上述方法各有侧重,在 LLM 知识冲突解决方面取得了初步进展,但大多数方法仍然过度依赖经验阈值以及固定的融合策略,难以在内外知识置信度高度近似时做出稳健决策。因此,为应对不确定性条件下的情境-记忆冲突问题,本文提出一种基于智能体的冲突消解框架,实现内外部知识的动态权重分配与知识融合,有效提升了 LLM 在复杂

决策场景中的稳定性与适应能力。

3　问题定义

本章节围绕情境-记忆冲突场景,依次给出核心概念的形式化定义,包括情境-记忆知识冲突的刻画、LLM 的生成不确定性描述、冲突消解任务的建模以及反事实稳定性定义,为后续方法设计奠定理论基础。表 1 对本文常用的符号表示及其对应含义进行汇总并作简要说明。

表 1　符号表示及含义

符号	含义	符号	含义
K_{int}	内部知识	p	答案置信度
K_{ext}	外部检索知识	S	反事实稳定率
A	LLM 生成的答案	M	采样次数
Q	查询问题	T	温度
U	LLM 输出不确定性	w	知识融合权重

3.1　情境-记忆知识冲突

给定一个大语言模型(LLM)和用户查询 Q,LLM 的知识来源可划分为内部固有参数化知识和外部检索知识两类。本文将固化于模型参数中的内部知识记为 K_{int},外部检索得到的知识记为 K_{ext}。LLM 利用内部知识可生成候选答案 $A_{int}=LLM(Q,K_{int})$,而基于外部知识可生成候选答案 $A_{ext}=LLM(Q,K_{ext})$。

定义 1. 情境-记忆知识冲突. 定义语义冲突判别函数:

$$\delta(A_{int},A_{ext})=\begin{cases}1, & A_{int}\perp A_{ext}\\ 0, & \text{其他}\end{cases} \tag{1}$$

其中,$A_{int}\perp A_{ext}$ 表示 A_{int} 和 A_{ext} 在语义上存在实质性矛盾或互斥。当且仅当 $\delta(A_{int},A_{ext})=1$ 时,情境-记忆知识冲突发生。

3.2　LLM 的生成不确定性

在 DB4LLM 架构中,大语言模型(LLM)的答案生成不仅依赖于查询 Q 本身,还同时受到其内部固有参数化记忆 $x\in X$ 和外部检索知识 $d\in D$ 的综合影响。我们将 LLM 答案生成过程形式化为:给定 (Q,x,d),LLM 输出序列 $A=\{y_1,y_2,\cdots,y_N\}$。

为刻画 LLM 在融合内外部知识条件下的输出不确定性,我们采用条件熵作为度量指标[32]。具体地,在第 n 个位置,已知上下文前缀 $y_{<n}$,LLM 在条件 (Q,x,d) 下对候选词 $v\in V$ 的概率分布为 $p(y_n=v\mid y_{<n},Q,x,d)$,其对应的条件熵定义为

$$H_n=-\sum_{v\in V}p(y_n=v\mid y_{<n},d,x,Q)\cdot \log p(y_n=v\mid y_{<n},d,x,Q) \tag{2}$$

将所有位置的条件熵求平均,即可得到 LLM 在输出序列时的生成不确定性:

$$U(A)=\frac{1}{N}\sum_{n=1}^{N}H_n \tag{3}$$

$U(A)$ 越高,表明 LLM 在多个候选词间难以抉择,即输出存在较强的不确定性。

然而,该定义并未区分不确定性的不同来源。为此,本文进一步给出两种互补的信息量度量:冲突信息量 I_c 和补充信息量 I_s。冲突信息量 I_c 用于衡量内部知识与外部检索知识之间的冲突强度,I_c 越大说明外部检索知识与内部知识冲突强度越高,会增加输出的不确定性。补充信息量 I_s 用于衡量外部检索知识对当前内部知识状态的增益,I_s 越大说明外部检索知识携带了显著的新信息,对生成过程有较高补充价值,会降低输出的不确定性。鉴于两种度量在不确定性上的作用方向相反,本文将其统一纳入一个具有可解释性的表达形式中,即采用两者差值的绝对值(即 $|I_c-I_s|$)作为衡量内部与外部知识差异的指标。随后将该差值映射为最终的不确定性度量,用以近似刻画 LLM 在知识融合过程中的生成不确定性:

$$U(A)\propto \exp(|I_c-I_s|) \tag{4}$$

其中,$\propto$ 表示生成不确定性 U 与冲突-补充信息差值 I_c-I_s 的响应函数成正比。$\exp(\cdot)$ 表示 LLM 输出分布对内外信息差异的敏感性函数。

公式(4)并非从公式(2)与(3)通过严格数学推导得到的,而是一种合理的近似表达[33],其目的是在内部知识与外部知识置信度差异较小时,为 LLM 的不确定性决策行为提供一个直观解释。在这种近似框架下,当 $I_c\gg I_s$ 或 $I_s\ll I_c$ 时,LLM 能基于明确信号偏向一侧信息源,不确定性较低。当 $I_c\approx I_s$(即 $|I_c-I_s|\leqslant\theta$)时,内部知识与外部知识都能给出高概率答案,内外知识置信度近似,导致模型对“该信哪边”难以形成偏好,最终生成的输出就呈现熵增高、生成波动性上升的现象,即陷入不确定性决策区[34]。因此,该区域应被识别为知识融合过程中的高风险区,需设计知识融合机制以抑制 LLM 输出波动,提升稳定性。

3.3　冲突消解任务

定义 p_{int} 和 p_{ext} 分别表示 LLM 对 A_{int}、A_{ext} 的置信度估计。

定义 2. 冲突消解. 在存在情境-记忆知识冲突的条件下,设计一种机制 $F(\cdot)$,对 A_{int} 和 A_{ext} 进

行动态权重分布，生成融合后的最终答案 A_{fusion}。最终答案 A_{fusion} 需满足以下条件：(1)可靠性最大化：最大程度依赖更可信的知识源；(2)稳定性增强：在置信度接近(即 $|I_c - I_s| \leqslant \theta$)时，避免输出剧烈波动。

本文将冲突消解任务定形式化建模为

$$A_{fusion} = F(A_{int}, A_{ext}, I_c, I_s, U, Q) \tag{5}$$

其中，U 为融合后的生成不确定性估计，用于辅助判断是否触发重检索。

3.4 反事实稳定性

为刻画答案的鲁棒性，本文给出反事实稳定性的形式化定义。对给定查询 Q，内部知识 K_{int}，外部知识 K_{ext}，设决策函数 f 输出答案分布 P：

$$P = f(Q, K_{int}, K_{ext}) \tag{6}$$

本文通过在关键前提上进行否定或实体替换等操作构建合法的反事实干预集合 $\mathcal{K}$，其输入 LLM 后的输出答案分布为

$$P^k = f(Q, K_{int}^k, K_{ext}^k) \tag{7}$$

由此定义反事实稳定率 $S(Q)$：

$$S(Q) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} 1[\text{argmax } p = \text{argmax } P^k] \tag{8}$$

其中，$S(Q) \in [0,1]$，表示在所有受限干预下最优答案保持不变的比例。当 $S(Q)$ 在统计意义上已显著接近 1 时，可认为该决策满足反事实稳定性。

4 智能体冲突消解框架 ACR

基于智能体的冲突消解框架(Agent Conflict Resolution Framework，ACR)如图 2 所示。该框架由两个智能体协同合作：置信度校准智能体(Evidence-Confidence Agent，E-Agent)负责从 LLM 内部知识与外部检索知识中提取、量化并统一建模各自的置信度分布，用于建模生成答案的可信程度与不确定性风险；知识融合智能体(Knowledge-Fusion Agent，K-Agent)则以置信度校准结果为输入，结合反事实稳定性检测结果，构建自适应权重解析函数，实现对内外部知识权重的动态分配与知识融合。两者之间通过反馈机制形成相互协同，当 K-Agent 识别出内外部知识置信度高度近似或答案生成不稳定时，可向 E-Agent 发起补充知识请求，实现答案生成的自我修正。接下来本文将详细介绍关键模块的设计细节与交互机制。

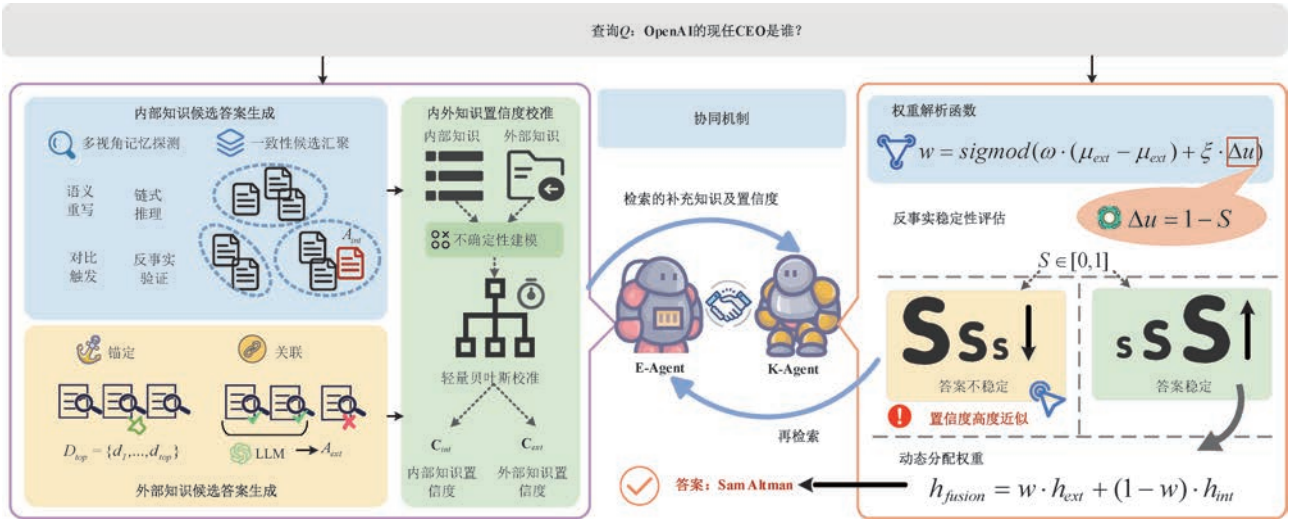

图 2 ACR 框架图

4.1 置信度校准智能体 E-Agent

置信度校准智能体 E-Agent 旨在统一建模 LLM 内部知识与外部检索知识的置信程度，为后续知识融合智能体 K-Agent 提供结构化的置信度表示。首先，E-Agent 从两个来源并行提取候选答案：(1)基于内部固有知识生成的候选答案 A_{int}；(2)基于外部检索知识生成的候选答案 A_{ext}。然后，E-Agent 进一步对候选答案的生成不确定性进行量化，并通过轻量贝叶斯校准将内部与外部知识的置信度分布归一至同一置信度空间，以实现跨源置信度的可比性并提升生成稳定性。

4.1.1 内部知识候选答案生成

对于内部知识候选答案生成，共包含多角度提示采样和一致性候选答案生成两个阶段。

(1)多角度提示采样

为了最大化内部知识空间的可达范围，并为后续不确定性估计提供充分样本，E-Agent 在接收查询 Q 后首先执行多角度提示采样。

本文从“语义重写-链式推理-反事实验证-对比触发”四个角度构造提示模板簇$\{t^{(k)}\}_{k=1}^{4}$,如图 3 所示。每个模板均以占位符 Q 接收查询。其中,语义重写[35]需要实现同义改写以及句式变换,提高表达的多样性。链式推理[36]要求 LLM 先思考,后回答,显示中间逻辑。反事实验证[37]将查询转写为判断题或是非题,使 LLM 进行自我验证,给出真假判定与理由。对比触发[38]会通过候选实体集合扩展的方式构造对照实体,并引导 LLM 在回答时明确指出查询实体与对照实体的差异并给出理由,以增强 LLM 建模语义边界,提升答案在事实判别上的鲁棒性与可解释性。

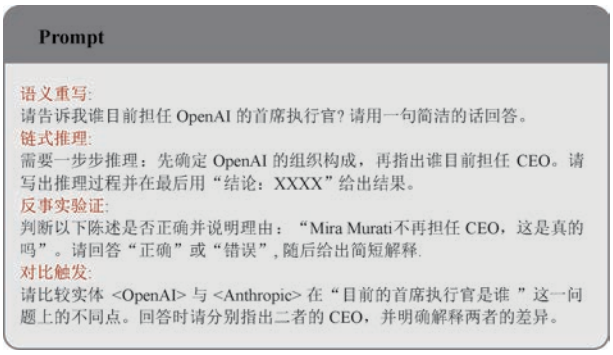

图 3　多角度提示采样示例

提示模板簇从不同角度激发 LLM 的内部知识检索潜能,确保生成过程在语义层面具备足够的多样性。为了进一步丰富输出的内容与知识粒度,我们在解码阶段引入随机化采样机制。具体而言,在相同提示模板与输入问题下,模型会进行 M 次独立采样,从而生成多份具有差异性的候选答案。这些答案构成候选池 $A_{raw}=\{A_{int}^{m}\}_{m=1}^{M}$,其中 A_{int}^{m} 为

$$A_{int}^{m}=LLM(Q;t;T;p),m=1,\cdots,M \quad (9)$$

T 为温度参数,p 为核采样阈值,t 为提示模板,Q 为输入的问题。

E-Agent 通过对查询 Q 进行多视角探测以及多次采样,能够充分检索 LLM 内部不同记忆知识片段,使暗藏的冲突、盲区在早期即被暴露。

(2)一致性候选答案生成策略

接着,E-Agent 通过语义归并选择唯一的内部主候选答案 A_{int}。具体的,首先对答案候选池 A_{raw} 中候选答案进行同义聚类,然后选择规模最大的语义簇,并在该簇中挑选平均对数概率最高的答案作为最终内部候选答案 A_{int}。

4.1.2　外部检索知识候选答案生成

对于外部知识候选答案生成,E-Agent 通过锚定检索和知识关联两个阶段将外部检索知识同化为候选答案 A_{ext}。

(1)锚定检索

在锚定检索阶段,E-Agent 首先对查询 Q 进行语义分解,得到问句实体集合 e_Q、关系集合 r_Q 以及谓词分析等生成若干子查询。然后针对外部知识库进行检索,得到证据集合 $D=\{d_1,d_2,\cdots,d_n\}$。最后会对检索出的外部知识片段消除冗余,得到精简证据集 D_{top}。

(2)知识关联

在知识关联阶段,E-Agent 把锚定检索阶段获得的原始证据 D_{top} 同化为一条答案 A_{ext}。

首先,对每条待处理证据 d_i 向 LLM 发送抽取提示,模板设计如下:“你是一名信息抽取助手,请从下列文本中提取与问句最相关的事实三元组,格式严格为〈s〉主语/谓语/宾语〈/s〉。文本:{d_i};问句:{Q}。”该提示保证输出的三元组集合可由正则表达式稳定解析。

考虑到 LLM 在抽取过程中可能产生幻觉,我们引入验证机制以提高结果可信度。针对每个候选三元组〈s,r,o〉发送语义对齐验证提示:“问句实体集合:{e_Q};问句关系短语:{r_Q};候选三元组〈s,r,o〉。请回答该三元组是否同时包含所有实体且谓语与关系短语同义? 仅回答‘yes’或‘no’。”LLM 若返回 yes,则保留该三元组进入候选集;否则丢弃,删除不匹配项。对多个候选三元组进行冗余抽取与合并,并使用投票策略选出最一致的核心事实。针对长文档或大规模知识库场景,为保证方法的可扩展性,我们可采用分块处理与并行化抽取的策略。最后,将经过验证与融合的三元组集合作为结构化输入,向 LLM 发送生成提示:“请依据以下事实三元组用自然语言回答问题,不增加额外信息。”LLM 输出的陈述将作为答案 A_{ext}。

4.1.3　内外知识置信度校准

为衡量这两类答案的置信度,本文设计置信度估计和不确定性建模两阶段校准流程。

首先,本文对每类答案计算初始置信度分数,采用对数似然形式的点估计:

$$p_{int}=\frac{1}{N}\sum_{n=1}^{N}\log p(y_n^{int}\mid y_{<n}^{int},Q),$$

$$p_{ext}=\frac{1}{N}\sum_{n=1}^{N}\log p(y_n^{ext}\mid y_{<n}^{ext},Q,D_{top}) \quad (10)$$

其中,y_n^{int} 和 y_n^{ext} 分别表示基于内部知识与外部知识生成的第 n 个 token,N 为生成长度。该指标反映了 LLM 对生成序列的整体置信程度。

但仅有点估计不足以全面反映输出的可靠性。这是由于 LLM 可能对某些错误答案给出高度自信的答案生成，尤其在受到幻觉干扰时，点估计会掩盖潜在的不确定性风险。

为了刻画生成答案的不确定性，本文设计了基于置信度的评估策略。首先 E-Agent 对 4.1.1 节和 4.1.2 节最终选定的候选答案 A_{int} 和 A_{ext} 进行 M 次采样，得到每条候选答案的置信度样本表示序列 $\{\hat{p}^{(1)},\hat{p}^{(2)},\cdots,\hat{p}^{(M)}\}$ 后，计算其样本均值与方差：

$$p=\frac{1}{M}\sum_{i=1}^{M}\hat{p}(i),$$

$$\sigma_p^2=\frac{1}{M-1}\sum_{i=1}^{M}(\hat{p}^{(i)}-p)^2 \tag{11}$$

其中，p 和 σ_p^2 仍可能受小样本波动影响。因此，本文进一步引入轻量贝叶斯校准，对置信度均值和方差构成的向量 (p,σ_p^2) 进行正则化调整，以获得统计稳健的置信度表示。具体地，本文将 p 映射至 log-odds 空间 $\ell=\log\frac{p}{1-p}$，并假设采样过程可由高斯-Gamma 先验生成：

$$\tau\sim Gamma(\omega,\xi),\ell\mid\tau\sim\mathcal{N}(\bar{\ell},1/\tau) \tag{12}$$

其中，$(\omega,\xi)=(1,1)$ 为弱信息先验。令 $\bar{\ell}$ 与 σ_ℓ^2 分别为 log-odds 样本均值、方差，则其后验精度期望为

$$\hat{\tau}=\omega+\frac{M}{2}/\left(\xi+\frac{M\sigma_\ell^2}{2}\right) \tag{13}$$

据此可得 log-odds 域的校准方差：

$$Var[\ell]=\frac{1}{(M+1)\hat{\tau}} \tag{14}$$

最后，将校准结果映射回概率空间：

$$\mu=\mathrm{sigmoid}(\bar{\ell})$$

$$\sigma=\sqrt{\mu(1-\mu)Var[\ell]} \tag{15}$$

于是可得校准后置信度表示：

$$c_{int}=[\mu_{int},\sigma_{int}],c_{ext}=[\mu_{ext},\sigma_{ext}] \tag{16}$$

其中，μ 为知识源的平均置信强度，σ 为刻画估计方差的稳定性。该校准过程使置信度估计随有效样本量自适应调整，有助于鉴别“均值相近而结构不稳”的高风险输出。

4.2 知识融合智能体 K-Agent

知识融合智能体（K-Agent）承担 ACR 框架中的核心角色，其目标是在基于内部知识生成的答案 A_{int} 和基于外部检索知识生成的 A_{ext} 存在情境-记忆知识冲突时，根据其置信度和稳定性评估生成动态融合权重，从而实现更精确的知识融合，使 LLM 生成更可靠的一致性答案。其包括两个关键组件：反事实稳定性评估和权重解析函数构造。

4.2.1 反事实稳定性评估

为了提升答案的可靠性，本文提出了反事实稳定性评估机制。该机制旨在揭示模型潜在的脆弱关联和假性相关，即判断“若事实稍有改动，结论是否仍然成立”，从而避免出现“微扰即翻转”的风险。此外，当外部知识存在噪声或错误时，E-Agent 的置信度校准可能受到干扰，导致判断结果不可靠。反事实稳定性评估机制在此情况下能够发挥缓解作用，通过识别答案是否对局部异常证据过度敏感，并在必要时触发再检索与再融合，从而提高系统在低质量外部知识下的稳定性。

具体实现上，该机制首先基于当前生成的中间答案 A^0，构造一组关键前提上的反事实假设，即语义否定（对命题型前提出现与非）、实体替换（将主体替换为同语义类别中不同但合逻辑的实体）和条件背景扰动（如改变事件发生时间等），如图 4 所示。随后，利用 LLM 判断反事实前提是否会改变答案结论，从而计算出答案对扰动的稳定率 $S\in[0,1]$。本文将其互补量 Δu 形式化定义反事实稳定性：

$$\Delta u=1-S \tag{17}$$

当 Δu 较大时，表明答案在反事实前提下容易翻转，稳定性不足；反之，则表明答案在多种扰动下保持稳定，可靠性更高。其将作为 4.2.2 节中权重解析函数中的重要组成部分，驱动后续的知识检索-再融合决策。

图 4 反事实假设构造示例

4.2.2 权重解析函数

为避免传统硬阈值策略在置信度差异较小时出现决策不稳定的现象，在 K-Agent 中设计了一个自适应权重解析函数，将内外知识置信度均值差与答案反事实稳定性综合映射为融合权重 $w\in(0,1)$：

$$w=sigmoid(\bar{\omega}\cdot(\mu_{int}-\mu_{ext})+\xi\cdot\Delta u) \tag{18}$$

其中,$sigmoid$ 函数用于平滑权重,$\mu_{int}-\mu_{ext}$ 为内外置信度均值差,Δu 为 4.2.1 节中估计的反事实稳定性,衡量当前答案的不确定程度,值越大表示越不可靠。为进一步增强模型的自适应性,$\bar{\omega}$ 和 ξ 不是固定常数权重,而采用相对归一化机制动态确定:

$$\bar{\omega}=\frac{|\mu_{int}-\mu_{ext}|}{|\mu_{int}-\mu_{ext}|+|\Delta u|},\xi=1-\bar{\omega} \quad (19)$$

模型能根据知识冲突强度与答案稳定性自动平衡两类信息源的影响。

4.3　智能体协同机制

在 ACR 框架中,E-Agent 负责知识置信度的校准与不确定性建模,K-Agent 负责融合权重的动态调控与答案稳定性评估。两者以不确定性感知为起点,通过动态融合与反馈校准实现生成稳定性的自适应调控。

在每一轮交互中,E-Agent 将基于内部与外部知识生成的置信度分布 $c_{int}=[\mu_{int},\sigma_{int}]$ 和 $c_{ext}=[\mu_{ext},\sigma_{ext}]$ 发送给 K-Agent。

K-Agent 首先会依据两者置信度分布的均值差与方差差异计算冲突 I_c 与补充信息量 I_s:

$$I_c=-\log\left(\frac{\mu_{ext}-\mu_{int}}{s}\right),$$

$$I_s=\frac{(\mu_{ext}-\mu_{int})^2}{2s^2},$$

$$s=\sqrt{\sigma_{int}^2+\sigma_{ext}^2} \quad (20)$$

然后 K-Agent 基于自适应权重解析函数(4.10)计算当前融合权重 w_t,并据此基于公式(4.13)得到融合知识表示 h_t:

$$h_t=w_t\cdot h_{ext}^t+(1-w_t)\cdot h_{int}^t \quad (21)$$

该融合表示被输入至 LLM 生成候选答案 y_t。随后对 y_t 评估反事实稳定性,得到新的不稳定度 Δu_t。

此时,K-Agent 将判断是否处于稳定收敛区或需进入下一轮调整。若 $|I_c-I_s|\leqslant\theta$ 且 $\Delta u_t<\Delta u_{t-1}$,说明内外置信度高度一致,难以做出决策,而且新增外部信息依然显著提升稳健性,则 K-Agent 向 E-Agent 发送再次检索请求,追加高相关外部知识并要求再校准。否则循环终止,基于当前知识融合表示 h_t 生成最终答案 $A_{fusion}=F(Q,h_t)$。

智能体间通过协同合作增强了 LLM 在复杂冲突场景下的生成稳定性和融合准确性,有效缓解了置信度高度近似场景下的随机输出困境。

5　实验分析

5.1　实验数据

本文选取五个公开的问答数据集评估提出方法的有效性,其覆盖多个领域以及不同难度层级。具体如下:

(1)ClashEval[39]:包含多领域世界知识问题,每个问题配备正确与扰动后的错误上下文,用于检验模型在冲突知识输入下的推理鲁棒性。

(2)NaturalQA(NQ)[40]:基于真实场景的开放域数据集,问题覆盖自然语言表述的复杂信息需求,侧重考察模型对非结构化外部知识的整合能力。

(3)PopQA(PQ)[41]:可用于评估知识冲突方法中模型处理低流行度实体的能力,如评估模型面对长尾问题或冷门实体时的推理与回答准确性等。

(4)TriviaQA(TQ)[42]:该数据集具有问题复杂、句法词汇多变且需跨句推理的特点,可用于评估知识冲突方法中模型对冲突信息的检测、推理解决能力,以及在复杂语境下给出准确合理答案的能力。

(5)2WikiMultiHopQA[43]:作为多跳推理相关数据集,用于评估模型在需要多步逻辑推导问题上的表现,可检验模型的复杂推理与知识整合能力。

5.2　评价指标

对于 ClashEval、PopQA、2WikiMultiHopQA 和 TriviaQA 数据集,本文使用精确度(Acc)作为评价指标,即当模型输出包含任一真实答案片段时判定为正确,侧重考察知识冲突下的答案召回能力;对于 NaturalQA 数据集,因其真实答案存在不完整性,采用"精确度+基于 LLM 度量"的策略,首先通过匹配精确度判断答案核心要素是否存在,若不满足则启用大模型评估,考察方法在复杂语境下对冲突知识的筛选与融合效果。所有指标均按比例抽样,数据集详细统计信息见表 2。其中,"采样率"与"近似冲突率"均是基于数据集的统计结果,而非方法环节的采样设置。前者指在数据集中随机抽取样本的比例,后者则指在该抽样数据上估计得到的知识冲突比例。

表 2　实验数据集

数据集	问答对数量	采样率	近似冲突率
ClashEval	1200	41.6%	23%
NaturalQA	3610	13.8%	17%
PopQA	1399	35.7%	12%
TriviaQA	11313	8.8%	14%
2WikiMHQA	12576	7.8%	19%

5.3　基准模型

为了充分评估提出方法的有效性,本文分别选择三类基准模型进行对比。具体如下:

(1)忠于记忆

该类方法仅依赖 LLM 参数记忆中的固有参数化知识进行推理问答,验证模型原生知识处理能力,包括 Direct QA、CoT[36] 和 RAAT[44] 方法。Direct QA 直接通过问答指令(如"请回答以下问题:[问题内容]")驱动 LLM 生成答案,依赖 LLM 本身内部固有参数化知识进行答案生成。CoT(Chain of Thought)则在 Direct QA 基础上,通过提示词引导模型显式生成推理步骤(如"请逐步思考并回答以下问题:[问题内容]"),模拟人类多步逻辑推导过程,增强复杂问题的内部知识调用能力。RAAT 通过削弱模型对外部情境的盲目依赖,从而更忠实于自身记忆与稳定知识。

(2)忠于情境

该类方法主要依赖外部检索的情境知识进行问答推理,检验 LLM 对外部知识的独立适配能力,以 CAD[24] 和 IRCAN[45] 为代表。CAD 遵循对比输出分布,当模型在有上下文和没有上下文的情况下使用时,它会放大输出概率之间的差异,在覆盖模型的先验知识方面特别有效。IRCAN 则通过识别并强化与上下文相关的神经元,使模型在生成时优先遵从外部情境知识。

(3)混合策略

该类方法协同 LLM 内部记忆知识与外部检索情境知识,聚焦知识冲突场景下的知识筛选与融合策略,包括 ASTUTE RAG[30]、CK-PLUG[31]、KAFT[27]、KnowPO[9]、Parenting[46] 及 ACTIVE RAG[47]。ASTUTE RAG 显式区分内外部知识来源并通过迭代式源感知整合机制融合多源信息,结合冲突信息筛选与最优答案选择策略。CK-PLUG 设计置信增益检测模块识别知识冲突,通过设置 LLM 参数权重与情境知识依赖阈值,实现内外部知识融合控制。KAFT 通过记忆控制门,让模型可调地激活或抑制不同来源的知识,使模型既能避免上下文噪声干扰,又能在需要时融入外部情境知识。KnowPO 构建知识冲突的数据集微调大模型,并引入偏好优化学习如何避免错误的知识选择。Parenting 通过参数解耦与微调优化模型,平衡模型对内部记忆与外部情境知识的依赖。ACTIVE RAG 通过模拟人类主动学习行为,校准 LLM 内部知识偏差,优化外部知识利用效率,有效降低内外部知识冲突。

5.4 实验设置

本文使用不同的大语言模型来评估提出方法的有效性,包括 Llama3-8B、ChatGPT-4o 以及 ChatGPT-4o-Mini. 对于开源的 Llama3-8B 模型,本文在 NVIDIA RTX A6000 进行部署。对于闭源大模型 ChatGPT-4o 和 ChatGPT-4o-Mini,本文通过调用硅基流动平台的官方 API 进行实验。值得注意的是,RAAT、IRCAN、KAFT、KnowPO 与 Parenting 五种基线方法需要访问模型参数以进行微调或涉及内部神经元级操作,因此我们仅在 Llama3-8B 这一开源模型上对其进行了相应实验。由于闭源模型 ChatGPT-4o 和 ChatGPT-4o-Mini 无法进行参数更新或结构改动,未在其中进行实验基线对比。

本文中超参数设置如下:采样次数 M 为 3,温度参数 T 为 0.5,随机核阈值 p 为 0.8,阈值 θ 为 0.05。本文将数据集中包含的上下文信息统一整理并存储于 FAISS 向量数据库中,用作模型的外部知识源。在推理过程中,模型从 FAISS 中检索并整合与输入问题最相关的上下文,以辅助答案生成。

5.5 实验结果与分析

表 3～表 5 分别展示基于不同 LLM 时本文提出方法与基线模型在 5 个公开数据集上的性能比较。根据实验结果可以看出:

表 3 基于 Llama3-8B 在不同数据集上的精确度对比

	方法	ClashEval	NaturalQA	PopQA	TriviaQA	2WikiMHQA
忠于记忆	Direct QA	26.4	39.4	24.8	64.6	29.2
	CoT	30.6	41.2	25.4	67.1	45.5
忠于情境	RAAT	39.3	40.5	25.0	64.9	41.0
	CAD	32.2	33.6	31.6	67.3	48.2
	IRCAN	47.7	37.2	57.8	70.6	42.3
混合策略	ASTUTE RAG	37.4	49.8	33.4	70.1	49.1
	CK-PLUG	34.8	46.4	28.8	71.4	47.9
	KAFT	49.6	32.8	45.3	59.4	42.1
	KnowPO	52.4	48.3	53.1	71.0	56.2
	Parenting	53.9	54.1	57.4	72.3	58.5
	ACTIVE RAG	55.6	65.0	55.8	79.8	52.2
	ACR	**58.6**	**74.0**	**61.2**	**84.2**	**63.3**

表 4　基于 ChatGPT-4o-mini 的在不同数据集上的精确度对比

	方法	ClashEval	NaturalQA	PopQA	TriviaQA	2WikiMHQA
忠于记忆	Direct QA	38.4	44.4	31.6	67.7	46.1
	CoT	43.0	48.2	36.4	73.6	49.7
忠于情境	CAD	42.8	34.0	32.8	70.1	50.3
混合策略	ASTUTE RAG	41.4	51.2	38.2	72.5	51.3
	CK-PLUG	44.2	53.6	37.2	74.3	50.2
	ACTIVE RAG	57.8	71.6	60.8	83.4	59.6
	ACR	**59.8**	**76.4**	**61.2**	**86.5**	**66.6**

表 5　基于 ChatGPT-4o 的在不同数据集上的精确度对比

	方法	ClashEval	NaturalQA	PopQA	TriviaQA	2WikiMHQA
忠于记忆	Direct QA	43.6	49.0	37.2	69.5	49.1
	CoT	45.2	54.2	42.4	72.0	52.6
忠于情境	CAD	44.5	53.8	39.9	72.3	53.6
混合策略	ASTUTE RAG	47.6	55.6	43.8	75.3	54.6
	CK-PLUG	55.8	59.4	43.2	77.5	54.3
	ACTIVE RAG	59.2	73.2	61.6	84.2	62.7
	ACR	**60.4**	**78.6**	**63.6**	**87.9**	**68.4**

(1)整体性能提升:总体而言,ACR 在所有数据集上表现出色,显著超越了当前的 SOTA 方法。忠于记忆的方法(如 CoT、RAAT)强调对模型内部知识的遵从性,能保持生成过程的一致性与逻辑连贯,但在知识陈旧的情境下易受局部偏差影响。忠于情境的方法(如 CAD、IRCAN)则通过引入外部检索信息增强事实的时效性,但往往忽视模型内部知识的约束,从而在生成阶段产生情境漂移的问题。基于混合策略的方法尝试在两者之间实现平衡,但是他们在知识整合上多采用静态或阶段性融合机制。如 ASTUTE RAG 与 CK-PLUG 通过固定权重注入外部知识,KAFT 与 KnowPO 仅在训练微调阶段对权重进行有限调整,Parenting 与 ACTIVE RAG 虽具有一定动态性,但并非自适应优化。相比之下,ACR 通过 E-Agent 与 K-Agent 的协同建模,实现生成过程中对置信度与知识融合权重的自适应动态调整,而且智能体间的协同机制实现的答案自我纠正,进一步增强 LLM 的鲁棒性和可靠性。此外,随着 LLM 模型能力的增强,各方法整体性能均有所提升,但 ACR 能在各模型中保持稳定领先,表现出良好的可迁移性,说明该方法协同机制并非依赖特定模型结构,而是通过动态融合与自校正过程普遍增强了大模型在知识冲突下的稳定性与一致性。

(2)动态权重分配的冲突消解能力:ACR 通过置信度校准与自适应权重分配的协同机制,显著提升了 LLM 在内外部知识冲突场景下的鲁棒性与稳定性。在 ClashEval 数据集(强调外部知识冲突)和 TriviaQA 数据集(强调内部知识冲突)上,ACR 性能均优于所有基线模型。以 Llama3-8B 为例,ACR 在 ClashEval(58.6)和 TriviaQA(84.2)上分别较次优方法 ACTIVE RAG 提升 3.0%和 4.4%,这是由于 E-Agent 的轻量贝叶斯校准模块可将内外知识的置信度映射至统一概率空间,避免现有静态阈值方法对不同知识源的偏置,同时 K-Agent 自适应权重解析函数在内外知识置信度差异不大情况下仍能保持梯度平滑分配,确保融合比例的动态调整。对于 TriviaQA 数据集中涉及多源争议的常识性问题,ACR 通过反事实稳定度建模动态修正外部知识置信度,避免模型对低置信度知识过度依赖。在 GPT-4o 上,ACR 较 ACTIVE RAG 提升 3.7%,进一步验证了 ACR 对情境-记忆知识冲突的消解能力。

(3)长尾知识的自适应融合能力:ACR 在应对低流行度实体问题时(PopQA 数据集)展现出优越的知识融合鲁棒性与跨规模适应性。在 Llama3-8B 上,ACR(61.2)性能优于现有混合策略(如 ACTIVE RAG,55.8),这是由于 E-Agent 在提取内外部知识时通过多次采样、多轮检索等机制有效降低了低质量检索片段对 LLM 决策的误导概率,并且 K-Agent 动态调整内外知识融合权重,使得 LLM 能更稳定地偏向于相对可信的知识来源。

此外,ACR 在 LLM 参数规模扩大后仍保持优势,表现出良好的可扩展性。例如,在 GPT-4o 上 ACR(63.6)较 ACTIVE RAG(61.6)提升 2.0%,证明 ACR 框架有很好的扩展性,既能缓解小规模 LLM 的知识覆盖瓶颈,又能适配大规模 LLM 的更高密度的知识分布特性,在长尾知识处理任务中具有更强的通用性与稳

定性。

(4)复杂语义的深度融合能力:在 NaturalQA 数据集中,LLM 需处理跨多个知识源、上下文模糊等复杂问答任务。ACR 在该场景下展现出显著优于现有方法的细粒度知识融合能力。这是由于 ACR 框架中,E-Agent 的轻量贝叶斯置信度校准策略将内外部知识映射至统一度量空间,配合 K-Agent 的反事实稳定度建模,对答案的不稳定性进行动态评估并据此调整内外知识的融合比例,从而实现语义层级上的深度整合。实验结果表明 ACR 在 GPT-4o 上精确度达到 78.6,相较 ACTIVE RAG 提升 5.4%,且在资源受限的 GPT-4o-mini 中精确度仍保持 76.4,相较于 ACTIVE RAG 提升 4.8%。进一步对比基线方法 ASTUTE RAG(75.3)和 CK-PLUG(77.5),ACR 通过自适应权重解析函数避免了静态阈值融合策略的随机决策问题,使知识整合过程更加平滑、连续、语义一致,实现了细粒度的知识互补。

(5)多跳推理能力:在 2WikiMHQA 数据集中,LLM 需基于检索多个分散外部知识片段完成多跳逻辑推理,该任务对知识融合过程的连贯性与推理链条的稳定性提出了更高要求。ACR 通过 E-Agent 和 K-Agent 的协同合作有效缓解了因内外部知识冲突导致的推理路径断裂问题。具体的,在 Llama3-8B 上,ACR 达到 63.3 的准确率,较 ACTIVE RAG(52.2)提升 21.6%。这是由于 K-Agent 的权重解析函数可在每一跳推理过程中根据局部置信度差异与反事实稳定性信号,动态调整内外知识源的融合比重,从而保证中间结论在逻辑上前后一致,避免因单一来源知识错误导致整体链式断裂。当模型规模扩大至 GPT-4o 时,ACR 仍然较 ACTIVE RAG 提升 5.7%,进一步验证 ACR 框架在推理路径较长、知识分布复杂的场景中,能够稳定引导模型完成推理问答,提升问答准确性。

5.6 消融实验

为进一步验证 ACR 框架中两个智能体的独立作用与协同增益,本文进行消融实验,结果如表 6 所示。实验结果表明 ACR 中 E-Agent 与 K-Agent 在不同类型的知识冲突任务中各具关键作用,且协同建构了 ACR 的整体性能优势。当移除 E-Agent(w/o E-Agent)时,ACR 在 TriviaQA 数据集上性能下降最显著(87.9→84.0,−3.9%),表明轻量贝叶斯校准机制对于复杂推理任务中的置信度度量一致性至关重要,能够有效修正语义偏差、提高置信度对比精度。而 ClashEval 数据集上,精确度仅下降 1.8%(60.4→58.6),表明外部知识冲突主导的场景中,其他组件对置信度缺失具一定补偿能力。当移除 K-Agent(w/o K-Agent)时,整体性能显著下降,尤其在多跳推理任务 2WikiMHQA 中精确度下降 8.5%(68.4→59.9),表明其自适应权重解析函数对动态决策的核心作用。K-Agent 缺失后 LLM 退化为静态融合策略,无法解决内外知识置信度近似时知识冲突问题。完整 ACR 框架在 TriviaQA(+3.9%)和 NaturalQA(+6.8%)上的性能提升最为显著,表明 ACR 通过校准-融合链式增强实现协同增效:E-Agent 消除度量偏差为 K-Agent 提供可靠输入,而 K-Agent 的反事实稳定度建模进一步优化权重分配。

表 6 消融实验

方法	ClashEval	NaturalQA	PopQA	TriviaQA	2WikiMHQA
w/o E-Agent	58.6↓	76.4↓	61.4↓	84.0↓	64.1↓
w/o K-Agent	54.2↓	71.8↓	59.2↓	80.5↓	59.9↓
w/o 语义重写	59.1↓	77.0↓	61.9↓	84.9↓	65.0↓
w/o 链式推理	59.0↓	76.7↓	61.8↓	85.0↓	64.6↓
w/o 反事实验证	59.5↓	77.5↓	62.2↓	85.5↓	65.6↓
w/o 对比触发	59.3↓	77.1↓	62.0↓	85.3↓	65.4↓
ACR	60.4	78.6	63.6	87.9	68.4

为验证四种提示策略(语义重写、链式推理、反事实验证、对比触发)的有效性,我们逐一移除各提示进行消融实验。结果如表 6 所示,去除任一提示策略都会导致性能下降,说明四个角度的设计在提升模型理解与推理方面均发挥了关键作用。其中,去除语义重写和链式推理的影响最为显著,表明这两类提示在增强多跳问答与复杂语义解析上贡献最大;反事实验证与对比触发则进一步提升了模型的稳健性与泛化能力。

此外,本文对 ACR 的知识自主选择能力进行评估,如表 7 所示。当外部检索知识中存在正确答案时,ACR 在 NaturalQA(NQ,82.6→+2.2)、PopQA(PQ,91.8→+1.6)、TriviaQA(TQ,95.4→+1.4)的性能均优于 ACTIVE RAG,表明其通过

内外部知识的权重动态调整,避免单一静态阈值导致 LLM 在置信度近似时陷入随机化困境。

表 7　内外知识权重动态调整能力评估

方法	外部知识存在答案			外部知识不存在答案			内部记忆知识		
	NQ	PQ	TQ	NQ	PQ	TQ	NQ	PQ	TQ
RAG	76.2	85.4	91.7	6.4	3.2	18.0	79.0	85.2	88.7
ACTIVE RAG	80.4	90.2	94.0	25.8	8.4	36.0	92.8	87.2	95.3
ACR	82.6	91.8	95.4	26.2	9.8	37.4	93.2	91.2	96.5

当外部检索知识不存在答案时,ACR 在 NaturalQA(NQ,26.2→+0.4)、TriviaQA(TQ,37.4→+1.4)的性能也优于 ACTIVE RAG,尤其 TQ 提升达 3.9%,表明其可通过动态调整内外部知识权重提升 LLM 性能;在仅依赖内部记忆知识场景下,ACR 在 PopQA(PQ,91.2→+4.0)的显著优势进一步验证其并非单向依赖外部知识,而是通过 K-Agent 动态分配内外部知识权重,从而实现比 ACTIVE RAG 更精准地自主决策。

5.7　案例研究

为进一步验证 ACR 在内外知识置信度高度接近时的知识冲突处理能力,本文选取了一个来自 NaturalQA 的代表性案例进行分析,如图 5 所示。在该案例中,LLM 内部记忆正确地关联至“Julius Wellhausen”,而外部检索返回的知识中包含错误信息:“Jahwist 的提出者是 14 世纪北非学者 Ibn Khaldun”,形成情境-记忆知识冲突。此时,现有方法如 ACTIVE RAG,因过度依赖外部检索知识,错误地采纳了 Ibn Khaldun 作为答案。而 ACR 则成功识别并消解该冲突。这是由于 ACR 中的 E-Agent 首先对内外知识进行置信度建模,利用轻量贝叶斯校准策略将两类知识统一映射至相同的度量空间,该机制不仅整合了多轮采样结果,还量化了答案生成的稳定性与不确定性,最终分别得到外部知识置信度 0.51 与内部知识置信度 0.49,体现出置信度极为接近、模型缺乏明显偏向的典型不确定性边界场景。在此基础上,K-Agent 使用自适应权重解析函数,对置信度差异与反事实稳定度进行联合建模,面对仅 2%差值的高不确定性区域,K-Agent 并未强行做出二元选择,而是通过梯度驱动的动态融合机制,动态偏向于稳定度更高的内部记忆响应,最终生成正确答案“Julius Wellhausen”。这一案例研究表明,ACR 相较于使用单一静态阈值决策策略的方法,不仅能规避在不确定性区域附近因误差扰动引发的“随机决策”,更能有效缓解当前主流 DB4LLM 框架(如 ACTIVE RAG)在面对错误外部检索知识时的“盲信外部知识”问题。综上所述,该案例充分验证了 ACR 框架在置信度近似的知识冲突区域中,凭借双智能体协同机制,具备更强的抗干扰能力与推理稳定性,从而提升复杂问答场景下的生成准确率与鲁棒性。

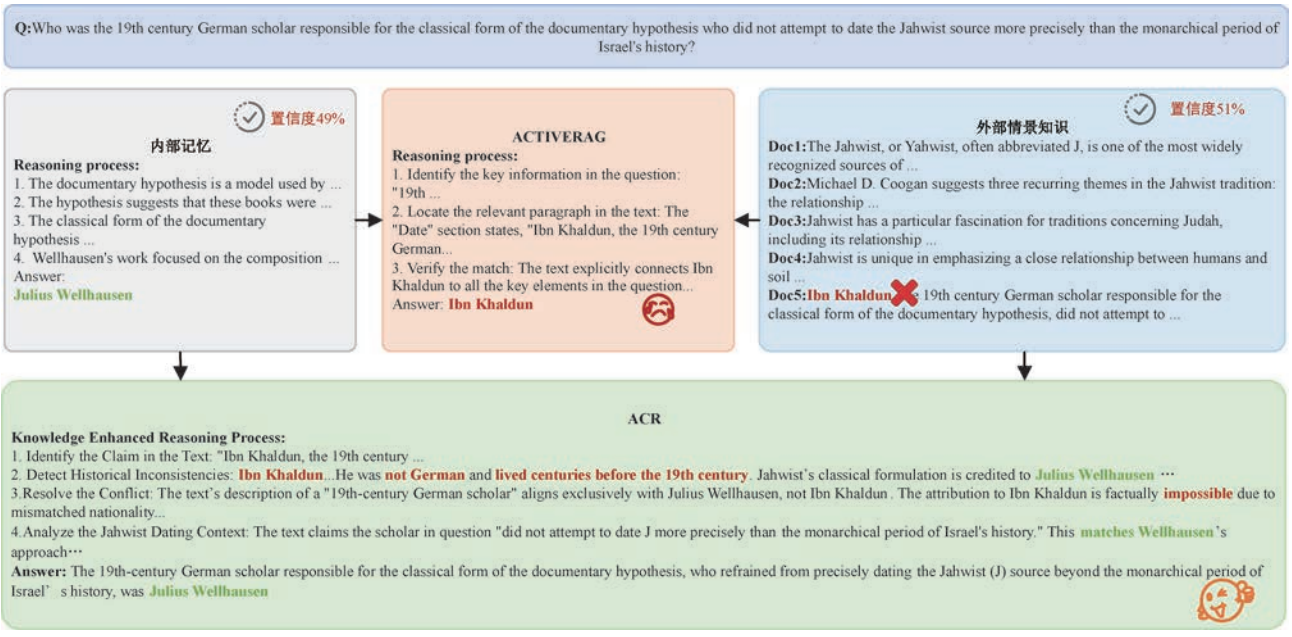

图 5　ACTIVE RAG 与 ACR 在置信度近似场景下的知识融合案例分析

我们对该案例在知识融合过程中的权重 w 的动态变化趋势进行可视化，如图 6 所示。可以观察到，在推理初期，内外置信度高度一致，且反事实不稳定度 Δu 较大，连续触发两次验证性再检索，以补充外部证据并重新校准置信分布。随着新的外部知识注入，外部知识置信度 μ_{ext} 逐步提升、反事实不稳定度 Δu 显著下降，融合权重缓慢偏向外部知识，说明 K-Agent 和 E-Agent 协同机制有效实现了对高一致性不确定区的自适应调控。最终迭代 3 次以后，置信度有了明显差异，可直接做出决策，推理过程结束。融合权重的自适应变化能够直观反映 ACR 在多源知识间的动态平衡过程，验证了我们方法在高不确定性决策区的可解释性与鲁棒性。

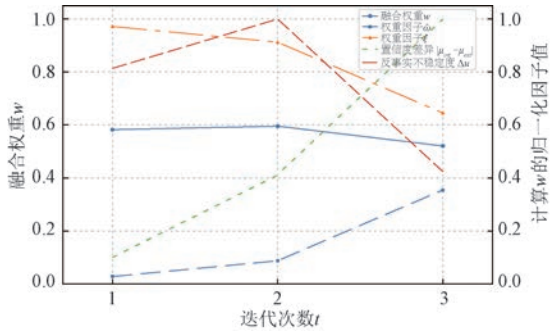

图 6 融合权重 w 的变化趋势

5.8 超参数敏感性分析

本文基于 Llama3-8B 在 NaturalQA 数据集上，对采样次数 M、温度 T、核阈值 p 和阈值 θ 做了超参数敏感性实验，实验结果如图 7～图 9 所示。

在图 7 中，温度和核阈值热图的高亮区稳定落在 $T\approx0.5$，$p=0.8$，此处精确度达 74.0%为最优；温度 T 再升至 0.9 或 p 提高到 0.95 时，多样性虽略增准确率反而降低。

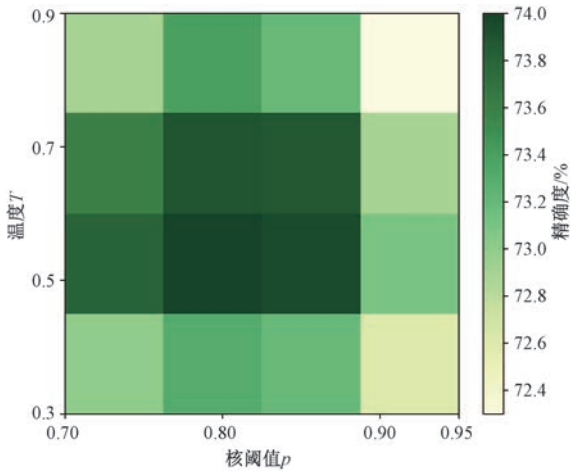

图 7 温度 T & 核阈值 p

图 8 中采样次数 $M=3$ 时，模型性能达到最优。当采样次数较少时，内部知识覆盖不足，导致模型无法充分探索多样化的推理路径。而当采样次数过多，虽然信息冗余增加，但噪声样本比例上升，反而削弱了融合效果。

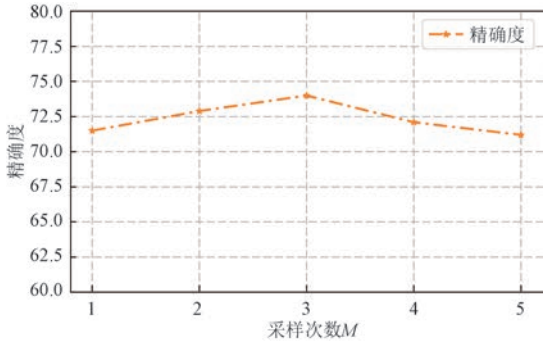

图 8 精确度随采样次数 M 的变化趋势

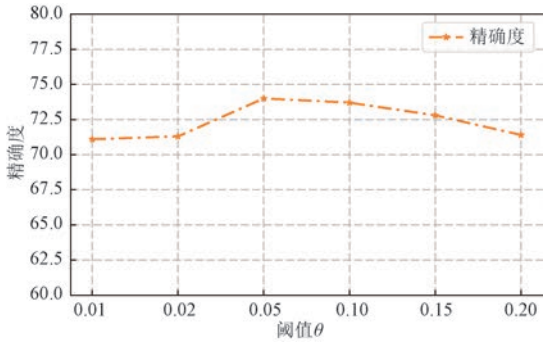

图 9 精确度随阈值 θ 的变化趋势

此外，为进一步验证模型在内外置信度接近（即高不确定性）场景下的表现，我们选取置信度差值满足的样本，并将其划分为 5 个区间：[0，0.02）、[0.02，0.05）、[0.05，0.10）、[0.10，0.20）和[0.20，1）。随后对各区间样本的性能进行统计分析，如图 10 所示。可以观察到，在置信度差异较小的区间中，

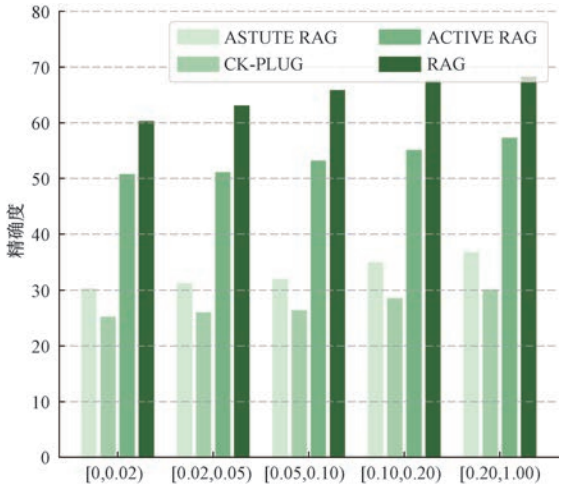

图 10 不同置信度差异区间的性能分析

ACR 相较基线模型性能提升显著,这说明本文方法在高不确定性决策区内能更好地融合内外知识。

5.9　时间开销分析

如图 11 所示,本文以 NaturalQA 数据集为例,基于 Llama3-8B 对比了 ASTUTE RAG、CK-PLUG、ACTIVE RAG 以及 ACR 方法在精度、推理时长和迭代次数三项指标下的综合性能。CK-PLUG 由于采用单步可插拔推理机制,推理时间最短,但精度较低,在复杂任务下表现受限。ASTUTE RAG 与 ACTIVE RAG 均为多阶段迭代框架,平均推理时间在一定程度上兼顾了精度与效率。相比之下,ACR 的推理时长略高,平均迭代次数与 ASTUTE RAG 和 ACTIVE RAG 与一致,但其精度显著优于其他方法。这一结果表明,ACR 在推理时间可接受的前提下,通过多轮内外知识融合获得了更高的准确性。尽管多次交互带来了轻微的时间开销增加,但该机制有效缓解了内外知识冲突,在精度-效率权衡上展现出更优的整体表现。

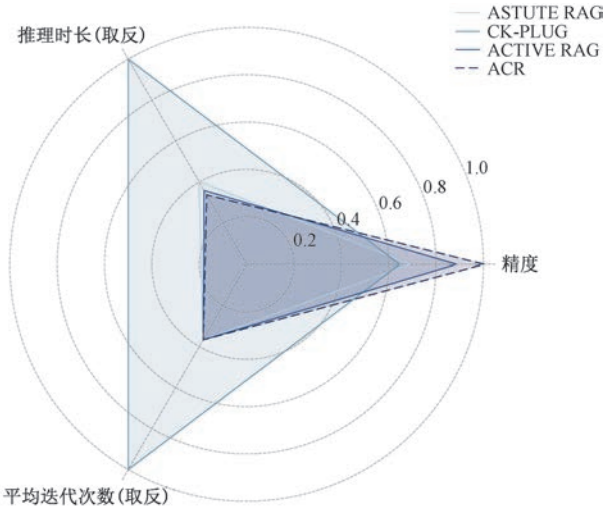

图 11　精度-效率分析

6　总　结

本文针对大语言模型(LLM)在 DB4LLM 框架中普遍面临的情境-记忆知识冲突问题,尤其是内外部知识置信度高度近似的情形下,提出了一种基于智能体的冲突消解框架 ACR。ACR 通过置信度校准与动态权重调控机制,实现了对内外部知识的细粒度融合。具体的,E-Agent 建模生成不确定性,并通过轻量贝叶斯校准策略将内部记忆与外部检索知识统一映射到同一置信度空间,有效提升了置信度可比性与生成稳定性;K-Agent 通过反事实稳定性判断和置信度差值建模自适应权重解析函数,实现知识融合权重的动态分配。两个智能体相互协作,在高不确定性区域触发反馈机制,实现答案生成的自我修正。大量实验结果表明,ACR 在多种复杂任务中性能均显著优于现有 SOTA 方法,特别是在置信度高度近似情形下的情境-记忆知识冲突消解能力得到明显提升。

参　考　文　献

[1] Liu Zeyuan, Wang Pengjiang, Song Xiaobin, Zhang Xin, Jiang Benben. Survey on hallucinations in large language models. Journal of Software, 2025, 36(3): 1152-1185 (in Chinese)
(刘泽垣,王鹏江,宋晓斌,张欣,江奔奔. 大语言模型的幻觉问题研究综述. 软件学报, 2025, 36(3): 1152-1185)

[2] Shu Wentao, Li Ruixiao, Sun Tianxiang, Huang Xuanjing, Qiu Xipeng. Large language models: Principles, implementation, and progress. Journal of Computer Research and Development, 2024, 61(2): 351-361 (in Chinese)
(舒文韬,李睿潇,孙天祥,黄萱菁,邱锡鹏. 大型语言模型:原理、实现与发展. 计算机研究与发展, 2024, 61(2): 351-361)

[3] Xu Rongwu, Qi Zehan, Guo Zhijiang, Wang Cunxiang, Wang Hongru, Zhang Yue, Xu Wei. Knowledge conflicts for LLMs: A Survey//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Florida, USA, 2024: 8541-8565

[4] Adam Roberts, Colin Raffel, Noam Shazeer. How much knowledge can you pack into the parameters of a language model? //Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Virtual. 2020: 5418-5426

[5] Adam Liska, Tomas Kocisky, Elena Gribovskaya, et al. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models//Proceedings of the 39th International Conference on Machine Learning. Maryland, USA, 2022: 13604-13622

[6] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Texas, USA, 2016: 2383-2392

[7] Liu Pengfei, Yuan Weizhe, Fu Jinlan, Jiang Zhengbao, Hiroaki Hayashi, Graham Neubig. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys, 2023, 55(9): 1-35

[8] Qian Cheng, Zhao Xinran, Wu Sherry Tongshuang. "Merge Conflicts!" Exploring the impacts of external distractors to

parametric knowledge graphs, 2023, arXiv preprint, abs/2309.08594

[9] Zhang Ruizhe, Xu Yongxin, Xiao Yuzhen, Zhu Runchuan, Jiang Xinke, Chu Xu, Zhao Junfeng, Wang Yasha. KnowPO: Knowledge-Aware preference optimization for controllable knowledge selection in retrieval-augmented language models//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2025, 39(24): 25895-25903

[10] Pan Liangming, Chen Wenhu, Kan Min-Yen, William Yang Wang. Attacking open-domain question answering by injecting misinformation//Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. Nusa Dua, Indonesia, 2023: 525-539

[11] Giovanni Spitale, Nikola Biller-Andorno, Federico Germani. Ai model gpt-3 (dis) informs us better than humans, 2023, arXiv preprint, abs/2301.11924

[12] Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 2023: 10056-10070

[13] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, Sameer Singh. Entity-based knowledge conflicts in question answering//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican, 2021: 7052-7063

[14] Chen Hung-Ting, Zhang Michael, Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: recalibrating models to reflect conflicting evidence//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 2292-2307

[15] Xie Jian, Zhang Kai, Chen Jiangjie, Lou Renze, Su Yu. Adaptive Chameleon or Stubborn Sloth: Revealing the behavior of Large Language models in knowledge conflicts//Proceedings of the Twelfth International Conference on Learning Representations. Vienna, Austria, 2024

[16] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui. Realtime QA: What's the answer right now? //Proceedings of the Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems. LA, USA, 2023, 36: 49025-49043

[17] Nicola De Cao, Wilker Aziz, Ivan Titov. Editing factual knowledge in language models//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican, 2021: 6491-6506

[18] Liao Huanxuan, He Shizhu, Xu Yao, Zhang Yuanzhe, Liu Shengping, Liu Kang, Zhao Jun. Awakening Augmented Generation: Learning to awaken internal knowledge of large language models for question answering//Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE, 2025: 1333-1352

[19] Pan Yikang, Pan Liangming, Chen Wenhu, Preslav Nakov, Min-Yen Kan, William Yang Wang. On the risk of misinformation pollution with large language models//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 1389-1403

[20] Xu Rongwu, Lin Brian, Yang Shujian, Zhang Tianqi, Shi Weiyan, Zhang Tianwei, Fang Zhixuan, Xu Wei, Qiu Han. The Earth is Flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024, 1:16259-16303

[21] Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, Benjamin Van Durme. Defending against disinformation attacks in Open-Domain question answering//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. St. Julian's, Malta , 2024: 402-417

[22] Hong Giwon, Jeonghwan Kim, Kang Junmo, Sung Hyon-Myaeng, Joyce Jiyoung Whang. Discern and answer: Mitigating the impact of misinformation in retrieval-augmented models with discriminators, 2023, arXiv preprint, abs/2305.01579

[23] Zhou Wenxuan, Zhang Sheng, Poon Hoifung, Chen Muhao. Context-faithful prompting for large language models//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 14544-14556

[24] Shi Weijia, Han Xiaochuang, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, Wen-tau Yih. Trusting Your Evidence: Hallucinate less with context-aware decoding//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Mexico City, Mexico, 2024, 2: 783-791

[25] Zhang Michael, Choi Eunsol. Mitigating temporal misalignment by discarding outdated facts//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 14213-14226

[26] Huang Pengcheng, Liu Zhenghao, Yan Yukun, Yi Xiaoyuan, Chen Hao, Liu Zhiyuan, Sun Maosong, Xiao Tong, Yu Ge, Xiong Chenyan. PIP-KAG: Mitigating knowledge conflicts in knowledge-augmented generation via parametric pruning, 2025, arXiv preprint, abs/2502.15543

[27] Li Daliang, Ankit Singh Rawat, Manzil Zaheer, Wang Xin, Michal Lukasik, Andreas Veit, Yu Felix, Sanjiv Kumar. Large language models with controllable working memory//Proceedings of the Association for Computational Linguistics. Toronto, Canada, 2023: 1774-1793

[28] Zhang Yunxiang, Muhammad Khalifa, Lajanugen Lo geswaran, Moontae Lee, Honglak Lee, Wang Lu. Merging generated and retrieved knowledge for Open-Domain QA//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 4710-4728

[29] Jin Zhuoran, Cao Pengfei, Chen Yubo, Liu Kang, Xiao jian Jiang, Xu Jiexin, Li Qiuxia, Zhao Jun. Tug-of-War between Knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. Torino, Italy, 2024: 16867-16878

[30] Wang Fei, Wan Xingchen, Sun Ruoxi, Chen Jiefeng, Sercan Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Vienna, Austria, 2025: 30553-30571

[31] Bi Baolong, Liu Shenghua, Wang Yiwei, Xu Yilong, Fang Junfeng, Mei Lingrui, Cheng Xueqi. Parameters vs. Context: Fine-Grained Control of Knowledge Reliance in Language Models, 2025, arXiv preprint, abs/2503.15888

[32] Wang Jiatai, Xu Zhiwei, Jin Di, Yang Xuewen, Li Tao. Accommodate knowledge conflicts in retrieval-augmented LLMs: Towards Reliable Response Generation in the Wild, 2025, arXiv preprint, abs/2504.12982

[33] Tyralis, H., Papacharalampous, G. A review of predictive uncertainty estimation with machine learning. Artificial Intelligence Review, 2024, 57(4): 94

[34] Li Yingjie, Luo Yun, Xie Xiaotian, Zhang Yue. Task Calibration: Calibrating large language models on inference tasks//Proceedings of the Association for Computational Linguistics. Vienna, Austria, 2025: 6937-6951

[35] Xin Chunlei, Lu Yaojie, Lin Hongyu, Zhou Shuheng, Zhu Huijia, Wang Weiqiang, Liu Zhongyi, Han Xianpei, Sun Le. Chain-of-Rewrite: Aligning question and documents for Open-Domain question answering//Proceedings of the Association for Computational Linguistics. Florida, USA, 2024: 1884-1896

[36] Wei Jason, Wang Xuezhi, Schuurmans Dale, Bosma Maarten, ichter brian, Xia Fei, Chi Ed, Le Quoc V, Zhou Denny. Chain-of-Thought prompting elicits reasoning in large language models//Proceedings of the Neural Information Processing Systems. New Orleans, USA, 2022, 35: 24824-24837

[37] Kyle Moore, Jesse Roberts, Thao Pham, Douglas Fisher. Chain of thought still thinks fast: APriCoT Helps with Thinking Slow, 2024, arXiv preprint, abs/2408.08651

[38] Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, Lidong Bing. Contrastive chain-of-thought prompting, 2023, arXiv preprint, abs/2311.09277

[39] Wu Kevin, Wu Eric, Zou James. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2024, 37: 33402-33422

[40] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, et al. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 2019, 7:452-466

[41] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023, 1: 9802-9822

[42] Mandar Joshi, Eunsol Choi, Daniel Weld, Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017, 1: 1601-1611

[43] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 6609-6625

[44] Fang Feiteng, Bai Yuelin, Ni Shiwen, Yang Min, Chen Xiaojun, Xu Ruifeng. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024, 1: 10028-10039

[45] Shi Dan, Jin Renren, Shen Tianhao, Dong Weilong, Wu Xinwei, Xiong Deyi. IRCAN: Mitigating Knowledge Conflicts in LLM Generation via Identifying and Reweighting Context-Aware Neurons//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2024

[46] Xu Yongxin, Zhang Ruizhe, Jiang Xinke, et al. Parenting: optimizing knowledge selection of retrieval-augmented language models with parameter decoupling and tailored tuning//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Vienna, Austria, 2025, 1: 11643-11662

[47] Xu Zhipeng, Liu Zhenghao, Liu Yibin, Xiong Chenyan, Yan Yukun, Wang Shuo, Yu Shi, Liu Zhiyuan, Yu Ge. ActiveRAG: Revealing the treasures of knowledge via active learning. 2024, arXiv preprint, abs/2402.13547

YAO Xin, Ph. D. candidate. Her main research interest is large language models.

SONG Mei-Ling, M. S. candidate. Her research direction is large language models.

SUN Bin-Hong, M. S. candidate. His research direction is large language models.

BI Xin, Ph. D., associate professor. His main research interests include large language models, big data management and analysis.

ZHAO Xiang-Guo, Ph. D., professor. His research interests include large language models, big data management and analysis.

ZHANG Ao-Qian, Ph. D., associate professor. His research interests include big data analysis.

LI Bo-Yang, Ph. D. His research interests include big data, artificial intelligence, distributed system.

YUAN Ye, Ph. D., professor. His research interests include big data management and analysis, artificial intelligence.

Background

The rapid development of large language models (LLMs) has greatly improved natural language understanding and generation. However, LLMs often suffer from hallucination issues. DB4LLM technology mitigates this issue by retrieving external knowledge during inference, but they introduce a new challenge: context-memory conflict, where external knowledge conflicts with the model's internal parametric knowledge, leading to unstable and unreliable outputs.

Existing methods typically rely on static confidence thresholds for conflict resolution, which become ineffective when the confidence levels of internal and external knowledge are highly similar, forming an uncertainty decision region. In this region, LLMs often exhibit unstable outputs, reduced accuracy, and increased entropy in the generation process.

This paper addresses the above challenge by proposing the Agent Conflict Resolution (ACR) framework. ACR employs two collaborative agents. The E-Agent calibrates uncertainty by a lightweight Bayesian method to unify confidence estimation. The K-Agent performs dynamic knowledge fusion through differentiable weight allocation based on confidence gaps and counterfactual stability. ACR achieves robust conflict resolution, particularly under high uncertainty.

Experiments on five public datasets show that ACR achieves an average performance improvement of 6.08% over existing methods across multiple base models, demonstrating its effectiveness, stability, and generalizability.

This work was supported by the Major Program of the National Natural Science Foundation of China (No. 62394332), the Young Scientists Found (Category A) of the National Natural Science Foundation of China (No. 62225203), the Joint Fund of the National Natural Science Foundation of China (No. U23A20297), the Hebei Provincial Program for Enhancing Innovation Capacity (No. 235A0101D), the Independent Research Project of State Key Laboratory of Intelligent Deep Metal Mining and Equipment (No. IDMEIRI2504), the Liaoning Revitalization Talents Program (No. XLYC2204005), the Beijing Natural Science Foundation (No. L241010), and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM108).