山西大学学报(自然科学版)49(2):189—198,2026
Journal of Shanxi University(Nat. Sci. Ed. )

**DOI**:10.13451/j.sxu.ns.2025101

# 基于多角度分析的大语言模型虚假信息检测

贾彩燕[1*],杨子琦[1],赵一[1],白祥意[2]
(1.北京交通大学 计算机科学与技术学院, 北京 100080;
2.北京交通大学 詹天佑学院, 北京 100080)

**摘　要**:虚假信息在社交媒体上的传播频繁,而现有基于深度学习和图神经网络的方法在适应复杂环境和提供可解释性方面仍存在局限。此外,大语言模型(Large Language Models, LLMs)虽具有强大语言理解能力,但在处理复杂文本线索和传播模式时,其潜力尚未被充分发挥。本文提出了一种基于大语言模型的多角度特征提取框架,旨在充分发挥LLMs的深层语义挖掘能力和可解释性优势。具体而言,本框架通过设计一套多维度提示指令,引导LLMs从写作风格、事实一致性、网友情绪等十个角度对社交媒体内容进行分析,并输出详细的解释性答复。接着,利用大语言模型的特征提取能力获取数据的会话链语义表示和解释性文本语义表示,并利用多层感知机进行分类训练,从而实现虚假信息的低成本高效检测。实验结果表明,本文方法的*F*1分数、准确率等关键指标显著超越了图神经网络与LLMs两类基线方法,在Weibo、Twitter15和Twitter16三个公开数据集上其*F*1分数与最优基线方法相比分别提高了1.4%、3.8%和3.5%。此外,该方法在保持高性能的同时,大幅降低了训练成本,提升了模型决策的透明度和可解释性。

**关键词**:社交媒体分析;谣言检测;人工智能;提示工程;可解释性
**中图分类号**:TP18;TP391　　**文献标志码**:A　　**文章编号**:0253-2395(2026)02-0189-10

# Large Language Models for Misinformation Detection Based on Multi-angle Analysis

JIA Caiyan[1*], YANG Ziqi[1], ZHAO Yi[1], BAI Xiangyi[2]
(1. *School of Computer Science and Technology, Beijing Jiaotong University, Beijing* 100080*, China;*
2. *Jeme Tienyow Honors College, Beijing Jiaotong University, Beijing* 100080*, China*)

**Abstract:** The spread of misinformation on social media is continual, however, existing methods based on deep learning and graph neural networks still face limitations in adapting to complex environments and providing interpretability. In addition, although large language models (LLMs) possess powerful language understanding capabilities, their potential in handling complex textual cues and propagation patterns has not been fully explored. This paper proposes a multi-angle feature extraction framework based on large language models, aiming to fully leverage the deep semantic mining and interpretability advantages of LLMs. Specifically, this framework guides LLMs to analyze social media content from ten perspectives, such as writing style, factual consistency, and netizen's sentiment, and outputs detailed explanatory responses by designing a set of multi-angle prompting instructions. The model then outputs detailed responses along with explanatory rationales for each question. Subsequently, the framework utilizes the feature extraction function of LLM to obtain the semantic representation of the session chain and the semantic representation of the interpreted text of the raw data, and further processes them with classification training using a multilayer perceptron, thus realizing low-cost and high-efficiency misinformation detection. Experimental results demonstrate that the proposed method in this paper significantly sur-

**收稿日期**:2025-08-03;**修回日期**:2025-09-19
**基金项目**:国家自然科学基金(62576026);中央高校基础科研业务项目(2024XKRC024)
* **通信作者**:贾彩燕(1976—),女,宁夏石嘴山人,教授,博士生导师,研究方向为计算机科学与技术。E-mail:cyjia@bjtu.edu.cn
**引文格式**:贾彩燕,杨子琦,赵一,等.基于多角度分析的大语言模型虚假信息检测[J].山西大学学报(自然科学版),2026,49(2):189-198. DOI:10.13451/j.sxu.ns.2025101.

passes two categories of baseline methods, Graph Neural Networks (GNNs) and Large Language Models (LLMs), in key metrics such as *F*1-score and accuracy. On the three public datasets of Weibo, Twitter15, and Twitter16, the proposed method in this paper improves the *F*1-score by 1.4%, 3.8% and 3.5%, respectively, compared to the strongest baseline. Furthermore, the method not only maintains high performance but also substantially reduces training costs while enhancing the transparency and interpretability of model decision-making.

**Key words:** social media analysis; rumor detection; artificial intelligence; prompt engineering; interpretability

## 0 引言

在社交媒体高度普及的时代,信息的传播方式发生了深刻变革。然而,与此同时,虚假信息(misinformation)或假新闻(fake news)也在全球范围内迅速扩散,给社会舆论、公共决策、政治稳定和经济发展带来了严峻挑战。虚假信息的快速传播不仅可能误导公众认知[1],影响社会共识,还可能加剧社会分裂,引发现实世界的危机事件。例如,新冠疫情大流行期间,大量未经证实的医疗建议和阴谋论在社交媒体上传播[2],严重影响了公共健康政策的执行。

针对虚假信息的自动检测已经成为计算机科学与社会科学交叉领域的一个核心问题。早期的研究主要依赖于基于规则的方法或传统的机器学习分类器,这些方法通常利用文本的语言特征、情感特征或传播模式进行检测。然而,这些方法在应对社交媒体动态变化和语境多样性方面存在明显局限性。

近年来,深度学习,尤其是大语言模型(Large Language Models, LLMs)的崛起,为虚假信息检测提供了新的思路。其中,深度学习方法,如循环神经网络[3]、卷积神经网络[4],图神经网络[5]、Transformer方法[6],不仅存在深度学习固有的可解释性差的问题,且难以应对真实世界不断变化的环境。而LLMs拥有强大的语义理解能力[7],能够基于大量文本数据学习语言模式,并在一定程度上进行逻辑推理,具有很强的泛化性和可解释性。然而, LLMs可能会受到训练数据的偏见(bias)影响,在处理虚假信息任务时容易生成错误的判断,甚至产生“幻觉”(hallucination)[8],即生成看似合理但实际错误的内容。并且,简单使用大语言模型对虚假信息进行直接判定[9],不能深入挖掘虚假信息本身在传播过程中的关键判定线索。

为了解决上述问题,本研究提出了一种基于可解释性大语言模型的虚假信息检测框架。该方法不仅利用大语言模型的强大文本理解能力,还结合可解释性技术,从多个角度分析文本内容,提供更透明、更可信的判断结果。具体而言,本研究的主要贡献包括:

1. 多角度解释生成:通过设计一套多维度提示指令,引导大语言模型从语言煽动性、事实一致性、逻辑矛盾、社交互动等多个维度对社交媒体文本进行深入分析,增强虚假信息检测的可解释性。

2. 特征融合策略:提出一种结合原始会话链语义表示和合理性解释语义表示的特征融合方法,并利用多层感知机进行分类,提升检测的鲁棒性和精度。

3. 高效可解释检测:相较于现有的使用LLMs直接分类方法,本文的方法不仅在准确率和*F*1分数等指标上取得领先,还提供了透明的决策依据,使检测结果更具可信度。

4. 消融实验验证:通过消融实验分析会话链语义表示和合理性解释文语义表示在检测任务中的贡献,进一步验证了多角度解释方法的有效性。

## 1 相关工作

社交媒体平台上的虚假信息泛滥对社会造成了诸多负面影响,促使学术界和工业界对虚假信息检测技术给予了广泛关注。为应对这一挑战,研究者提出了多种方法[10]。

近年来,基于深度学习的谣言检测方法主要集中于新闻文本的内容特征提取上。例如,基于循环神经网络(Recurrent Neural Network, RNN)的技术[3],通过建模微博帖子的时间序列信息以实现对谣言的检测,或结合卷积神经网络(Convolutional Neural Network, CNN)技术进行研究[4],从而达到早期检测的目的。上述

方法虽然取得了一定成效,但其难以全面捕捉谣言传播的复杂性和多维度机制。

随着研究进一步深入,一些研究者开始关注社交媒体内容的传播模式。例如,利用传播的树状结构对谣言扩散过程进行建模[5],对数据建立异构信息网络并使用图对抗学习的方式建立模型[11]。这些方法能够有效捕捉谣言传播的时空特性,但仍然存在其局限性。例如,大多数方法呈现出“黑箱”[12]特性,其内部决策过程缺乏透明度,导致可解释性不足。其次,这些方法往往伴随着较高的训练成本,包括计算资源和时间的大量消耗,且其基于封闭世界假设进行训练,难以应对真实世界复杂的情境变化,从而限制了其在实际应用中的广泛推广。

最近,LLMs的出现为本领域开辟了新的研究方向。研究者充分利用LLMs卓越的语言理解与生成能力,通过链式思维提示(Chain-of-Thought Prompting)技术,引导LLMs生成更具逻辑性的推理过程,或通过设计提示策略,使其能够直接对新闻的真实性进行预测[13]。此外,LLMs还被用于生成新型虚假新闻样本[14],以辅助检测系统的训练与优化,或被用于立场检测[15]。这些方法在提升虚假信息检测效果方面展现了一定潜力,为该领域的技术发展注入了新的活力[16-26]。但现有方法中,由于LLMs的训练数据可能存在偏见,易产生“幻觉”现象[8],即生成看似合理但实际上错误的内容,从而导致误判。并且,现有方法缺乏对传播结构的有效利用。

总体而言,现有方法在不同方面各有优势,但也存在局限性。本文提出的框架旨在充分挖掘大语言模型在虚假信息检测领域的潜能,发挥其强大的理解能力和可解释性优势,使用较少的计算资源消耗提高检测的准确性和泛化性。

## 2 方法

本节中将介绍本研究采用的方法,所提出的方法整合了来自社交媒体帖子和用户评论的上下文信息,并利用大模型知识推理技术提高了模型决策的透明度、可解释性和准确性。

### 2.1 方法框架

本研究方法分为三个主要模块:传播结构会话链提取及指令设计、解释生成和语义提取以及分类训练。

系统框架图如图1所示。

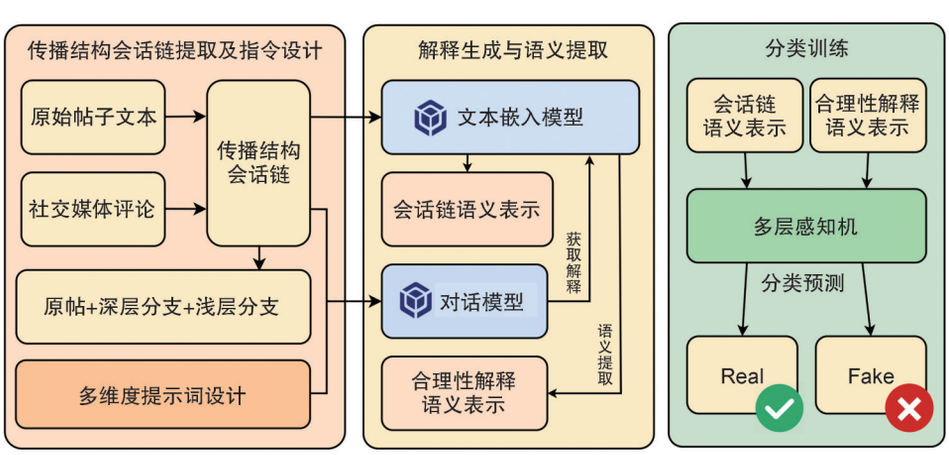


**图1** 基于多角度分析的大语言模型虚假信息检测系统框架

**Fig. 1** Framework of misinformation detection using multi-angle analysis with LLMs

### 2.2 详细步骤

#### 2.2.1 传播结构会话链提取及指令设计

(1)传播结构会话链提取

虚假信息在传播过程中的转发及回复关系可以建构为以源帖为根节点的树型传播结构。在该传播结构中存在着对虚假信息有重要判定作用的关键线索,如大众对当时帖子的观点,大众关注的与事实相符或不相符的方面等。因此,如何在传播结构中提取出对话题判定有重要作用的关键节点、子路径、子结构,送给大语言模型,是利用LLMs实现谣言检测的关键。

本方法采用先深后浅的策略提取传播节点

以构成完整的会话链。深层的传播结构通常蕴含丰富的语义信息,这些信息对于理解虚假信息的本质和传播路径至关重要。因此,在分析过程中,我们优先考虑将深层传播结构中的关键节点紧随原始帖子之后进行处理,以保持其语义上的紧密联系。相对而言,浅层的传播结构尽管讨论度较低,但它们覆盖的内容更为广泛,表现出较大的宽度。此类结构将在深层传播结构分析完成后进一步分析,以便全面评估虚假信息的影响范围及其在不同受众间的变异情况。

通过该方法,从原帖出发,经过深层传播结构的细致分析,再到浅层传播结构的广泛探讨,形成完整的对话链路结构,实现全面剖析虚假信息的传播动态。

(2)多维度提示指令设计

为充分挖掘LLMs在多角度特征提取方面的能力,本研究设计了一套多维度提示指令清单,并在后续步骤中将提示指令输入大语言模型,以获取其对给定会话链数据的详细回答和解释性理由。这些指令旨在借助LLMs的强大语义理解能力,提取文本中的多维度隐含信息,从而生成对虚假信息检测有益的信息。指令的设计涵盖了煽动性语言分析、事实一致性验证、逻辑矛盾识别以及用户反应评估等多个关键维度。

这些指令使LLMs避免直接作出简单的“是”或“否”的判断,而是从多角度对数据进行深入分析,将复杂的问题进行拆解并逐一评估,充分挖掘LLMs的可解释性与分析能力,提升结果的准确性。此外,这种多角度评估的方法能够在一定程度上减少LLMs因“幻觉”现象而生成错误信息的可能性,从而进一步增强结果的可靠性和可信度。

完整的多维度提示指令设计参见表1。

将原始帖子文本和社交媒体评论构成的会话链传播结构与多维度提示指令相结合,构建成完整的提示词输入大语言模型,生成相应的合理性解释文本。

完整的提示词流程设计参见图2。

### 2.2.2 解释生成和语义提取

(1)合理性解释生成

本研究通过将会话链与设计的提示词相结合,将其输入到大语言模型中,为每个数据生成了详细的解释性文本***explaination***。此过程充分利用了大语言模型的语义分析能力,结合提示词的引导作用,确保生成的文本能够从多个角度对数据进行深入分析和合理解释。

(2)语义提取

本方法使用LLMs对每条数据提取了两个语义表示向量:

① 会话链语义提取:将构造好的完整会话链结构输入大语言模型,获取原始数据的会话链语义表示,记为 $\boldsymbol{Embedding}_{\text{semantic}}$,维度为 $d_1$。

② 合理性解释语义提取:将合理性解释文本***explaination***送入大语言模型提取合理性解释语义表示,记为 $\boldsymbol{Embedding}_{\text{explaination}}$,维度为 $d_2$。

**表1** 大语言模型多维度提示指令清单

**Table 1** List of multi-dimensional prompt instructions for LLMs

| 序号 | 提示指令 |
|---|---|
| 1 | 帖子的内容是否违背了生活常识或常理? |
| 2 | 帖子是否提供了可靠的资料来源或证据? |
| 3 | 帖子内容是否存在刻意夸大或不切实际的表述? |
| 4 | 评论中是否有专业人士或权威人士的反驳或澄清? |
| 5 | 帖子是否涉及容易引起恐慌或误解的敏感话题? |
| 6 | 评论区的互动是否理性,是否有人质疑内容的真实性? |
| 7 | 帖子是否含有明显的情绪煽动性语言? |
| 8 | 帖子内容是否与已知真实事件相符? |
| 9 | 帖子是否有可能利用恐惧或愤怒等负面情绪来吸引点击或关注? |
| 10 | 帖子是否存在误导读者的可能性? |

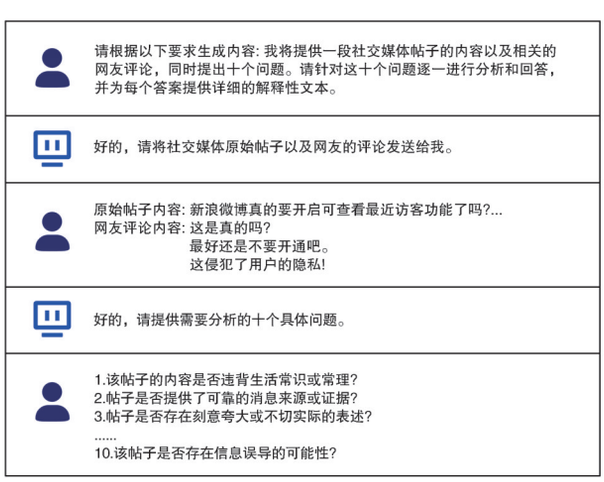


**图2** 多维度提示词设计流程

**Fig. 2** Workflow of multi-dimensional prompt design

2.2.3 分类训练

本方法采用多层感知机对前述获取的会话链语义表示和合理性解释语义表示进行特征映射和融合,最后完成分类任务,以识别虚假信息。

设会话链语义表示$\boldsymbol{E}_{\mathrm{t}}\in\boldsymbol{R}^{d_1}$,合理性解释语义表示为$\boldsymbol{E}_{\mathrm{x}}\in\boldsymbol{R}^{d_2}$,其中$d_1$和$d_2$分别表示两个语义表示的维度。采用两支特征提取模块分别对$\boldsymbol{E}_{\mathrm{t}}$和$\boldsymbol{E}_{\mathrm{x}}$进行特征映射,并使用全连接层进行特征融合,模型结构如公式(1)至公式(4)所示:

$$\boldsymbol{H}_{\mathrm{t}}=\mathrm{ReLU}(\boldsymbol{W}_{\mathrm{t}}\boldsymbol{E}_{\mathrm{t}}+\boldsymbol{b}_{\mathrm{t}}),\tag{1}$$

$$\boldsymbol{H}_{\mathrm{x}}=\mathrm{ReLU}(\boldsymbol{W}_{\mathrm{x}}\boldsymbol{E}_{\mathrm{x}}+\boldsymbol{b}_{\mathrm{x}}),\tag{2}$$

$$\boldsymbol{H}_{\mathrm{fusion}}=\mathrm{ReLU}\left(\boldsymbol{W}_f[\boldsymbol{H}_{\mathrm{t}};\boldsymbol{H}_{\mathrm{x}}]+\boldsymbol{b}_f\right),\tag{3}$$

$$\hat{\boldsymbol{y}}=\mathrm{Softmax}(\boldsymbol{W}_o\boldsymbol{H}_{\mathrm{fusion}}+\boldsymbol{b}_o),\tag{4}$$

其中$\boldsymbol{W}_{\mathrm{t}},\boldsymbol{W}_{\mathrm{x}},\boldsymbol{W}_f,\boldsymbol{W}_o$为可学习参数,$\boldsymbol{b}_{\mathrm{t}},\boldsymbol{b}_{\mathrm{x}},\boldsymbol{b}_f,\boldsymbol{b}_o$为偏置项,$[\boldsymbol{H}_{\mathrm{t}};\boldsymbol{H}_{\mathrm{x}}]$表示特征拼接操作。最后一层使用Softmax归一化得到分类概率,输出类别$\hat{\boldsymbol{y}}\in\{0,1\}$。

采用交叉熵损失函数(Cross-Entropy Loss)进行训练,定义如公式(5)所示:

$$\mathcal{L}=-\sum_{i=1}^{N}\boldsymbol{y}_i\log P_i+(1-\boldsymbol{y}_i)\log(1-P_i),\tag{5}$$

其中$\boldsymbol{y}_i$为真实标签,$P_i$为模型预测概率,$N$为样本总数。

训练采用Adam优化器,并采用5折交叉验证进行模型评估。本研究使用Optuna对五个关键超参数进行调优,最优超参数配置详见附录A。

## 3 实验

### 3.1 数据集

本方法选择了3个虚假信息检测领域的经典数据集进行实验和评估,分别是中文数据集Weibo[27],英文数据集Twitter15和Twitter16[28]。Twitter15和Twitter16的原始数据集为四分类标签,本研究对其进行了筛选,保留了False-Rumor和True-Rumor两类数据,构成二分类任务。详见表2。

**表2** 实验数据集基本统计信息

**Table 2** Basic statistics of experimental datasets

| 统计 | Weibo | Twitter15 | Twitter16 |
|---|---|---|---|
| 语言 | zh | en | en |
| 原帖数量 | 4 664 | 1 244 | 658 |
| 假谣言数量 | 2 351 | 379 | 140 |
| 真谣言数量 | 2 313 | 865 | 518 |

### 3.2 大语言模型选择

对于中文虚假信息检测任务,本研究选择ZhipuAI提供的大语言模型GLM-4-flash和嵌入模型embedding-3(输出维度为2 048)进行实验。对于英文虚假信息检测任务,本研究选择OpenAI提供的大语言模型ChatGPT-4o-mini和嵌入模型text-embedding-ada-002(输出维度为1 536)进行实验。详见表3。

**表3** 大语言模型选择

**Table 3** LLMs selection

| 语言 | 数据集 | 开发公司 | 对话模型 | 嵌入模型 |
|---|---|---|---|---|
| 中文 | Weibo | ZhipuAI | GLM-4-flash | embedding-3 |
| 英文 | Twitter15、Twitter16 | OpenAI | ChatGPT-4o-mini | text-embedding-ada-002 |

### 3.3 实验环境

所有实验均使用NVIDIA® GeForce RTX 4070 Laptop GPU进行,利用单GPU配置训练,模型使用PyTorch框架实现。

### 3.4 评估指标

本研究选择如下标准分类指标评估实验结果。

(1)准确率:正确分类实例的比例,如公式(6)所示:

$$\mathrm{Acc}=\frac{TP+TN}{TP+TN+FP+FN},\tag{6}$$

其中$TP$、$TN$、$FP$和$FN$分别代表真阳性、真阴性、假阳性和假阴性的数量。

(2)精确度:在所有正面预测中,真正的正面预测所占比例,如公式(7)所示:

$$P=\frac{TP}{TP+FP}。\tag{7}$$

精确度衡量正面预测的准确性。

(3)召回率:真阳性预测占所有实际阳性

实例的比例,如公式(8)所示:

$$R=\frac{TP}{TP+FN}。\tag{8}$$

召回率反映了模型检测所有正向实例的能力。

(4) $F1$ 分数:精确度和召回率的调和平均值,是对模型性能的均衡衡量,如公式(9)所示。

$$F1=2\times\frac{P\times R}{P+R}\tag{9}$$

### 3.5 基准方法

本文选择两类基准方法进行了比较,即大语言模型方法(LLMs)与图神经网络方法(GNNs),详见表4。

### 3.6 实验结果

表5列出了本方法和基准方法的实验结果。

**表4** 虚假信息检测基准方法汇总

**Table 4** Summary of baseline methods for misinformation detection

| 分类 | 方法 | 介绍 |
|---|---|---|
| LLMs | Zero-Shot[29] | 该方法使用大语言模型直接进行检测和判断。 |
| | Few-Shot | 该方法先给大语言模型提供一些样例及标签,再执行检测任务。 |
| | Z-CoT[9] | 该方法通过将标准指令与简单的短语"Let's think step by step",即链式思维(Chain of Thoughts, CoT)相结合利用了大语言模型自动生成的推理过程。 |
| | LeRuD[13] | 该方法通过设计提示词指导带有盲目性对新闻和评论中的重要线索进行推理,并将整个传播信息划分为一个传播链以减轻大语言模型的负担。 |
| | ARG[30] | 该方法使用大语言模型从常识和文本线索的角度生成两个真实性依据,并对两个独立的 BERT 模型进行微调,通过注意力机制对这些理由和帖子进行编码。该方法还结合了几项专门定义的任务来训练谣言检测神经网络。 |
| GNN | BiGCN[5] | 该方法代表了图神经网络在谣言检测中模拟原始推文双向传播的初步应用。 |
| | GACL[13] | 它包括为每个实例生成对抗训练样本,并在预训练的对抗判别器基础上采用监督学习,以动态区分谣言和非谣言。 |

注:Z-Cot 指 Zero-shot chain-of-Thought(零样本思维链);LeRuD 指 LLM-empowered Rumor Detection(大模型赋能谣言检测);ARG 指 Adaptive Rationable Guidance(自适应基本原理导引);BiGCN 指 Bi-Directional Graph Convolutional Networks(双向图卷积网络);GACL 指 Graph Adversarial Contrastive Learning(图对抗对比学习)。

**表5** 不同方法在 Weibo、Twitter15 和 Twitter16 数据集上的性能对比

**Table 5** Performance comparison of different methods on weibo, Twitter15, and Twitter16 datasets

| 方法 | Weibo | | | | Twitter15 | | | | Twitter16 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | $P$ | $R$ | $F1$ | Acc | $P$ | $R$ | $F1$ | Acc | $P$ | $R$ | $F1$ |
| Zero-Shot | 0.751 | 0.693 | 0.892 | 0.780 | 0.622 | 0.615 | 0.653 | 0.633 | 0.596 | 0.587 | 0.639 | 0.612 |
| Few-Shot | 0.797 | 0.749 | 0.886 | 0.812 | 0.638 | 0.654 | 0.587 | 0.618 | 0.638 | 0.656 | 0.576 | 0.613 |
| Z-CoT | 0.699 | 0.665 | 0.811 | 0.730 | 0.592 | 0.638 | 0.418 | 0.506 | 0.607 | 0.644 | 0.468 | 0.542 |
| LeRuD | 0.800 | 0.734 | **0.938** | 0.821 | 0.611 | 0.685 | 0.405 | 0.509 | 0.634 | 0.711 | 0.444 | 0.546 |
| ARG | 0.904 | 0.902 | 0.909 | 0.906 | 0.880 | 0.884 | 0.891 | 0.888 | 0.929 | 0.911 | 0.927 | 0.919 |
| BiGCN | 0.919 | 0.919 | 0.923 | 0.921 | 0.843 | 0.848 | 0.835 | 0.854 | 0.859 | 0.879 | 0.823 | 0.842 |
| GACL | 0.915 | 0.900 | 0.930 | 0.915 | 0.913 | 0.903 | 0.936 | 0.919 | 0.921 | **0.969** | 0.886 | 0.925 |
| ours | **0.929** | **0.928** | 0.929 | **0.928** | **0.954** | **0.953** | **0.954** | **0.954** | **0.959** | 0.956 | **0.962** | **0.958** |

注:加粗表示同列指标中最优结果。

### 3.7 结果分析

通过与多种基准方法的对比实验,本文提出的方法展现出其在虚假信息检测任务中的优势。首先,与其他使用大语言模型的方法相比,本方法展现出了强大的性能。采用大语言模型直接进行判断并得出结果的方法(如Z-CoT、LeRuD)的实验结果并不理想,原因在于大语言模型没有发挥其多角度分析能力,容易受限于单一的推理路径,难以捕捉虚假信息的复杂特征。这使得这些方法在实际应用中可能无法直接对真伪信息进行准确区分,尤其是在信息源复杂、情境多变的网络社交媒体平台上。相较之下,ARG方法仅仅生成了常识和文本线索两个方面的解释,并不具备足够的多样性和全面性,容易忽略传播结构中的细节或复杂的隐含内容,故其结果与本文提出的方法存在一定差距。本方法通过设计一系列多角度的提示指令,充分挖掘大语言模型的可解释性能

力,从多个维度生成丰富的合理性解释,提高了模型的可靠性和透明度,减少了“幻觉”现象的发生,从而增强了判断的准确性和可靠性,具体案例分析详细附录B。

此外,与一些使用图神经网络的方法(如BiGCN、GACL)相比,本方法仍具有较高的性能。图神经网络方法基于封闭世界假设进行训练,难以应对真实世界复杂的情境变化,且由于其“黑箱”特性,缺乏可解释性。而本方法通过嵌入大语言模型的强大语言理解能力,在较低计算成本的情况下仍能有效地捕捉虚假信息的多层次特征,实现更加高效和准确的检测,展现了其在实际应用中广泛推广的潜力。

### 3.8 消融实验

为评估模型中不同组成部分的贡献和实验方法的有效性,本节通过消融实验研究对模型进行深入分析。本研究分别对会话链语义表示和合理性解释语义表示进行单独训练,并评估其在虚假信息检测任务中的性能。实验结果如表6所示。

**表6** 消融实验结果

**Table 6** Ablation Results

| 方法 | Weibo | | | | Twitter15 | | | | Twitter16 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | *P* | *R* | *F*1 | Acc | *P* | *R* | *F*1 | Acc | *P* | *R* | *F*1 |
| Full version | **0.929** | **0.928** | **0.929** | **0.928** | **0.954** | **0.953** | **0.954** | **0.954** | **0.959** | **0.956** | **0.962** | **0.958** |
| w/o $\boldsymbol{Embedding}_{\text{semantic}}$ | 0.907 | 0.895 | 0.922 | 0.908 | 0.943 | 0.939 | 0.948 | 0.943 | 0.937 | 0.929 | 0.947 | 0.937 |
| w/o $\boldsymbol{Embedding}_{\text{explaination}}$ | 0.867 | 0.868 | 0.862 | 0.865 | 0.946 | 0.951 | 0.941 | 0.945 | 0.956 | 0.952 | 0.962 | 0.956 |

注:实验设置包括三种模型:Full Version(完整模型,同时使用两类语义表示),w/o $\boldsymbol{Embedding}_{\text{semantic}}$(仅使用解释语义表示)、以及w/o $\boldsymbol{Embedding}_{\text{explaination}}$(仅使用会话链语义表示)。加粗表示同列指标中最优结果。

实验结果表明,融合使用两种语义表示的完整模型在所有数据集上均取得最优性能,验证了语义互补带来的显著性能增益。在Weibo数据集上,完整模型的*F*1分数为0.928,相较于仅使用 $\boldsymbol{Embedding}_{\text{explaination}}$(0.908)或 $\boldsymbol{Embedding}_{\text{semantic}}$(0.865),分别提升了约2.0%和6.3%。在Twitter15和Twitter16数据集中,完整模型*F*1分数分别为0.954和0.958,均优于任一单独特征表示的模型,最高提升幅度达2.2%。值得注意的是,在Twitter16数据集中w/o $\boldsymbol{Embedding}_{\text{explaination}}$的表现(*F*1=0.956)已非常接近完整模型,说明该数据集中会话链语义表示具有更高的信息密度;而Weibo数据集中,$\boldsymbol{Embedding}_{\text{explaination}}$对性能的提升最为显著,表明中文语境下解释性语义对虚假信息的识别尤为关键。总体来看,消融实验定量验证了两个语义模块在不同语境下的独立有效性与互补性,同时也凸显了多角度提示生成机制在提升虚假信息检测可解释性和判别力方面的重要价值。

## 4 结论

本文提出了一种基于多维度分析的可解释性大语言模型框架,用于社交媒体平台上的虚假信息检测任务。该方法通过多维度提示指令设计,引导大语言模型从煽动性语言、事实一致性、逻辑矛盾等多个维度对会话链传播结构进行深入分析与推理,从而生成详细的合理性解释和判断依据。此外,本研究利用大语言模型的特征提取能力提取出数据的会话链语义嵌入和合理性解释语义表示,使用多层感知机进行分类训练,实现了高效且透明的虚假信息检测。实验结果表明,本方法在Weibo、Twitter15和Twitter16三个公开数据集上取得了优异的表现,特别是在*F*1分数方面,本方法分别达到了0.928、0.954和0.958,展现了强大的检测能力和鲁棒性。同时,消融实验进一步验证了不同语义表示在提升模型性能中的重要作用,证明了多角度特征融合的有效性。

尽管本研究取得了一定成果,但仍存在一些潜在的研究方向值得进一步探索。未来工作还将致力于扩展本方法的应用场景,例如跨语言虚假信息检测和多模态数据处理,为构建更加智能、可靠的虚假信息检测系统提供支持。

**参考文献:**

[1] FARIS R, ROBERTS H, ETLING B, *et al*. Partisanship, Propaganda, and Disinformation: Online Media and the

2016 US Presidential Election[R/OL]. Cambridge: Berkman Klein Center Research Publication, 2017: 6. https://dash.harvard.edu/handle/1/33759251.

[2] ZOU W X, TANG L. What Do we Believe In? Rumors and Processing Strategies During the COVID-19 Outbreak in China[J]. *Public Underst Sci*, 2021, **30**(2): 153−168. DOI: 10.1177/0963662520979459.

[3] DONG X S, QIAN L J. Semi-supervised Bidirectional RNN for Misinformation Detection[J]. *Mach Learn Appl*, 2022, **10**: 100428. DOI: 10.1016/j.mlwa.2022.100428.

[4] CHEN Y X, SUI J, HU L, *et al*. Attention-residual Network with CNN for Rumor Detection[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Barcelona: ACM, 2019: 1121−1130. DOI: 10.1145/3357384.3357950.

[5] BIAN T, XIAO X, XU T Y, *et al*. Rumor Detection on Social Media with Bi-directional Graph Convolutional Networks[J]. *Proc AAAI Conf Artif Intell*, 2020, **34**(1): 549−556. DOI: 10.1609/aaai.v34i01.5393.

[6] MA J, GAO W. Debunking Rumors on Twitter with Tree Transformer[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 5455−5466. DOI: 10.18653/v1/2020. coling-main.476.

[7] BANG Y, CAHYAWIJAYA S, LEE N, *et al*. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity[EB/OL]. (2023−02−08)[2026−01−20]. 2023. https://arXiv.org/abs/2302.04023.

[8] TONMOY S M, ZAMAN S M, JAIN V, *et al*. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models[EB/OL]. (2024−01−02)[2026−01−20]. https://arXiv.org/abs/2401.01313.

[9] LUCAS J, UCHENDU A, YAMASHITA M, *et al*. Fighting Fire with Fire: The Dual Role of LLMS in Crafting and Detecting Elusive Disinformation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2023: 14279−14305. DOI: 10.18653/v1/2023.emnlp-main.883.

[10] ZHOU X Y, ZAFARANI R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities[J]. *Acm Comput Surv*, 2020, **53**(5). DOI: 10.1145/3395046.

[11] SUN T N, QIAN Z, DONG S J, *et al*. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning[C]//Proceedings of the ACM Web Conference 2022. New York: ACM, 2022: 2789−2797. DOI: 10.1145/3485447.3511999.

[12] LOYOLA-GONZÁLEZ O. Black-box Vs. White-box: Understanding Their Advantages and Weaknesses from a Practical Point of View[J]. *IEEE Access*, 2019, **7**: 154096−154113. DOI: 10.1109/ACCESS.2019.2949286.

[13] CAO Y, NAIR A M, EYIMIFE E, *et al*. Can Large Language Models Detect Misinformation in Scientific News Reporting [EB/OL]. (2024−02−22)[2026−01−20]. https://arXiv.org/abs/2402.14268.

[14] HUANG Y, SUN L. FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models [EB/OL]. (2023−10−08) [2026−01−20]. https://arXiv.org/abs/2310.05046.

[15] ZHANG M, GONG H, Liu Q, *et al*. Breaking Event Rumor Detection via Stance-separated Multi-agent Debate [EB/OL]. (2024−12−06) [2026−01−20]. https://arXiv.org/abs/2412.04859.

[16] WU G Y, WU W J, LIU X H, *et al*. Cheap-fake Detection with LLM Using Prompt Engineering[C]//2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). New York: IEEE, 2023: 105−109. DOI: 10.1109/ICMEW59549.2023.00025.

[17] WAN H R, FENG S B, TAN Z X, *et al*. DELL: Generating Reactions and Explanations for LLM-based Misinformation Detection[C]//Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg, PA, USA: ACL, 2024: 2637−2667. DOI: 10.18653/v1/2024.findings-acl.155.

[18] CHEN C Y, SHU K. Combating Misinformation in the Age of LLMS: Opportunities and Challenges[J]. *AI Mag*, 2024, **45**(3): 354−368. DOI: 10.1002/aaai.12188.

[19] CHEN J Y, LIU L Y, ZHOU F. Do Not Wait: Preemptive Rumor Detection with Cooperative LLMS and Accessible Social Context[J]. *Inf Process Manag*, 2025, **62**(3): 103995. DOI: 10.1016/j.ipm.2024.103995.

[20] ZHONG N J, JIANG X C, YAO Y. From Detection to Explanation: Integrating Temporal and Spatial Features for Rumor Detection and Explaining Results Using LLMS[J]. *Comput Mater Continua*, 2025, **82**(3): 4741−4757. DOI: 10.32604/cmc.2025.059536.

[21] WANG B, MA J, LIN H Z, *et al*. Explainable Fake News Detection with Large Language Model *via* Defense among Competing Wisdom[C]//Proceedings of the ACM Web Conference 2024. New York: ACM, 2024: 2452−2463. DOI: 10.1145/3589334.3645471.

[22] LAI J Q, YANG X R, LUO W Y, *et al*. RumorLLM: A

Rumor Large Language Model-based Fake-news-detection Data-augmentation Approach[J]. *Applied Sci*, 2024, **14**(8): 3532. DOI: 10.3390/app14083532.

[23] YANG R C, GAO W, MA J, *et al*. Reinforcement Tuning for Detecting Stances and Debunking Rumors Jointly with Large Language Models[C]//Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg, PA, USA: ACL, 2024: 13423-13439. DOI: 10.18653/v1/2024.findings-acl.796.

[24] AJWANI R, JAVAJI S R, RUDZICZ F, *et al*. LLM-generated Black-box Explanations can be Adversarially Helpful[EB/OL]. (2024-05-10) [2026-01-20]. https://arXiv.org/abs/2405.06800.

[25] YUE Z R, ZENG H M, ZHANG Y, *et al*. MetaAdapt: Domain Adaptive Few-shot Misinformation Detection via Meta Learning[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: ACL, 2023: 5223-5239. DOI: 10.18653/v1/2023.acl-long.286.

[26] QIN C W, ZHANG A, ZHANG Z S, *et al*. Is ChatGPT a General-purpose Natural Language Processing Task Solver? [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2023: 1339-1384. DOI: 10.18653/v1/2023.emnlp-main.85.

[27] MA J, GAO W, WONG K F. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: ACL, 2018: 1980-1989. DOI: 10.18653/v1/p18-1184.

[28] MA J, GAO W, WONG K F. Detect Rumors in Microblog Posts Using Propagation Structure viaKernel Learning[C]//Proceedings of the 55th Annual Meeting of the Association forComputational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: ACL, 2017: 708-717. DOI: 10.18653/v1/p17-1066.

[29] LIN H Z, YI P Y, MA J, *et al*. Zero-shot Rumor Detection with Propagation Structure *via* Prompt Learning[J]. *Proc AAAI Conf Artif Intell*, 2023, **37**(4): 5213-5221. DOI: 10.1609/aaai.v37i4.25651.

[30] HU B Z, SHENG Q, CAO J, *et al*. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection[J]. *Proc AAAI Conf Artif Intell*, 2024, **38**(20): 22105-22113. DOI: 10.1609/aaai.v38i20.30214.

## 附录 A

本研究针对多层感知机(MLP)分类器的关键训练超参数进行了系统性的自动化调优。具体而言,我们采用 Optuna 提供的贝叶斯优化框架对以下五个超参数进行搜索:隐藏层维度(hidden_dim)、dropout 概率(dropout_prob)、训练轮数(num_epochs)、批量大小(batch_size)以及学习率(learning_rate)。

超参数优化过程中,以 $F1$ 分数作为优化目标,并针对每个数据集(Weibo、Twitter15 和 Twitter16)分别独立执行搜索过程。最终获得的最优超参数组合如表 A1 所示。

**表 A1** 各数据集对应的最优超参数配置

**Table A1** Optimal Hyperparameter Settings for Each Dataset

| 超参数 | Weibo | Twitter15 | Twitter16 |
|---|---|---|---|
| hidden_dim | 512 | 512 | 512 |
| dropout_prob | 0.7 | 0.3 | 0.3 |
| num_epochs | 60 | 50 | 60 |
| batch_size | 32 | 32 | 32 |
| learning_rate | 0.000 1 | 0.001 | 0.001 |

## 附录 B

本节将展示一个具体的案例并进行分析,以使读者可以清晰地看到本文提出的多维度提示指令如何指导 LLMs 从不同角度评估一条帖子的真实性和可信度,并最终生成相应的解释性回复。详细案例如表 B1 所示。

**表B1** 案例分析:多维度提示指令对榴莲可乐谣言的分析示例

**Table B1** Case Analysis: Example of Multi-Angle Prompt Instructions on the Durian-Coke Rumor

| 类型 | 文本内容 |
| --- | --- |
| 原帖 | 吃榴莲后,喝可口可乐,毒过眼镜蛇!又一游客,客死泰国异乡!一位中国游客在泰国旅游的时候,吃了很多榴莲之后又喝了可乐,致咖啡因中毒血压飙升,结果引发心脏病猝死,年仅28岁,泰国有明确规定,食用大量榴莲后,8小时之内不能喝可口可乐!圈里爱榴莲的朋友很多,转发提醒一下!让所有人都知道! |
| 解释性回复 | 1. 该帖子内容是否违背生活常识?<br>该帖子内容违背了基本的生活常识。咖啡因中毒通常是由于摄入过量的咖啡或含咖啡因的饮料,而不是因为食用榴莲后立即饮用可乐。榴莲与可乐同时食用是否会导致严重的健康问题,并没有确凿的科学证据。<br>2. 该帖子是否提供了可靠的来源或证据?<br>帖子没有提供可解释性<br>回复靠的来源或证据。帖子中没有提及任何科学研究的引用,也没有提到官方健康机构的声明。<br>3. 帖子内容是否有夸大或不实的表述?<br>是的,帖子中存在夸大和不实的表述。将榴莲和可乐同时食用的风险描述为“毒过眼镜蛇”,并且提到会导致“咖啡因中毒血压飙升,引发心脏病猝死”,这些表述缺乏科学依据。<br>4. 评论中是否有专业人士或权威机构的反驳或澄清?<br>评论中并没有出现专业人士或权威机构的反驳或澄清,大部分评论是表达惊讶、担忧或转发提醒。<br>5. 该帖子是否涉及敏感话题,容易引起恐慌或误解?<br>是的,该帖子涉及了健康和旅行安全等敏感话题,可能会引起一些读者的恐慌或误解。<br>6. 评论区的互动是否理性,有没有人质疑该内容的真实性?<br>评论区的互动并不完全理性,有些评论表达了担忧和恐慌,但也有一部分评论者对内容的真实性提出了质疑,要求求证。<br>7. 该帖子是否有明显的情绪煽动性语言?<br>是的,帖子中使用了一些情绪煽动性的语言,如“毒过眼镜蛇”“客死泰国异乡”等,这些语言可能会激发读者的情感反应。<br>8. 该帖子内容是否与已知的真实事件相符?<br>帖子描述的事件(游客因食用榴莲后饮用可乐而猝死)可能并不符合已知的真实事件,其缺乏可靠的证据。<br>9. 该帖子是否利用恐惧、愤怒等负面情绪来吸引点击或关注?<br>是的,该帖子通过描述一个可能引起恐惧的情景来吸引点击和关注。<br>10. 该帖子是否存在信息误导的可能性?<br>是的,该帖子存在信息误导的可能性,因为它没有提供科学依据,并且使用了夸大和不实的表述。 |