

DOI: 10.3969/j.issn.1671-0673. 2026. 03. 003

置信度引导迭代检索与证据图推理的 事实核查方法

韩 箫¹, 屈 丹^{1,2}, 常禾雨¹, 陈 琦¹, 许旻辰¹

(1. 信息工程大学, 河南 郑州 450001; 2. 先进计算与智能工程(国家级)实验室, 江苏 无锡 214083)

摘要:为解决现有自动事实核查方法中效率与推理深度难以平衡的问题,提出融合迭代检索与深度证据图推理的事实核查框架。该框架以置信度驱动迭代机制为基础,首先通过语言模型内部知识判断声明置信度,低置信度时生成精准搜索查询获取初始证据,并结合早期终止策略避免冗余检索;其次同时融入3层证据图结构,对检索证据进行层级扩展,替代单一内部知识判断以提升低置信度场景精度;最后通过动态迭代检索控制平衡推理深度与成本。在两个公开数据集上的实验结果表明,该框架在 $F1$ 指标较基线方法 FIRE 提升 1~6 个百分点,同时,模型调用搜索次数较另一基线方法 SAFE 最多降低约 86.5%。故该框架在维持高准确性的同时,降低了计算与检索成本。

关键词:事实核查;检索增强生成;大语言模型;智能体;证据图推理

中图分类号:TP391

文献标识码:A

文章编号:1671-0673(2026)03-0267-08

Confidence-Guided Iterative Retrieval and Evidence Graph Reasoning for Fact-Checking

HAN Xiao¹, QU Dan^{1,2}, CHANG Heyu¹, CHEN Qi¹, XU Minchen¹

(1. Information Engineering University, Zhengzhou 450001, China;

2. Laboratory for Advanced Computing and Intelligence Engineering, Wuxi 214083, China)

Abstract: To balance efficiency and reasoning depth in automated fact-checking, a framework of couples iterative retrieval with deep evidence graph reasoning is proposed. Anchored in a confidence-driven iterative mechanism, the confidence of input claims is first assessed by leveraging the intrinsic knowledge of large language models in the framework. For claims assigned low confidence, precise search queries are formulated to retrieve initial evidence and an early termination is employed to avoid redundancy. Then, a three-layer evidence graph is constructed to hierarchically integrate and reason over the retrieved evidence, replacing the single internal knowledge judgment to enhance verification accuracy in low-confidence scenarios. Finally, the trade-off between reasoning and cost is balanced by dynamically controlling retrieval depth. Experimental evaluations on two public datasets demonstrate that the proposed framework achieves a $F1$ score improvement of 1 to 6 percentage points over the FIRE baseline and reduces search times by up to 86.5% against those of SAFE baseline. In summary, this framework demonstrates feasibility in terms of reduced retrieval costs alongside consistent high accuracy.

Key words: fact-checking; retrieval-augmented generation; large language model; agent; evidence graph reasoning

收稿日期:2026-01-03;修回日期:2026-03-17

基金项目:河南省科技攻关项目(252102211040);河南省自然科学基金(252300420990)

作者简介:韩 箫(2000—),女,硕士生,主要研究方向为自然语言处理。E-mail:2018211632@bupt.cn

随着互联网的不断发展,社交媒体成为主要新闻来源,而其中隐藏着的部分虚假信息,会对社会造成恶劣影响。社交媒体数据量大,自动的虚假新闻检测十分必要。事实核查能够准确且直观地反映信息的真实性,自动事实核查(Automated Fact-Checking)是应对虚假信息传播的关键技术。大语言模型(Large Language Model, LLM)驱动的智能体能够对其所处环境进行推理,并通过调用外部工具或执行内部操作来做出决策^[1],近期研究赋予了智能体系统调用搜索引擎等外部工具的能力^[2-4]。Quelle等^[5]探索了利用LLM调用搜索工具,结合外部知识进行事实核查。检索增强生成(Retrieval-Augmented Generation, RAG)模型由FAIR(Facebook AI Research)团队提出^[6],通过从外部知识库中检索相关信息,并将其作为提示输入LLM,以增强模型处理知识密集型任务的能力。LLM指令微调(Instruction Tuning)是对预训练模型进行二次训练,使其能够理解并执行自然语言指令的技术。低秩适应(Low-Rank Adaptation, LoRA)^[7]是一种模型的轻量微调方法,通过向模型的部分层添加可训练的低秩矩阵模块,实现模型在特定任务上的能力调整。

在验证阶段,Sun等^[8]提出了一种基于马尔可夫链的多智能体辩论方法,让智能体基于检索到的证据进行协同决策。Wang等^[9]利用LLM,将待检测信息转化为一阶逻辑子句,再转化为事实核查的具体问题,通过调用外部工具得到对应知识。Zheng等^[10]提出了一种专注于证据检索阶段的事实核查框架,由混合证据检索和联合事实验证两个主要模块组成。Niu等^[11]介绍了一种通过从给定新闻中提取核心事实,并进行全网搜索以识别支持或冲突的报道,利用来源的可信度进行验证的事实核查方法。Xie等^[12]提出的迭代检索框架FIRE通过动态决策降低计算与搜索成本。Zhang等^[13]提出一个可强化跨源推理的3层证据图与证据条件提示机制。陈静等^[14]提出一种知识增强的异质图注意力虚假信息检测方法,通过异质文本图与双层图注意力机制,增强模型对帖文的内容细粒度语义感知和多视角捕捉虚假线索的能力。现有方法常面临效率与推理深度难以兼顾的问题,固定推理流程导致效率低下,而单一的检索获取信息缺乏对证据多层次关联的深度挖掘。

为解决上述问题,本文提出一种融合迭代检索与多层次证据图推理的事实核查框架。该框架

采用置信度引导的迭代检索机制,首先通过LLM内部知识对原声明做出判断,同时生成一个置信度,当置信度较低时,LLM将自动生成一个搜索查询并调用外部工具获取初始证据,避免单一依赖LLM内部知识。然后,框架对检索到的证据进行层级扩展,融入决策模块,提升低置信度场景的核查准确率。此外,对本地部署的模型进行LoRA^[7]微调处理,以提高LLM针对事实核查任务的内部知识储存,进一步提高核准确率以及降低模型搜索调用次数。本文在Factcheck-Bench^[2]和Bing-Check^[15]数据集上进行了实验,结果表明,相比FIRE,所提框架在Macro F1指标上实现了1%~6%的提升,并且模型调用搜索次数进一步减少;与SAFE^[16]相比,搜索次数最多降低7.4倍,实现了高效检索与深度推理的协同优化,为大规模复杂声明的事实核查提供了实用的解决方案。

1 所提方法

已有的迭代检索方法虽能有效控制搜索成本,但对已获取证据的语义关联与上下文理解不足,难以支撑复杂声明所需的深度推理;而基于图推理的深度验证方法虽能提升复杂声明的判断准确性,却缺乏对检索过程的动态调控,易导致资源浪费。针对上述局限,提出一个融合迭代检索与上下文增强推理的统一框架,通过引入数值化置信度引导机制,将模型对当前证据的判断确信程度量化为具体分数,并通过预设阈值实现检索深度的动态控制;同时构建3层证据图,对证据间关系进行建模,从而在低置信度场景下实现证据的深度语义理解与上下文增强推理。通过将检索停止时机与证据结构化处理统一优化,在维持高效查询的同时,显著提升对复杂声明的核查准确性。

1.1 问题定义与框架概览

评估长文本的事实准确性面临着巨大挑战^[17]。为解决这一问题,可将文本分解为单个可核查声明^[4],称为“原子声明”,逐一进行事实核查,随后汇总其事实性得分,以评估原始文本的整体事实准确性。自动化事实核查的任务是给定需要验证的原子声明 c ,输出一个二分类标签 $y \in \{\text{“True”}, \text{“False”}\}$,分别表示声明被支持或反驳。该过程不仅依赖LLM内部知识 k ,更需要从开放域,如互联网中动态检索相关外部证据集 $E = \{e_1, e_2, \dots, e_n\}$ 。为实现上述过程,提出一种置信度引导迭代检索与证据图推理的事实核查方法,具体流程如图1所示。

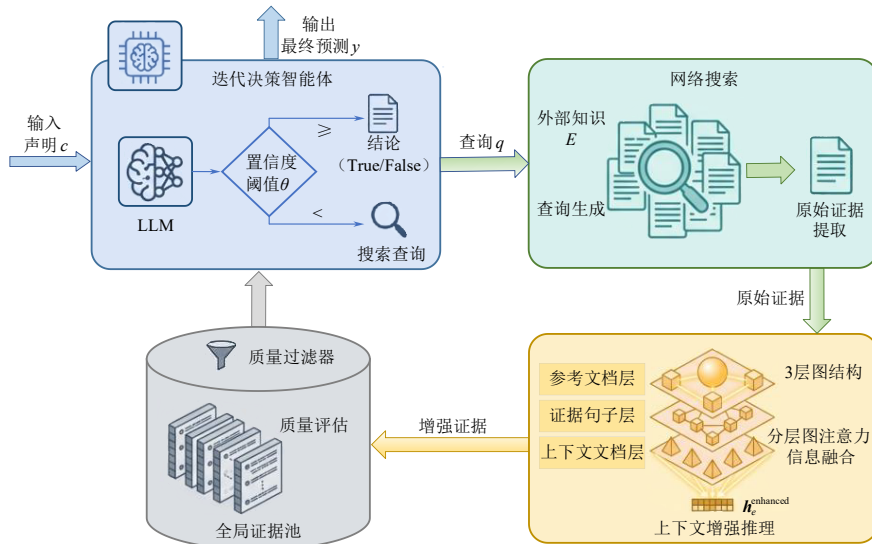


图1 置信度引导迭代检索与证据图推理的事实核查方法

其目标是通过一个智能体 A ,依托自身置信度决策终止验证并输出判断的时机,在信息不足时生成搜索查询以获取外部证据。这一过程需在验证准确性、LLM调用与搜索API费用之间取得平衡。

首先,智能体接收输入声明后,将基于当前证据(初始为空)和内部知识进行初步评估。若智能体对判断具有高置信度,则直接输出结果,否则生成一个优化的搜索查询。其次,新检索到的证据先经由一个上下文增强推理模块进行处理。该模块为证据构建包含其来源上下文与相关引用的3层图结构,并通过图神经网络进行语义增强与融合,输出增强的证据表示。更新后的证据池再被送入决策模块,开启下一轮迭代。最后,上述循环持续进行,直至满足置信度阈值或达到预设的最大迭代次数。这种设计使每一次检索都建立在已有证据的深度理解之上,而每一次推理又为下一次检索提供了更精准的方向。

1.2 迭代决策智能体模块

该模块为框架的核心,其原理是将声明输入智能体模块,LLM将对能否判断声明标签生成一个置信度评分,若置信度高于/等于设定的阈值,则直接得出判断结论;若低于设定的阈值,则LLM将生成一个搜索查询,并调用搜索工具进行搜索。置信度分数由LLM依据对证据的充分性、内部一致性与冲突程度的评估,将该评估映射为 $[0, 1]$ 区间的概率性度量,该分数反映了模型对自身判断的确信程度。

图2为在每一步迭代中,模型基于当前知识判断是否足够自信做出最终结论的提示词。

```
FINAL_ANSWER_OR_NEXT_SEARCH_FORMAT = f"""
Instructions:
1. You are provided with a STATEMENT and relevant KNOWLEDGE points.
2. Based on the KNOWLEDGE, assess the factual accuracy of the STATEMENT.
3. Before presenting your conclusion, think through the process step-by-step.
   You MUST estimate a calibrated confidence score in  $[0, 1]$  and include it in your final JSON output.
   - 0.0 means the current KNOWLEDGE is clearly insufficient or contradictory.
   - 1.0 means the current KNOWLEDGE is sufficient and highly consistent for a final verdict.
   - Use intermediate values when evidence is partial, ambiguous, or weak.
   - Do not use only 0 or 1 unless certainty is extreme.
   - Output "confidence" as a numeric field (for example: 0.73).
4. If the KNOWLEDGE allows you to confidently make a decision (i.e., your confidence score  $\geq$  {DEFAULT_CONFIDENCE_THRESHOLD}), output the final answer as a JSON object in the following format:
   {{
     "confidence": <0-1 float>,
     "final_answer":
       "{_FACTUAL_LABEL}" or "{_NON_FACTUAL_LABEL}"
   }}
5. If the KNOWLEDGE is insufficient to make a judgment (i.e., your confidence score  $<$  {DEFAULT_CONFIDENCE_THRESHOLD}), issue ONE Search query that could provide additional evidence. Output the search query in JSON format, as follows:
   {{
     "confidence": <0-1 float>,
     "search_query": "Your Google search query here"
   }}
6. The query should aim to obtain new information not already present in the KNOWLEDGE, specifically helpful for verifying the STATEMENT's accuracy.
7. Output exactly one JSON object only, with no additional text, and do not include keys outside the required format.

KNOWLEDGE:
{KNOWLEDGE_PLACEHOLDER}

STATEMENT:
{STATEMENT_PLACEHOLDER}
"""
```

图2 基于当前置信度模型能否做出结论的提示词

迭代搜索过程依赖决策函数 $f_{\text{decision}}(c, E, k)$,该函数用于评估系统针对现有信息做出最终事实判断的总体置信度 t_{conf} 。决策逻辑遵循简单的阈值规

则, θ 为预设的置信度阈值:

$$f_{\text{decision}}(c, E_t, k) = \begin{cases} \text{Answer}, t_{\text{conf}} \geq \theta; \\ \text{Search}, \text{其他}. \end{cases}$$

当搜索查询次数为 q_i , 若置信度达到或超过 θ , 系统终止检索过程, 输出当前的验证结果。反之, 系统生成一个新的搜索查询 q_{i+1} , 并从外部知识源获取进一步证据。此迭代过程持续进行, 直至达到置信度要求或达到预设的最大检索轮次上限。通过设立这一机制, 确保搜索行为仅在模型对当前证据的解释存疑时才触发, 从而实现该框架成本优化。

1.3 上下文增强推理模块

该模块负责对迭代中检索的原始证据 e_{raw} 进行深度加工, 其目标是将证据句子转化为连接了相关背景知识的向量表示, 并映射为一个质量分数。证据及其背景信息构成的异构图包含证据节点、上下文节点和参考文献节点 3 类节点, 初始特征由预训练模型对相应文本进行编码。证据一上下文边依据文档结构建立, 每个证据节点与其所属上下文文档节点直接相连; 证据一参考文献边基于显式引用或超链接建立, 连接证据节点与其引用的参考文献节点; 证据层内边将同一声明所有证据节点全连接, 以实现多证据推理。所有边均为无向边, 构成消息传递的基础结构。表 1 总结了节点类型与边连接方式。

表 1 三级证据图的节点类型及边连接方式

节点/边类型	描述
证据节点	原始证据句子, 提供直接事实依据
上下文节点	证据的上下文段落, 提供证据的背景信息和语境
参考文献节点	证据引用的链接, 提供额外佐证信息
证据一上下文边	通过共同语义关联连接
证据一参考文献边	通过引用关系连接
证据层内边	通过共同声明主题连接

首先, 构建 3 级证据图, 以原始证据 e_{raw} 作为核心节点, 自动提取其所在文档或页面的相邻段落作为上下文层节点。节点之间通过引用关系和语义相似度进行连接, 形成结构化的异质信息图。

其次, 通过分层图注意力推理进行信息融合^[18]。采用分层的消息传递机制, 首先在每一层内部进行节点聚合, 捕捉局部上下文; 然后进行跨层注意力聚合, 采用分层图注意力机制进行多轮推理。模块包含 12 层 Transformer, 图消息传递从第 1 层开始逐层进行, 让证据节点同时参与到相关的

上下文和参考文献节点。该过程使用优化的图注意力网络实现, 在注意力权重计算中引入了一个可学习的先验项, 该先验项基于节点类型进行偏置, 确保核心证据在信息整合中的主导地位。层内与跨层聚合后, 通过不对称多头自注意力将上下文与参考文献信息融合进证据节点的表示。经过多轮迭代得到编码该证据及其相关背景联合信息的每个证据节点的增强嵌入 h_e^{enhanced} , 为后续的证据质量过滤与事实核查模块提供结构化输入。

最后, 引入证据质量过滤。所有检索到的内容并非同等可靠, 通过一个质量评估器, 根据证据来源的权威性、与声明的直接相关性以及自身陈述的明确性, 为每个增强后的证据赋予一个质量分数 $q(e) \in [0, 1]$ 。将证据加入全局证据池 E 之前, 分数低于阈值 η 的证据将被丢弃, 从而抑制低质量或无关信息的干扰。最终, 通过质量过滤的增强证据嵌入 h_e^{enhanced} 与其对应的质量分数 $q(e)$ 共同构成用于事实核查的证据表示 $(h_e^{\text{enhanced}}, q(e))$, 该表示既包含深度融合的语义信息, 也反映了证据的可信度。

2 实验及分析

2.1 实验设置

本文实验采用了事实核查研究领域两个数据集 Factcheck-Bench^[2] 和 BingCheck^[15] 进行验证。Factcheck-Bench 是一个全面的端到端评估框架, 基于此框架构建了开放领域的事实性评测数据。BingCheck 是一个专门针对大语言模型生成文本设计的事实核查数据集, 其原始分布存在严重的类别不平衡, 包含 3 581 条正确声明和仅 42 条错误声明。为构建一个公平有效的评估集, 采用无放回简单随机抽样从正确声明中抽取 100 条, 同时保留所有 42 条错误声明。该抽样策略缓解原始数据中类别不平衡问题的同时能够确保模型在少数类上的性能也能被准确评估, 且随机抽样保证所选样本在统计意义上能够代表原始正确声明的整体分布。

本文基于 Qwen3-8B 模型以及 DeepSeek-V3.2 模型进行实验。评价指标包括真实/虚假样本的精确率、召回率、F1 值以及模型调用搜索次数。其中 Qwen3-8B 模型进行了 LoRA 微调操作。针对 Qwen3-8B 模型的微调使用 Alpaca^[19] 格式的指令数据进行训练^[20], Alpaca 以 (instruction, input, output) 为结构, 微调数据集为 LIAR-RAW^[21], 该数据集是一个专为可解释性假新闻检测构建的基准数据集。

为优化迭代检索过程的效率,有关参数控制 如表 2 所示。

表 2 检索说明		
实验设置	策略/参数值	策略/参数说明
数据来源	Tavily API (互联网实时搜索)	确保证据的时效性与广泛性
单次返回结果数	3	在提供足量信息与避免处理负担间取得平衡
结果处理方式	使用 Tavily 返回的智能摘要	利用其优化特性,自动提取核心内容
最大搜索迭代轮数	3	避免无限循环,控制总体计算成本
知识上下文保留	最近 3 条结果,总长度≤2 000 字符	防止历史信息过载,确保 LLM 聚焦于最新、最相关的证据
重复搜索次数容忍度	2	当连续搜索查询或结果高度相似时提前终止,提升效率

模型判断是否得出答案的置信度阈值 θ 为 0.7,针对该阈值设定的实验结果如表 3 所示,使用 DeepSeek-V3.2 在 BingCheck 数据集上进行了相关实验。

表 3 小范围内置信度阈值效果对比							
置信度 阈值	搜索 调用 次数	真实样本识别率			虚假样本识别率		
		精确率	召回率	F1	精确率	召回率	F1
0.6	149	0.86	0.89	0.88	0.72	0.67	0.69
0.8	157	0.87	0.91	0.89	0.76	0.67	0.71
0.7(本文)	147	0.87	0.91	0.89	0.76	0.67	0.71

从表 3 可以看出,当阈值设定为 0.6 时,虽模型调用搜索次数相当,但在真实和虚假样本识别率上均有一定下降;而当阈值设定为 0.8 时,其正负样本识别率无明显提升,但增加了模型调用搜索次数。采用 Sentence-BERT(all-MiniLM-L6v2 版

本^{[22])}量化连续生成的搜索查询或检索结果之间的相似程度,借鉴 Shashavali 等^[23]在为控制误报率而设定的 0.9 阈值,在本任务中,0.9 的阈值能精准捕捉重复查询,同时避免误判,实现精确率与召回率的平衡。当查询相似度超过该阈值时,触发迭代过程终止,避免向搜索引擎重复提交相似查询。

2.2 实验结果

为评估本文框架的综合性能,在 BingCheck 与 Factcheck-Bench 数据集上进行实验,并与原始 FIRE^[12]以及 SAFE^[16]等基线方法进行对比。FIRE 是通过动态决策降低计算与搜索成本的迭代检索框架。SAFE 是基于 LLM 的自动化事实核查方法,核心在于通过多轮搜索进行推理对事实进行相关性判断。

基于 Factcheck-Bench 与 BingCheck 数据集的真实/虚假样本类精确率、召回率、F1 值如表 4 所示。

表 4 Factcheck-Bench 与 BingCheck 效果对比								
数据集	框架	检测器模型	真实样本识别率			虚假样本识别率		
			精确率	召回率	F1	精确率	召回率	F1
Factcheck-Bench	SAFE ^[16]	DeepSeek-V3.2	0.92	0.81	0.86	0.58	0.79	0.67
	FIRE ^[12]	Qwen3-8B	0.91	0.78	0.84	0.54	0.77	0.63
		DeepSeek-V3.2	0.90	0.80	0.85	0.55	0.74	0.63
	FIRE+LoRA	Qwen3-8B	0.93	0.85	0.89	0.65	0.81	0.72
	本文方法	Qwen3-8B	0.93	0.86	0.89	0.66	0.81	0.73
		DeepSeek-V3.2	0.92	0.81	0.86	0.58	0.78	0.67
BingCheck	SAFE ^[16]	DeepSeek-V3.2	0.86	0.81	0.84	0.60	0.69	0.64
	FIRE ^[12]	Qwen3-8B	0.86	0.83	0.84	0.62	0.67	0.64
		DeepSeek-V3.2	0.86	0.82	0.84	0.62	0.69	0.65
	FIRE+LoRA	Qwen3-8B	0.88	0.90	0.89	0.74	0.71	0.72
	本文方法	Qwen3-8B	0.89	0.90	0.90	0.76	0.74	0.75
		DeepSeek-V3.2	0.87	0.91	0.89	0.76	0.67	0.71

在 Factcheck-Bench 上,本文方法在真实样本识别方面取得了与 FIRE 相当的表现,在 DeepSeek-V3.2 模型上精确率、召回率与 F1 值分别为 0.92、

0.81 与 0.86,较 FIRE 方法有一定的提升,而在虚假样本识别方面,同样如此。在 Qwen3-8B 模型上,经过 LoRA 微调后表现出更优的虚假识别能力,本文

方法在该模型上的虚假样本识别指标分别达到 0.66、0.81 与 0.73, 对比其他基线方法, 虚假样本的识别指标有明显提高。在 BingCheck 数据集上, 本文方法继续表现出良好的适应性, 在 DeepSeek-V3.2 模型上真实样本识别精确率、召回率与 $F1$ 值分别为 0.87、0.91 和 0.89, $F1$ 值较 SAFE 和 FIRE 方法都有提高。在 Qwen3-8B 模型上也展现出相同趋

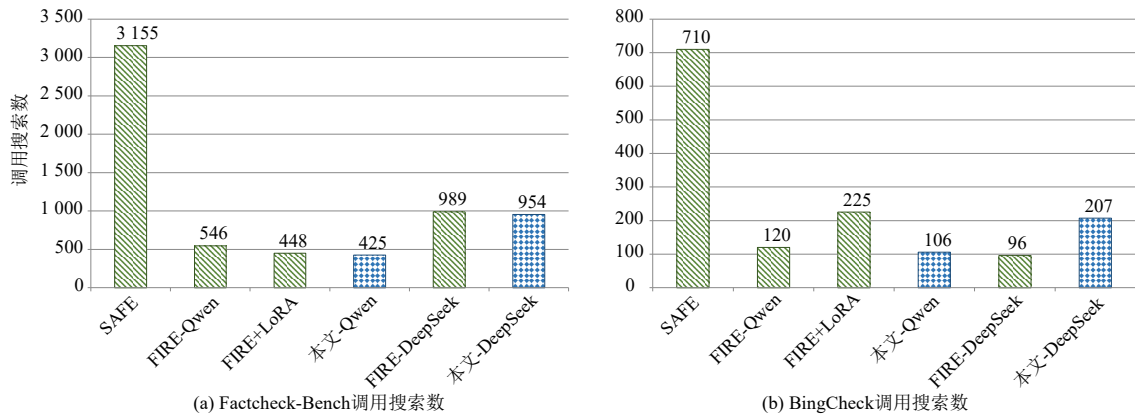


图3 调用搜索数对比图

使用 Qwen 模型时, 在 Factcheck-Bench 上 SAFE 方法共触发 3 155 次搜索, 而本文方法仅触发 425 次, 即 SAFE 方法调用搜索数约为本文方法的 7.4 倍。在 BingCheck 上趋势同样明显。而在使用 DeepSeek 模型时, 相较 SAFE 方法本文方法的搜索次数分别降低约 69.7% 和 70.6%。同时, 在两个模型和数据集上, 本文方法的搜索次数均较 FIRE 方法进一步降低。本文方法在保持较高识别性能的同时, 显著降低了对外部检索的依赖, 说明将迭代检索与深度证据推理相结合的策略有效, 且动态决策机制在成本控制上拥有优势。

表 5 列出了部分模型判断成功与失败的案例。从案例可见, 该方法在具有高共识度的声明上表现稳健, 如“放松训练有助于缓解压力”属普遍共

势的提高。本文方法在两个模型上的真实召回率均达到 0.90 以上, 显示出较强的真实样本覆盖能力。

本文与基线方法在两个数据集上调用搜索数对比柱状图如图 3 所示, 其中: 横坐标代表各个方法; 纵坐标代表搜索次数; 蓝色格纹柱形表示本文方法。

识性判断, 模型在不进行检索的情况下亦能给出正确结论; 当声明包含明确线索时, 如“AAP”与“成人”同时存在, 模型能够构造针对性查询并据此识别该声明中指南适用范围的不匹配, 从而做出否定判断。

相对地, 方法在涉及数值精确度与表述粒度的反例上易出现误判, 例如“马里亚纳海沟深度为 11 km”在公开资料中多以“约 11 km”呈现, 检索结果的近似一致性会被模型吸收为正向证据, 未能区分测量不确定性与标签标准的严格等同性要求。

同时, 由总体识别结果可知, 模型针对虚假样本识别率偏低, 而真实识别率较高, 可见模型存在将声明判为真的系统性偏向。

表5 部分成功与失败案例

声明	原始标签	模型搜索	模型回答
Relaxation techniques such as deep breathing, meditation, mindfulness, or progressive muscle relaxation can help cope with stress.	True		True
The American Academy of Pediatrics recommends no more than two hours of screen time per day for adults.	False	American Academy of Pediatrics screen time recommendations for adults	False
The deepest known area of the ocean is the Mariana Trench, which has a depth of 11 km.	False	Mariana Trench depth in kilometers verified measurements	True

2.3 消融实验

为验证本文各模块的有效性, 使用 DeepSeek-V3.2 在 BingCheck 数据集上进行消融实验, 实验设

置无检索、无上下文语义增强和无质量过滤 3 个条件, 并与本文完整框架效果对比, 以评估各模型在事实核查任务中的必要性, 实验结果如表 6 所示。

表6 消融实验效果对比

方法	真实样本识别率			虚假样本识别率		
	精确率	召回率	F1	精确率	召回率	F1
无搜索	0.83	0.90	0.87	0.71	0.56	0.63
无上下文语义增强	0.86	0.82	0.84	0.62	0.69	0.65
无质量过滤	0.85	0.86	0.85	0.65	0.64	0.65
本文	0.87	0.91	0.89	0.76	0.67	0.71

在无搜索的情况下,LLM 倾向于将更多样本判断为真实,虚假样本的识别受到严重影响,其精确率与召回率均显著下降,表明 LLM 在缺乏外部实时证据支持时,对于虚假声明的识别能力有限。无上下文语义增强模块时,框架缺乏对检索结果与待核查声明之间深层次语义的分析,框架的精确率与 F1 值均未达到最优水平,说明仅依靠检索获取的证据,难以充分支撑模型对声明做出判断。无质量过滤模块时,框架虚假精确率低于完整模型,说明未经筛选的检索结果中可能包含无关或低可信度的信息,干扰模型决策,降低判断的准确性。本文集成迭代检索、上下文语义增强与证据质量过滤的完整框架在各项指标上实现了最佳平衡。消融实验验证了各模块设计的有效性,3 个关键模块共同提升了本文自动事实核查框架对于声明的识别能力。

3 结束语

本文提出一种融合动态迭代检索与上下文增强推理机制的自动事实核查框架,旨在系统性解决开放域自动化事实核查中效率与准确性难以兼顾的问题,提供一种平衡自动化事实核查中的效率与准确性问题的方案。该框架通过多阶段自适应的证据获取与深度语义理解策略,提升了 LLM 推理可靠性与决策可解释性。实验结果表明,本框架在 Factcheck-Bench 与 BingCheck 两个数据集上均表现出稳定的综合性能,在维持高准确性的同时降低了对外部检索资源的依赖与调用搜索开销,验证了其维持高准确性的同时降低计算与检索成本,为进一步探索融合多模态证据、引入长期记忆机制以及面向开放域复杂声明的可扩展验证系统提供参考。

参考文献:

[1] WANG Z R, CHENG Z J, ZHU H, et al. What are tools anyway? a survey from the language model perspective [DB/OL]. (2024-03-18) [2026-03-17]. <https://arxiv.org/abs/2403.15452>.

[2] WANG Y X, GANGI REDDY R, MUJAHID Z M, et al. Factcheck-bench: fine-grained evaluation benchmark for automatic fact-checkers [C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, USA: ACL, 2024:14199-14230.

[3] CHENG X X, LI J Y, ZHAO W X, et al. Small agent can also rock! empowering small language models as hallucination detector [C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2024:14600-14615.

[4] CHERN I C, CHERN S, CHEN S Q, et al. FacTool: factuality detection in generative AI: a tool augmented framework for multi-task and multi-domain scenarios [DB/OL]. (2023-07-26) [2026-01-20]. <https://arxiv.org/abs/2307.13528>.

[5] QUELLE D, BOVET A. The perils and promises of fact-checking with large language models [J]. *Frontiers in Artificial Intelligence*, 2024, 7: No.1341697.

[6] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York, USA: ACM, 2020:9459-9474.

[7] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models [DB/OL]. (2021-10-16) [2026-01-25]. <https://arxiv.org/abs/2106.09685>.

[8] SUN X X, LI J P, ZHONG Y, et al. Towards detecting LLMs hallucination *via* Markov chain-based multi-agent debate framework [C]//Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, USA: IEEE, 2025:1-5.

[9] WANG H R, SHU K. Explainable claim verification *via* knowledge-grounded reasoning with large language models [C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, USA: ACL, 2023: 6288-6304.

[10] ZHENG L W, LI C Z, ZHANG X, et al. Evidence retrieval is almost all you need for fact verification [C]//Findings of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2024:9274-9281.

[11] NIU C, GUAN Y, WU Y H, et al. VeraCT scan: retrieval-augmented fake news detection with justifiable reasoning [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2024:266-277.

[12] XIE Z H, XING R, WANG Y X, et al. FIRE: fact-checking with iterative retrieval and verification [C]//Findings of the Association for Computational Linguistics: NAACL 2025. Stroudsburg, USA: ACL, 2025:2901-2914.

[13] ZHANG D C, LEE D. CORRECT: context- and reference-

- augmented reasoning and prompting for fact-checking[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: ACL, 2025:3007-3019.
- [14] 陈静,周刚,但文皓.知识增强的异质图注意力虚假信息检测方法[J].信息工程大学学报,2025,26(3):323-329.
- [15] LI M R, PENG B L, GALLEY M, et al. Self-checker: plug-and-play modules for fact-checking with large language models[C]//Findings of the Association for Computational Linguistics: NAACL 2024. Stroudsburg, USA: ACL, 2024:163-181.
- [16] DU C, HU N, HUANG D, et al. Long-form factuality in large language models[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Cambridge, USA: NeurIPS, 2024:80756-80827.
- [17] 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, USA: ACL, 2023:12076-12100.
- [18] LIU Z H, XIONG C Y, SUN M S, et al. Fine-grained fact verification with kernel graph attention network[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2020:7342-7351.
- [19] TAORI R, GULRAJANI I. Alpaca: a strong, replicable instruction-following model [EB/OL]. (2023-03-13) [2026-01-20]. <https://crfm.stanford.edu/2023/03/13/alpaca.html>.
- [20] CHEUNG T H, LAM K M. FactLLaMA: optimizing instruction-following language models with external knowledge for automated fact-checking[C]//Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. Piscataway, USA: IEEE, 2023:846-853.
- [21] YANG Z W, MA J, CHEN H C, et al. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection[C]//Proceedings of the 29th International Conference on Computational Linguistics. Stroudsburg, USA: ACL, 2022:2608-2621.
- [22] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: ACL, 2019:3982-3992.
- [23] SHASHAVALI D, VISHWJEET V, KUMAR R, et al. Sentence similarity techniques for short vs variable length text using word embeddings[J]. Computación y Sistemas, 2019,23(3). DOI:10.13053/cys-23-3-3273.

(编辑:冯春)

(上接第266页)

- [15] WEI W, FU S, TANG Y J, et al. Joint optimization of UAV aided covert edge computing *via* a deep reinforcement learning framework[J]. Chinese Journal of Aeronautics, 2025,38(10):No.1103218.
- [16] 方维维,王云鹏,张昊,等.基于多智能体深度强化学习的车联网通信资源分配优化[J].北京交通大学学报,2022,46(2):64-72.
- [17] SHAKYA A K, PILLAI G, CHAKRABARTY S. Reinforcement learning algorithms: a brief survey[J]. Expert Systems with Applications, 2023,231:No.120495.
- [18] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015,518(7540):529-533.
- [19] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[J]. AAAI Press, 2016,30(1). DOI:10.1609/aaai.v30i1.10295.
- [20] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning. [J]. Computerence, 2015. DOI:10.1016/S1098-3015(10)67722-4.
- [21] BELLVER M, GIRO-I-NIETO X, MARQUES F, et al. Hierarchical object detection with deep reinforcement learning [DB/OL]. (2016-11-11) [2026-04-09]. <https://arxiv.org/abs/1611.03718>.
- [22] WANG Y, WU J D, HE H W, et al. Data-driven energy management for electric vehicles using offline reinforcement learning[J]. Nature Communications, 2025,16:No.2835.

(编辑:冯春)