证据链增强的多智能体分组讨论事实核查框架研究*

相蔓蔓[1,2]　纪泽旭[1]　郭佳鸣[1]　王秀琨[1]
（1.吉林财经大学人工智能与计算机学院，长春 130117；2.吉林财经大学大数据与交叉科学研究院，长春 130117）

摘要：事实核查是甄别虚假信息的重要手段，而单一大语言模型在事实核查中存在“幻觉”问题、可解释性不足、准确率低等局限，本研究旨在提出一种证据链增强的多智能体分组讨论事实核查（Evidence Chain Enhanced Multi-Agent Group Discussion for Fact-Checking，ECEMA-GD-FC）框架，以期有效提升事实核查任务的准确性、鲁棒性与可解释性。ECEMA-GD-FC框架由主持人智能体负责任务调度、结果整合与共识裁决，下设科学证据、媒体证据与事实证据三个核查智能体小组，各小组内部遵循证据检索、证据链构建、真伪判定、报告生成的标准化流程，实现“统筹—分工—聚合”闭环核查机制。基于事实核查数据集的实验结果表明，本研究所提ECEMA-GD-FC框架在真伪判定准确率与核查报告生成质量上展现出良好的性能，能显著提升事实核查结果的可信度。

关键词：虚假信息；事实核查；多智能体；证据链；大语言模型
中图分类号：TP18　DOI：10.3772/j.issn.1673-2286.2026.05.003
引文格式：相蔓蔓，纪泽旭，郭佳鸣，等. 证据链增强的多智能体分组讨论事实核查框架研究[J]. 数字图书馆论坛，2026，22（5）：20-33.

在数字化与智能化深度融合的时代背景下，信息传播的即时性与广度空前提升，虚假信息与错误陈述的生成与扩散也呈现低成本、高速度、跨平台的特征，对公众认知、社会信任乃至国家安全构成持续性风险。根据微博发布的《2025微博社区信任报告》[1]，2025年微博共有效处置清理谣言信息31万余条。在此背景下，国家在多项政策文件中持续强调加强网络信息治理与谣言防控，其中《网络信息内容生态治理规定》[2]明确要求网络信息内容服务平台建立网络谣言处置制度。虚假信息和错误信息在公共卫生[3]、政治决策[4]与社会治理[5]等领域尤为突出，极易造成认知偏差，不仅诱发个体健康风险，还会干扰政策落地、加剧社会分歧，带来多重治理风险。因此，构建能够动态识别、精准核实并清晰解释的自动化事实核查系统已成为信息治理与情报支撑领域亟待突破的关键课题。

事实核查（Fact-Checking）指基于证据验证公共领域主张准确性的方法[6]。传统事实核查高度依赖专业人员的经验判断与手工操作[7]，在面对高速产生的海量信息流时，其效率瓶颈日益凸显。基于规则、统计机器学习及深度学习方法的自动化事实核查相关研究因而不断涌现[8-12]。此类方法虽显著提升了处理效率，但在语义深度理解、复杂语境推理及过程可解释性等方面仍存在局限。大语言模型（Large Language Model，LLM）凭借强大的语义理解与生成式推理能力，为事

收稿日期：2026-04-08
*本研究得到国家社会科学基金一般项目“认知和心理账户视角下社交媒体用户隐私悖论行为及干预机制研究”（编号：21BTQ057）资助。

实核查提供了新的技术路径。通过证据检索、链式推理与结论生成的一体化建模，LLM能够处理原始声明并生成解释文本[13]。然而，单一LLM在事实核查任务中常受“幻觉”问题困扰，即生成缺乏事实依据或证据相互矛盾的内容，其“黑箱”式的推理过程也削弱了结果的可信度与可追溯性。

目前，已有研究[14]构建了大模型驱动的事实核查多智能体系统框架，通过角色分工与协作流程设计，实现了从单一模型到多智能体协同的初步探索。然而，该框架侧重于智能体间的角色分工与讨论流程设计，其协同机制主要依赖智能体的自由讨论与主持人引导，对证据本身的实时检索、逻辑组织与结构化约束尚未建立明确的机制，且多智能体的自由讨论削弱了多视角独立校验的客观性。故本研究提出一种证据链增强的多智能体分组讨论事实核查（Evidence Chain Enhanced Multi-Agent Group Discussion for Fact-Checking，ECEMA-GD-FC）框架，通过引入证据链机制，设计证据检索、证据链构建、真伪判定、报告生成的标准化流程以约束各智能体的推理轨迹，提升结论的可追溯性；同时采用分组讨论，设立科学证据、媒体证据与事实证据三个核查小组并行工作，使不同证据维度的校验在独立状态下完成，避免群体讨论中常见的“从众趋同”现象对事实核查客观性的干扰。

1　相关研究

1.1　事实核查方法

作为应对虚假信息、错误陈述等信息失序现象的核心治理技术，早期事实核查方法以特征工程与规则匹配为主，通过构建词典、提取句法特征或风格特征，并结合传统机器学习模型实现对文本的真伪分类[8–9]。这类方法在结构规整、特征显性的文本中具有一定效果，但由于严重依赖专家知识，难以捕捉深层语义关联，无法应对刻意规避检测的隐性虚假信息或复杂修辞。随着神经网络、预训练语言模型（如BERT、RoBERTa）的兴起，事实核查研究迈入了以语义表征学习为核心的技术阶段。研究者通过构建大规模标注数据集（如SciFact、MR^2），将任务形式化为“声明–证据”（Claim-Evidence）对的自然语言推理问题，并利用模型的上下文理解能力进行端到端训练[10–12,15]。这一阶段显著提升了事实核查模型在多领域、复杂句式下的判别性能，但其本质仍是一种基于语义相关性的模式匹配，缺乏对证据进行显性逻辑组织与因果推理的能力，模型决策过程不可追溯，且可解释性较弱。近年来出现的以GPT（Generative Pre-trained Transformer）系列为代表的LLM凭借强大的语义理解、逻辑推理与检索能力，为自动化事实核查构建了全新技术范式。通过设计思维链（Chain-of-Thought，CoT）提示、引入检索增强生成（Retrieval-Augmented Generation，RAG）技术，LLM能够自主执行证据检索、组织与结论生成，输出具备一定解释性的事实核查报告[16–17]。然而单一LLM存在“幻觉”，成为高风险事实核查场景下难以突破的瓶颈。针对这一不足，本研究从证据组织与协同验证的角度出发，引入多智能体协同机制，以期提升事实核查准确性，同时增强推理过程的可解释性。

1.2　多智能体系统在事实核查中的应用

多智能体系统（Multi-Agent System，MAS）源于人工智能（Artificial Intelligence，AI）与复杂系统科学两个领域，是多个自主智能体协作解决复杂任务的系统[18]，已成为研究群体决策、协作行为与演化规律的重要方法论工具[19]，被应用于地震应急指挥[20]、网络舆情治理[21]等现实场景中。早期MAS研究多基于博弈论与复杂系统理论，通常将智能体设定为遵循固定规则的理性决策者，通过定义收益函数与约束条件，分析个体决策在群体层面的均衡状态。相关成果揭示了MAS在信息共享[22]、竞争协作[23]与监管约束[24]等情境中的策略演化，为理解多主体交互提供了理论框架，但在处理主体间复杂交互与非线性机制方面仍存在局限。与此同时，元胞自动机理论被引入MAS以模拟信息传播、情绪扩散等宏观涌现现象，通过定义局部交互规则与主体状态转移，刻画舆情演化、空间扩散等动态过程[25–26]，并在模型中引入异质主体与动态网络结构，以增强模型对现实场景的刻画精度。但该类方法规则依赖性强，对文本深层语义与高阶认知推理的表征能力较为局限。近年来，LLM驱动的智能体构建成为研究热点。LLM凭借强大的语义理解、逻辑推理与生成能力，为智能体赋予了在开放环境中进行对话交互、任务规划与知识融合的潜力[27]，在领域问答、信息理解与决策支持等任务中展现出更高的可靠性与适应性[28–29]。由于单一智能体的能力仍显不足，研究者开始探索将LLM与MAS相结合，

通过角色分工、任务分解与协同辩论，模拟人类专家团队的核查流程，在应急情报[30]、国防科技[31]、食品安全[32]与工程仿真[33]等领域展现出了良好的应用前景。

综上所述，尽管LLM驱动的MAS已在诸多领域展现出潜力，但现有研究也存在一定局限。首先，部分研究关注智能体角色分工与协作流程设计而缺乏系统性能评估，或追求判别准确性而忽视推理过程的可解释性；其次，现有多智能体协作多侧重观点交互与结果融合，对协作过程中的多源证据组织与逻辑约束关注仍显不足，导致证据与判定结果之间的关联不够清晰。因此，本研究将证据链与MAS融入事实核查任务中，在提升事实核查准确率与鲁棒性的同时，增强核查过程的可解释性与结果可信度。

2 研究框架

本研究所述的事实核查旨在验证公共领域主张的准确性，核查对象涵盖虚假信息、错误信息、谣言这三类信息中具有可验证性的事实性陈述。本研究构建的事实核查框架（见图1）由一个主持人智能体和三个专业化核查智能体小组构成，形成“统筹—分工—聚合”的协同工作模式。主持人智能体负责全局任务调度、流程控制与结果整合；三个小组分别聚焦于科学证据、媒体证据和事实证据三个维度开展并行核查，每个小组内置4个智能体，分别是证据搜集智能体、证据链搭建智能体、真伪判定智能体、解释报告生成智能体。该分组方式能够覆盖事实核查中主要的证据来源类型及其对应的验证逻辑，构成具有代表性且近似完备的证据空间划分。各小组在提示词中通过偏差参数（bias）控制引导检索与分析视角，在统一流程下实现多源证据的交叉验证与视角互补。其中，科学证据小组强调学术共识与可重复性，依赖同行评议、实验数据等权威来源；媒体证据小组注重信源可信度与报道一致性，需要对比主流媒体、官方通报等多渠道信息；事实证据小组依托原始记录与确凿材料（如官方记录、公开数据库、档案等），通过核验材料中的关键事实要素来判定真伪。各小组依据其专业维度，遵循证据检索、证据链构建、真伪判定、报告生成的标准化流程开展工作，最终由主持人智能体汇总各小组结论，通过一致性整合机制形成兼具结构完整与解释性的事实核查报告。

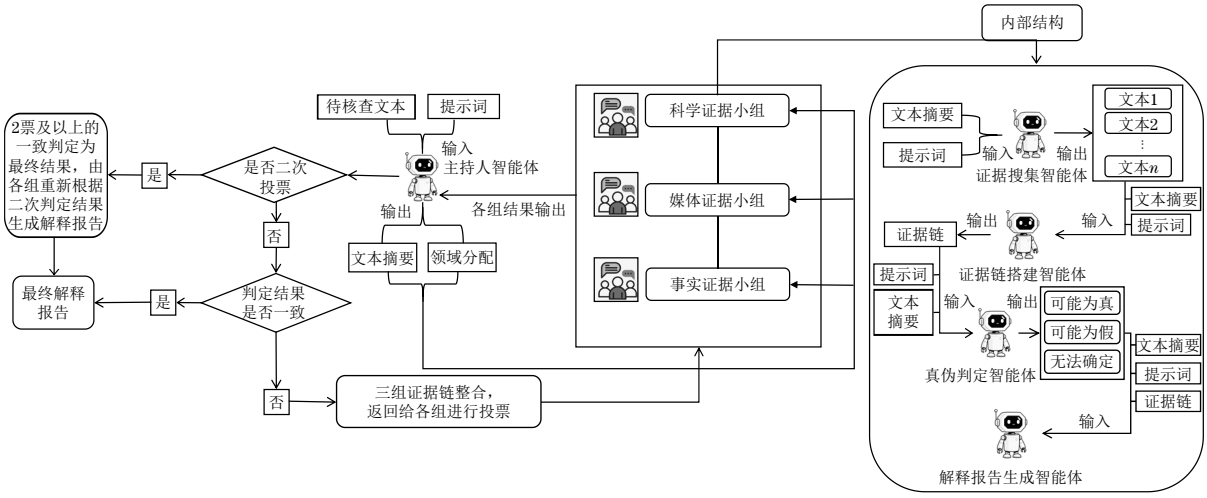

图1 事实核查框架

在事实核查框架中，智能体作为一类能通过灵活自主行动（情境性、自主性、响应性、主动性、社会性）实现设计目标的计算机系统[34]，遵循统一的构建流程，具体步骤如下。①明确目标与能力范围。为每个智能体定义唯一的角色定位与任务边界，例如：证据搜集智能体仅负责检索与提取原始证据；真伪判定智能体仅基于证据链输出判定标签。职责分离旨在降低单一智能体的认知负载。②构建工具箱。为智能体配备可调用的外部工具集合，具体包括：用于证据检索的网络搜索工具（通过智谱AI提供的search_std标准化接口调用）；用于存储本小组已获取候选证据的内部证据缓存；存放各环节结构化提示词的提示词模板库。③集成LLM。所有智能体底层由统一的LLM实例驱动，通过应用程序编程接口（Application Programming Interface，API）调用完成语义理解与文本生成。④添加记忆机制。为智能体设计短期记忆机制，在同一核查任务的

生命周期内，智能体将前序环节的输出（如证据检索结果、证据链中间状态）暂存于会话上下文，供后续环节调用。⑤部署与迭代。将所有智能体部署于本地服务器，通过API对外提供服务。使用者可在框架部署后根据日志分析，持续优化提示词模板、调整检索策略。

2.1　文本摘要与领域分配

文本摘要与领域分配是事实核查框架的初始环节，由主持人智能体执行，主要完成以下任务：首先，将输入的待核查文本提炼为一句简要的文本摘要，作为核查主张；其次，判断该主张所属领域（如医学、社会、科技等），并将该领域信息作为上下文提示词分发给各核查小组，以增强核查针对性。为更符合事实核查的逻辑结构，充分利用LLM的语义理解和内容生成功能，本研究设计主持人智能体初始环节提示词模板（见图2）。

2.2　分组讨论机制

三个专业化核查智能体小组内部均包含四类智能体，各智能体依次负责证据检索、证据链构建、真伪判定、报告生成等任务。

（1）证据搜集智能体。证据搜集智能体基于主持人智能体提供的文本摘要信息，通过调用大模型集成的搜索工具进行网络搜索，各专业化核查智能体小组通过偏差参数（bias）引导检索方向。具体而言，科学证据小组优先检索同行评审的研究报告、实验设计等学术证据；媒体证据小组侧重官方声明、主流媒体和专业核查机构意见等信息；事实证据小组关注原始记录、案例报告等一手材料。证据搜集完毕后，证据搜集智能体再进行结构化处理，提取标题、摘要、URL等关键字段，并以此结构将证据传递给证据链搭建智能体，用于支持证据链的构建与逻辑推理。

（2）证据链搭建智能体。证据链搭建智能体将证据搜集智能体提供的碎片化证据组织为语义连贯、结构严谨的证据链，为后续真伪判定任务提供系统化证据支撑。首先，深度理解待核查主张，识别其中的核心断言、关键实体（如人物、地点、事件、时间）和关系；其次，依据检索结果的来源权威性与待核查任务的语义相关性，结合代码逻辑与提示词中内置的置信度评分机制，为每条证据赋予评分（confidence），筛选出高质量证据；最后，将筛选得到的证据按照逻辑关系（如因果关系、时序关系、支持关系）组织成链式结构，构建证据链。为避免大模型“幻觉”问题，本研究限定证据链的搭建仅可以使用已检索到的证据语句，禁止模型引入外部知识或自行生成证据内容。为实现上述功能，本研究设计证据链搭建智能体提示词模板（见图3）。

（3）真伪判定智能体。真伪判定是事实核查任务的核心决策阶段，真伪判定智能体直接对核查内容的真实性作出权威性判断。基于证据链搭建智能体构建的证据链，真伪判定智能体对主张的真实性进行判定，输出以下三类标签之一，并附置信度评分（0~1）。①可能为真（likely_true）。当证据链中包含充分、一致且可信度较高的佐证信息，与待核查内容形成高度逻辑契合时，判定为该类别。②可能为假（likely_false）。当证据链中存在明确、可靠的反驳或矛盾信息，足以质疑或否定待核查主张的关键要素时，判定为该类别。③无法确定（uncertain）。当证据链不完整、证据之间存在冲突或缺乏足够可靠的信息支撑时，判定为该类别。为确保任务执行过程的精准性，本研究设计真伪判定智能体提示词模板（见图4）。

（4）解释报告生成智能体。报告生成是事实核查任务的最终环节，解释报告生成智能体基于前期的证据链和真伪判定结果，生成一段逻辑清晰、内容真

```
你是主持人智能体，收到待核查信息如下：
{raw_text}
任务：
1.用一句话概括该信息的核心待核查内容，用肯定句进行描述。原文有时间信息则纳入概括，不要使用“是否”等字样，不要改变原有的意思，禁止自行增删内容造成指向性误导。
2.判断所属领域（医学/科技/社会/政治等），只输出简短领域标签，不要附加多余解释性文字。
请返回严格可解析的JSON对象，不要输出JSON以外的思考、说明、注释等任何额外文本：
{
  "summary": "…",
  "domain": "…"
}
```

图2　主持人智能体初始环节提示词模板

```
你是证据链搭建智能体。以下是搜索结果（按索引编号）：{search_text}
任务：
-基于这些结果构建一条逻辑清晰的证据链，禁止引入外部知识或自行生成证据内容。
-对每条证据评估其与原始主张的关系，只能三选一：支持（support）/反驳（contradict）/
中立（neutral）。
-请不要在此步骤直接给出最终真假判断（真伪判定智能体会完成判断）。
-若传入bias参数，请优先采纳该bias指定的证据类型；bias为空时，不设置证据类型偏向。
Bias (if any): {bias}
请返回严格可解析的JSON对象，不要输出JSON以外的思考、说明、注释等任何额外文本：
{
  "evidence_chain_text": "…",
  "structured": [
    "index": null,
    "title": "…",
    "claim_excerpt": "…",
    "support": "support/contradict/neutral",
    "confidence": 0.00]
}
```

图3　证据链搭建智能体提示词模板

```
你是真伪判定智能体。根据以下证据链判断信息真伪（请直接给出对原始声明的结论）。
原始声明：{original_claim}
证据链：{evidence_text}
请仔细分析每条证据对于原始问题的支持 / 反驳关系，基于多数证据倾向做出判断。
请返回严格可解析的JSON对象，不要输出JSON以外的思考、说明、注释等任何额外文本：
{
  "verdict": "likely_true|likely_false|uncertain",
  "confidence": 0.00
}
```

图4　真伪判定智能体提示词模板

实、语义连贯的自然语言解释报告，旨在清晰、客观地阐述事实判定背后的依据与完整推理过程，从而增强事实核查结论的可解释性与可信度。输出报告格式为Markdown或纯文本，内容包括：report_text（完整解释报告文本，内容涵盖主张摘要、证据链和判定结果的综合分析）、key_points（关键要点列表，提取报告中的核心观点和重要信息）以及recommendation（基于分析结果给出的建议或后续行动指导）。为确保报告内容的客观性、平衡性与事实准确性，本研究设计解释报告生成智能体提示词模板（见图5）。

```
你是解释报告生成智能体，请根据以下信息生成解释报告（以便主持人智能体汇总）：
主张摘要：{summary}
证据链：{evidence}
判定结果：{judgement_result}
Bias: {bias}
任务要求：
基于上面全部输入信息生成事实核查解释报告，禁止引入输入之外未经核实的外部知识。
report_text输出完整可读纯文本报告，不要使用Markdown语法；key_points为字符串数组，
输出本次核查的核心要点；recommendation输出面向使用者的核查提示语建议。
请返回严格可解析的JSON对象，不要输出JSON以外的思考、说明、注释等额外文本：
{
  "report_text": "…",
  "key_points": […],
  "recommendation": "…"
}
```

图5　解释报告生成智能体提示词模板

2.3　协同决策与共识生成

科学证据小组、媒体证据小组、事实证据小组完成核查后，由主持人智能体负责汇总结果，并通过两轮协同决策与仲裁机制处理分歧，最终达成共识。若三个小组判定结果完全一致，主持人智能体将直接对各小组提交的解释报告进行整合，形成最终解释报告。若三个小组判定结果不完全一致，由主持人智能体整合各小组提交的解释报告，将其反馈给各小组，启动第二轮审议与投票。在第二轮投票中，若三个小组达成共识，则根据共识重新由主持人智能体整合输出最终解释报告。若经过两轮审议仍未达成共识，则启动多数裁决规则，采纳至少2票支持的结果作为最终结论，由主持人智能体生成最终解释报告。

为确保上述三个小组协同判定流程的顺利执行，系统通过内部程序逻辑实现对三个小组间结果一致性的自动判断与流程控制。因此，本研究设计主持人智能体裁决环节提示词模板（见图6）。该环节一致性校验机制由代码层面实现，故未直接体现在LLM的提示词设计中。

```
你是主持人智能体，负责融合多组独立报告观点并生成统一立场的终极分析报告。
输入各组独立报告内容，请整合：{merged_input}；二轮投票结果（若有）：{{g['group']:
g.get('d_out_second_round', g['d_out']) for g in group_reports}}，基于该多数结果生成其
解释报告。
任务：
-开头给出最终结论（Final verdict），仅可三选一：likely_true|likely_false|uncertain；并
输出0~1的置信度（confidence）。
-接着用简洁段落撰写综合理由（50~200字）。
-最后给出简短的建议（行动或标注建议）。
-报告要面向读者，可直接发布（避免使用内部标注格式）。
-不需要输出三组的逐条细节（可在appendix字段简短列出三组verdict和置信度）。
-禁止使用输入以外未经核实的外部知识。
请返回严格可解析的JSON对象，不要输出JSON以外的思考、说明、注释等任何额外文本：
{
  "final_report_text": "…",
  "final_verdict": "likely_true|likely_false|uncertain",
  "final_confidence": 0.00
}
```

图6　主持人智能体裁决环节提示词模板

3　实证研究

3.1　数据集和评价指标

为验证ECEMA-GD-FC框架的有效性，本研究构建了一个兼具历史代表性与时效性的中文事实核查实验数据集。以公开的CHEF[35]中文事实核查数据集（约10 000条样本）为基础，考虑该数据集发布时间相对较早，为反映近期互联网可核查信息特征，从中国互联网联合辟谣平台爬取2024年1月1日—2025年12月31日辟谣数据作为补充。将所有数据合并后，本研究采取分层抽样方法，从“真实”与“虚假”两个类别中分别抽取样本500份构成实验集。该数据集通过人工分类，主要涵盖社会、政治、文化与公共卫生等多个领域，文本噪声程度高、表达形式多样，能较全面地反映真实场景下事实核查任务的复杂性。

本研究设计了一个综合性双维度评价指标体系，分别用于衡量ECEMA-GD-FC框架的性能和解释报告质量。其中，性能评估指标使用准确率、精确率、召回率和F1值，基于样本的真实标签与模型的判定结果进行计算；解释报告质量评估指标采用人工主观评分法，从证据支撑充分性、逻辑性与条理性、语言可读性、整体说服力四个维度进行评价。每项分值范围为1~10分，代表从“差”到“优”。需要说明的是，ECEMA-GD-FC框架的真伪判定输出包含“likely_true”“likely_false”“uncertain”三类，在计算准确率、精确率、召回率和F1值等分类指标时，由于“uncertain”样本极少故予以剔除，将判定为“likely_true”或“likely_false”的样本纳入计算。

为提升评价的稳健性与客观性，本研究采用人机协同的评估机制。首先，邀请3位长期从事信息治理与事实核查研究的专家独立进行评分。相关方法学研究

表明，在专家评审或共识评估研究中，专家数量为3人即可形成较为稳定的评价结果[36]。其次，引入LLM作为辅助评价主体，选取豆包doubao-seed-1.6-lite（以下简称“豆包”）进行评分，因为豆包在中文语义理解与文本评价任务中具有较强表现，并在开放环境中表现出良好的稳定性与可复现性。专家与豆包分别独立完成评分后，最终解释报告质量得分取3位专家评分均值与豆包评分的综合平均值，以降低单一评价主体可能带来的主观偏差，从而提升整体评价结果的可靠性。经计算得出，专家评分与豆包评分的平均Pearson相关系数为0.79，说明人机协同评分具有可接受的线性相关性，主观偏差风险可控。

3.2 实验设置

本研究提出框架的核心智能体均通过调用LLM应用程序接口驱动。为评估框架在不同LLM下的普适性及性能差异，本研究遵循以下原则选择LLM开展实验：①属于当前主流的中文大语言模型系列，在公开评测基准（如CFLUE[37]等中文语言理解基准）中表现出较强的语义理解与推理能力；②模型分别来自不同研发机构，在一定程度上避免单一技术路线带来的偏差；③模型均提供稳定的API并支持多轮对话生成，能够满足多智能体协同推理场景下的实验需求。经过筛选，最终选择qwen-flash、Baichuan4-Air、glm-4.5-flash和deepseek-v3.2这4个具有代表性的LLM。

通过在统一实验框架下对不同模型进行测试，可以更客观地评估本研究所提出方法对模型能力差异的适应性。在实验过程中，所有智能体均采用同一款LLM实例驱动，以确保交互逻辑的一致性并排除模型自身能力对框架性能的影响。所有模型调用均采用统一的超参数设置以保障实验的可重复性：温度系数为1，以维持适度的生成多样性；Top_p核采样阈值为1，以保证生成结果的全面性；频率惩罚和存在惩罚均为0，以避免对词项重复进行额外干预，保持不同模型在默认生成策略下的自然输出特征；同时设置生成的最大词元（token）为4 096，以满足复杂语义生成与多轮对话的需求。

证据检索阶段统一调用智谱AI提供的标准化网络搜索接口search_std，因其能够在不同实验运行周期中提供结构化且稳定的外部信息检索API，从而保证证据来源形式与返回结果结构的一致性，降低外部检索环境变化对实验结果的影响。

3.3 成本计算

ECEMA-GD-FC框架单次核查平均调用大语言模型API共14次（主持人智能体2次，科学证据小组、媒体证据小组、事实证据小组各4次），调用网络搜索工具3次（科学证据小组、媒体证据小组、事实证据小组各1次）。基于qwen-flash模型定价（输入0.15元/百万词元，输出1.5元/百万词元），由于单次核查平均输入2 000~5 000词元、输出1 000~1 500词元，单次调用API成本为0.001～0.003元；搜索工具调用成本为0.030元。

3.4 整体性能评估与分析

不同LLM驱动下ECEMA-GD-FC框架的实验结果如表1所示。qwen-flash模型驱动下ECEMA-GD-FC框架的综合性能表现优秀，F1值（0.848）和准确率（0.842）最高，且精确率（0.822）与召回率（0.876）均保持在较高水平，尤其在召回率上领先，表明其漏报率最低。相比之下，glm-4.5-flash模型驱动下ECEMA-GD-FC框架在精确率指标上表现突出（0.926），说明该模型在事实核查过程中具有更高的判别严格性，但其召回率最低。deepseek-v3.2与Baichuan4-Air模型驱动下ECEMA-GD-FC框架均呈现出高精确率、低召回率的特征，表明这两个模型在事实核查过程中更倾向于保守策略，从而导致部分虚假信息未能被有效识别。

在解释报告质量方面，各组别的解释报告综合评

表1 不同LLM驱动下ECEMA-GD-FC框架的实验结果

组别	准确率	精确率	召回率	F1值	解释报告综合评分
Baichuan4-Air	0.757	0.877	0.566	0.688	32.355
qwen-flash	**0.842**	0.822	**0.876**	**0.848**	37.269
glm-4.5-flash	0.738	**0.926**	0.510	0.658	33.668
deepseek-v3.2	0.815	0.906	0.711	0.797	**37.340**

注：表中加粗数值为对应指标下的最优性能值。

分与F1值排名不一致：qwen-flash的F1值最高，但其综合评分略低于deepseek-v3.2，表明LLM在事实判定任务中的推理能力与其解释文本生成能力之间存在一定的正相关关系，但综合评分还受到证据支撑充分性、逻辑性与条理性、语言可读性以及整体说服力等多维度因素的影响。

3.5　消融实验

为验证ECEMA-GD-FC框架中关键组件的有效性，本研究选取综合性能最优的qwen-flash模型作为驱动模型，设计了三组消融实验：实验组1移除证据链构建环节，智能体直接基于检索到的原始证据进行判定；实验组2不使用专业化核查智能体小组协同，即仅保留科学证据小组进行单一路径核查；实验组3不采用第二轮共识裁决，即当多智能体结果不一致时，直接采用简单多数决定，不进行第二轮审议反馈。

消融实验结果（见表2）显示，移除证据链导致模型所有性能指标均有所下降，F1值下降较为明显，表明缺乏逻辑化证据组织的支持，模型更易产生误判，证据链是保障判定准确性与稳健性的基础；移除多智能体小组协同，模型各个性能指标均有所降低，说明多智能体小组从不同视角进行的并行推理与交叉验证，能有效发现被单一视角忽略的疑点，从而提升系统识别虚假信息的能力，降低漏报率；移除第二轮共识裁决，精确率与召回率均有所降低，导致F1降低，表明该机制通过对分歧案例的聚焦与理由交互，起到了重要的校准作用，有助于过滤误报、提升判定精准性。各消融版本的解释报告质量评分与完整框架相近，说明生成高质量解释文本更多依赖于底层大语言模型的基础能力，而框架本身主要优化的是判定过程的可靠性与结论的准确性。另外，采用5折交叉验证对上述结果进行稳健性检验，各消融实验的F1值与表2基本一致，标准差均小于0.03，且t检验显示完整框架与其他消融版本的差异均具有统计学意义（$p<0.01$）。

4　案例分析

本研究选取三个典型样本分别呈现ECEMA-GD-FC框架在不同核查情境下的表现：成功案例为在各小组证据一致的情况下，展示框架在常规情境下的标准工作流程；失败案例为判定错误的情况，揭示框架的失效边界与改进方向；边界案例为各小组证据冲突的情况下，检验框架在决策临界点上的共识形成机制。

4.1　成功案例

为更直观地展示ECEMA-GD-FC框架的运行机制，本研究以中国互联网联合辟谣平台的真实案例为例，对实验流程进行说明。

在这一环节中，为主持人智能体输入以下待核查文本：“眼下，聚酯纤维已经成为重要的服饰面料。近日，‘聚酯纤维是用塑料瓶做的’等说法在网上流传。有网友调侃：‘穿上摇粒绒，用一身塑料过冬’‘打开衣柜，一柜子都是塑料瓶做的衣服’。还有网友分析：‘用塑料做衣服，怪不得冬天会出现静电’。更有自称国外某毒理学专家分析：‘穿着聚酯纤维面料，会导致微塑料透过皮肤进入血液之中’”。

主持人智能体通过自然语言理解模块从输入文本中提炼出文本摘要，并进行领域分配，以JSON格式输出结果（见图7），随后将任务分发给三个专业化核查智能体小组。

以科学证据小组为例，其内部的证据搜集智能体将文本摘要内容作为证据检索的输入，通过调用智谱AI的网络搜索工具search_std进行网络搜索，获取与该命题相关的事实性信息与权威来源，信息筛选约束设置为：优先采纳同行评审研究，重点考察研究的实验设计、样本规模、对照组设置和统计显著性，以研究质量和证据强度作为信息取舍的重要依据。证据搜集智能体输出部分证据检索结果如图8所示。

证据链搭建智能体对每一条证据语句进行语

表2　消融实验结果

组别	准确率	精确率	召回率	F1值	解释报告综合评分
实验组1	0.770	0.755	0.771	0.763	36.560
实验组2	0.812	0.812	0.824	0.818	37.367
实验组3	0.820	0.793	0.875	0.832	**37.460**
本方法	**0.842**	**0.822**	**0.876**	**0.848**	37.269

注：表中加粗数值为对应指标下的最优性能值。

"summary": "2025年12月，有自称国外某毒理学专家声称穿着聚酯纤维面料会导致微塑料透过皮肤进入血液。",
"domain": "科技"

图7　主持人智能体文本摘要与角色分配输出结果

义分析与关联性评估，判断其与原始待核查主张之间的逻辑关系类型，划分为支持（support）、反驳（contradict）、中立（neutral）三类。证据链搭建智能体输出部分结果如图9所示。

真伪判定智能体对证据链进行一致性与逻辑完整性分析，形成对原始主张的总体判断。该环节的输出不仅包括最终的真伪判定结果（likely_true、likely_false、uncertain），还要求生成简要的判定理由说明，以体现系统推理过程的可解释性与透明性。真伪判定智能体输出结果如图10所示。

title="聚酯纤维是用塑料瓶做的，穿了会导致微塑料进入血液？耸人听闻（发布时间：2025-12-05 15:37:12)", link="https:/www.163.com/dy/article/KG1BFF7A055040N3.html?spss=dy_author", content="眼下，聚酯纤维已经成为重要的服饰面料。近日，聚酯纤维是用塑料瓶做的等说法在网上流传。有网友调侃：穿上摇粒绒，用一身塑料过冬打开衣柜，一柜子都是塑料瓶做的衣服。还有网友分析：用塑料做衣服，怪不得冬天会出现静电。更有自称国外某毒理学专家分析：穿着聚酯纤维面料，会导致微塑料透过皮肤进入血液之中。这些说法都是真的吗？聚酯纤维是用塑料瓶做的，容易产生静电的说法在网上广泛流传。穿聚酯纤维会导致微塑料透过皮肤进入血液之中的说法也一度引发恐慌。聚酯纤维是用塑料瓶做的吗？一部分是的。塑料瓶和聚酯纤维（涤纶）服装，本质上来源于同一种石油基原料——PET（聚对苯二甲酸乙二醇酯）。目前，市场上的聚酯纤维面料来源主要有两方面，一是直接由石油化工原料生产，称为原生涤纶。二是用废旧塑料瓶等（主要是PET材质）生产，称为再生涤纶、环保涤纶或rPET。消费者在购买纺织品时，如果看到再生涤纶、rPET这样的标签，就代表面料是由塑料瓶变身而来的。将塑料瓶回收利用，生产出高品质的再生涤纶，其强度、耐用性、手感等方面与原生涤纶几乎无异。这项技术已有几十年历史，工艺稳定，技术成熟，还具有重要的环保价值……"

图8　证据搜集智能体输出部分证据检索结果

"evidence_chain_text": "聚酯纤维面料的来源主要包括原生涤纶（由石油化工原料直接生产）和再生涤纶（rPET，由废旧塑料瓶等PET材质回收再加工而成）。目前，全球多项研究表明，聚酯纤维在正常穿着条件下不会显著释放微塑料。健康、完整的皮肤作为人体第一道屏障，具有强大的防御功能，能有效阻挡微塑料透皮吸收。即便存在微塑料接触，其进入血液的可能性极低。此外，聚酯纤维耐老化性良好，在使用寿命内基本不会因断裂或降解而释放微塑料。因此，所谓‘穿着聚酯纤维面料会导致微塑料透过皮肤进入血液’的说法缺乏科学依据，属于耸人听闻的误传。"
"structured": [
{"index": 1, "title": "聚酯纤维是用塑料瓶做的，穿了会导致微塑料进入血液？（发布时间：2025-12-13 18:31:01）", "claim excerpt": "目前，市场上的聚酯纤维面料来源主要有两方面，一是直接由石油化工原料生产，称为原生涤纶。二是用废旧塑料瓶等（主要是PET材质）生产，称为再生涤纶、环保涤纶或rPET。消费者在购买纺织品时，如果看到再生涤纶、rPET这样的标签，就代表面料是由塑料瓶变身而来的。", "support": "contradict", "confidence": 0.95},
{"index": 2, "title": "聚酯纤维是塑料衣服？穿久了微塑料会进入血液？真相是……（发布时间：2025-12-23 12:07:24)", "claim_excerpt": "主流的生产方法，是直接用聚酯切片，将切片加热到熔融状态，再通过精细的纺丝工艺，拉成比头发丝还细的纤维，最终编织成聚酯纤维面料。", "support": "neutral", "confidence": 0.85},
{"index": 3, "title": "聚酯纤维是用塑料瓶做的，穿了会导致微塑料进入血液？耸人听闻（发布时间：2025-12-05 15:24:21)", "claim_excerpt": "目前，科学界的主流观点和证据都表明，在正常穿着聚酯纤维的情况下发生微塑料进入血液的可能性极低。", "support": "contradict", "confidence": 0.96}……]

图9　证据链搭建智能体结果输出部分结果

"verdict": "likely false", "confidence": 0.95

图10　真伪判定智能体输出结果

解释报告生成智能体综合分析各智能体输出的信息生成解释报告。报告内容包括三个部分：生成的完整解释报告文本、关键要点列表、基于分析结果给出的建议或后续行动指导。解释报告生成智能体输出结果如图11所示。

鉴于三个小组在功能结构与执行流程上的一致性，其他两个小组的具体分析过程不逐一展示。最终由主持人智能体将三个专业化核查智能体小组生成的解释报告进行汇总。由于三个智能体小组的判定结果均为“likely_false”，因此未触发用于处理分歧结论的第二轮投票机制。主持人智能体最终输出结果如图12所示。

4.2　失败案例

失败案例待核查主张为“2018年12月，一千多吨病死猪肉流入11省区，千吨病死猪肉最近流入市场”，经过核查判定，失败案例中主持人智能体输出结果如图13所示。

主持人智能体给出结果为“likely_true”（可能为真），但实际为“假”。进一步分析发现，该错误并非源于单条证据的明显偏差，而是由于模型在多源信息整合过程中产生了跨事件语义错配与推理层级偏移。错误源于大模型在多源数据整合中发生认知偏差：时间锚定失效（模型未能区分“2018年12月”与历史中其他类似事件的时间边界）、因果边界模糊（将“历史上存在病死猪肉事件”这一普遍事实直接等价为“当前主张所指的具体事件已被证实”），最终导致事实判定偏离真实标签。针对上述问题，未来可从以下方面进行改

"report_text": "根据现有科学证据，关于‘聚酯纤维面料可能使微塑料透过皮肤进入血液’的主张被判定为极不可能成立。多项全球研究一致表明，在正常穿着条件下，聚酯纤维（包括原生涤纶和再生涤纶，即rPET）不会显著释放微塑料。聚酯纤维具有良好的耐老化性和结构稳定性，其在使用寿命内基本不发生断裂或降解，因此几乎不会产生可释放的微塑料颗粒。此外，人体健康、完整的皮肤作为天然生理屏障，具备强大的防御能力，能够有效阻挡微小颗粒如微塑料的透皮吸收。目前尚无可靠临床证据或重复性研究支持微塑料可通过完整皮肤进入血液循环的路径。尽管微塑料在环境中的存在已受到广泛关注，但其与人体穿用聚酯纤维衣物之间的因果关系尚未得到科学证实。综合来看，该说法缺乏实证支持，与当前医学、材料科学及毒理学共识相悖，属于未经证实且带有误导性的误传。建议公众理性看待相关议题，避免因信息过载而产生不必要的健康焦虑。", "key_points": ["聚酯纤维在正常穿着条件下不会显著释放微塑料。", "健康完整的皮肤具有强大屏障功能，能有效阻止微塑料透皮吸收。", "聚酯纤维耐老化性强，使用寿命内基本不降解或断裂。", "多项独立研究未发现微塑料通过衣物接触进入血液的可靠证据。", "该主张缺乏科学依据，属于耸人的误传。", "recommendation": "建议主持人在汇总时强调：当前科学证据不支持‘聚酯纤维导致微塑料进入血液’的说法，应基于可重复、实证的研究结论进行传播，避免传播未经证实的健康担忧。同时可提醒公众关注真实风险（如环境污染），而非过度担忧日常衣物材质带来的假想健康威胁。"]

图11　解释报告生成智能体结果输出结果

"final report text": "综合多方权威证据，关于‘有研究指出聚酯纤维面料可能导致微塑料透过皮肤进入血液’的说法被判定为极不可能成立。现有科学共识明确指出，聚酯纤维（包括原生涤纶与再生涤纶，rPET）在正常穿着条件下具有优异的耐老化性，几乎不会发生降解或断裂释放微塑料；同时，健康完整的皮肤作为天然屏障，能有效阻止微小颗粒如微塑料的透皮吸收。多个主流媒体平台及权威辟谣机构已对这一说法进行核查并定性为‘耸人听闻’且缺乏实证支持。目前无可靠研究、临床案例或重复性数据证实微塑料可通过衣物穿用进入人体血液循环。该主张不符合当前材料科学、毒理学与医学领域的主流认知，属于未经证实的误传。公众应理性看待纺织品安全问题，避免因信息过载引发不必要的健康焦虑。",
"final_verdict": "likely_false",
"final confidence": 0.94

图12　主持人智能体最终输出结果

"final report text": "综合三组分析，尽管原始声明中‘2018年12月’的时间点存在明显偏差，但其所描述的核心事实——即千余吨病死猪肉跨11个省区流通——与公安部于2014年挂牌督办、2015年告破的重大食品安全案件高度一致。该案件覆盖12个省区，查获问题肉品超1 000吨，由央视、公安部通报及司法文书等权威渠道共同确认，具备完整证据链和一致性来源。虽2018年12月期间安徽、广东等地确有地方性排查，但未发现大规模跨省流通证据，仅属局部事件。三组报告均指向同一真实事件被错误归因于2018年，属于典型的时间记忆错位。结合证据权威性、现实可复现性及多源印证，该主张本质属实，仅为时间表述不准确。建议在传播时明确事件真实发生时间为2015年，并强调历史案件的警示意义，避免误导公众。", "final_verdict": "likely_true", "final confidence": 0.93

图13　失败案例中主持人智能体输出结果

进：①在证据链构建环节引入时间戳强制比对，要求模型显示标注证据与主张的时间关系；②设计因果强度分级提示词，引导模型区分“直接因果”“背景信息”等不同证据层级。

4.3　边界案例

该案例待核查主张为“2021年，金星上没有发现生命。金星生命‘时有时无’？有些‘乌龙’让科学更接近真理”，在该任务的第一轮判定中，科学证据小组和事实证据小组判定为“可能为真”，媒体证据智能体判定为“可能为假”，初轮判定存在分歧，由主持人智能体整合后重新发给各智能体进入第二轮重判。在第二轮判定中，三个小组结论都为“可能为真”，案例判定过程如表3所示。

该案例结果表明，当多源证据之间存在冲突时，智能体小组之间的分歧能够有效触发不确定性判断，从而进入第二轮评估促进结论收敛，并在一定程度上降低误判风险。

综合对比上述成功案例、失败案例、边界案例结果可以看出，所提ECEMA-GD-FC框架在证据一致性较高时能够得出正确结论并体现出良好的可靠性；在证据冲突时，多小组之间的分歧机制能有效触发二次判定机制，在一定程度上避免误判，但有时受限于跨事件的语义对齐能力而产生“幻觉”，出现判定结果偏差。

5　总结与展望

本研究针对事实核查任务中单一LLM存在的“幻觉”问题、可解释性不足及鲁棒性较弱等问题，提出并验证了一种证据链增强的多智能体分组讨论事实核查框架——ECEMA-GD-FC框架。该框架通过构建一个由主持人智能体统筹、多组智能体小组按证据检索、证据链构建、真伪判定、报告生成的流程并行协作的闭环核查机制，实现了从单模型判别到多智能体协同验证的范式演进。实验结果表明，ECEMA-GD-FC框架不仅能显著提升事实核查的准确性，并通过结构化的证据链与解释报告增强了核查过程的解释性与可信度。本研究的核心创新在于：①证据链驱动的流程刚性，每个小组均遵循证据检索、证据链构建、真伪判定、解释生成的标准流程，以外显的逻辑链约束推理轨迹，提升过程的可追溯性；②多维证据的协同校验，设立三个小组，分别从学术共识、信源交叉与原始材料等维度并行核查，通过视角互补抑制单一证据偏差；③主持人导引的共识形成，引入主持人智能体统筹任务分发、汇总中间结果，并基于多轮反馈与裁决机制化解分歧，形成最终解释报告。

从实践应用角度看，本研究提出的ECEMA-GD-FC框架对提升网络信息治理能力具有一定的管理启示。首先，监管部门和网络平台可在现有内容审核体系中引入ECEMA-GD-FC框架，对传播范围较大或被用户举报频率较高的信息进行自动化核查，并生成结构化证据报告供人工审核人员参考，以提升谣言识别效率并降低人工审核成本。其次，建议相关管理部门推动建设权威事实核查数据资源库，例如整合政府公告、主流媒体报道及专业机构辟谣信息等可信数据源，为自动化核查系统提供稳定的证据检索来源，并逐步形成标准化的事实核查数据接口，以支持不同平台和机构之间的信息共享与协同治理。

本研究仍存在若干局限需要在未来工作中进一步完善。首先，当前框架主要面向文本信息，未来可扩展至图像、视频等多模态信息的核查，研究跨模态证据的检索、对齐与融合机制。其次，引入跨事件语义对齐机制，增强该框架在证据冲突场景下对不同事件间实体关系、因果链条的判别能力，降低因语义错位引发的幻

表3　案例判定过程

判定过程	科学证据小组	媒体证据小组	事实证据小组
第一轮	"verdict": "likely_true", "confidence": 0.92, "reasoning": 尽管2020年9月麻省理工学院萨拉西格教授团队在金星大气中检测到磷化氢，该气体在地球上主要由生物活动产生，引发关于金星可能存在大气生命的推测，但后续研究指出该信号存在争议，且无法排除非生物来源的可能性……	"verdict": "likely_false": "confidence": 0.92, "reasoning": 2024年7月利用詹姆斯·克拉克·麦克斯韦望远镜的重复观测不仅确认了磷化氢的存在，且检测值为最初的140倍，并发现氨，进一步强化了非生物过程无法解释的化学异常，因此原声明“未发现生命”在科学语境下已不成立……	"verdict": "likely_true", "confidence": 0.85, "reasoning": 研究揭示了金星云层中存在支持生命化学基础的环境可能性。所有提到的研究均属于对生命化学条件的模拟与推测，尚未证实生命存在……
第二轮	"verdict": "likely_true", "confidence": 0.92, "reasoning": 研究虽指出金星大气对阳光的异常吸收可能与生物过程有关，但同样缺乏直接证据支持。此外，金星表面极端环境（高温、强酸雨）被广泛认为不适合地球型生命生存……	"verdict": "likely_true", "confidence": 0.92, "reasoning": 2021年伦敦帝国理工学院团队首次在金星云层中探测到磷化氢，磷化氢在地球上主要由生物过程产生，而金星缺乏自然生成该化合物的条件，因此其存在强烈暗示可能存在非地球型生命活动，但未被证实……	"verdict": "likely_true", "confidence": 0.85, "reasoning": 尽管该时间段内没有直接观测到生命迹象的报告，但后续研究揭示了金星云层中存在支持生命化学基础的环境可能性，虽然存在潜在支持生命的化学条件，但尚未达到发现生命的程度……
各小组解释报告生成	"report_text": "2021年，科学界在金星上未发现确凿的生命迹象，但一项具有争议性的发现推动了天体生物学与行星科学研究的深入发展……", "key_points": "2021年，《自然·天文学》报告称在金星大气中检测到磷化氢（PH_3），被视为潜在生命迹象。": "后续多项同行评审研究通过对原始数据重新分析，未发现显著的磷化氢信号……", "recommendation": "主持人A可基于本报告内容，总结科学界在金星生命探测上的进展：虽未发现生命，但争议性发现激发了更严谨的研究实践，反映了科学探索的本质……"	"report_text": "麻省理工学院（MIT）与卡迪夫大学的研究团队在2022年初发表论文指出，该信号更可能源于非生物过程，尤其是二氧化硫（SO_2）在金星云层中的复杂化学反应，而非生命活动产物……", "key_points": "2022年更新数据支持非生物成因，主流科学界普遍质疑生命假说。", "其他潜在生命指标（如硫醇、甲烷、氨）尚未被独立证实，缺乏确凿证据……", "recommendation": "主持人智能体可基于此报告进行汇总陈述，强调科学界对早期发现的审慎态度以及当前的重要性，并引用权威机构结论以增强公信力……"	"report_text": "2021年，科学界在金星大气中检测到磷化氢（phosphine）的光谱信号，这一发现曾引发关于金星可能存在生命迹象的广泛讨论。磷化氢在地球上主要由生物过程产……", "key_points": ["2021年金星大气中检测到磷化氢信号，引发生命存在假说。", "后续研究指出该信号可能由非生物过程（如火山活动或大气化学反应）解释……", "recommendation": "主持人智能体可以基于此报告总结为：尽管2021年金星大气中发现了疑似生命标志的磷化氢信号，但经过多轮独立验证与深入分析，科学界已排除其作为生命直接证据的可能性……"
最终解释报告	综合三组报告的科学评估，金星在2020年11月23日至2021年12月31日未发现确凿的生命证据。尽管麻省理工学院团队于2020年9月首次探测到磷化氢，该气体在地球上多与生物活动相关，引发关于金星大气可能存在生命的可能性讨论，但后续研究未能重复该信号，且主流科学界普遍认为其可能源于非生物过程或观测误差；同时，金星极端表面环境（高温、强酸雨）被广泛视为不适合地球型生命生存。尽管2024年有新观测显示磷化氢浓度显著上升并伴随氨的发现，但这些数据发生于2021年之后，不构成对原始时间段内结论的有效反证。此外，虽然实验室模拟表明氨基酸可在金星云层硫酸环境中稳定存在，但这仅说明化学可行性，并非生命存在的实证。因此，在所涉时间范围内，所有证据均缺乏可重复性、可验证性及直接生命信号支持。综合研究质量、权威来源共识与实践可重复性标准，当前无充分依据推翻未发现生命的结论		

觉风险。再次，探索证据链的动态校验与回溯机制，当最终判定与多数小组结论不一致时，自动回溯证据链构建环节，识别导致偏差的关键证据或推理步骤。最后，可针对不同领域（如医学、政治、科技）的特点设计领域自适应的判定逻辑，以提升框架在跨领域场景中的泛化能力。

参考文献

[1] 2025微博社区信任报告[EB/OL]. [2026−03−11]. https://weibo.com/ttarticle/p/show? id=2309405260880345956798.

[2] 网络信息内容生态治理规定[EB/OL]. [2026−03−11]. https://www.moj.gov.cn/pub/sfbgw/flfggz/flfggzbmgz/202101/t20210105_146435.html.

[3] SUAREZ-LLEDO V, ALVAREZ-GALVEZ J. Prevalence of health misinformation on social media: Systematic review[J]. Journal of Medi-cal Internet Research, 2021, 23 (1): e17187.

[4] BRIAN E W. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on suscepti-bility to political misinformation[J]. The Journal of Communication, 2015, 65 (4): 699–719.

[5] WARDLE C. Misinformation has created a new world disorder-scientific american[EB/OL]. [2025−04−27]. https://www.scientifi-camerican.com/article/misinformation-has-created-a-new-world-disorder/.

[6] Code of standards[EB/OL]. [2025−04−27]. https://efcsn.com/code-of-standards/.

[7] VLACHOS A, RIEDEL S. Fact checking: Task definition and dataset construction[C]//Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science.

Baltimore，MD：Association for Computational Linguistics，2014：18–22.

[8] MIHALCEA R，STRAPPARAVA C. The lie detector: Explorations in the automatic recognition of deceptive language[C]//Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec，Singapore：Association for Computational Linguistics，2009：309–312.

[9] FENG S，BANERJEE R，Choi Y. Syntactic stylometry for deception detection[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics（Volume 2：Short Papers）. Jeju Island，Korea：Association for Computational Linguistics，2012：171–175.

[10] WANG W Y. “liar, liar pants on fire”: A new benchmark dataset for fake news detection[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver：Association for Computational Linguistics，2017：422–426.

[11] 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 Language Technologies. New Orleans：Association for Computational Linguistics，2018：809-819.

[12] WILLIAMS E，RODRIGUES P，NOVAK V. Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models[J]. arXiv preprint，2020. DOI:10.48550/arXiv.2009.02431.

[13] 兰恭瑞，杨建林. 基于证据链和大语言模型的描述性情报核查方法研究：以健康信息学领域为例[J]. 情报理论与实践，2026，49（1）：181–188，209.

[14] 朱梦蝶，管振凯，王艺，等. 大模型驱动的事实核查多智能体系统构建研究[J]. 图书与情报，2025（4）：61–72.

[15] HU X，GUO Z，CHEN J，et al. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：Association for Computing Machinery，2023：2901–2912.

[16] 郭海玲，卫金金，刘仲山. 生成式人工智能虚假信息协同共治研究[J]. 情报杂志，2024，43（9）：121–129，165.

[17] 林荣鑫，李硕豪，董力铭，等. 基于视觉语言大模型的多模态虚假新闻检测[J/OL]. 计算机工程，1–13[2025–11–12]. https://doi.org/10.19678/j.issn.1000-3428.0252354.

[18] DORRI A，KANHERE S，JURDAK R. Multi-agent systems: A survey[J]. Ieee Access，2018，6：28573–28593.

[19] WOOLDRIDGE M，JENNINGS N R. Intelligent agents: Theory and practice[J]. The Knowledge Engineering review，1995，10（2）：115–152.

[20] 郭翔，胡伟彪. 基于多Agent模型的应急指挥对地震应急救援的影响仿真研究[J]. 清华大学学报（自然科学版），2025，65（6）：1027–1039.

[21] 刘继，陈艳. 基于多主体协同治理的网络舆情观点演化分析[J]. 情报杂志，2026，45（3）：131–138.

[22] 张萌，牟冬梅，邓君，等. 医疗数据共享利益相关者协同机制研究：基于四方演化博弈分析[J]. 情报学报，2025，44（1）：61–74.

[23] 支凤稳，彭兆祺，赵梦凡，等. 中小企业间竞争情报共享策略演化博弈与仿真研究[J]. 情报学报，2024，43（6）：733–746.

[24] 张金灿，张雪莹，郭可莹. 动态奖惩机制下数据质量协同监管多方博弈分析 [J]. 图书情报工作，2025，69（19）：13–25.

[25] 张煊，韩一士，孔方正. 一种基于元胞自动机的微信涉警网络舆情传播模型[J]. 情报科学，2025，43（6）：42–48.

[26] 郭耿，陈海汉，陈福集，等. 基于改进元胞自动机的微博舆情演化研究[J]. 情报理论与实践，2019，42（5）：165–170.

[27] 高彦杰，陈跃国. 大语言模型预训练系统关键技术综述[J]. 软件学报，2026，37（1）：200–229.

[28] 赵学东，林民，唐邦洋，等. KG和RAG协同知识增强的数学古籍领域智能体[J]. 计算机工程与应用，2026，62（4）：124–136.

[29] KARUNANAYAKE N. Next-generation agentic AI for transforming healthcare[J]. Informatics and Health，2025，2（2）：73–83.

[30] 殷跃，张海涛，刘彦辉，等. 基于大模型智能体的粮食安全应急情报体系构建研究[J]. 情报理论与实践，2025，48（1）：20–30.

[31] 李白洋，张欣宇，赵英潇，等. 基于大模型的国防科技情报研究多智能体系统构建与应用研究[J]. 情报理论与实践，2025，48（9）：178–184.

[32] 丁志坚，张文杰，徐天阳，等. 基于DeepSeek的食品安全法律法规多智能体检索问答[J]. 中国食品学报，2025，25（6）：449–464.

[33] NI B，BUEHLER M J. MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data，and integrate knowledge[J]. Extreme Mechanics Letters，2024，67：102131.

[34] JENNINGS N R，SYCARA K，WOOLDRIDGE M. A roadmap of agent research and development[J]. Autonomous agents and multi-agent systems，1998，1（1）：7–38.

[35] HU X，GUO Z，WU G Y，et al. CHEF: A pilot Chinese dataset

for evidence-based fact-checking[J]. arXiv preprint，2022. DOI:10.48550/arXiv.2206.11863.

[36] KELLERHUIS B E，JENNISKENS K，KUSTERS M P T，et al. Expert panel as reference standard procedure in diagnostic accuracy studies: a systematic scoping review and methodological guidance[J]. Diagnostic and Prognostic Research，2025，9（1）：12.

[37] ZHU J，LI J，WEN Y，et al. Benchmarking large language models on CFLUE-a Chinese financial language understanding evaluation dataset[C]//Findings of the Association for Computational Linguistics：ACL 2024. Bangkok：ACL，2024：5673−5693.

作者简介

相薏薏，女，博士，副教授，研究方向：信息隐私、用户信息行为，E-mail：xiangmengm123@sina.com。
纪泽旭，男，硕士研究生，研究方向：信息隐私、网络舆情。
郭佳鸣，女，硕士研究生，研究方向：信息隐私。
王秀琨，女，本科生，研究方向：网络舆情。

Research on an Evidence Chain Enhanced Multi-Agent Group Discussion Fact-Checking Framework

XIANG MengMeng[1,2]　JI ZeXu[1]　GUO JiaMing[1]　WANG XiuKun[1]
(1. School of Artificial Intelligence and Computer Science, Jilin University of Finance and Economics, Changchun 130117, P. R. China;
2. Institute of Big Data and Interdisciplinary Sciences, Jilin University of Finance and Economics, Changchun 130117, P. R. China)

Abstract: Fact-checking serves as a critical approach for identifying and mitigating misinformation. However, individual large language models (LLMs) suffer from inherent limitations in fact-checking tasks, including hallucination issues, insufficient interpretability, and limited accuracy. This study proposes an Evidence Chain Enhanced Multi-Agent Group Discussion Fact-Checking Framework to improve the accuracy, robustness, and explainability of fact-checking systems. The proposed framework consists of a moderator agent responsible for task orchestration, result aggregation, and consensus adjudication, along with three specialized fact-checking agent groups focusing on scientific evidence, media evidence, and factual evidence, respectively. Within each group, agents follow a standardized fact-checking workflow involving evidence retrieval, evidence chain construction, truthfulness assessment, and report generation, thereby establishing a closed-loop mechanism of “Coordination−Specialization−Integration” Experimental results on benchmark fact-checking datasets demonstrate that the proposed framework achieves competitive performance in both veracity classification accuracy and fact-checking report generation quality, substantially enhancing the credibility of fact-checking results.

Keywords: Misinformation; Fact-checking; Multi-agent Systems; Evidence Chain; Large Language Models